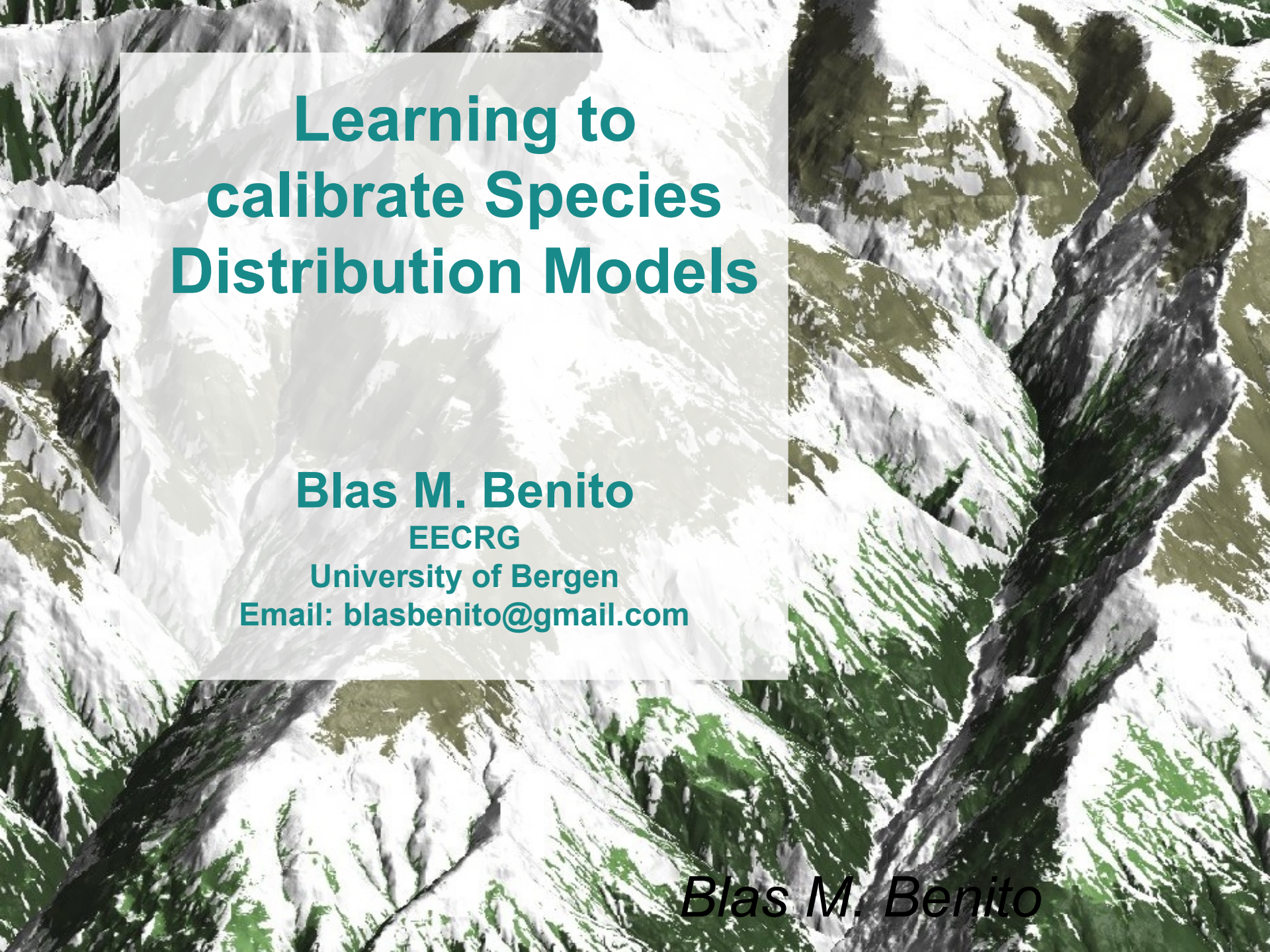
Learning to calibrate Species Distribution Models

Blas M. Benito
EECRG
University of Bergen
Email: blasbenito@gmail.com



CONTENTS

- MODELLING METHODS: THEORY AND PRACTICE
- MODEL EVALUATION
- APPLYING THRESHOLDS
- PROJECTION IN TIME AND SPACE

MODELLING ALGORITHMS

BIOCLIM

AND

DOMAIN

(explanation on code)

REGRESSION METHODS: GLM, GAM, and MARS

GENERALIZED LINEAR MODELS

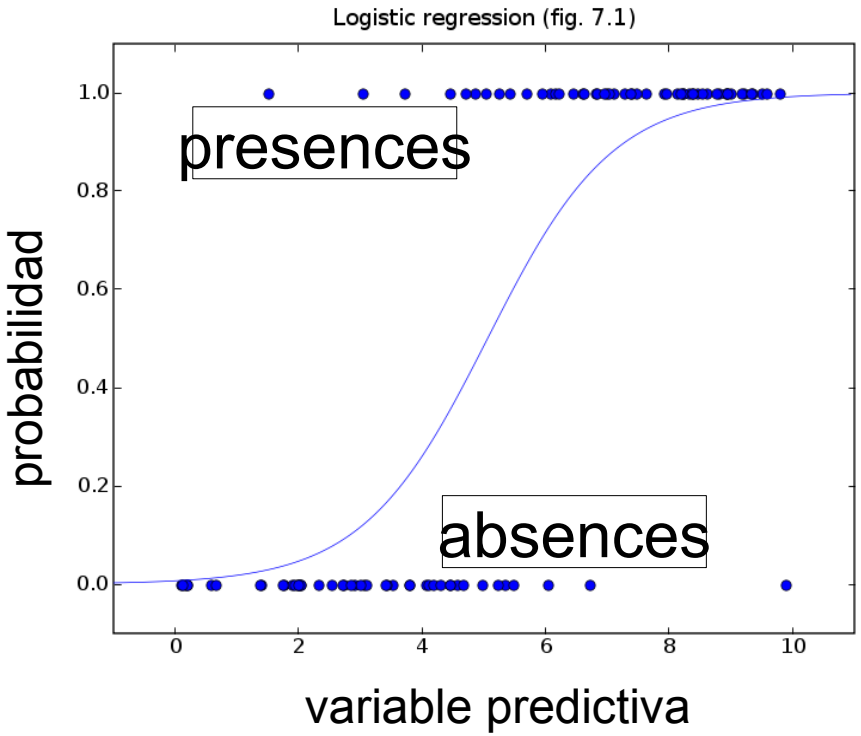
- Allow to fit curvilinear responses
- Residuals can follow different distributions: normal, **binomial**, **Poisson**, negative binomial, gamma.

LOGISTIC REGRESSSION

Particular case of GLMs

- Binomial data (1 - presence; 0 - absence)
- Family of errors is binom
- Link function
  - Logit

$$v\,dep = \ln\left(\frac{PL}{1-PL}\right)$$



FLEXIBILITY

- Curve complexity
  - Logistic
  - Polynomials (2nd, 3th degrees)
- Presence data
  - Absence
  - Pseudoabsence
  - Background
- Variable interactions
  - Additive model only (no interactions)
  - Variable interactions

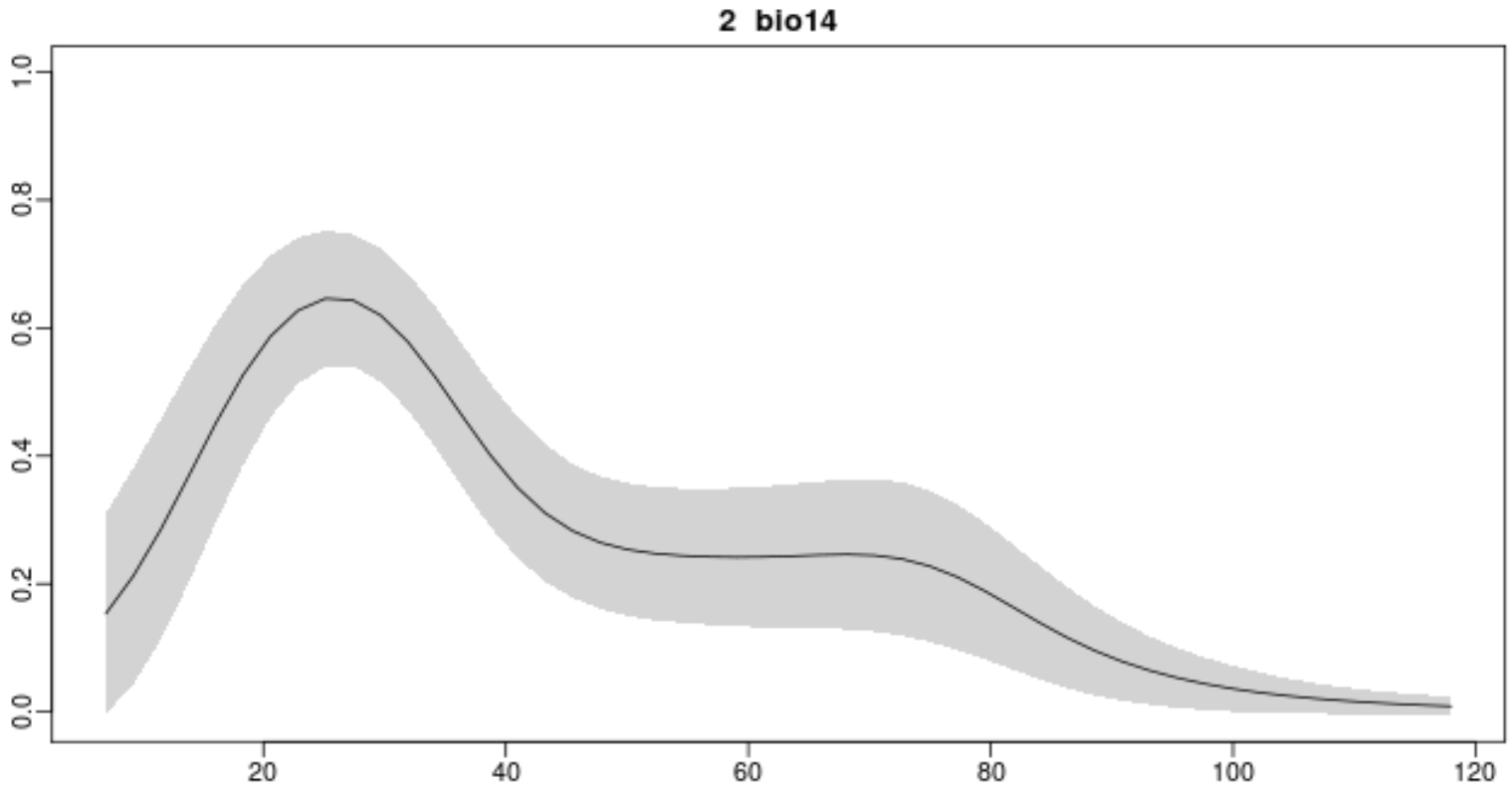
MINIMUM NUMBER OF PRESENCES

- 5 presences and y 5 absences (if so) for each predictor, polynomial term (for second degree polynomials every predictor has 2 polynomial terms), and interaction term (interactions among predictors).

GENERALIZED ADDITIVE MODELS (GAM)

- Extension of GLMs
- Non parametric regression
- Fits splines (piecewise polynomial curves) locally to the predictors: smoothed predictors
- Splines can be more or less flexible depending on the degrees of freedom (k)
- Useful to model non-linear responses
- Requires more input data than GLM
- To interpret variable contribution is trickier than in GLMs

GENERALIZED ADDITIVE MODELS



MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

- Non parametric regression method
- Handles non-linear data perfectly
- Accounts for (partial or complete) variable interactions
- It's resulting equations are pretty easy to understand
- Fast on big datasets
- Frequently used for the forecasting of time series in economy

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

- How does it work?
  - **Hinge functions** are key to fit non linear data
  - Hinge functions form **piecewise linear functions**
  - **Variable interactions** are represented by multiplicating hinge functions, which generates **non-linear functions**

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

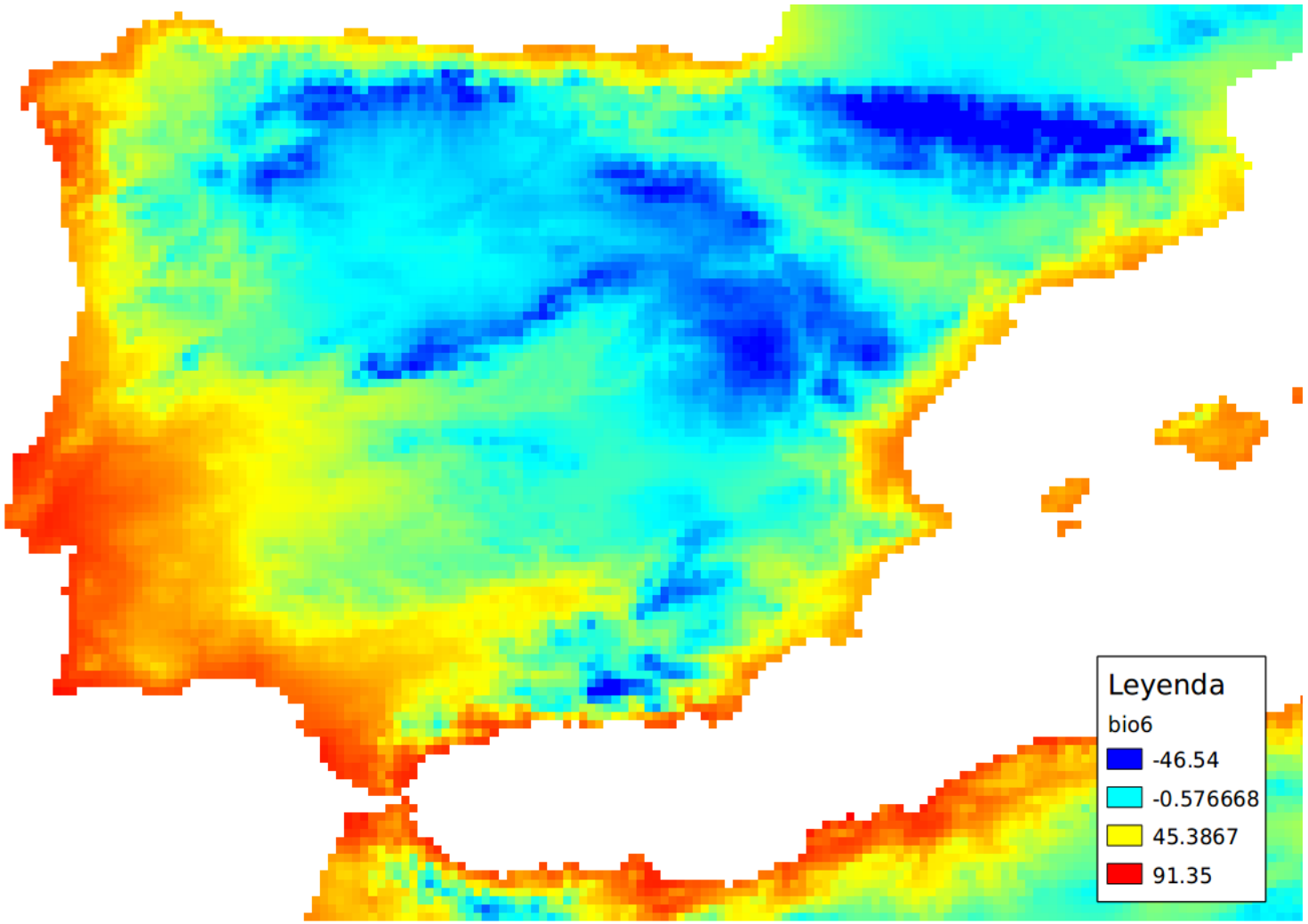
- Building the model
  - Forward pass
    - It fits all the hinge functions with the minimum residual error (overfitting!)
    - The process stops when the residual error cannot be minimized anymore, or the maximum number of terms has been reached
  - Backwards pass
    - Examinates the contribution of each individual term to the model, and removes the non-significant ones (model prunning and generalization)

MAXENT

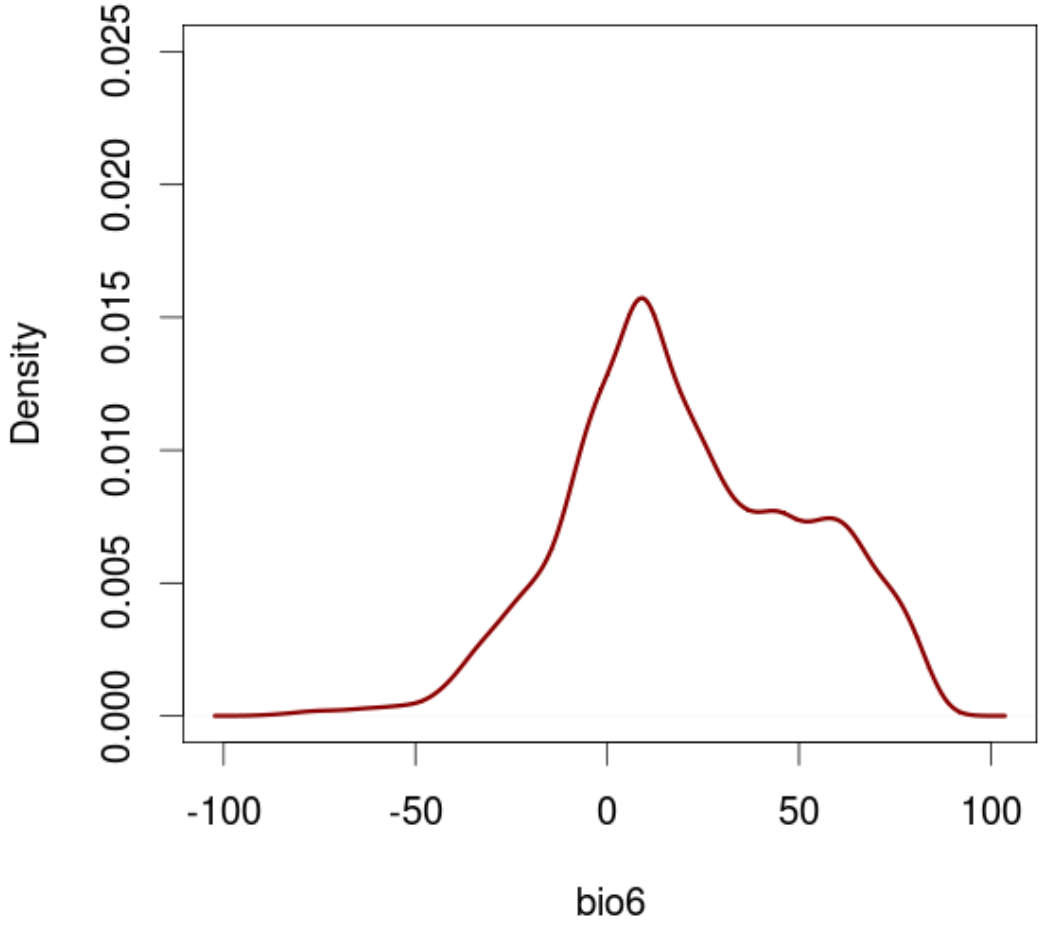
- Poisson regression with Lasso penalization
- Can work with a low number of presences
- Requires background data
- Model complexity can be controlled through the regularization multiplier
- It has a Java application that allows to fit and analyze the models
- We will be using the “maxnet” package

PREDICTOR

Bio6 → temperature of the coldest month



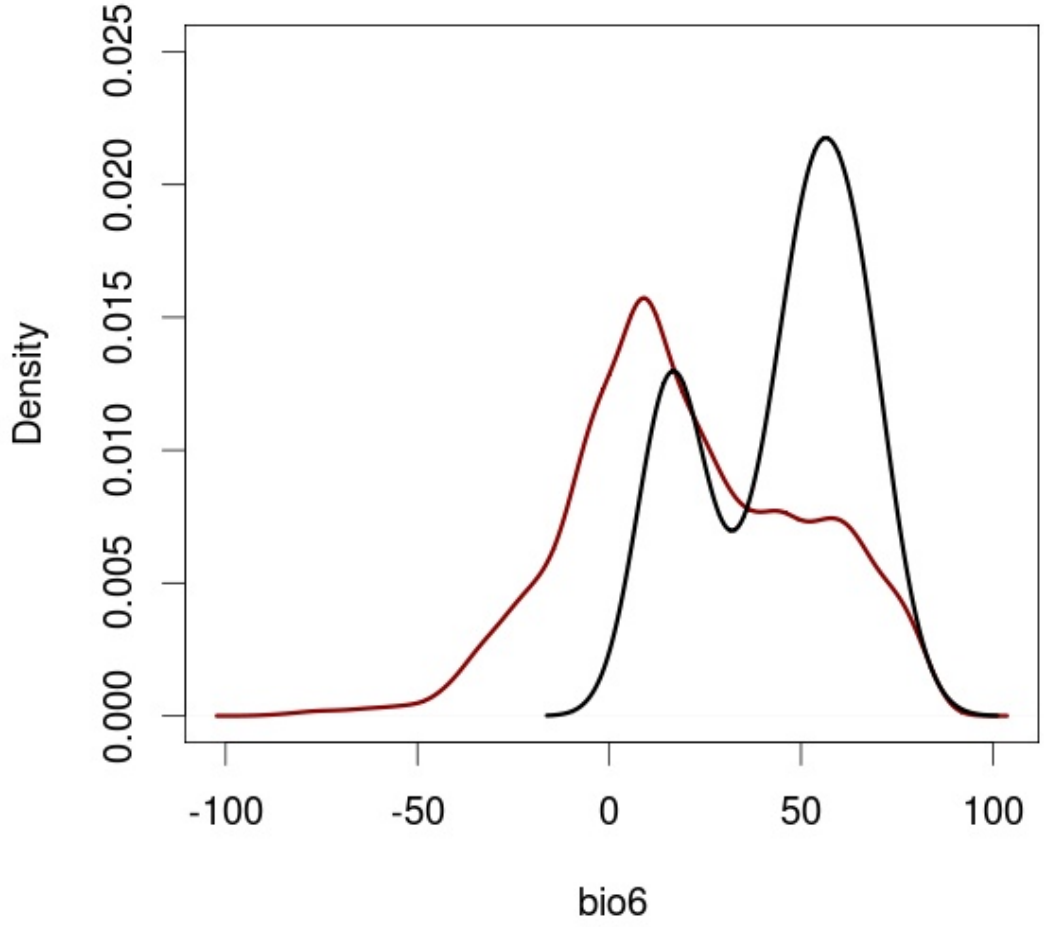
DENSITY OF THE BACKGROUND



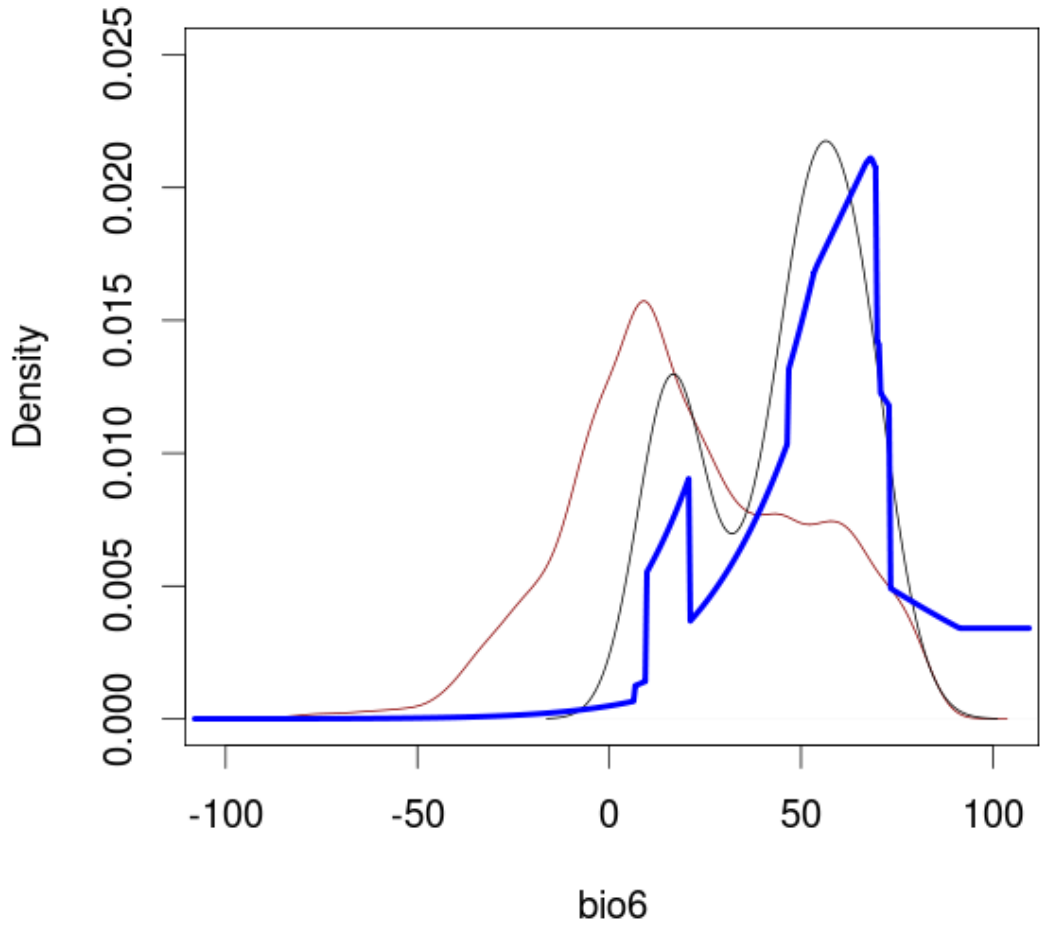
PRESENCE

Atelerix algirus

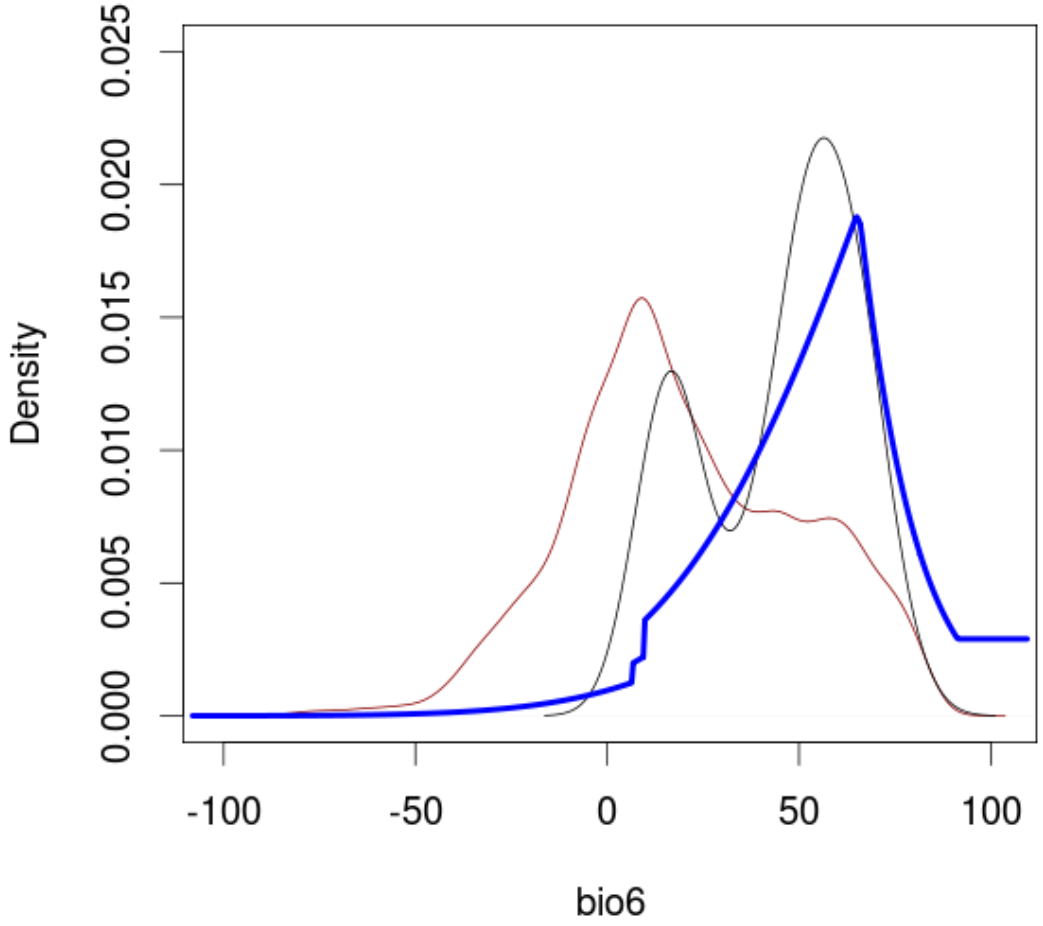
DENSITY OF THE PRESENCE



MAXENT FIT (max complexity)

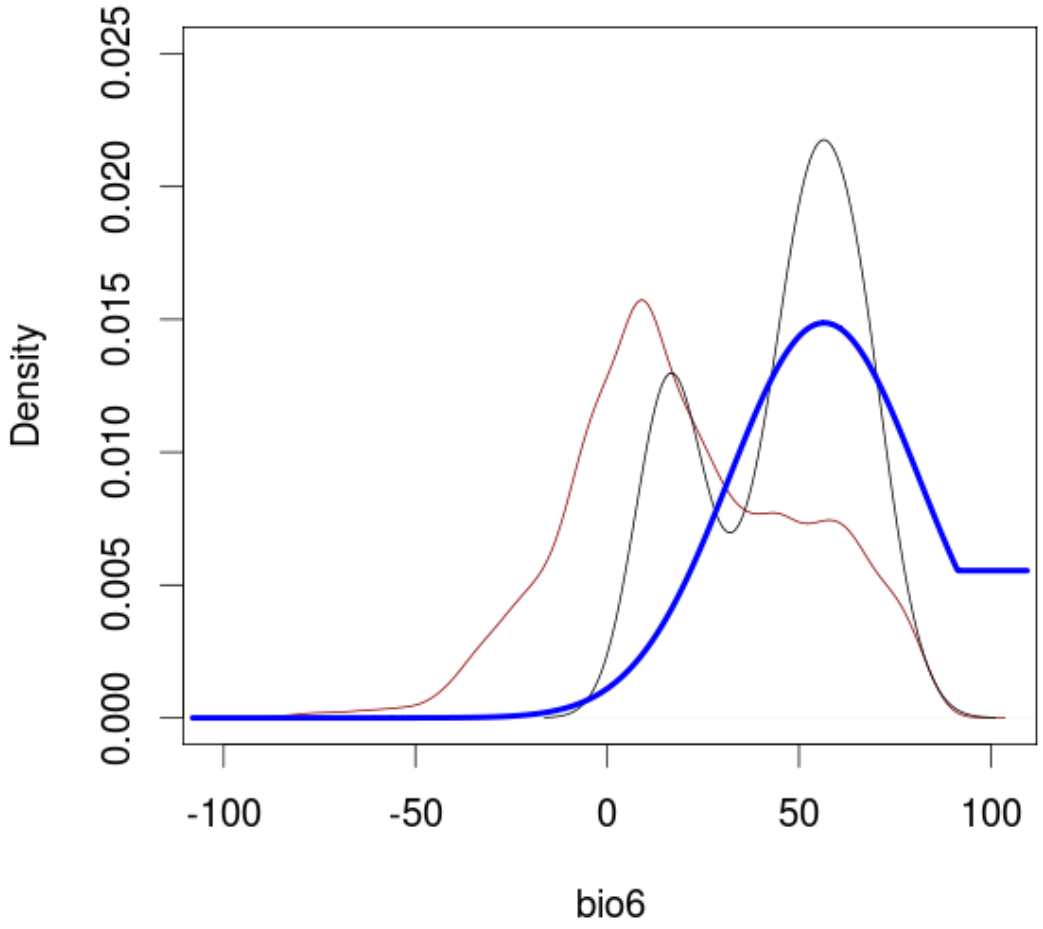


MAXENT FIT



Regularization multiplier = 3

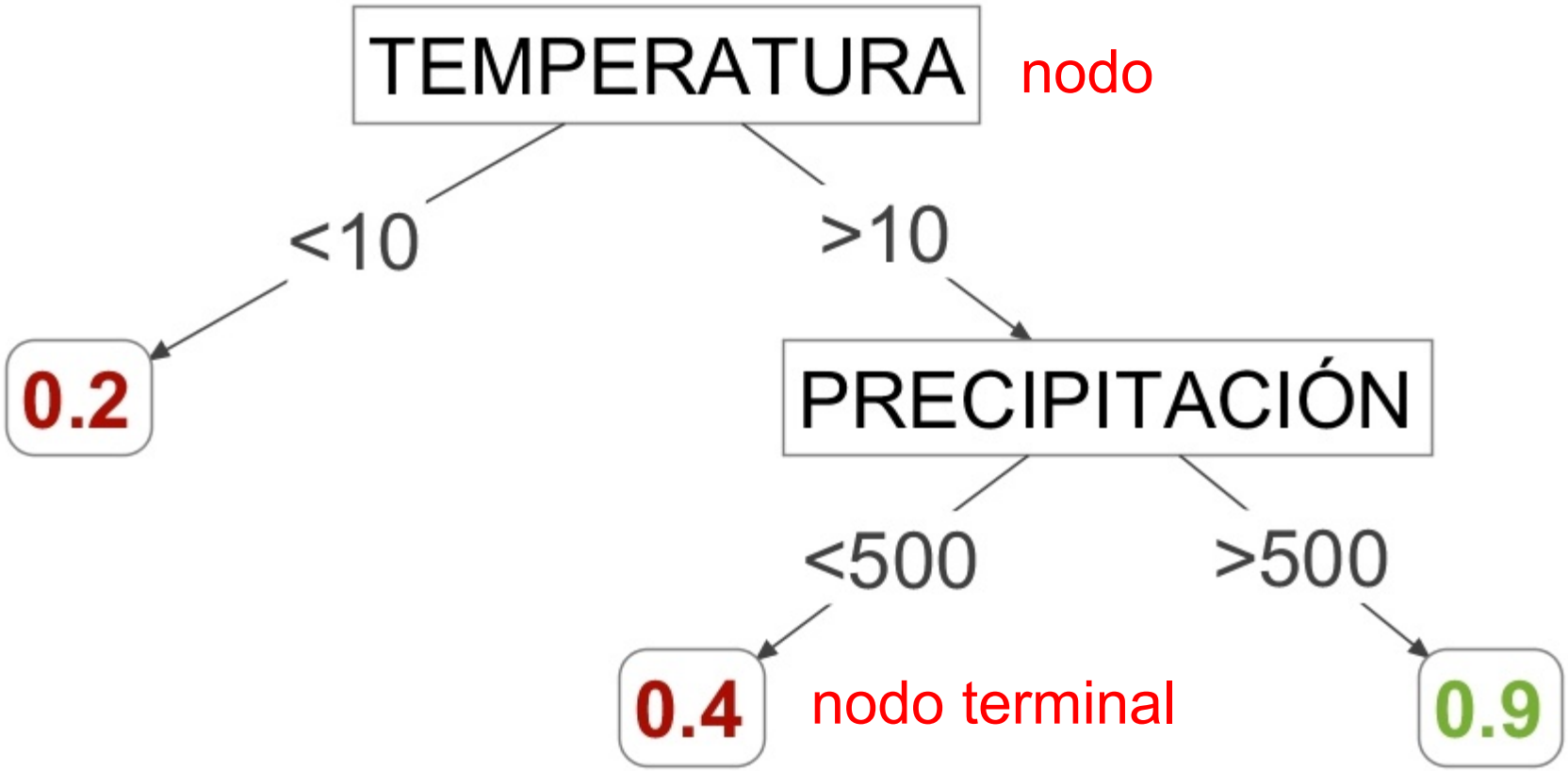
MAXENT FIT



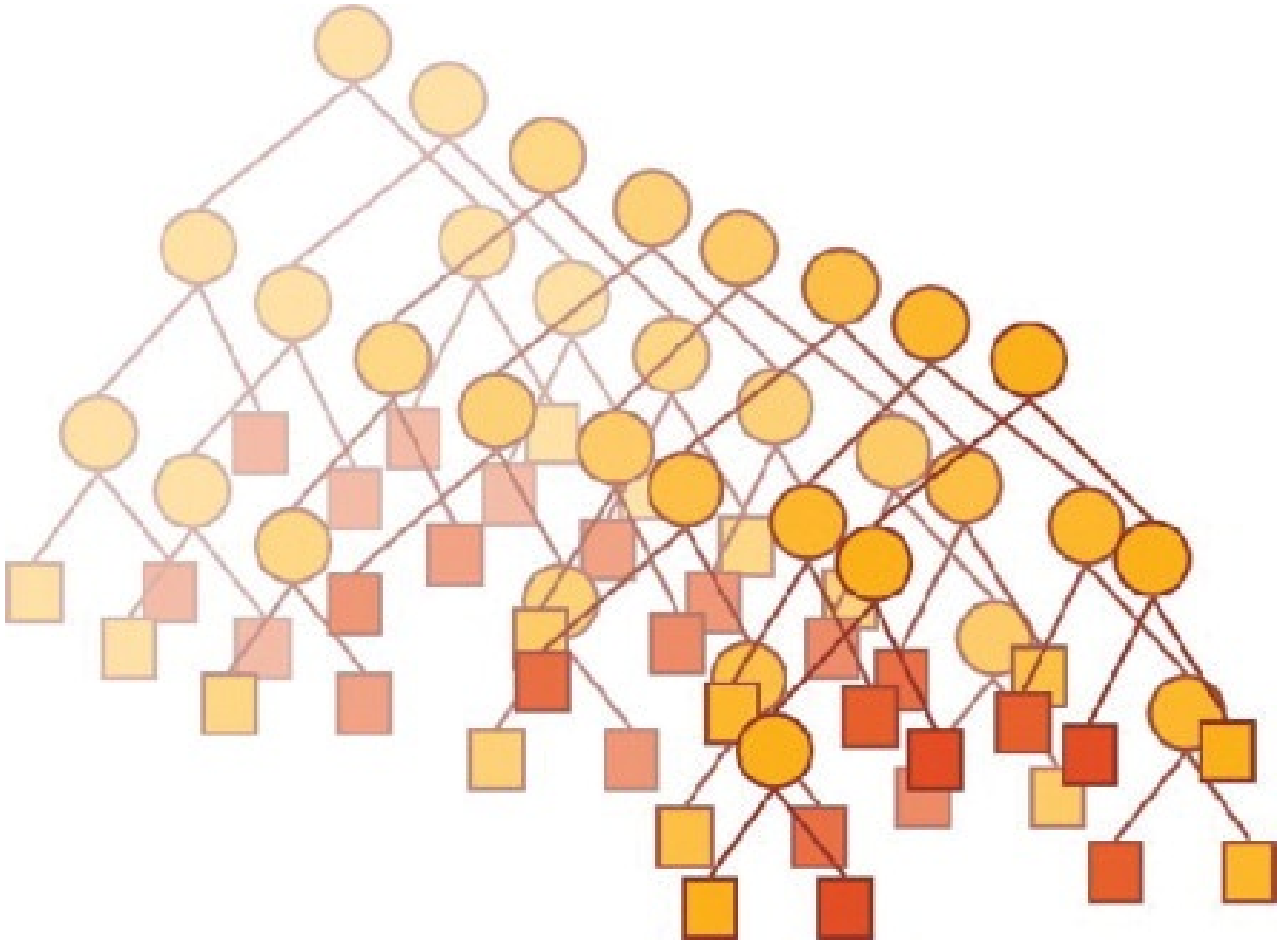
Regularization multiplier = 6

TREE-BASED MODELS

REGRESSION TREES



RANDOM FOREST



Fuente: Gedeck et al. 2010 Progress in Medicinal Chemistry

RANDOM FOREST

- Relevant parameters:
  - **ntree**: number of trees in the forest
  - **mtry**: number of predictors used to fit each tree
  - **nodesize**: minimum number of cases per terminal node
  - **maxnode**: maximum number of terminal nodes

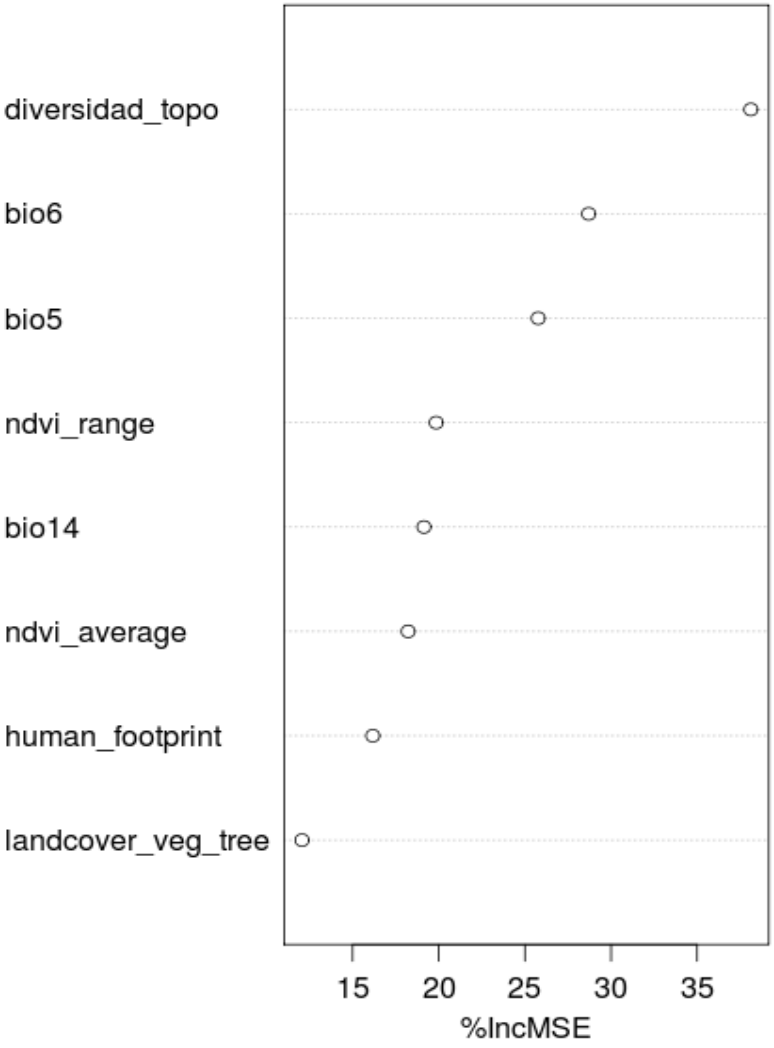
RANDOM FOREST

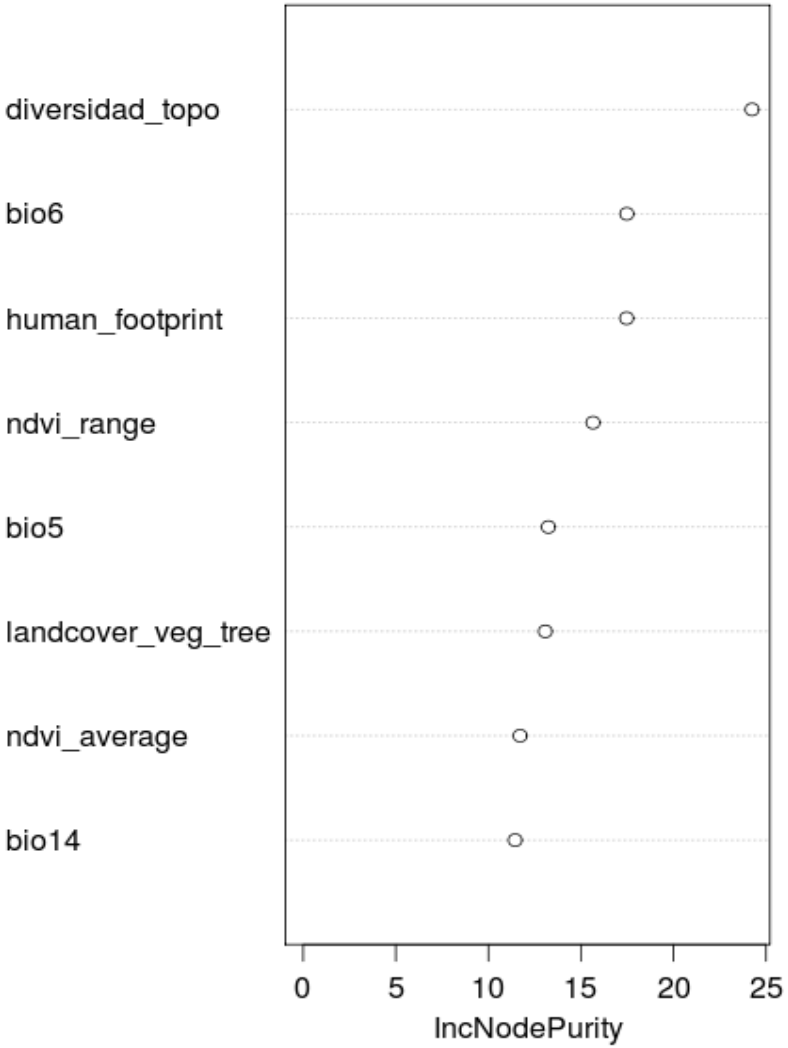
1. Per tree:
    1. Select **n** predictors randomly
    2. Select 60% of the data randomly
    3. Fit a regression tree
    4. Evaluate the tree with the 40% of data not used ot fit it (out-of-bag data).
2. Once all trees have been fitted:
    1. Use all trees to classify a sample (new data)
    2. The statistical mode across trees for that sample will be the final prediction

RANDOM FOREST

- Pros
  - Wide range of applications
  - Can handle big datasets
  - Great tool to study variable importance
- Cons
  - Overfitting
  - Interactions are hidden
  - Non linear responses are difficult to interpret
  - Stochastic (slightly different results each time you run it)

IMPORTANCE OF PREDICTORS





BOOSTED REGRESSION TREES

- Also named “gradient boosting”.
- The acronyms are BRT and GBM.
- Check “A working guide to Boosted Regression Trees” (Elith_2008.pdf in the papers folder).
- The vignettes of the “dismo” package are also a good starting point to work with this method.

BOOSTED REGRESSION TREES

- Some features:
  - It grows regression trees
  - Stochastic (as Random Forest)
  - “Boosting” (whoa!) optimization method to reduce error (explained or so in next slide)
  - Selects relevant predictors
  - Final model is linear combination of many trees
  - Allows to evaluate interactions among predictors
  - Allows to define error distributions and link functions (like GLM)

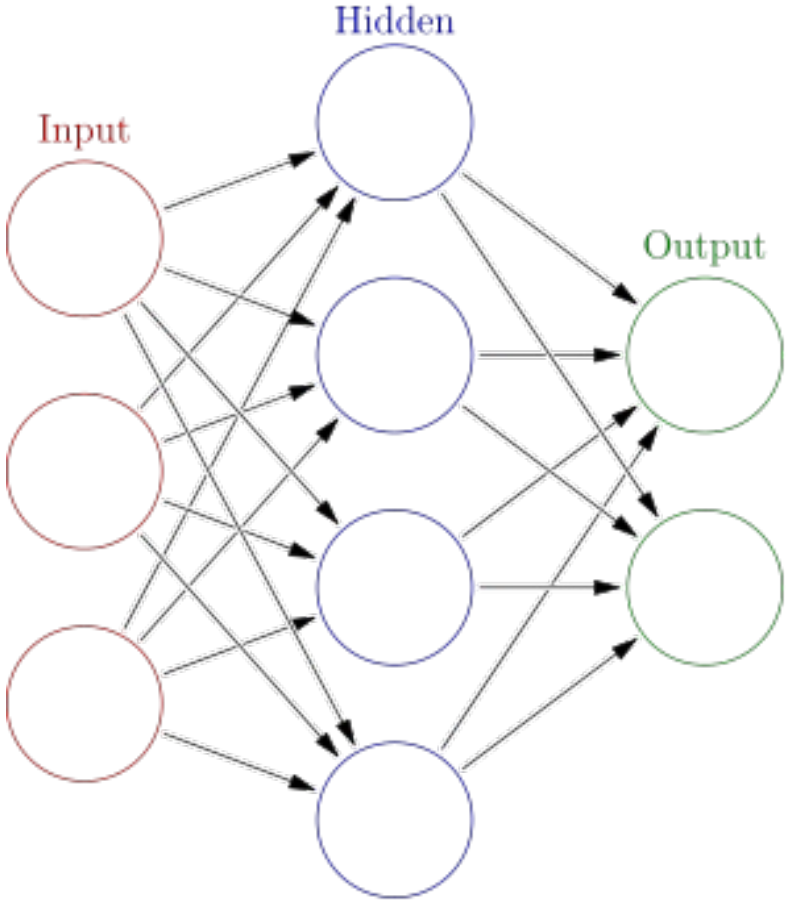
BOOSTING

- **Loss function:** measures lost in predictive performance due to suboptimal models
- **Boosting**: minimizes the loss function. HOW?
  - Grows tree (**t1**) of given size that minimizes the loss function the most
  - Grows the tree (**t2**) that better fits to the residuals of **t1** (variation not explained by the model yet).
  - Adds **t2** to the model, and the residuals of **t1 + t2** are computed.
  - Grows the tree (**t3**) that better fits to the residuals of **t1 + t2.**
  - Rinse and repeat until the loss function reaches an asymptote.

ARTIFICIAL NEURAL NETWORKS

ARTIFICIAL NEURAL NETWORKS

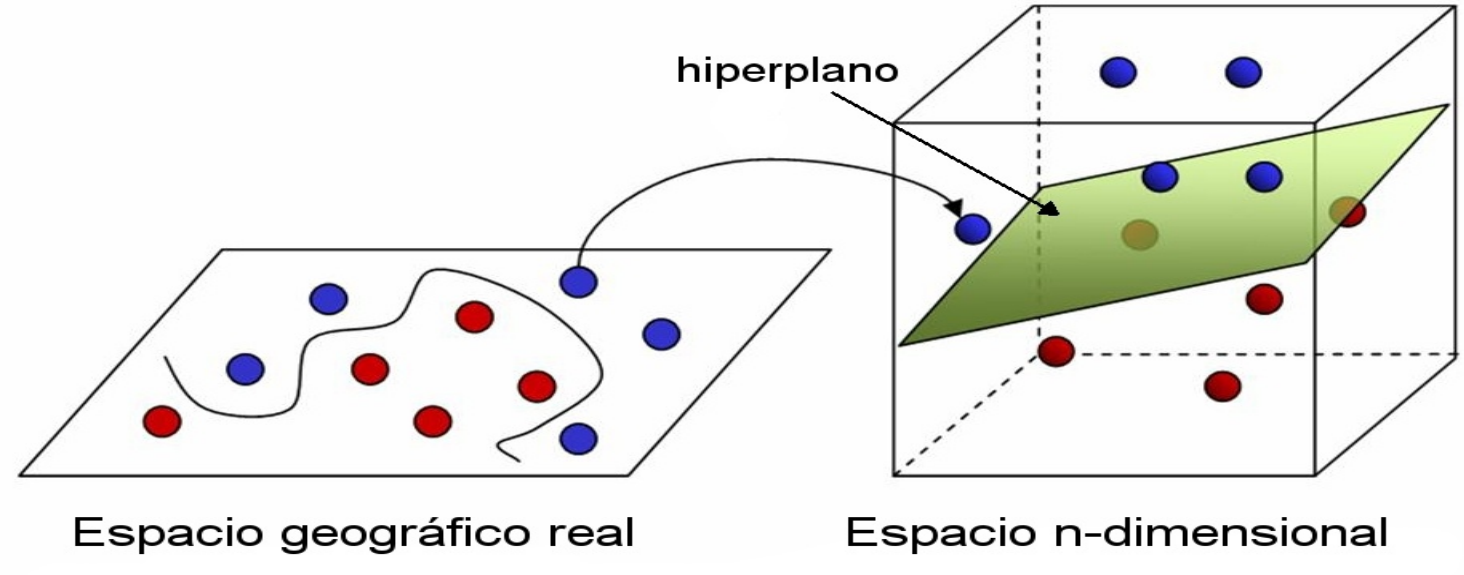
- Organized in layers
  - Input: predictors
  - Output: response
  - Hidden: linear functions relating both
- Each link has a weight (represents the intensity and sign of the relationship)
- Learning: computation of weights that minimize error (coefficients!)



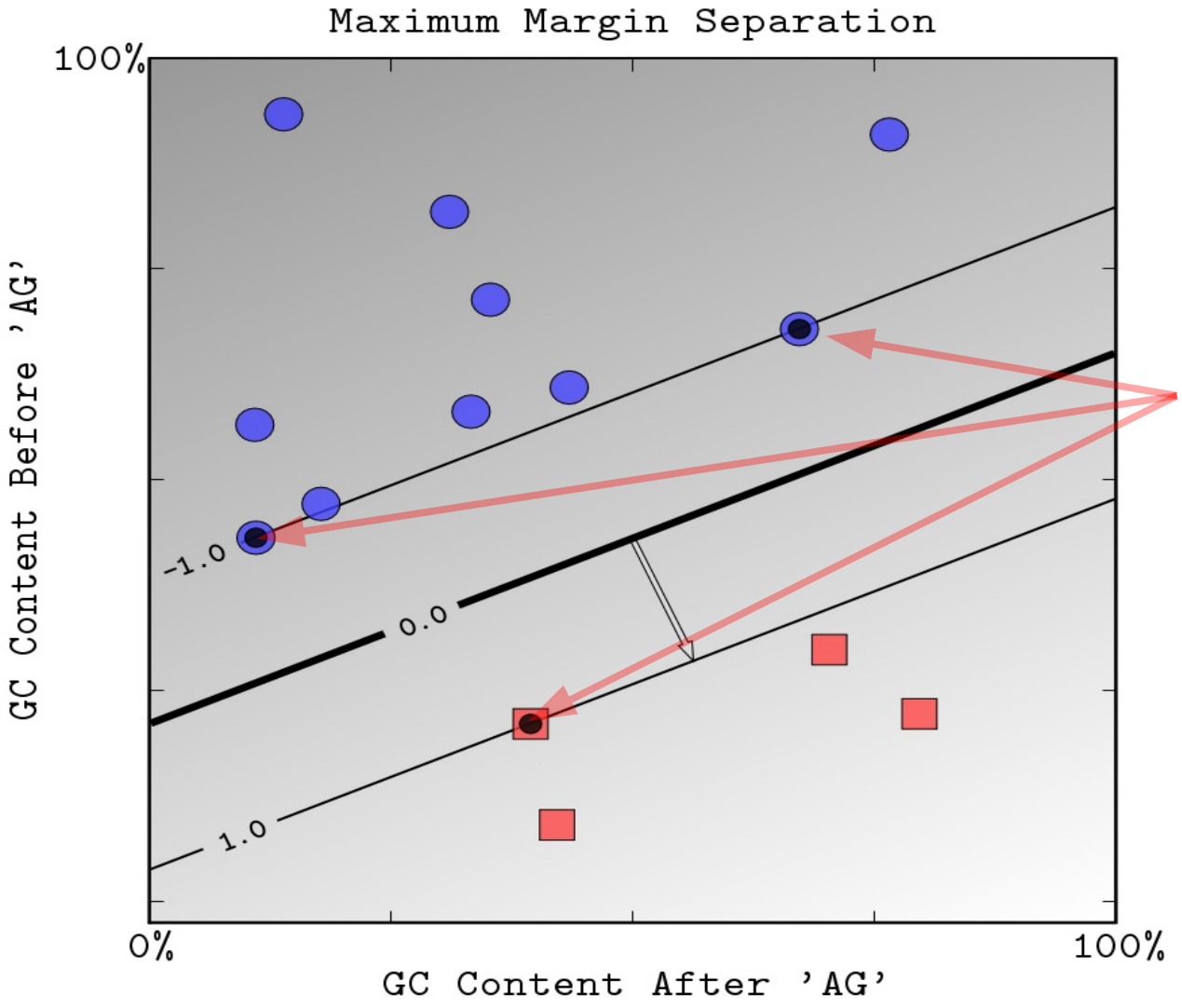
SUPPORT VECTOR MACHINES

SUPPORT VECTOR MACHINES

- Classification in n-dimensional spaces
- Separates groups using hiperplanes
- Data points get ranked depending on their distance to the hiperplane



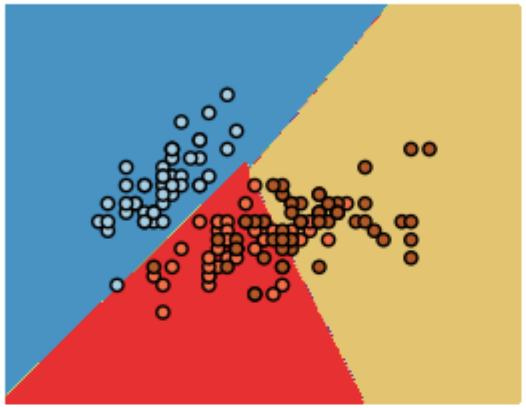
SUPPORT VECTOR MACHINES



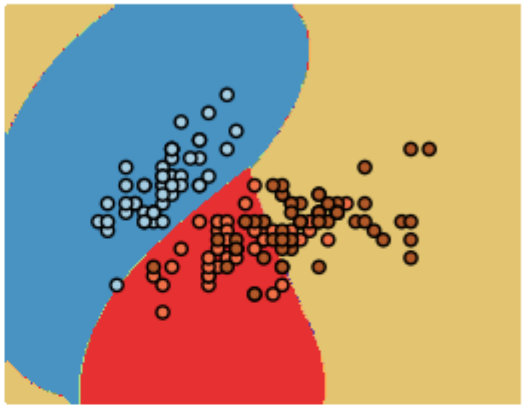
Frontier points

SUPPORT VECTOR MACHINES

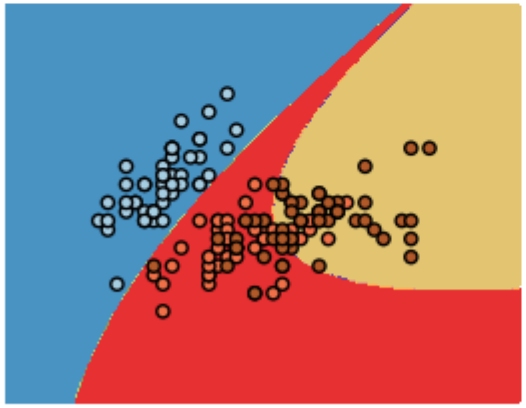
SVC with linear kernel



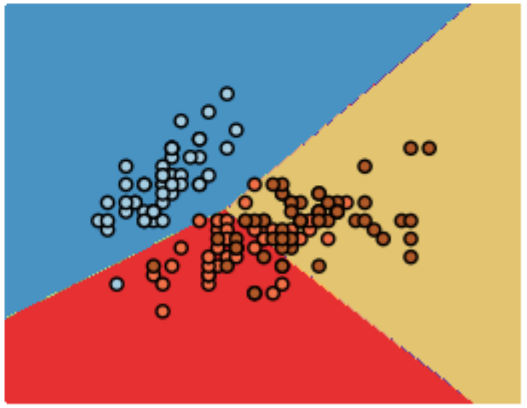
SVC with RBF kernel



SVC with polynomial (degree 3) kernel



LinearSVC (linear kernel)



ENSEMBLE MODEL FORECASTING

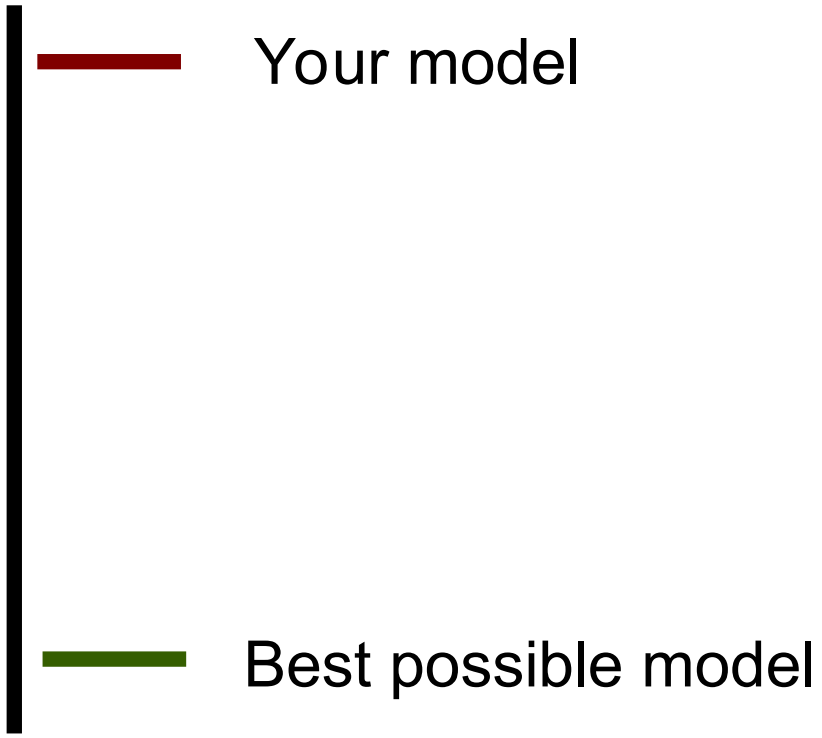
ENSEMBLE

- JW Gibbs (1878): Many “copies” of a system can be considered at once, representing each copy a possible state of the system.
- JM Bates y CWJ Granger (1969): An ensemble has a lower error probability than any of its individual components.
- Araújo & New 2006: When averaging SDMs, the target signal emerges above the noise coming from the errors and uncertainties of the individual models.

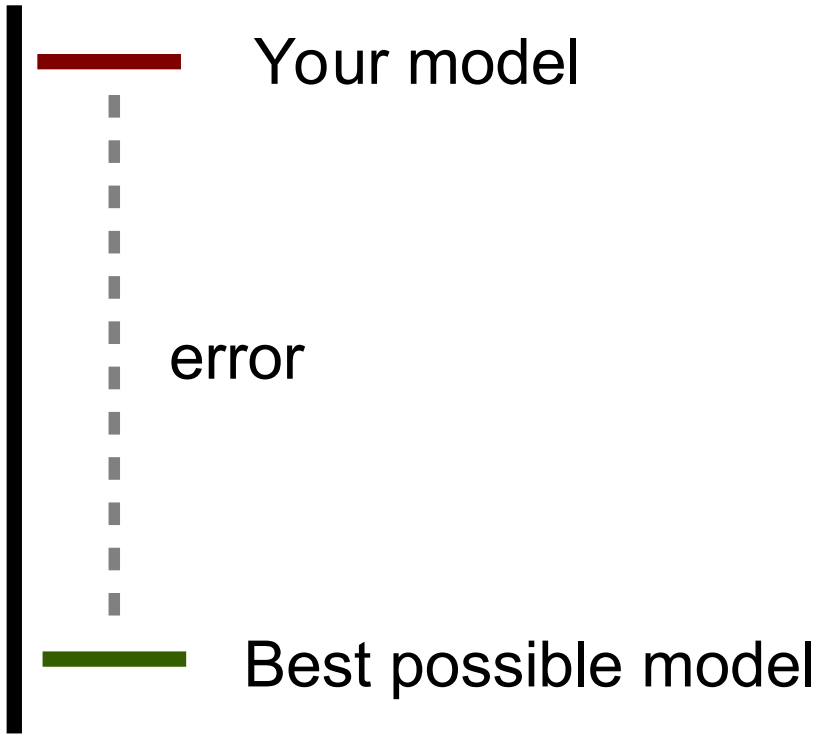
ENSAMBLADO

ENSAMBLADO

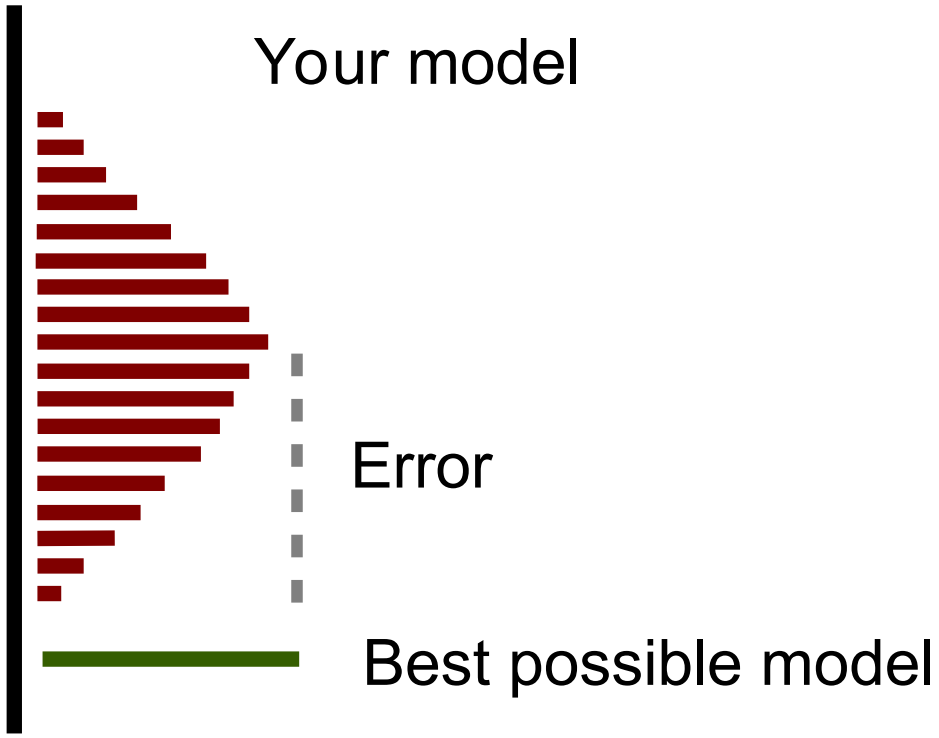
ENSAMBLADO



ENSAMBLADO



ENSAMBLADO



METHODS

- Median **
- **Mean** ** (buen método, Marmion 2009)
- Weighted mean (according AUC values) **
- Selection of models with highest AUC
- PCA: median of the models more correlated with the first component.

WARNING: when using several models, make sure that all them have the same output scale [0, 1], [0, 100], etc.

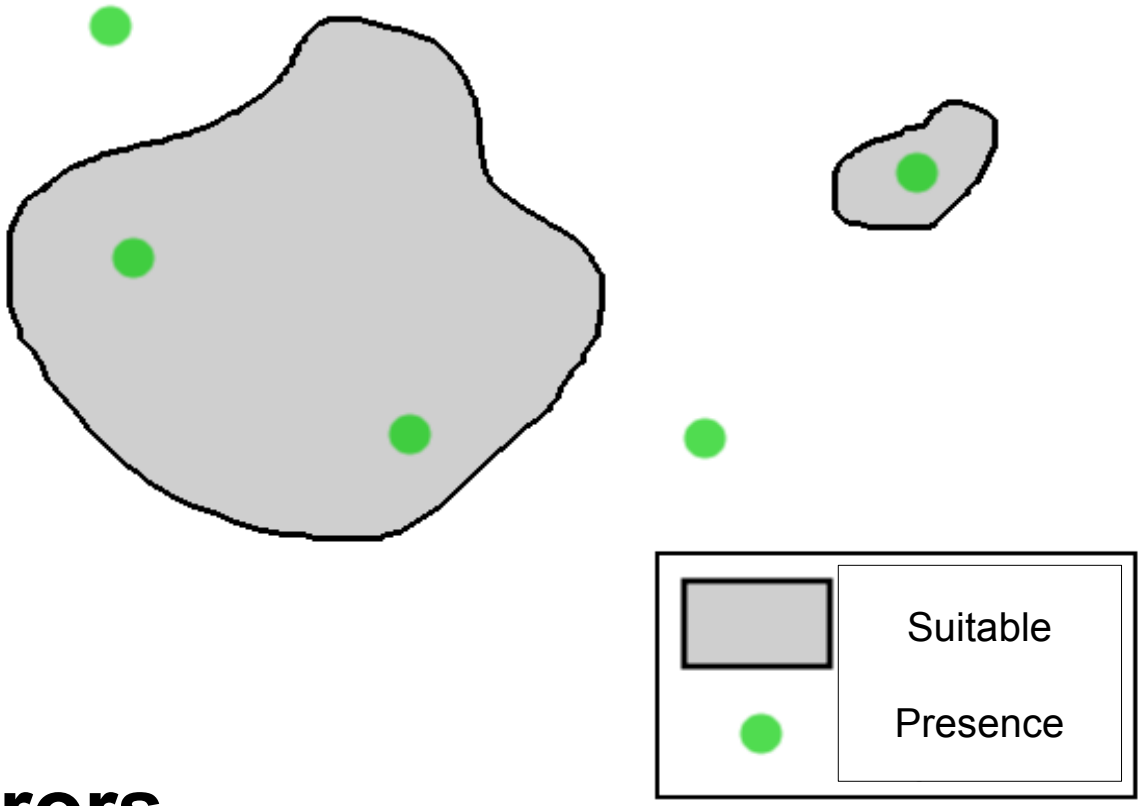
MODEL EVALUATION

KEY PAPER

Fielding AH y Bell JF 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24(1), 38-49

EVALUATING PRESENCE-ONLY BINARY MODELS

EVALUATION



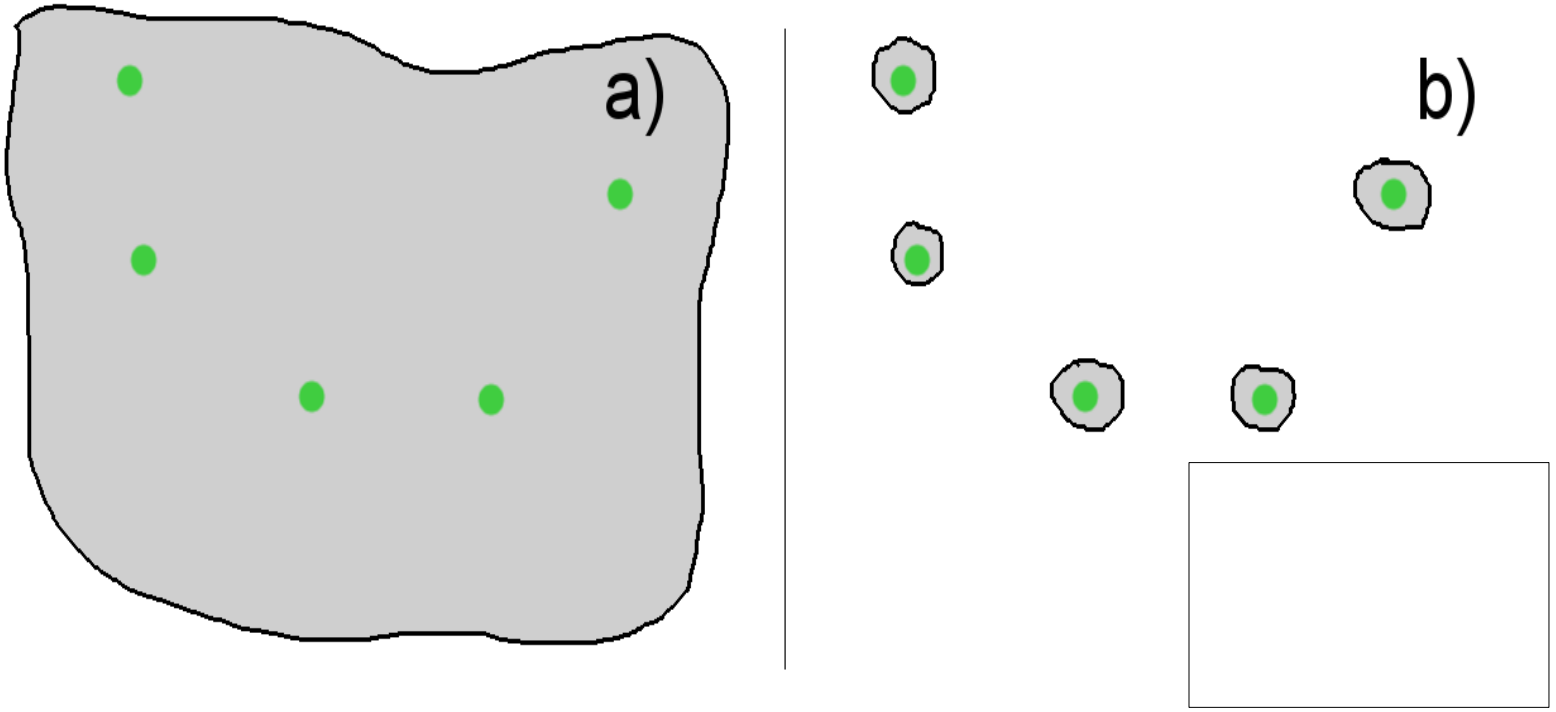
5 presences

3 successes

Sensitivity=0,6

2 **omission errors**

EVALUATION



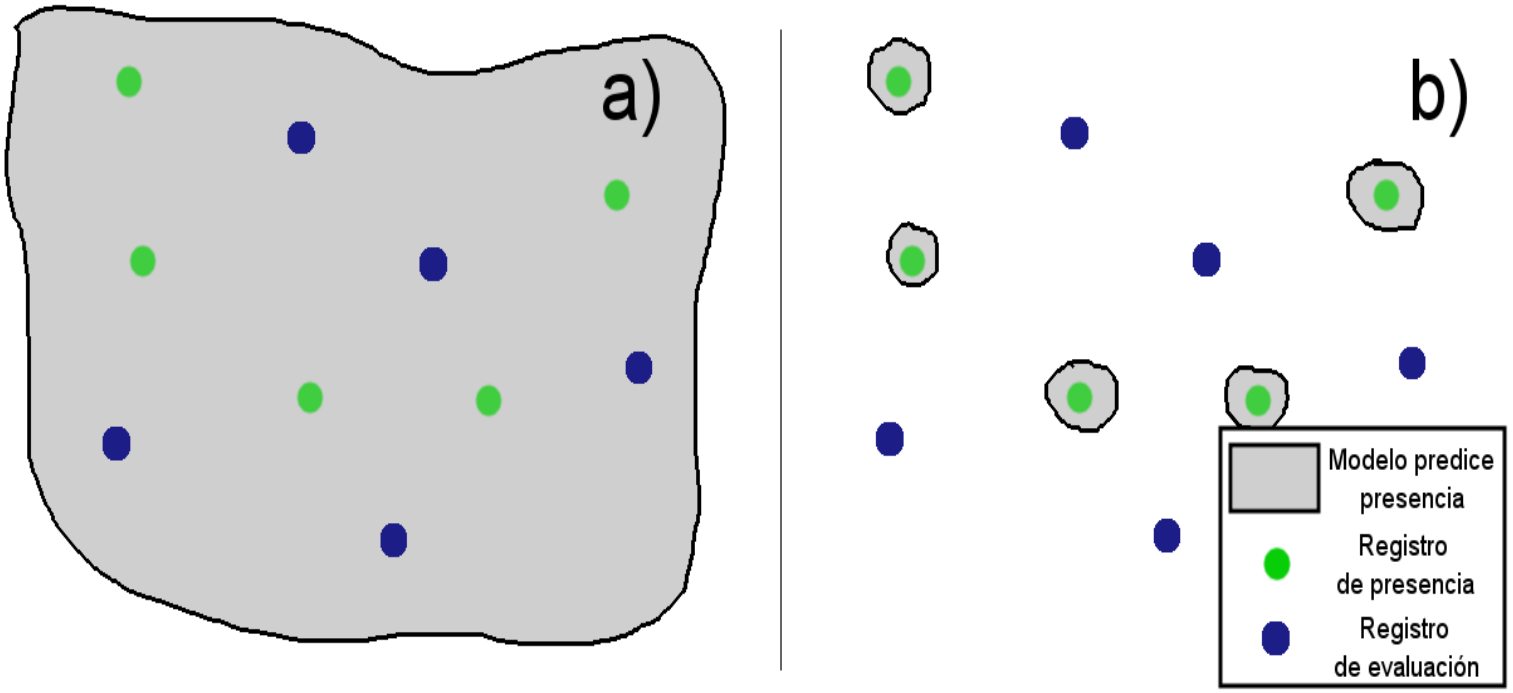
Sensitivity=1

Commission error?

Sensitivity=1

Overfitting?

EVALUATION



Sensitivity=1

¿?

Sensitivity=0

Overfitting!

PRESENCE – ABSENCE IN BINARY MODELS

CONFUSION MATRIX

A → correctly predicted presences
D → correctly predicted absences
B → missed absences (false positives or commission error)
C → missed absences (false negatives or omission error)

		PRESENCE DATA	
		Presence	Absence
MODEL PREDICTION	Presence	**A**	**B**
	Absence	**C**	**D**

sensitivity → S = A/(A+C)

specificity → E = D/(B+D)

true skill statistic → TTS = S + E – 1

CONFUSION MATRIX

A → correctly predicted presences
D → correctly predicted absences
B → missed absences (false positives or commission error)
C → missed presences (false negatives or omission error)

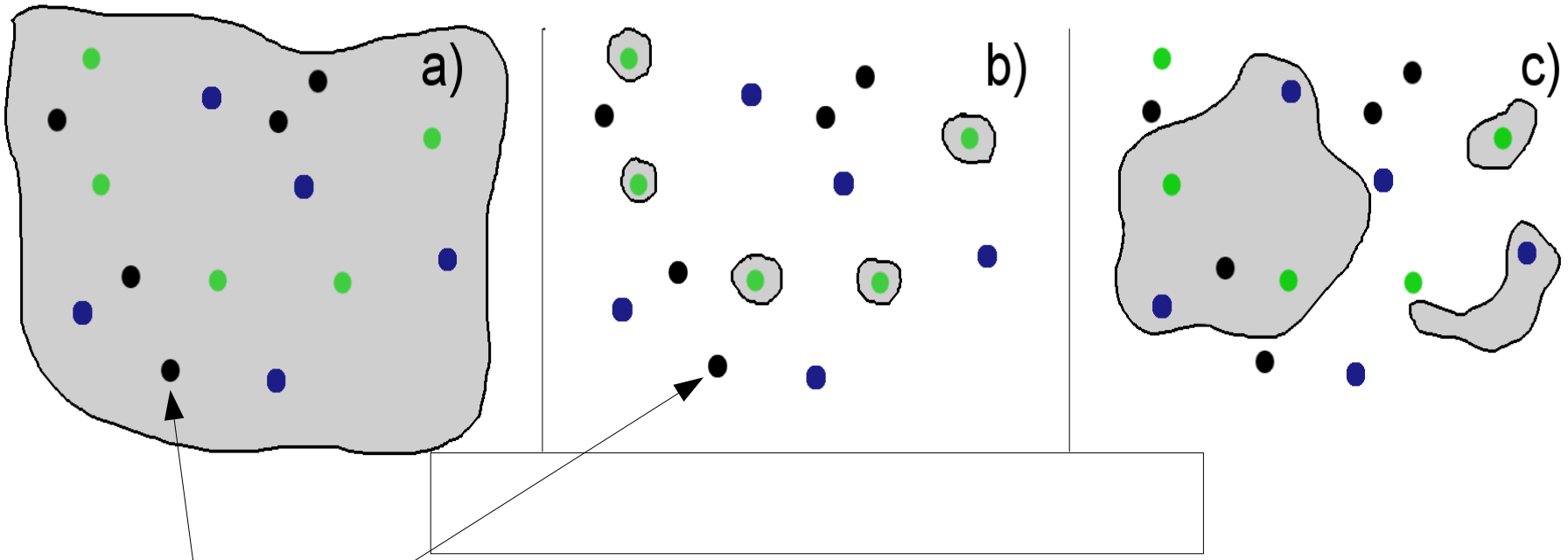
		PRESENCE DATA	
		Presence	Absence
MODEL PREDICTION	Presence	**A**	**B**
	Absence	**C**	**D**

sensitivity → S = A/(A+C)

specificity → E = D/(B+D)

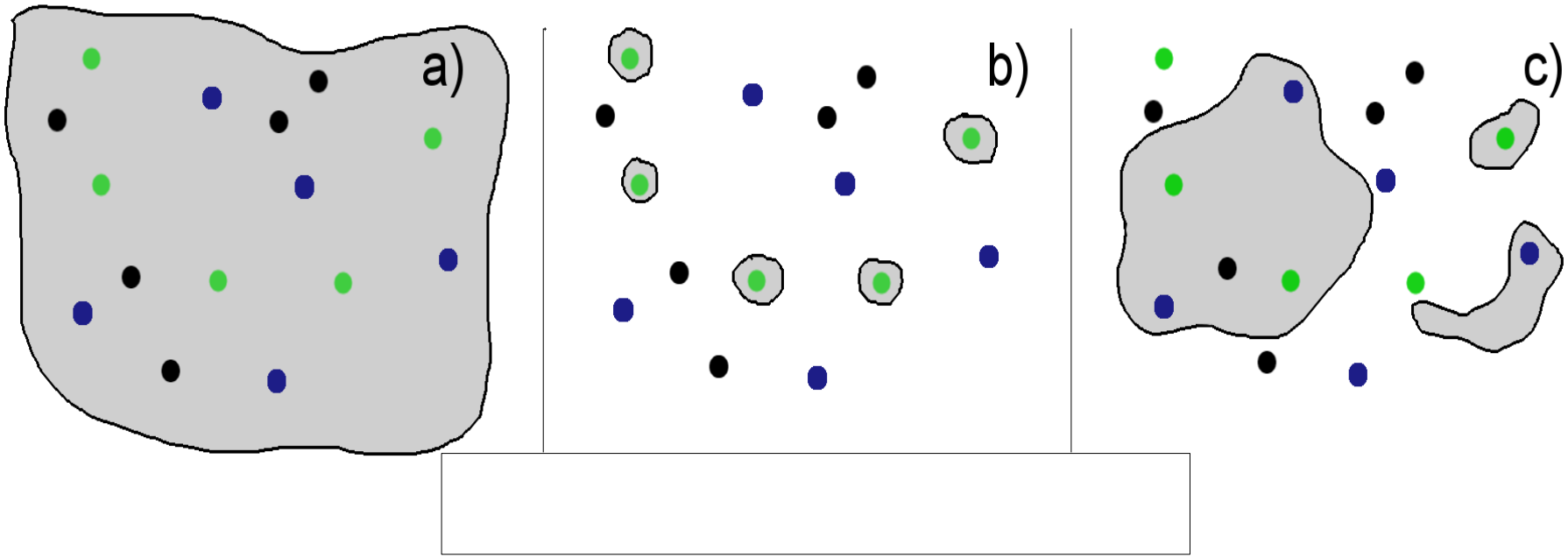
true skill statistic → TTS = S + E – 1

EVALUATION



Model		Data					
		a		b		c	
		pres.	aus.	pres.	aus.	pres.	aus.
	presencia	5	5	0	0	3	1
	ausencia	0	0	5	5	2	4

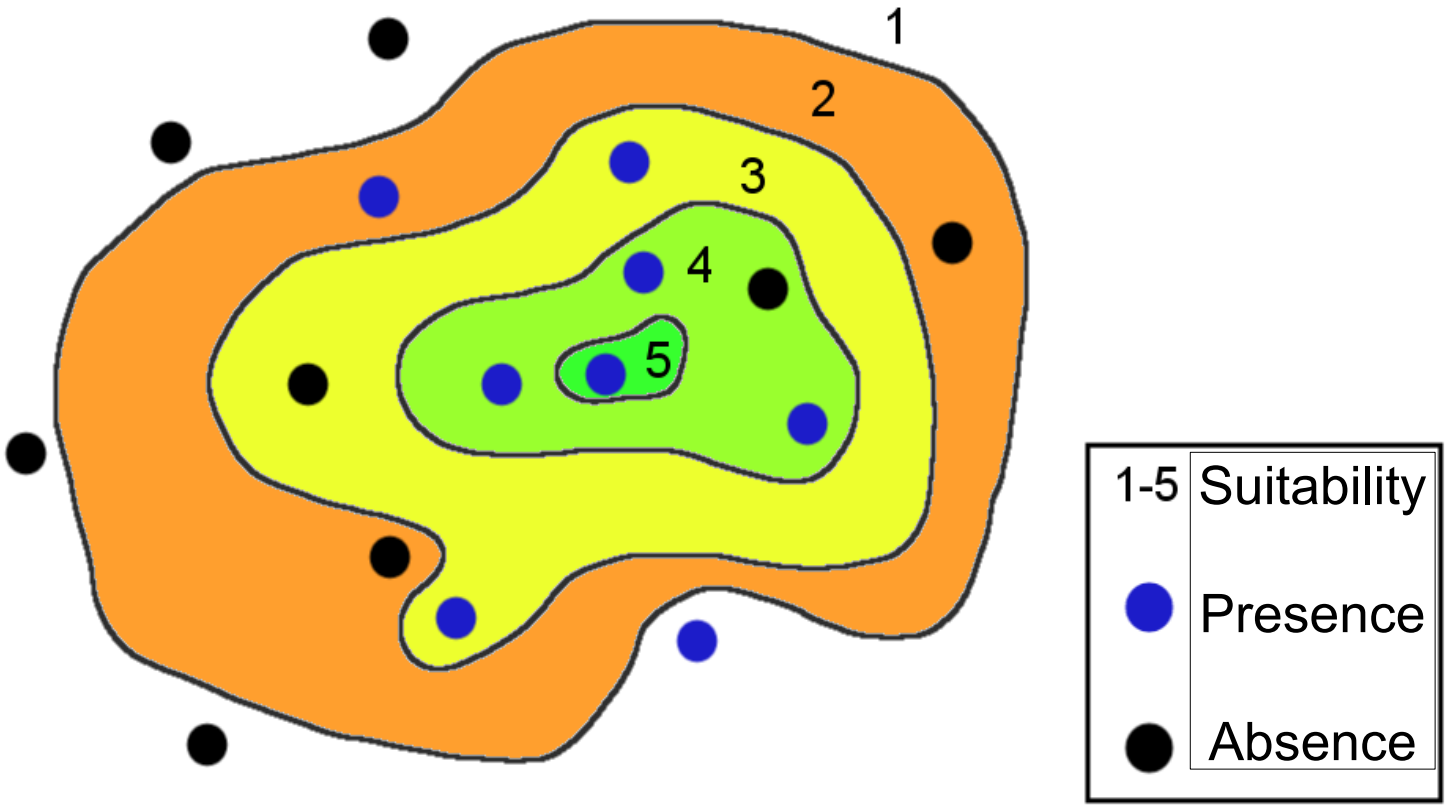
EVALUACIÓN



Model	a	b	c
sensitivity	1	0	0.6
specificity	0	1	0.8

PRESENCE – ABSENCE IN MODELS WITH CONTINUOUS HABITAT SUITABILITY VALUES

COMPUTATION OF A ROC CURVE



CONFUSION MATRIX

A → correctly predicted presences
D → correctly predicted absences
B → missed absences (false positives or commission error)
C → missed absences (false negatives or omission error)

		PRESENCE DATA	
		Presence	Absence
MODEL PREDICTION	Presence	**A**	**B**
	Absence	**C**	**D**

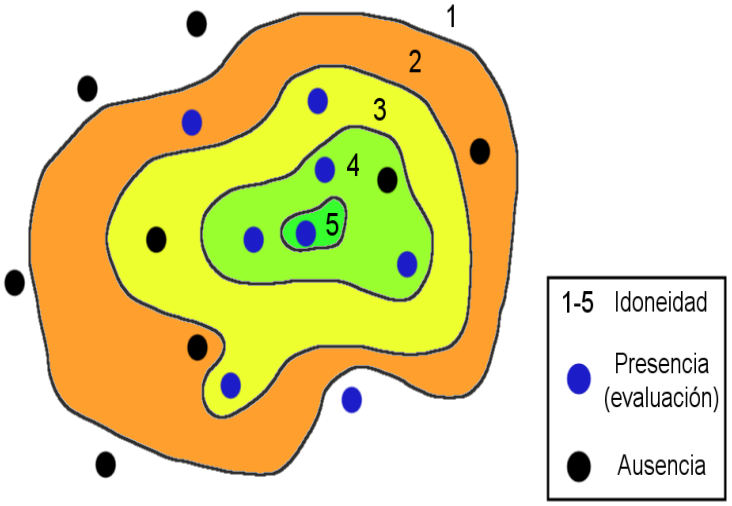
sensitivity → S = A/(A+C)

specificity → E = D/(B+D)

true skill statistic → TTS = S + E – 1

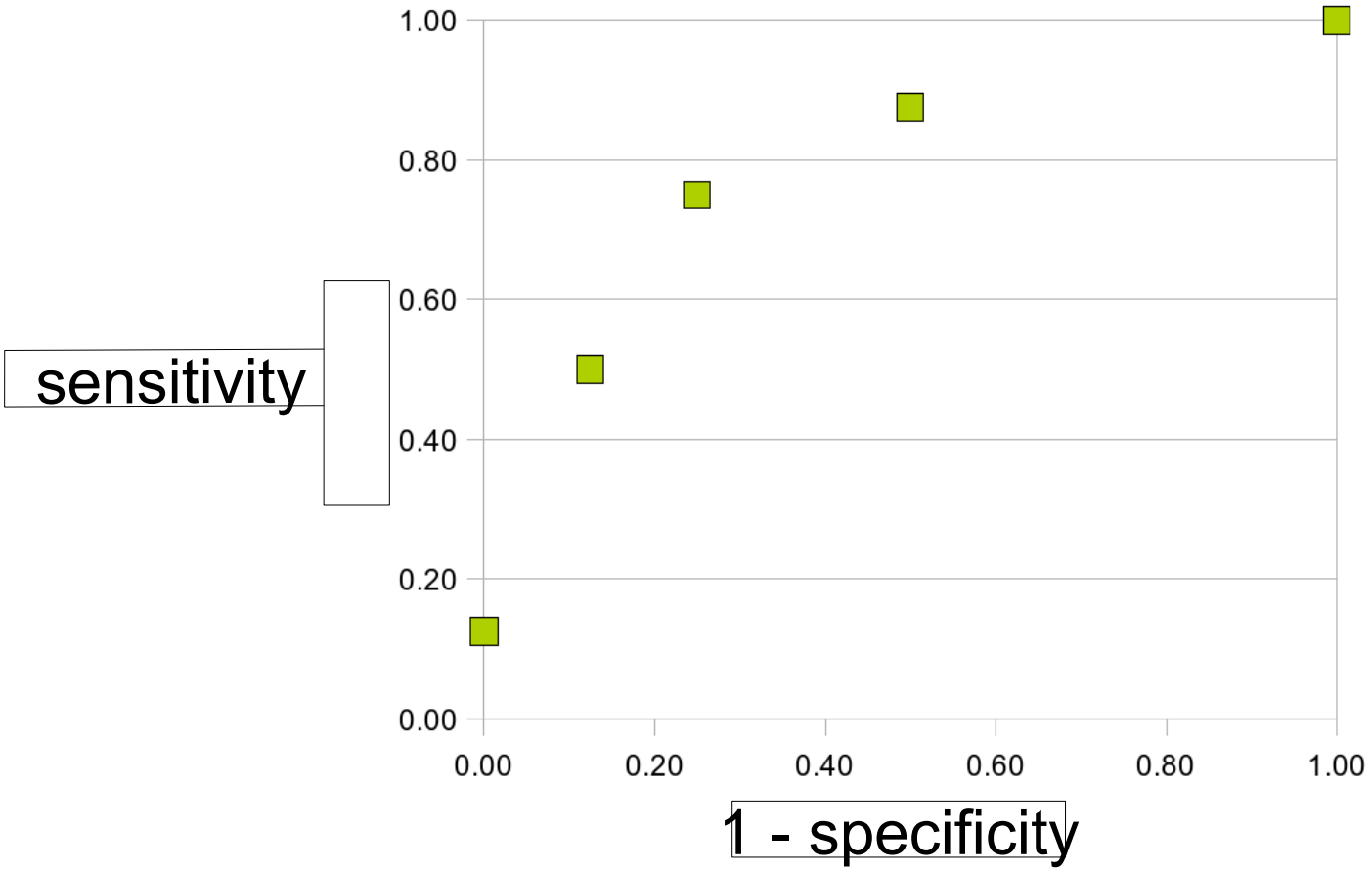
COMPUTATION OF A ROC CURVE

Notice that we are using **1-specificity** rather than specificity

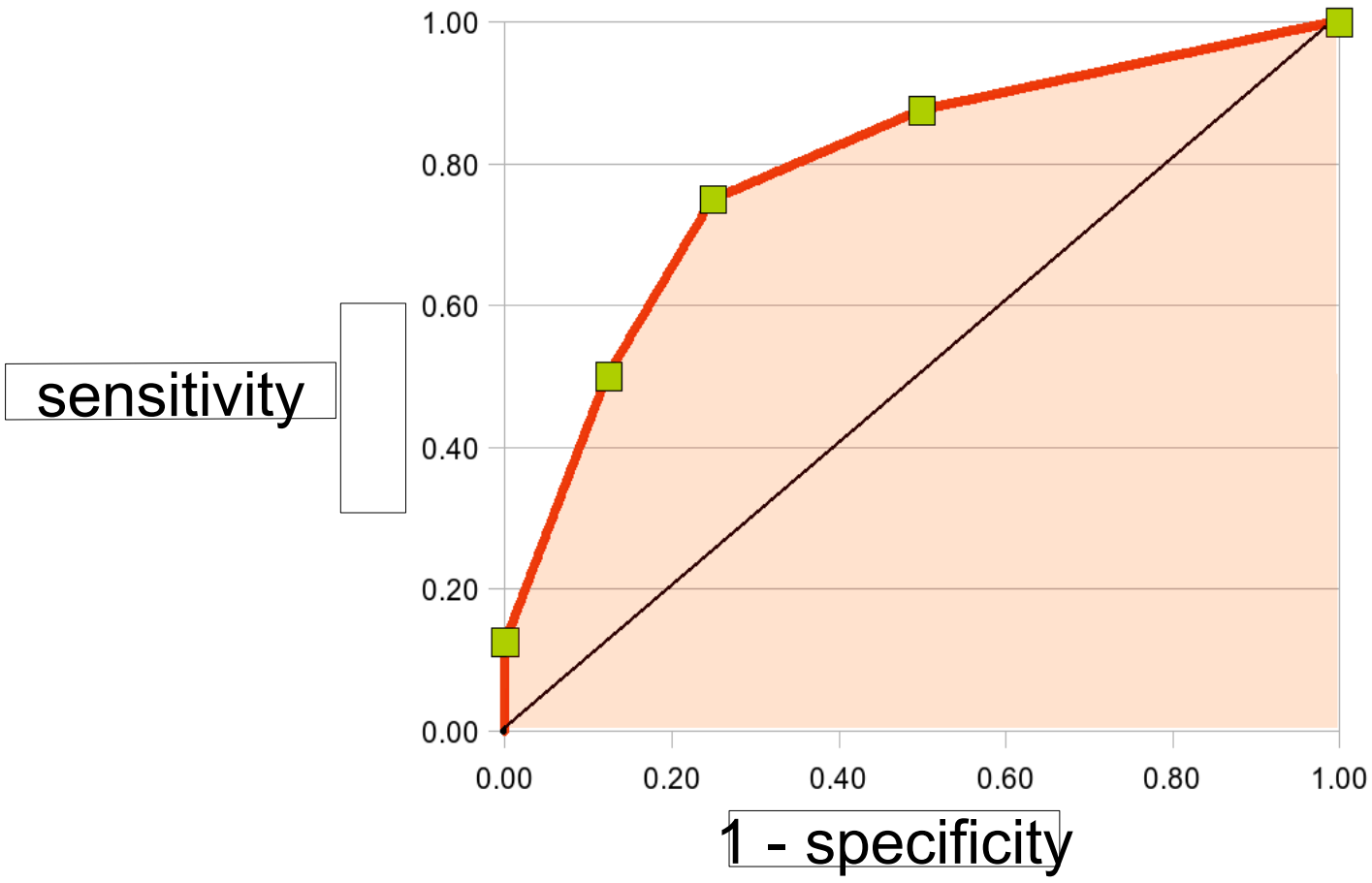


Predicted presences	Missed absences	Missed presences	Predicted presences	Sample size			
A	**B**	**C**	**D**	**N**	suitability	SENSITIVITY	1-SPECIFICITY
8	8	0	0	16	1	**1.00**	**1.00**
7	4	1	4	16	2	**0.88**	**0.50**
6	2	2	6	16	3	**0.75**	**0.25**
4	1	4	7	16	4	**0.50**	**0.13**
1	0	7	8	16	5	**0.13**	**0.00**

ROC CURVE


1.00
0.80
0.60
0.40
0.20
0.00
sensitivity
0.00
0.20
0.40
0.60
0.80
1.00
1 - specificity

ROC CURVE


sensitivity
1.00
0.80
0.60
0.40
0.20
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1 - specificity

AREA UNDER THE ROC CURVE

Given a presence and an absence randomly selected, AUC is the probability for the presence to have a higher habitat suitability than the absence.

Example:

If AUC = 0.74, the model will assign a higher habitat suitability to presences 74 out of 100 times.

PRESENCE – PSEUDOABSENCES IN CONTINUOUS HABITAT SUITABILITY MODELS

CONFUSION MATRIX

A → correctly predicted presences
D → correctly predicted absences
B → **NOT AN ERROR ANYMORE!!**
C → missed absences (false negatives or omission error)

		PRESENCE DATA	
		Presence	Absence
MODEL PREDICTION	Presence	**A**	**B**
	Absence	**C**	**D**

IT CHANGES THE MEANING OF AUC!!

- Now, AUC = probability for any presence to have a higher suitability value than any pseudo-absence.
- But pseudo-absences are not absences, and many of them fall within suitable areas, and therefore, maximum AUC values are always lower than 1.

AUC: a misleading measure of the performance of predictive distribution models

Jorge M. Lobo[1]*, Alberto Jiménez-Valverde[1] and Raimundo Real[2]

[1]*Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), Madrid, Spain,* [2]*Laboratorio de Biogeografía, Diversidad y Conservación, Departamento de Biología Animal, Facultad de Ciencias, Universidad de Málaga, Spain*

*Correspondence: Jorge M. Lobo, Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), Madrid, Spain.
E-mail: mcnj117@mncn.csic.es

ABSTRACT

The area under the receiver operating characteristic (ROC) curve, known as the AUC, is currently considered to be the standard method to assess the accuracy of predictive distribution models. It avoids the supposed subjectivity in the threshold selection process, when continuous probability derived scores are converted to a binary presence–absence variable, by summarizing overall model performance over all possible thresholds. In this manuscript we review some of the features of this measure and bring into question its reliability as a comparative measure of accuracy between model results. We do not recommend using AUC for five reasons: (1) it ignores the predicted probability values and the goodness-of-fit of the model; (2) it summarises the test performance over regions of the ROC space in which one would rarely operate; (3) it weights omission and commission errors equally; (4) it does not give information about the spatial distribution of model errors; and, most importantly, (5) the total extent to which models are carried out highly influences the rate of well-predicted absences and the AUC scores.

Keywords

AUC, distribution models, ecological statistics, goodness-of-fit, model accuracy, ROC curve.

SOME AUC ISSUES

- Gives equal weight to commission and omission errors (and depending on the case, the latter could be more important than the former, or viceversa).
- Doesn’t provide information on the spatial distribution of error.
- With pseudo-absences, AUC values depend on the relation presence surface vs study area surface (larger study areas increase AUC values).
- With pseudo-absences, AUC values of different species cannot be compared.

STILL...

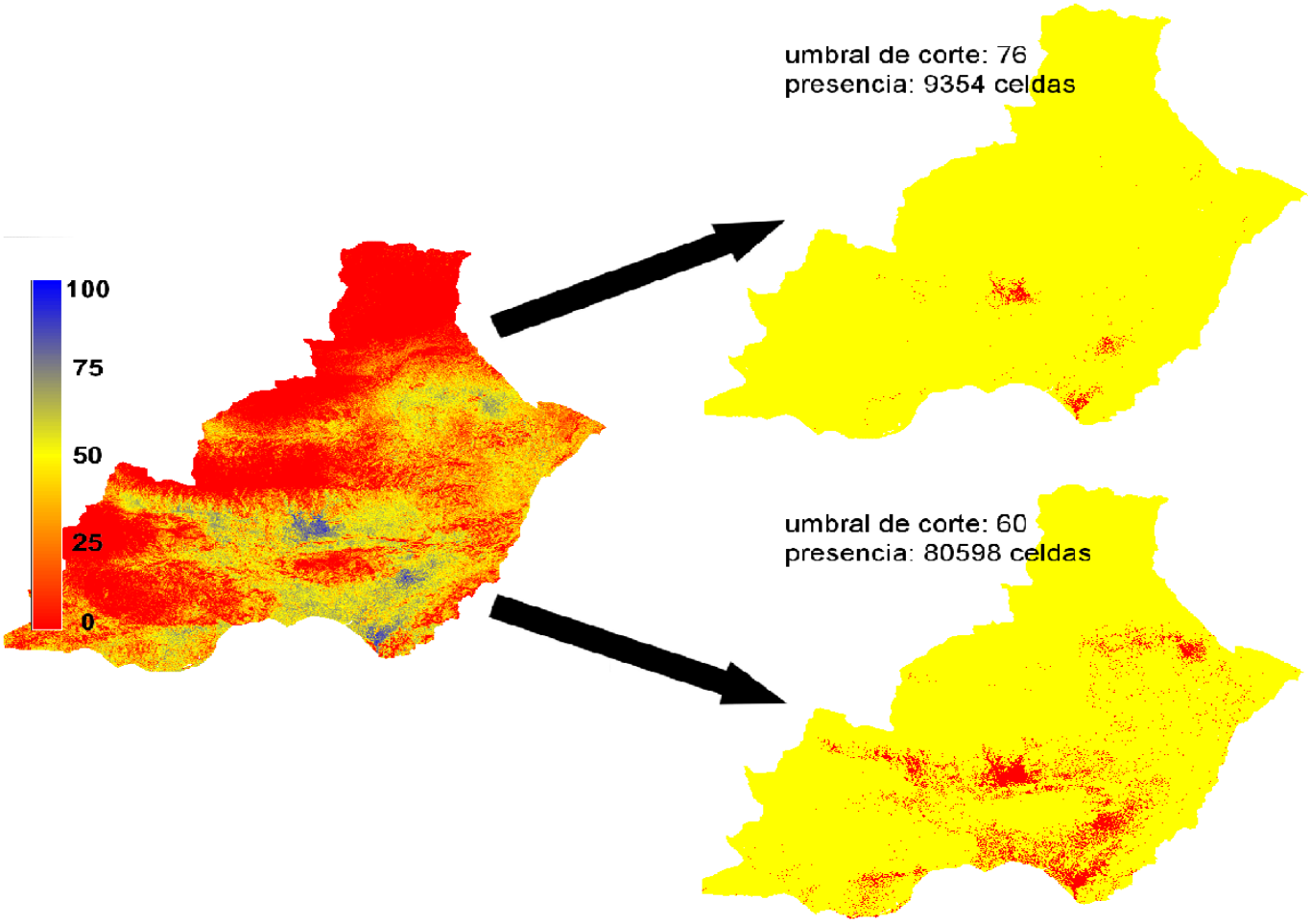
AUC values computed with pseudo-absences are useful to select the best model from a pool of models of the same species.

SAMPLING STRATEGIES TO COMPUTE AUC VALUES

- If you have **independent presence-absence data** (that's the pink unicorn of SDMs!): a single AUC value can be computed and reported.
- If you are a data-poor fellow: **Cross validation**
  - Data splitting
  - Bootstrap
  - K-fold
  - Leave-one-out

COMPUTING THRESHOLDS

CONTINUOUS TO BINARY



SOME RELEVANT PAPERS

- Liu et al. 2005
- Jiménez-Valverde y Lobo 2007
- Freeman y Moisen 2008

SUBJECTIVE SELECTION

"Arbitrary choices with no ecological basis" (Osborne et al. 2001)

- Fixed probability values: 0.5, 0.3, ...
- Fixed commission percentage: 95%, 90%, ...

OBJECTIVE SELECTION

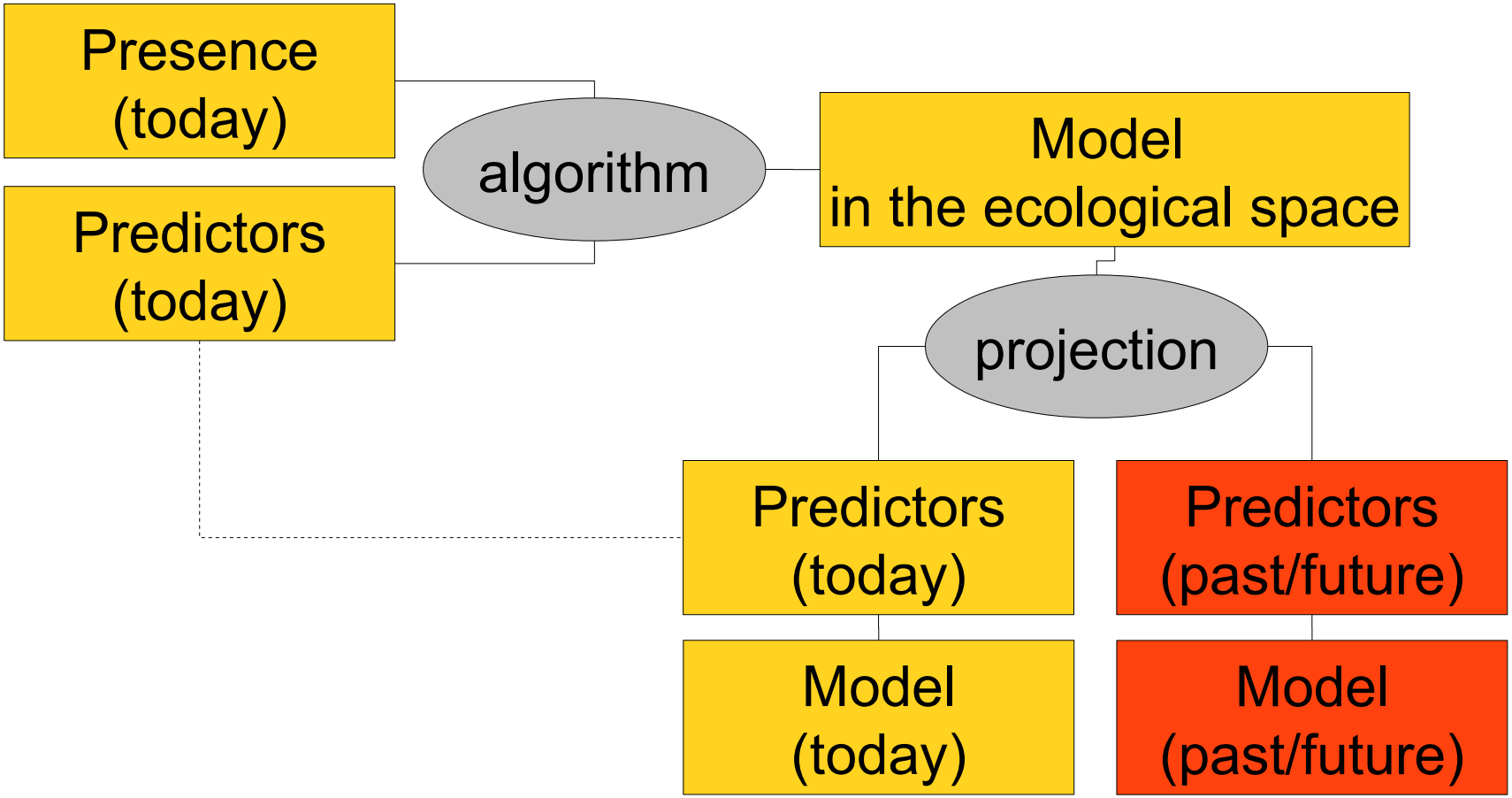
"Selecting the threshold maximizing the concordance between observed and modeled distributions" (Liu et al 2005)

- Kappa maximization
- Point of ROC curve with slope equal to 1
- Equal sensitivity and specificity
- And many others in Freeman y Moisen 2008

WARNING: all these criteria require true absences to be meaningful!

PROJECTING SDMs IN SPACE AND TIME

MODEL PROJECTION



PROJECTION INTO SPACE

- Requirements
  - Same predictors available for the calibration and projection areas
  - Same names and units
  - Same resolution (or so)
- Applications: invasion assessment, biogeographical hypotheses, finding new populations

PROJECTION INTO TIME

- Same requirements as before but:
- Some predictors are not available for the past or the future (ndvi, human footprint, etc), therefore climate and topography are the main predictors used for these projections
- Every projection is an scenario, NOT a prediction!
- Applications: assessment of range shift under climate change, palaeodistribution modelling

¿COMO PODEMOS EVALUAR ESTOS MODELOS?

- El AUC de un modelo actual no representa la capacidad predictiva del modelo en el pasado o el futuro
- Los modelos de paleodistribución de plantas se pueden evaluar con polen fósil y macrorrestos
- Los modelos de paleodistribución de animales se pueden evaluar con datos de registro fósil
- Los datos de evaluación y los modelos deben ser coetáneos.

ALGUNAS PREMISAS

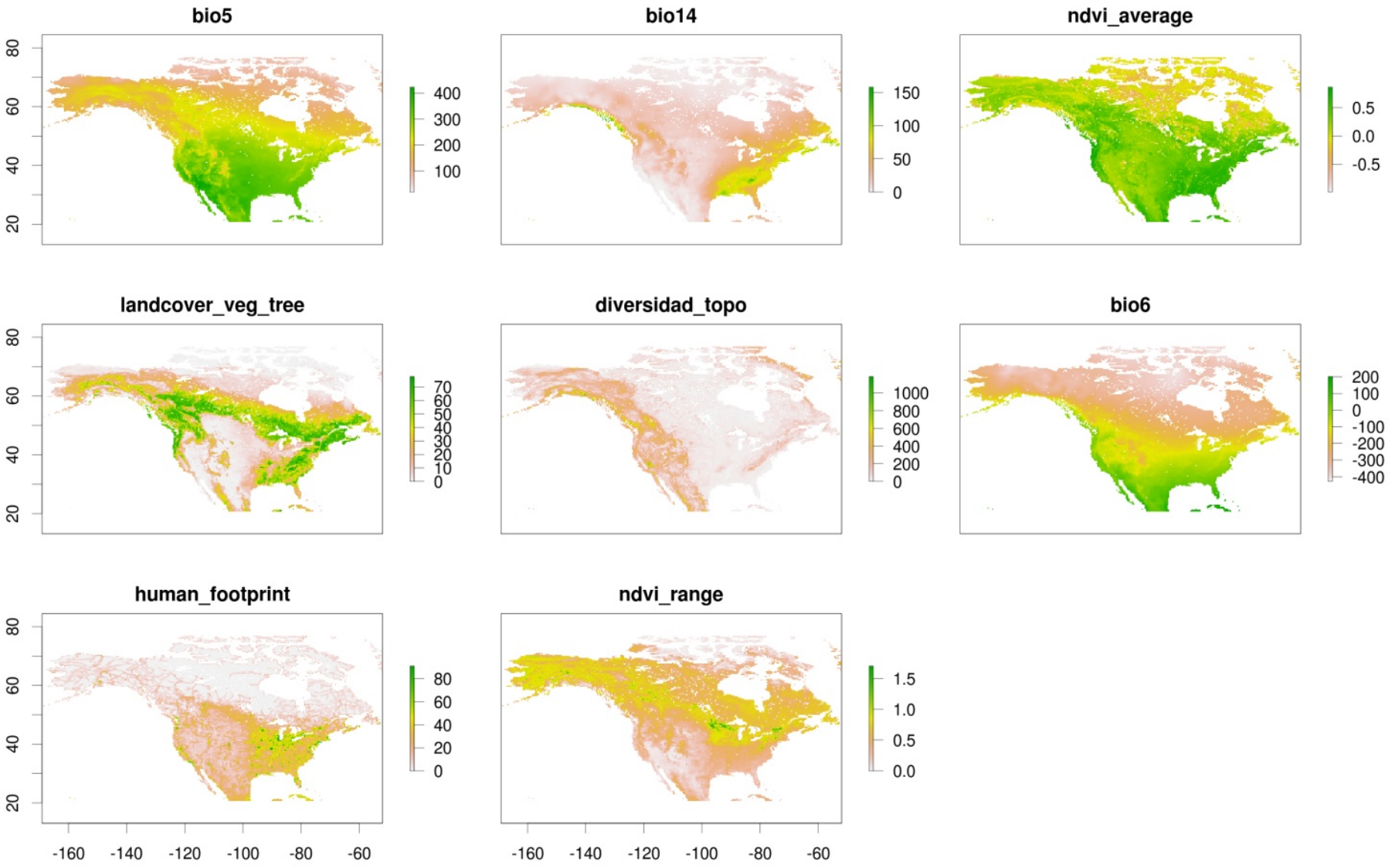
- Una proyección NO representa la distribución futura o pasada (o en otro lugar) de una especie.
- Una proyección SOLO representa donde habrá condiciones ecológicas similares a aquellas en las que se ha observado la especie (¡siempre que el modelo no extrapole!) .
- Las proyecciones asumen que el nicho ecológico de las especies es constante.
- Los mapas climáticos del pasado o futuro son ESCENARIOS, no representan la realidad.

BUT… NOVEL CLIMATES!

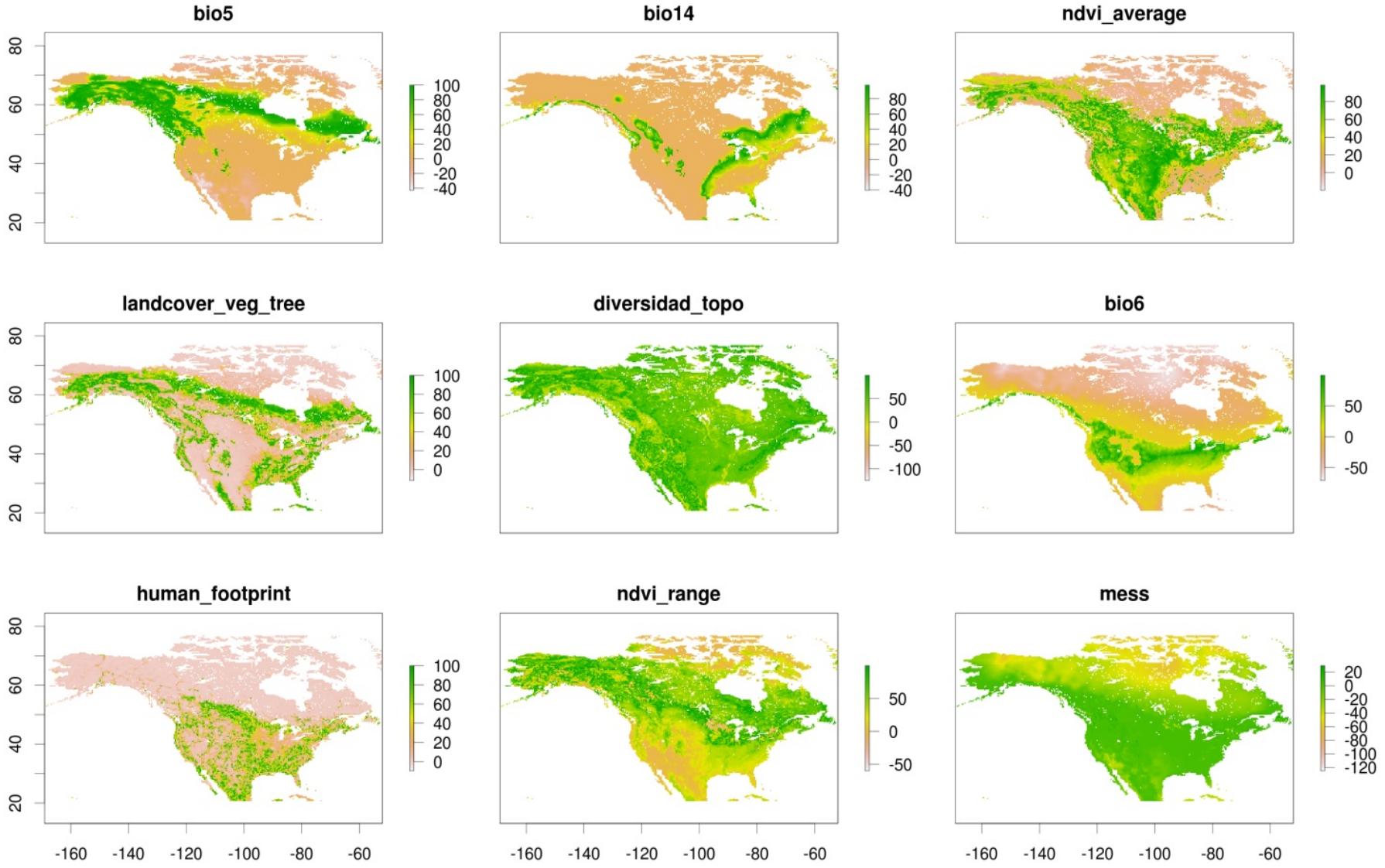
MULTIVARIATE ENVIRONMENTAL SIMILARITY SURFACES (MESS)

- Measures similarity between the ecological space of the presences and the ecological space available in the projection layers.
- The more different they are, the more we'll extrapolate, and you don't want to extrapolate.
- Key reference: Elith J., Kearney M., & Phillips S. 2010. The art of modelling range shifting species. Methods in Ecology and Evolution, 1 :330-342.

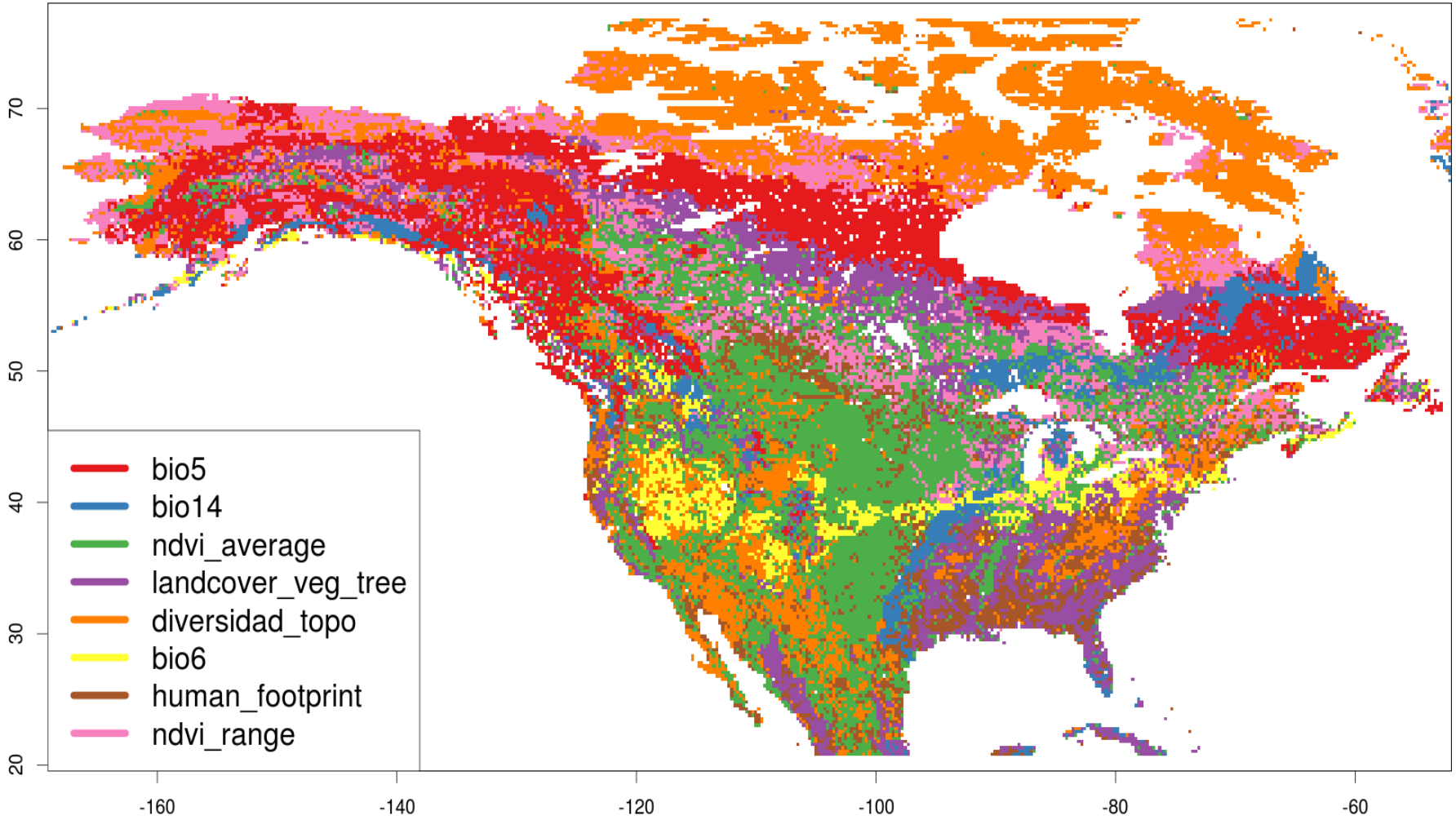
PREDICTORS


bio5
400
300
200
100
bio14
150
100
50
0
ndvi_average
0.5
0.0
-0.5
landcover_veg_tree
70
60
50
40
30
20
10
0
diversidad_topo
1000
800
600
400
200
0
bio6
200
100
0
-100
-200
-300
-400
human_footprint
80
60
40
20
0
ndvi_range
1.5
1.0
0.5
0.0
80
60
40
20
-160
-140
-120
-100
-80
-60

MESS


bio5
bio14
ndvi_average
landcover_veg_tree
diversidad_topo
bio6
human_footprint
ndvi_range
mess
80
60
40
20
-160
-140
-120
-100
-80
-60

MESS (maximum)



ALGUNAS PREMISAS

- Una proyección NO representa la distribución futura o pasada (o en otro lugar) de una especie.
- Una proyección SOLO representa donde habrá condiciones ecológicas similares a aquellas en las que se ha observado la especie (¡siempre que el modelo no extrapole!) .
- Las proyecciones asumen que el nicho ecológico de las especies es constante.
- Los mapas climáticos del pasado o futuro son ESCENARIOS, no representan la realidad.


That's all Folks!