CONCEPTS AND PRACTICE FOR ECOLIGICAL NICHE MODELLING

CICGE SBL Spatial Biology Lab
U.PORTO
FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Neftalí Sillero
CICGE/SBLab/FCUP
neftali.sillero@gmail.com

DECIDING WHAT TO MODEL

1. Think about what do you want to model.

2. Which is the ecological niche you want to model?
   Potential or *realised* niche (*sensu* Sillero 2011)?

3. Choose between *mechanistic* or *correlative* modelling methods depending in the ecological niche you want to model.

- If you choose ***mechanistic models***, you will need only species' physiological data.
- If you choose ***correlative models***, you will need species' **presence-absence records** or species' **presence-only records**.

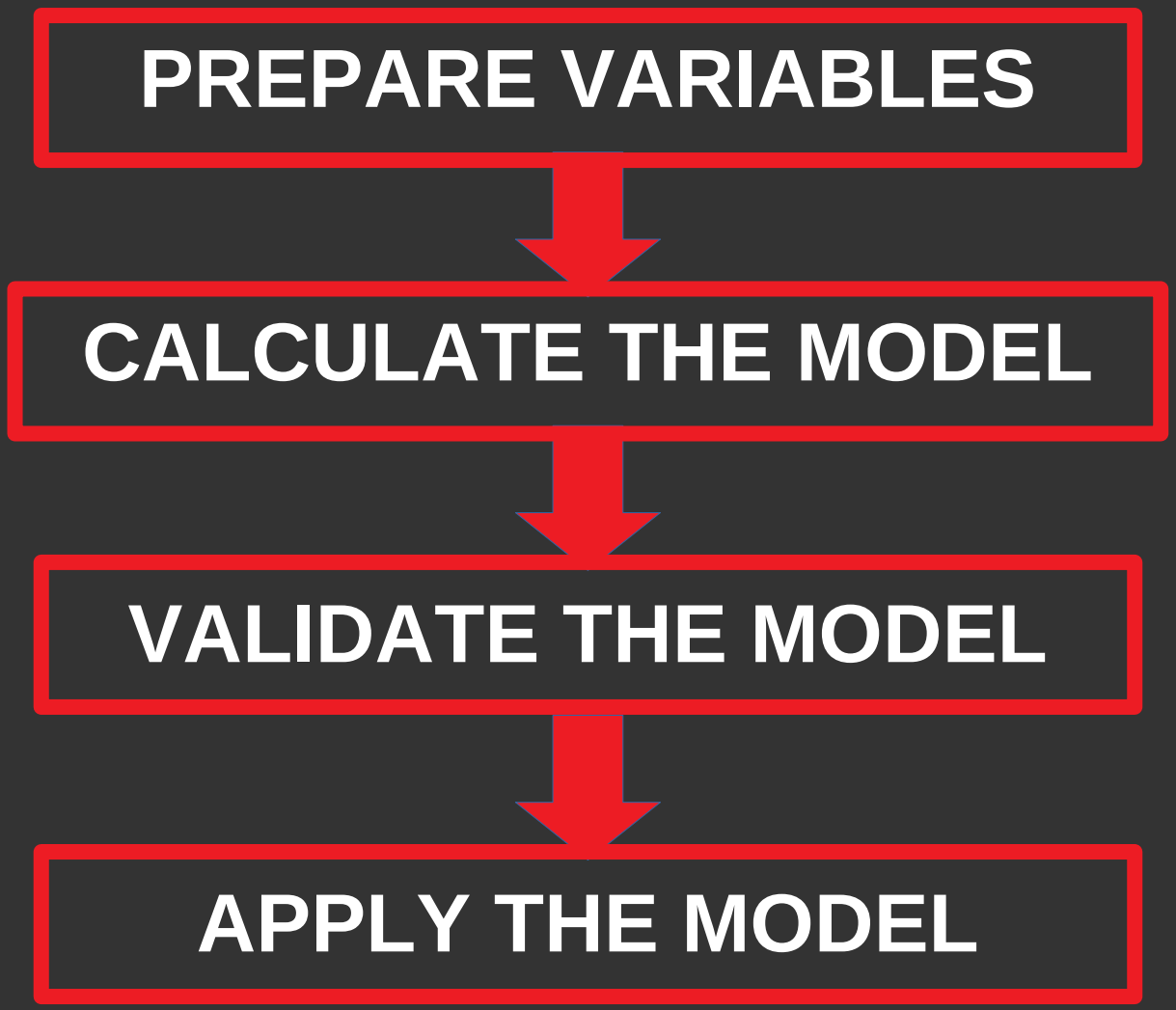
Guisan and Zimmermann 2000
Robertson et al. 2003
Elith et al. 2006
Jiménez-Valverde et al. 2008
Kearney et al. 2008
Kearney and Porter 2009
Sillero et al. 2010

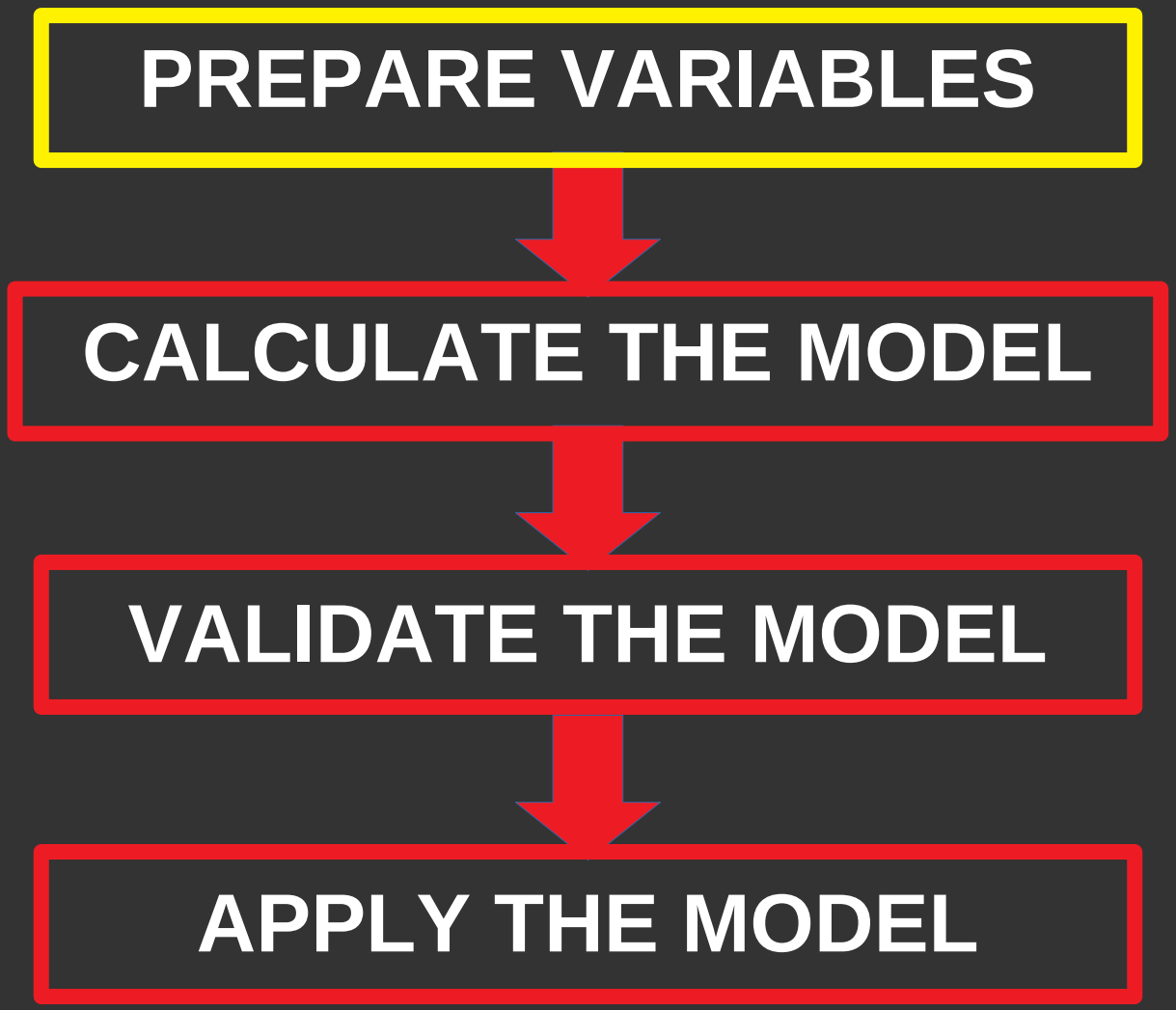
Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2015) **24**, 276–292

RESEARCH REVIEW

Is my species distribution model fit for purpose? Matching data and models to applications

Gurutzeta Guillera-Arroita[1]*, José J. Lahoz-Monfort[1], Jane Elith[1], Ascelin Gordon[2], Heini Kujala[1], Pia E. Lentini[1], Michael A. McCarthy[1], Reid Tingley[1] and Brendan A. Wintle[1]


PREPARE VARIABLES
CALCULATE THE MODEL
VALIDATE THE MODEL
APPLY THE MODEL


PREPARE VARIABLES
CALCULATE THE MODEL
VALIDATE THE MODEL
APPLY THE MODEL

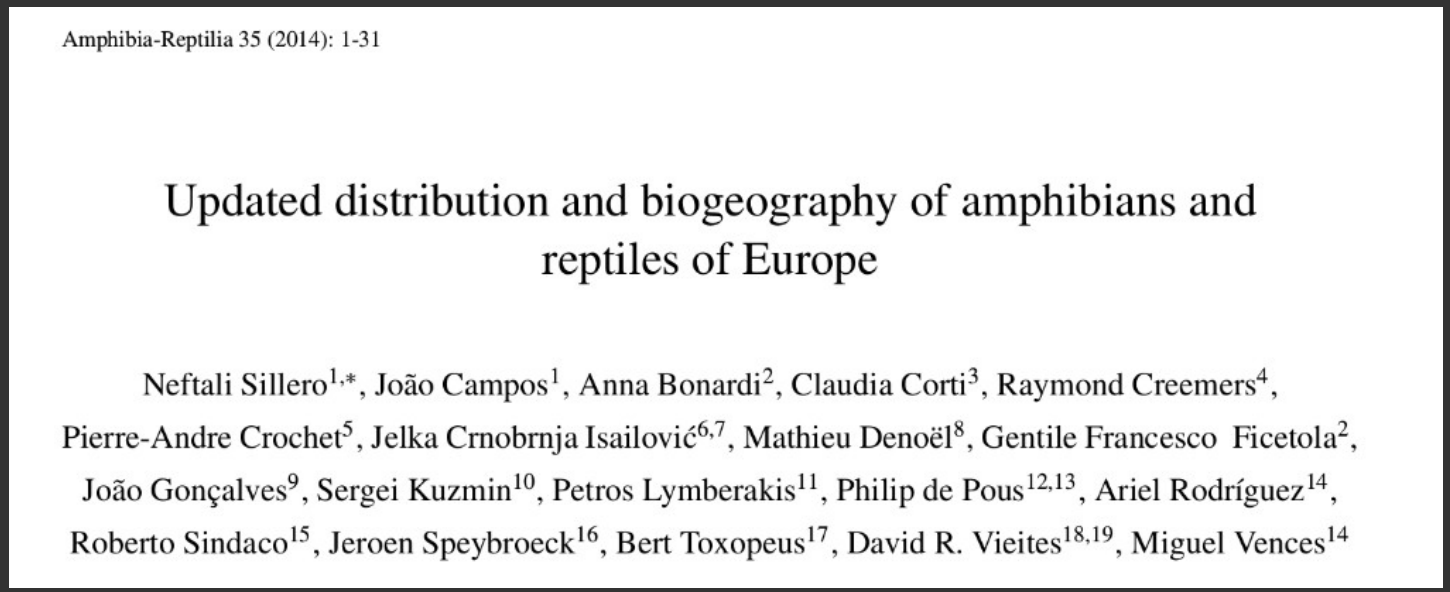
4. Check the following assumptions:

- **Niche Space Assumption**: The study contains the full range of conditions that the species can inhabit (for the examined abiotic variables).
- **Dispersal/demographic Noise Assumption**: Factors related to dispersal, establishment, and persistence do not cause the species to occupy an environmentally biased subset of the abiotically suitable areas.
- **Biotic Noise Assumption**: Biotic interactions do not cause the species to occupy an environmentally biased subset of the abiotically suitable areas.
- **Human Noise Assumption**: Human modifications of the environment do not cause the species to occupy an environmentally biased subset of the abiotically suitable areas.

Anderson 2012

5. Obtain the species' records.

- You can get species' records from several sources like on-line atlases, GBIF, fieldwork (GPS).
- You need a minimum sample size of species' records. Maxent is a good software for very small datasets.


Amphibia-Reptilia 35 (2014): 1-31

Updated distribution and biogeography of amphibians and reptiles of Europe

Neftali Sillero[1,*], João Campos[1], Anna Bonardi[2], Claudia Corti[3], Raymond Creemers[4], Pierre-Andre Crochet[5], Jelka Crnobrnja Isailović[6,7], Mathieu Denoël[8], Gentile Francesco Ficetola[2], João Gonçalves[9], Sergei Kuzmin[10], Petros Lymberakis[11], Philip de Pous[12,13], Ariel Rodríguez[14], Roberto Sindaco[15], Jeroen Speybroeck[16], Bert Toxopeus[17], David R. Vieites[18,19], Miguel Vences[14]

- **Contingent absences**
- **Environmental absences**
- **Methodological absences**

Ecography 33: 103–114, 2010
doi: 10.1111/j.1600-0587.2009.06039.x

Subject Editors: Núria Roura-Pascual and Nathan K. Sanders. Accepted 30 November 2009

The uncertain nature of absences and their importance in species distribution modelling

Jorge M. Lobo, Alberto Jiménez-Valverde and Joaquín Hortal

J. M. Lobo (mcnj117@mncn.csic.es), Dept Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales, c/ José Gutiérrez Abascal 2, ES-28006, Madrid, Spain. – A. Jiménez-Valverde, Natural History Museum and Biodiversity Research Center, The Univ. of Kansas. Lawrence, KS 66045, USA. – J. Hortal, NERC Centre for Population Biology, Div. of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK.

TYPES OF ABSENCES

Contingent absences: from climatically suitable areas but not occupied for different reasons (e.g. historical and dispersion).

- Outside of the occupied but inside the realised/fundamental niche.
- More probable in spatially distant localities with favourable environmental conditions.

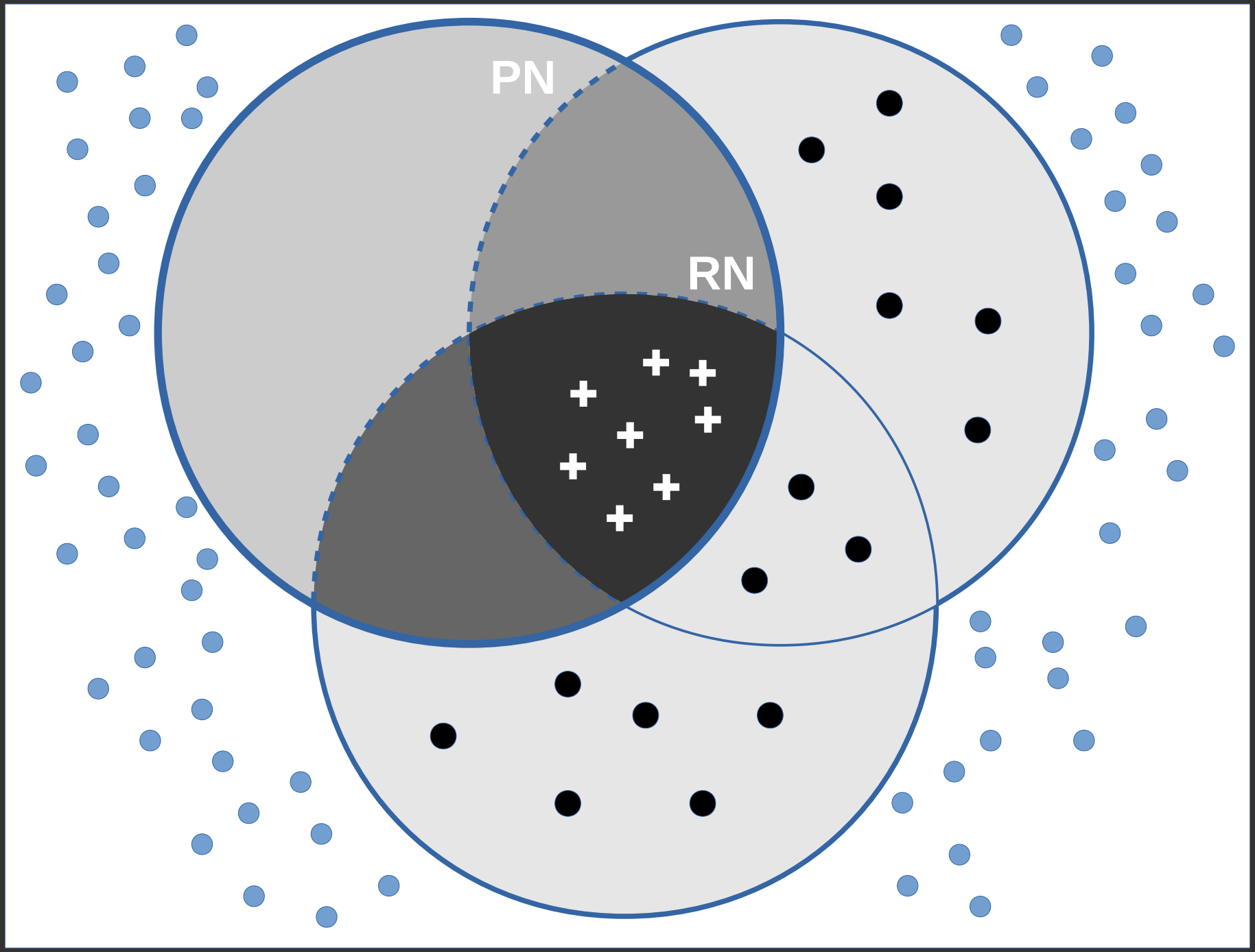
Environmental absences: from climatically unsuitable areas.

- Outside of both the realised and fundamental niche.
- More probable in those localities showing environmental conditions far away from the environmental universe defined by the presence localities.

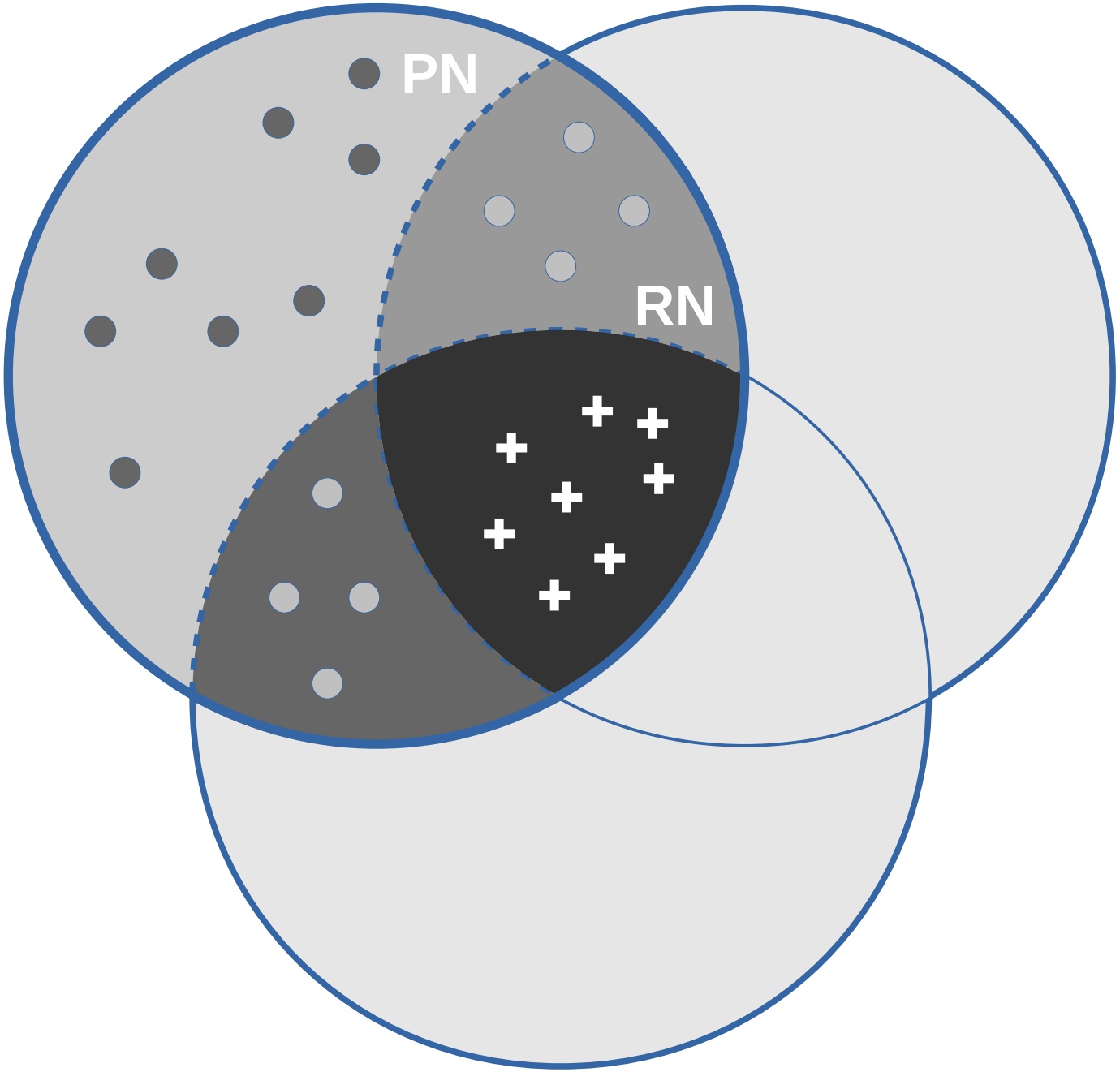
Methodological absences: created by errors in surveys.

- Inside both the realised and the fundamental niche.
- More probable in the environmentally favourable localities placed nearest to the known presence points.

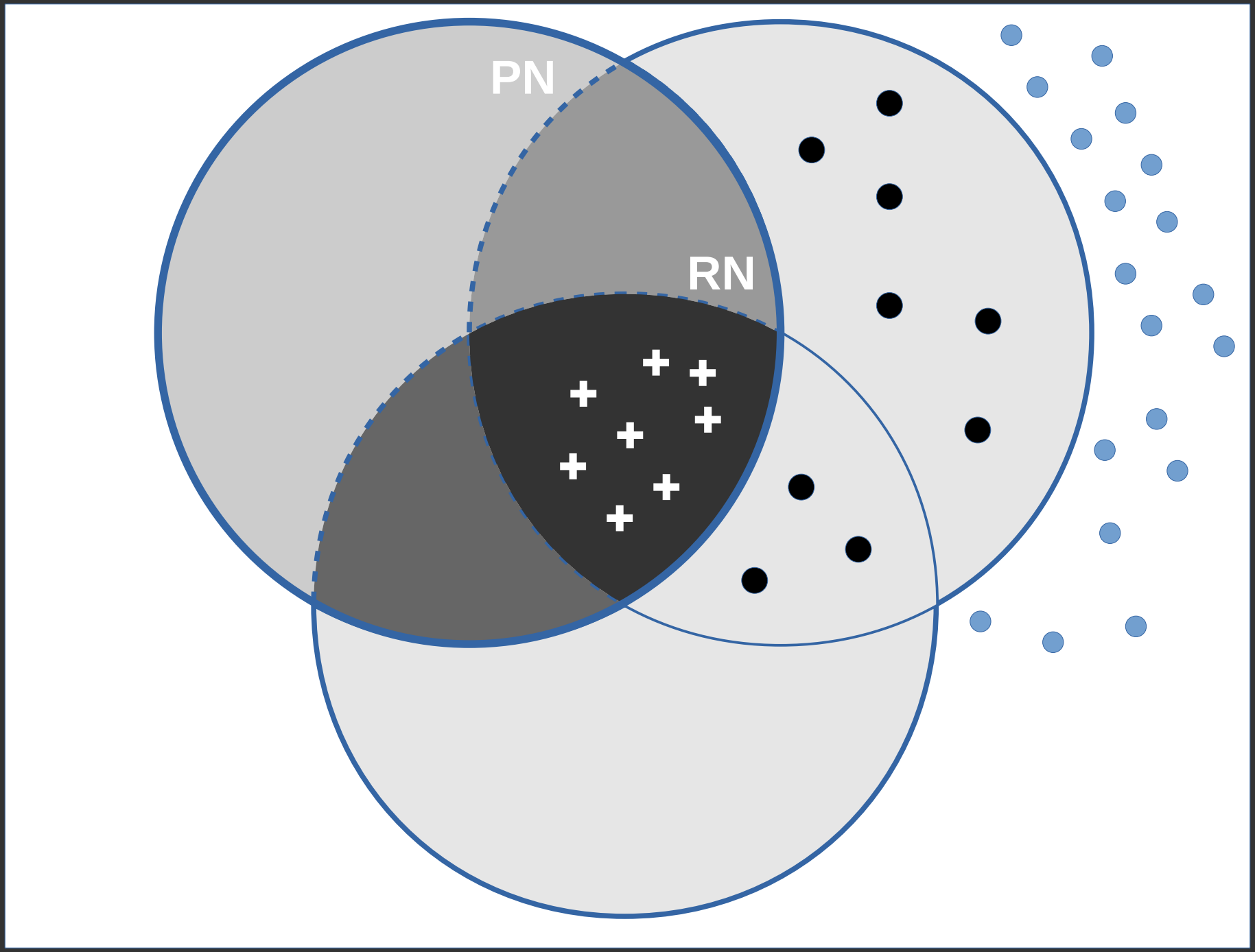
ENVIRONMENTAL ABSENCES



CONTINGENT ABSENCES



METHODOLOGICAL ABSENCES



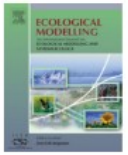
TYPES OF PSEUDO-ABSENCES

ECOLOGICAL MODELLING 220 (2009) 589–594

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Short Communication

Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know?

Jeremy VanDerWal[a,*], *Luke P. Shoo*[a], *Catherine Graham*[b], *Stephen E. Williams*[a]

[a] Centre for Tropical Biodiversity and Climate Change, School of Marine and Tropical Biology, James Cook University, Townsville, Queensland 4811, Australia

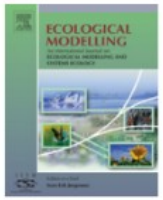
[b] Department of Ecology and Evolution, 650 Life Sciences Building, Stony Brook University, NY 11794, USA

ECOLOGICAL MODELLING 210 (2008) 478–486

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Assessing the effects of pseudo-absences on predictive distribution model performance

*Rosa M. Chefaoui, Jorge M. Lobo**

Dpto. de Biología Evolutiva y Biodiversidad, Museo Nacional de Ciencias Naturales, c/ José Gutiérrez Abascal 2, 28006 Madrid, Spain

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 327–338 doi: 10.1111/j.2041-210X.2011.00172.x

Selecting pseudo-absences for species distribution models: how, where and how many?

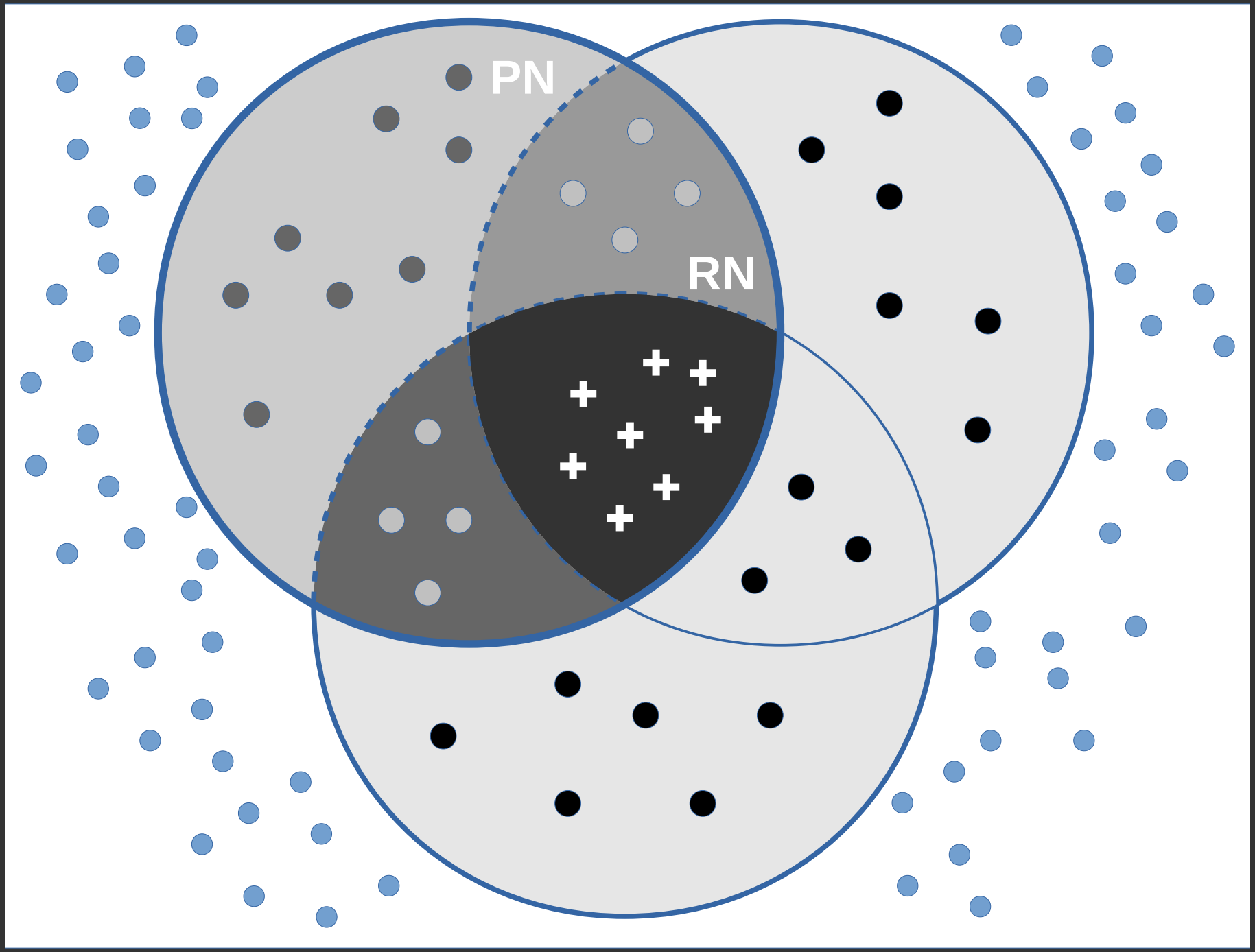
Morgane Barbet-Massin[1]*, Frédéric Jiguet[1], Cécile Hélène Albert[2,3] and Wilfried Thuiller[3]

[1]Muséum National d'Histoire Naturelle, UMR 7204 MNHN-CNRS-UPMC, Centre de Recherches sur la Biologie des Populations d'Oiseaux, CP 51, 55 Rue Buffon, 75005 Paris, France; [2]Department of Biology, McGill University, 1205 Docteur Penfield, Montréal, QC, Canada; and [3]Laboratoire d'Ecologie Alpine, UMR-CNRS 5553, Université Joseph Fourier, Grenoble I, BP 53, 38041 Grenoble Cedex 9, France

TYPES OF PSEUDO-ABSENCES

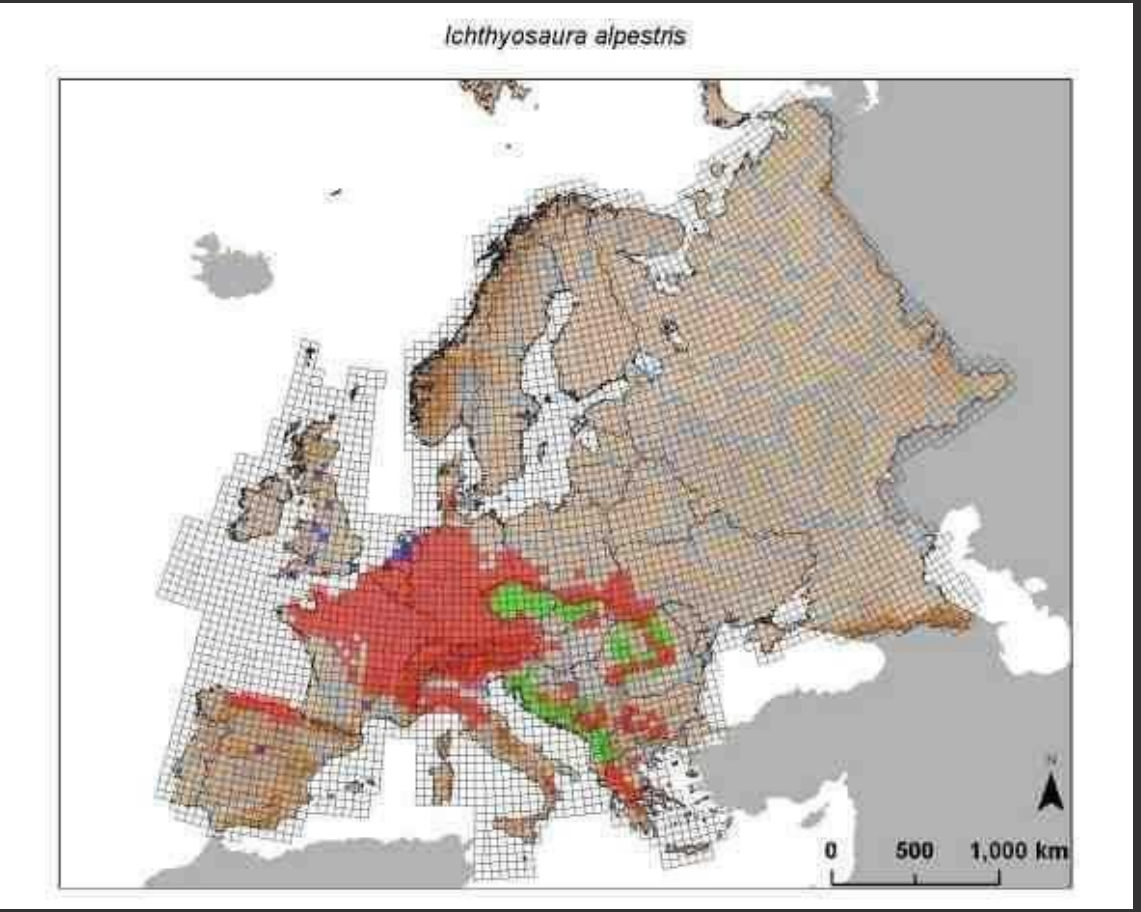
- **Allopatric species**: from sister species with an excluding distribution
- **Absence areas**: from verified climatically unsuitable areas.
- **Random**: random points all over the study area.
  - For regression models: **not stratified random selection**
  - For classification and machine learning algorithms: **stratified geographical or environmental random selection**

TYPES OF PSEUDO-ABSENCES

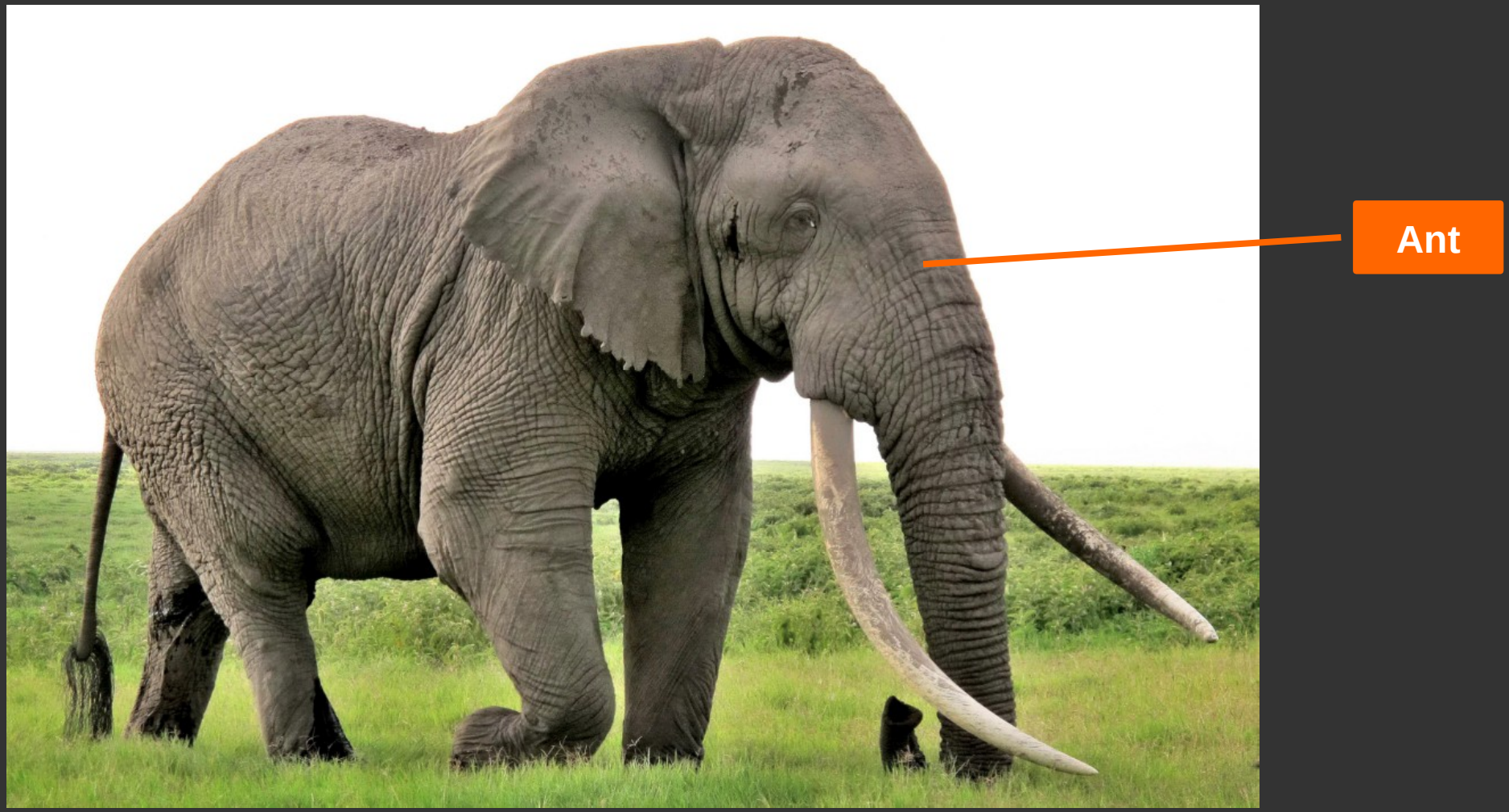


6. Check errors in the species' records.

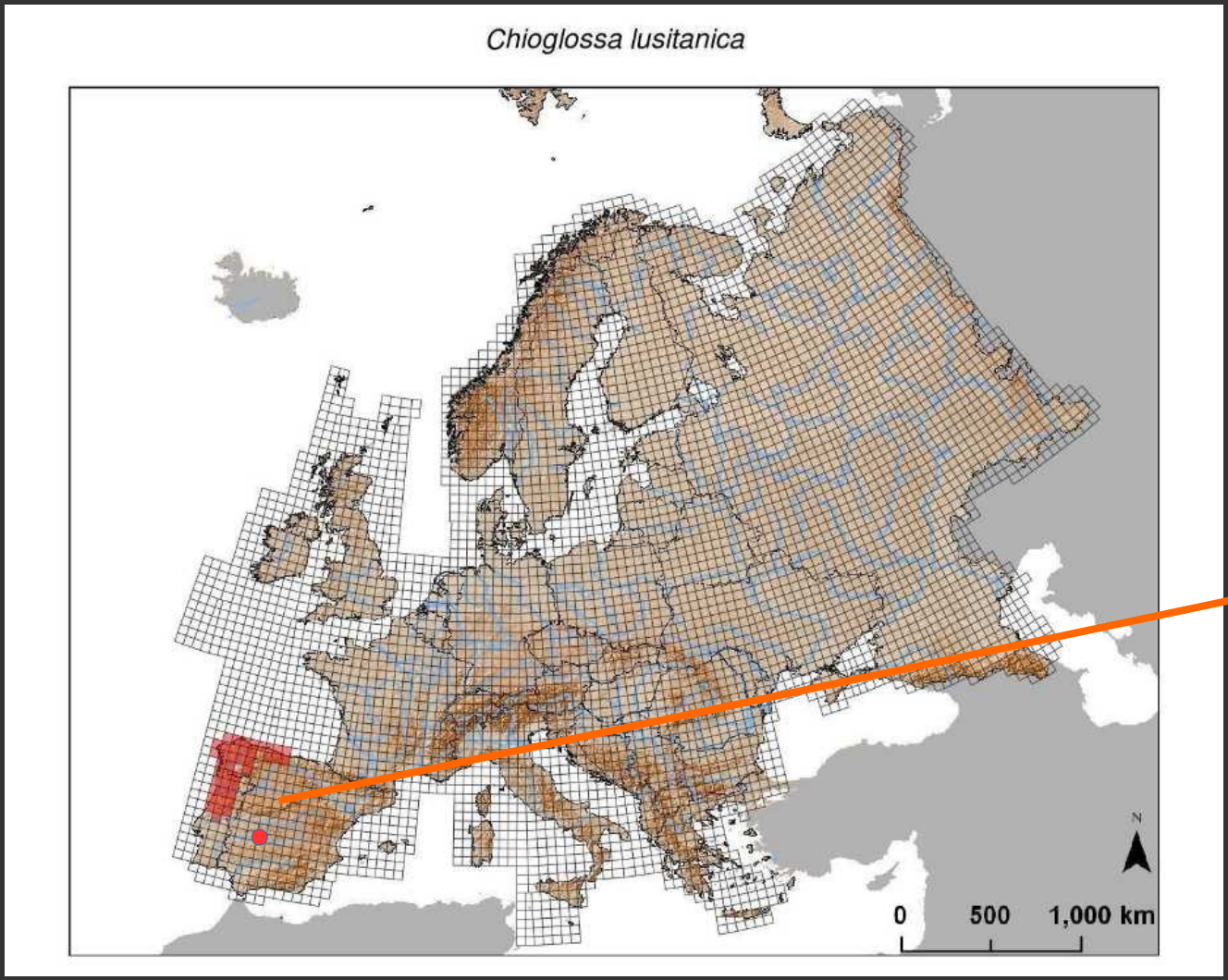
- Check for errors in the coordinates.
- Visualise always the species' records in a map in order to identify possible errors.



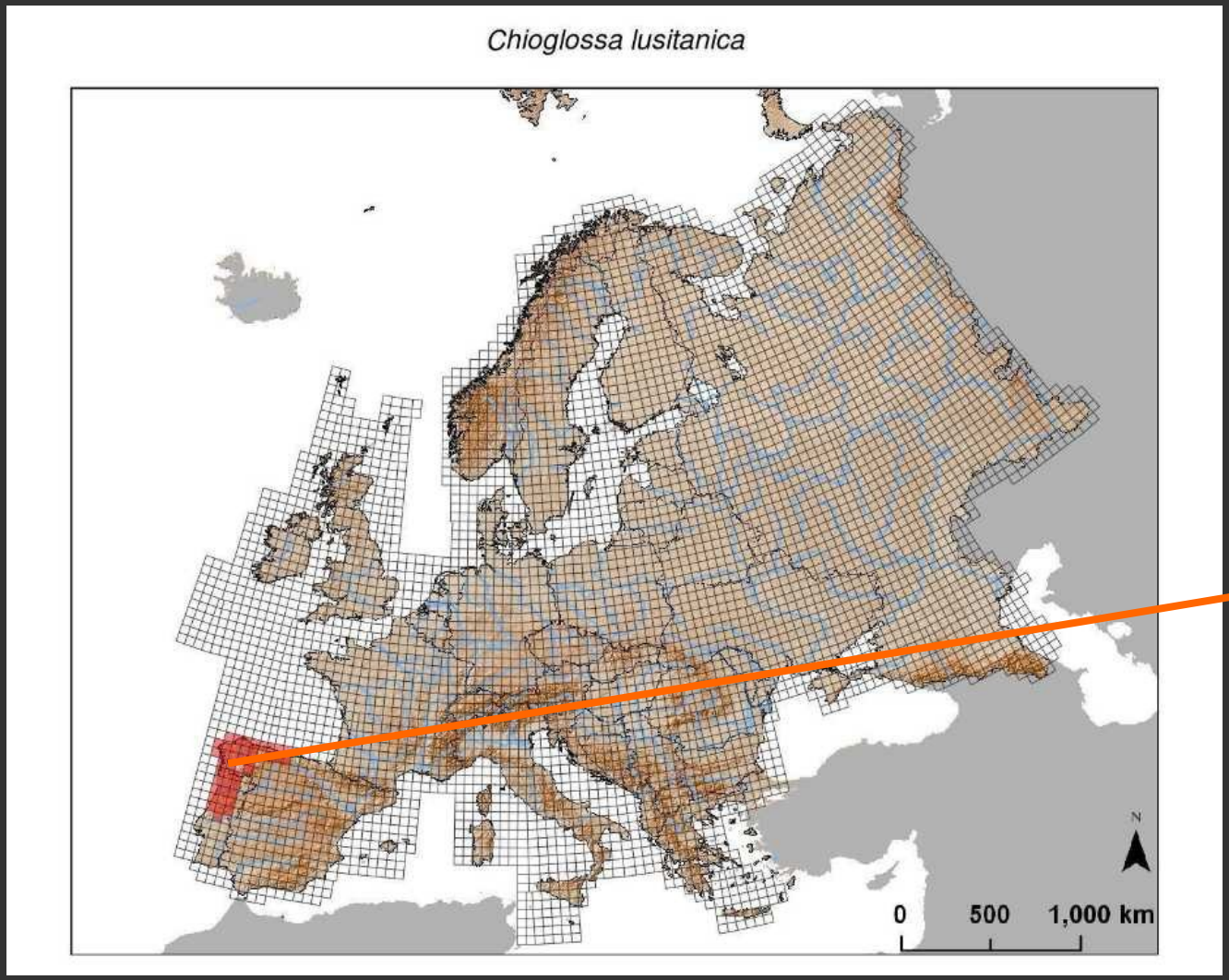
- **Species determination errors**: these errors are almost impossible to correct, at least if you are not familiar with the database.



- **Species determination errors**: these errors are almost impossible to correct, at least if you are not familiar with the database.



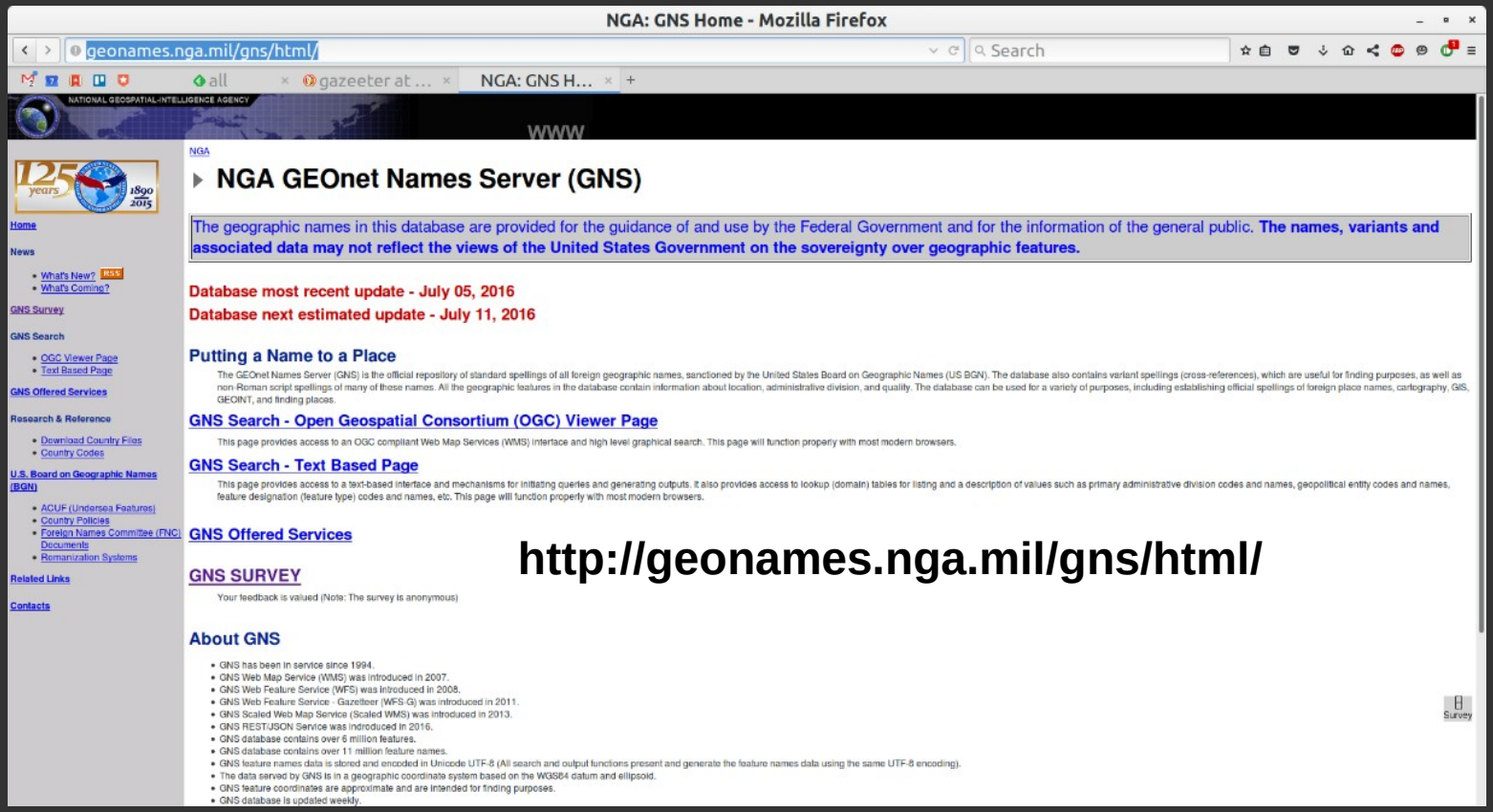
- **Species determination errors**: these errors are almost impossible to correct, at least if you are not familiar with the database.



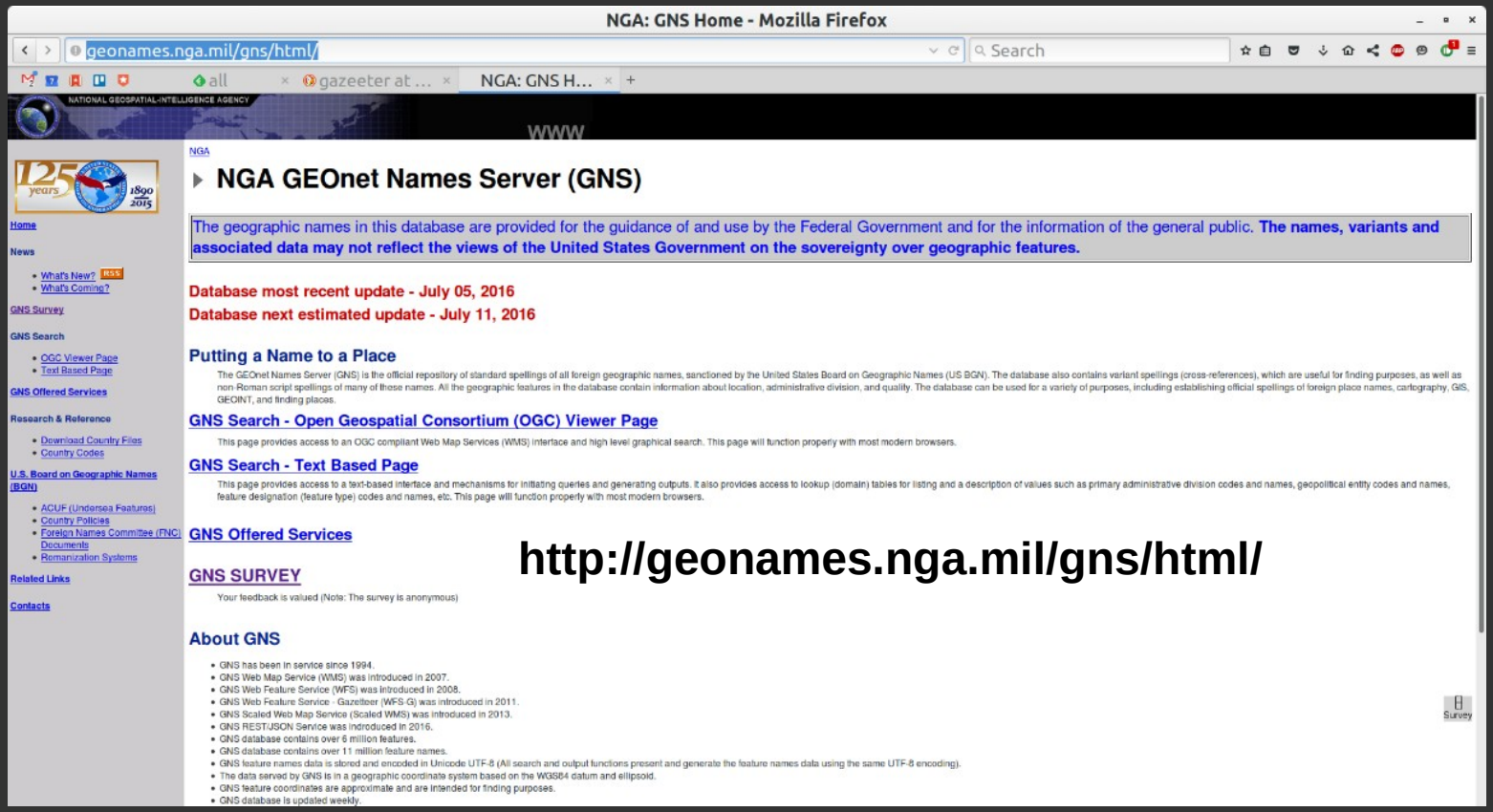
- **Records without coordinates**: you can associate coordinates to these records if you have some references to the localities.

  → You can use online gazetteer services to obtain the coordinates of the localities.

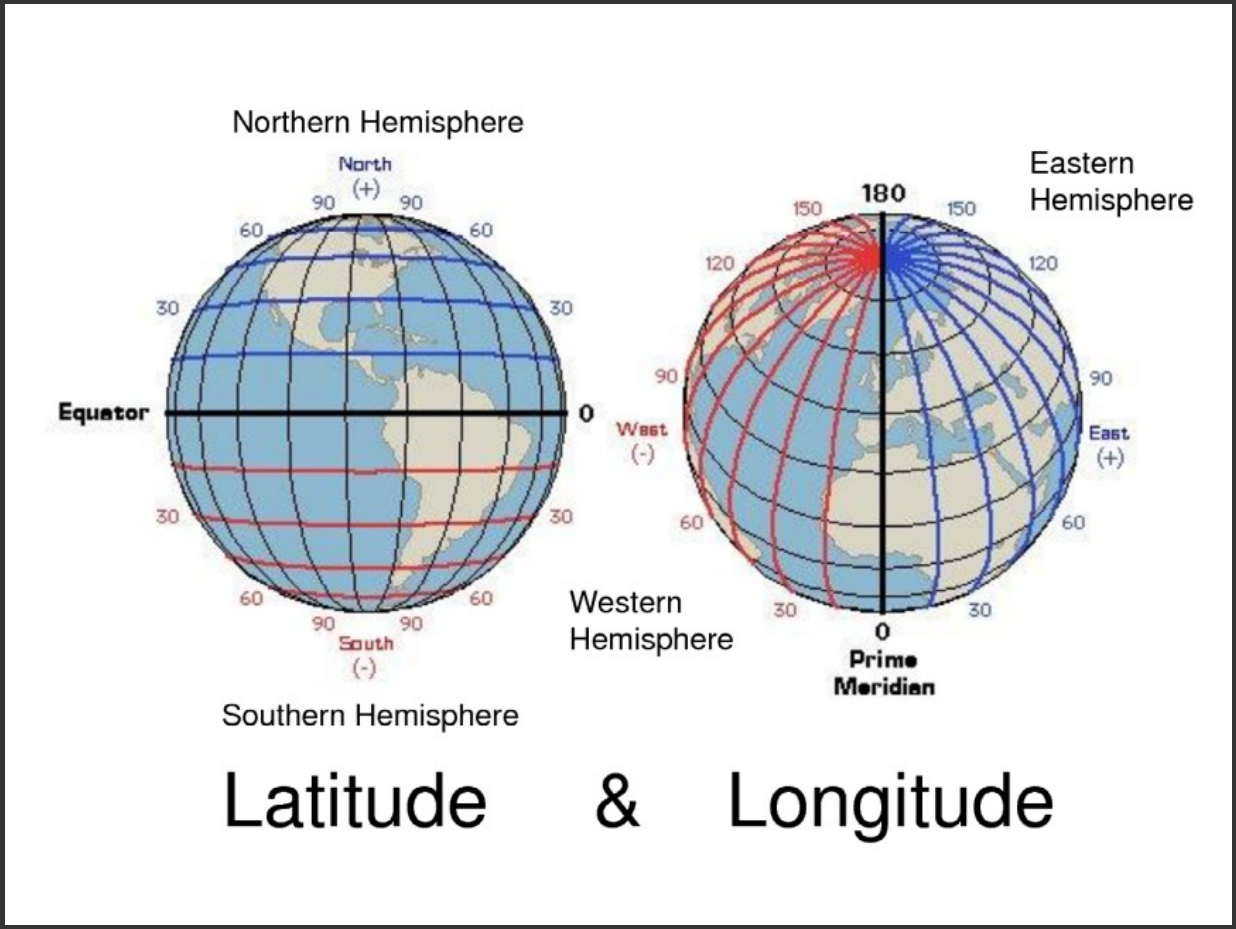
  → Remember that you must maintain a similar accuracy in the geo-referencing to the rest of the records.



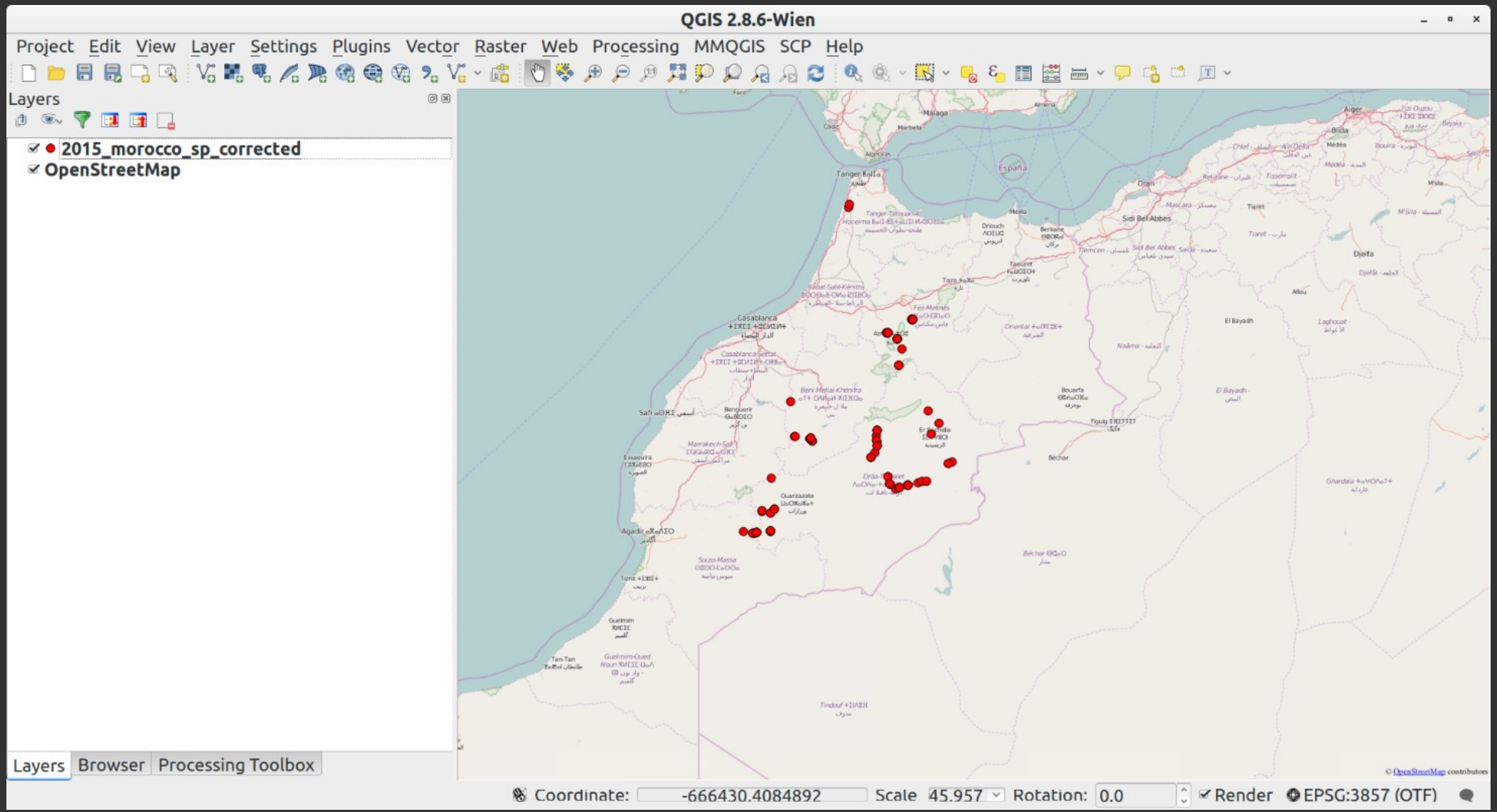
- **Records without one coordinate**: sometimes, records lack one of the coordinates or the value of one of the coordinates is zero.
  → You can correct this if you have some indications to the species' localities.
  → Use gazetteer services to associate coordinates to the localities.



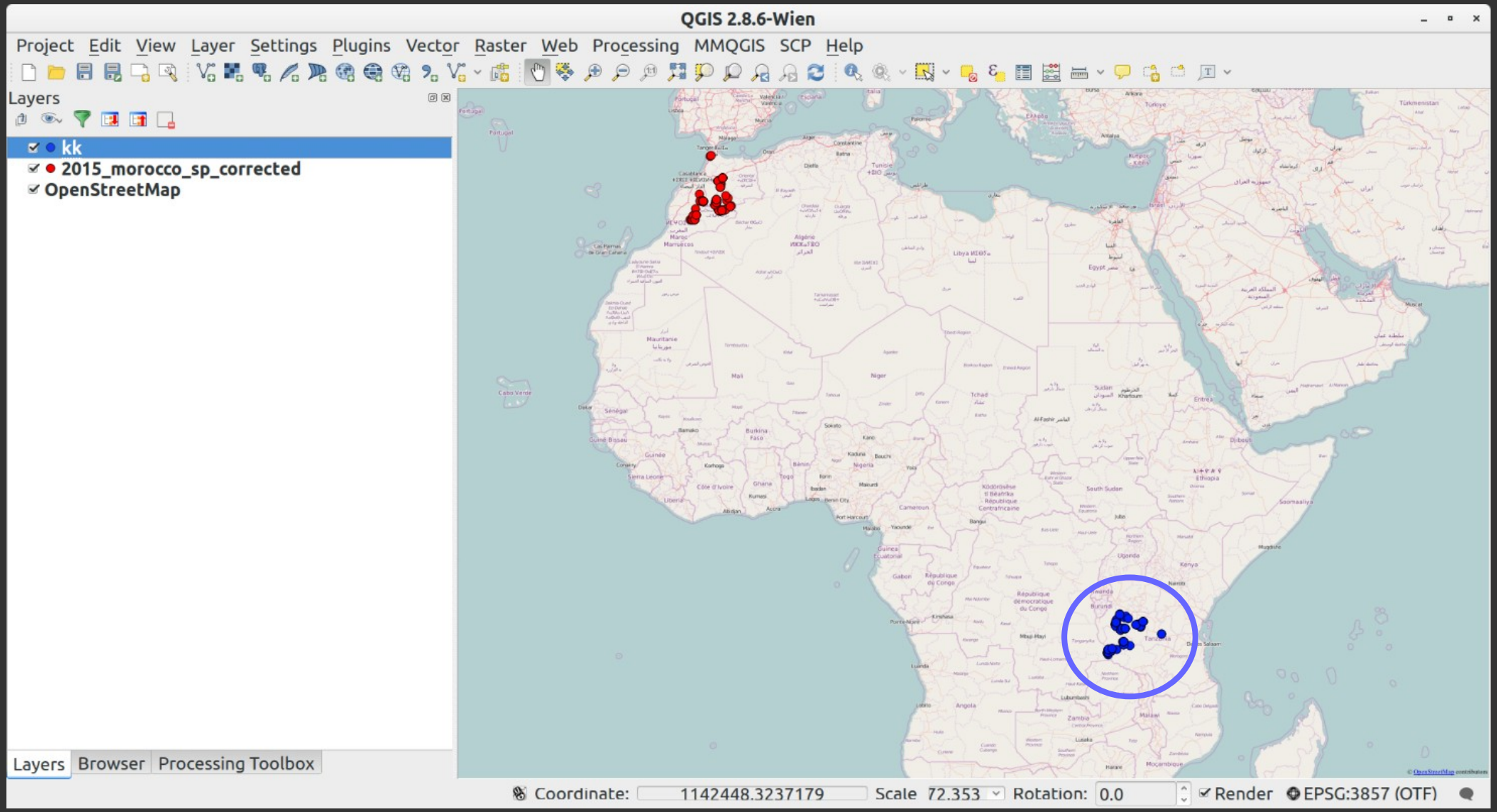
- **Errors in records' coordinates: the most common error is to associate the latitude to X coordinate and the longitude to the Y coordinate, producing an inverted distribution of the species. → Check if the X coordinate corresponds to the longitude and the Y coordinate to the latitude.**



COMMON ERRORS ON SPECIES' RECORDS



COMMON ERRORS ON SPECIES' RECORDS

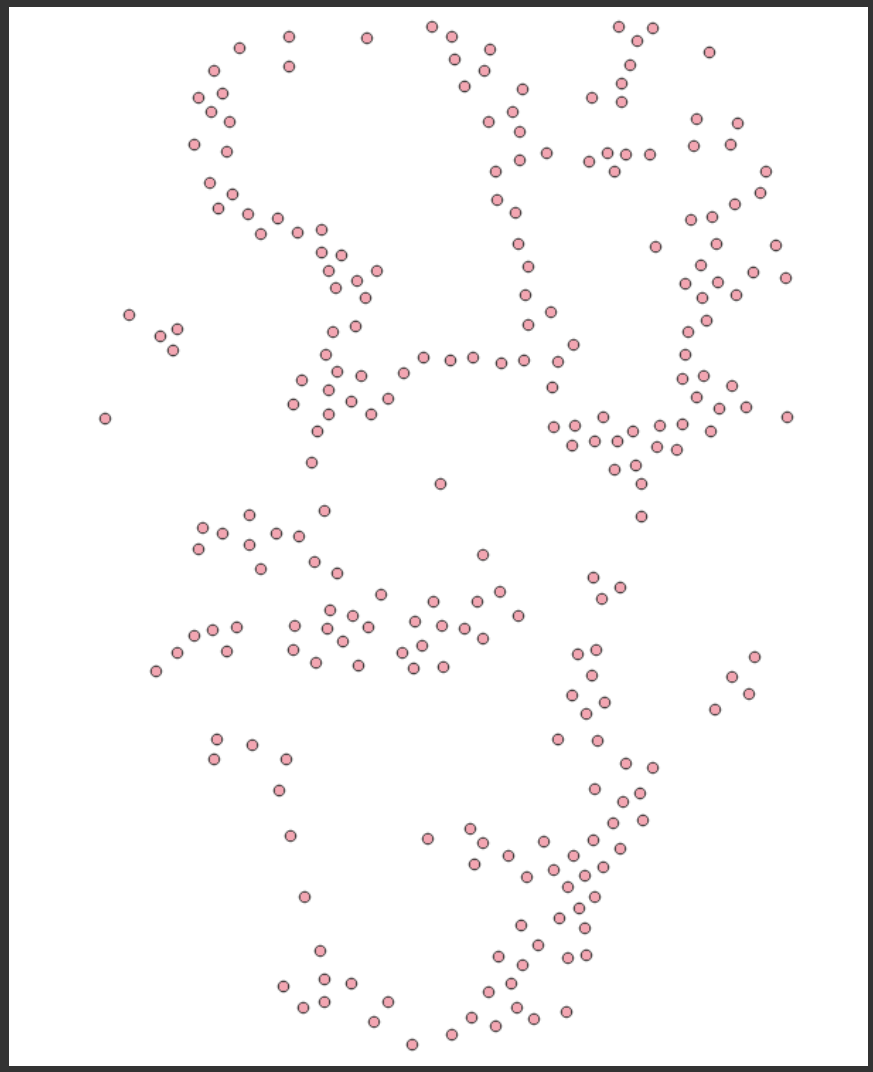


- **Duplicated errors: some modelling methods are sensible to the abundance per pixel of species records.**
  → If this abundance is consequence of biased sampling, you must delete the duplicated records.
  → Some modelling methods can do this automatically.

ID	Species	X	Y
1	*Podarcis bocagei*	-8.744444	41.352778
2	*Podarcis bocagei*	-8.744444	41.352778
3	*Podarcis bocagei*	-8.745678	41.353421
4	*Podarcis bocagei*	-8.766892	41.368239

7. Filter if necessary the species' records.

The intensity on the species point records must correspond to the species' observed distribution (Sillero et al. 2010).



AUTOCORRELATION ON SPECIES' RECORDS

Many people thinks that it is necessary to reduce the autocorrelation among the species point records → THIS IS WRONG!!

‘‘Given a set S containing n geographical units, spatial autocorrelation refers to the relationship between some variable observed in each of the n localities and a measure of geographical proximity defined for all n (n-1) pairs chosen from n’’.
[*In* Getis 2008]

Spatial correlation is a consequence of Tobler's First Law of Geography:

“Everything is related to everything else, but near things are more related than distant things”.

AUTOCORRELATION ON SPECIES' RECORDS

Spatial autocorrelation is a necessary property in ENM.

Without spatial autocorrelation
→ no relationship between the species distribution and the variable

We must no mistake spatial autocorrelation with intensity (density).

Intensity: Mean number of events per unit area.

- No spatial autocorrelation between training and test data.
- Both sets of data must be independent.

Plant Biosystems, Vol. 146, No. 4, December 2012, pp. 789–796

Taylor & Francis
Taylor & Francis Group

NEW TRENDS IN BIODIVERSITY INFORMATICS

Integrating fundamental concepts of ecology, biogeography, and sampling into effective ecological niche modeling and species distribution modeling

A. T. PETERSON & J. SOBERÓN

Biodiversity Institute, The University of Kansas, 1345 Jayhawk Blvd., Lawrence 66045, USA

If you have clusters, you must reduce the clusters.

Nearest Neighbour Index (Clark and Evans 1954) → until you obtain a random distribution of points

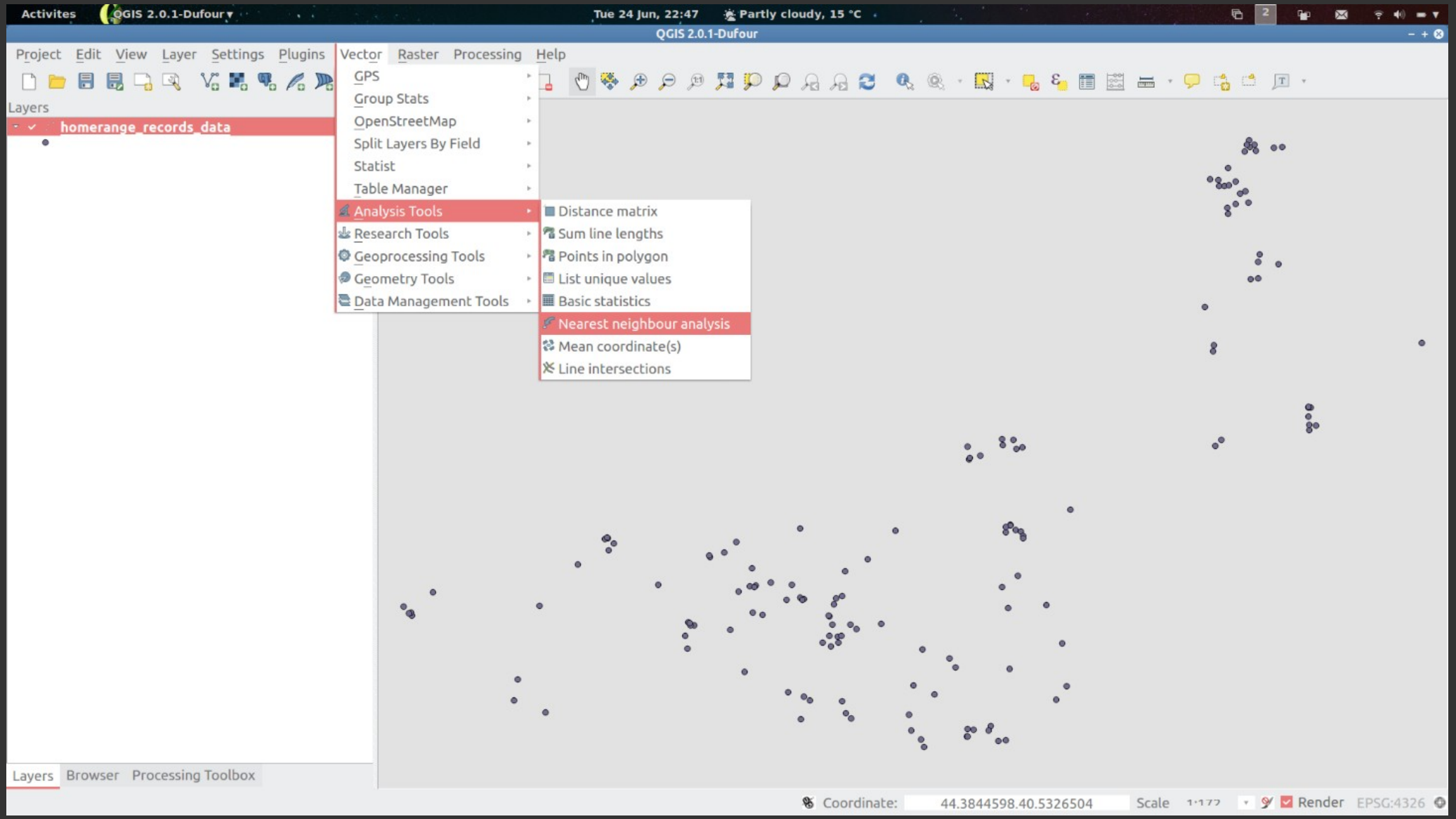
$$E_{(d_i)} = \left[\left(0.5 * \sqrt{\frac{A}{N}} \right) + \left(0.0514 + \frac{0.041}{\sqrt{N}} \right) * \frac{B}{N} \right]$$

E>1 suggests ordering; E<1 suggests clustering

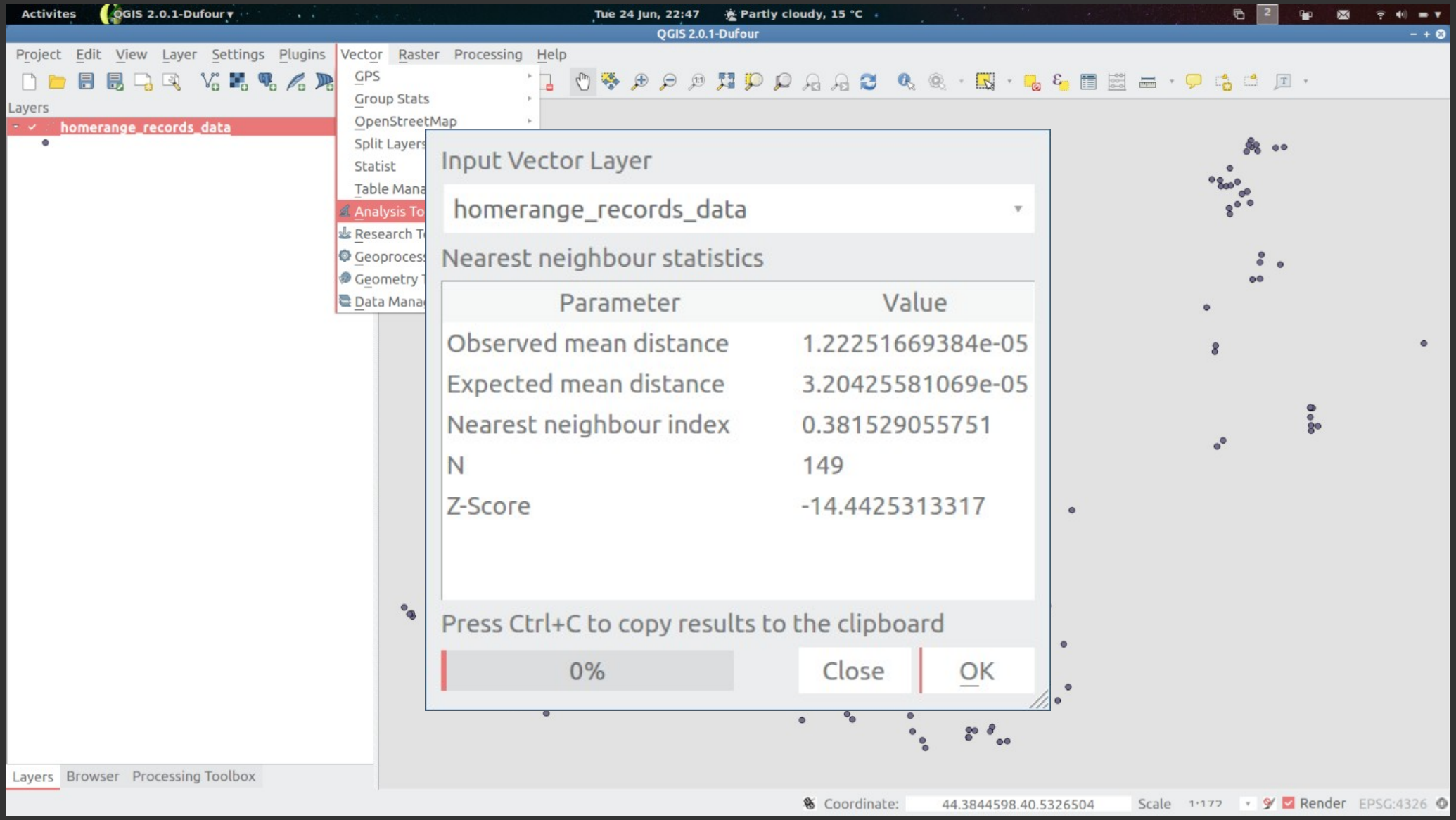
Local Indicators of Spatial Association (LISA; Anselin 1995) → for selecting the points to be cleaned

$$I_i = n(y_i - \bar{y}) \sum_{j \neq i} w_{ij} (y_j - \hat{y})$$

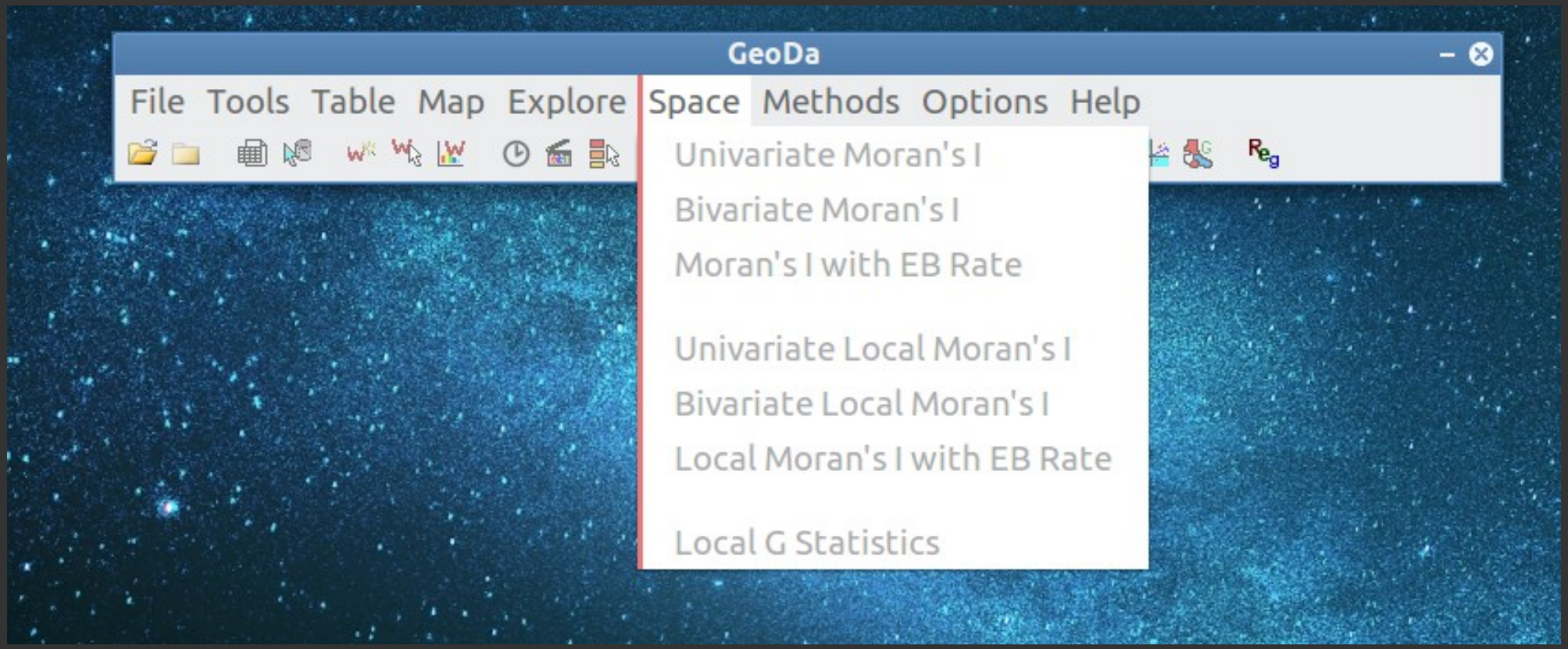
Nearest Neighbour Index



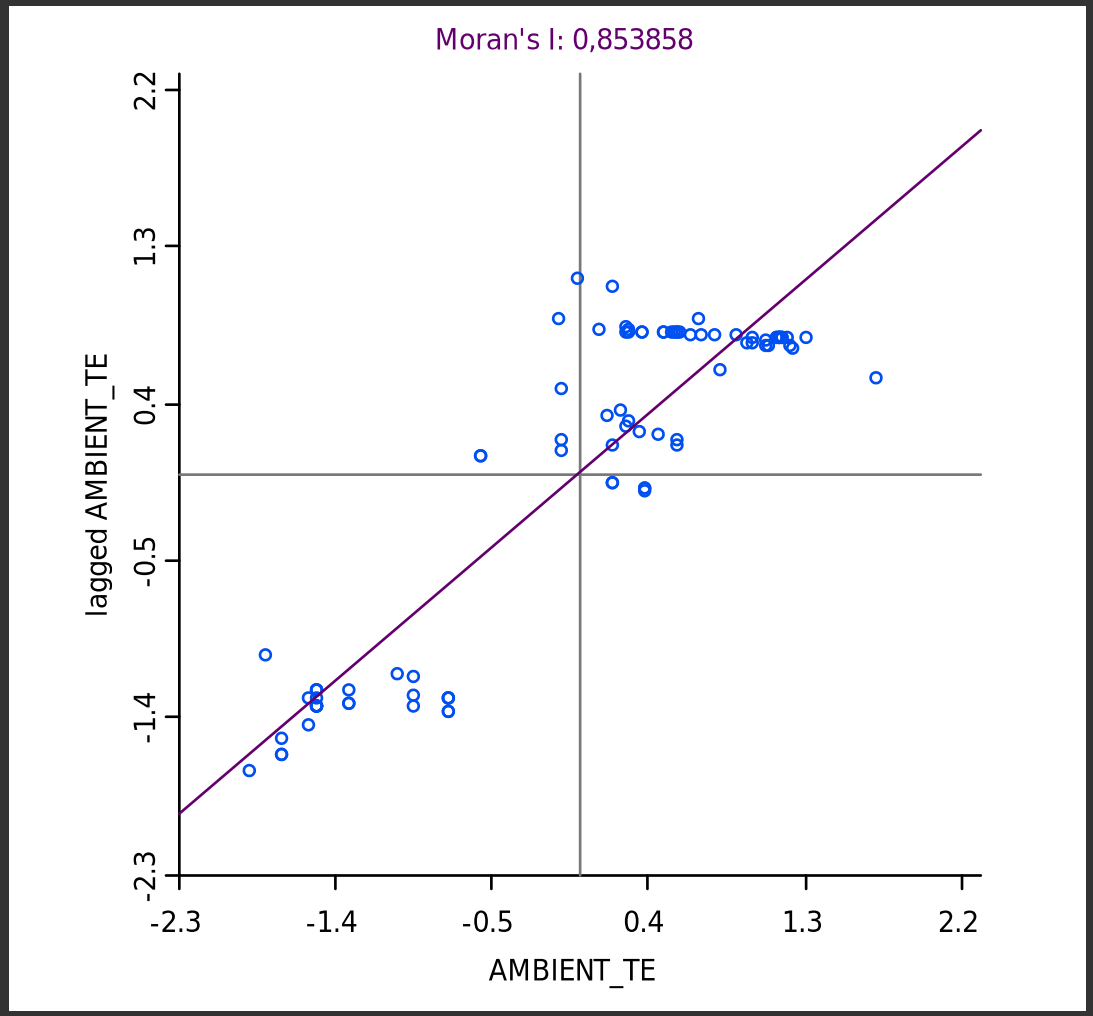
Nearest Neighbour Index

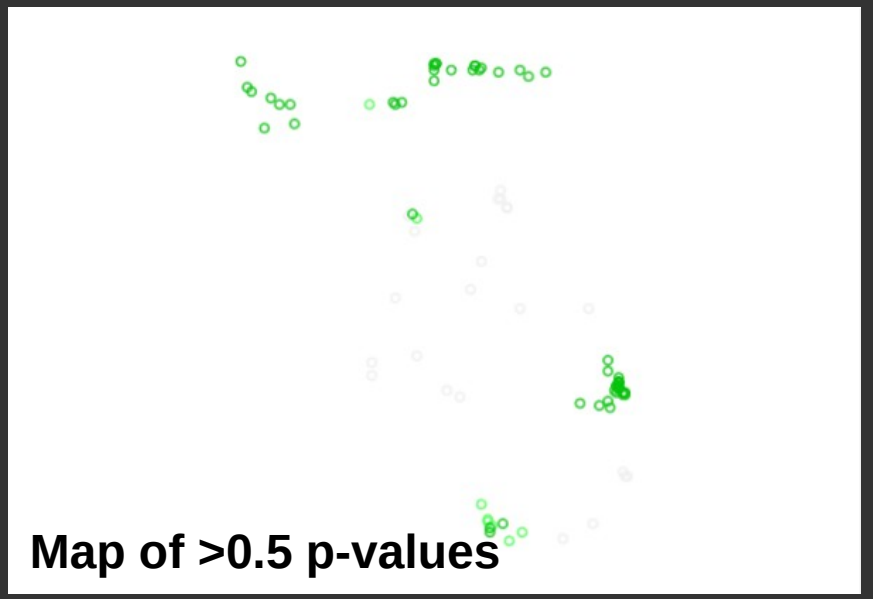


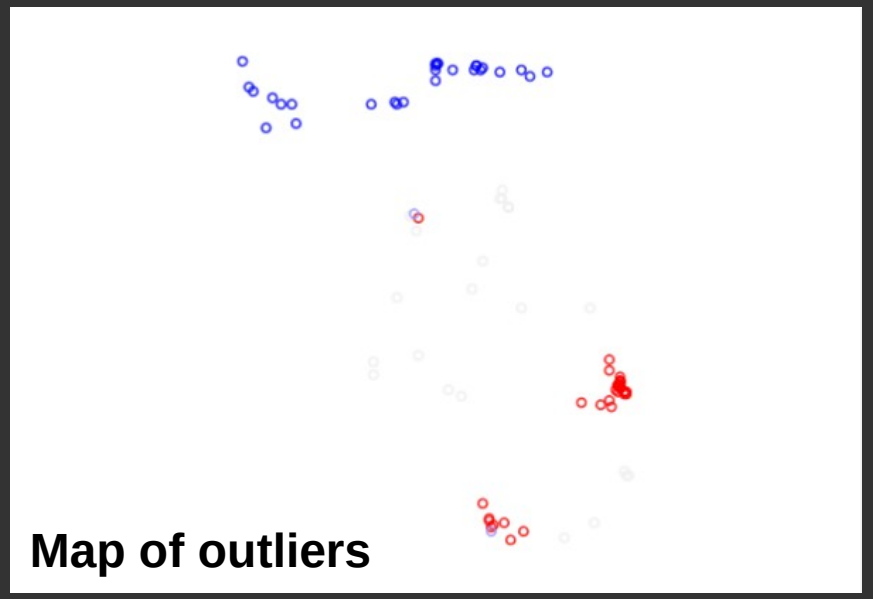
Local Indicators of Spatial Association



Local Indicators of Spatial Association







The intensity on the species point records must correspond to the species' observed distribution.

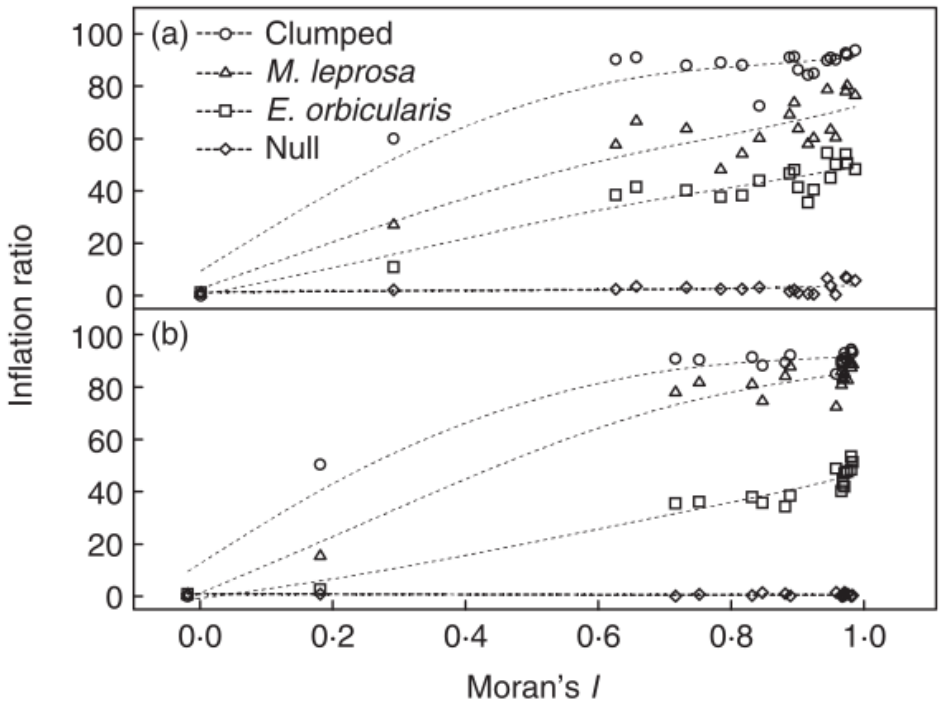


Fig. 5. Variation of the inflation ratio (number of times a significant test was found in relation to the expected number of significant tests, according to the significance level adopted, in this case $P < 0{\cdot}01$, with the Moran's I-values of the original variables). Dashed lines are spline fits. (a) Eastern rectangle; (b) western rectangle.

Segurado et al 2006

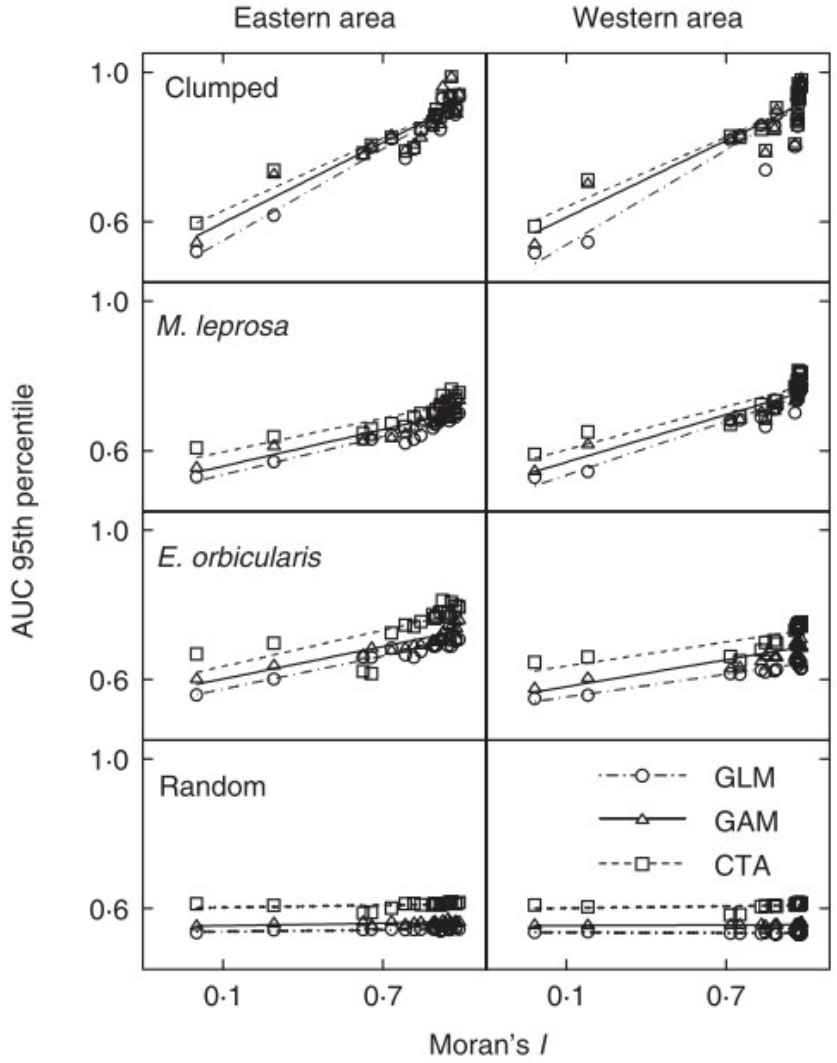


Fig. 2. Variation of the 95th percentile of AUC with Moran's I of environmental variables using three modelling techniques (GLM, GAM and CTA), four distributions (*M. leprosa*, *E. orbicularis*, a simulated distribution with totally clumped occurrences and a random distribution) and 1000 simulated surfaces (lines represent linear fits).

SPECIES' RECORDS SAMPLE SIZE

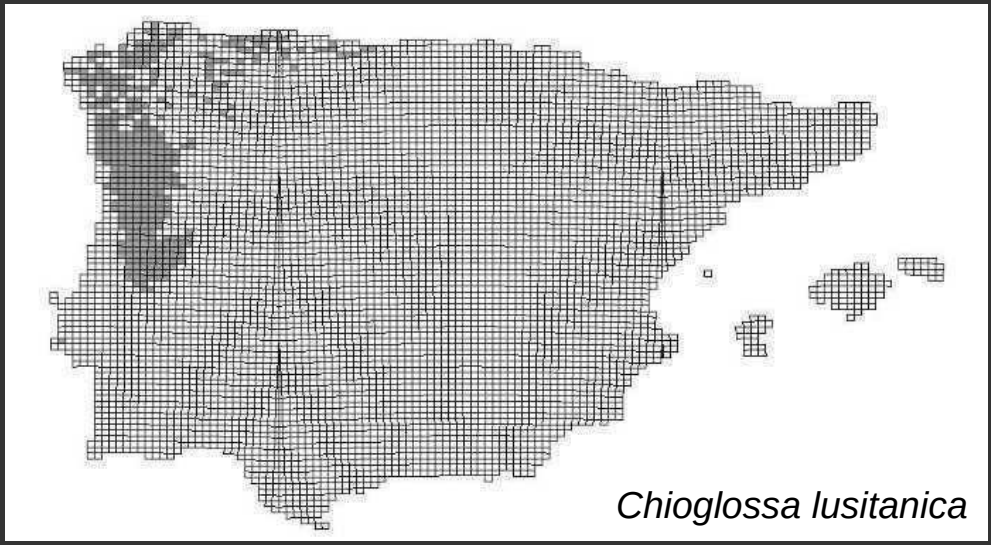
There is no rule to know which is the **minimum sample size**.

The sample size depends on the distribution of the species:

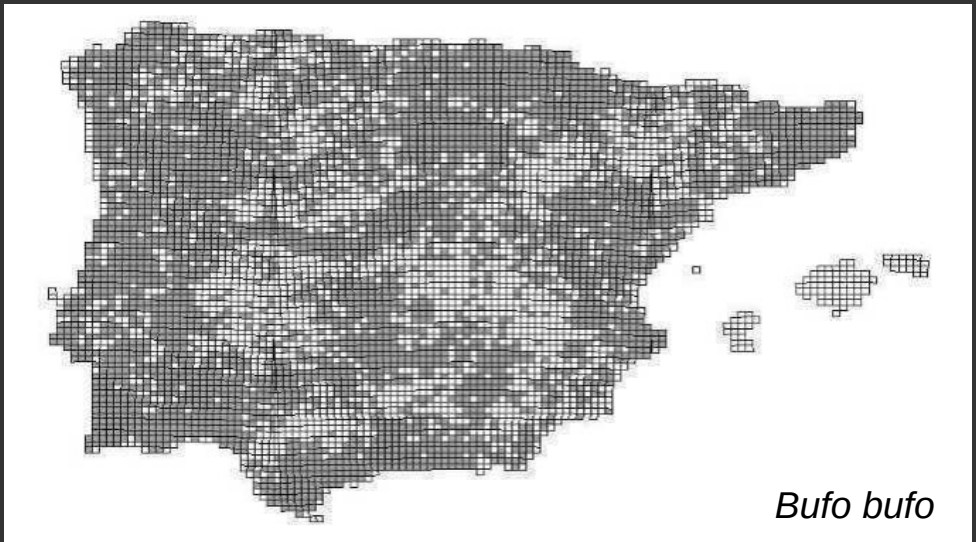
Widespread species or it is distributed across a large environmental area → you will need more samples

Specialised species, and it occurs in a small range → it will be easier to be modelled and you will need less records

Maxent is a good modelling technique for small samples.



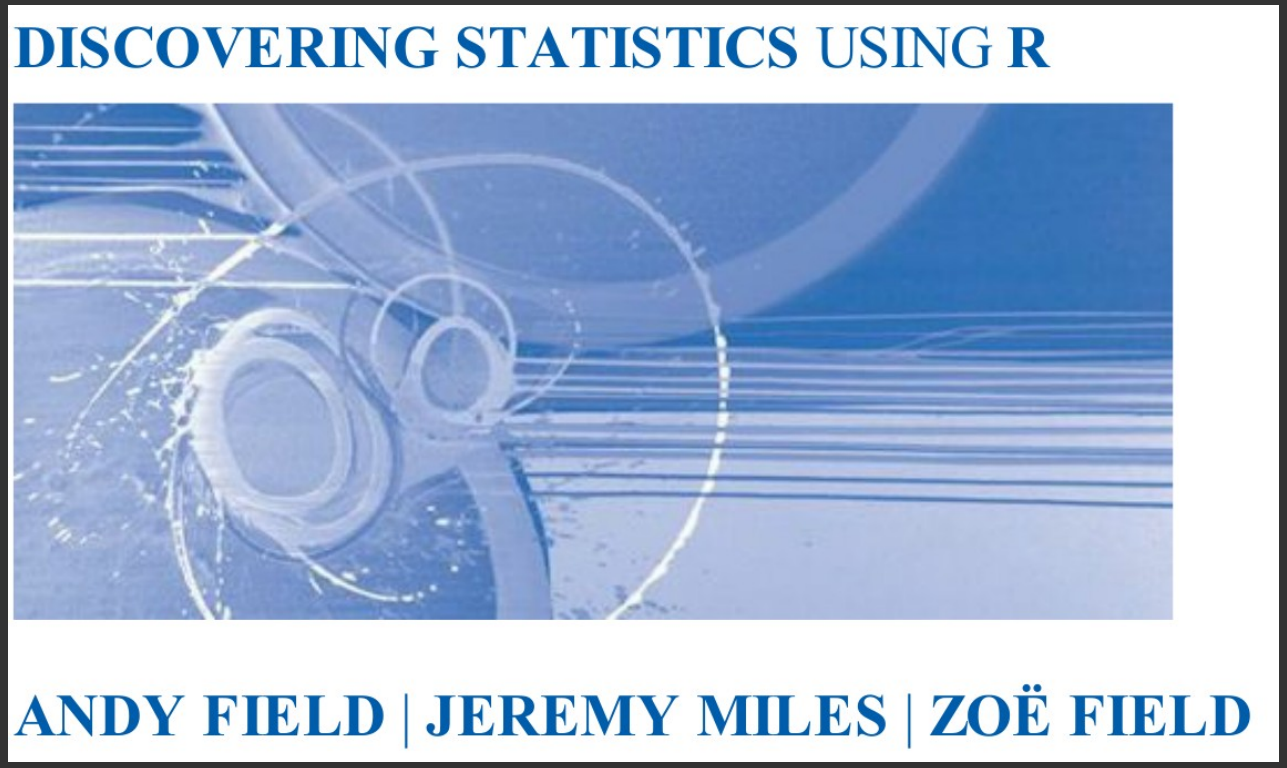
Chioglossa lusitanica



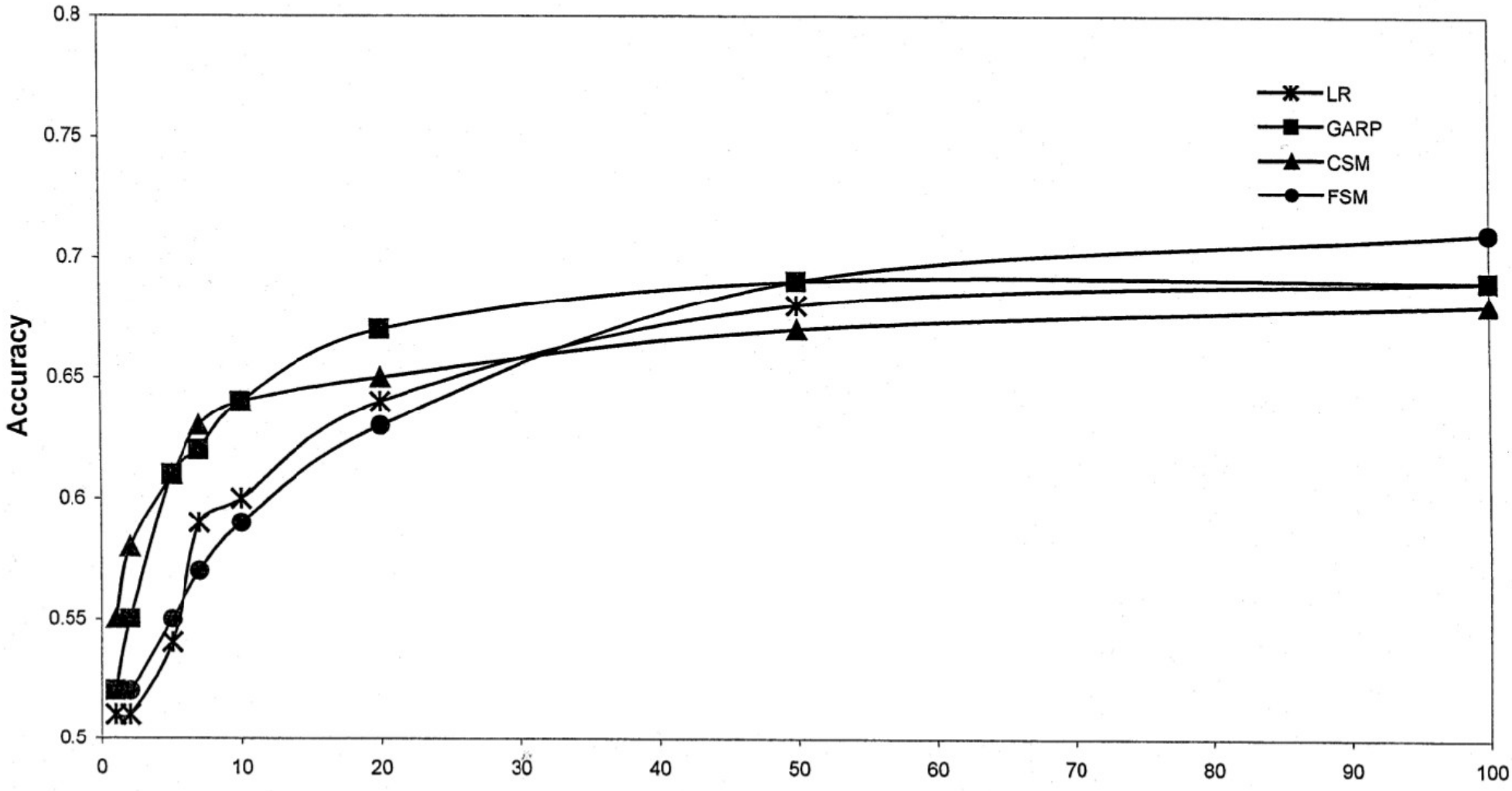
Bufo bufo

→ In GLMs:

- number of records: 50 + 8k
- number of records: 104 + k
  - k is the number of variables



The accuracy of the models increase with the sample size until the asymptote is reached.



Stockwell & Peterson 2002

8. Choose carefully the variables you need for calculating the models.

- The variables must be meaningful for the species and must have a spatial resolution according with your objectives.
- You can use continuous or categorical datasets depending on the modelling methods.

  → Not all methods can deal with categorical datasets.

Table 3.1. *Spatial scales of environmental resources and response scale of biological organisms in a hierarchical framework linking species distributional behavior to models of their spatial distribution*

Environment	Organisms
Global (temperature)	Species
Meso (geology)	Species
Topography	Population
Micro (canopy/gap)	Group
Nano (soil patches)	Individual

From Mackey and Lindenmayer, 2001.

8. Choose carefully the variables you need for calculating the models.

- The variables must be meaningful for the species and must have a spatial resolution according with your objectives.
- You can use continuous or categorical datasets depending on the modelling methods.

  → Not all methods can deal with categorical datasets.

- Exclude variables Bio 3, Bio 14, Bio 15 → Low correlation between current and future variables.

RESEARCH ARTICLE

A Short Guide to the Climatic Variables of the Last Glacial Maximum for Biogeographers

Sara Varela[1,2]*, Matheus S. Lima-Ribeiro[3], Levi Carina Terribile[3]

1 Department of Ecology, Faculty of Science, Charles University, Praha, Czech Republic, **2** Museum für Naturkunde. Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany, **3** Departamento de Ciências Biológicas, Universidade Federal de Goiás–UFG, Jataí, GO, Brazil

* svarela@paleobiogeography.org

Global and Planetary Change 107 (2013) 1–12

Contents lists available at SciVerse ScienceDirect

Global and Planetary Change

journal homepage: www.elsevier.com/locate/gloplacha

Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections

Joaquín Bedia [a,*], Sixto Herrera [b], José Manuel Gutiérrez [a]

[a] Instituto de Física de Cantabria, Universidad de Cantabria-CSIC, 39005 Santander, Spain
[b] Predictia Intelligent Data Solutions, S.L. CDTUC Fase A, Planta 2-203, Avda. los Castros s/n, 39005 Santander, Spain

9. Obtain the variables for calculating the models.

- Several global datasets available in the Internet at different spatial resolutions (Sillero and Tarroso 2010).
- The most famous is WorldClim (Hijmans et al. 2005).
- Remote sensing as data source (Sillero et al. 2009, 2012).

10. Export the variables to the format needed by the modelling method programme.

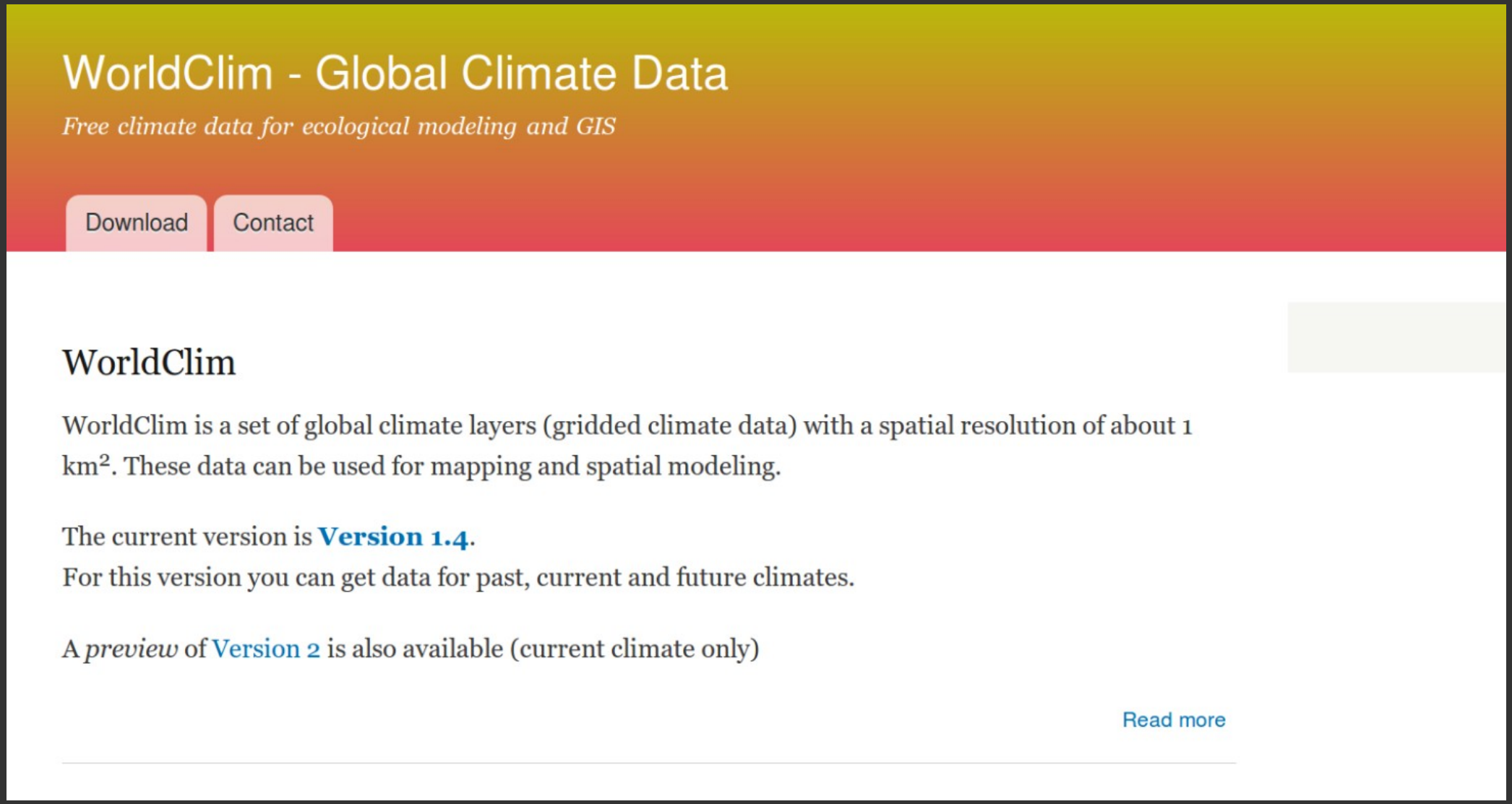
ELEVATION

SLOPE or any other topographical index

- Calculate SLOPE in a metrical projected system
- It is not possible to calculate the slope in a degree geographical system
- Check if you want slope in degrees or percentage

ASPECT OR CARDINAL ORIENTATION

- Calculate ASPECT
- Change 360º to 0º
- 0º = 360º = NORTH orientation



www.worldclim.org

Home About Bioclim **Last Glacial Maximum Climate** Timeseries **Future (CMIP5)**

Downloads Known issues CHELSAcruts (1901-2016)

CHELSA

Climatologies at high resolution for the earth's land surface areas

CHELSA – Free climate data at high resolution

http://chelsa-climate.org

High-resolution gridded datasets (and derived products)

This page contains four sections:

- Current Datasets and Static Climatologies
- Legacy Datasets
- Superseded Datasets
- General Information

Current Datasets and Static Climatologies	
The CRU TS dataset was developed and has been subsequently updated, improved and maintained with support from a number of funders, principally the UK's Natural Environment Research Council (NERC) and the US Department of Energy. Long-term support is currently provided by the UK National Centre for Atmospheric Science (NCAS), a NERC collaborative centre. **CRU gratefully acknowledges the support of all these funding agencies.** ***Always read the relevant documentation and publications***	
CRU TS v. 4.01 The current version of CRU TS, using the new process introduced with version 4.00, which it supersedes. Access at BADC If BADC is down, there is a Local Copy Google Earth Interface (TMP and PRE only)	A gridded time-series dataset This version, released 20 September 2017, covers the period 1901-2016 *Dataset DOI:* http://doi.org/10/gcmcz3 *Coverage:* All land areas (excluding Antarctica) at 0.5° resolution *Variables:* pre, tmp, tmx, tmn, dtr, vap, cld, wet, frs, pet Note that the following reference only partly applies to v4.01 (please read the Release Notes) *Reference:* Harris et al. (2014) doi:10.1002/joc.3711 (click to access) *Correction to the above paper:* Revised Appendix 3 (CLD)

https://crudata.uea.ac.uk/cru/data/hrg/

frontiers in Ecology and Evolution

DATA REPORT
published: 04 April 2018
doi: 10.3389/fevo.2018.00031

Check for updates

The EuMedClim Database: Yearly Climate Data (1901–2014) of 1 km Resolution Grids for Europe and the Mediterranean Basin

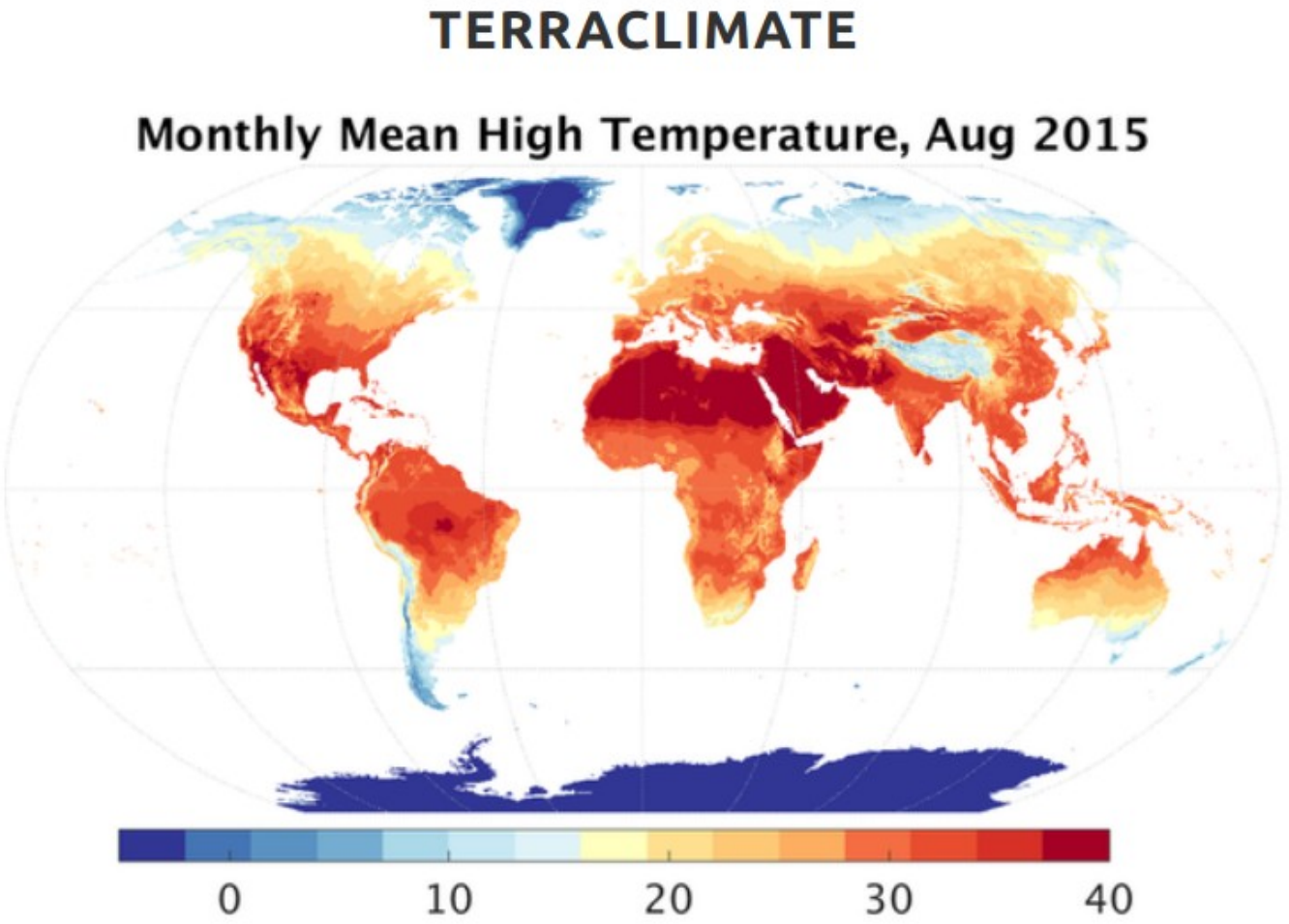
Thibaut Fréjaville and Marta Benito Garzón*

Biodiversité Gènes et Communautés (UMR 1202), Institut National de la Recherche Agronomique, Université de Bordeaux, Pessac, France

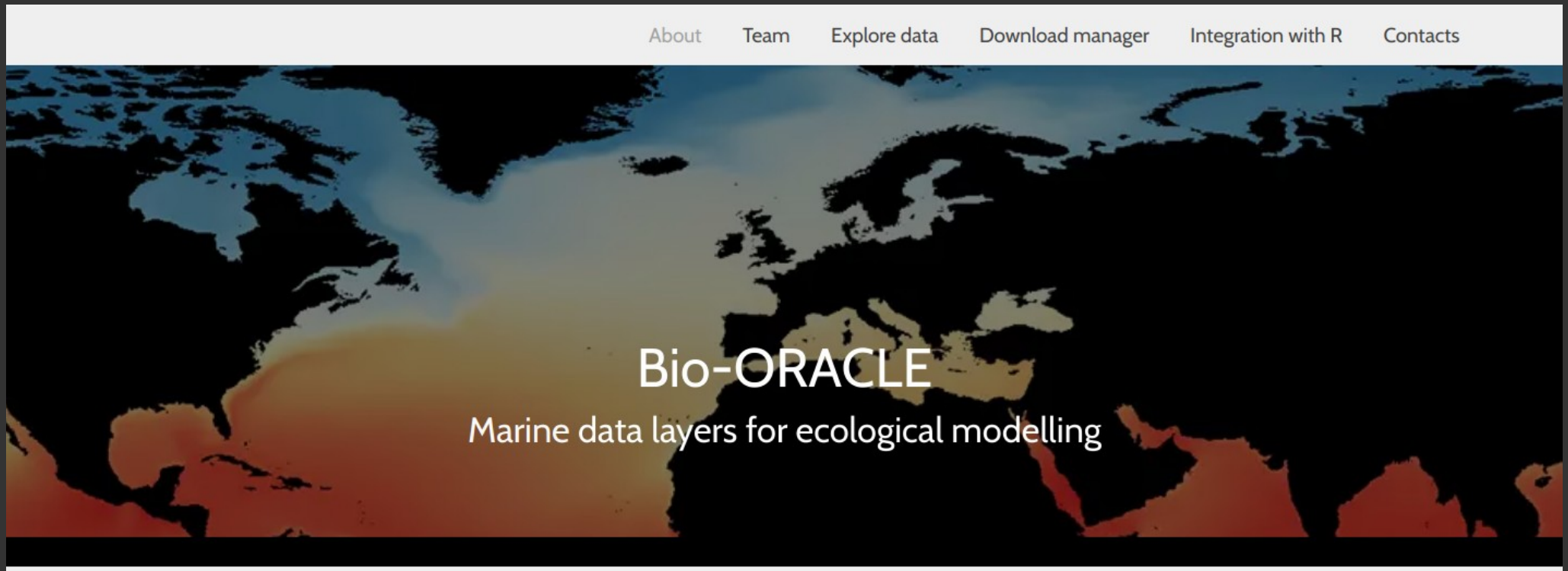
Keywords: anomaly, bioclim, climatic extremes, CRU, interpolation, precipitation, temperature, WorldClim

http://gentree.data.inra.fr/climate/

About | Download | Updates | References



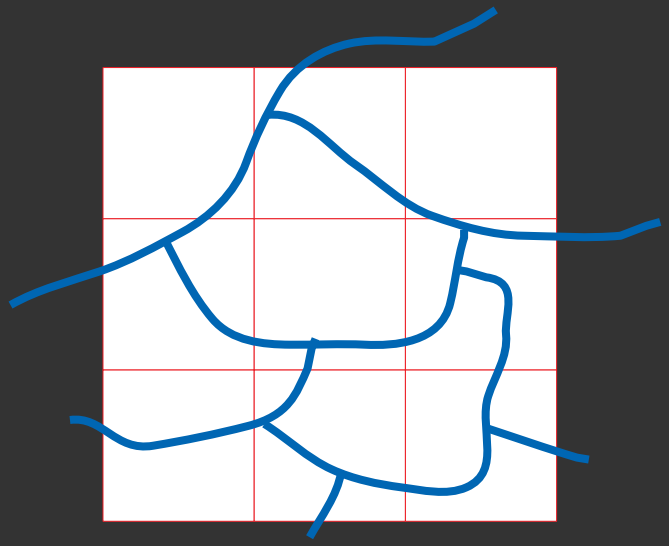
www.climatologylab.org/terraclimate.html

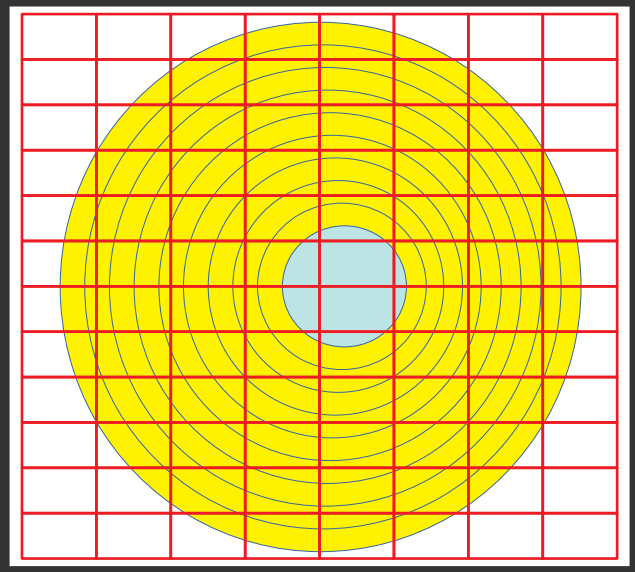


http://bio-oracle.org

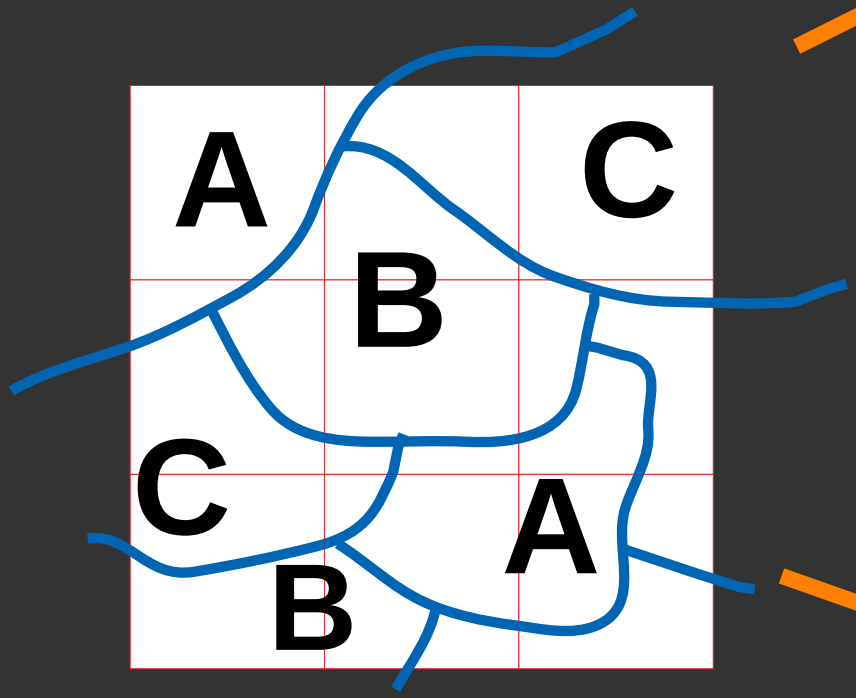
HOW TO USE CATEGORICAL VARIABLES

- Many modelling methods can use categorical variables
- Interpretation is more difficult
- The best solution is to create continuous variables from the categorical ones
- Export each class to a different layer
- Measure the area occupied by each class per sampling unit (pixel)
- Calculate distances of each pixel to each class

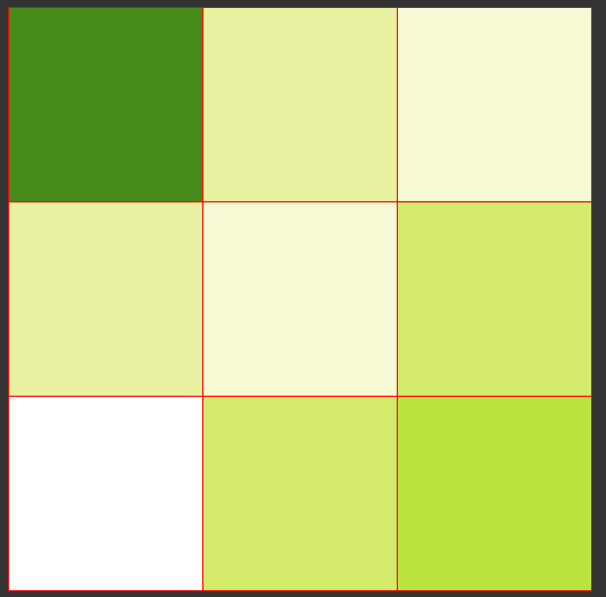




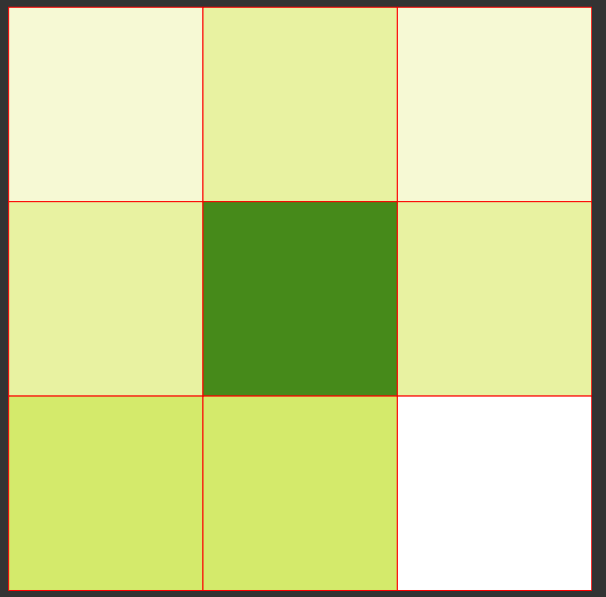
HOW TO USE CATEGORICAL VARIABLES



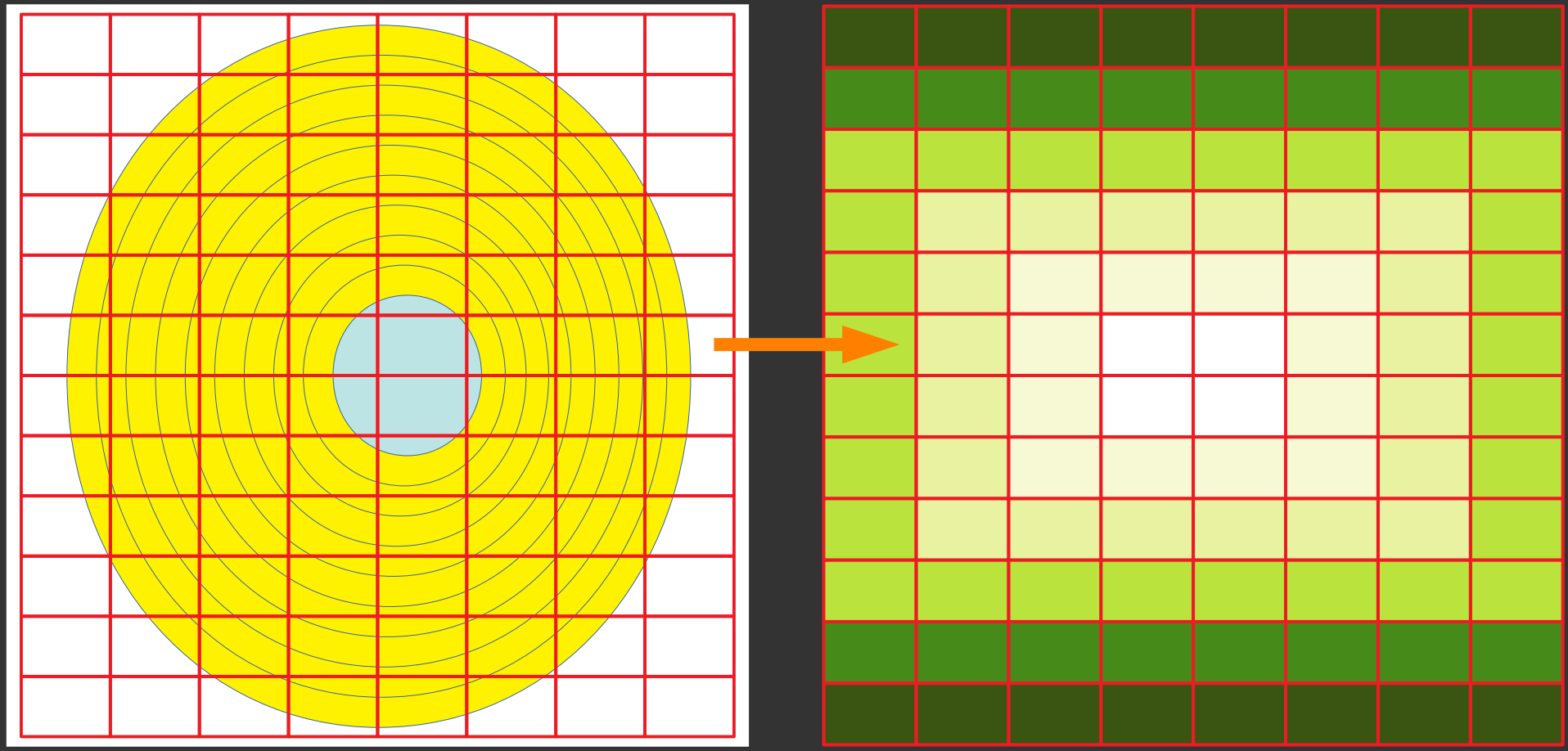
A



B



HOW TO USE CATEGORICAL VARIABLES



11. Choose the study area where you will calculate the models.

- **Make a shapefile of the study area.**

The size and form of the study area may affect the models' outputs but currently it is not completely understood.

→ Guisan & Thuiller 2005 recommend to increase the study area if the response curves of the variables are truncated.

→ Sillero 2010 recommend to model always using coherent biogeographical regions as study areas.

→ Anderson & Raza 2010 recommend to exclude those areas where the species cannot disperse.

12. Clip all the variables according to the chosen study area.

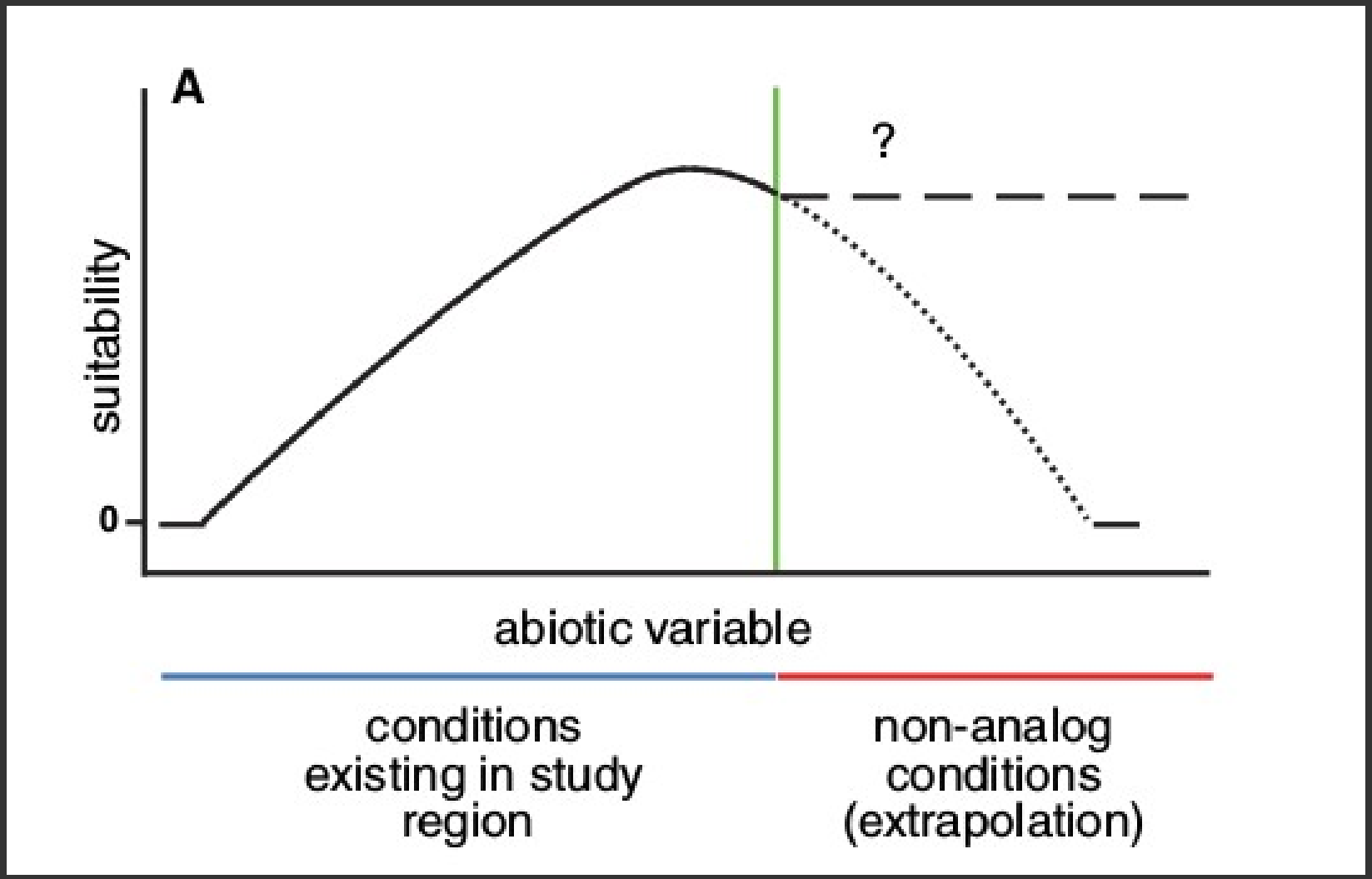
13. Check that all the variables have the same pixel size and shape.

14. Check that variables and species have the same projection system.

- If not, consider to project the species' records to the same projections system of the variables
  → it is easier in this direction

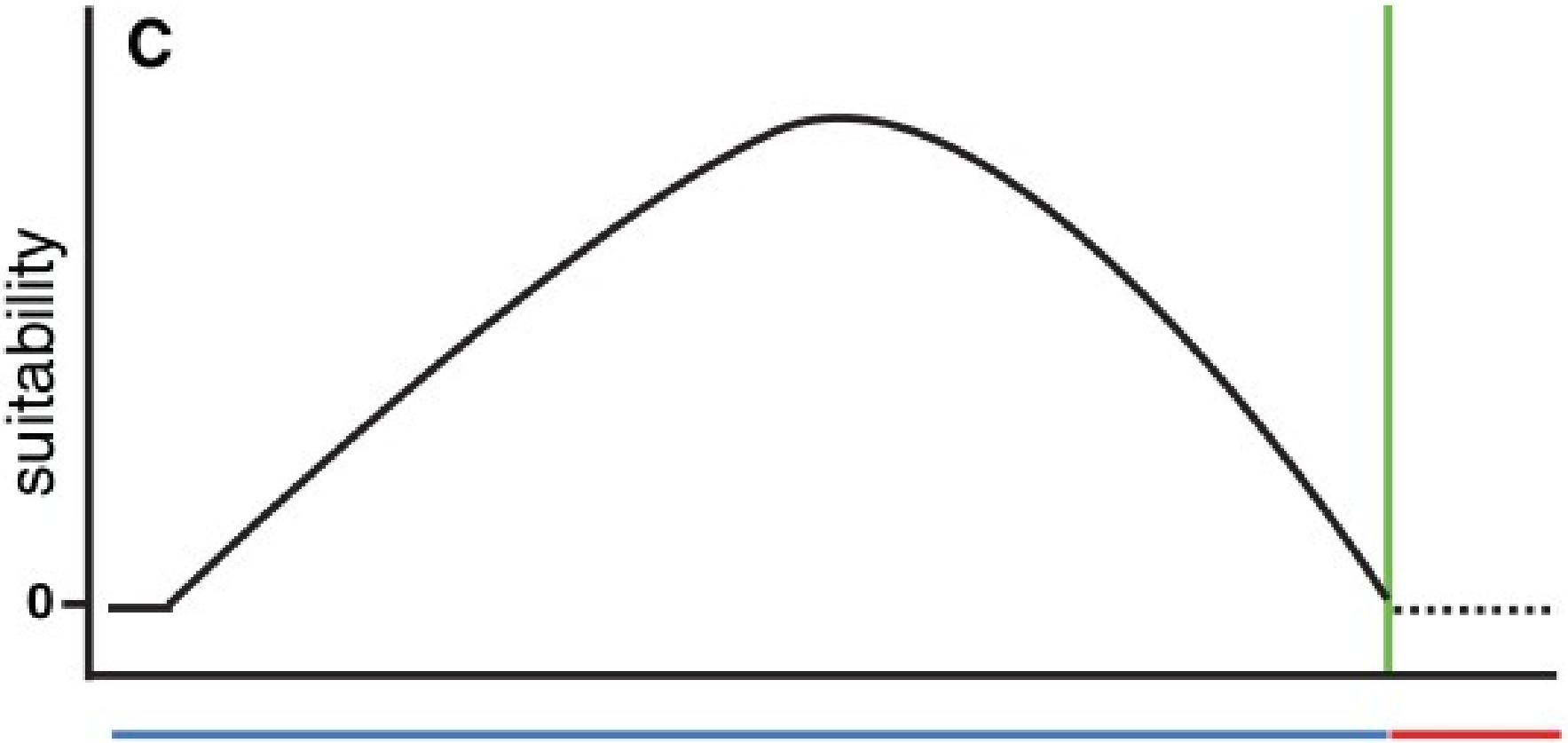
15. Analyse the relationship (in the environmental space) between the species points and the variables.

- **Check if the relationship is truncated or not.**



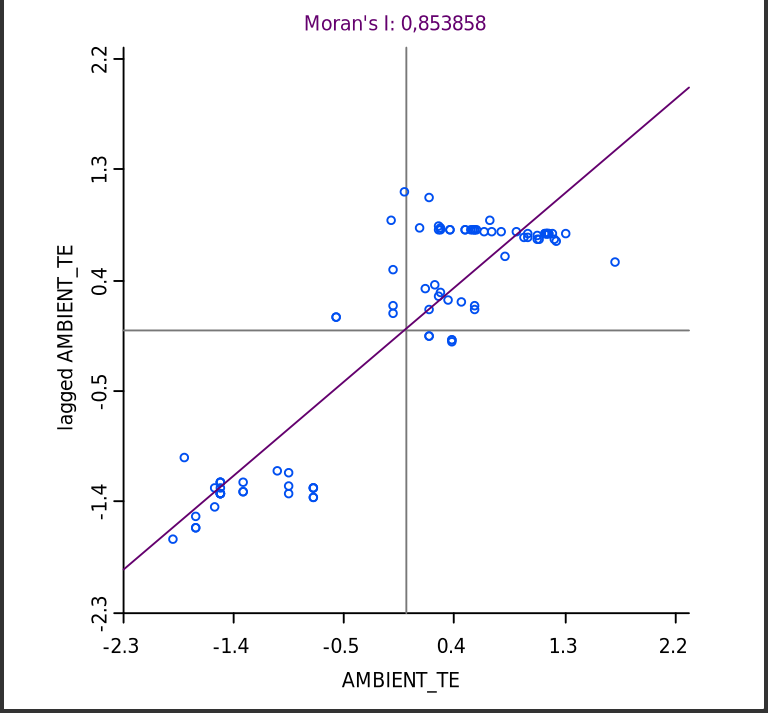
15. Analyse the relationship (in the environmental space) between the species points and the variables.

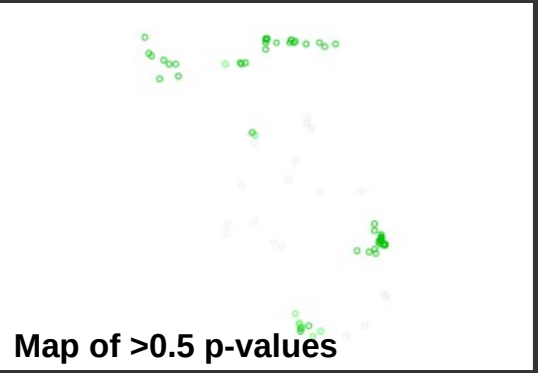
- **Check if the relationship is truncated or not.**

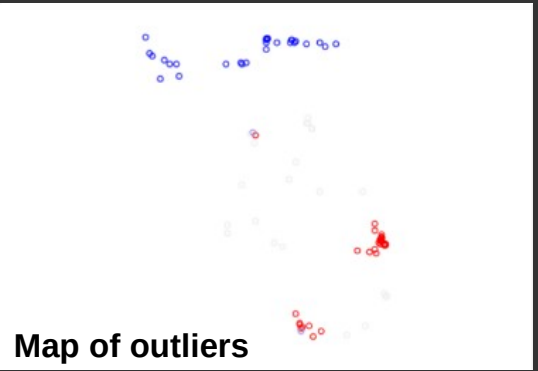


16. Calculate the Moran's I (Moran 1950) for each variable using the species' records in order to measure if you have autocorrelation.

- If you do not have autocorrelation, consider to filter the points (#6) or to change the variables (#7).
- The lack of autocorrelation can mean that you have local spatial patterns or that variable is meaningless to the species' distribution.







17. If you obtained truncated relationships in some of the variables, consider to change the study area.

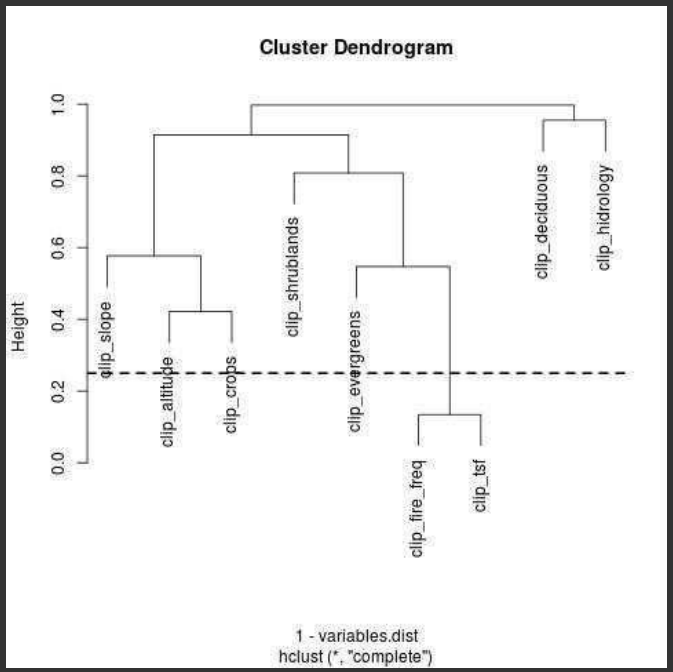
- **Repeat steps 10-14 until you have not-truncated relationships on the variables.**

18. Check that all species' records fall inside the study area.

- **If not, consider to change the study area or to loss those species' records.**
- **Sometimes, depending on the spatial resolution of the variables, you will be able to move the species' records to the nearest pixel.**

19. Measure the Pearson's correlation of the variables.

- Reject those variables with an absolute correlation value > 0.75.
- If you have enough variables, try to reject variables with an absolute correlation value > 0.70.
- No rule for how many variables you need to calculate the models, but around 5-8 is a good number.
- Pearson's correlation is parametric.
- Spearman's correlation is non-parametric.



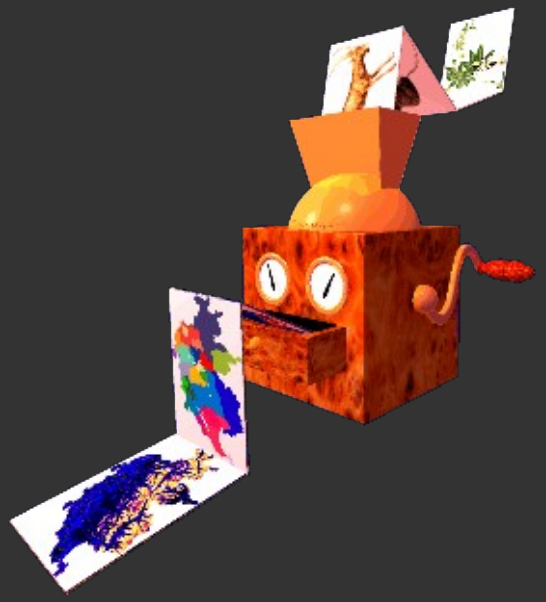
VARIABLES CORRELATION

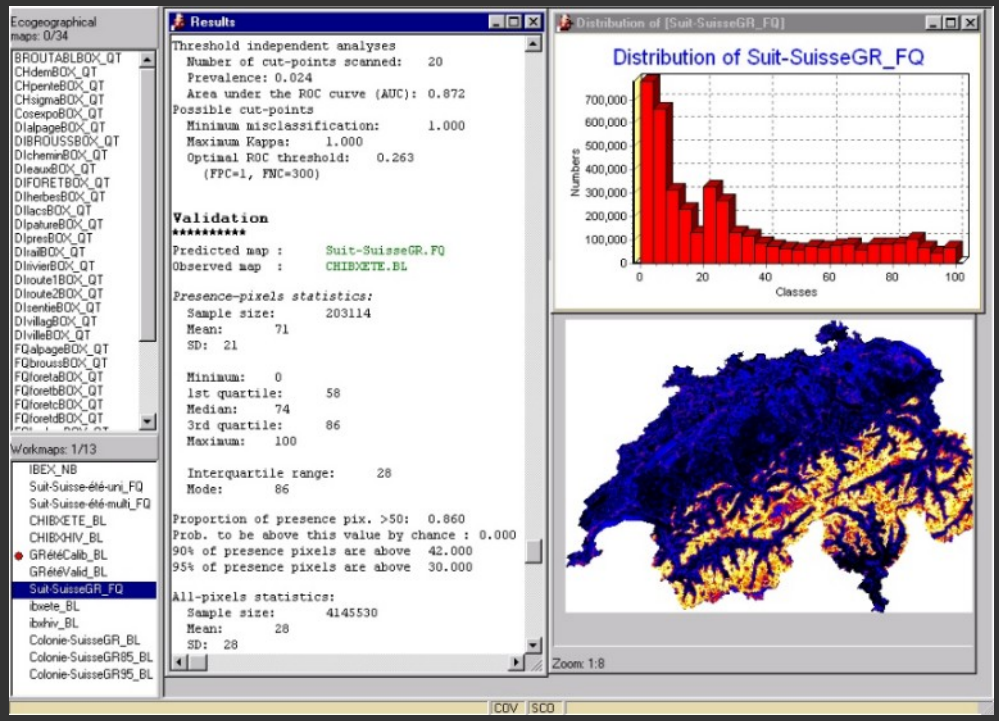
Most of the modelling techniques are sensible to high levels of correlation among variables.

- It is impossible to have uncorrelated variables
- It is difficult to identify correctly relationships among highly correlated variables

The model can be better than in reality.

We decide a maximum degree of correlation (positive or negative), normally at 0.75.





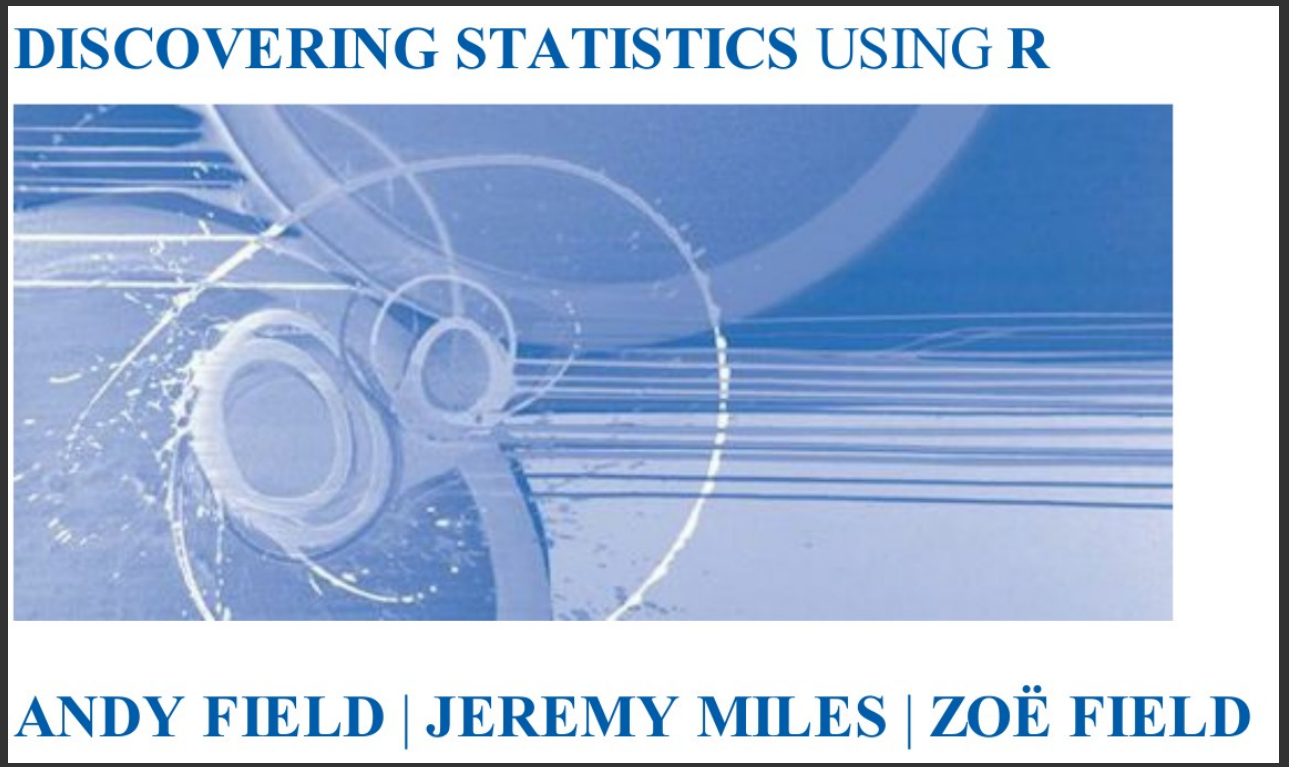
→ A solution is to calculate a Principal Component Analysis (PCA) and to use the main components instead of the variables.

→ The Ecological Niche Factor Analysis (ENFA) transforms the variables into several factors through a similar process to a PCA.

→ Less variables than species points

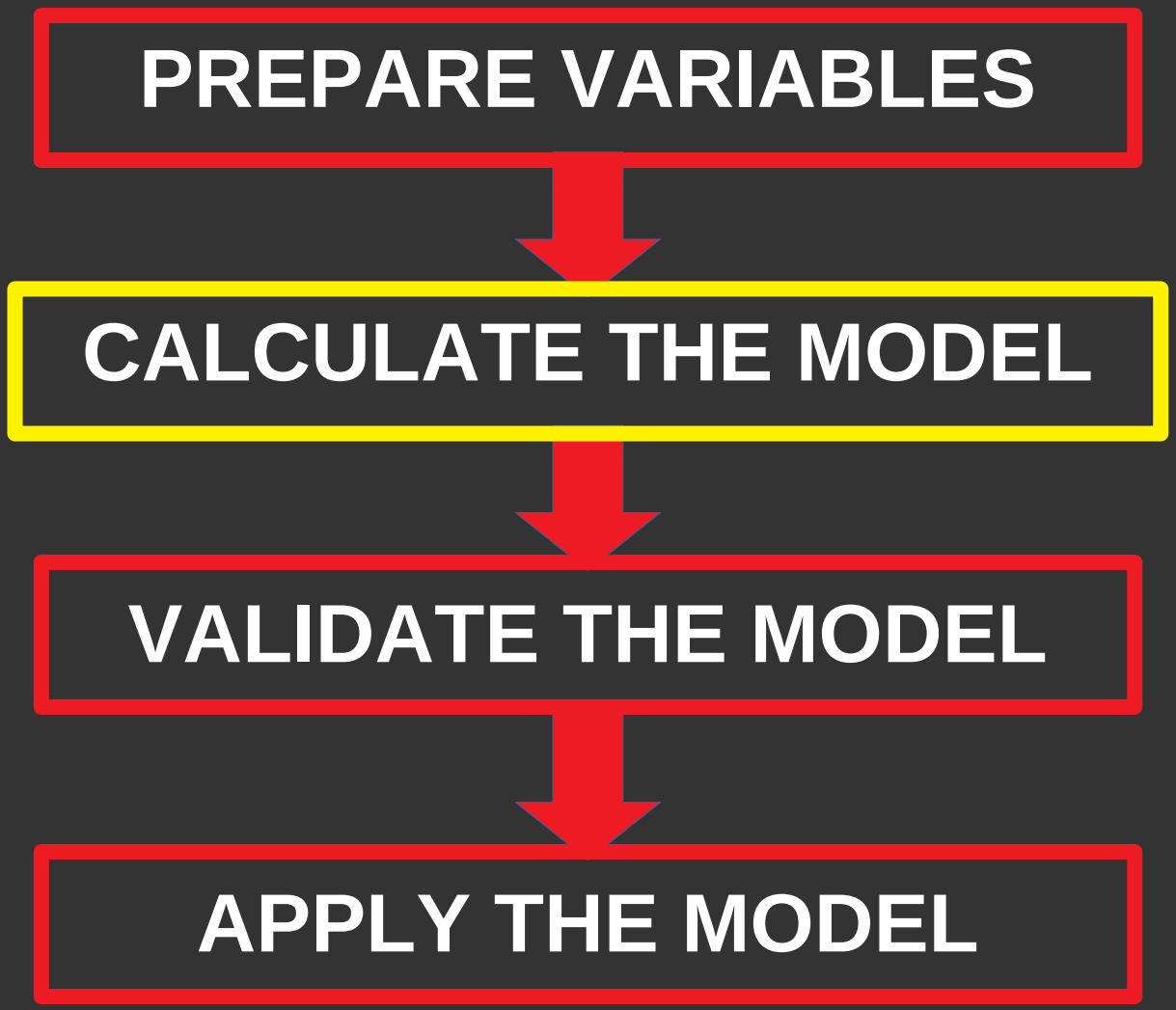
→ In GLMs:

- number of records: 50 + 8k
- number of records: 104 + k
  - k is the number of variables



ALL PROCESSES RELATED WITH OBTAINING AND PREPARING THE ENVIRONMENTAL VARIABLES ARE THE MOST TIME CONSUMING PART OF A MODELLING EXERCISE.

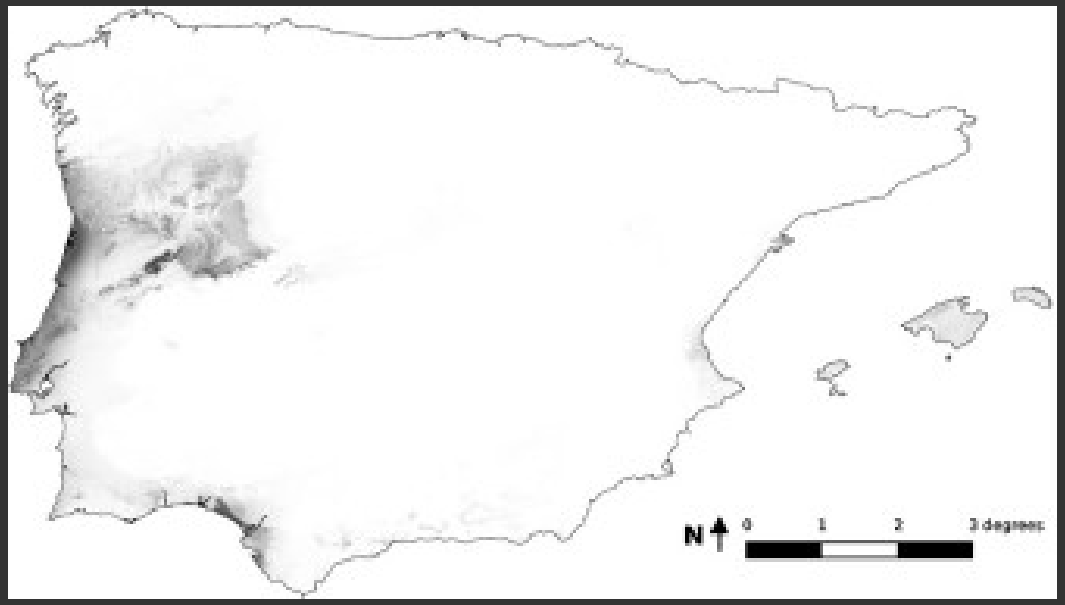
YOU WILL PROBABLY SPEND IN THIS PROCESS AROUND 95% OF YOUR TIME, AND ONLY 5% IN CALCULATING THE MODELS.

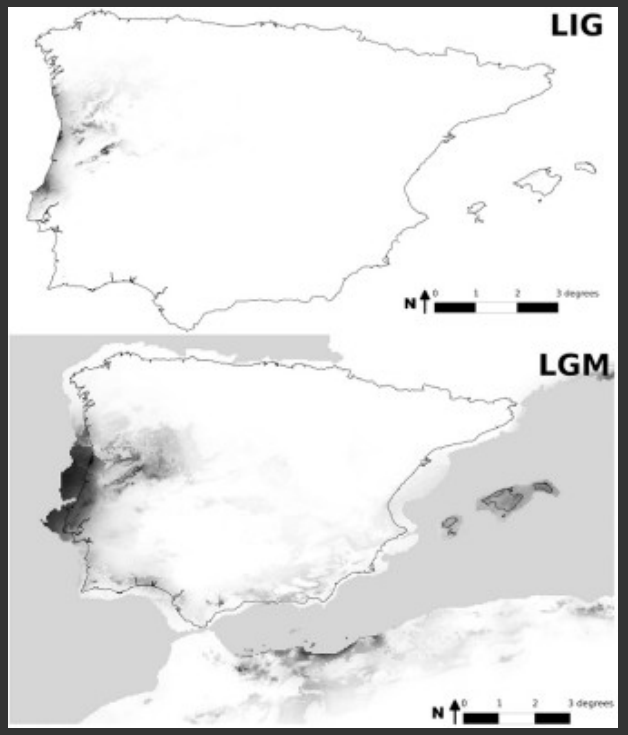

PREPARE VARIABLES
CALCULATE THE MODEL
VALIDATE THE MODEL
APPLY THE MODEL

20. If you want to project your models in space or time, you need the same set of variables (with the same name) for those scenarios of different space or time.

- **Some modelling methods can project the model to only one scenario.**
- **Other methods are able to project the model to multiple scenarios.**

Model

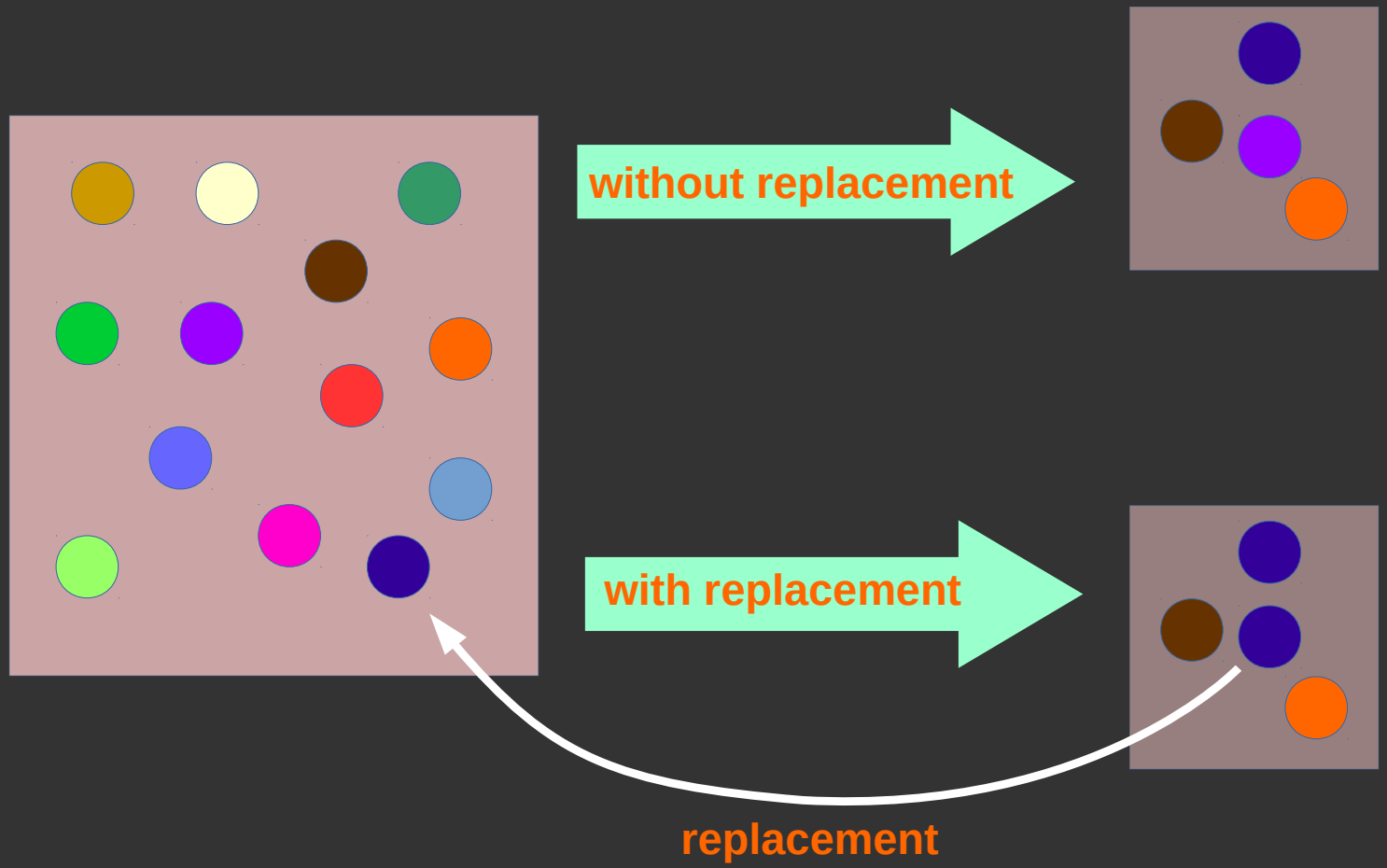




Projections

21. Normally, you must split your data in training records and testing records.

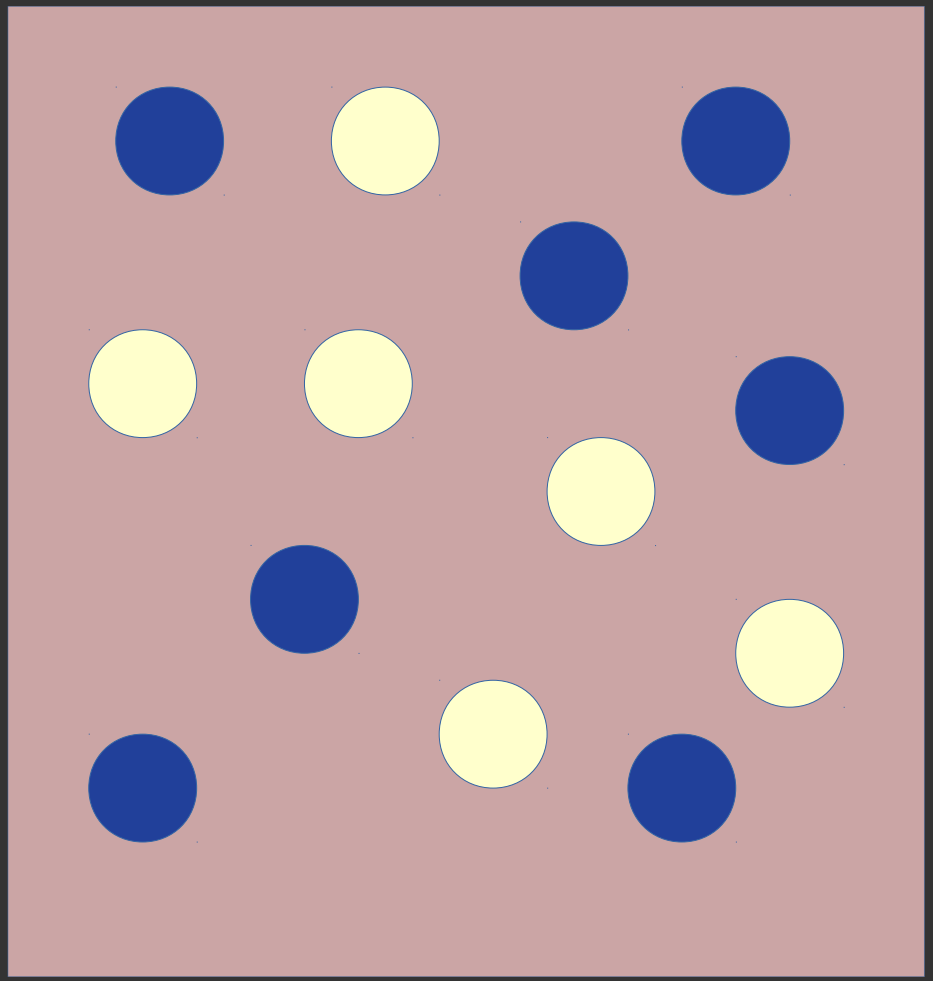
- You can use 70% for training records and 30 for testing.
- You have several splitting methods like K-fold partitioning, cross-validation, or bootstrapping.



→ Test the model with independent data.

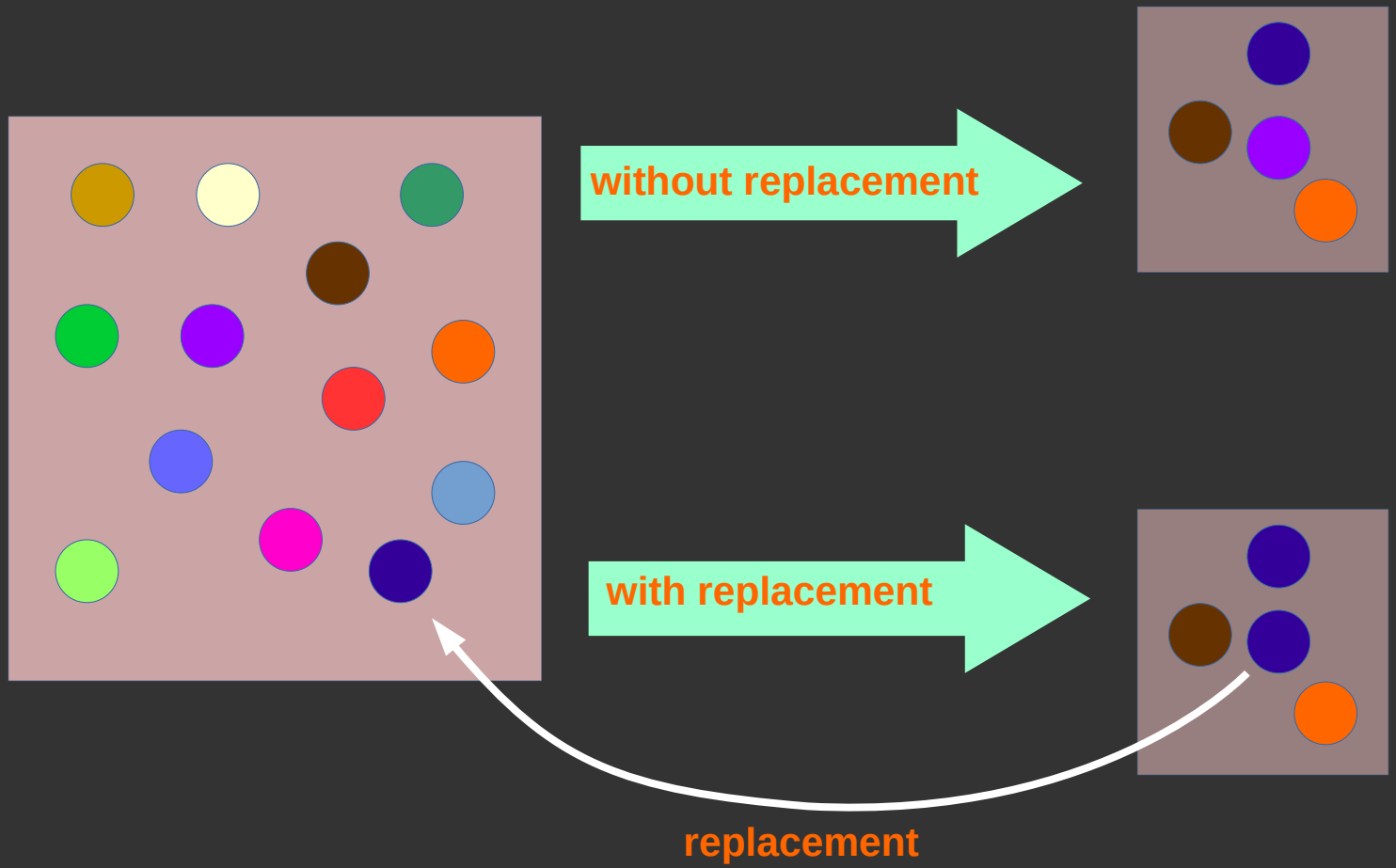
→ We do not usually have independent data.

→ TRAINING DATA
→ TEST DATA



Sampling without replacement → no element can be selected more than once in the same sample.

Sampling with replacement → an element may appear multiple times in the one sample.



Fielding & Bell 1997
Manel et al 2001

DATA PARTITIONING METHODS

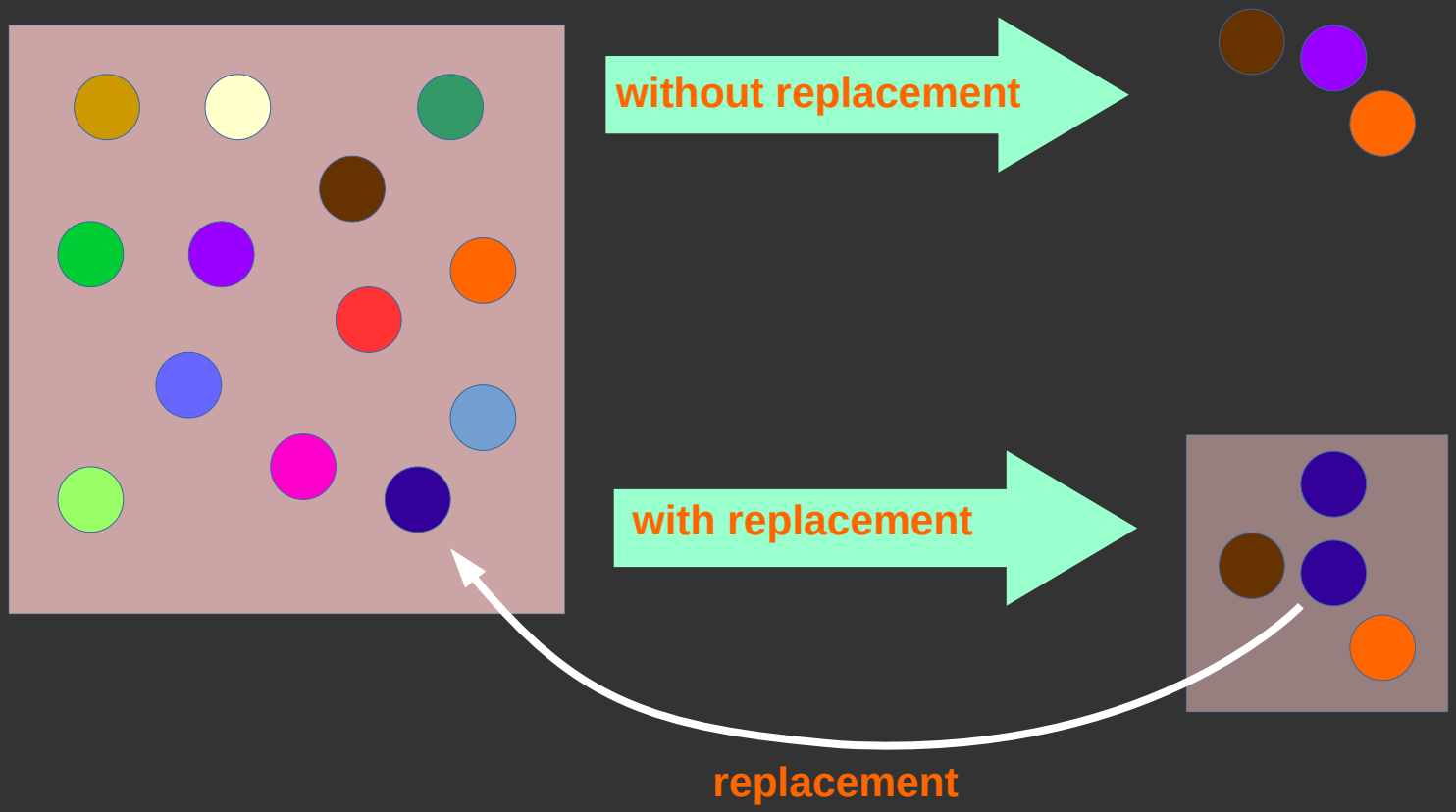
Table 1 Data partitioning methods for the allocation of cases to training and testing data sets.

Method	*Examples*	*Notes*
Resubstitution	Stockwell (1992) Osborne & Tigar (1992)	No partitioning is carried out, the same data are used for training and testing. This tends to provide optimistic measures of prediction success.
Bootstrapping	Buckland & Elston (1993) Verbyla & Litaitis (1989)	Bootstrap samples (sampling with replacement) are used to assess prediction success. Accuracy is usually reported as a mean and confidence limits.
Randomization	Capen *et al*. (1986)	Random samples are obtained by sampling without replacement. Accuracy is usually reported as a mean and confidence limits.
Prospective sampling	Capen *et al*. (1986) Fielding & Haworth (1995) Morrison *et al*. (1987)	A new sample of cases is obtained after the model has been developed. These could be from a different region or time.
k-fold partitioning	Stockwell (1992)	The data are split into k ($k > 2$) sets, only one of which is used for training. The remaining $k - 1$ sets are pooled for testing purposes. Also known as the hold-out or external method. Accuracy is usually reported as a mean and confidence limits.
Special cases of k-fold partitioning		
Leave-One-Out (L-O-O)	Capen *et al*. (1986) Osborne & Tigar (1992)	Also known as jackknife sampling, n samples of 1 case are tested sequentially, the remaining $n - 1$ cases forming the training set.
$K = 2$	Smith (1994)	Data are split into one training set and one testing set. A variety of strategies may be employed to determine the split.

Fielding & Bell 1997
Manel et al 2001

DATA PARTITIONING METHODS: BOOTSTRAPPING

- Training data is selected by sampling with replacement from the presence points, with the number of samples equaling the total number of presence points.
- The number of presence points in each set equals the total number of presence points, so the training data sets will have duplicate records.



Fielding & Bell 1997
Manel et al 2001

21. Normally, you must split your data in training records and testing records.

- **Rule of thumb described by Huberty (1994)**
- **Optimum partitioning of training and testing data depending on the number of variables**

Proportion of testing data

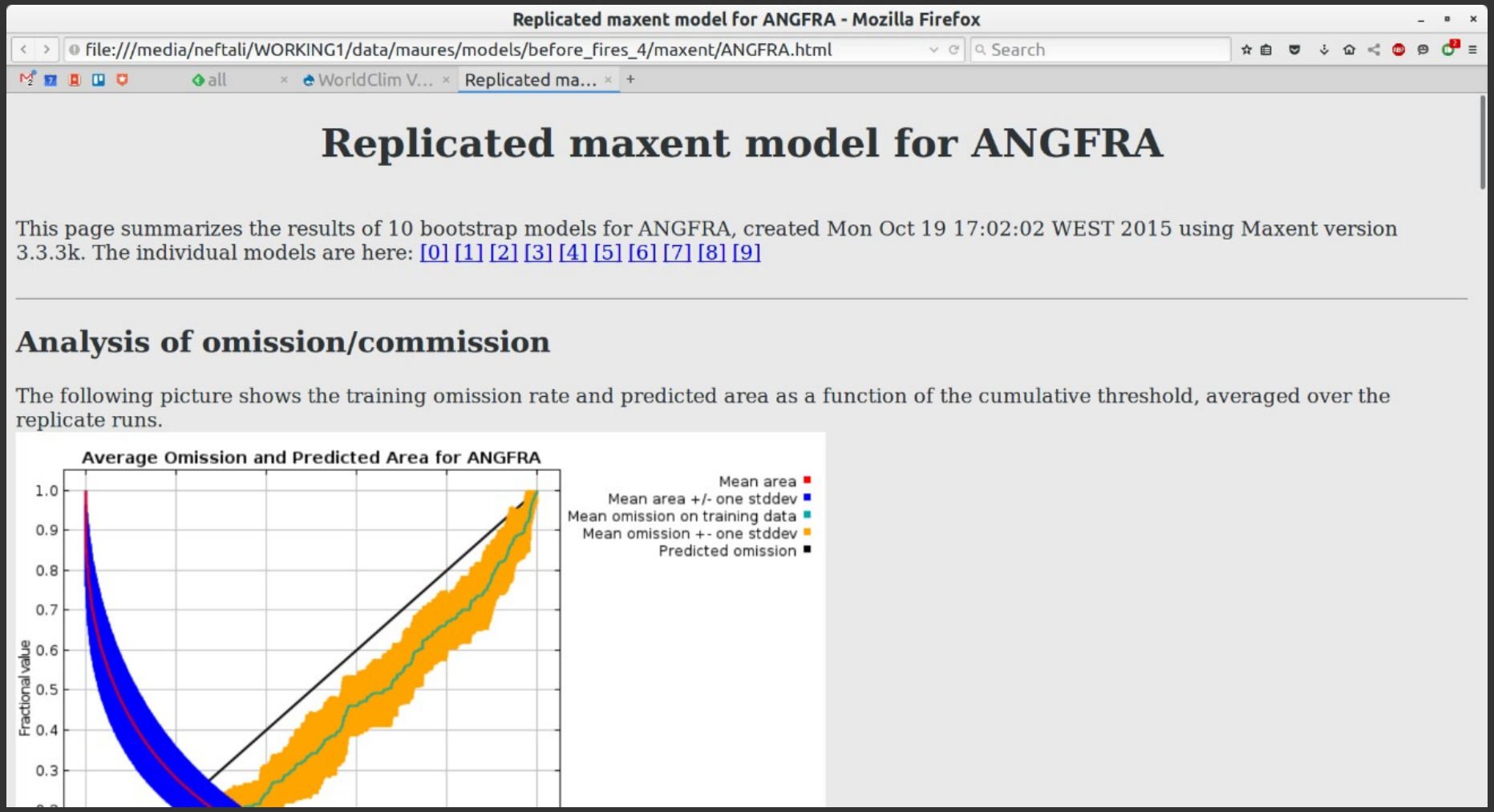
$$1/\left(1+\sqrt{p-1}\right)$$

- **p is the number of predictors**
  - **2 predictors → 50:50**
  - **5 predictors → 67:33**
  - **>10 predictors → 75:25**

Huberty, C. J. (1994) Applied Discriminant Analysis. New York, USA: Wiley Interscience.

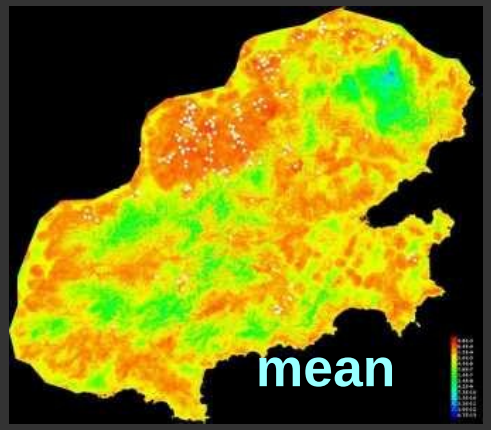
22. Consider if you need to calculate replicates for your models.

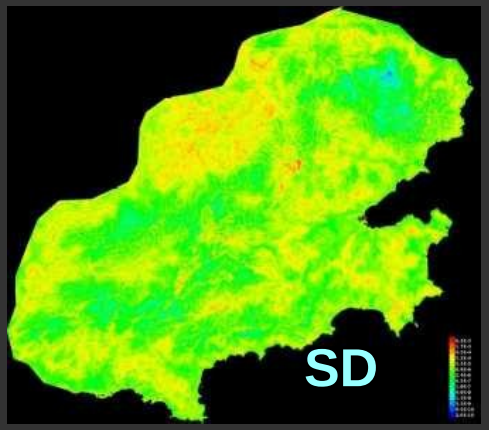
- For methods using probabilistic components.
- Any time you calculate the model, the result is going to be slightly different.

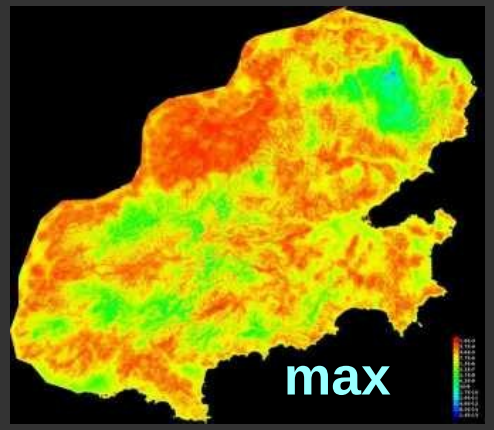


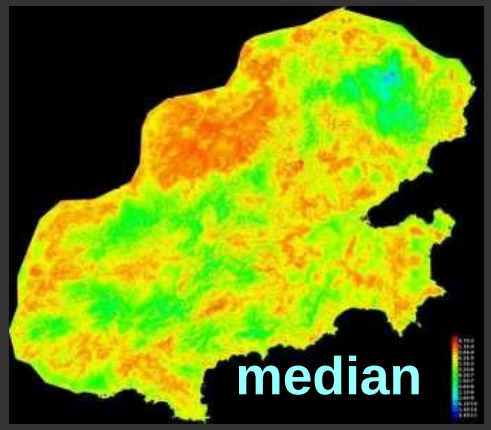
23. Calculate the models. This depends on the method used.

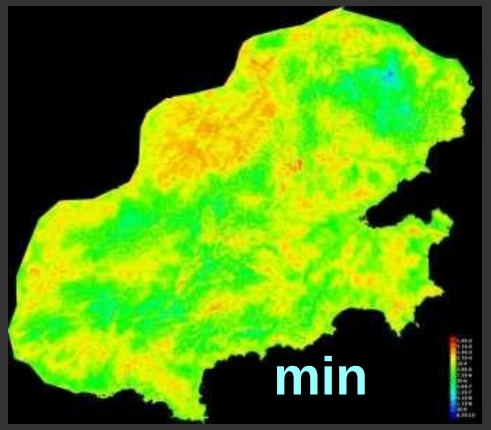
24. If you calculated several replicates of your models, you need to calculate the mean and SD of them.





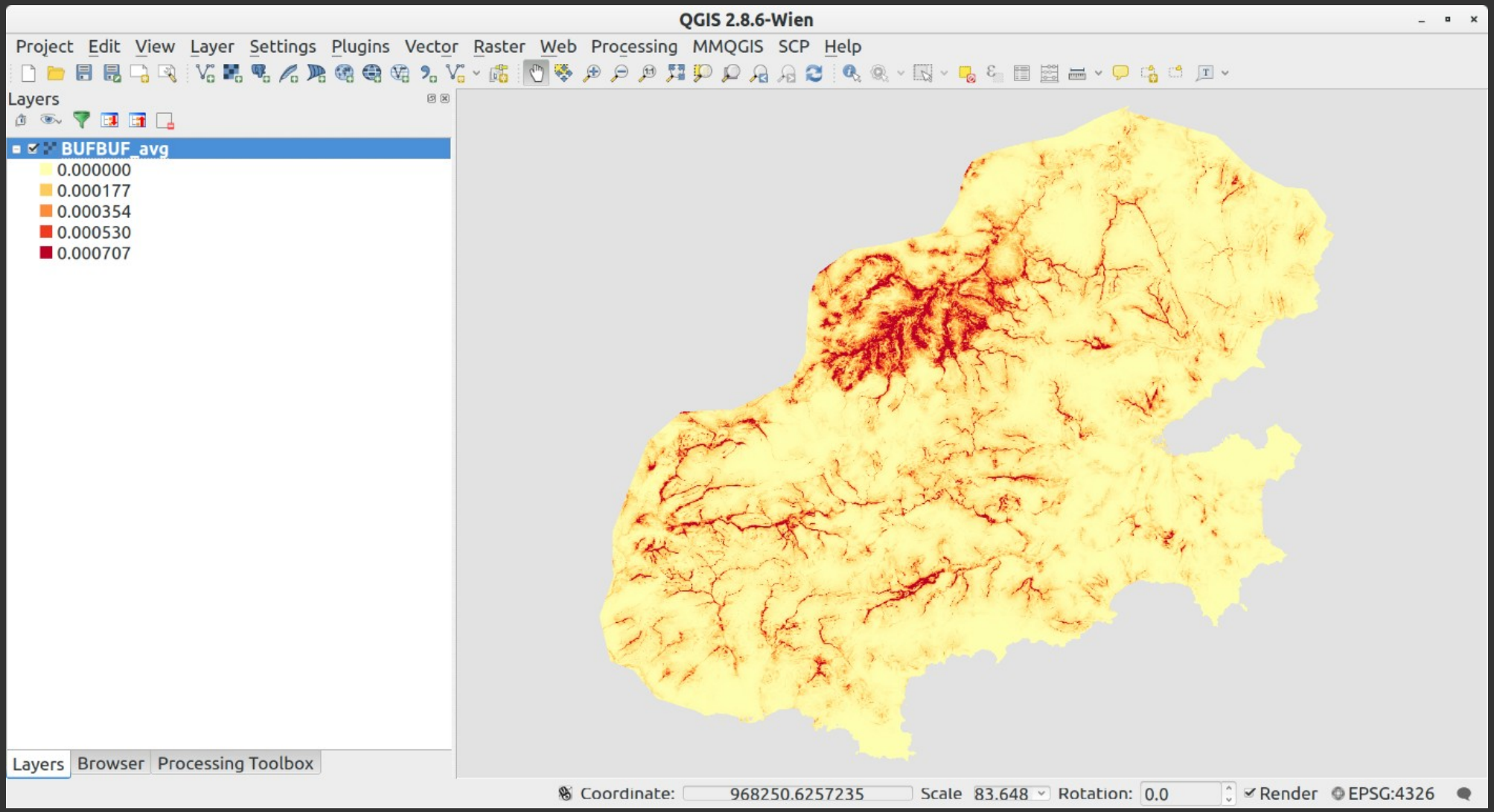


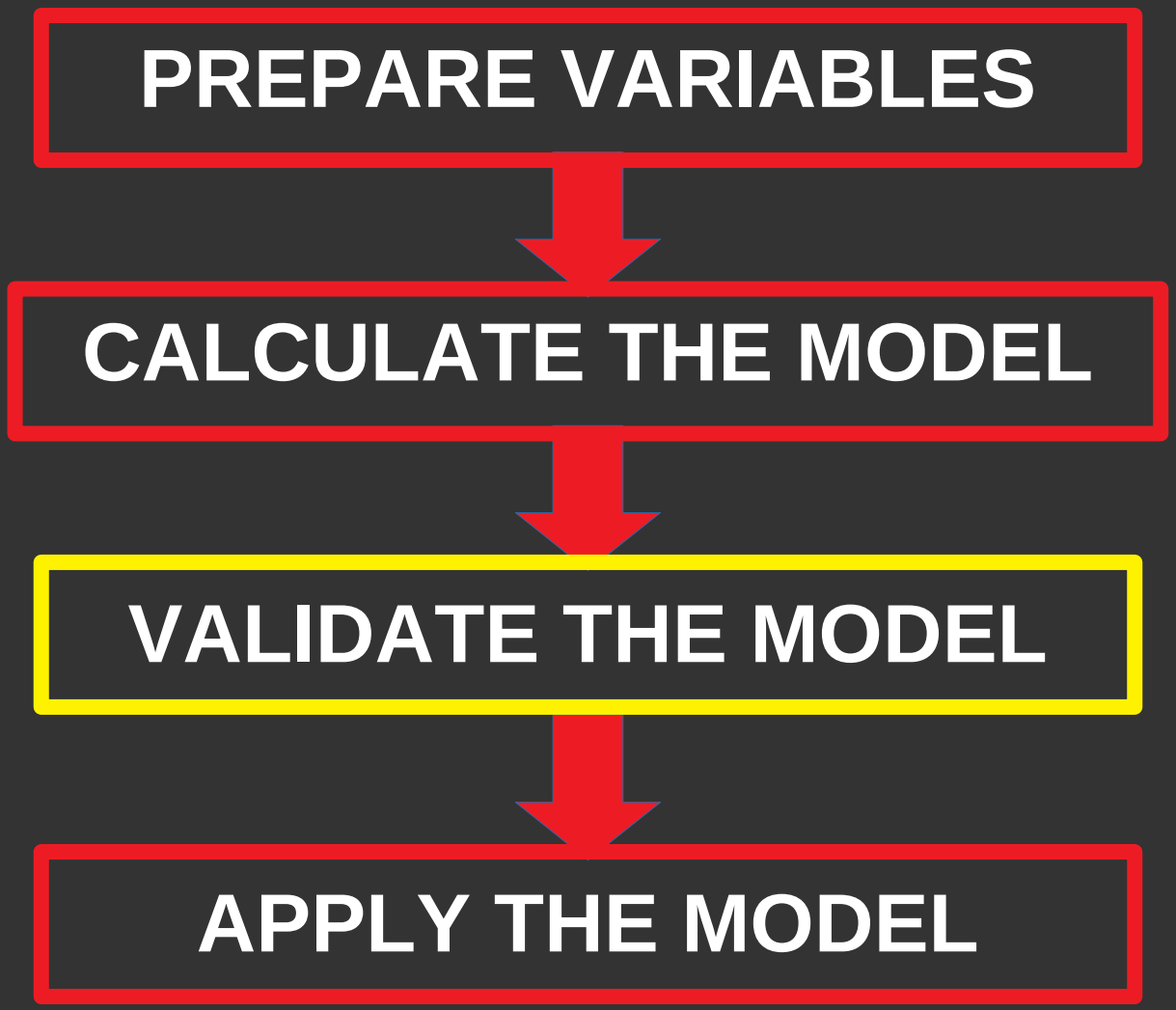




25. Think about the output format of the resulting models.

- Probably you will need to export this format to another one that can be read by a GIS.




PREPARE VARIABLES
CALCULATE THE MODEL
VALIDATE THE MODEL
APPLY THE MODEL

26. Check for validation measures provided by the modelling software.

- To validate a model is a very complex task.
  - All validation methods are designed for presence/absence models, not to presence-only methods.
  - All validation methods use a threshold to define unsuitable and suitable areas → the model must not include any presence record inside an unsuitable area.
    - not all individuals live in the best of the habitats.
  - The source-sink hypothesis proved that it is possible to have species living in unsuitable habitats (Pulliam 2000).

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Figure 1 A confusion matrix.

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Prevalence: percentage of presences correctly and incorrectly predicted; ratio of true and false presences

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Overall diagnostic success: percentage of all cases incorrectly predicted

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Correct classification rate: percentage of all presences and absences correctly predicted

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Sensitivity: percentage of true positives (presences) correctly predicted

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Specificity: percentage of true negatives (absences) correctly predicted

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

False positive rate: ratio of incorrectly assigned absences to all absences

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

False negative rate: ratio of incorrectly assigned presences to all presences

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Positive predictive power: percentage of predicted presences that were real

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Negative predictive power: percentage of predicted absences that were real

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Odds ratio: ratio of correctly assigned cases to incorrectly assigned cases

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Kappa: proportion of specific agreement

Fielding & Bell 1997
Manel et al 2001

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/ N
Overall diagnostic power	(b + d)/ N
Correct classification rate	(a + d)/ N
Sensitivity	a/ (a + c)
Specificity	d/ (b + d)
False positive rate	b/ (b + d)
False negative rate	c/ (a + c)
Positive predictive power (PPP)	a/ (a + b)
Negative predictive power (NPP)	d/ (c + d)
Misclassification rate	(b + c)/ N
Odds-ratio	(ad)/ (cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/ N)]/ [N − (((a + c)(a + b) + (b + d)(c + d))/ N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/ [N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Normalised mutual information: proportion of specific agreement

Fielding & Bell 1997
Manel et al 2001

Diversity and Distributions, (Diversity Distrib.) (2013) **19**, 1333–1338

BIODIVERSITY LETTER

New measures for assessing model equilibrium and prediction mismatch in species distribution models

A. Márcia Barbosa[1,2*], Raimundo Real[3], A.-Román Muñoz[4,5] and Jennifer A. Brown[6]

$$\text{Under} - \text{Prediction Rate(UPR)} = \frac{\text{unsuitable \& occupied}}{\text{unsuitable}} = \frac{c}{c+d}$$

$$\text{Over} - \text{Prediction Rate(OPR)} = \frac{\text{suitable \& unoccupied}}{\text{suitable}} = \frac{b}{a+b}$$

$$\text{Potential Presence Increment(PPI)} = \frac{\text{suitable}}{\text{occupied}} - 1 = \frac{a+b}{a+c} - 1$$

$$\text{Potential Absence Increment(PAI)} = \frac{\text{unsuitable}}{\text{unoccupied}} - 1 = \frac{c+d}{b+d} - 1$$

Barbosa et al 2013

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Figure 1 A confusion matrix.

Fielding & Bell 1997
Manel et al 2001

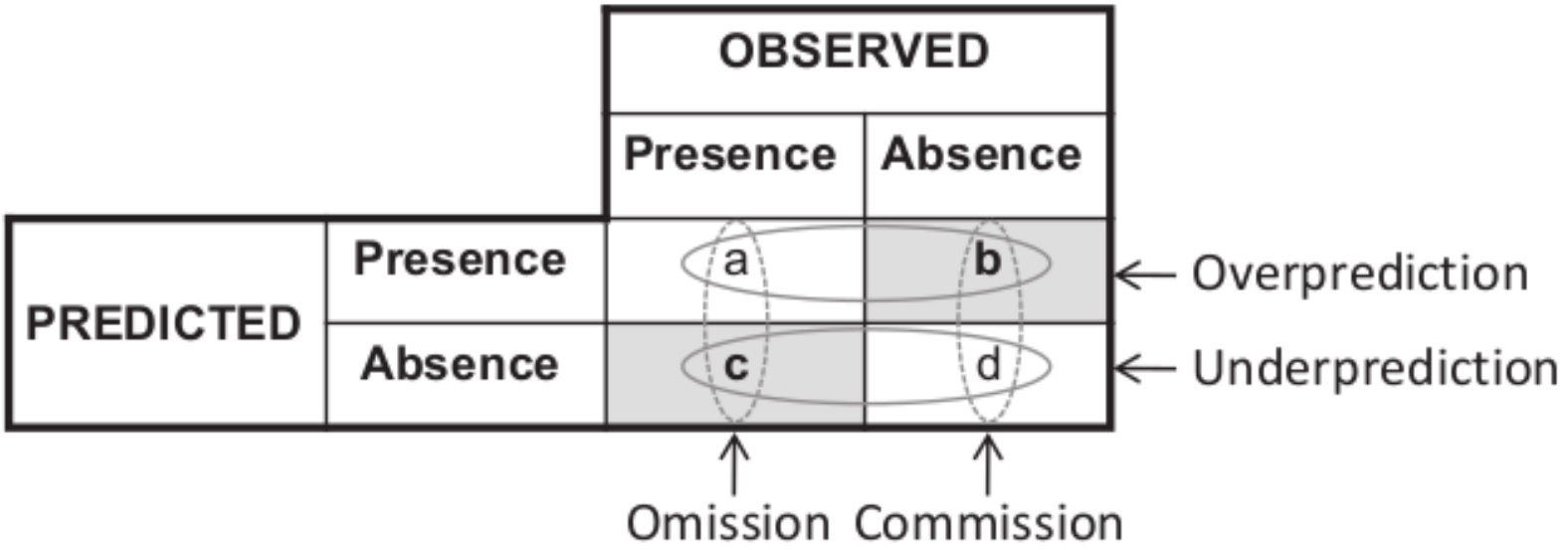


Figure 1 Confusion matrix showing match (white background) and mismatch (grey background) between observed and predicted presences and absences of a modelled species' distribution. Encircled are the elements used to calculate the omission and commission rates (dashed lines; Anderson *et al.*, 2003) and the proposed under- and over-prediction rates (solid lines).

Barbosa et al 2013

RECEIVER-OPERATING CHARACTERISTIC AREA UNDER THE CURVE

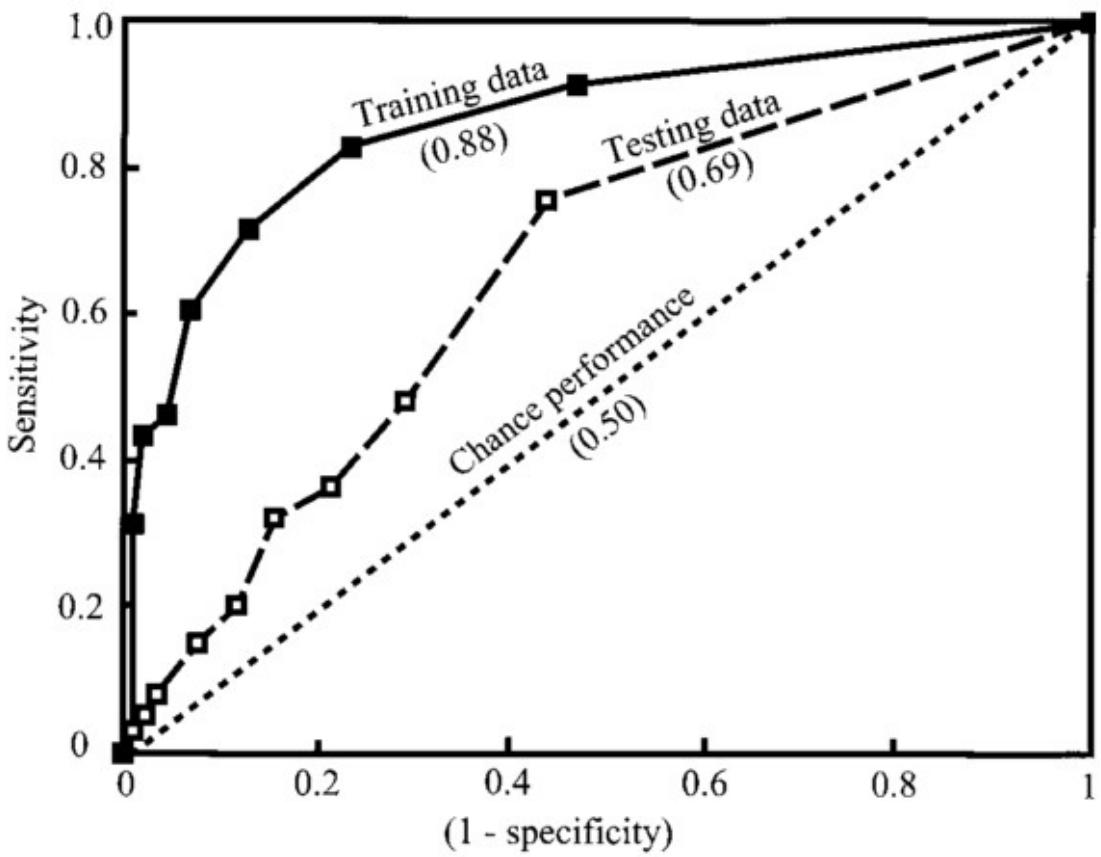


Figure 4 An example ROC plot. The data relate to Golden Eagle distribution using a logistic regression model. Figures in parentheses are the areas under the curves.

Fielding & Bell 1997
Manel et al 2001

RECEIVER-OPERATING CHARACTERISTIC
AREA UNDER THE CURVE

ROC PLOT is build with all possible thresholds.

The Sensitivity and 1-Specificity is calculated for each threshold.

AUC is the integral of the ROC plot.

- **A random model has an AUC of 0.5.**
- **Less than 0.5 is worse than a random model.**
- **Good models must be above 0.5 and close to 1.**

Sensitivity	a/ (a + c)
Specificity	d/ (b + d)

Fielding & Bell 1997
Manel et al 2001

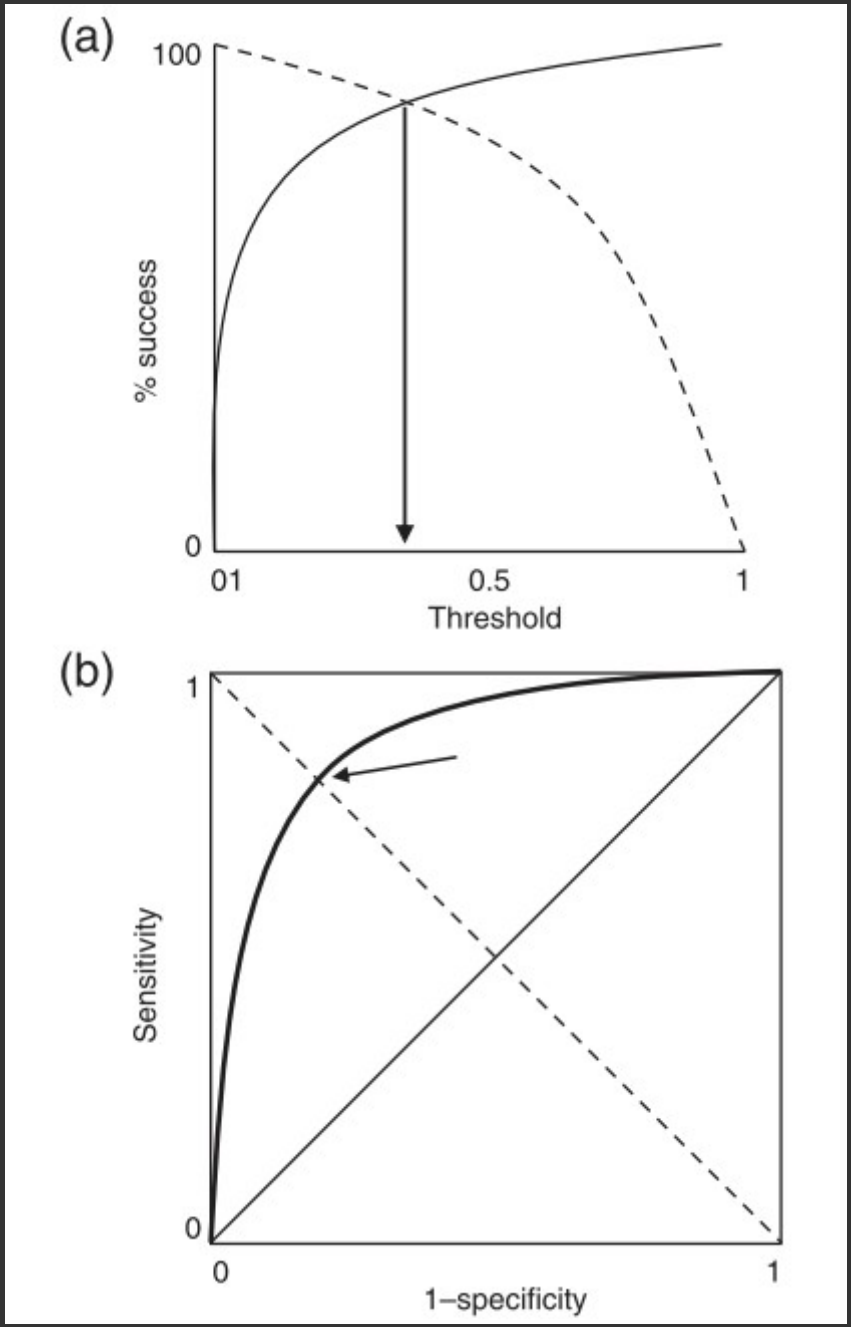


Figure 1 (a) Variation in the percentage of success in the prediction of presences (continuous line) and absences (broken line) with the change in the threshold used to discriminate both states from a continuous probability variable. The arrow represents the threshold that minimizes the difference between sensitivity and specificity. (b) ROC plot in which an arrow shows the 'most north-western' point.

Lobo et al 2008

RECEIVER-OPERATING CHARACTERISTIC
AREA UNDER THE CURVE

Lobo et al 2008 rejected the use of ROC as validating methods because AUC values depend on the size of the study area.

Specialised species are easier to model and thus models have better validation results.

→ When you increase the study area, you are increasing the specialisation degree of the species and thus you will obtain higher AUC values → BUT THIS NOT A PROBLEM!!

This only hampers to compare models of different species if you are using different study areas (VanderWal et al 2009).

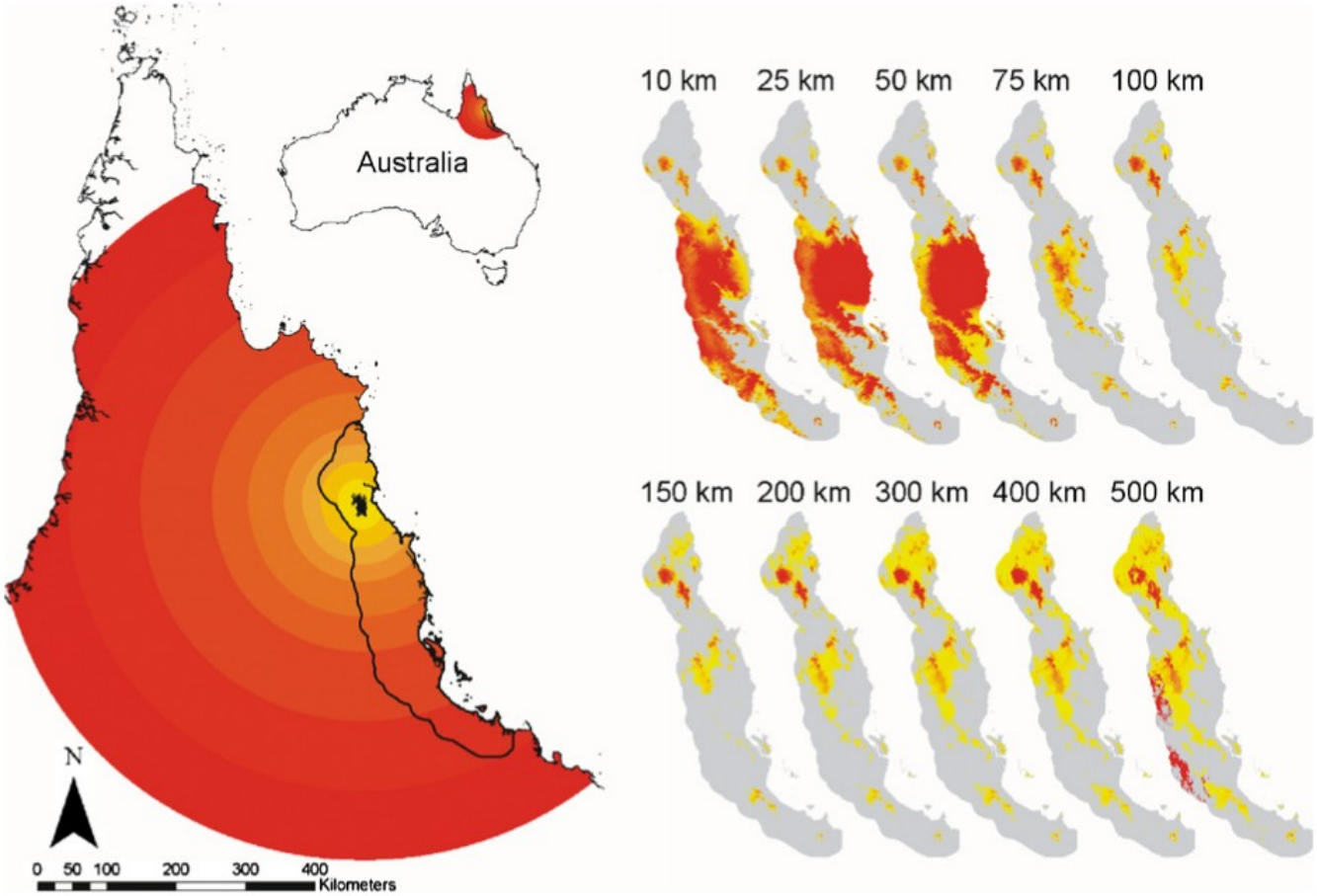


Fig. 1 – Backgrounds from which pseudo-absences were drawn for C. *hosmeri* and predicted distributions in the Australia Wet Tropics region (outlined in black) given the different background sizes. Increasing background size corresponds to darkening of buffering bands surrounding the occurrence points of the species (represented as x symbols here) in the left image; these regions represent increasing distances from 10 to 500 km from the occurrence points. Warmer colors on the right images infer greater predicted suitability for the C. *hosmeri*. Grey areas fall below the threshold of suitability and are assumed to not be part of the distribution.

VanderWal et al 2009

TSS = sensitivity + specificity - 1

$$\frac{a*d-b*c}{(a+c)*(b+d)}$$

TSS is independent of prevalence

- **Ranges from −1 to +1**
- **+1 indicates perfect agreement**
- **Zero or less indicate a performance no better than random**

Journal of Applied Ecology 2006
43, 1223–1232

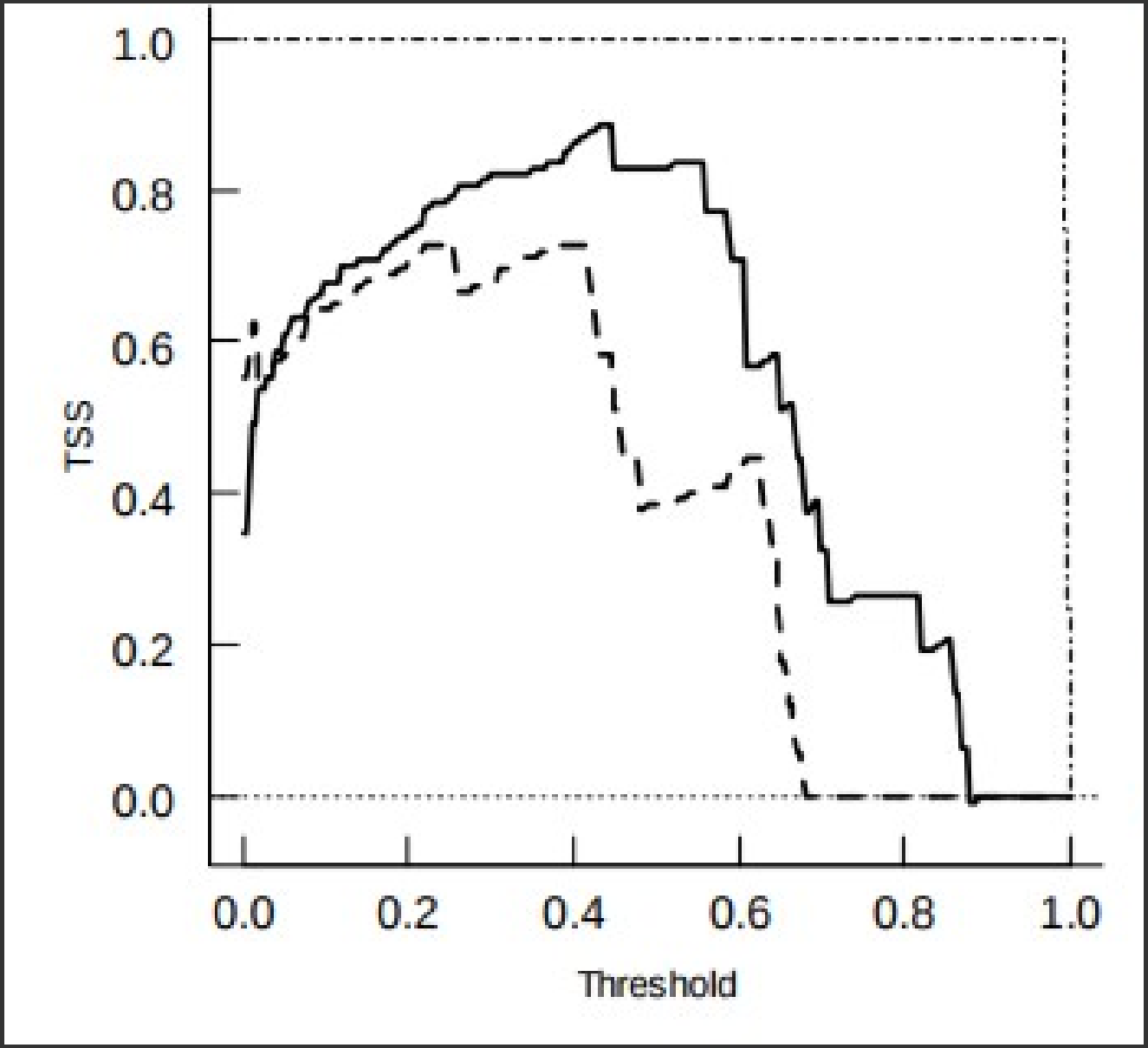
METHODOLOGICAL INSIGHTS

Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)

OMRI ALLOUCHE, ASAF TSOAR and RONEN KADMON
Department of Evolution, Systematics and Ecology, Institute of Life Sciences, The Hebrew University, Givat-Ram, Jerusalem 91904, Israel

Allouche et al 2006

TSS: TRUE SKILL STATISTICS



Allouche et al 2006

IS IT POSSIBLE TO VALIDATE A MODEL?

- The main conclusion is there is no good methods for validating a ENM.
- You can obtain a very high AUC value and the model can be very bad.
- Statistical meaning may be not linked with biological meaning.
- Everything depends on the quality of the species' records.

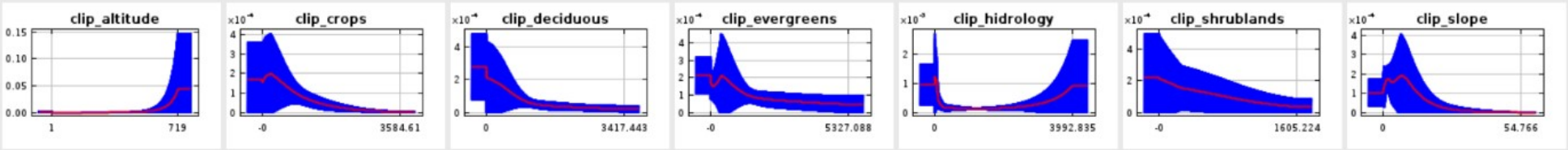
26. Check for validation measures provided by the modelling software.

- Calculate Null Models following Raes & ter Steege (2007).

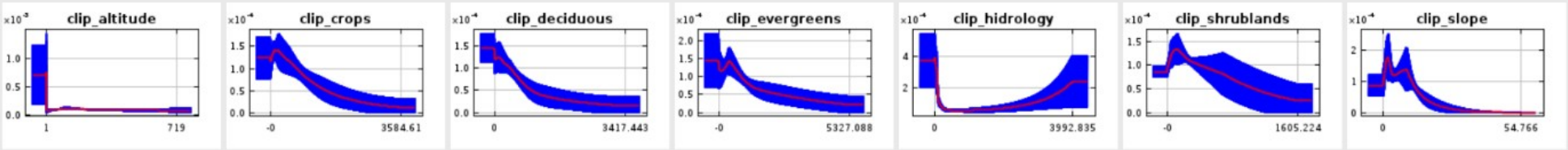
Raes, N., & ter Steege, H. (2007). A null-model for significance testing of presence-only species distribution models. Ecography, 30(5), 727–736.

27. Analyse the variables' response curves if available.

These curves show how each environmental variable affects the Maxent prediction. The curves show how the logistic prediction changes as each environmental variable is varied, keeping all other environmental variables at their average sample value. Click on a response curve to see a larger version. Note that the curves can be hard to interpret if you have strongly correlated variables, as the model may depend on the correlations in ways that are not evident in the curves. In other words, the curves show the marginal effect of changing exactly one variable, whereas the model may take advantage of sets of variables changing together. The curves show the mean response of the 10 replicate Maxent runs (red) and and the mean +/- one standard deviation (blue, two shades for categorical variables).



In contrast to the above marginal response curves, each of the following curves represents a different model, namely, a Maxent model created using only the corresponding variable. These plots reflect the dependence of predicted suitability both on the selected variable and on dependencies induced by correlations between the selected variable and other variables. They may be easier to interpret if there are strong correlations between variables.

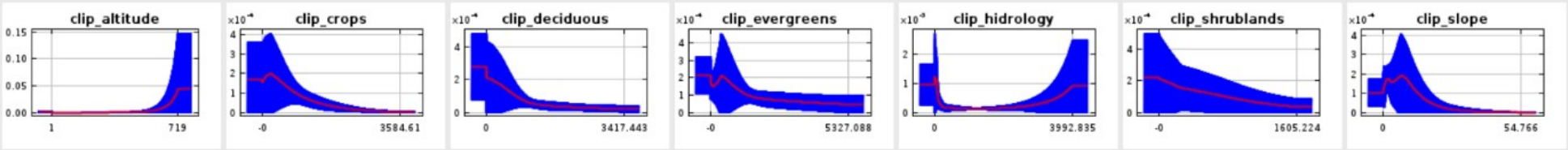


28. Analyse the contribution importance of each variable to the final model.

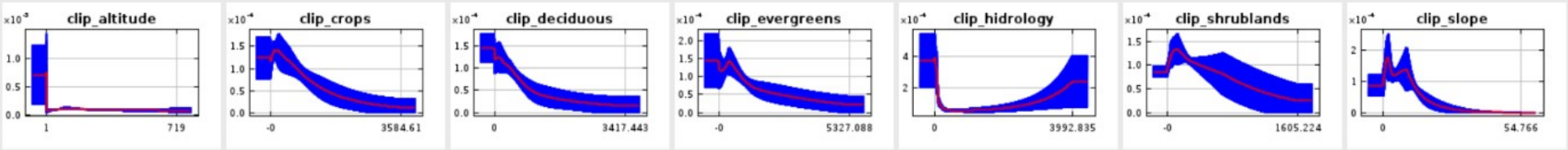
Variable	Percent contribution	Permutation importance
clip_hidrology	36.7	23.6
clip_crops	13.9	15.4
clip_slope	13.1	16.8
clip_altitude	12.9	14.8
clip_deciduous	11.6	11.8
clip_evergreens	8.9	11.2
clip_shrublands	2.8	6.4

29. Check for correlation effects on the variables.

These curves show how each environmental variable affects the Maxent prediction. The curves show how the logistic prediction changes as each environmental variable is varied, keeping all other environmental variables at their average sample value. Click on a response curve to see a larger version. Note that the curves can be hard to interpret if you have strongly correlated variables, as the model may depend on the correlations in ways that are not evident in the curves. In other words, the curves show the marginal effect of changing exactly one variable, whereas the model may take advantage of sets of variables changing together. The curves show the mean response of the 10 replicate Maxent runs (red) and and the mean +/- one standard deviation (blue, two shades for categorical variables).

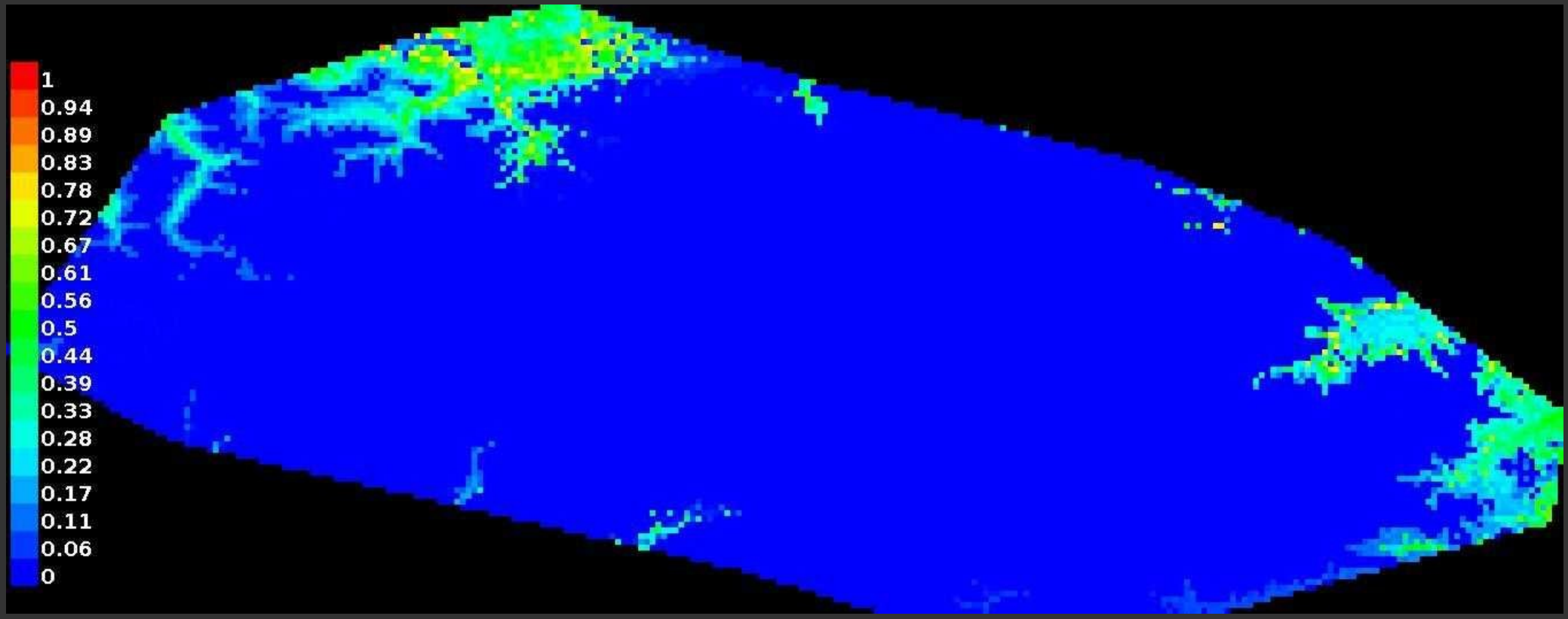


In contrast to the above marginal response curves, each of the following curves represents a different model, namely, a Maxent model created using only the corresponding variable. These plots reflect the dependence of predicted suitability both on the selected variable and on dependencies induced by correlations between the selected variable and other variables. They may be easier to interpret if there are strong correlations between variables.

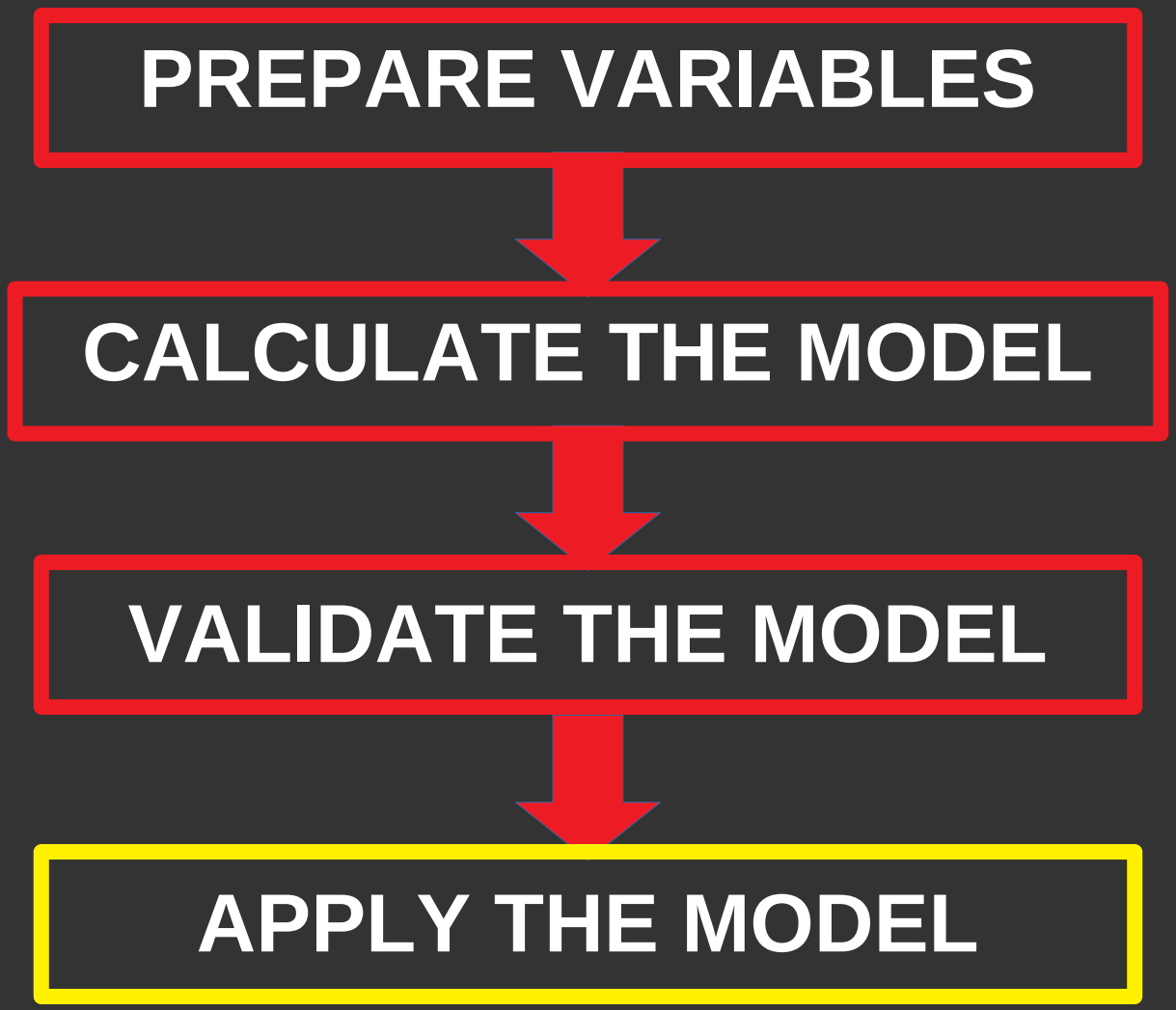


30. Analyse the projected models, if available.

31. Identify those areas outside the variable values of your models with Clamping maps.
    MESS: Multivariate Environmental Similarity Surface



- REMEMBER THAT THE MODEL IS AN APPROACH TO THE REALITY
- DO NOT GET CONCLUSIONS THAT ARE NOT SUPPORTED BY YOUR RESULTS
- DO NOT ASK QUESTIONS THAT YOUR MODEL CANNOT ANSWER

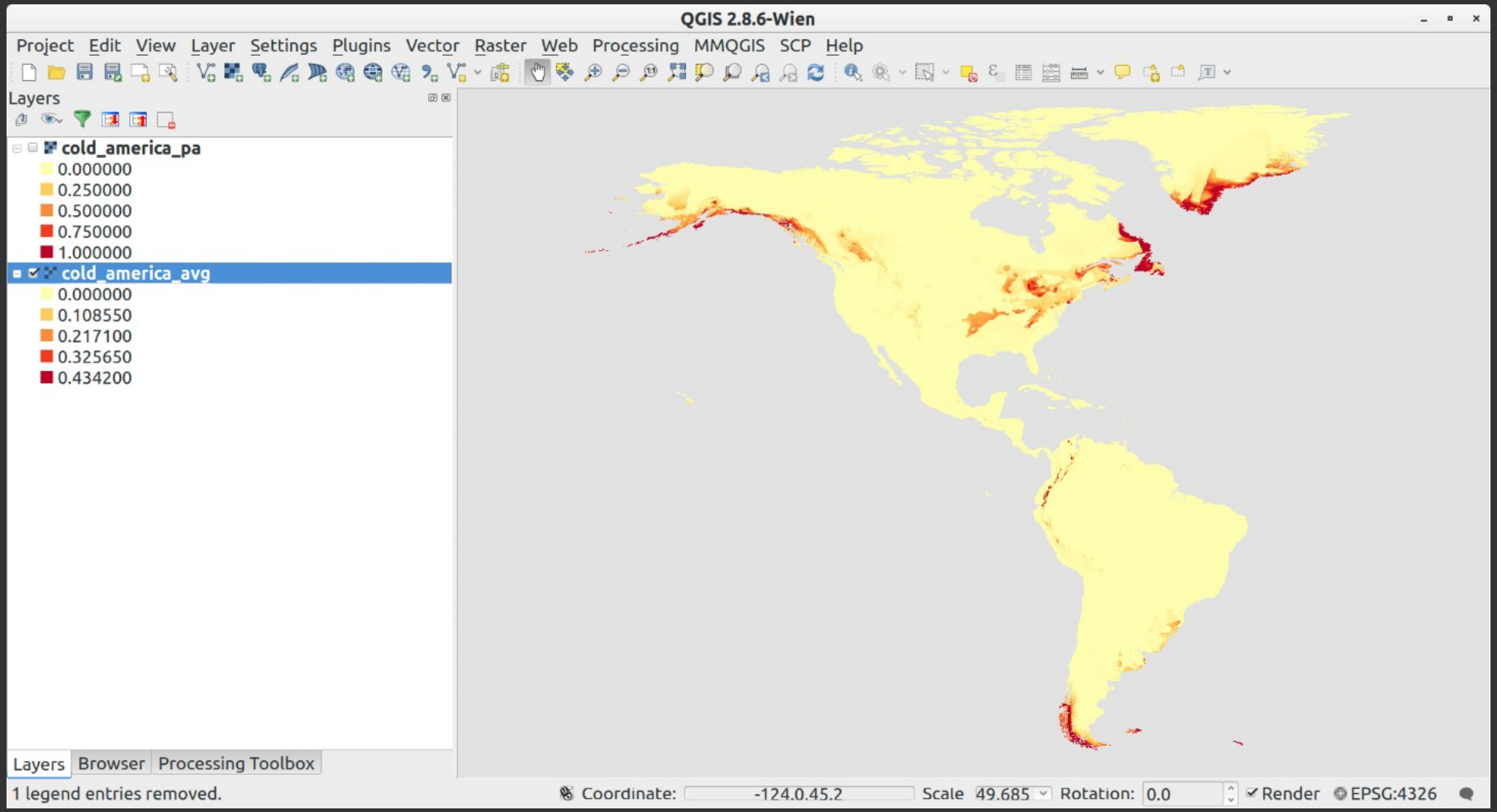

PREPARE VARIABLES
CALCULATE THE MODEL
VALIDATE THE MODEL
APPLY THE MODEL

32. If you need habitat suitability maps (*sensu* Sillero 2011), you will need to choose a threshold to split your models in suitable and unsuitable areas (Liu et al. 2005, 2013, Jiménez-Valverde and Lobo 2007, Freeman and Moisen 2008, Nenzén and Araújo 2011).

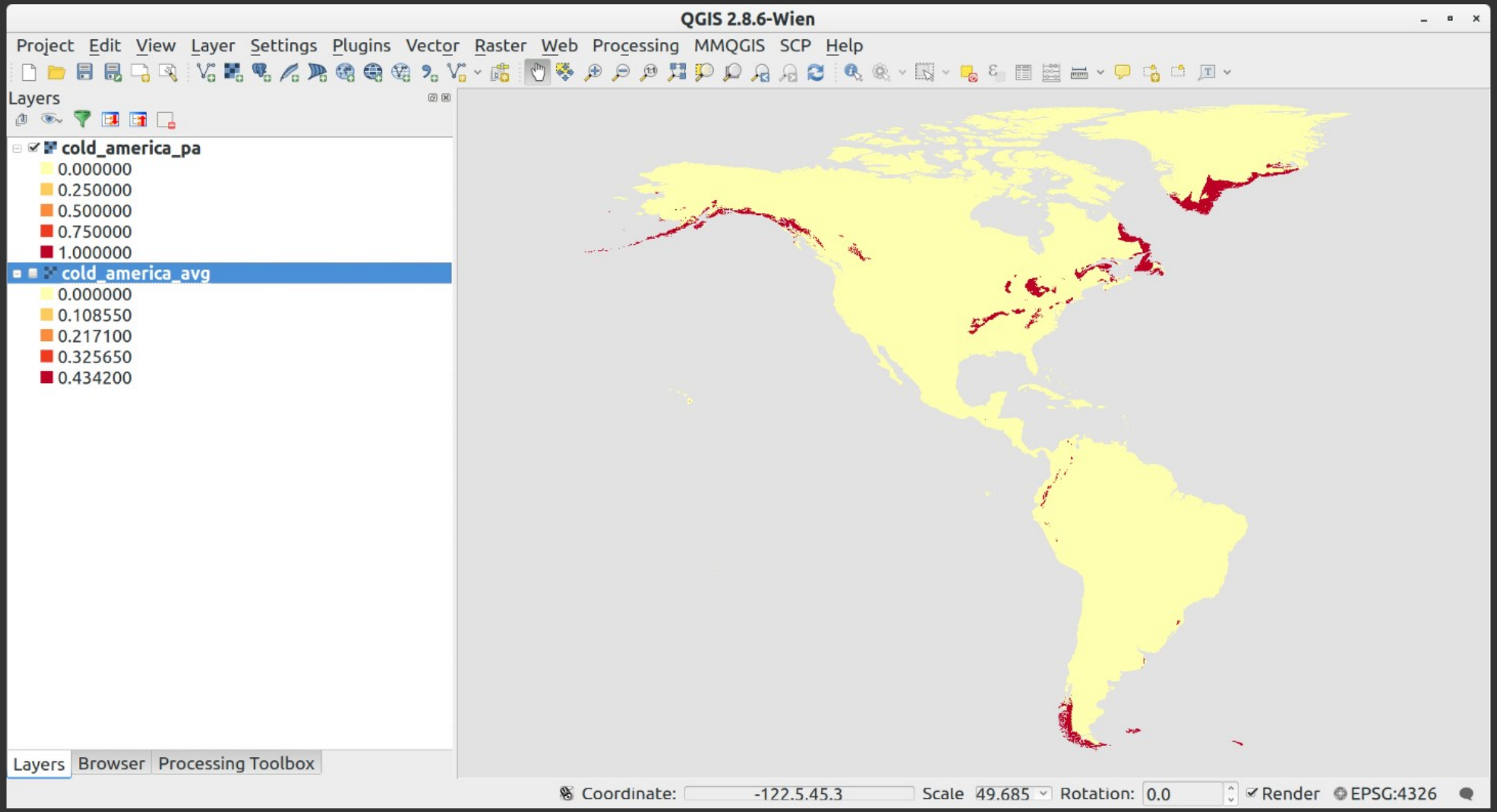
- The threshold is a subjective value and there are many available (10% presence percentile, lowest presence habitat suitability value, etc).

33. Apply the chosen threshold to the models with a GIS.

HABITAT SUITABILITY MAPS



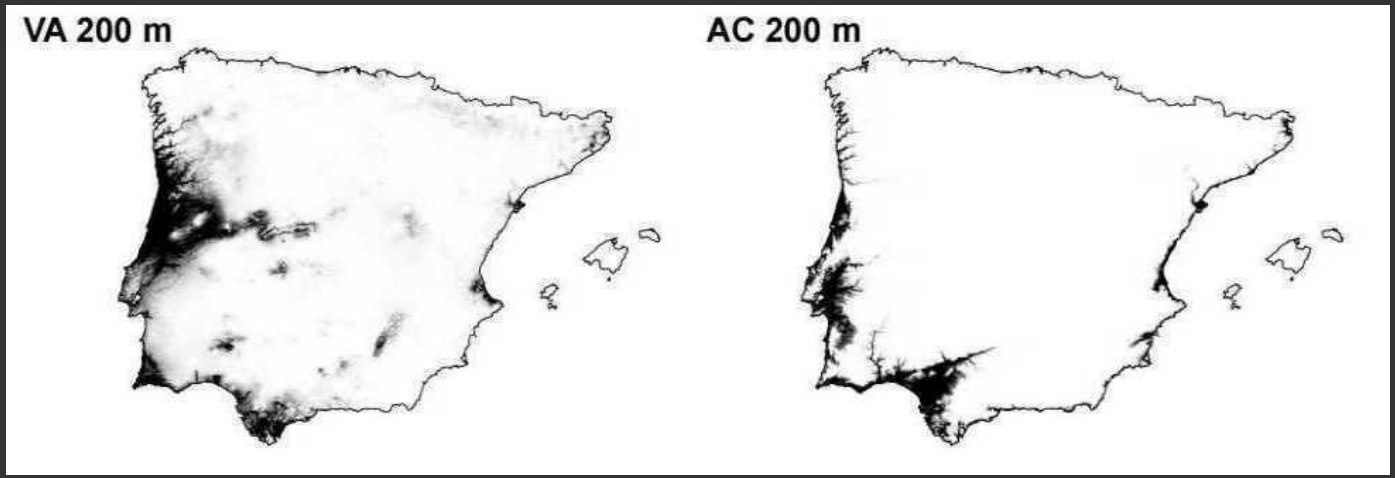
HABITAT SUITABILITY MAPS



Continuous model

Between 0 and 1

0 and 100

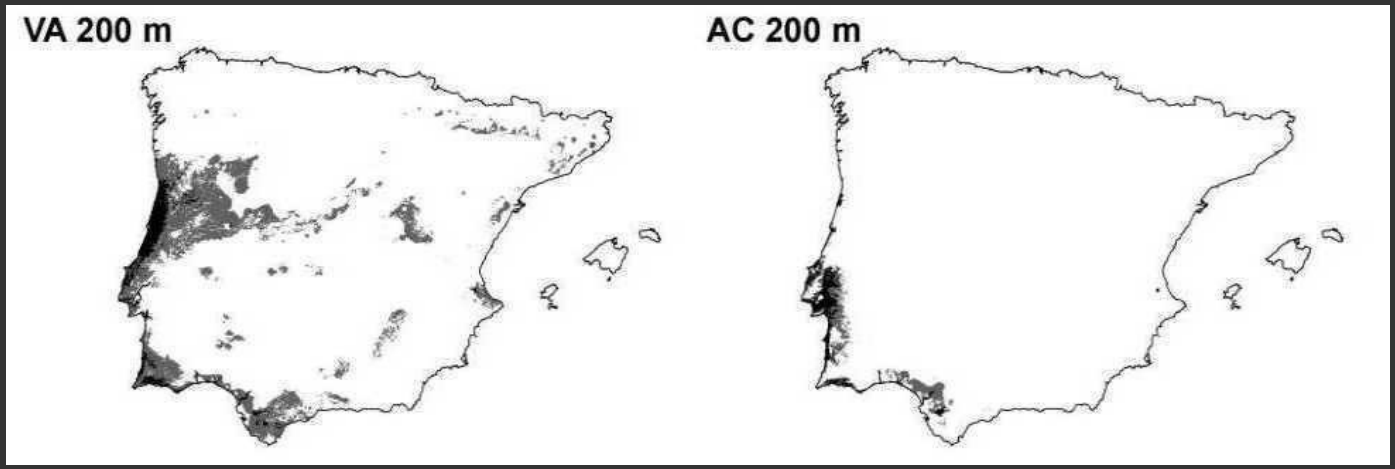


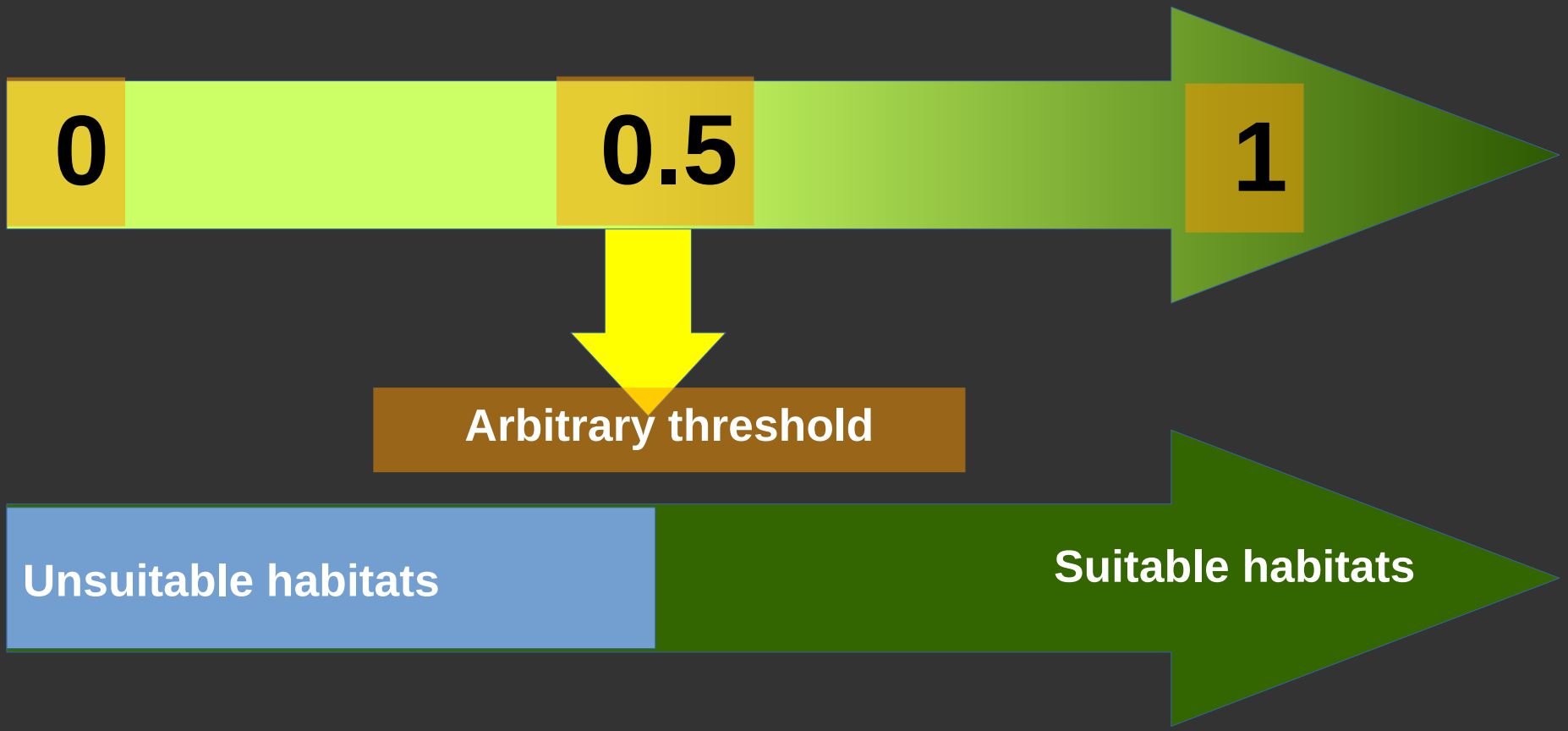
Arbitrary threshold

Categorical model

Suitable habitats

Unsuitable habitats





Presence-absence models

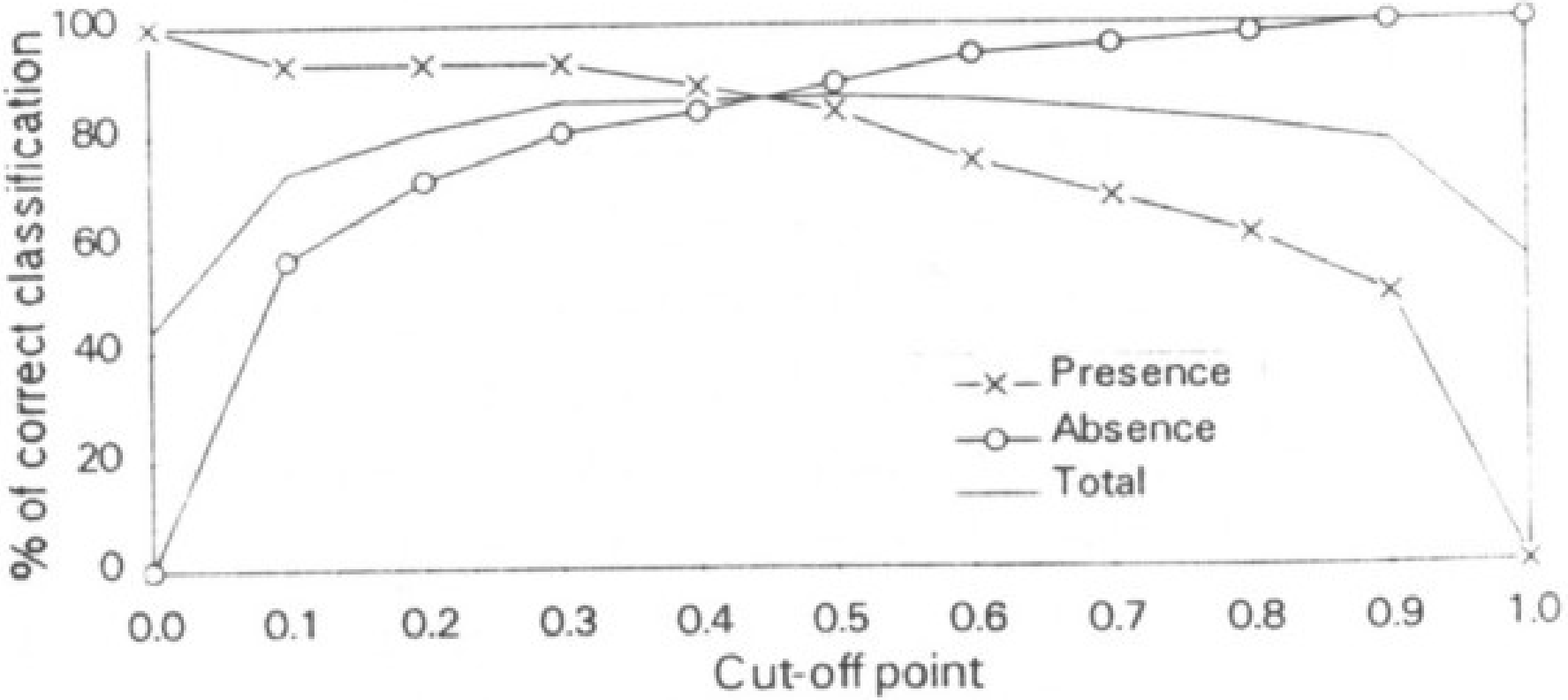
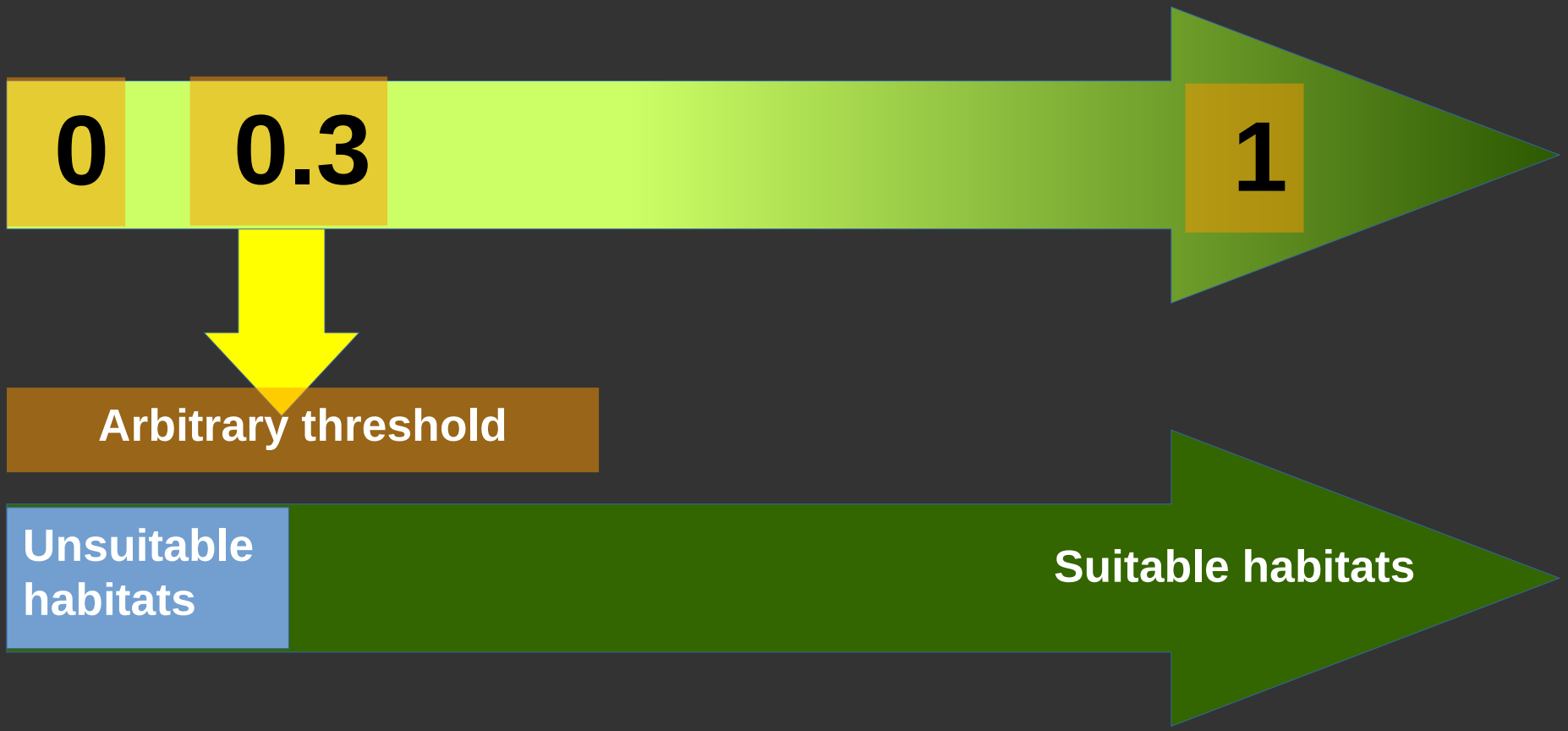


Fig. 3. Correct classification rates, for model 3, considering all possible cut-off points, at 0.1 intervals.

Brito et al 1999



Presence-only models

TYPES OF THRESHOLDS

Maxent

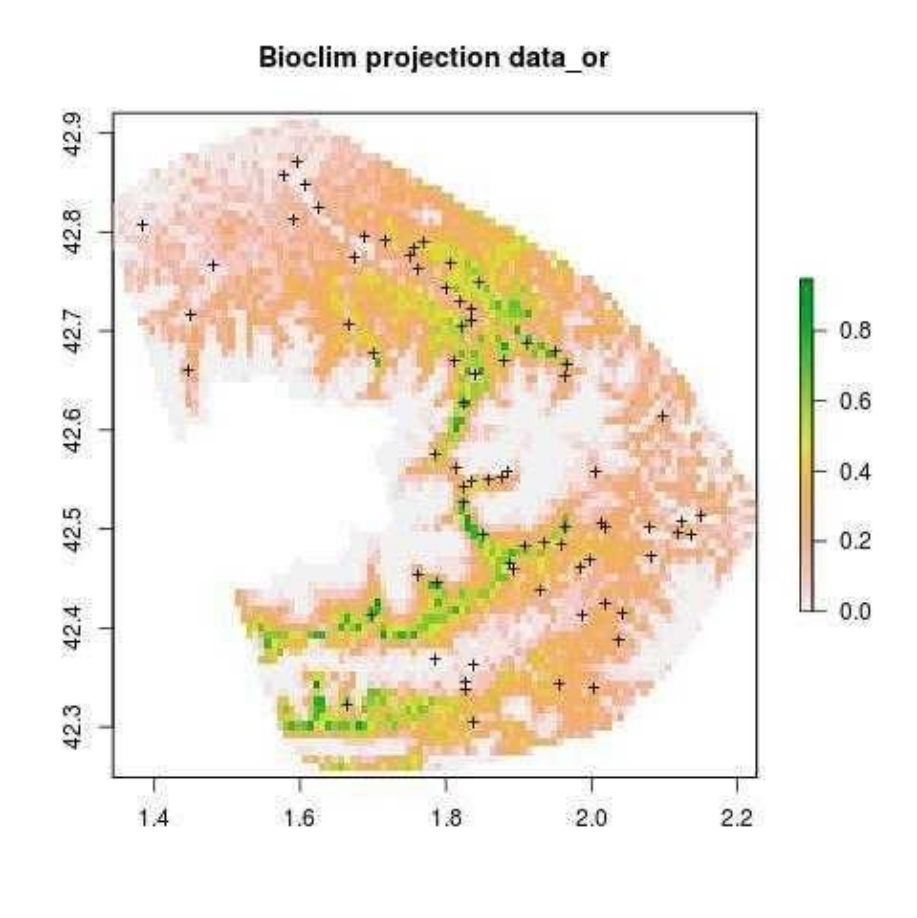
- 10 percentile training presence cumulative threshold
- 10 percentile training presence logistic threshold
- 10 percentile training presence area
- 10 percentile training presence training omission
- 10 percentile training presence test omission
- 10 percentile training presence binomial probability
- Equal training sensitivity and specificity cumulative threshold
- Equal training sensitivity and specificity logistic threshold
- Equal training sensitivity and specificity area
- Equal training sensitivity and specificity training omission
- Equal training sensitivity and specificity test omission
- Equal training sensitivity and specificity binomial probability
- Maximum training sensitivity plus specificity cumulative threshold
- Maximum training sensitivity plus specificity logistic threshold
- Maximum training sensitivity plus specificity area
- Maximum training sensitivity plus specificity training omission
- Maximum training sensitivity plus specificity test omission
- Maximum training sensitivity plus specificity binomial probability

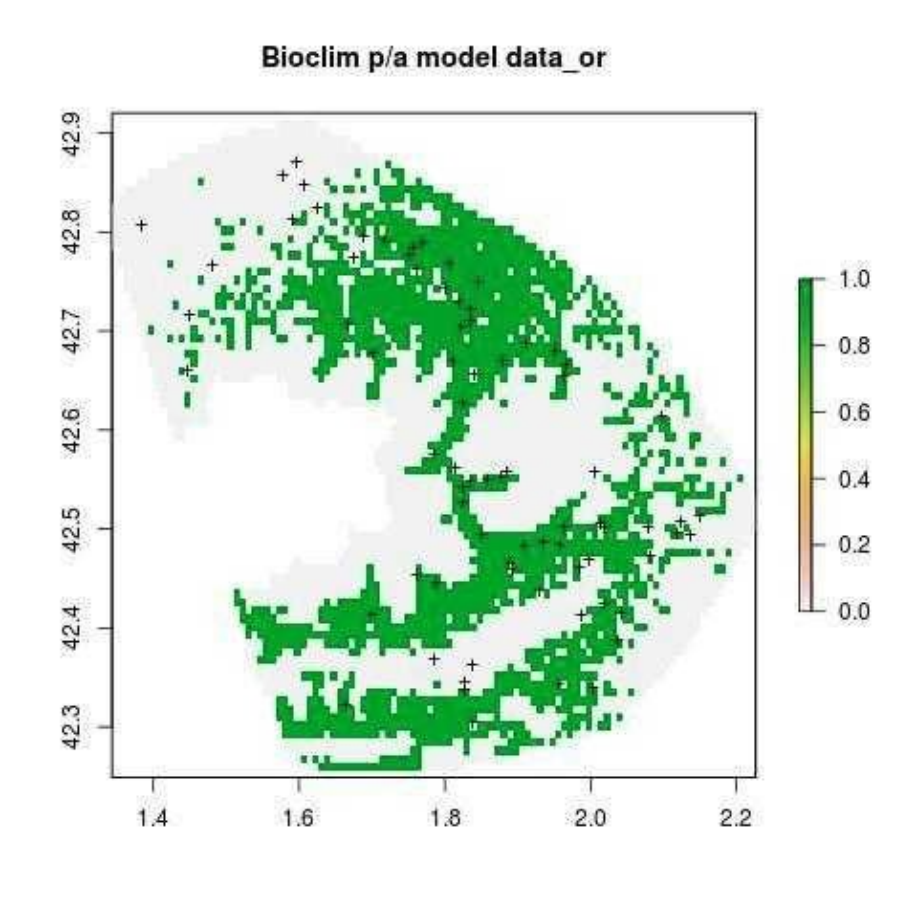
TYPES OF THRESHOLDS

Maxent

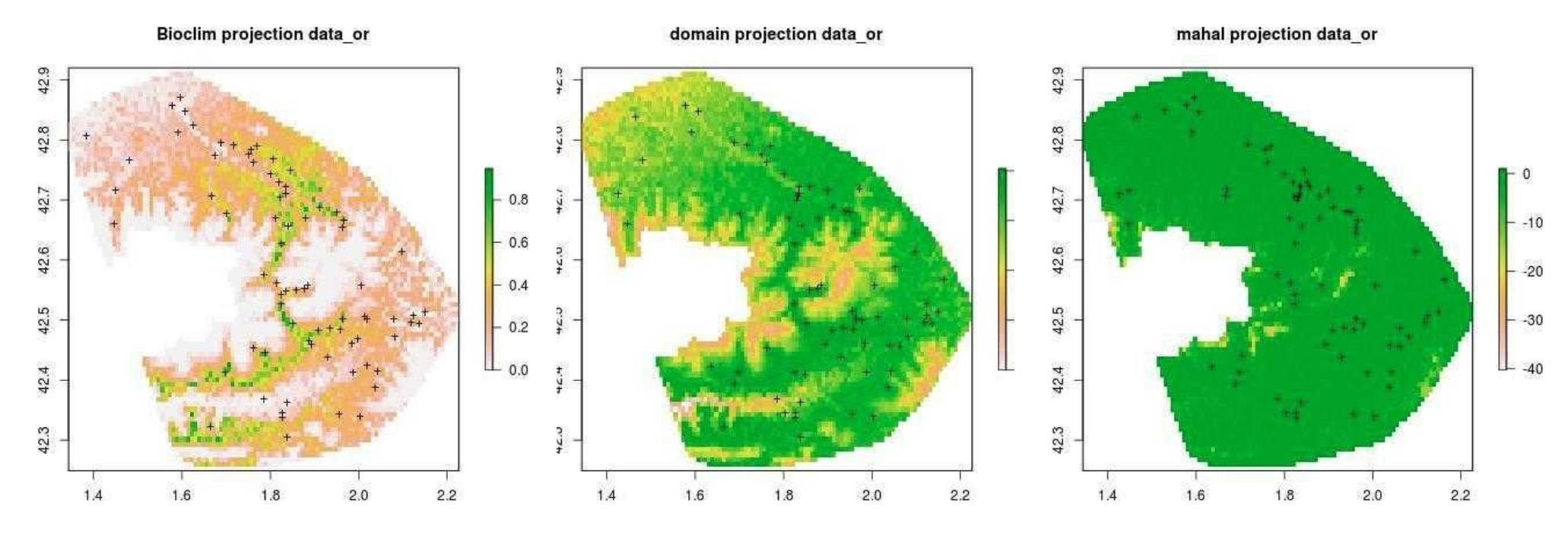
- Equal test sensitivity and specificity cumulative threshold
- Equal test sensitivity and specificity logistic threshold
- Equal test sensitivity and specificity area
- Equal test sensitivity and specificity training omission
- Equal test sensitivity and specificity test omission
- Equal test sensitivity and specificity binomial probability
- Maximum test sensitivity plus specificity cumulative threshold
- Maximum test sensitivity plus specificity logistic threshold
- Maximum test sensitivity plus specificity area
- Maximum test sensitivity plus specificity training omission
- Maximum test sensitivity plus specificity test omission
- Maximum test sensitivity plus specificity binomial probability
- Balance training omission. predicted area and threshold value cumulative threshold
- Balance training omission. predicted area and threshold value logistic threshold
- Balance training omission. predicted area and threshold value area
- Balance training omission. predicted area and threshold value training omission
- Balance training omission. predicted area and threshold value test omission
- Balance training omission. predicted area and threshold value binomial probability

HABITAT SUITABILITY MAPS

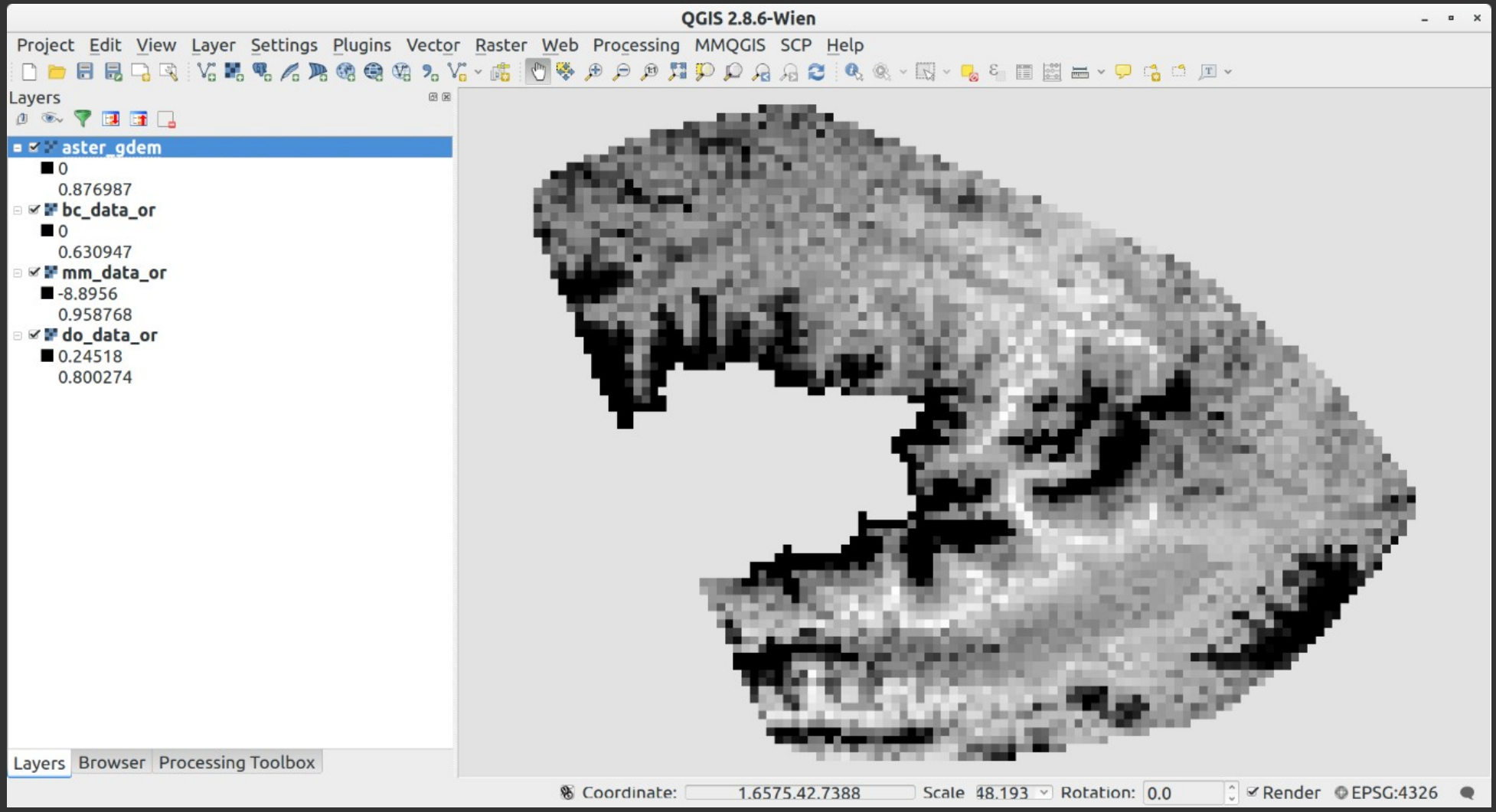




34. If you performed several models, you can calculate an ensemble model by averaging them (Araujo and New 2007, Marmion et al. 2009).



34. If you performed several models, you can calculate an ensemble model by averaging them (Araujo and New 2007, Marmion et al. 2009).



35. The outputs can be considered under two dispersion models: null and total dispersion scenarios.

- **Null dispersion model: the species is not able to disperse to new areas. Thus, you must not consider those new suitable areas without observed presence of the species.**

- **Total dispersion model: the species has an unlimited capacity to disperse to new areas. Thus, you can consider those new suitable areas without observed presence of the species.**

QUESTIONS?