IX Taller GBIF de Modelización de Nichos Ecológicos (sesión 2)

Aprendiendo a modelizar

Blas M. Benito

gbif.es Nodo Nacional de Información en Biodiversidad

CONTENIDOS

- TEORÍA Y PRÁCTICA DE MÉTODOS DE MODELADO:
  - GLM
  - GAM
  - RANDOM FOREST
  - MAXENT
- EVALUACIÓN DE MODELOS
- APLICACIÓN DE “THRESHOLDS”
- PROYECCIÓN

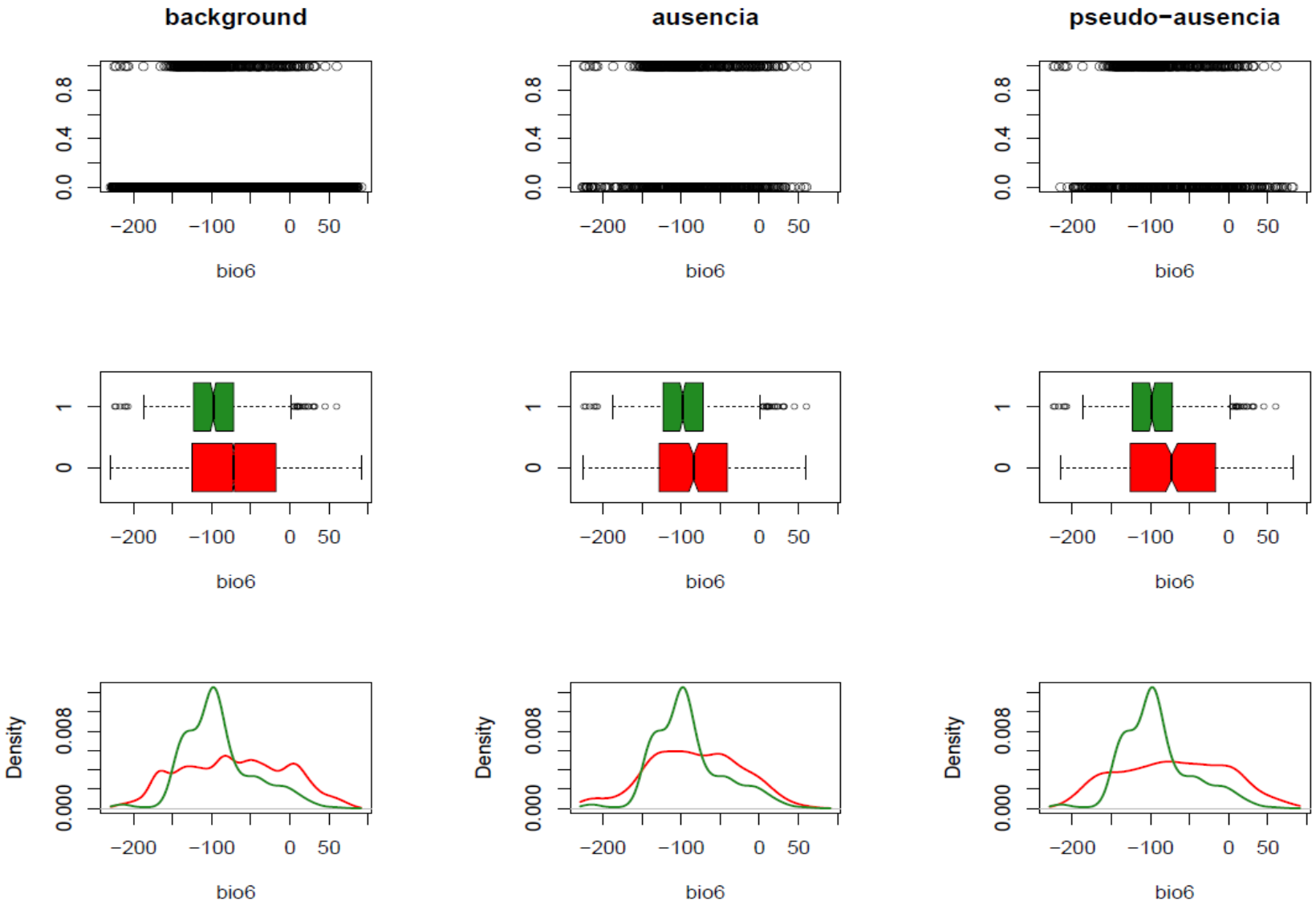
SCRIPT DE R

Reinicia Rstudio y abre

2_modelos.R

ANÁLISIS EXPLORATORIO

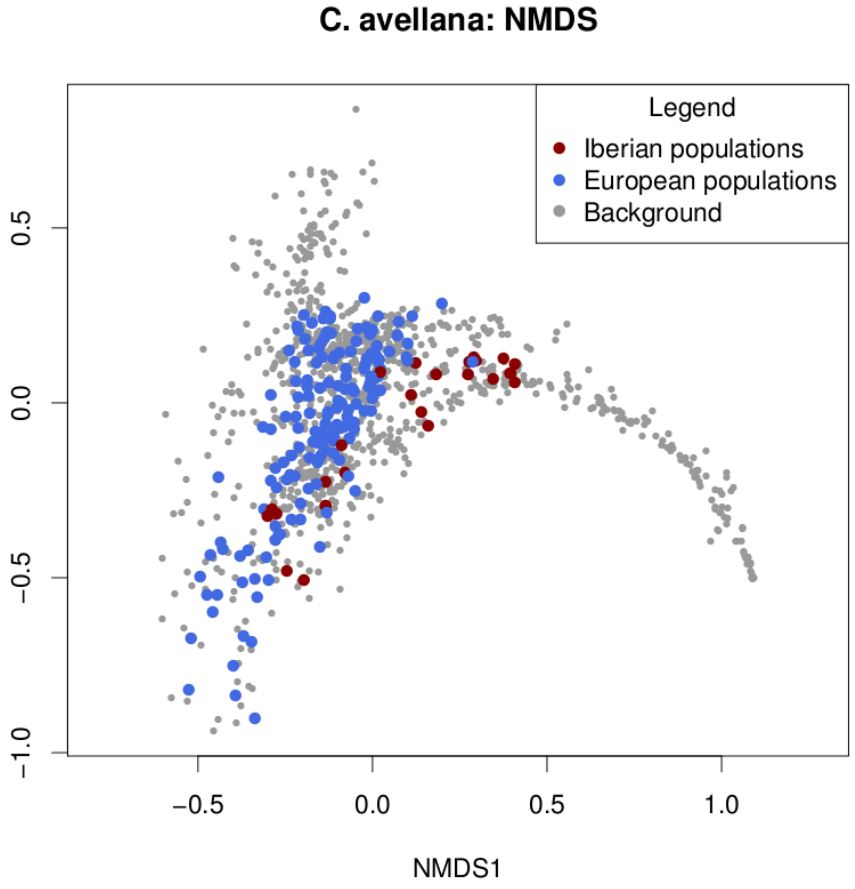
BOXPLOTS Y DENSITY PLOTS

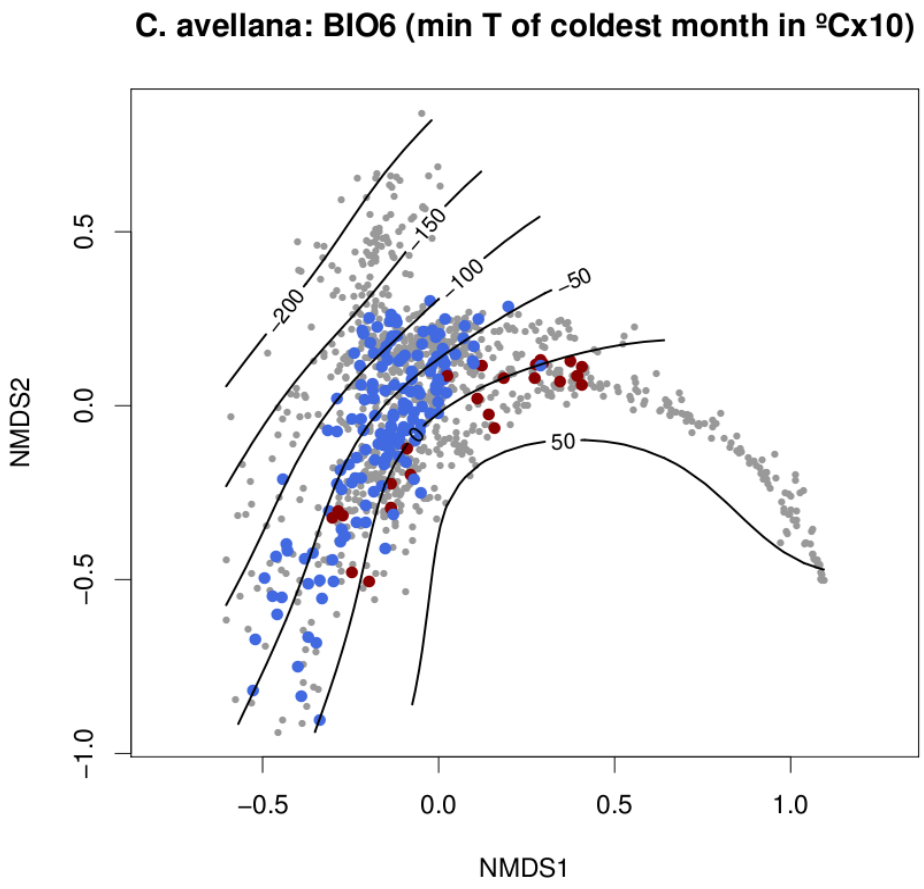


Non Metric Multidimensional Scaling NMDS

- Los puntos de presencia con los valores de las variables representan una nube de puntos de n dimensiones (nD).
- Un NMDS representa en 2D este espacio de nD.
- Las distancias entre puntos en 2D son proporcionales (aproximadamente) a las distancias entre puntos en nD.
- En R se hacen con la función 'metaMDS' de la librería 'vegan'.
- La función 'ordisurf' permite ajustar al NMDS isolíneas representando las variables ambientales

Non Metric Multidimensional Scaling (NMDS)



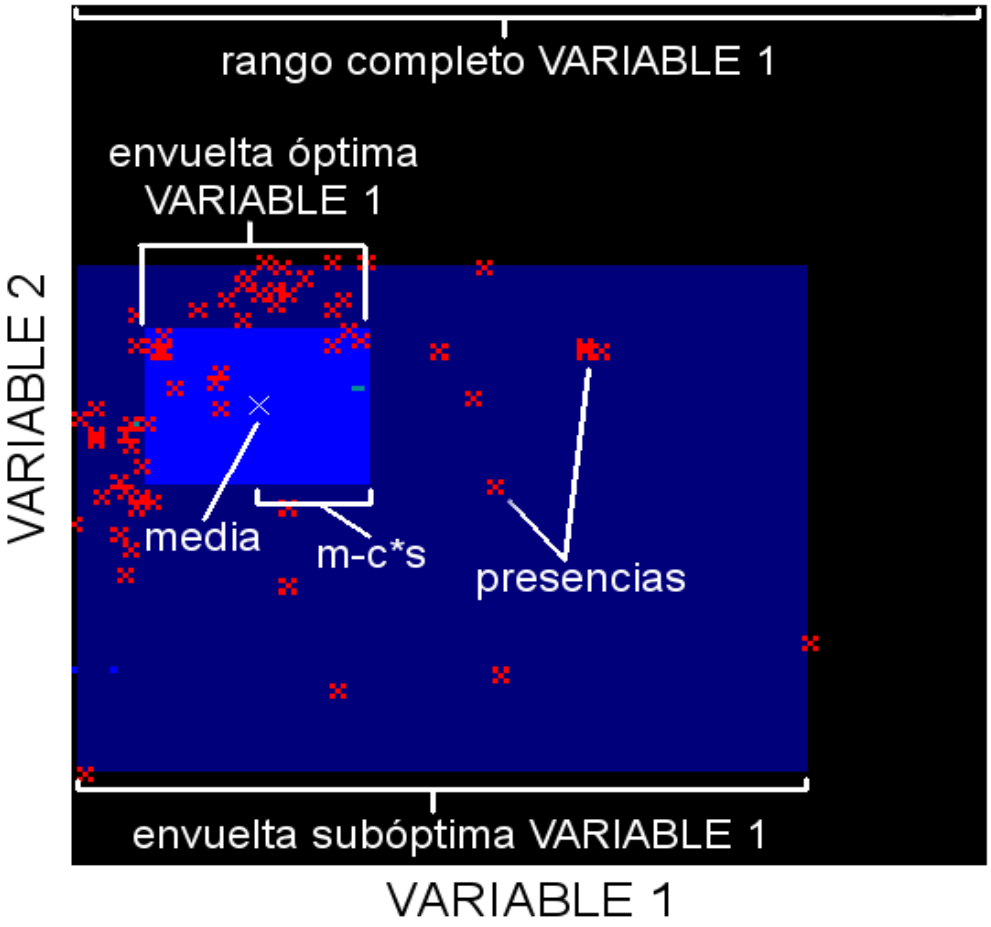


CALIBRANDO MODELOS

BIOCLIM

Envuelta bioclimática cuadrangular

- [m-c*s, m+c*s]
  - m = media
  - c = % desviación
  - s = desviación
- Solo requiere presencias



MÉTODOS DE REGRESIÓN

GENERALIZED LINEAR MODELS

- **Permiten modelar respuestas no lineales**
- Los residuos pueden seguir distintas distribuciones de probabilidad: normal, **binomial**, **Poisson**, binomial negativa, gamma

¿COMO FUNCIONA?

- Según complejidad de las curvas
  - Logística
  - Polinomio 2º, 3º, 4º, ... grado
- Según los datos de ausencia
  - Ausencia
  - Pseudo-ausencia
  - Background
- Según las interacciones entre variables
  - Sin interacción
  - Con interacción

DIBUJEMOS UN POCO PARA ENTENDERLO!

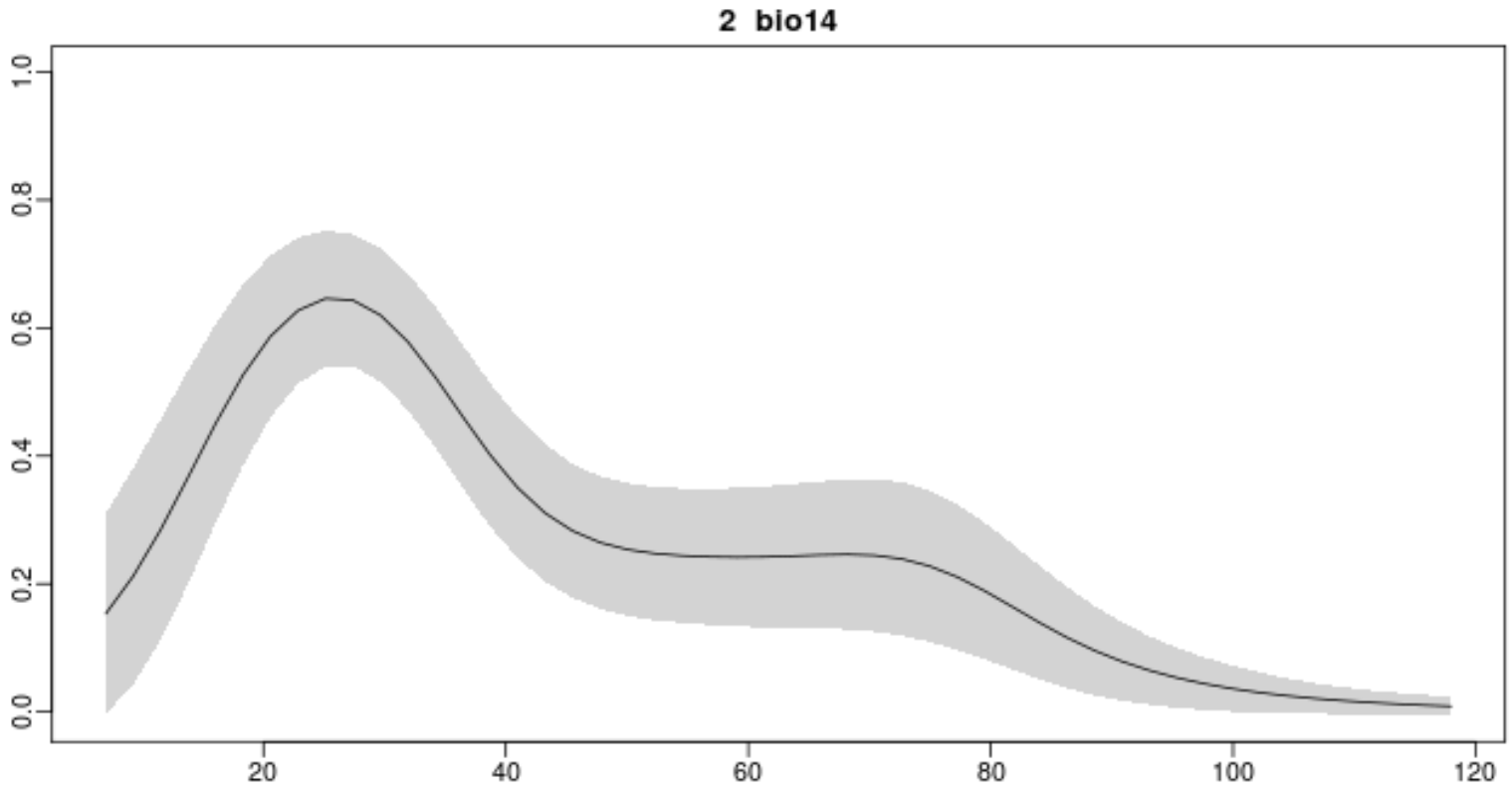
NÚMERO MÍNIMO DE PRESENCIAS

- Necesitamos al menos 5 presencias (y 5 ausencias, si el modelo es de presencia-ausencia) por cada variable.
- Necesitamos otras 5 presencias por cada término polinomial:
  - 1º grado: 5 presencias
  - 2º grado: 10 presencias
  - And so on.

GENERALIZED ADDITIVE MODELS (GAM)

- Método de regresión no paramétrico
- Variables predictivas suavizadas (smoothing)
- Modelado de respuestas no lineales
- Requiere tamaños de muestra grandes (más que GLM)

GENERALIZED ADDITIVE MODELS



MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

- Trabaja bien con datos no lineales.
- Tiene en cuenta interacciones parciales y completas entre variables.
- Las ecuaciones resultantes son fáciles de interpretar.
- Muy rápido con conjuntos de datos grandes.
- Usado para la predicción de series temporales en economía.

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

- Como funciona?
  - **Hinge functions** (bisagra) encadenadas.
  - Interacción entre variables representada por la multiplicación de hinge functions.

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

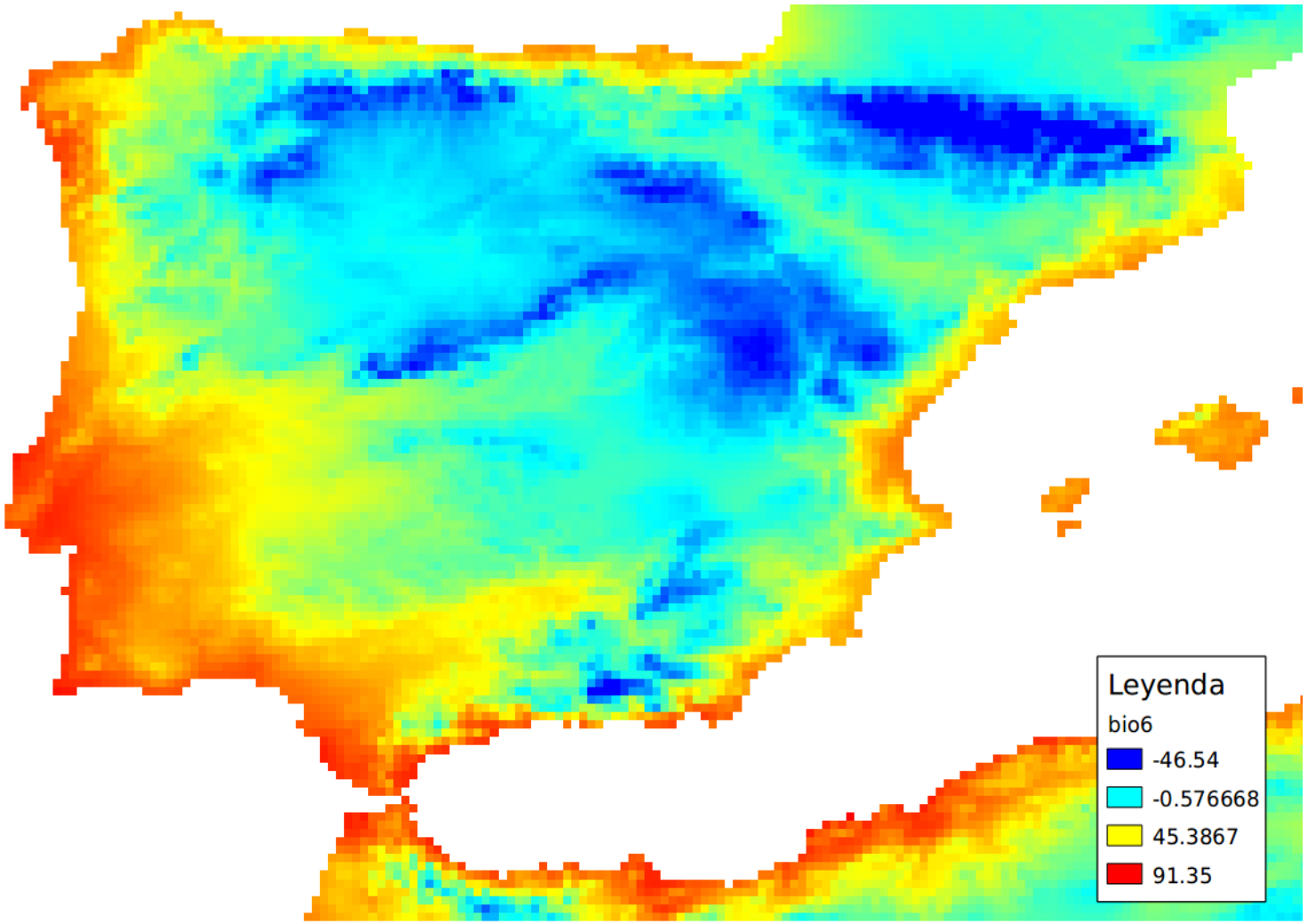
- Construyendo el modelo
  - Forward pass
    - Ajusta todas las funciones bisagra posible a los datos (sobreajuste!)
    - El proceso termina cuando los residuales no se pueden minimizar más, o se alcanza el máximo número de términos de la ecuación.
  - Backwards pass
    - Examina la contribución de cada término individual, y elimina los que no son significantes (model prunning: generalización)

MAXENT

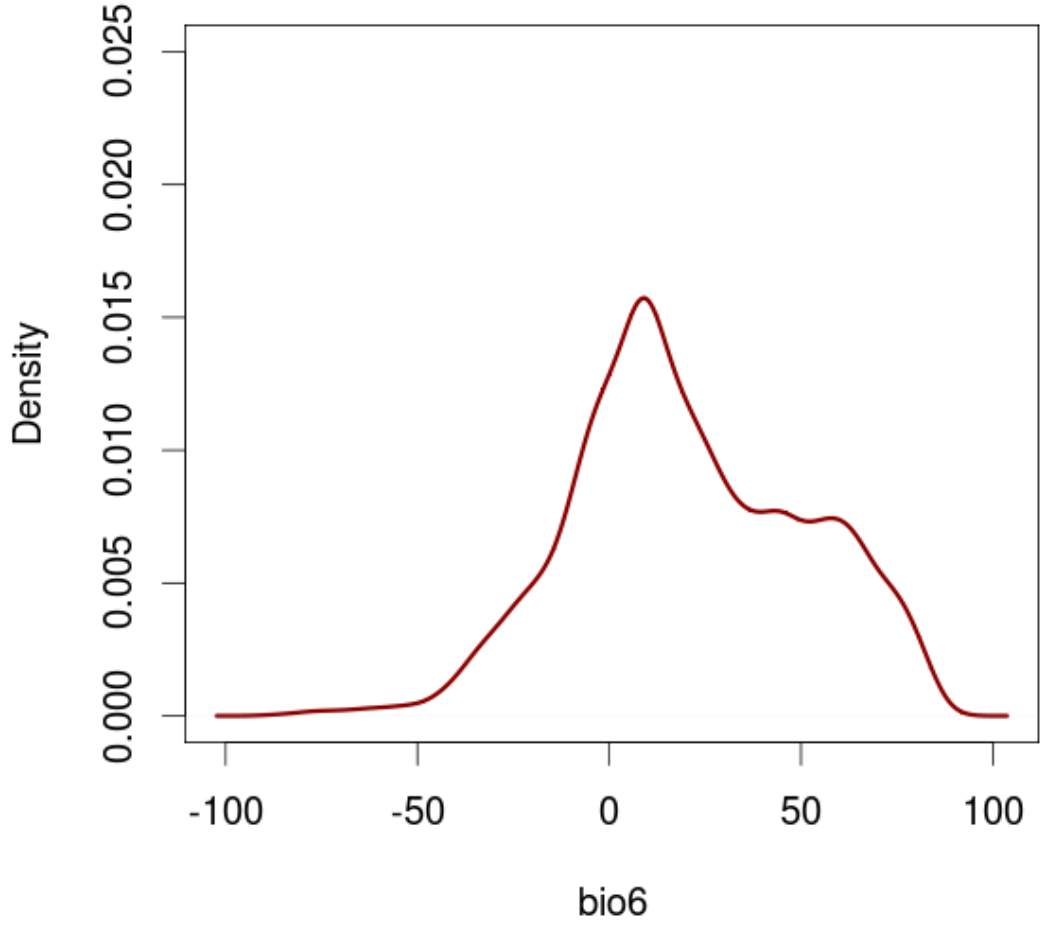
- Regresión de Poisson con penalización Lasso.
- Puede trabajar con un número bajo de presencias.
- Requiere background.
- La complejidad del modelo se controla con el **regularization multiplier.**
- Está disponible en una aplicación Java con interfaz gráfico.
- Nosotros usamos el paquete “maxnet”.

PREDICTOR

Bio6 → temperatura del mes más frío



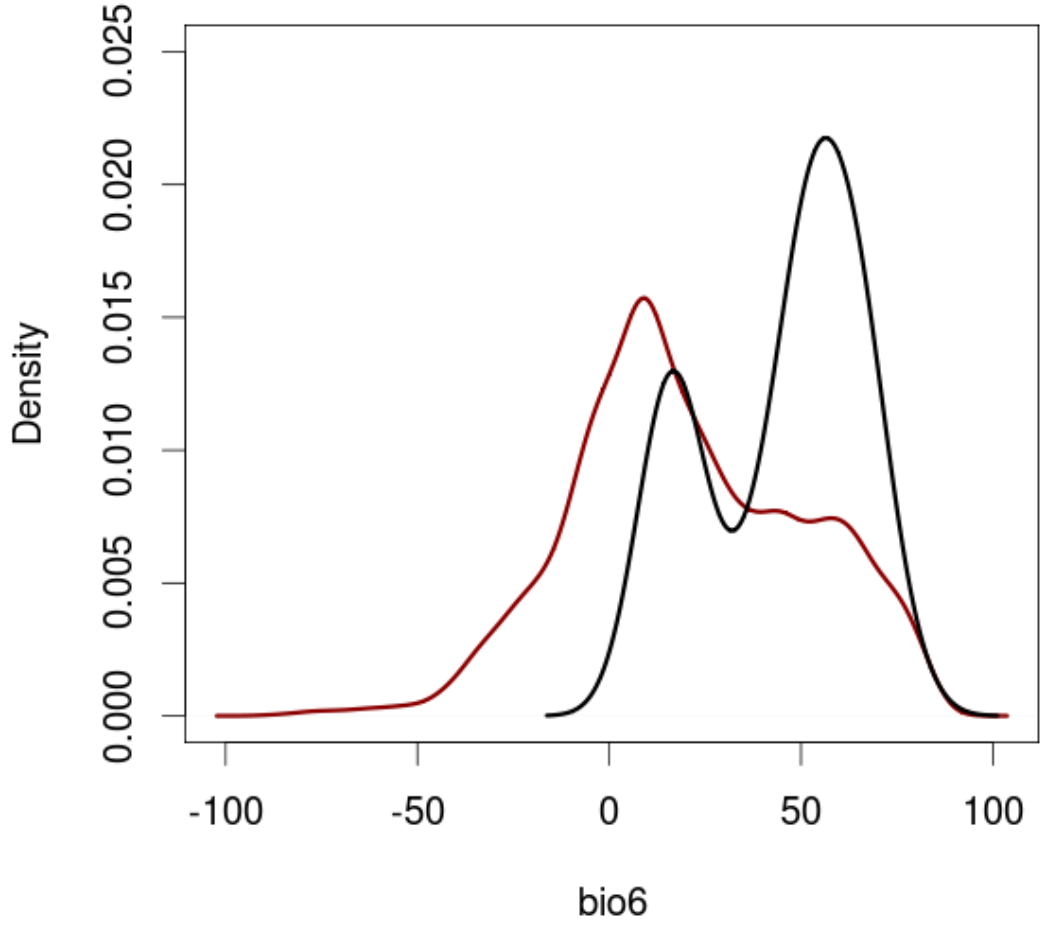
DENSIDAD DEL BACKGROUND



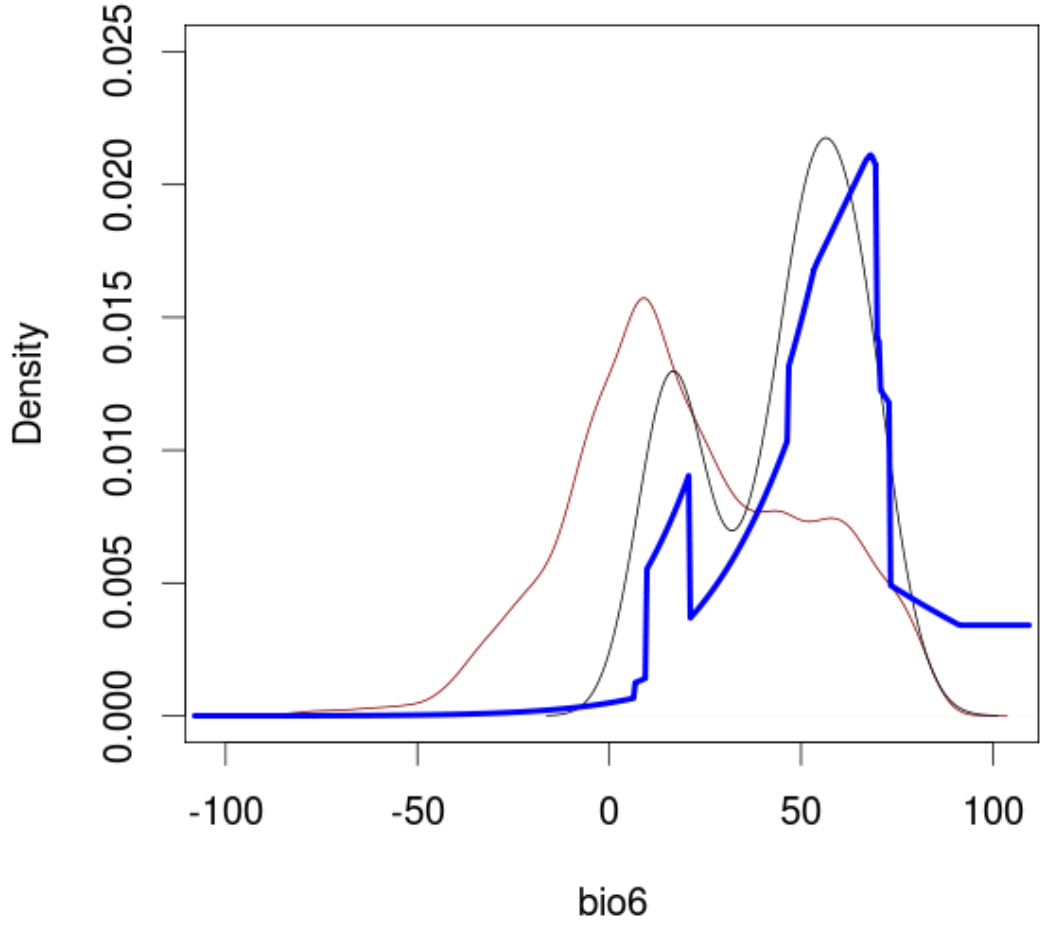
PRESENCIA

Atelerix algirus

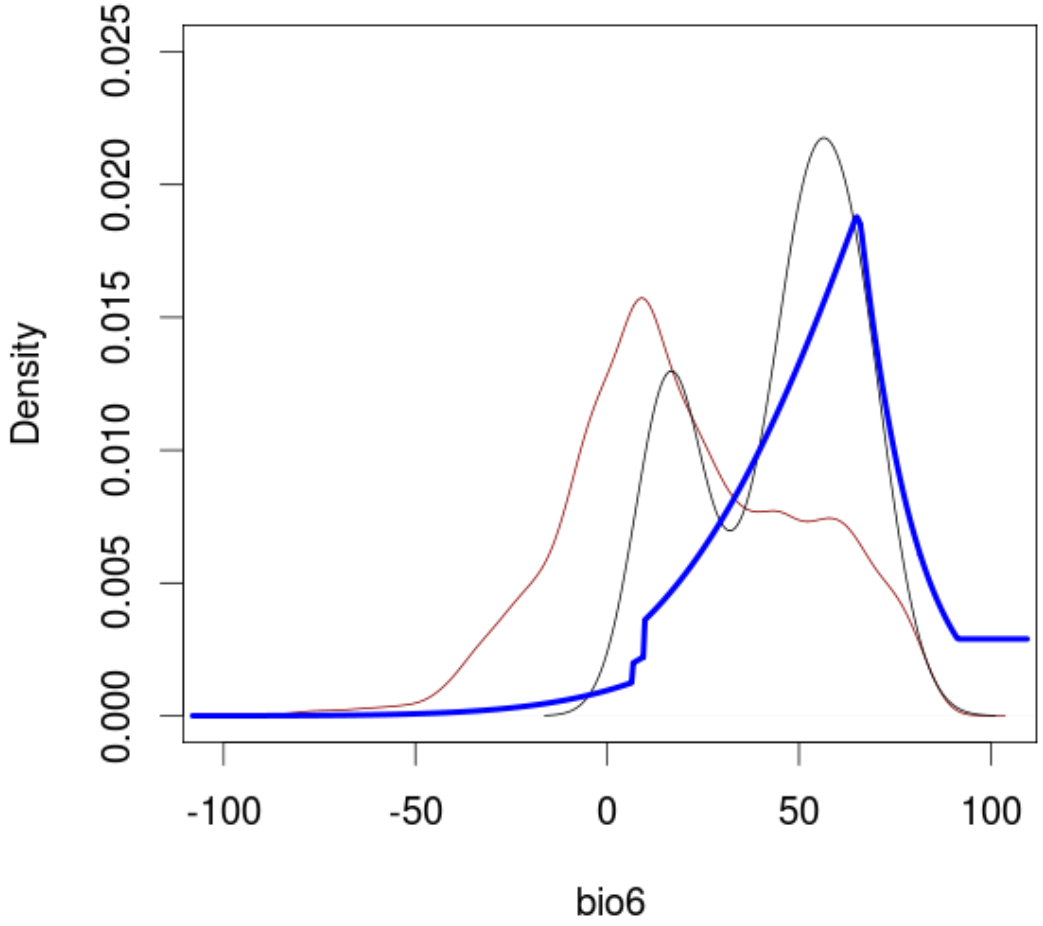
DENSIDAD DE LA PRESENCIA



MAXENT FIT (max complejidad)

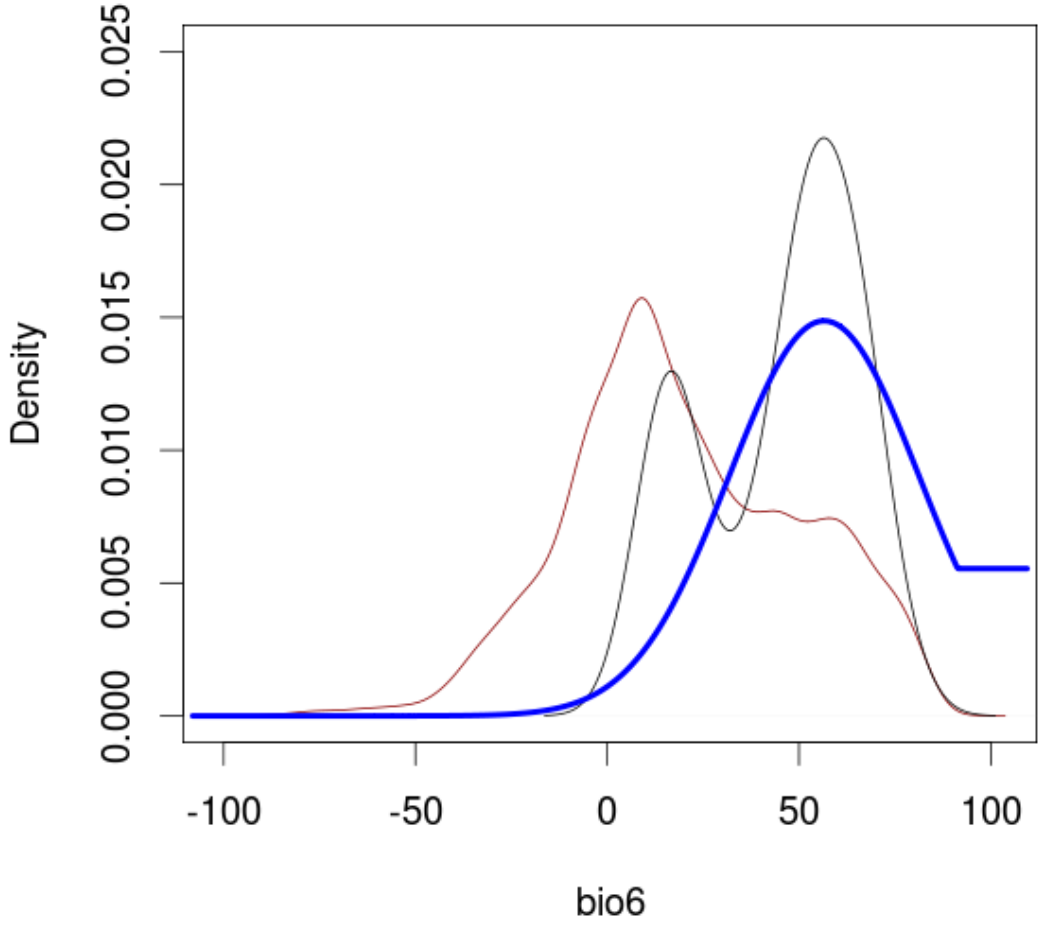


MAXENT FIT



Regularization multiplier = 3

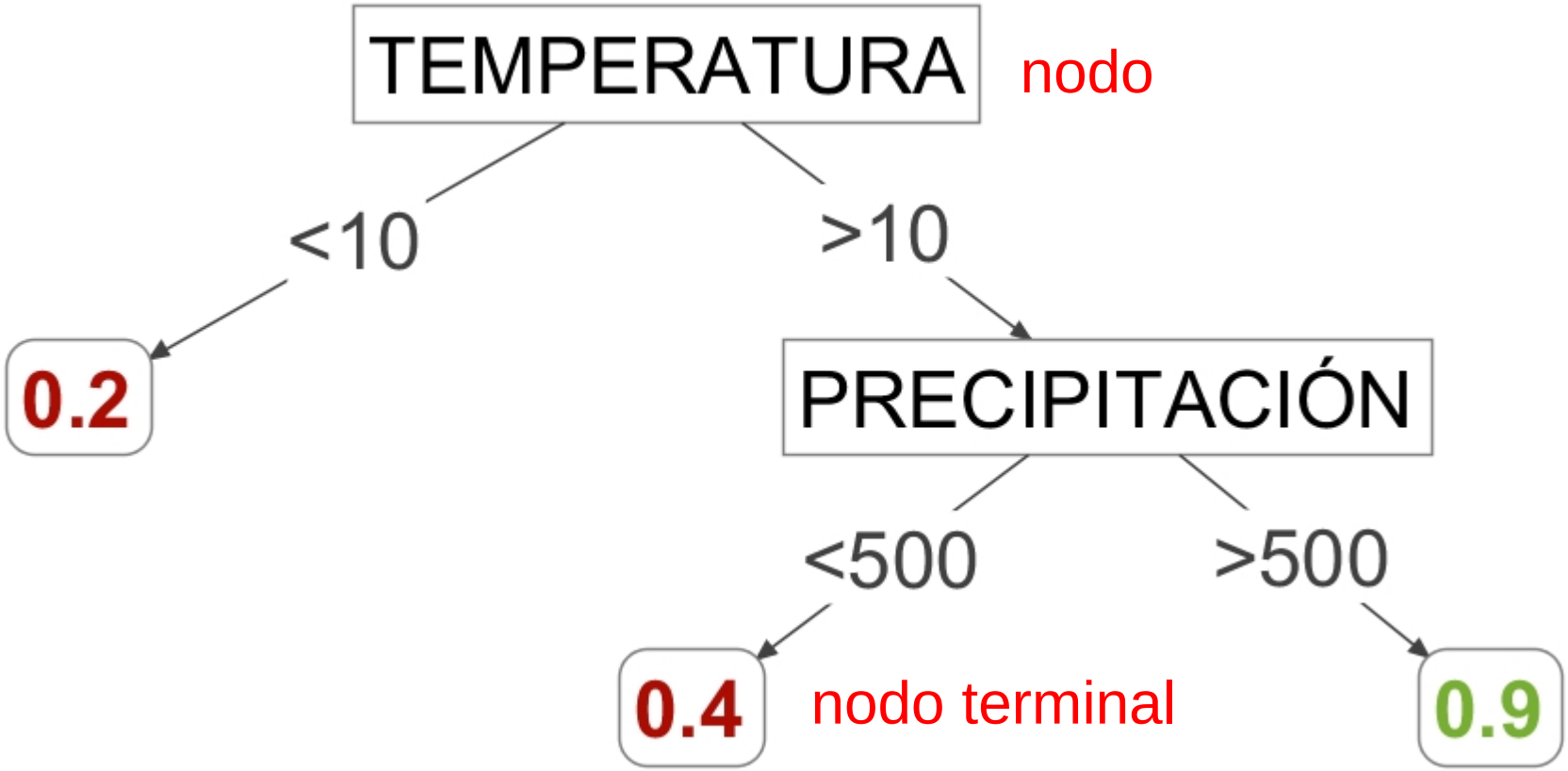
MAXENT FIT



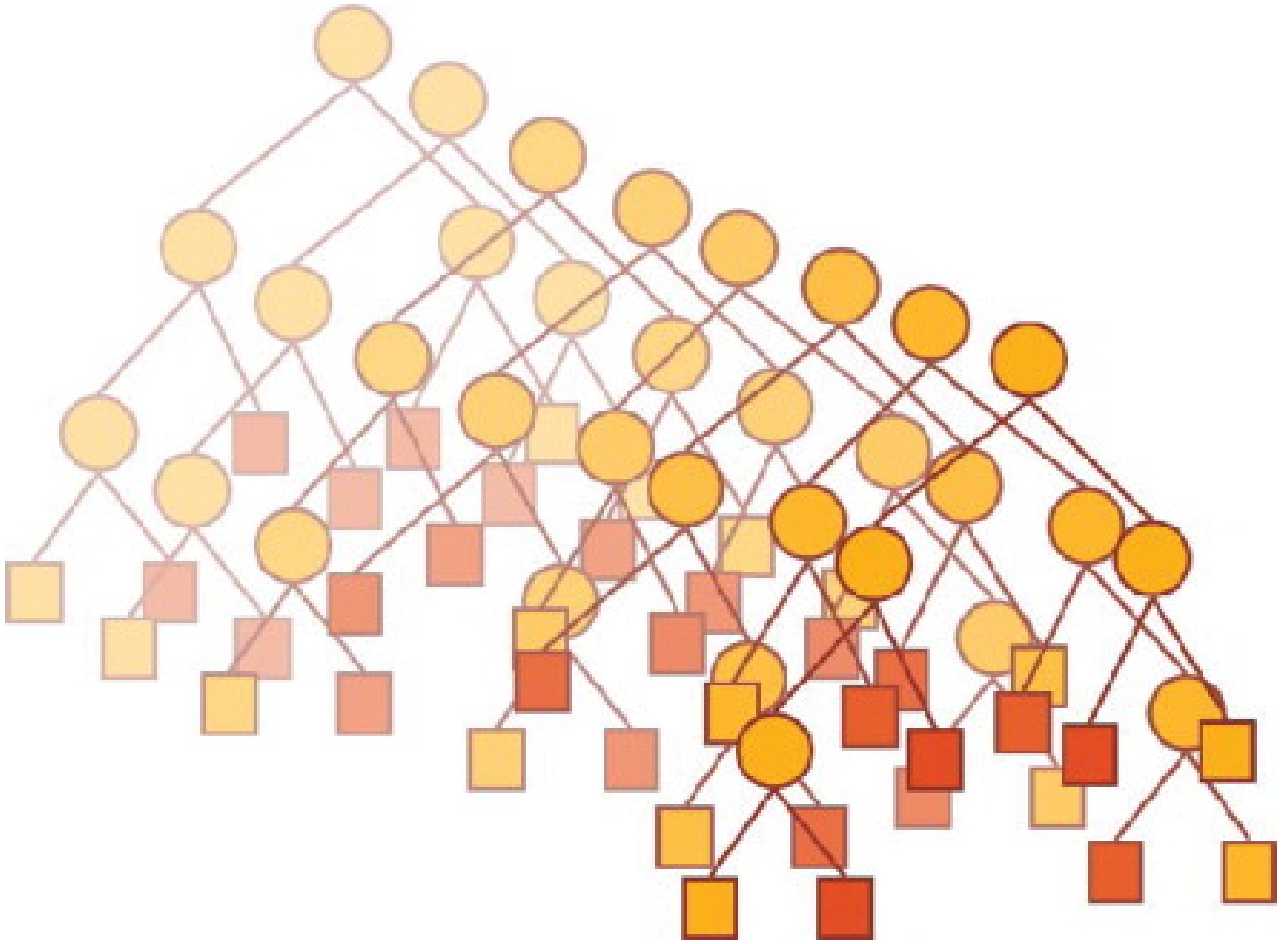
Regularization multiplier = 6

ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN

ÁRBOL DE REGRESIÓN



RANDOM FOREST



Fuente: Gedeck et al. 2010 Progress in Medicinal Chemistry

RANDOM FOREST

- Parámetros importantes:
  - **ntree**: número de árboles a calibrar
  - **mtry**: número de variables usadas en cada árbol
  - **nodesize**: número mínimo de casos en cada nodo terminal
  - **maxnode**: número máximo de nodos terminales

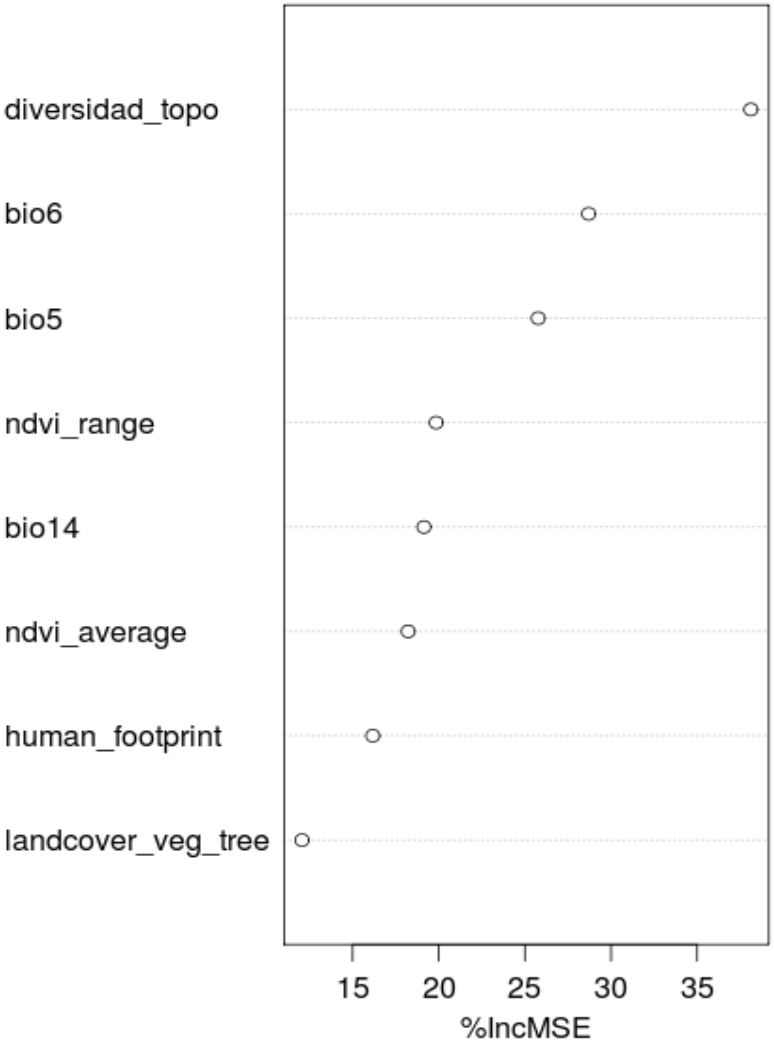
RANDOM FOREST

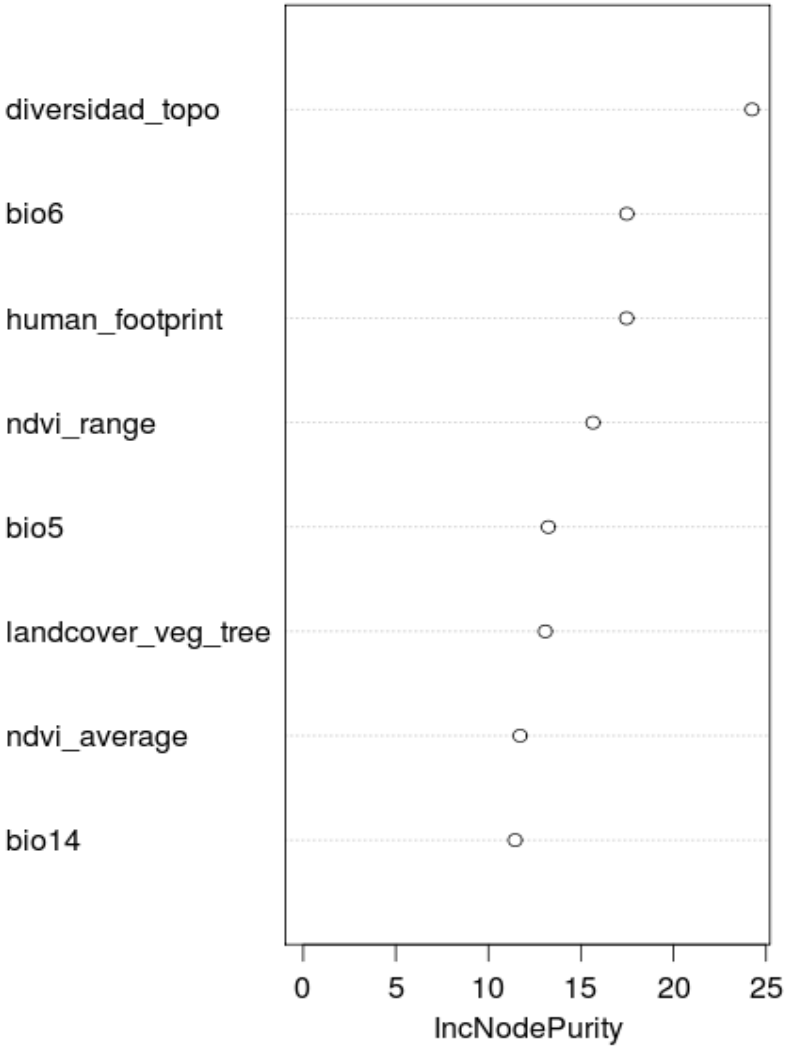
1. Por cada árbol:
    1. Selecciona **n** variables al azar
    2. Selecciona 60% de datos al azar
    3. Calibra un árbol de regresión
    4. Evalúa el árbol con el 40% de los datos no usados para calibrarlo
2. Una vez calibrados todos los árboles
    1. Calcula el resultado de un nuevo caso (celda) para cada uno de los árboles
    2. Calcula la moda del resultado de todos los árboles

RANDOM FOREST

- Ventajas
  - Muy potente
  - Puede manejar gran cantidad de datos
  - Analiza interacción de variables
- Inconvenientes
  - Potencial sobreajuste a los datos
  - El resultado es difícil de interpretar

IMPORTANCIA VARIABLES





BOOSTED REGRESSION TREES

- Gradient boosting Modelling (GBM).
- Importante mirar “A working guide to Boosted Regression Trees” (Elith_2008.pdf en la carpeta de artículos).
- Las “vignettes” del paquete “dismo” también son un buen punto de partida.

BOOSTED REGRESSION TREES

- Características:
  - Genera árboles de regresión
  - Componente estocástico (como Random Forest)
  - “Boosting”: método de optimización para reducir el error del modelo.
  - Selecciona las variables más relevantes y la cantidad necesaria de árbiles.
  - El modelo final es una combinación lineal de muchos árboles
  - Permite evaluar la interacción entre variables.

BOOSTING

- **Loss function:** mide la pérdida en capacidad predictiva debido a modelos subóptimos.
- **Boosting**: minimiza esa función:
  - Genera árbol (**t1**) de forma que minimize la loss function lo máximo posible.
  - Genera el árbol (**t2**) que mejor explica los residuales de **t1** (los residuales indican la varianza no explicada por el modelo).
  - Añade **t2** al modelo, y se calculan los residuales de **t1 + t2**.
  - Genera el árbol (**t3**) que mejor se ajusta a los residuales de **t1 + t2.** Repite hasta que la loss function no se puede minimizar más.

ENSAMBLADO DE MODELOS

ENSAMBLADO

- JW Gibbs (1878): Muchas copias de un sistema consideradas simultáneamente. Cada copia representa un estado posible del sistema.
- JM Bates y CWJ Granger (1969): Un ensamblado tiene una probabilidad de error menor que cualquiera de sus constituyentes individuales.
- Araújo y New 2006: Al promediar varios modelos la señal objetivo emerge del ruido asociado a los errores e incertidumbres de los modelos individuales.

ENSAMBLADO

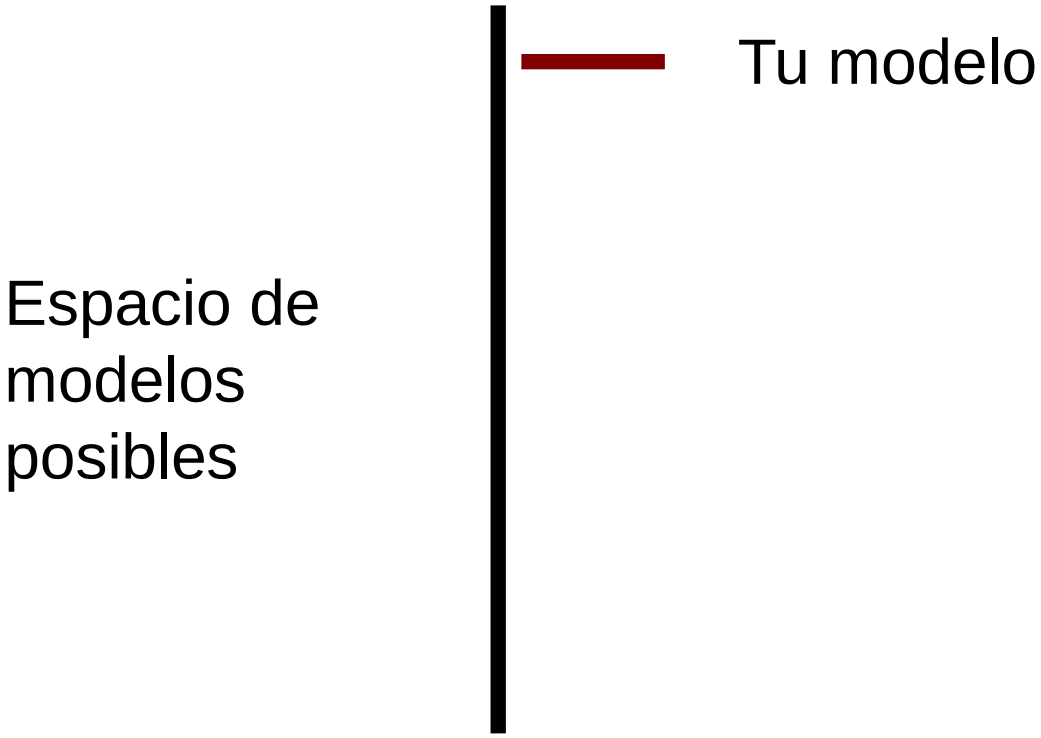
Las “copias” varían a lo largo de varios ejes

- Condiciones iniciales (tanto presencias como variables)
- Tipos de modelos
- Parámetros de los modelos

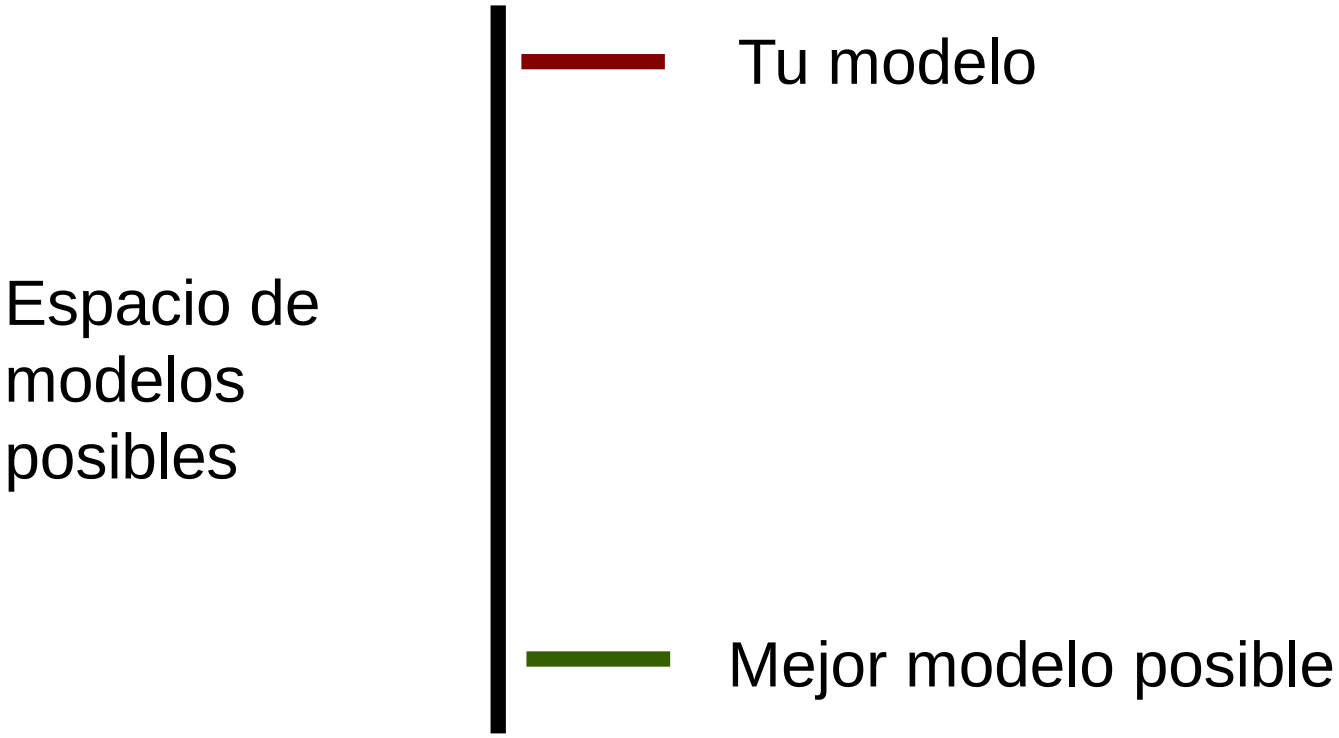
ENSAMBLADO

Espacio de modelos posibles

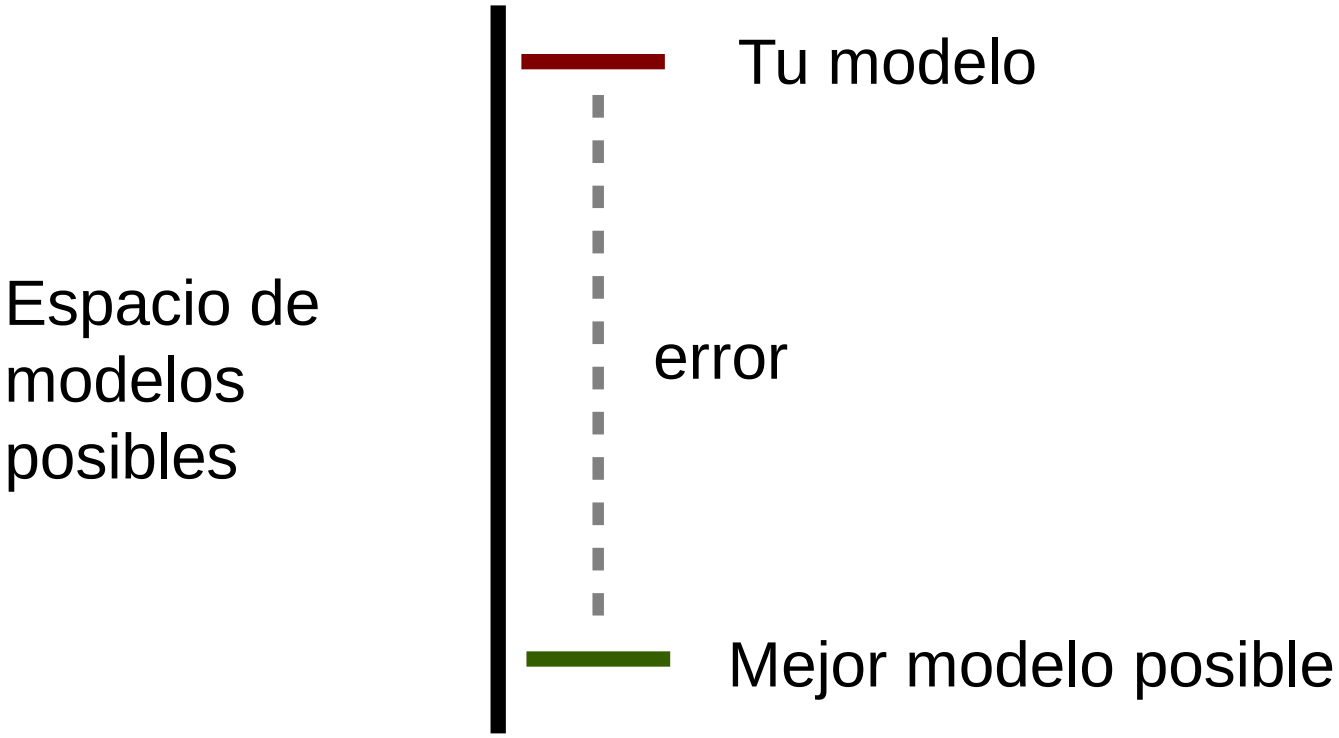
ENSAMBLADO



ENSAMBLADO

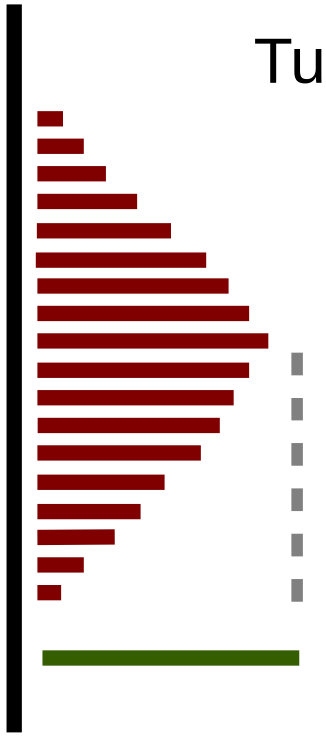

Tu modelo
Espacio de modelos posibles
Mejor modelo posible

ENSAMBLADO


Espacio de modelos posibles
Tu modelo
error
Mejor modelo posible

ENSAMBLADO

Espacio de modelos posibles

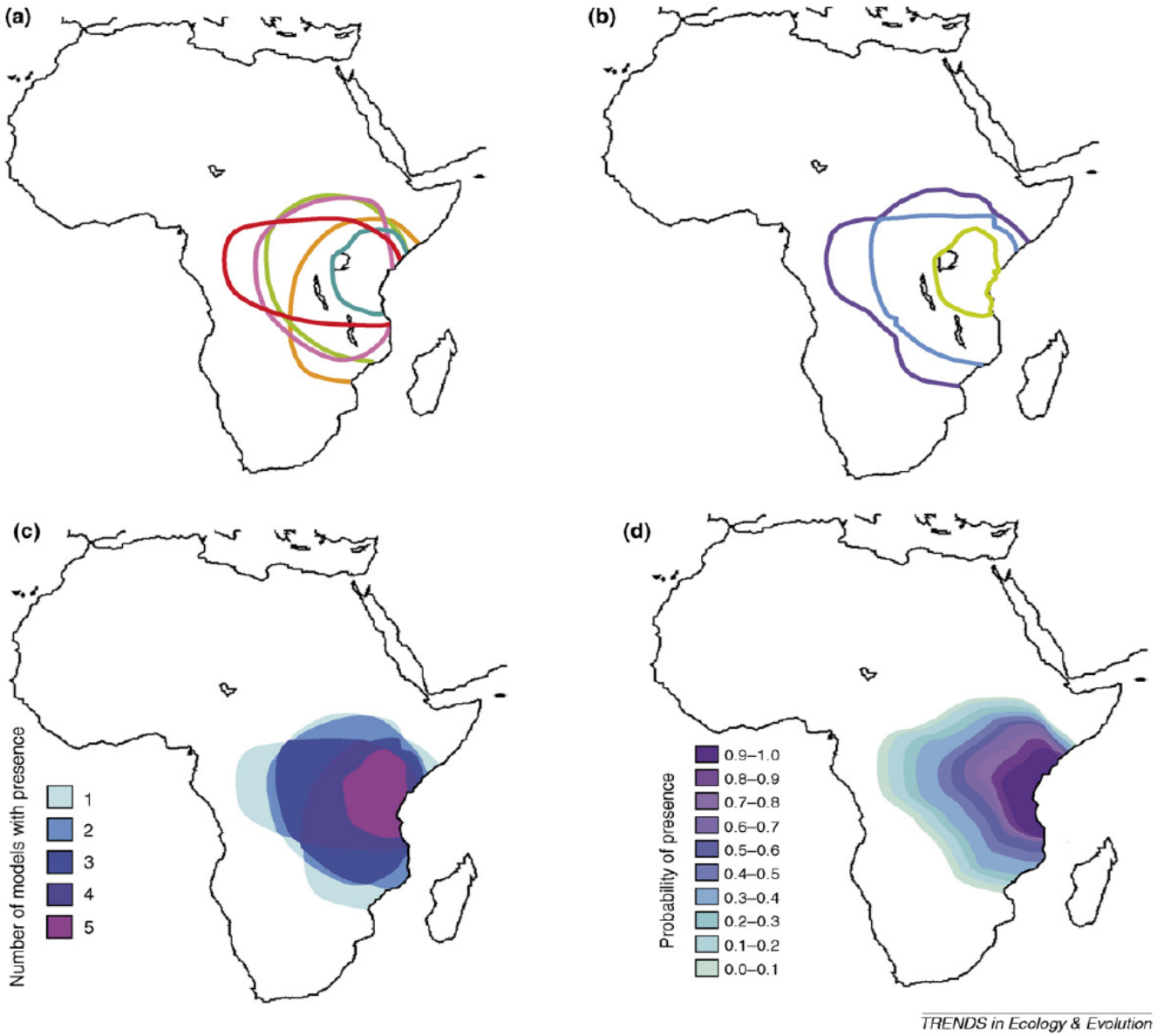


Tu modelo

Error más reducido

Mejor modelo posible

MÉTODOS DE ENSAMBLADO



MÉTODOS DE ENSAMBLADO

- Mediana **
- Media aritmética ** (buen método, Marmion 2009)
- Media ponderada según valores de AUC **
- Selección de modelos con mayor AUC
- PCA: primer componente refleja la tendencia general. Se seleccionan los modelos más relacionados con este componente, y se les calcula la mediana

Ojo con las escalas de valores de los modelos, deben ser todas iguales!

EVALUACIÓN DE MODELOS

EVALUACIÓN

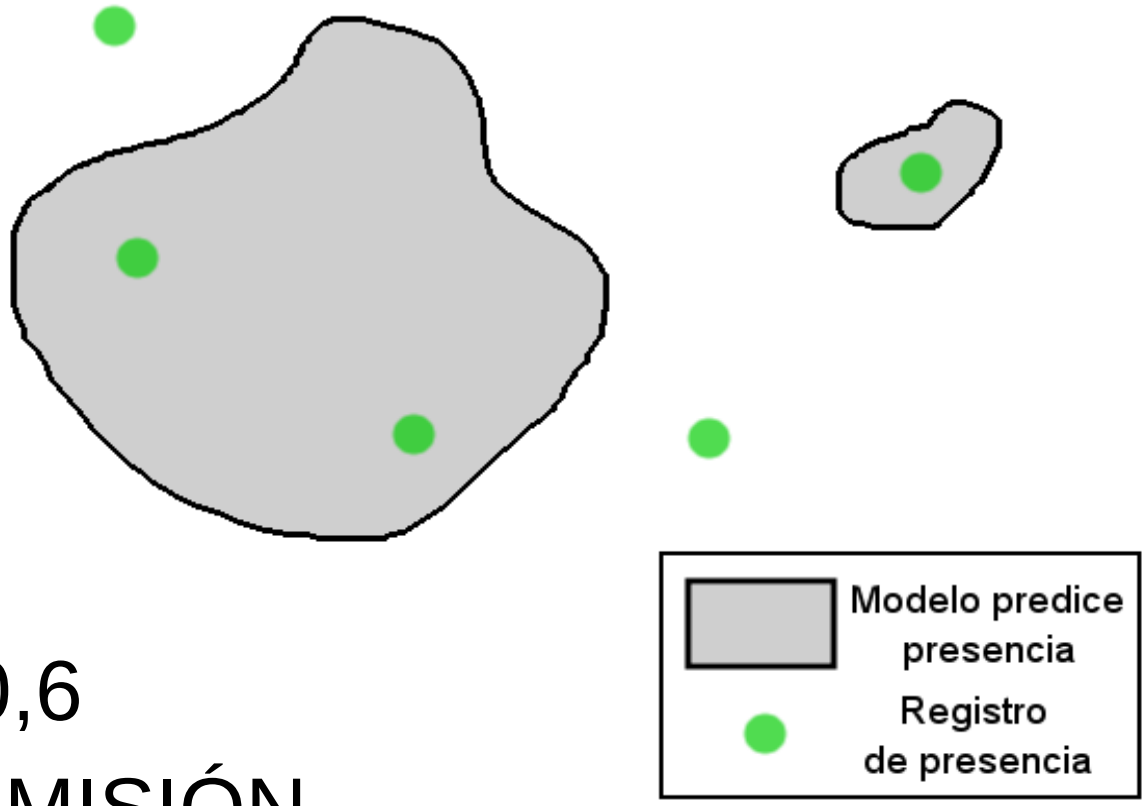
Artículo clave:

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

(2856 citas en abril de 2014)

SOLO-PRESENCIA EN MODELOS BINARIOS

EVALUACIÓN



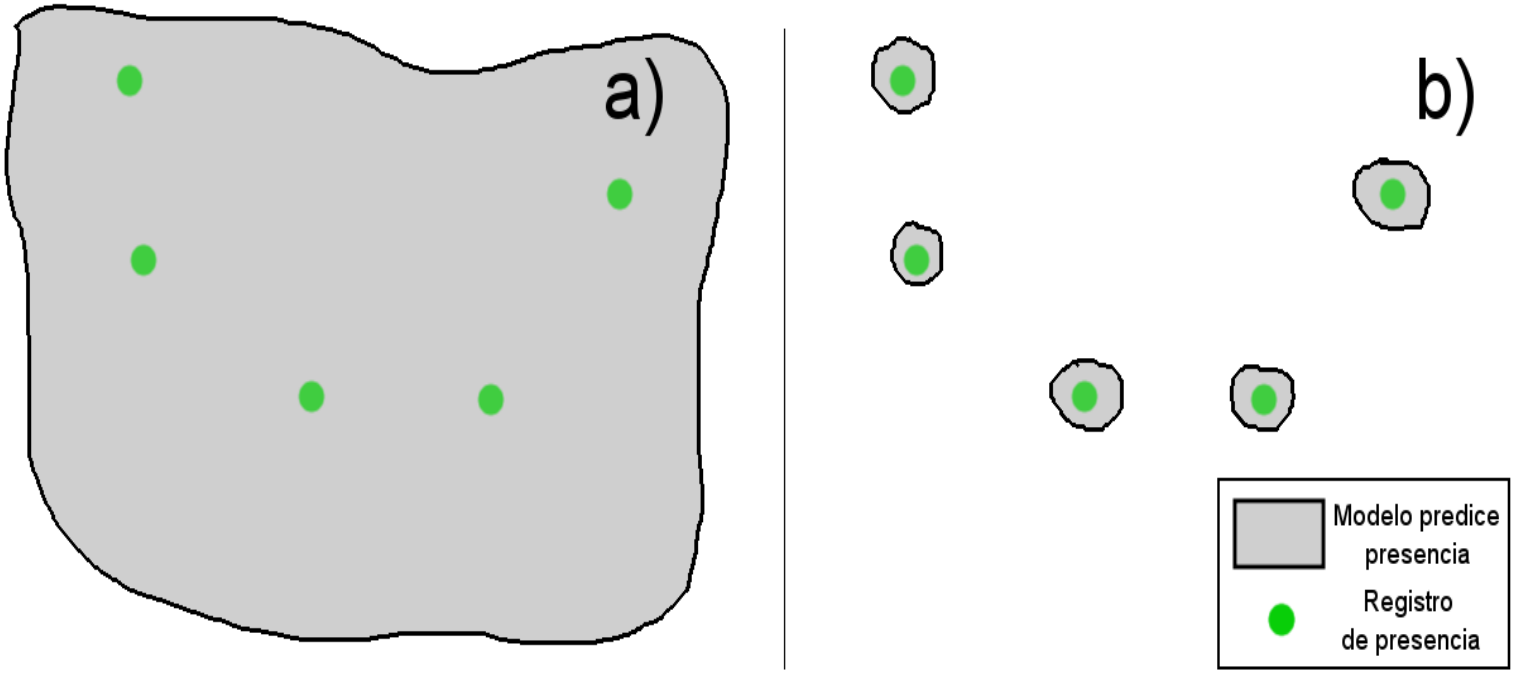
5 presencias

3 aciertos

Sensibilidad=0,6

2 errores de OMISIÓN

EVALUACIÓN



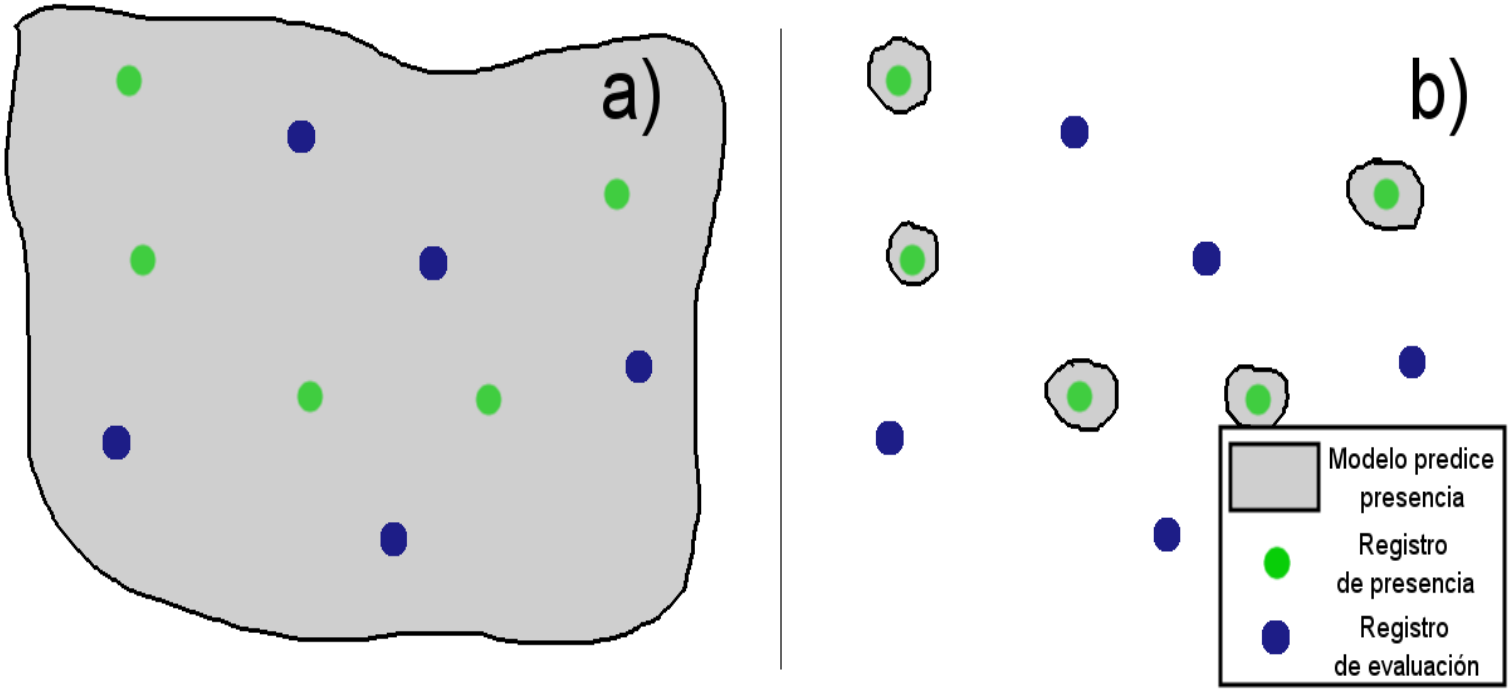
Sensibilidad=1

¿Error de comisión?

Sensibilidad=1

¿Sobreajuste?

EVALUACIÓN



Sensibilidad=1

¿?

Sensibilidad=0

¡Sobreajuste!

PRESENCIA – AUSENCIA EN MODELOS BINARIOS

MATRIZ DE CONFUSIÓN

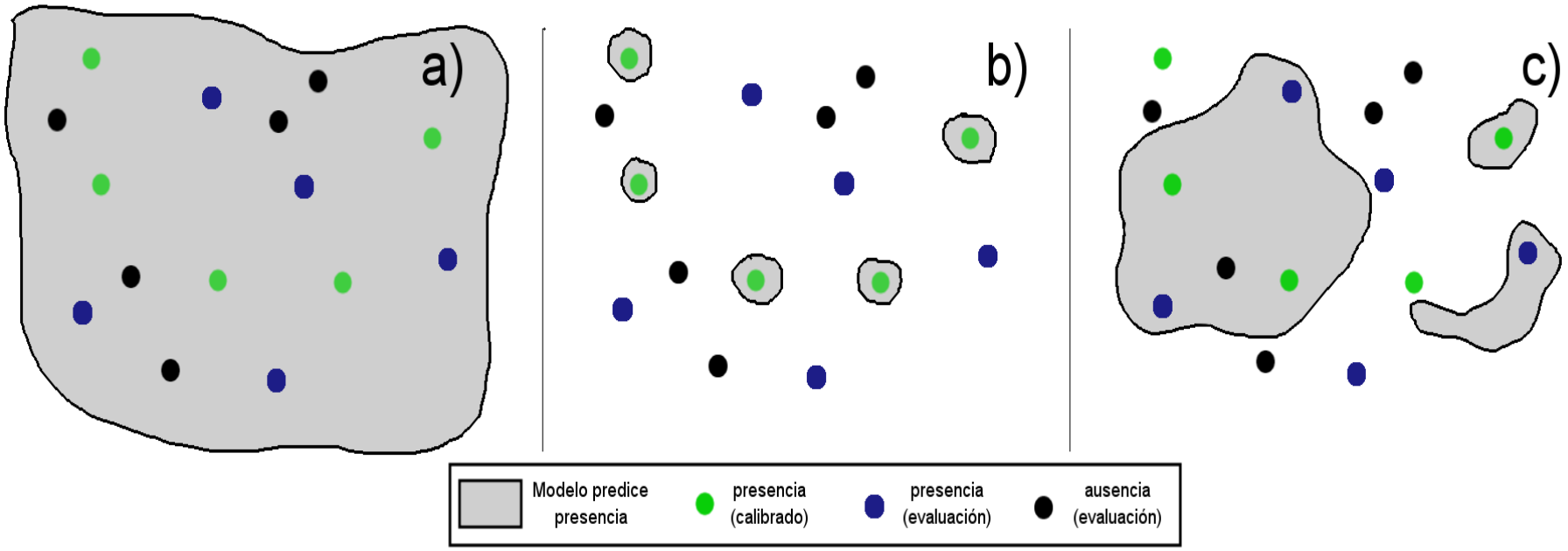
A → presencias acertadas
D → ausencias acertadas
B → ausencias fallidas (falsos positivos o **error de comisión**)
C → presencias fallidas (falsos negativos o **error de omisión**)

		Datos reales (registros de presencia y ausencia)	
		presencia	ausencia
Datos simulados (modelo de distribución)	presencia	**A**	**B**
	ausencia	**C**	**D**

SENSIBILIDAD = A/(A+C)

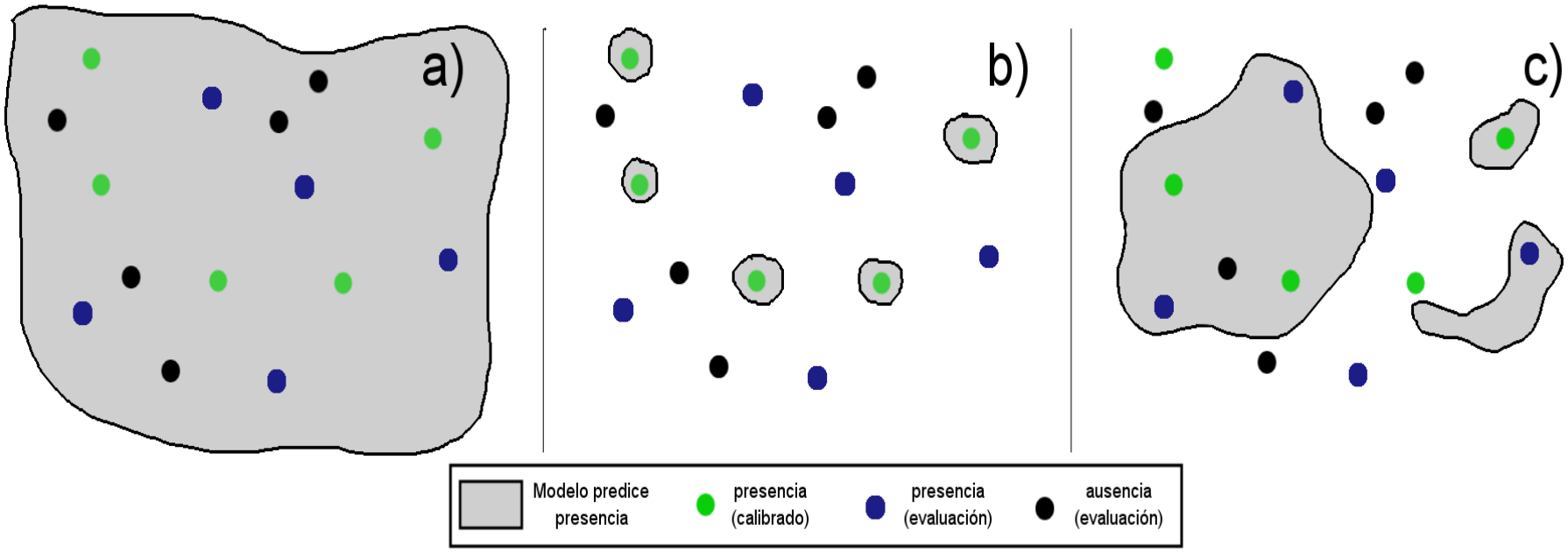
ESPECIFICIDAD = D/(B+D)

EVALUACIÓN



		Datos reales (registros de presencia y ausencia)					
		modelo a		**modelo b**		**modelo c**	
		pres.	aus.	pres.	aus.	pres.	aus.
Datos simulados (modelo de distribución)	presencia	5	5	0	0	3	1
	ausencia	0	0	5	5	2	4

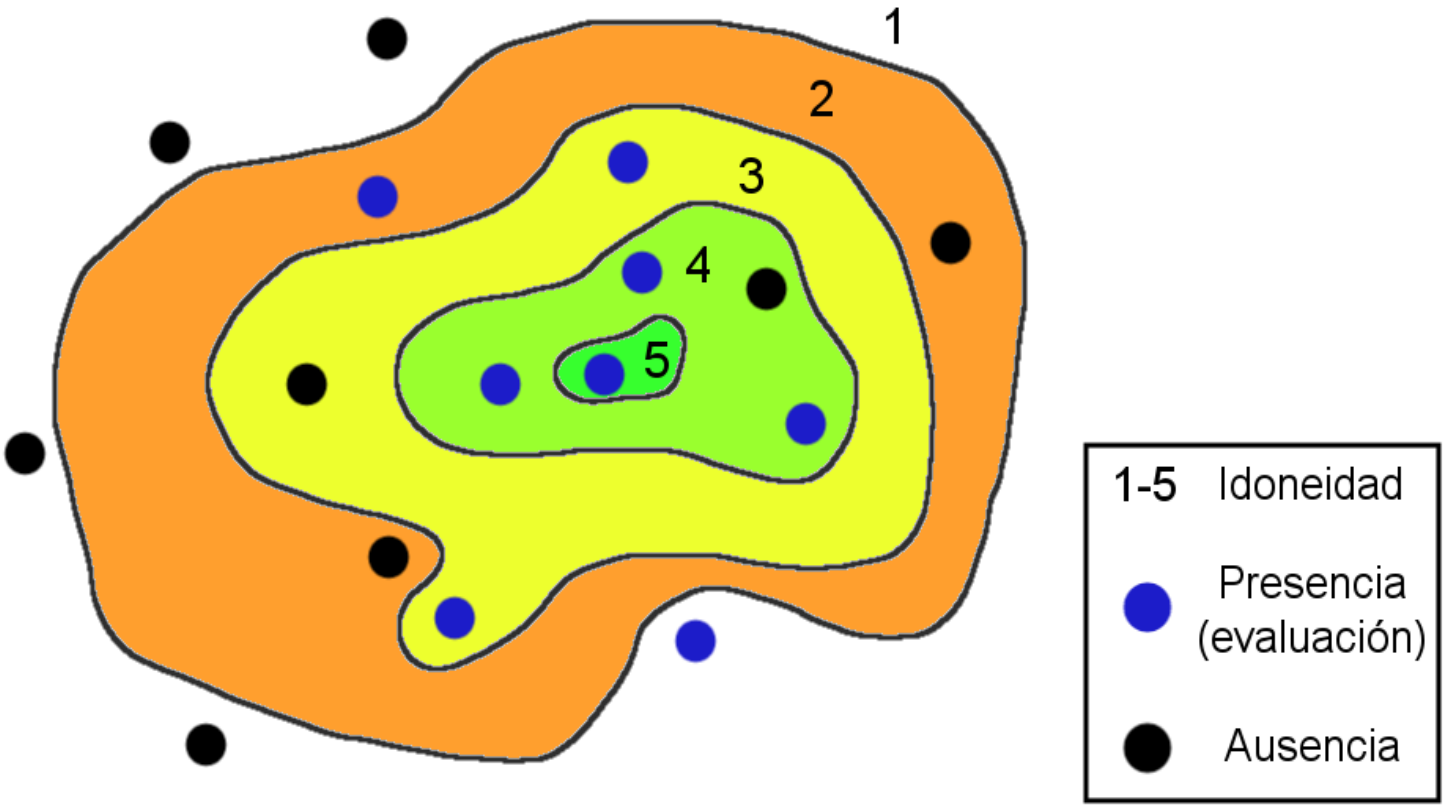
EVALUACIÓN



modelo	a	b	c
sensibilidad	1	0	0.6
especificidad	0	1	0.8

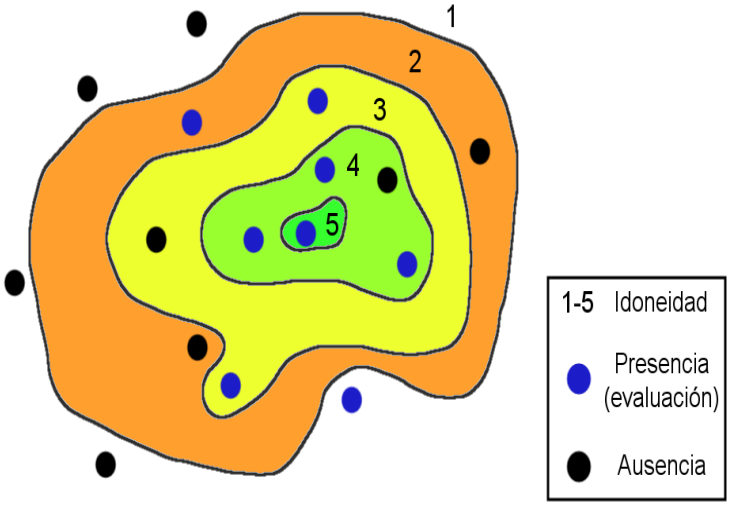
PRESENCIA – AUSENCIAS EN MODELO CONTINUO

CURVA ROC



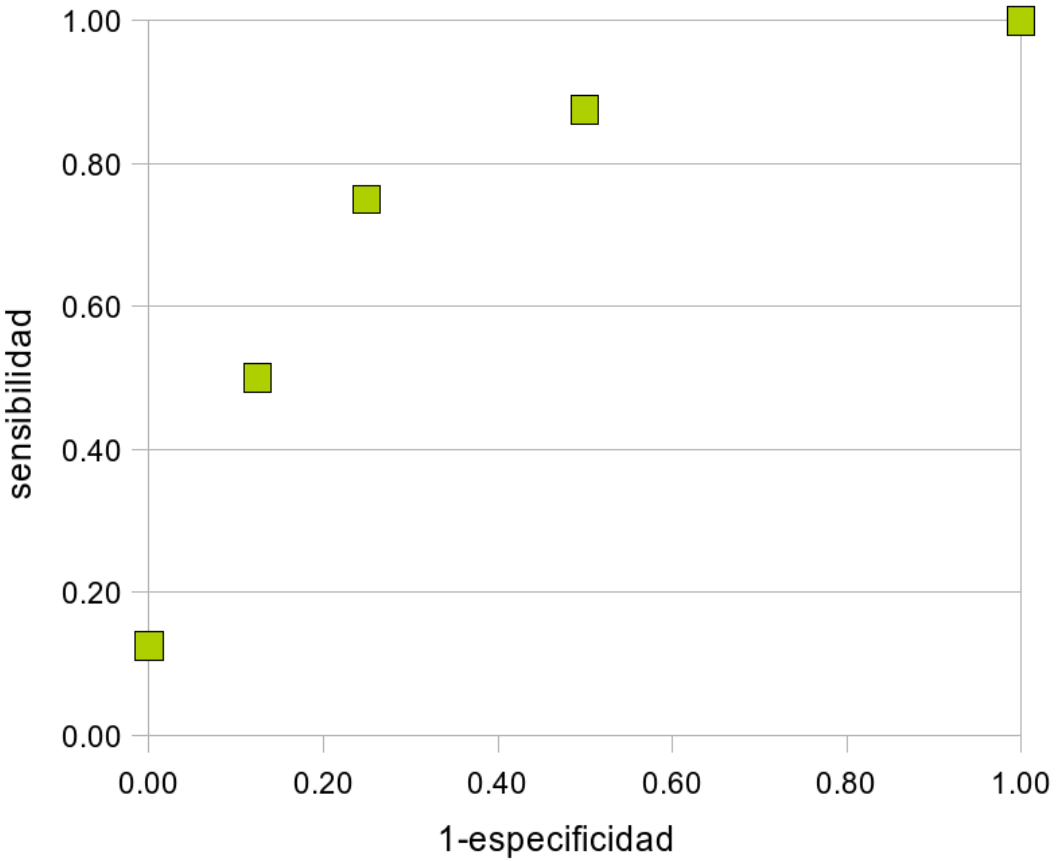
CURVA ROC

Observa que en lugar de la especificidad, usamos 1-especificidad

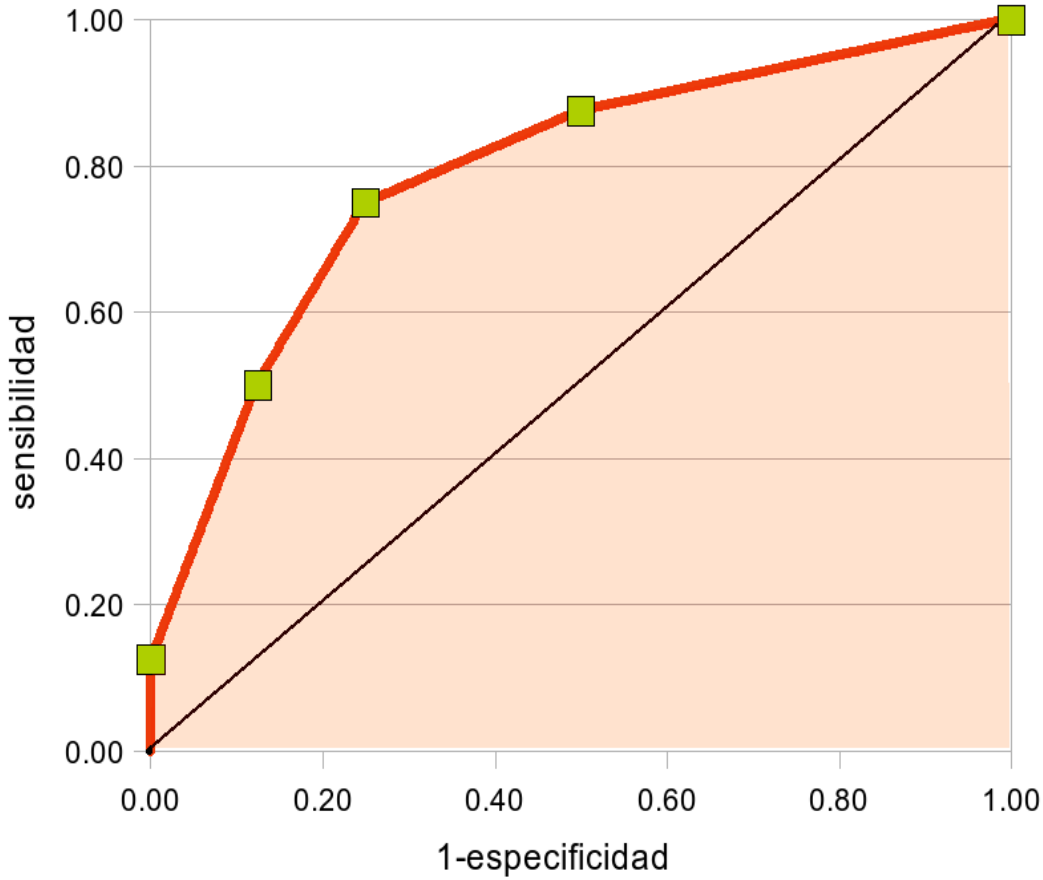


presencias acertadas	ausencias falladas	presencias falladas	ausencias acertadas	tamaño de muestra			
A	**B**	**C**	**D**	**N**	**UMBRAL CORTE**	**SENSIBILIDAD**	**1-ESPECIFICIDAD**
8	8	0	0	16	1	1.00	1.00
7	4	1	4	16	2	0.88	0.50
6	2	2	6	16	3	0.75	0.25
4	1	4	7	16	4	0.50	0.13
1	0	7	8	16	5	0.13	0.00

CURVA ROC


1.00
0.80
0.60
0.40
0.20
0.00
sensibilidad
0.00
0.20
0.40
0.60
0.80
1.00
1-especificidad

CURVA ROC


1.00
0.80
0.60
0.40
0.20
0.00
sensibilidad
0.00
0.20
0.40
0.60
0.80
1.00
1-especificidad

CURVA ROC

Area Under the Curve ROC (“Receiver Operating Characteristic”) -> probabilidad de que, seleccionando al azar una presencia y una ausencia, el modelo clasifique con un valor de idoneidad mayor a la presencia que a la ausencia.

Suponiendo AUC = 0.74, el modelo dará mayor valor de idoneidad a las presencias un 74% de las veces

PRESENCIA – ALEATORIOS EN MODELO CONTINUO

MATRIZ DE CONFUSIÓN MODIFICADA

A → presencias acertadas
D → ya no es un acierto
B → ya no es un error
C → presencias fallidas (falsos negativos o **error de omisión**)

		Datos reales (registros de presencia y puntos aleatorios)	
		presencia	aleatorio
Datos simulados (modelo de distribución)	presencia	**A**	**B**
	ausencia	**C**	**D**

CURVA ROC CON PUNTOS ALEATORIOS

- Cambia el significado: AUC es la probabilidad de que un punto de presencia seleccionado al azar tenga un valor de idoneidad más alto que el de un punto aleatorio seleccionado al azar.
- Pero ahora AUC siempre será menor que 1, porque siempre habrá puntos aleatorios sobre áreas de hábitat idóneo.

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.

ALGUNOS PROBLEMAS DE AUC

- Considera regiones del espacio ROC en los que no se trabaja, como los extremos de la curva, en los que las tasas de error son elevadas.
- Pondera por igual los errores de comisión y omisión.
- No informa de distribución espacial de los errores.
- Las áreas de trabajo amplias resultan en valores de AUC más altos.
- No pueden compararse modelos de distintas especies.

MÉTODOS DE EVALUACIÓN DE MODELOS

- Datos independientes: 1 solo valor de AUC por modelo
- Sin datos independientes: Cross validation
  - Data splitting: separas un conjunto de datos para calibrar el modelo, y otro para evaluar
  - K-fold: separación de los datos en n grupos. Calibras con n-1 y evalúas con 1.
  - Bootstrap: partición iterativa de los datos, para calibrar con unos, y evaluar con otros.
  - Leave-one-out: para muestras pequeñas.

AÚN ASÍ...

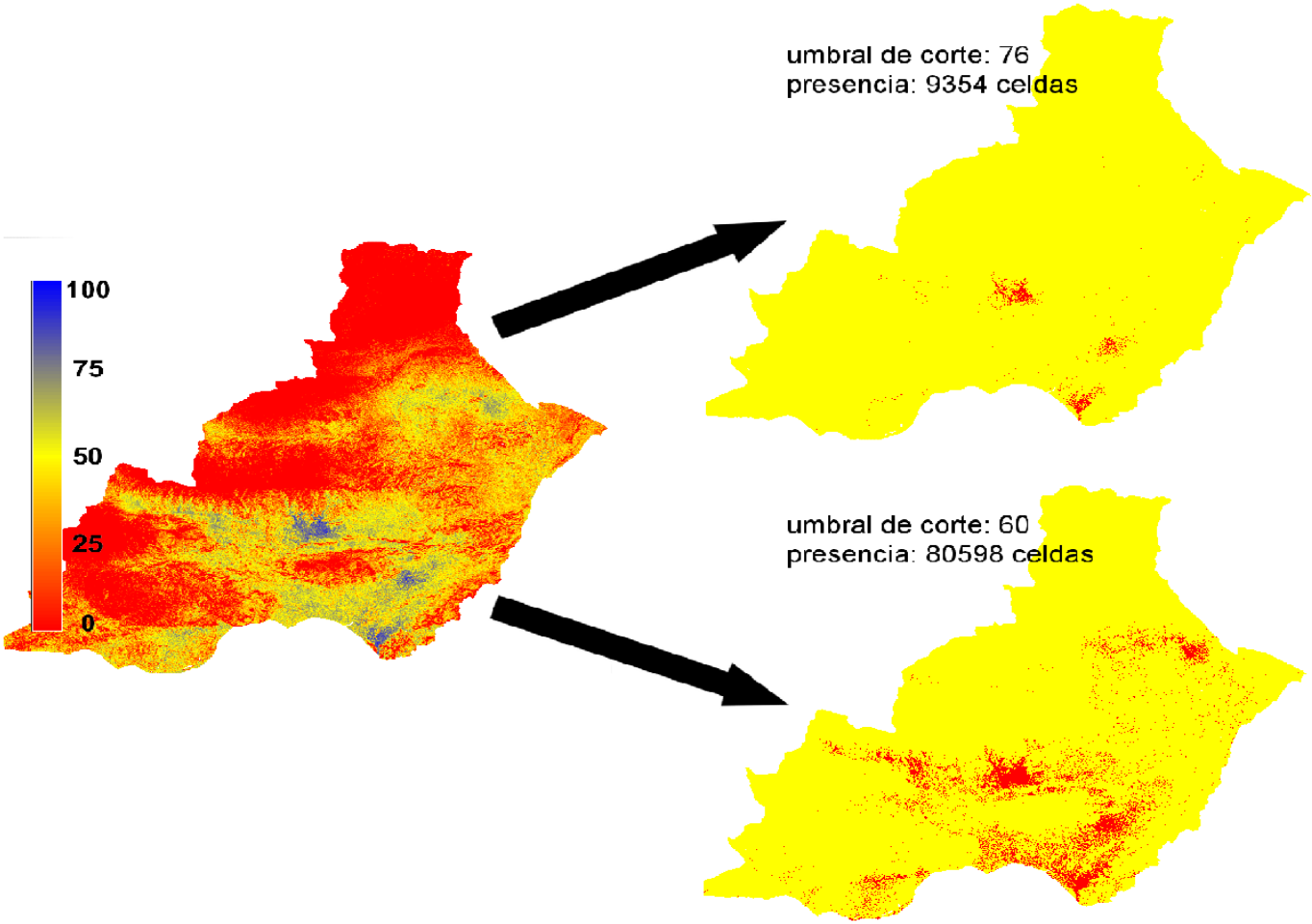
Es una buena herramienta para comparar modelos para la misma especie y área de trabajo

APLICACIÓN DE THRESHOLDS

DE CONTINUO A BINARIO

- Un mapa de valores binarios (0 o 1) es más fácil de comprender que uno continuo (0 a 1)
- Para transformar MDE continuos en binarios:
  - seleccionamos un valor de referencia: umbral (threshold)
  - asignamos valor 1 a todas las celdas por encima del umbral
  - asignamos valor 0 a todas las celdas por debajo del umbral

DE CONTINUO A BINARIO



DE CONTINUO A BINARIO

- ¿Cómo seleccionamos el umbral?...
  - Liu et al. 2005
  - Jiménez-Valverde y Lobo 2007
  - Freeman y Moisen 2008
    - Selección subjetiva
    - Selección “objetiva”

SELECCIÓN SUBJETIVA

“Elecciones arbitrarias sin base ecológica” (Osborne et al. 2001)

- Valores fijos: 0.5, 0.3, ...
- Porcentaje de comisión: 95%, 90%, ...

SELECCIÓN OBJETIVA

"El umbral se selecciona para maximizar la concordancia entre la distribución observada y la modelada" (Liu et al 2005)

- maximización de Kappa (no recomendado)
- punto de curva ROC con pendiente = 1
- valor con igual sensibilidad y especificidad
- y muchos más en Freeman y Moisen 2008

Estos criterios requieren datos de presencia - ausencia!

'COSAS' IMPORTANTES

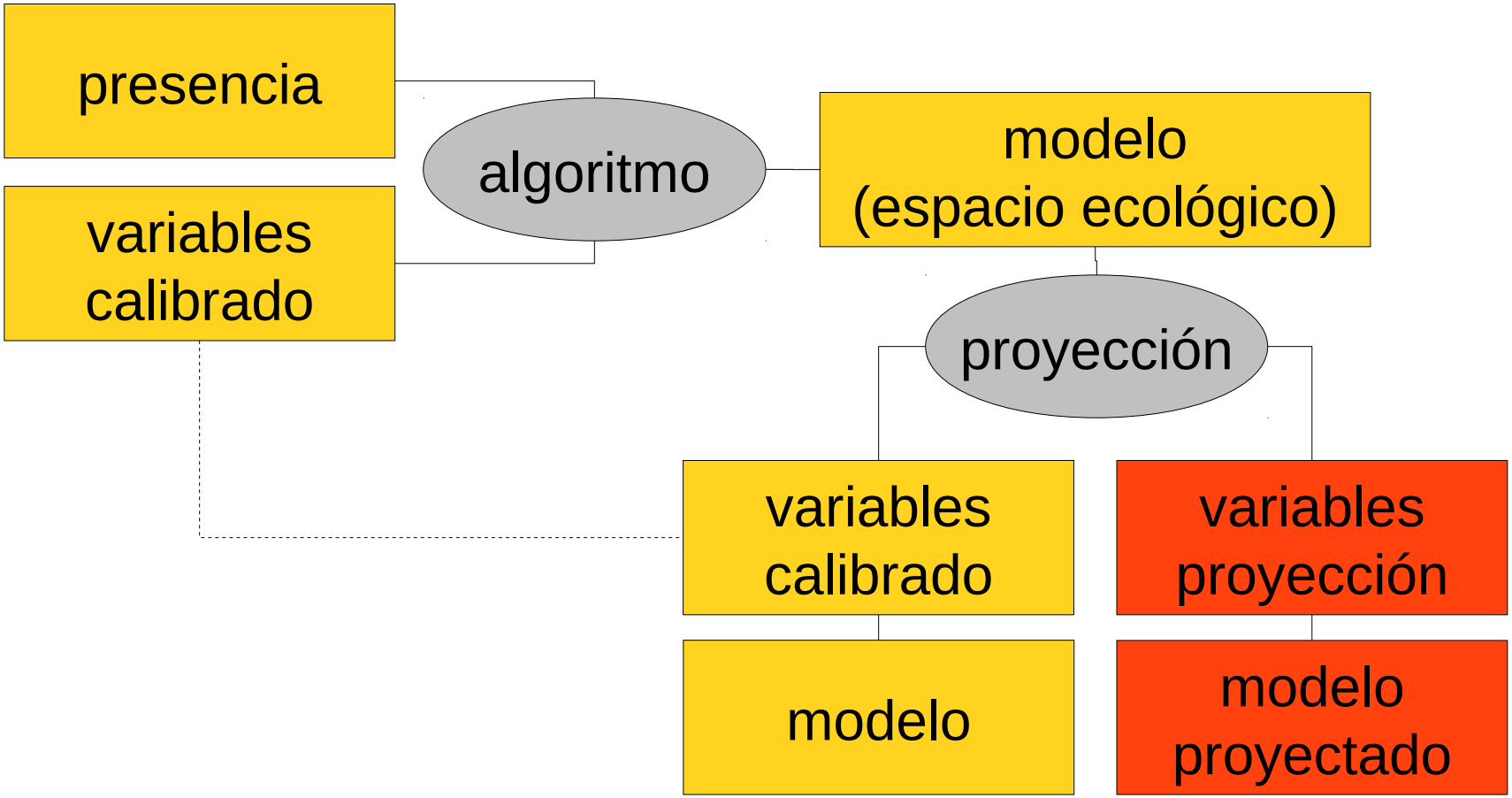
- La elección del umbral depende del objetivo del modelo, no hay una norma fija
- Los modelos de especies con pocas presencias o mal ajuste son muy sensibles a la elección del threshold
- No hay obligación de aplicar un threshold a un modelo, salvo que tu análisis concreto lo requiera
- Siempre es mejor usar la versión continua de un modelo

THRESHOLD CON R

- Con R podemos calcular las estadísticas de las presencias sobre el modelo para tomar decisiones
- También podemos ver el gráfico de densidad los valores de las presencias sobre el modelo con 'extract', 'density' y 'plot' para decidir manualmente un punto de corte
- La función 'evaluate' de dismo ofrece herramientas para calcular thresholds

PROYECCIÓN DE MODELOS DE DISTRIBUCIÓN EN EL ESPACIO Y EN EL TIEMPO

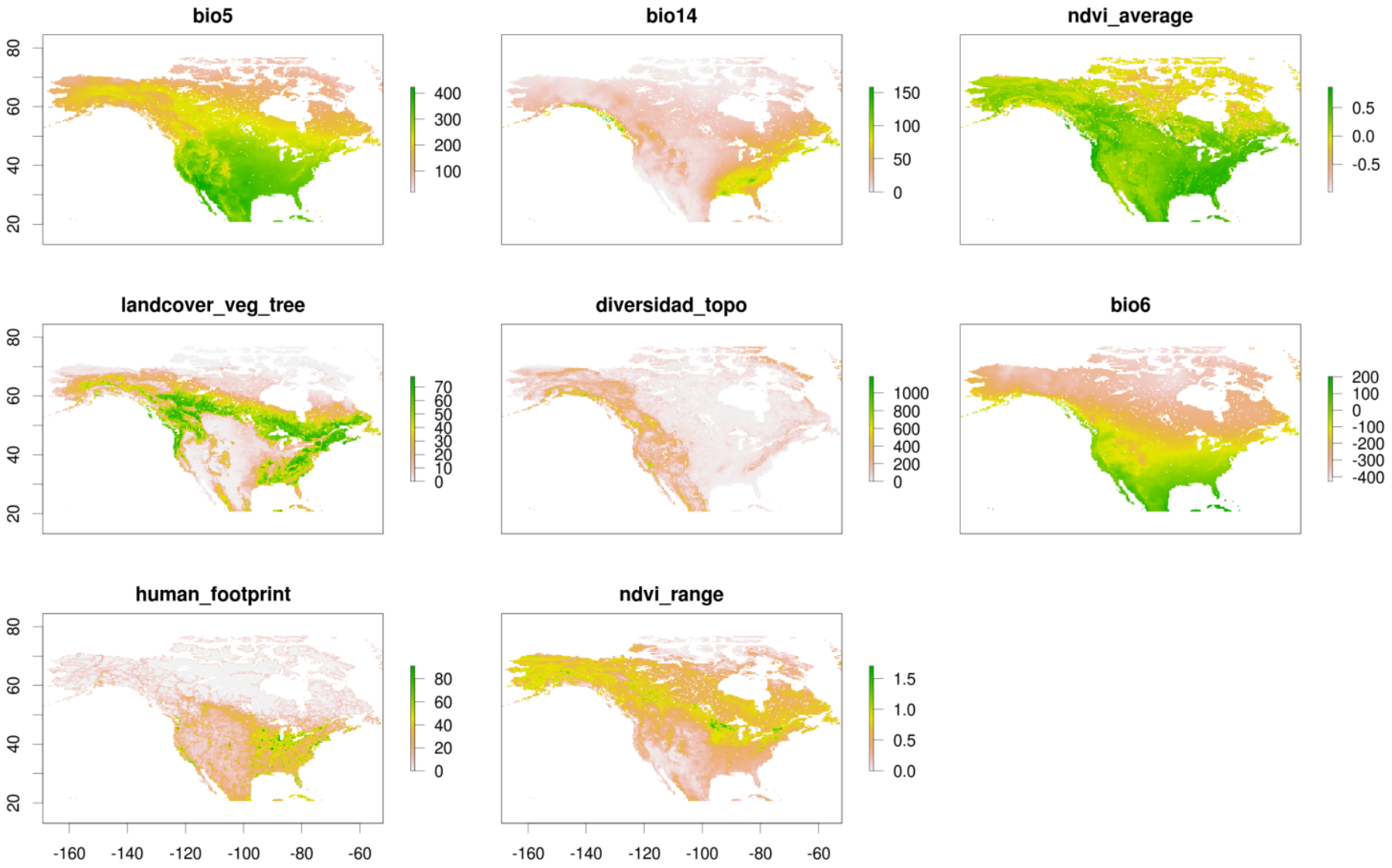
PROYECCIÓN DE MODELOS



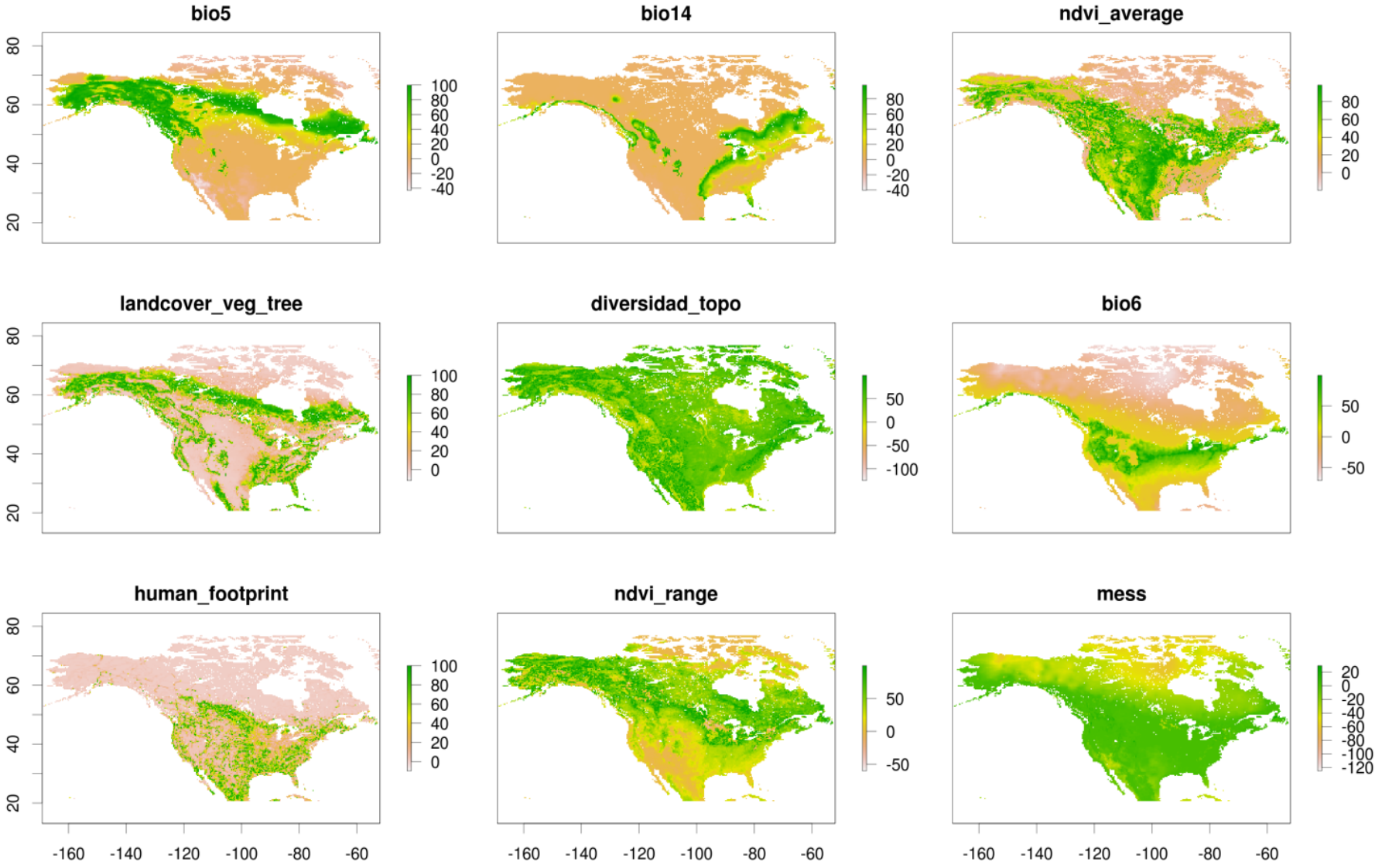
MULTIVARIATE ENVIRONMENTAL SIMILARITY SURFACES (MESS)

- Índice de similaridad entre el espacio ecológico que ocupan las presencias, y el espacio ecológico que hay en las variables de proyección.
- Cuanto más diferente sean, más estaremos extrapolando, ¡¡Y EXTRAPOLAR ES MALO!!
- Referencia: Elith J., Kearney M., & Phillips S. 2010. The art of modelling range shifting species. Methods in Ecology and Evolution, 1 :330-342.

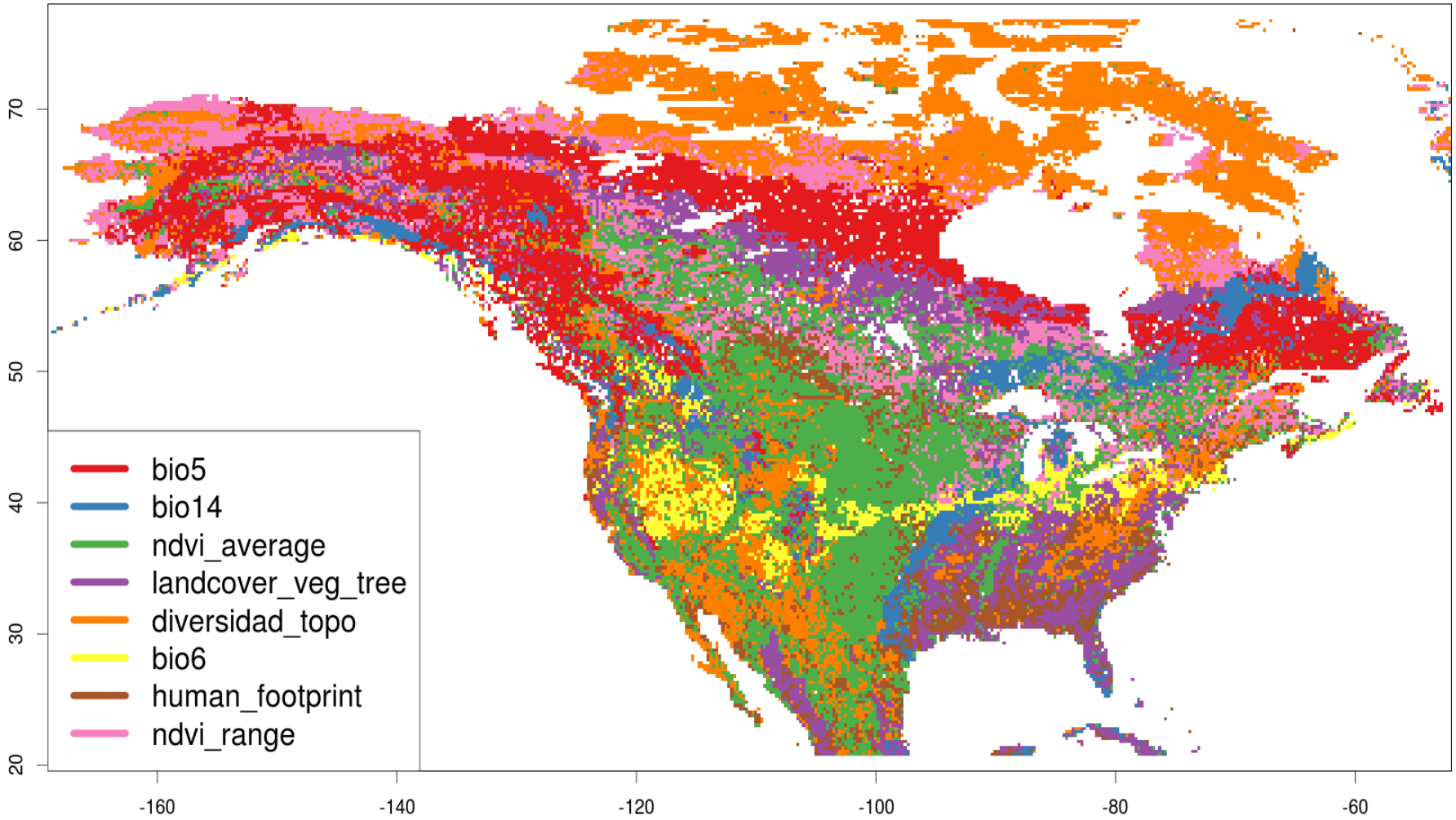
VARIABLES


bio5
bio14
ndvi_average
landcover_veg_tree
diversidad_topo
bio6
human_footprint
ndvi_range
80
60
40
20
400
300
200
100
150
50
0
0.5
0.0
-0.5
70
10
1000
800
600
200
-100
-200
-300
-400
-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 (máximo)


bio5
bio14
ndvi_average
landcover_veg_tree
diversidad_topo
bio6
human_footprint
ndvi_range
70
60
50
40
30
20
-160
-140
-120
-100
-80
-60

PROYECCIÓN EN EL ESPACIO

- Una región origen y una de destino
- Calibramos el modelo en la de origen
- Proyectamos el modelo en la de destino
- Necesitamos tener las mismas variables CON LOS MISMOS NOMBRES para ambas regiones, idealmente a la misma resolución.
- Uso más extendido: invasibilidad
- Problemas: a million...

PROYECCIÓN EN EL TIEMPO

- Una región y tiempo de origen y un tiempo de destino (pasado o futuro)
- Calibramos el en origen y proyectamos en destino
- Mismos nombres de las variables
- Algunas variables no están disponibles para pasado o futuro (ndvi, human footprint, etc)
- Suelen hacerse solo con clima y topografía
- Uso más extendido: cambio climático, paleodistribuciones
- Problemas: a million...

FUENTES DE PALEOCLIMA

- **Último interglacial** (120 – 140 kyr BP) disponible en www.worldclim.org/past (resolución: 1 km)
- **Último máximo glacial** (21 kyr BP) según los modelos CCSM y MIROC, procedente de PMIP3 (pmip3.lsce.ipsl.fr) y disponible en www.worldclim.org/past (resolución: 5km)
- **Holoceno medio** (6 kyr BP) solo disponibles en PMIP3 (pmip3.lsce.ipsl.fr)
- **TraCE-21ka** (21 kyr BP → presente) disponible en www.cgd.ucar.edu/ccr/TraCE a resolución grosera (2º, solo para escalas continentales)

¿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.


That's all Folks!