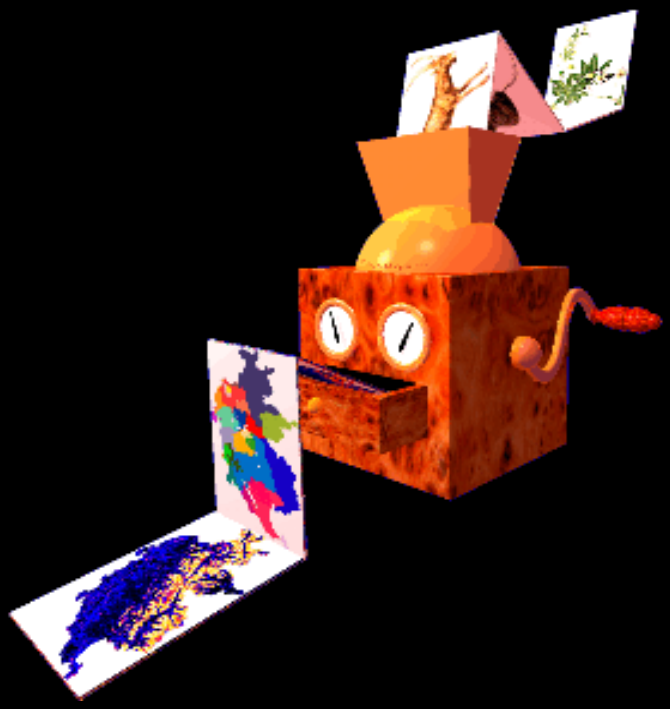
Frequently Asked Questions

FAQ version: 02.12.2005



How to use this FAQ

This file collects the common questions Biomapperians are asking before, during and after use of Biomapper. The questions and their answer are classed by order of appearance in a classical Biomapper analysis.That means that if this is the first time you use Biomapper, you may as well read this document sequentially. For veteran Biomapperians, this structure will still help you finding a particular answer; just think about it "chronologically". A few entries are referenced in two sections; however, the answers themselves occur only once and have been placed in the most relevant section.

- You may also want to use the "search" or "find" function of your browser to look for key-words.
- Along this document I am using the following abbreviations:

EGV	=	EcoGeographical variable (= environmental descriptor)
ENFA	=	Ecological Niche Factor Analysis
GLM	=	Generalised Linear Models
HS	=	Habitat Suitability

Biomapper

Before using Biomapper: Should I? Can I?

How much costs Biomapper?

Biomapper is a postcardware, meaning it is free to try it (download it), but you must eventually send me a postcard.

Does Biomapper work with Idrisi32?

Yes, Biomapper can read both Idrisi16 and Idrisi32 file format.

I'm working with Arcview. How can I use Biomapper?

Several of Biomapper's users are using Arcview. You will have to prepare your maps in Arcview before converting them to Idrisi/Biomapper format. Once the whole process has been completed, you can reimport the resulting maps into Arcview for further analysis or display.

If you are using Arcview 3.0, you can use the Biomapper module Grid Convertor.

In which programming language is Biomapper developped?

I'm using Borland Delphi for all my programming work. It allows me to quickly conceive fast running procedures wrapped in a neat user-friendly interface. The source code is not open-source but you will find the crucial procedures in the annexes of my PhD thesis.

On which platform does Biomapper run?

Biomapper is developed for Windows32 platforms, i.e. on all Windows since Win95. I have personally tested it on Win95, Win98, Win NT4, Win2000 and WinXP. I have been told that it worked well on a Mac with a PC-card, but there is no version specially developed for Mac, Linux, Unix or any other platforms.

Is Biomapper better suitable for plants or animals modelling?

So far, Biomapper has been applied on a wide range of the life tree: Mammals, Birds, Insects, tropical Trees and Ferns. Most of them are animals but it's mainly because I use to meet more zoologists than botanists...
In fact, absence data tend to be poorer for animals and so Biomapper could be particularly adapted to this case.

Where to begin? How to get help?

I would like to look for any word in the Biomapper help file. How can I do it?

The Windows help-file system allows you to generate a word search from any help file. Open the help file and click on the "Find" tab (rightmost tab). The first time you do this, you can choose the size of the database you want to use; generally I use the smallest size, but if you want more words to be included in the database, you can select a larger size. Then, the "Find" tab will provide you with the kind of tool you are looking for. Just type a word and the you will see a list of the

"chapters" containing this word.

Some checkboxes, buttons and options are not documented in the help file.

The help file is actually written along the thread of the modus operandi (or step-by-step guide) to guide the user through Biomapper to allow her/him to compute and validate a HS map easily and understand the main outputs. I agree there are a lot of undocumented small controls (checkboxes, buttons, etc.) I added for the comfort of the user. Whilst only a few are described in the help file or in this FAQ on the website, all of them have a built-in help message that appears when you hover the mouse cursor on the control; the help message - or "mouse hint" - appears both in the bottom bar of Biomapper main window and in a floating temporary panel; they are present for most controls in all modules. These hints try to explain what's the use of the control. When a more complete explanation was needed, I generally put it either in the help file or in the FAQ. If I did not, please send me an e-mail and I shall clarify the situation.

I found inconsistencies and undocumented options between the software and the help file.

There are some minor inconsistencies between the help file and the program as Biomapper is evolving continuously and I don't modify the help file each time. I hope you understand that I am a scientific researcher and not a professional programmer. My main interest is in solving ecological questions by developing more efficient algorithms and applying them to real cases. Biomapper is free and I'm alone to program it, therefore, I cannot devote - and am not interested in devoting - too much time and energy in help-file writing and updating work. So far, Biomapper development is a pleasure; I try to keep it like that. Should it become a pain, I would probably have to find a way to get some money-compensation.

However, I am always happy to answer the questions of the Biomapperians and do my best to fulfil their needs. All their questions and my answers are filed in this FAQ, which is therefore a good knowledge-base. This FAQ is updated regularly, each time I answer a new question, in fact. The help-file is more painful and time-consuming to update, and so I update it only about once a year, if there is enough matter to add and modify.

Do you have a step-by-step explanation of how doing a whole habitat suitability analysis?

Yes. You will find it in the menu Help/Modus operandi. The whole process is outlined there and by clicking on the main titles you access a more detailed step-by-step description of the process.

Could you kindly send me the user manual.

There is a user manual included with the Biomapper package. This is actually a PDF version of the help file. The FAQ (www.unil.ch/biomapper/faq.html) is also a good source of information.

Where else can I get support for Biomapper?

The main ressources are:

- The *Modus Operandi* (found in Biomapper's help menu)
- This FAQ
- The Biomapper-List discussion group on Yahoo: http://groups.yahoo.com/group/Biomapper-List/ where experienced Biomapperians (and myself) will answer your questions.
- Your last hope is to e-mail me.

Do you have an example data set to test Biomapper?

Alas, none of our data are in the public domain and we cannot give them away.

Biomapper miscellanea

How can I print a map from Biomapper? I guess I can do it from Idrisi but I´d like to use the rainbow palette.

Display the map and save it as a BMP file. This file you can insert in a word document or print with any picture software. Alternatively, you will find the rainbow palette in the Biomapper package: Biomapper.smp.

Error messages

I get an error message "... is not a valid floating point value"

It happens because your computer is not set to use "." as decimal separator. It is mandatory if you want to use Biomapper and greatly advised if you want to use Idrisi. You can change this setting in the Start Menu/parameters/Control panel/Regional settings/Numbers.

When I try to open a project, I get a message "Documentation file not found".

This message means that Biomapper could not find a map. This happen generally because the map is not in the same directory as the project file (*.bio). All EGV maps must be in the same directory. This constraint was introduced in autumn 2000 in order to make the transfer of project between different computers transparent. By now, when you create a new project, Biomapper will verify that all your files are in the same directory. However, you could have problem with project created with previous versions.

I get error messages when I add EGV maps, or try to normalise them, or mask them, etc.

This kind of problems generally happen when the metadata file is incorrect.

- If you are working with ArcView grids that you have imported into Idrisi format, your maps probably suffer from the infamous -9999 background bug. Check in the metadata file (*.rdc) what's the flag value: if it is -9999, this bug is probably at work. There is a tool in the MapManager module that will help you to fix this problem.
- If this does not solve your problmen, check the metadata file (*.rdc. You can open it in the NotePad, for example). Check for inconsistancies between your EGV maps and species maps, in particular in the following fields: columns, rows, Max X, Min X, Max Y, Min Y and ref. system.

If this doesn't help, please contact me.

Map preparation

Collecting data, sampling design

What is the minimum amount of presence points needed to get a good HS map?

Mmmh... This is a question I'm often asked. Alas I have no easy answer. It depends on the variance of the study area, the specialisation of the focal species, the kind and accuracy of the sampling design, etc...

I have generally used several hundreds of points, but I found that I could as well have used far less without decreasing significantly the accuracy of the model.

If you have a carefully designed sampling work, covering all possible habitats, I guess that the number of points needed would be minimal. Perhaps even as few as 20 or 30 points.

File format: Conversion / importation

How to convert an ESRI grid into an Idrisi/Biomapper raster map and conversely?

There are several possibilities:

- *ArcView 3.x extension*
  There is an extension doing the job on the ESRI site, made by Holger Schäuble. Look for Grid Converter (av2idrisi.zip) on http://arcscripts.esri.com/. It works for ArcView 3.x.
- *Biomapper's GridConvertor*
  If you have Arcview 3.0 or 3.1 and Spatial Analyst, you can use the Biomapper module *Grid Convertor*. It allows to convert several files in only one operation.
- *Manual conversion*
  If you have another version of ArcView, GridConvertor may not work (always problems due to ESRI proprietary policy).You will therefore have to convert your grids manually. In the Biomapper help file you will find a description of the Idrisi/Biomapper file format that should help you to transfer your data.You can try this: in Arcview, go into the file menu and select Export Data Source and export your image as a binary raster (has an flt extension). Exit Arcview and then change the file extension from .flt to .rst. You should then be able to create a document file (*.rdc) for this image with the information supplied in the help file. The flt and rst format are identical. You will probably have to choose "real" data type.
- *Otherwise...*
  You can also find useful programs at http://www.pierssen.com/idrisi/grid.htm

What are the first steps for an ArcView user who want to get into Biomapper?

By Tanya Leverette:

I will assume you have the Spatial Analyst extension for AV. First, you want to convert your shapefiles into grids (under Theme>convert to grid). Do this for both the EGVs and the species data. Once that is done, you can use the Av2Idrisi script that can be found in the script section of www.esri.com to convert to .rst format. It will appear in Theme>Convert to Idrisi. Remember to save those new files all together in a folder that you will be accessing from Biomapper. For

Biomapper to work happily, all your newly made Idrisi files need to be in one folder.

Then, things should move smoothly from there. In Biomapper, first go into Map Manager (under File) and convert your Idrisi species map (.rst) to Boolean (click on the Data Type Management tab). Then it is just a matter of adding the EGV maps and the species work map. The Modus Operandi can lead you from here.

Error while converting maps from Biomapper/Idrisi to ArcView 3.x using the AV2Idrisi extension. What is going wrong?

A problem I've found when converting Idrisi files to AV3.2 is caused by the "comments" created by Biomapper in the .rdc file: Open the .rdc in Idrisi or in Notepad or any other text editor, and erase all the lines "comment" and "lineage". Sometimes, there is also problems in the "resolution" information: if it is "unknown", give it the correct information (pixel size). Also, in AV, grids are stored into folders whose names must be at most 12 characters (some special characters are not allowed). (Advice kindly provided by Gwenaëlle Le Lay)

Another problem might be caused by the name of the directory in which your file resides: ArcView might be crashing / having problems if you're using any spaces in your folder, especially when working with grids. I'm using AV 3.2 under Win2000 and I don't have any problems with long file names. (Advice kindly provided by Jakob Fahr)

Why is it not possible to use most of the modules of Biomapper with Idrisi16 (.img) files?

Only the main Biomapper program has a button to switch from one format to the other. The other module use one of them according to the folder in which they are placed (it is not very elegant, I agree, and I will have to modify that some time in the future).

When the module is launched, it checks if there is an "idrisi.env" file in the same folder as the module executable. If there is one, it switches to the Idrisi16 mode, else it switch to Idrisi32 mode.

You can therefore enforce one mode or the other by placing a fake "idrisi.env" in your Biomapper folder.

How can I use satellite imagery pictures (or other file format) with Biomapper?

Biomapper has no importing capabilities. It can build a map from raw data but cannot convert alien file formats. It works only with the Idrisi (16 and 32) file format but Idrisi itself has an extensive set of conversion tools.

I get an "Out of memory error" in one of the auxiliary modules. What should I do?

- For some spatial analyses (e.g. DistAn and BigGroup) the module has to maintain entire maps into memory (in contrast with ENFA analyses where the maps are scanned line by line) and this may cause out of memory errors. I can see several solutions to this problem:
- Close all other programs when you run DistAn so as to allot it all the available memory.
- Increase the virtual memory of your computer. This can be done by clicking your way to Start/Settings/System/Advanced/Performance settings/Advanced/Change/Paging File Size (on Windows XP. You must have administrator access to your computer. Ask an administrator if you are not familiar with this kind of things)
- Use the distance module of a GIS software (Idrisi, ArcView, GRASS, etc.)
- Decrease the resolution of your maps, i.e. increase the size of the map cells. You will need Idrisi

to do that, through the menu Reformat/CONTRACT. This can probably also be done in ArcView, but I don't know the procedure.

After computing the ENFA, Biomapper complains that some eigenvalue is too large, or negative. What does it mean?

After the ENFA, the eigenvalues are the first thing to check: they must all be greater than 0. It may happen that one or more are negative (check the last eigenvalues as they are sorted by decreasing order), or one of them is very huge. This means that either the global or the specie correlation matrix was nearly singular and that the inversion algorithm produced absurd results. A matrix is singular is a matrix where one of the columns (or rows) is a linear function of one or several other, or where one of the columns is filled with zeros. Generally this is the latter happens in our case.

It means that two or more EGV maps are too highly correlated and one of them must be removed. To decide which map must to remove, examine the correlation tree (View/Correlation tree) to see which are the most correlated maps. Remove one of the most redundant pair (Right click on the map then Remove) and launch the ENFA again (as the covariance matrix is already computed, the process will be far shorter). Repeat these check-remove-compute operations until all the eigenvalues are null or positive. Don't fear a loss of information: the removed maps contain mostly redundant data.

I get the error message "Module cannot be found or is corrupt". What does it mean?

This message means that Biomapper can't find one of the module that are shipped with Biomapper (e.g. MapManager.exe, CircAn.exe, etc.). It is probably because it is not in the same directory as Biomapper.exe. These modules come with the Biomapper package and should be unzipped in the same directory.

Data structure: Raster / Vector

How to convert a point vector map into a species map?

In Idrisi32:

- Menu Reformat/Raster-vector conversion/POINTRAS
- Select your vector map, e.g. "Species.vct"
- Choose a name for the "image file to be updated". It must be a new name (not an existing raster) (e.g. "Species_bl.rst")
- Choose "Change cells to record the presence of 1 or more points"

Idrisi asks you if you want to bring up INITIAL: Answer YES.

- In the INITIAL dialog box:
- Select "Copy spatial parmeters..."
- Select one of your EGV maps in "Image to copy parameters from"
- In "Output data type", select "Byte"
- Click OK

And this should do the job. Species_bl can now be used as species map. You may want to partition it into calibration and validation data sets. You can do this with the Biomapper module Sampler.

Rememember that all the EGV maps and the species map must be in the same directory.

How to convert a polygon vector map into a species map?

In Idrisi32:

- Menu Reformat/Raster-vector conversion/POLYRAS
- Select your vector map, e.g. "Species.vct"
- Choose a name for the "image file to be updated". It must be a new name (not an existing raster) (e.g. "Species.rst")

Idrisi asks you if you want to bring up INITIAL: Answer YES.

- In the INITIAL dialog box:
- Select "Copy spatial parmeters..."
- Select one of your EGV maps in "Image to copy parameters from"
- In "Output data type", select "Byte"
- Click OK

Then, you must booleanise the rasterised polygons:

- Menu Analysis/Database query/image calculator
- Select "Logical expression"
- Type Species_BL = [Species]>0
- Click OK

Species_BL can now be used as species map. You may want to partition it into calibration and validation data sets. You can do this with the Biomapper module Sampler.

Rememember that all the EGV maps and the species map must be in the same directory.

File format: miscellanea

What are these "Biomapper extensions" used for? Can I ignore them?

These extensions are made to simplify the browsing and help the user to select among all the maps, those having the right data type. But these extensions are not used by Biomapper to verify the maps; it uses the raster documentation file (*.rdc). Thus you can as well ignore the biomapper extensions.

The species map

How can I create the species-presence map?

There are several possibilities depending on what kind of data you have at hand:

- List of observations coordinates: Put them in an ASCII file, using a structure x-coordinate tabulation y-coordinate (You can do this with Microsoft Excel) and the use the "Convertor" module to create a boolean presence map from this file.
- Observation map in a point-vector-file: Simply rasterise it (with Idrisi function "PointRas"), using the same resolution and window as your ecogeographical maps.
- Population map in a polygon-vector-file: First, rasterise it (with Idrisi function "PolyRas"), using the same resolution and window as your ecogeographical maps. Then make this map boolean (1=inside populations). Finally, use the module "Sampler" to divide this map into a calibration and a validation data sets.

How to insert the species-presence map?

Once the species-presence map has been created, you must insert it in Biomapper in order to use it

in the analyses. The species-map must be inserted in the Work maps list (NOT the EGV maps). This can be done in the `Files/Work maps/Add maps` menu. Once inserted in this list, you must declare it as the current species map by selecting it, right-clicking on it and selecting "Mark as species map". The current species map is then marked by a red circle in front of its name.

I have very good absence data. Can I use them with Biomapper?

If you are really sure that your absence data are good and that no historical or spatial factors could have biased them, you will probably get a better model by using a presence/absence-based method, like Generalized Linear Model (GLM) or Generalized Additive Model (GAM). But you can ever put your absence data aside and apply ENFA on presence data only.

How to use abundance data with Biomapper?

You can use it to weight the presence data (weighted boolean map). Just use a map with integere weights in place of 0 and 1 as species map.

The monitoring resolution is larger than the EGV resolution I have. What are my options?

Let's say you have two different resolutions in your data: the EGVs have a much finer resolution (say one hectare) than the species monitoring plan (say 10 hectares).

This is a quite common situation. The problem is that you don't know where in the 10 ha the bird has been seen and therefore there is no sense in keeping a too fine spatial resolution. There are several way of dealing with this problem:

A) My favourite one is to keep the finer resolution for all analyses, but to use EGVs that take the lack of monitoring resolution into account. For each 10-ha square, I consider there is only one presence, placed at the centre of the square. All EGVs are then derived from the existing data by using CircAn with a radius equal to the half of the 10-ha square. You can use the average value, maximum, minimum or other statistics, depending on the kind of variable and the species ecology.

B) Another way would be to use the coarser resolution and to convert all EGVs to this resolution, again by averaging (or maximising, minimising, etc.) values on the whole 10-ha square.

C) If you are sure that your species is present in the whole 10-ha square, you can also fill all the square with presences. However, if you are not sure of that, such a method could be misleading and could bias the model. However, I have never tested that. There mighty be a spatial autocorrelation issue too.

EGV maps

Why have the EGV maps to be quantitative?

The ENFA is based on quantitative computations. For example, the marginality factor is based on means. Therefore, if you use a purely categorical map these computations will be misleading. Example: say you have a vegetation map with 1.grassland, 2.forest, 3.agriculture field, 4.bushes, but which is mainly constitute of types 1 and 4. The ENFA will computes that the global mean for this map is about 2.5 (0.5*(4+1)), which at least doesn't mean anything and at worst is strongly misleading (could be taken as forest or agriculture land).

I have a categorical map (say a vegetation type map). How would you handle preparing this map?

I would use BOOLEANISATOR to get a boolean map of each relevant category and then feed these

boolean maps into DISTAN and CIRCAN to get distance and frequency maps of them. You could even choose several radius for the moving window in order to take into account several influence distances.CIRCAN is useful for all resources (food, shelter, etc.) variables, which the species could need in its home range. DISTAN is useful to makes disturbance (mainly human impact like tourism, noise, pollution,etc.) variables. Generally, I compute both distance and frequency maps for all my boolean variables, compute the ENFA and look at the score to see if there is any difference between how the species is sensible to both aspects.

You could also consider to use some fragmentation index according to the species you are mapping, as "Border length" in CIRCAN (It works well for animals/plants living or feeding near forest boundaries, like Ungulates)

In the module DistAn there are several types of distances. What are the differences?

In fact, I never used harmonic and geometric distance maps in the context of building EGV maps. However they could be useful (something to investigate, tell me if you find something interesting).

The practical difference between the various distances is the weight they give to individual observations. There is a gradient from minimal distance, to harmonic then geometric and finally arithmetic mean. Let's take the case of influence of buildings on the habitat of some species of bat:

- Arithmetic mean is probably useless in any situation.
- Use minimal distance if you feel that any single building can be a good habitat for bats, independently of the proximity of other buildings. Note that the minimal distance is the same as Idrisi's module *Distance*.
- Use harmonic mean distance if you feel that any single building can be a good habitat but that the bats prefer to live in high building density areas.
- Use geometric mean distance if you feel that building concentration is more important than single buildings.

How to choose the radius of the moving window in CircAn?

This strongly depend on the ecology of your species. You should try to put your mind into its head and see the world through its eyes. Ask yourself these questions:

- What ecological features are you expecting to influence the choice for an animal to stay somewhere or to leave?
- What is the distance to which these features have a significant effect?

It depends also on the kind of presence data you have. If they are nesting sites, the radius of the frequency window should be taken as large as the area explored by the bird when foraging, or its home range. If they are just casual observations, the radius should probably be larger as the animals are probably exploring sometimes further from optimal habitat. In some instances, an animal can travel long distances just to have access to a punctual resource (water, ungulates like to lick salty rocks, etc.) and thus, this resource should be modeled with a larger radius. Of course, the resolution of your EGV and presence map is also an issue.

When I was modelling the Bearded Vulture habitat, I was confronted to the following problem: its home range was so huge, it just couldn't be used for practical reasons. Therefore I tried two radii: one was approximately its field of vision size (500 m) (as this is what can influence it to soar in some place or to fly away) and another was the size of the area it generally explores when looking for food (2000 m). At the end, we found that the 2000-m radius was a better predictor and we

discarded the 500-m radius. You may want to try something like that.

Why is it not possible to use EGV maps having different scales, or not overlayable?

The ENFA requires to have for every cell of the map a value for each EGV. If the maps are not overlayable, they would have to be made such internally. That would mean to include into Biomapper a complete interpolation algorithm. Although it would be possible, interpolation is a whole world in itself and I could not implement it exhaustively into Biomapper nor keep up-to-date with all the new development in the domain. My strategy is therefore to leave this aspect to the softwares dedicated to this problem. Idrisi for example has a whole geostatistic module allowing to do all kind of interpolations, from the simpler to the most complex ones.

The function of Biomapper is to go where other GIS softwares do not go, not to repeat what they are already doing. There is of course some overlap but I keep it minimal.

Technicalities: Mask, Box-Cox transformation, verify maps, etc.

What is a Box-Cox transformation and why is it needed?

Box and Cox (1964) developed a procedure for estimating the best transformation to normality within the family of power transformations:

$$Y' = (Y^L - 1)/L \qquad \text{(for } L \neq 0\text{)}$$
$$Y' = \ln Y \qquad \text{(for } L = 0\text{)}$$

The numerical process consists to find the L (referred as lambda inside Biomapper) that optimise the normality of the variable distribution.

See "Biometry", Sokal & Rohlf, 1995, pp.417-419 for further explanations.

Biomapper uses the Box-Cox algorithm to normalise as well as possible the ecogeographic variables. Empirically, we have found that normality was not a crucial factor and this step could as well be ignored.

Box-Cox normalisation fails with some EGV maps. Should I discard them?

For myself, when the Box-Cox fails, I keep the original map. A "Box-Coxised" map gives better results than a "brute" map, but a "brute" map is still better than no map at all.

Anyway, you may include it at the beginning (to compute the big time-consuming covariance map. Once it is computed, you can remove easily variables from it, but if you add new variables, the whole matrix will have to be recomputed) and then try to remove it to see how it affects the result.

What is "background"?

For ESRI users, background is the equivalent to *no data*. Let's imagine your study area is a national park with some complex shape. To map it, you have to draw a rectangle containing it completely. There are now two kinds of cells in this map: those inside the study area, for which you have got a lot of information, and those outside about which you know nothing. Inside cells will be "information" and outside cells "background".

Assigning a background / information status to cells is achieved by the process of masking.

What is Masking and what background value to choose?

In Biomapper, masking is not simply assigning a 0 value to background cells. In fact, in some cases, 0 can convey information (e.g. for slope). When you mask a map in Biomapper, you must choose a background value that is 1° outside the range of possible information values and 2° inside the type-permitted values (e.g. byte:0-255). This value will be assigned to all cell outside the boolean mask (0). In the rdc file, the "flag def'n" field will be set to "background" and the "flag value" field to the background value just chosen. Note that you can do this manually by editing the rdc file. Biomapper just make it easier by allowing you to mask several maps at a time.

In Biomapper, it is important that all EGV maps have the same background/information pattern or you will get discrepancies.

Masking (Menu *Ecogeographical Variables/Formatting/Mask...*) will assign a background value to all the cell defined as zero in a boolean map. If the target map already has a background value defined, this value will be used. If not, you will have to choose a background value. You can define a different background value for every map, depending on its range and type. I usually stick to the following policy:

- Byte maps: 255
- Integer categorical maps: -1
- Count, distance maps: -1
- Positive-negative maps: -999

Once a background value has been set, you can forget it. Biomapper will simply ignore all cells having this value in further statistical analyses. Setting a background value doesn't change anything in the map data (in the rst file). Only the rdc file is modified.

If you want to assign a new value to all background cells (e.g., shifting from -99 to -999), use the menu *Ecogeographical Variables/Formatting/Modify background...*)

What are "discrepancies" and how to get rid of them?

Every map cell may have one of two status: either *background* or *information*. A discrepancy arises when one cell has not the same status throughout all the EGV maps. This may appear in several cases:

- One of the map has not been assigned a background (with Mask)
- Some maps were not masked with the same mask

In both cases, fixing the discrepancy when Biomapper proposes you to do it will achieve the correct result.

But there is also another possibility:

- The background value interferes with the information (e.g. the background value is 0 and the information ranges from -1 to 1)

In this case, if you fix the discrepancies with Biomapper, you will lose data in every cell where the information has a value of 0 (false background).

Therefore, I usually do not fix the discrepancies on the first map verification. Cancel the operation and look at the *~discrepancies_CL* map. This map is built by Biomapper during the verification

process and will show you where are located the discrepancies. You will then be able to determine in which of the situations enumerated above you are.

Not getting rid of the discrepancies will make the covariance matrix computation to stop with an error message, possibly after a long computation time and bring you back here. Therefore you MUST get rid of them NOW.

What is the best practice for masking EGV maps?

With maps coming from various sources and having followed different kind of processes, getting rid of all discrepancies can be harduous. Here is a step by step procedure for masking and verifying your maps in a way that minimises the risk to get discrepancies.

1 Put all your maps into Biomapper (*File/EGV/.Add maps...*)
2 Create a mask map (boolean), (0 = outside, 1 = inside study area).
3 Use the *EGV/Formatting/Mask...* menu to mask them. Do this in several steps, grouping EGV maps to which you will assign the same background value (following the above policy). You may want to keep a backup of your original maps, in case you make a mistake.
4 Verify your EGV maps (*File/EGV/Verify maps...)*. If some EGV map was originally "smaller" than the mask, there still could be discrepancies. In this case:
5 Look at the *~discrepancies_CL* map to understand what are these discrepancies. You will perhaps find that you made a mistake when masking (you chose a wrong background value). In that case, take your backuped maps and mask them again and go to step 4.
6 Verify your EGV maps again and now that you are sure that your discrepancies can be discarded, use the "fix" option.
7 Verify them a last time: in some case (typically when working with maps imported from ArcView), some maps are write-protected and could not be remasked by the "fix" option. You will have to unprotect them and redo the step 6.

When I verify the EGV maps, I get a warning message telling "Map "xxxxx" is not continuous enough. What does it mean and how to fix this problem?

This indicates that the EGV maps is either boolean or nearly boolean, i.e. the map contains of almost only two values. This may happen if you took too small a radius in *Circan* with too sparse an original boolean map; sometimes, the Box-Cox algorithm may have this effect too; in this case, just take the map withou Box-Coxing it. Feeding "nearly boolean" maps into the ENFA could cause "negative eigenvalue" or "very large eigenvalue" problems. This warning is just to make you aware of a possible cause of problems. In fact, visually, the EGV-maps should be beautiful rainbows of colors and not almost black and white to get the best results. To fix this problem you have four options:

- Don't change anything and try to perform the ENFA. If you get negative or very large eigenvalues, Biomapper will protest and you will know this is probably due to these maps. But perhaps it will work anyway. It's worth a try.
- Remove the faulty map(s) and perform the ENFA, but, alas! with less information.
- Try to increase the CircAn radius for the faulty map(s).

Use DistAn for these maps.

ENFA

ENFA: Computation

How is computed the score matrix?

The full mathematical details of this operation are described in the main paper in *Ecology*. You can get an intuitive understanding by looking at the help file or on this site at http://www.unil.ch/biomapper/enfa.html. Here is a short view of this process:

The eigenvalues and eigenvectors are extracted as follows: Compute the matrix $\mathbf{W}=\mathbf{Rs}^{-1} * \mathbf{Rg}$ where **Rg** is the global correlation matrix and **Rs** the species covariance matrix. From **W**, extract the marginality factor (I don't give here the mathematical procedure), which gives us the matrix $\mathbf{W}^*$. The specialisation factors are computed by extracting the eigensystem from $\mathbf{W}^*$.

In the Options dialog box, it is possible to switch between correlation and covariance matrix, and to change the norming of the eigenvectors. However, this doesn't seem to affecte ENFA outputs.

These options are intended for the Principal Components Analysis. They do not affect ENFA indeed.

ENFA: Marginality, Specialisation and Tolerance

What are global marginality, specialisation and tolerance?

The global marginality takes into account all the EGVs and gives you a summary of how much the species habitat differs from the available conditions. A low value (close to 0) indicates that the species tends to live in average conditions throughout the study area. A high value (close to 1) indicates a tendency to live in extreme habitats.

The global tolerance does the same with the specialisation factors. A low value (close to 0) indicates a "specialist" species (or stenoecious) tending to live in a very narrow range of conditions. A high value (close to 1) indicates a species that is not too picky on its living environment.

The global specialisation is the inverse of global tolerance, but as it varies between 1 and infinity, it is less easy to interprete.

Note that both values depend highly on the study area. A same species can be highly marginal if the study area has a large extent and include many different regions, but will show almost no marginality if the study area fit closely to its spatial distribution. These values should therefore only be used when comparing several species in a same study area, or to study how a species' ecology evolves with time in a given region.

Mathematically, we have:

Global Marginality = M = Sqrt($\sum_{i=1,v}[M_i^2]$)/1.96

Global Specialisation = S = Sqrt($\sum_{i=1,v}(\lambda_i)/V$)

Global tolerance = $T = 1/S$

where M_i are the coefficients of the marginality factor, Sqrt() is the square root function, V is the number of variables and λ_i are the eigenvalues.

Can I compare the global marginality and specialisation coefficients of different species if I use the same area but different set of ecogeographical maps, in particular if have had to discard different correlated maps in the process?

Strictly speaking, you cannot. Practically, the main biasing effect is the study area. When you have to remove a variable, it is because it does not contain more information than is already included into the model, so removing it should not alter significantly the global marginality and specialisation coefficients. Thus, provided the map sets do not differ drastically, you can still comparer them by these statistics. Anyway, do not assume too much significance to small differences in marginality or specialisation between species

To be true, I never tested this. You could test it by building a common minimal EGV set and apply it to all your species. You will then see how the marginality et specialisation differ between the common data-set and the species-optimal one. Tell what you get, should you decide to try this.

You say that sometimes the marginality factor takes also a part of the specialisation into account. Where can I find how much?

The amount of variance explained by the first factor is in fact the amount of specialisation. It generally ranges from 10% to 70%. This value is given in the eigenvalue table.

To summarise, the marginality factor explains always 100% of the marginality and some part of the specialisation. It is why it has always a great weight (minimum 0.5) for HS computation.

ENFA: Result interpretation

How to interpret factor biological meaning?

Look at the scores matrix. The first column of this matrix is the marginality factor. The other columns are the V-1 specialisation factors. (V is the number of variables). The rows are the EGV contributions to each factor. Look in the published papers to see some examples.

Two tools allow you to display the score matrix in a way making interpretation easier. You can access them through the menu *Multivariate Analysis/Factor computation/Sort scores* and *Multivariate Analysis/Factor computation/Score table*

How is the score table built?

Let C be some specialisation coefficient. The number of stars is computed by N=Round(Abs(C)*10) (Round=rounding to the closest integer, Abs=absolute value). This is the same for the + and - of the marginality factor.

Variable selection in Biomapper.

There is no such thing as variable selection in Biomapper. At least, not in the sense of regression or stepwise based methods. Variable selection is here replaced by factor interpretation.

The ENFA gives you two different information sets:

1 The eigenvalues give you and indication of how much variance is explained by the factors. The larger they are, the more information each factor is conveying (if your species was distributed randomly throughout the study area, the eigenvalues would be all close to 1, marginality would be close to 0 and tolerance would be close to 1).

2 The score matrix (eigenvectors) tells you how the factors are correlated with the variables. Providing that a factor explains enough information (you can use the broken-stick criterion to select the significant factors), those variables that show the highest coefficient (in absolute value) are the more important to explain the species distribution. Should one variable have values close to zero on all relevant factors, you could as well jettison it. Note that keeping it in the model should have no other effect than slowing down the computations.

This is a major difference with stepwise regression analyses (GLM, GAM, logistic regression, ...). In these analyses, those variables that do not explain a significant amount of variance are rejected and do not appear anymore in the final model. Therefore, if two variables suffer from some correlation, only one of will be kept by the model. The problem is that these analyses need the variables to be reasonably independent (uncorrelated) for the algorithms to work reliably. With correlated variables, variables will be rejected arbitrarily. With ENFA, if two variables are correlated, they will both appear in the model with similar coefficient. The decision of keeping them or rejecting one of them is left to the ecologist and not to a blind algorithm.

Habitat Suitability map

Habitat Suitability map: Options

What is the "broken-stick advice"?

The distribution of the eigenvalue is compared to the distribution of Mac Arthur's broken-stick. It is the expected distribution when breaking a stick randomly. Therefore, the eigenvalues that are larger than what would have been obtained randomly may be considered "significant". You can also keep only the factors with an eigenvalue larger than 1. These are objective means to choose how many factors to keep for HS map computing. These are just indications designed as a support when selecting the factors.

The number of categories per factor for the making of the HS map changes by steps of two. I mean, you can only chose 2, 4, 6 and so on classes per factor. Is this normal?

Yes. It is because the HS computation is based on the median of the factor distributions and the median must fall between two classes and so the number of classes must be even.

What's the difference between "explained variance" and "explained information"?

The Ecological Niche factors are conveying two kinds of information: marginality and specialisation. In fact, the first factor explains always 100% of the marginality and some varying part of specialisation; the subsequent factors explain no marginality and the rest of the specialisation.

Historically, I was using the concept of "explained variance", which in fact was related only to "explained specialisation". Marginality was not used in this value. Therefore, when using this index for deciding how many factors had to be kept for the HS analysis, the user could be misled by this value. Accordingly, I introduced the concept of "explained information" to cope with this by giving marginality and specialisation the same "information power".

Mathematically, if $L1,L2,...,Lf$ are the eigenvalues of the f retained factors, and SL is the sum of all n eigenvalues ($= L1+L2+...+Lf+...+Ln$), we have :
Explained variance = Explained specialisation = $Se = (L1+L2+...+Lf)/SL$

In the explained information index, the marginality gets the same weight as the specialisation and we have therefore:

Explained information = $Ie = (Se+1)/2$

Ie is therefor always greater than Se and can never be smaller than 0.5. If the user chooses to keep only the first factor, he will explain at least half of the information, the one included into marginality. This Ie value is to be seen as a decision support value to choose how many factors are to be included into the HS analysis.

Is it possible to obtain a better model by reducing the amount of explained variance (i.e. the number of factors used) and, consquently, increasing the number of classes per factor?

Yes. You must find the best trade-off between explained variance and smoothness of the HS model.

Note that it is generally not useful to select more than 10 classes.

Habitat Suitability map: Other issues

How do you compute the HS value for each cell from the scores matrix?

This is a rather complex procedure. The full mathematical details of this operation are described in the main paper in *Ecology*.

Shortly said, for each retained factor, a frequency histogram is computed over all the cells of the map. The median of this distribution is computed. The further a cell is from this median, the lower is its suitability for this factor. The global suitability is then obtained by computing a weighted mean on these "partial suitabilities". Marginality has a weight of 1, the sum of specialisation factors has a weight of 1, proportionally to their eigenvalue. Mathematically this means that the weights are as follows:

Marginality factor: 0.5 + 0.5*L1/SL
Specialisation factor i: 0.5*Li/SL

where Li = i th eigenvalue and SL = Sum of the Li

I have computed an ENFA model and I would like to extrapolate it on a wider / other area. Is there some equation I can use to do it?

Presence-only models are difficult to extrapolate to other areas. Indeed they are based on the comparison between the locations where the species has been observed settling down and the available habitat. Although it can accurately make prediction on the study area, exporting the model to another place can be very tricky. In particular if you are comparing areas very far from each other.

Even extrapolation in a closer area can be tricky if the environmental layers have not been collected in the same way. An environmental variable as simple as mean water temperature can be done in very different ways (time of the day, depth, season, etc.) which could prevent the model to make good predictions. Or it is sufficient to move the study window a few kilometers to make it cover a different habitat distribution, which will bring unpredictable disturbances.

Finally, Biomapper has not been done for that purpose and there is presently no easy way to conduct such an extrapolation.

What is the difference between unidimensional multidimensional histogram algorithm for the computation of HS maps and how strong is the impact on the resulting HS maps?

The ful detail of the unidimensional algorithm are in the main paper in *Ecology*. The multidimensional is by now obsolete but I let here the explanation for history's sake. I give here a "feeling" of what they do and how they differ:

Unidimensional algorithm:

Once the ENFA factors have been computed, it is possible to compute for every cell of the map a value along each of them (in fact, usually, one computes it only for a few of the first factors). The distribution histogram of each factor is then computed taking into account only the presence points. Each histogram will be used to attribute a "partial suitability" value to every cell (the more the cell

factor value depart from the median of the histogram, the lower its "partial suitability". Then, a weighted sum of these partial suitability values is made for each map cell, and finally these sum are stretched in order to have their maximum at 100.

Multidimensional algorithm:

Here we do not address the selected factors independently. By crossing all factors together, the factor space is divided in small units (hypercubes). Then we count how many cells belong to each hypercube : this is the multidimensional histogram. This multidimensional distribution is computed both for all cells (global distribution) and the presence cells (species distribution). Finally, the HS value is computed for each hypercube by dividing the species hypercube by the global hypercube (and multiplied by 100).

The problem with the multi algorithm is that it need very huge amount of presence data to be accurate, in particular if you want to include more than two factors, which is generally the case. It is also very memory-consuming. So far, I never got good results with this algorithm and it is why it will not be described in your paper. We strongly advise Biomapper users to use only the unidimensional algorithm. In fact, I could well have removed it from the interface...

I am currently working on new kinds of algorithms, but it is another story... Stay tuned! Biomapperians will be informed of all the new developments.

I know that the habitat suitability of my species is linearly related to some variable. Nevertheless, in the HS map, the optimum for this variable seems not to be at an extremum. Why?

The HS computing algorithm is not linear but bases itself on the observed distribution. Your problem arises generally on the marginality factor. Let's imagine a species linearly related to the frequency of forest and that this variable is strongly correlated to the marginality factor (but the reasoning hold also for specialisation factors) ; it means that, the more forest there is around a given cell, the more suitable it is for the species, the maximum being a frequency of 100%. This optimum is what we know from our knowledge of the species, field studies, etc. Now, let's see the Biomapper's point of view: To it, the optimal frequency is the one where the species is the most frequent (More precisely, the median of the species distribution along the marginality factor. As the distribution of this factor is generally unimodal and more or less symmetrical, the median corresponds also to the most frequent.) Therefore, if large forests are rare in the study area, points with 100% forest freq. will be rare too, and the optimum for the species will not be 100% but lower (say 80%). Then, when computing the HS index, 80% freq. will provide the highest partial suitability value and this will decrease when freq. is either increasing or decreasing. The rarer the large forests, the steeper the rate of decrease.

Sometimes this effect is welcome (when dealing with median optima) and sometimes counterproductive (with extreme optima). I am currently working at new algorithms which will hopefully address this problem.

Does Biomapper produces any kind of formula for the habitat suitability?

Habitat suitability maps produced in Biomapper are based on an environmental envelope algorithm. These envelopes are fitting to the observed distribution in the niche space. That means there is no simple formula that can be used but rather a set of frequency histograms. Actually, the maps are produced directly within Biomapper and no formula is needed.

As for interpreting the ecological requirements of the studied species, it is done on the coefficients

of the ecological niche factors. coefficients of the ecological niche factors They will tell you how marginal and specialised the species is on the various relevant environmental variables.

Validation

What are the statistics provided after one leaves the Area-Adjusted Cross-validation box?

As for now, here are the one you may use:

Bin #: Area-adjusted frequency of bin #.
Rs: Spearman correlation coefficient of the AAF curve (cf Boyce et al's paper)
P(Rs=0): Significance of Rs, i.e. probability that Rs=0. This is provided for compliance with Boyce's paper, but I would suggest you don't use it.
AVI: Absolute Validation Index: the proportion of validation points whose HS>=50. This is the old validation method, before AAF came into play. It is used in some papers and this is why I'm going to keep providing it.
CVI: Contrast Validation Index: AVI-AVIchance, i.e. the AVI minus the AVI that would be expected from chance alone. This is actually similar to the AAF of a bin HS>=50. Again, this is an old index.

For now, just ignore any other statistics you may see. I'll sort them out in right time.

I don't understand what mean all the old validation statistics

I agree that this validation part is not yet very well documented... It is the fruit of hard work to find a way to evaluate a model without absence data. I tried thus many methods that are still present in the output, without any further explanation... I shall try here to unveil a few details on this subject. This will finally be incorporated into the help file.

You must have parted your sample into two sets. For validation purpose, you must use as "validation map" the set you did NOT used for model calibration. This observed map is therefore a boolean map indicating species presence. You can then evaluate the habitat suitability map produced with the ENFA model.

When no reliable absence data are available, evaluation consist to compare various statistics computed on the "predicted map" (the habitat suitability map): 1st On the whole study area; 2nd Only on the validation points.

The best way to understand this is to look at the box-plot displayed after the validation process. A good model should produce high species HS-value (80-100). The global box-plot gives you insight on how marginal the species is in the studied area and thus how much these results could have been obtained by chance only.

The results window gives a few statistics to resume this box-plot. It is composed of three parts: 1° species, 2° global statistics and 3° Comparisons.

The two first part give identical information: first, common distribution statistics are computed (mean, median, quartiles, etc.). Then a few more specifics statistics: the one I find the most useful is the "proportion of presence cells >50": this is the proportion of validation points that have a predicted HS-value over 50. The higher this value, the better your model. The "Prob. to be above this value by chance" statistic use a bootstrap procedure to assess how much this value could have

been obtained by chance. Practically, it is not very interesting as I have always got here a 0.000 probability (good news)... Then you can see the 90th percentile and 95th percentile. You can compare all the previous values between global and validation sets, but only the validation set gives a really objective idea of the quality of your model.

The last part give three comparison values between the two sets: The "Kappa" coefficient is a modified kappa statistics that integrate both how good is the model and how far from random it is. "Prop.of pix being significantly above 50" is the difference between the two "Prop.>50" values. And the "Probability to be over 50%" is the probability to be over 50% by chance, computed on the global set distribution histogram.

These three last values are interesting to assess how the model is different from what could be achieved by a random model but it says nothing about the absolute quality of your model. They are highly related to the global HS of the study area and thus, if your species is not very marginal nor very specialised, the model could be very good but get a very bad "far from random" score.

What is the "modified Kappa coefficient"?

I gave this name to a home-made statistic whose behaviour was to be similar to the Cohen's Kappa coefficient. It is computed as follows:

$K_{mod} = (M_S - M_G) / (1 - M_G)$

What does mean the bootstrap statistics?

This statistic is related to the proportion of validation points that have a HS value larger than 50. It gives the chance to get such a result by chance. It does so by bootstrapping the global distribution of HS value. In pseudo-code it is done as follows:

Let's P^* be the above proportion.

1. Draw randomly (with replacement) a HS value from the global distribution.
2. Repeat step 1 100 times and count the proportion P the value is larger than 50
3. Repeat steps 1 to 2 1000 times, counting the proportion of P being larger than or equal to P^*.

This proportion is the probability given by Biomapper's bootstrap test. It explains the probability to get an number higher than P^* by pure chance.

How are the ROC curves and kappa calculated if Biomapper does not use absence data?

The kappa and the ROC curves that you find in the validation dialog box assume that blank values are true absences. They are therefore not suitable for most data sets where absences are unreliable. I put them here for the cases where you can rely upon absences.

How can I compare results from GLM and ENFA?

Comparing ENFA and GLM is a tricky stuff. In my recent paper (Hirzel et al, 2001, Ecological Modelling 145), I was able to compare them because I was using virtual data and so I had access to the "reality", the "truth". But in the general case, we do not know it and so we are constrained to use the standard statistics (Kappa, ROC, etc.). There are three main problems when comparing presence/absence to presence-only methods:

- If absences are thought to be unreliable to build a model, there is no sense in using them to validate it afterwards. So, the standard statistics are not useful. I tried to develop a few statistics to replace them (see the FAQ on the Biomapper site) but the perfect statistic has still to be invented.
- As it is based on presence data only, the ENFA is more efficient to model the areas with average to high suitability; its predictions for low-suitability regions should be taken with prudence. By contrast, presence/absence methods and in particular GLMs, will tend to model good versus bad areas causing such a kind of "stepped" response which is different from the linear one of the ENFA.
- Without absences - i.e. bad habitat points - to "fix the floor", ENFA must scale its suitability index to the ceiling. That means that, on an ENFA HS map, you will always, by construction, have at least one cell with a HS of 1 (or 100) ; with GLM, it is generally not the case as it is computing "probability of presence" and so the maximum values are generally quite lower than 1.

Thus, when comparing visually GLM and ENFA maps (computed on well-known species at equilibrium, i.e., when absence data are largely reliable), the results are obviously quite similar. But if you try to compare them statistically, you get strange results biased either for one or for the other depending on which base hypothesis you use. To compare them you must correct both results to make them comparable:

- Remove the "ceiling effect" by stretching the GLM results between 0 and 1.
- Synchronise the "step effects" by transforming both GLM and ENFA results into boolean maps (by choosing a threshold).
- Then you can compare these results with standard presence/absence statistics.

How to compare models obtained by various factor number/class number trade-off?

You can use the validation module of Biomapper. Look at the box-plot it generates. Focus on the species box; it must be as high as possible. The higher and the narrower the better. (The global-box is not useful to assess model quality; it is here to see how different from randomness is your model. If you built your model on an area globally good for your species, you will get a good model simply because the species can live everywhere.)

Do you have a good reference about cross-validation methods?

A good paper about these methods is

> Fielding, A.H. & J.F. Bell (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38-49.

Note however that most methods described in this paper need presence/absence data. However, the k-fold cross-validation methods and their advantages are well presented.

How many partitions should I use for the cross-validation?

So far, I've found that 10 partitions was a good number, providing enough information about the predictive power variance. With few species data, this number may be too large and you should try a lower number.
Look at the Huberty's rule answer too.

What's the Huberty's rule?

The Huberty's rule is a rule of thumb for determining the ratio of calibration and validation points. It suggests that this ratio should be 1/(1 + sqrt(V - 1)), where V is the number of EGVs. (see p. 40 in Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24, 38-49.)

However, I'm not very happy with it. It does not depend on the number of observations, but only on the number of EGV and above 4 or 5 of them, it converges to a ratio of 75%, which corresponds to 4 partitions. Empirically, I 've found it better to use about 10 partitions.

What's random seed?

The partitions are built randomly. This causes each cross-validation to be different. However, a same random seed will always produce a same set of partitions, making cross-validation reproducible. That's useful when you want to compare two models.

How to interpret the curves resulting from the cross-validation?

The curves give you the response of the HS algorithm. It is important to look both at their shape (linear, sigmoid, exponential, etc.) and at their variance.

The **first** thing to **look at the variance**. The smaller the variance, the better the predictive power. If all the curves are close together, that means that you model is highly reproducible. If the variance is too high you can't say much more.

The **second** thing to look at is **how the variance is distributed** among the bins. The bin with a low variance are better predicted than the other. You can then tell if your model is best at predicting high-quality habitats or low-quality habitats or both.

The **third** thing is to **look at the shape of the curves**. The shape is not related to the predictive power but to the response "resolution". We can imagine several shapes:

- **Linear**: That's the best case. The model gives a good information for all HS values and you can draw a map with many shades.
- **Exponential**: That's good too. You have a low plateau at the start of the curve telling you that the model can't discriminate between low-quality habitats, but it gives a good resolution for the high-quality one. For the HS map, make a large "unsuitable" category encompassing the whole plateau (below 1) and a few narrow categories in the growing part of the curve.
- **Sigmoid**: Here, you have two plateaus, connected by a transition curve. This kind of results gives less resolution in the HS map. Makes 3 or 4 categories: one for "unsuitable habitat" (low plateau), one fore "core habitat" (high plateau) and one or two in the transition part for "marginal habitat".
- **Saw-toothed**: When you have saw-teeth in some part of the curve, that means there is a large variance in that region on the curve, or too many points in those bins. Select a lower number of bins and make the shaky bins wider (you can modify their width in the histogram plot) until there is no saw-teeth anymore and you should fall down to one of the above shapes (probably a sigmoid).
- **Flat line**: Sorry, there is no hope anymore. The line is probably very close to 1 and it means that your model does not much better than just giving random HS values to the cells in the map.

How to reclassify the HS map into suitable/unsuitable according to the cross-validation graphs?

In the cross-validation windows, select the number of bins you want to have (in the suitable/unsuitable case, that would be 2). Then you can change the bin boundaries by clicking in between two bars on the histogram and drawing in around. Once you have placed the bins as you want them, click on the "Classify" button. This will open a new window where you can select which map you want to reclassify according to these bins and voilà! You'll get a new map, reclassified into 2 classes according to your bins.

How to place the bin boundaries is another problem, which depend on the species:

- With a very mobile species (like a bird), you may select the point where the AAF curves cut the 1 threshold, thus separating the areas where the species is found more frequently than is expected by chance, from the areas where it is found more rarely than by chance.
- With a more sedentary species (like an insect, or a small mammal), which is not often found outside of its suitable habitat, then you may want to keep most of the points and thus choose a low HS threshold. Typically, choosing HS=10 (with ispleth scaling) means that 90% of the observed presences will be included into the suitable habitat and 10% will be outside.
- Personally, I prefer a smoother approach by making three classes (or more, if the data allow it): a core habitat where the AAF curves are well above the 1-line, a marginal habitat where the AAF curves are about 1, and an unsuitable habitat below 10 (so as rejecting 10% of the presence points).

Whenever I increase the number of bins, I get more rugged AAF curves. Why?

This is a common pattern: as you increase the number of bins, the number of validation points per bin necessarily decreases, thus increasing the variance in the AAF curves. The less presence points you have, the more it is going to happen. You can precisely use this information to choose the correct number of bins and reclassify the map accordingly.

And after that...

Publication

How to quote Biomapper / ENFA?

You can quote the main paper:

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: How to compute habitat- suitability maps without absence data? Ecology 83, 2027-2036.

and the Biomapper software itself:

Hirzel, A., Hausser, J., Perrin, N., 2002. Biomapper 3.1. Lausanne, Lab. for Conservation Biology. URL: http://www.unil.ch/biomapper.

Do you have any published paper about ENFA?

There are already several papers. More are coming. You can get a list of related publications on the Bibliography page. There, you can also download a PDF version of my PhD thesis. Registered users will be informed when new papers are coming out.