

**Ecological Niche Modelling:
Inter-model Variation.
Best-subset Models Selection**

Enrique Martínez-Meyer

The problem:

We want to represent the geographic distribution of species under the following circumstances:

Most occurrence data available for the vast majority of species are asymmetric (*i.e.* presence-only)

Sampling effort across most species' distributional ranges is uneven, thus occurrence datasets are eco-geographically biased

Environmental variables encompass relatively few niche dimensions, and we do not know what variables are relevant for each species

More problems:

Many algorithms do not handle asymmetric data (*e.g.* GLM, GAM)

Some of the algorithms that does handle asymmetric data do not handle nominal environmental variables (*e.g.* soil classes) [*e.g.* Bioclim, ENFA]

Many stochastic algorithms present different solutions to a problem, even under identical parameterization and input data (*e.g.* GARP)

We do not know the ‘real’ distribution of species, so we do not know when models are making mistakes (mainly over-representing distributions), and when are filling knowledge gaps

We have to live with all those problems, so we need a way to make the best decision possible

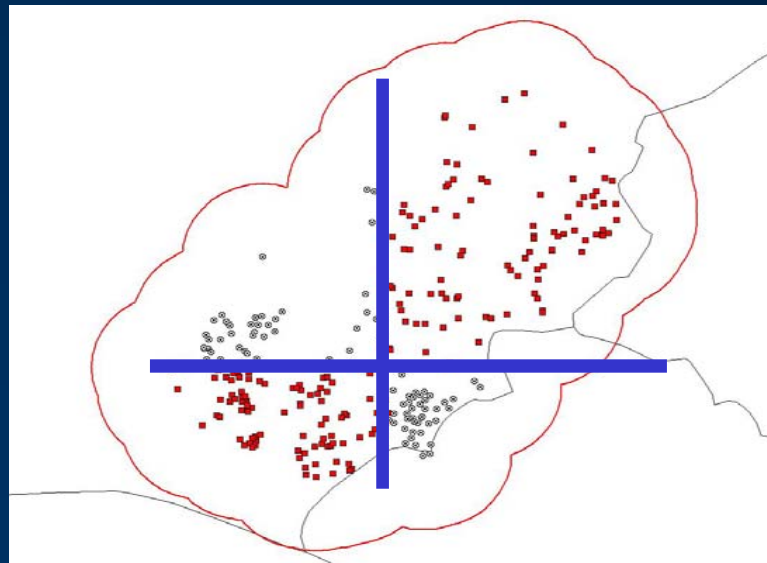
Anderson, Lew and Peterson (2003) developed a procedure to detect the best-subset models among a given amount of varying models

To evaluate model quality you need to:

1. Generate an 'independent' set of data

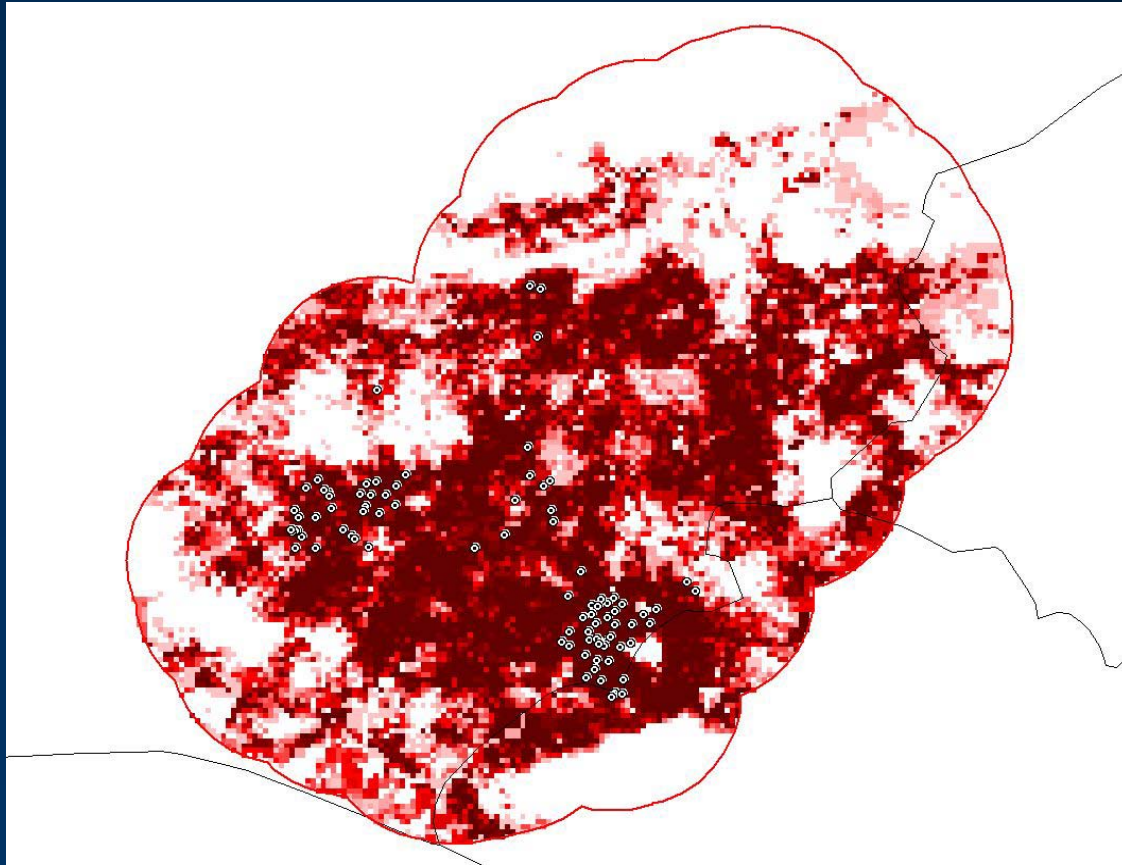
There are at least two strategies to do so:

- Collect new data
- Split your data into two sets

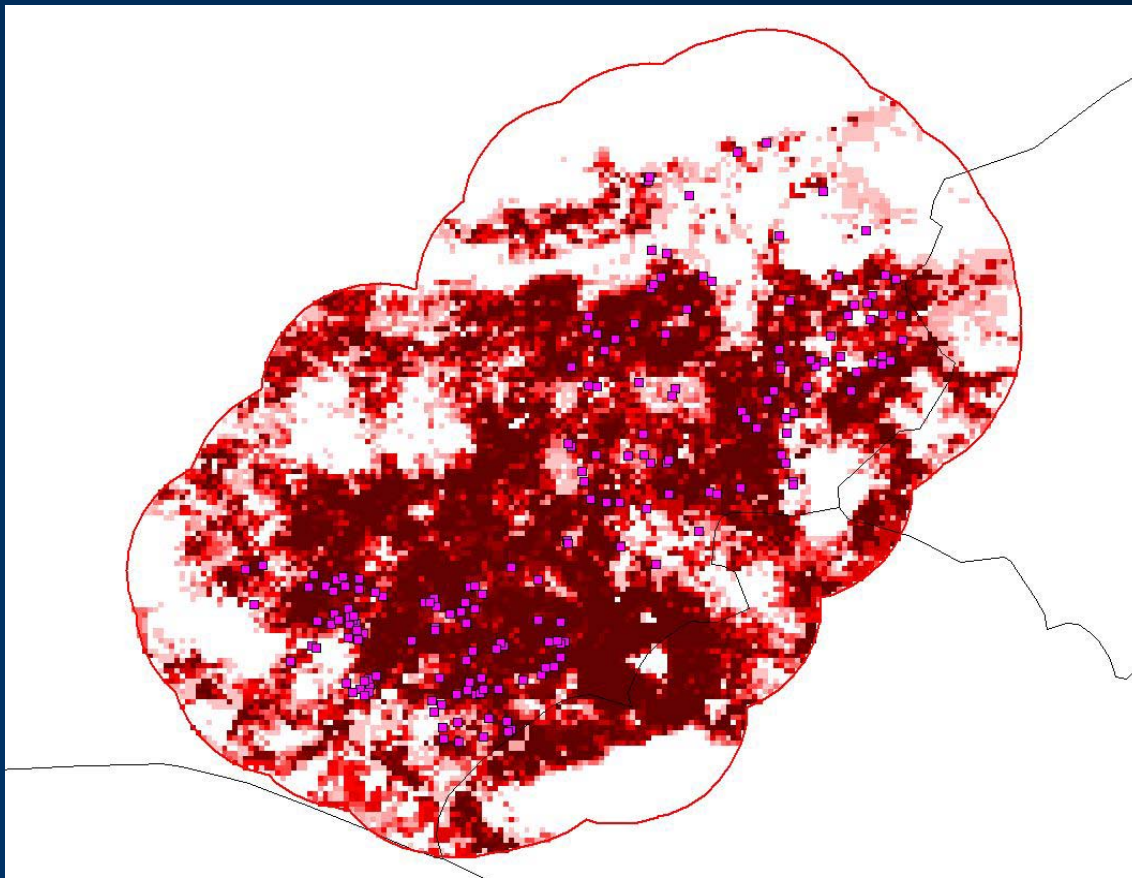


Regardless of the method, you end up with two data sets, one for *training* the model, and one for *testing* the model

2. Generate a model with the *training* data



3. Quantify error components with a confusion matrix



	Actually Present	Actually Absent
Predicted Present	<i>a</i>	<i>b</i>
Predicted Absent	<i>c</i>	<i>d</i>

a & *d* = correct predictions

b = commission error
(false positives,
overprediction)

c = omission error
(false negatives,
underprediction)

What does *omission* mean?

In general, omission error can be considered a 'hard' (true) error. However, under some circumstances, a presence record might not be so:

1. Identification of the taxon is wrong
2. Georeferencing of the locality is wrong
3. A record represents individuals outside of their ecological niche; *e.g.* 'sink' (rather than 'source') populations; individuals in transit or vagrant, etc.

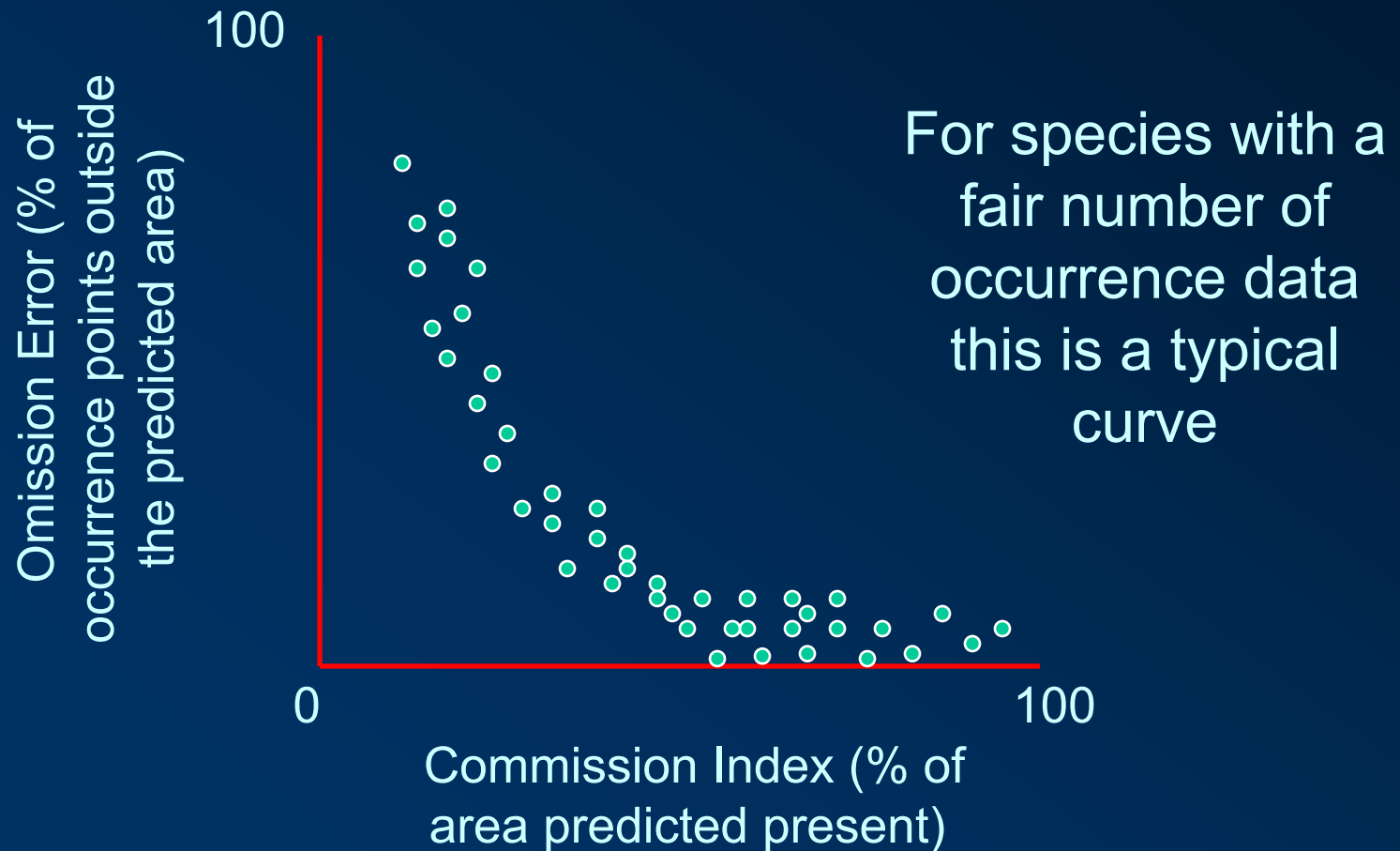
What does *commission* mean?

Model overprediction may or may not be an error, so commission is not a 'hard' error.

Prediction in areas where a species does not have a confirmed record is caused by different factors:

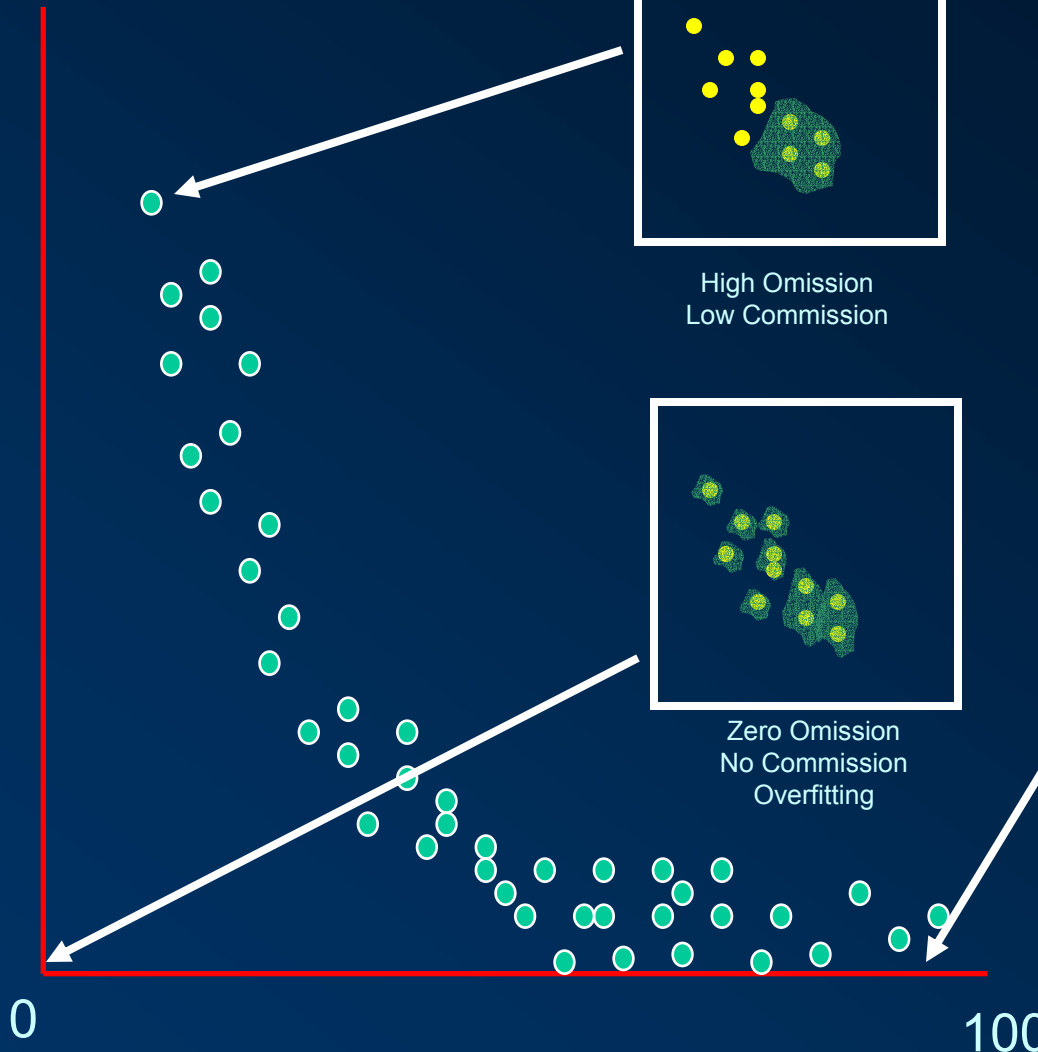
1. The area is suitable for the species, but no sampling effort has been made. The species may be there.
2. The area is suitable for the species, but historical (barriers, dispersal capability) or biotic (competition, predation) factors have impeded the species to occupy it, or went extinct.
3. The area is unsuitable: True commission error

Some stochastic algorithms (like GARP) produce somehow different models with the same input data. If we produce several models, we can calculate their errors and plot them in an omission/commission space



Omission Error (% of occurrence points outside the predicted area)

100



High Omission
Low Commission

Zero Omission
No Commission
Overfitting

Distribution of a
species in an area

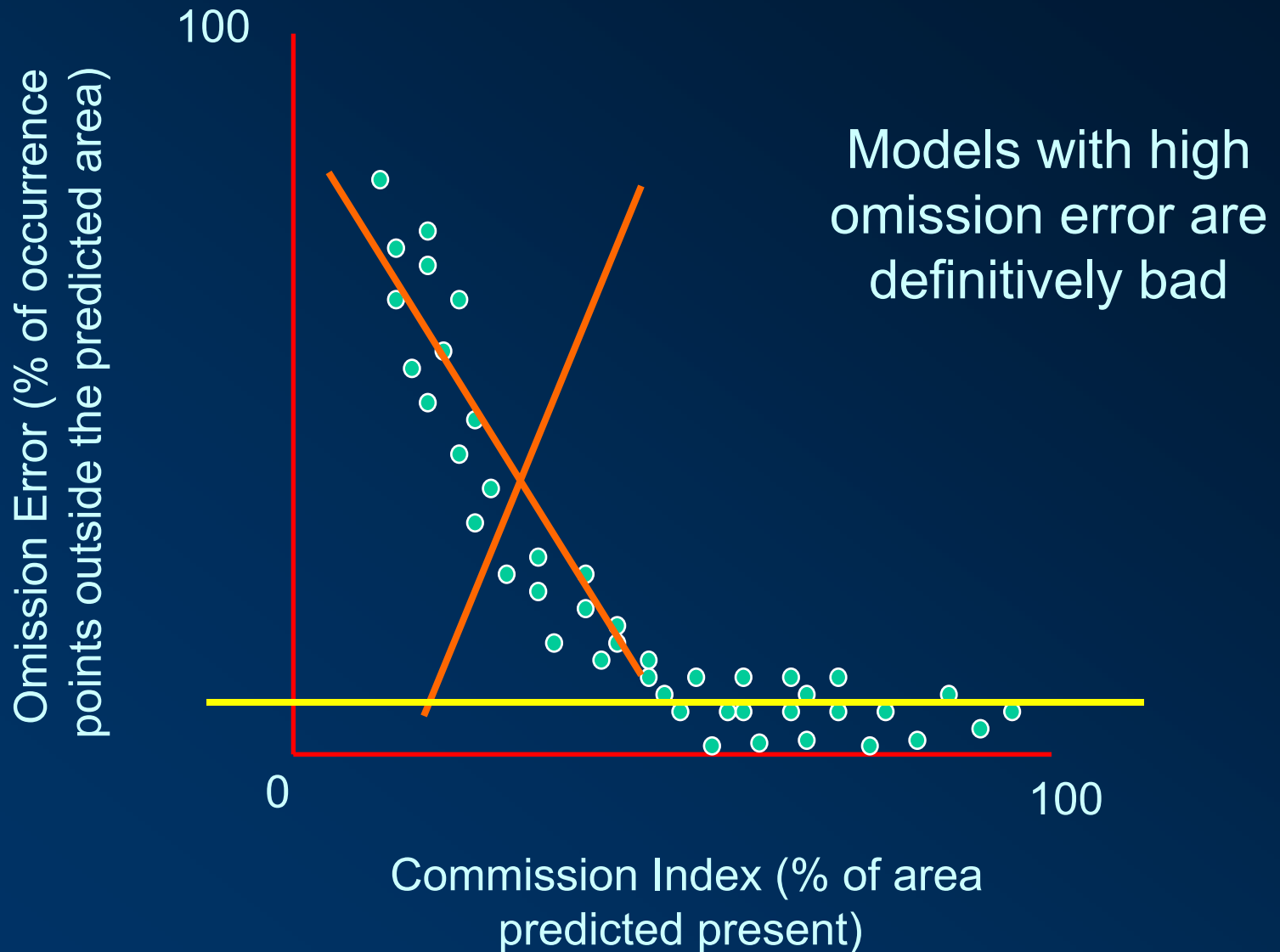
Zero Omission
High Commission

0

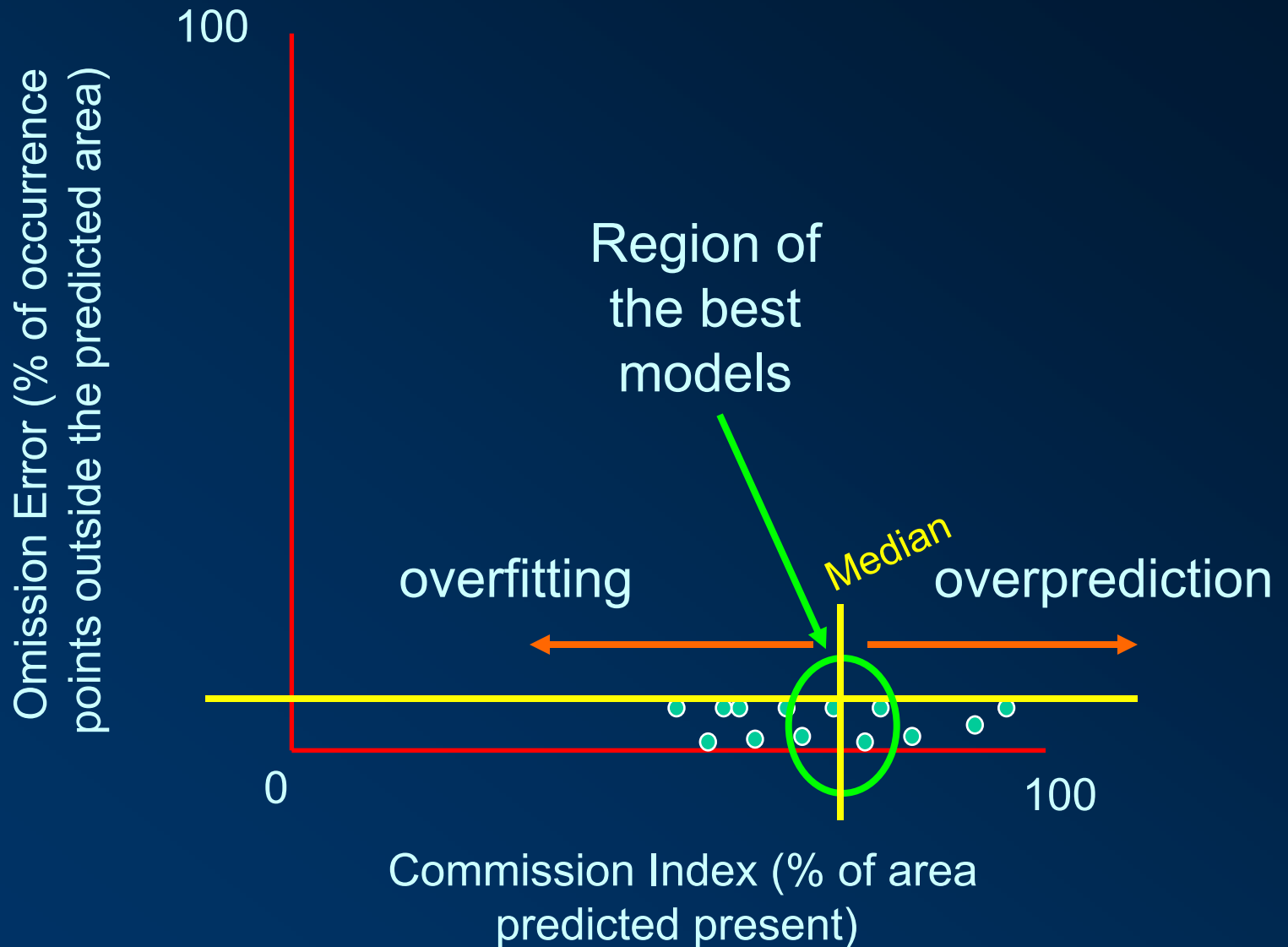
100

Commission Index (% of area
predicted present)

The question now is, which of these models are good and which ones are bad?



The question now is, which of these models are good and which ones are bad?



Implementation in Desktop GARP

Having enough occurrence data, you can split them into *training* and *testing* datasets. When this is the case, it is convenient to select *Extrinsic* in the *Omission Measure* option. Otherwise, if you have *100% for training*, you have to select *Intrinsic*

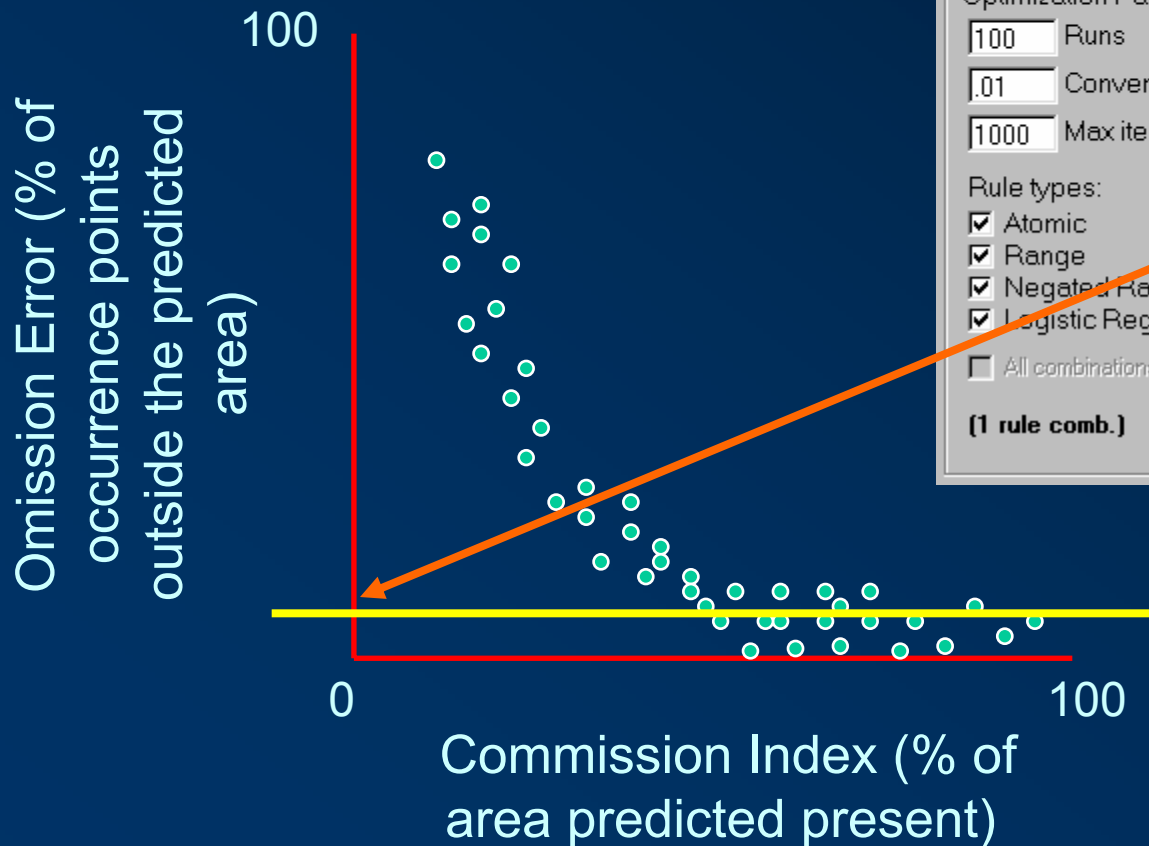
The screenshot shows the Desktop Garp software interface with the following settings:

- Species Data Points:**
 - Species List: [2 selected]
 - L.callotis (77)
 - S.cunicularius (41)
 - Upload Data Points button
 - Options:
 - Use % for training
 - At least training points
- Optimization Parameters:**
 - Runs:
 - Convergence limit:
 - Max iterations:
 - Rule types:
 - Atomic
 - Range
 - Negated Range
 - Logistic Regression (Logit)
 - All combinations of the selected rules
 - [1 rule comb.] [100 total runs]
- Best Subset Selection Parameters:**
 - Active
 - Omission measure: Extrinsic, Intrinsic
 - Omission threshold: Hard, Soft
 - % omission
 - Total models under hard omission threshold:
 - Max models per spp.: 100
 - Commission threshold: % of distribution

An orange arrow points from the 'At least 30 training points' checkbox to the 'Extrinsic' radio button in the 'Omission measure' section.

Implementation in Desktop GARP

In the *Omission threshold* section, if you select *Hard* means that you will use an absolute value in the omission axis of the plot. You set that value in the *% omission* box

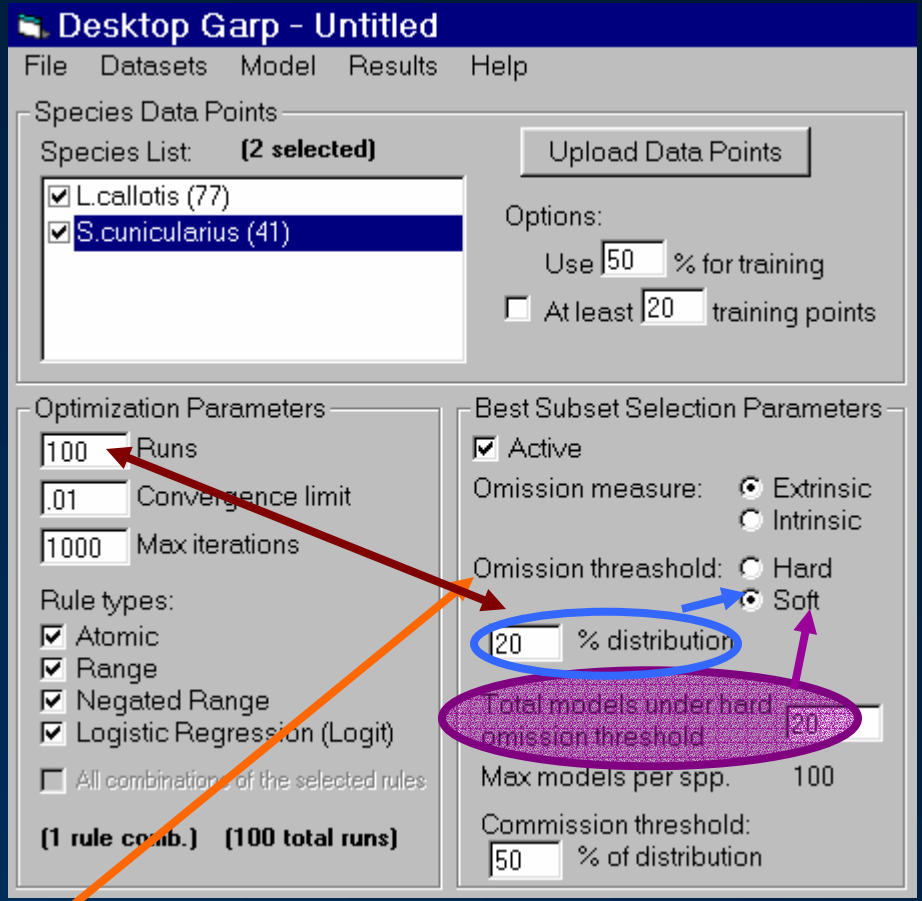
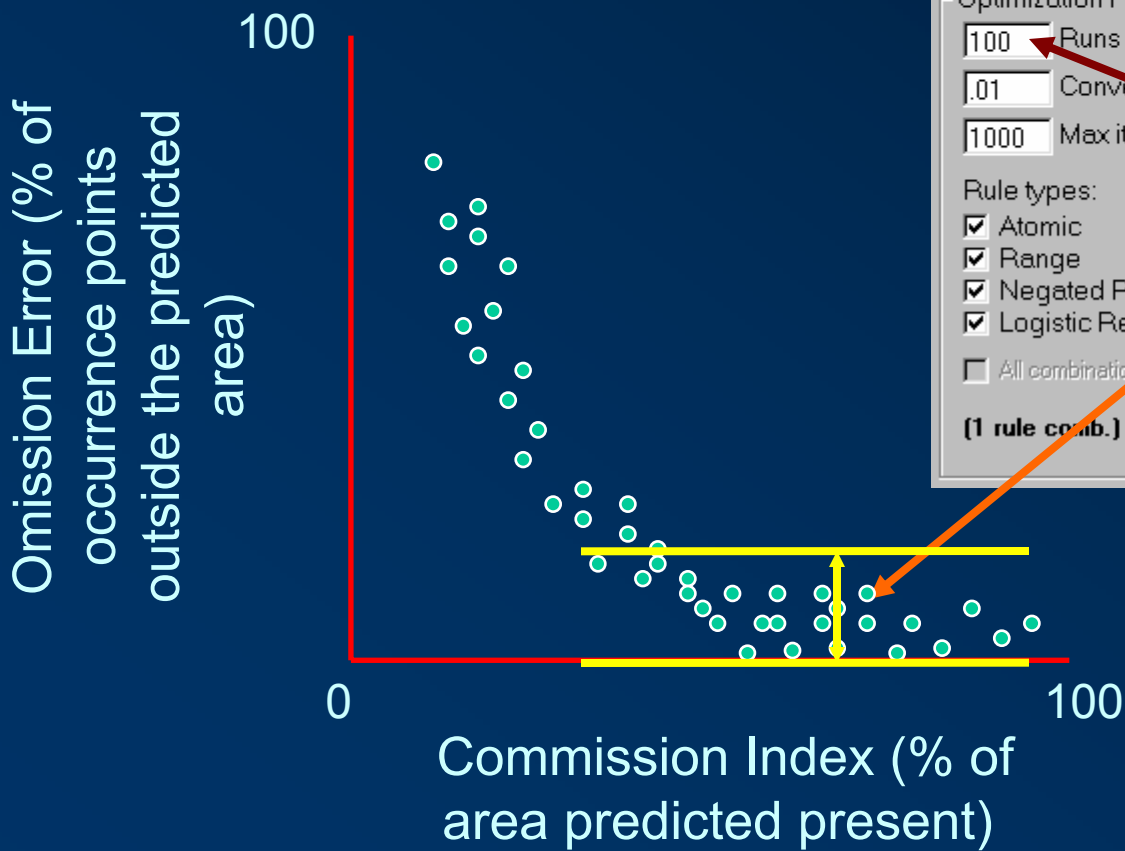


The screenshot shows the Desktop Garp software interface. The title bar reads 'Desktop Garp - Untitled'. The menu bar includes 'File', 'Datasets', 'Model', 'Results', and 'Help'. The 'Species Data Points' section shows a 'Species List' with two selected species: 'L.callotis (77)' and 'S.cunicularius (41)'. There is an 'Upload Data Points' button and 'Options' for training, including 'Use 50 % for training' and 'At least 20 training points'. The 'Optimization Parameters' section includes 'Runs' (100), 'Convergence limit' (.01), 'Max iterations' (1000), and 'Rule types' (Atomic, Range, Negated Range, Logistic Regression (Logit)). The 'Best Subset Selection Parameters' section is active, showing 'Omission measure' (Extrinsic), 'Omission threshold' (Hard), and 'Total models under hard omission threshold' (20). A blue circle highlights the '10 % omission' box, and a purple circle highlights the '20' box. An orange arrow points from the '10 % omission' box to the yellow line in the plot, and a purple arrow points from the '20' box to the plot.

Then you have to select the number of models that you want DG to select under that hard omission threshold

Implementation in Desktop GARP

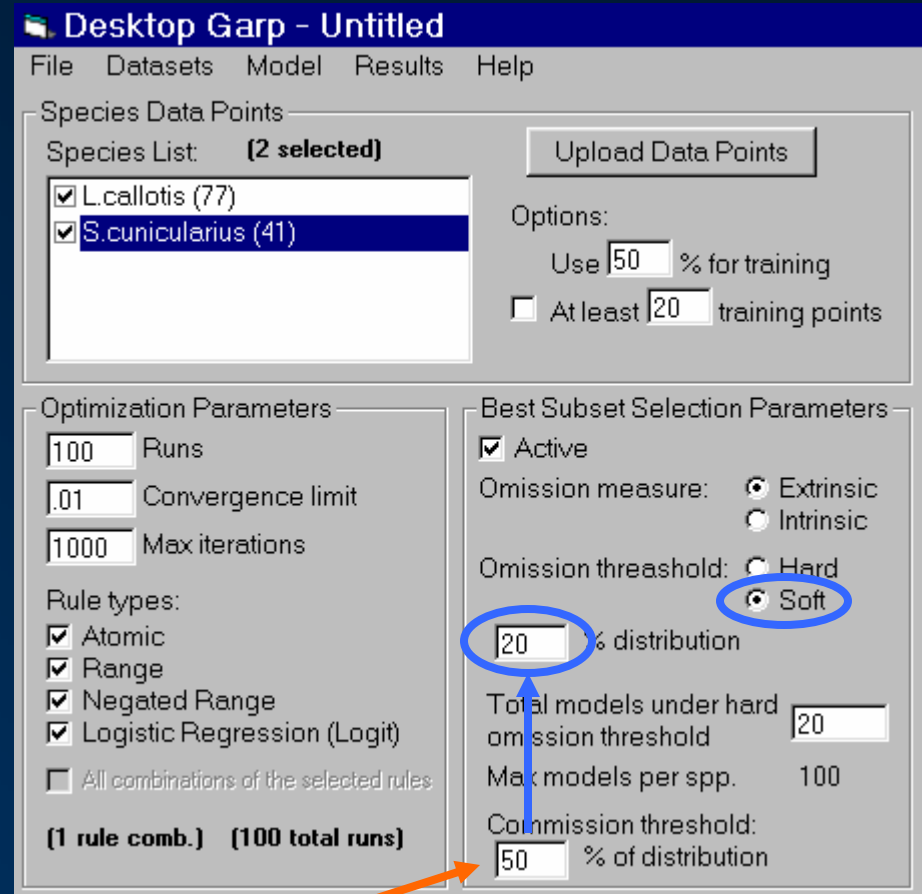
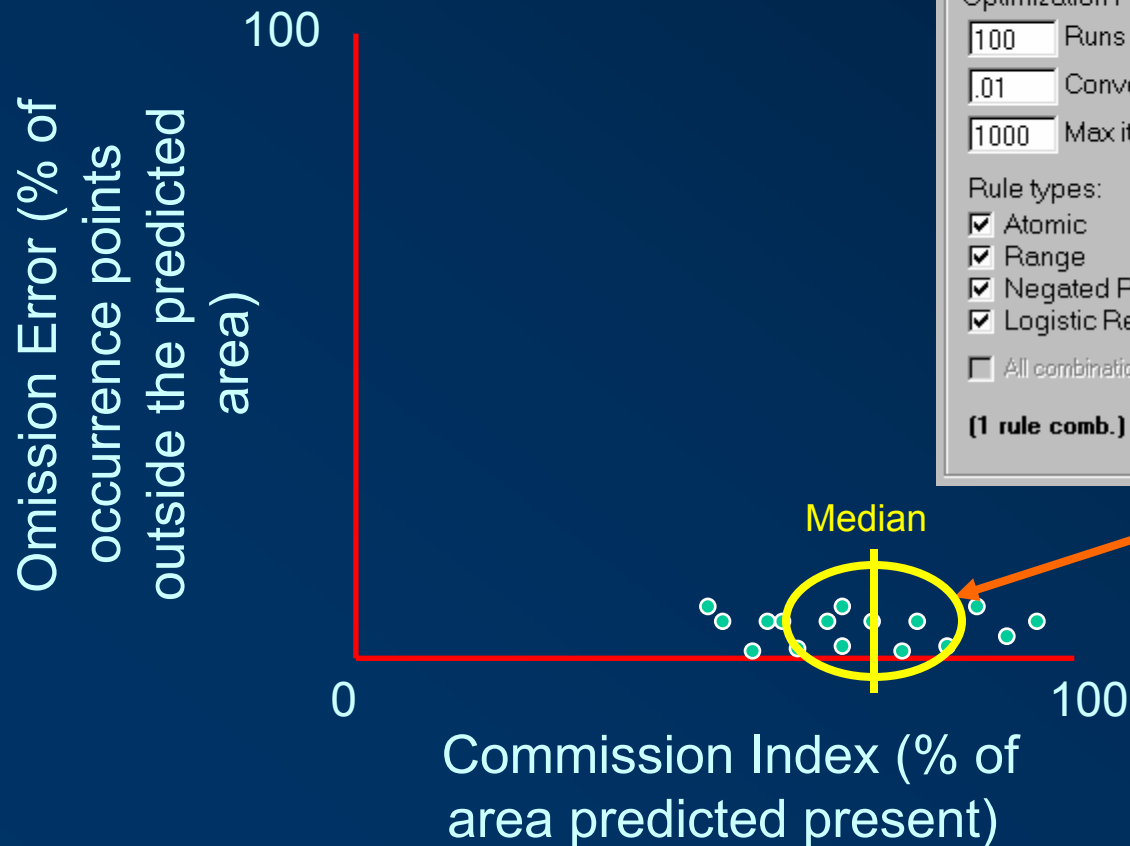
When you select *Soft* means that you will select certain number of models (in percentage), indicated in the % distribution box, with the least omission. This is useful when you are running more than one species at a time



In this case, the *Total models under hard omission threshold* box does not apply

Implementation in Desktop GARP

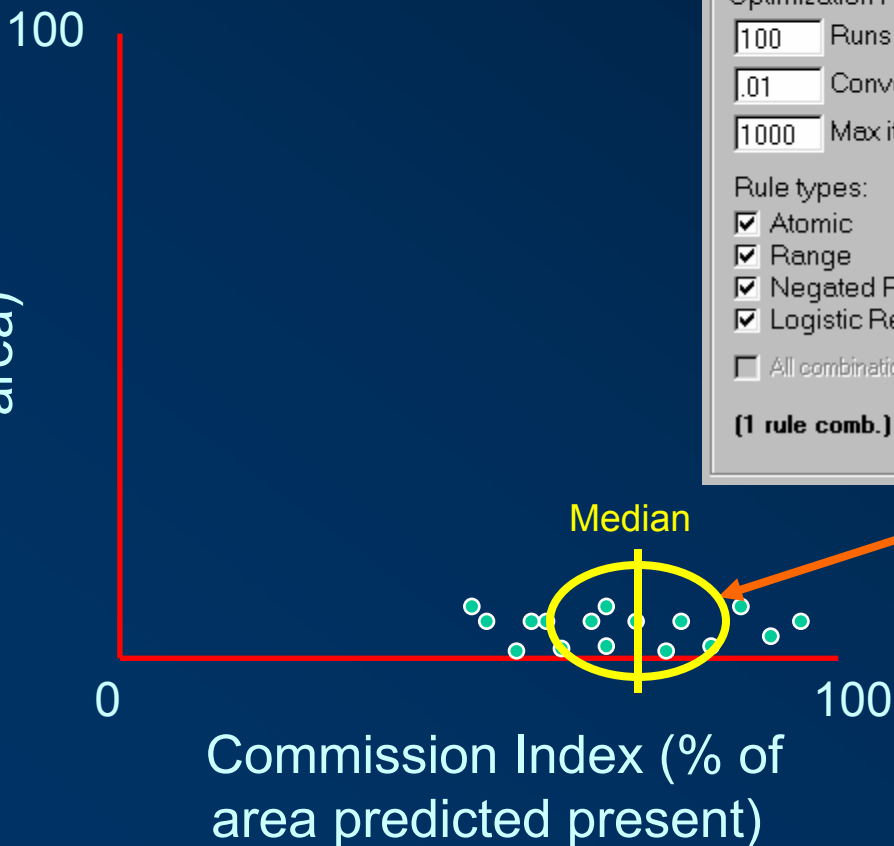
Finally, in the *Commission threshold* box you indicate the number of models (in percentage) closer to the Median in the Commission Index axis that you want to be selected from the remaining models, after filtering with the omission criteria



When the *Omission threshold* is in *Soft*, the *Commission threshold* value is relative to the % *distribution* value

Implementation in Desktop GARP

Omission Error (% of occurrence points outside the predicted area)



Desktop Garp - Untitled

File Datasets Model Results Help

Species Data Points

Species List: **(2 selected)**

- L.callotis (77)
- S.cunicularius (41)

Upload Data Points

Options:

Use % for training

At least training points

Optimization Parameters

Runs

Convergence limit

Max iterations

Rule types:

- Atomic
- Range
- Negated Range
- Logistic Regression (Logit)
- All combinations of the selected rules

(1 rule comb.) (100 total runs)

Best Subset Selection Parameters

Active

Omission measure: Extrinsic Intrinsic

Omission threshold: Hard Soft

% omission

Total models under hard omission threshold:

Max models per spp. 100

Commission threshold: % of distribution

When the Omission threshold is in *Hard*, the Commission threshold value is relative to the Total models under hard omission threshold value