UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE ESTUDIOS ESTADÍSTICOS

TRABAJO FIN DE MASTER EN MINERIA DE DATOS E INTELIGENCIA DE NEGOCIOS

Distribución observada y potencial del género *Aphodius (Illiger 1798)* de la Península Ibérica (Coleoptera, scarabaeoidea)

JOSÉ LUIS AGUILAR COLMENERO
DNI: 77342390Z

Curso Académico 2015/2016

Índice

Índice de Tablas

INTRODUCCIÓN

Si hay alguna expresión artística de la sociedad que enmarque las preguntas que todos nos hemos planteado alguna vez ese es el cuadro de *Paul Gauguin* "¿De dónde venimos? ¿Quiénes somos? ¿Adónde vamos?"

Conocer por qué estamos asentados en un lugar, o hacia dónde vamos es algo que no solo trasciende a los humanos, también este tipo de pensamiento lo podemos trasladar al resto de especies del planeta ¿De dónde vienen?, ¿Quiénes son?, ¿A dónde van?, en el presente trabajo nos planteamos este tipo de cuestiones para un género de coleóptero en concreto, Aphodius (Illiger, 1798).

Tener conocimiento de su distribución y su distribución potencial en base a una serie de características puede ser de utilidad para la sociedad en la que vivimos actualmente la cual está aumentando a favor nuestro y en contra de este y otros géneros, además de poder ser de utilidad como herramienta para preservar la biodiversidad. El estudio de la relación entre la distribución y condiciones ambientales o topográficas es necesario para conocer la distribución geográfica de la biodiversidad (Jorge M. Lobo, 2000).

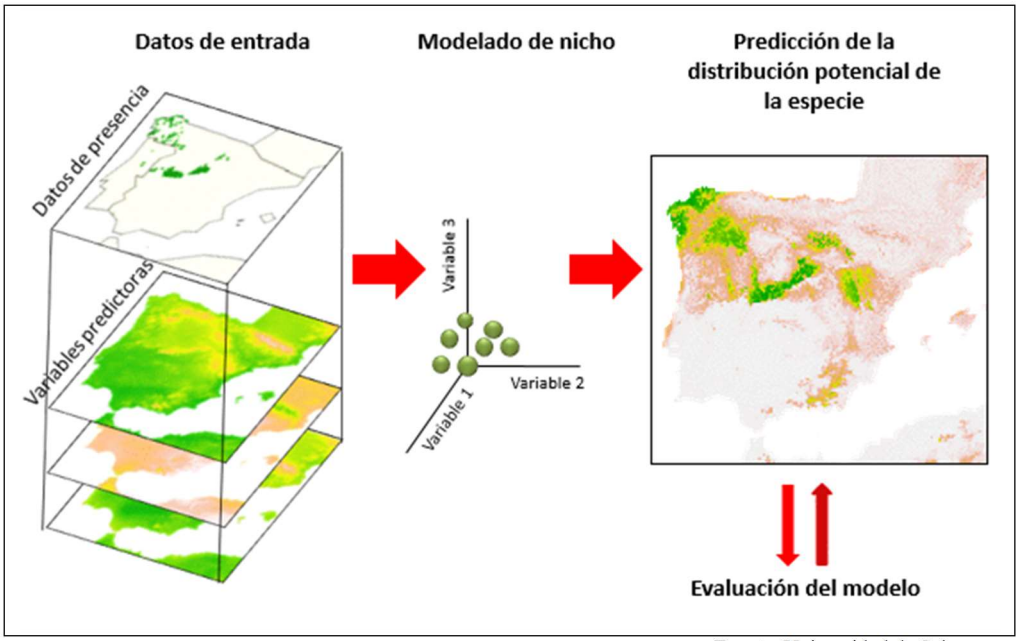
Dado los escasos muestreos y por tanto la escasa información sobre la presencia de muchas especies, es necesario utilizar modelos estadísticos para predecir la ocurrencia de dichas especies (Stockwell y Peters 2009; Phillips et al. 2006). Además, a la vez que se utilizan estos modelos para predecir hacia dónde o de qué forma los Aphodius pueden distribuirse, sirven para evaluar la importancia de variables de tipo ambiental en la ocurrencia de las especies (Cassini 2011).

Cuando nos referimos al análisis de la distribución de una especie en realidad estamos tratando de evaluar el **nicho ecológico**. Si combinamos una serie de condiciones ambientales a la distribución de una especie, nos referimos al concepto fundamental de nicho ecológico de Evelyn Hutchinson 1944-58. Sin embargo, mucho antes que Hutchinson, Joseph Grinnell en 1924 definió el concepto de nicho ecológico como "…la unidad de distribución más pequeña, dentro de la cual, cada especie se mantiene debido a sus limitaciones instintivas y estructurales". Por lo contrario, Charles Elton, expone un concepto en el que le da más importancia a la cadena trófica que a las condiciones o factores abióticos. Como se aprecia, desde hace tiempo se viene intentando consensuar una definición para el concepto de nicho ecológico, en la

actualidad el más nombrado o el que más repercusión ha tenido en ecología es el de Hutchinson, sin embargo hay autores recientes como Diego P Vazquez 2005 que indica algunas limitaciones en el concepto Hutchinsoniano.

En este trabajo utilizaremos variables ambientales para el modelado del nicho ecológico, por lo que seguiremos el concepto de Hutchinson; un esquema de este modelado se puede ver en la siguiente ilustración.

Ilustración 1. Esquema del modelado y predicción de la distribución potencial



Fuente: Universidad de Salamanca

Los Aphodius representan un género muy numeroso dentro del orden de los Coleópteros, en concreto en España se cuenta con 102 especies repartidas por todo el territorio nacional. Pertenecen a la familia Aphodiidae que en la mayoría de los casos tienen un comportamiento coprófago (Hortal, Lobo y Del Rey, 2006), aunque también son dados a la carroña.

Los datos geográficos o de presencia de las distintas especies de Aphodius han sido extraídos de la base de datos GBIF (Global Biodiversity Information Facility), es una base de datos "Open Access" en la que se recompila toda la información disponible a partir de literatura, museos y colecciones privadas, así como tesis y otros datos no publicados a nivel mundial y para toda la Península Ibérica (Lobo y Martín-Piera 1991). Actualmente contiene información de 1.643.699 especies y aumentando día a día, con respecto a la especie seleccionada para el estudio estamos hablando de unas 34.820 observaciones, de las cuales 3.256 son registros provenientes de España.

En la actualidad hay diversos trabajos relacionando factores geográficos de la biodiversidad de escarabeidos coprófagos (Lobo & Martín- Piera, 2002; Hortal *et al*., 2001, 2003; Lobo *et al*. 2002, 2004; Verdú & Galante, 2002; Martín-Piera & Lobo, 2003; Cabrero-Sañudo & Lobo, 2003), dichos trabajos son perfectamente extrapolables al género de los Aphodius. En el artículo publicado por (J. Hortal, J. M. Lobo y L. del Rey, 2006) exponen que los efectos más importantes parecen provenir de las variables climáticas y de los factores de índole histórico-geográfica. Las variaciones climáticas tienen un importante efecto sobre la riqueza de los escarabeidos en general, lo que evidencia un probable control climático sobre los ensamblajes de Scarabaeoidea, resultado de sus limitaciones y requerimientos ambientales. Dada la importancia que las variables climáticas tienen en la biodiversidad de esta familia y de género que atribuye este trabajo se seleccionan como variables explicativas una serie de datos bioclimáticos extraídos de la web worldclim-Global climate data, en esta web se almacenan datos climáticos a nivel mundial en formato grd con una resolución de alrededor de $1Km^2$.

Esta web es un repositorio con variables climáticas, opensource y libre, que ha permitido un desarrollo exponencial de los trabajos sobre biogeografía, macroecología y cambio climático en los últimos 10 años. Worldclim permite la descarga de 19 variables climáticas, a diferentes resoluciones espaciales y en diferentes formatos raster (que son formatos SIG, es decir, una matriz de datos georeferenciada). Además de la interpolación espacial para el presente, con datos de estaciones meteorológicas de entre 1950 a 2000 (Hijmans et al. 2005), Worldclim también dispone de capas SIG con información sobre las mismas variables climáticas en el pasado y en el futuro (para el 2070). Para generar estas capas combinan información sobre cambio climático proveniente de modelos de circulación global (AOGCMs (atmosphere-ocean coupled general circulation models), modelos físicos sobre dinámica climática) y su capa para el presente (proveniente de una interpolación).

OBJETIVOS

Con la información obtenida tanto a nivel geográfico como climático los objetivos marcados para este trabajo pretende ampliar el conocimiento biogeográfico sobre los Aphodius de España mediante:

1- Obtener el o los mejores modelos que estimen la probabilidad de que la especie se encuentra presente en una localización condicionada a variables ambientales.

2- la creación de mapas que representen la distribución observada y predicha de todas aquellas especies con información georreferenciada disponible;

3- cuáles han podido ser los principales factores bioclimáticos causales que han propiciado la actual distribución de la diversidad biológica de este género y,

4- predecir la distribución de los Aphodius en España para el año 2070.

METODOLOGÍA

Zona de estudio

El trabajo se centra en la región de España incluyendo las Islas Baleares, y excluyendo las Islas Canarias. Su extensión es de 504.645 km², es el cuarto país más extenso del continente europeo, y tiene una altitud media de 650 metros sobre el nivel del mar y es también uno de los países más montañosos.

España presenta una alta diversidad animal y vegetal albergando la mayor de toda la Unión Europea, con un gran número de especies endémicas, siendo especialmente vulnerable al cambio climático. Se encuentran más del 80% del total de especies de plantas vasculares que hay en Europa y más del 50% de la especies de animales (LA BIODIVERSIDAD EN ESPAÑA, Pilar Álvarez-Uría Tejero y Cristina Zamorano Chico, 2000). Esta elevada biodiversidad queda reflejada en la gran extensión de territorio que forma parte de la Red Natura, que ocupa en la actualidad el 25% de la superficie de España.

Por la parte que ocupa a este trabajo, la entomofauna en general está teniendo una gran repercusión como herramienta para la estimación de la biodiversidad, algunas comunidades autónomas están utilizando este grupo como estimador, por lo que da a conocer el grado de su importancia. Trabajos como este pueden ayudar a la zonificación

y predicción de distribuciones potenciales de especies y a la protección de sus localizaciones presenciales o futuras.

Variables ambientales

Como se ha mencionado en la introducción las variables explicativas seleccionadas para el trabajo se corresponden a 19 variables bioclimáticas extraídas de worldclim.

Tabla 1. Variables bioclimáticas

COD	Variable
Bio01	Temperatura media anual
Bio02	Rango de temperatura diurno medio (Temp. Máxima – Temp. Mínima)
Bio03	Isotermalidad (Bio2 / Bio7) (* 100)
Bio04	Estacionalidad de temperatura (desviación estándar * 100)
Bio05	Temperatura máxima del mes más caliente
Bio06	Temperatura mínima del mes más frío
Bio07	Rango de temperatura anual (Bio5 – Bio6)
Bio08	Temperatura media del trimestre más húmedo
Bio09	Temperatura media del trimestre más seco
Bio10	Temperatura media del trimestre >más caliente
Bio11	Temperatura media del trimestre más frío
Bio12	Precipitación total anual
Bio13	Precipitación del mes más húmedo
Bio14	Precipitación del mes más seco
Bio15	Estacionalidad de la precipitación (coeficiente de variación)
Bio16	Precipitación del trimestre más húmedo
Bio17	Precipitación del trimestre más seco
Bio18	Precipitación del trimestre más caliente
Bio19	Precipitación del trimestre más frío

Fuente: www.worldclimb.com

Selección de variables

La pre-selección de variables candidatas para el modelado de la distribución de las especies se debe hacer antes del modelado (Elith *et al*., 2011), por lo que es necesario realizar un análisis previo para eliminar la autocorrelación entre variables, y por lo tanto, evitar la inestabilidad en los modelos.

Los resultados que se presentan en este apartado no entran en mucho detalle con respecto a la modelización de las diferentes técnicas ya que esto podría ser en sí solo otro trabajo de master.

Para todos los modelos de distribución, que serán definidos más adelante, se estimara la distribución potencial con todas las variables bioclimáticas sin hacer una selección

previa. Estos resultados se compararán con los obtenidos con los mismos modelos utilizando las variables preseleccionadas en este apartado.

Para elegir las variables explicativas idóneas se comparan varias técnicas o varios modelos con validación cruzada, y se elige como óptimo aquel que menor *tasa de fallos* presenta y/o *AUC* sea mayor.

De entre los diferentes métodos para seleccionar variables en función de una variable binaria, se ha decidido utilizar estos tres métodos: Regresión logística, Random Forest y Redes neuronales. Cuando se utiliza regresión logística es conveniente hacer árboles de decisión con el fin de comparar resultados, las redes neuronales artificiales también se han utilizado para pre-seleccionar las variables (p.eg. Gustavo Cruz-Cárdenas; José Luis Villaseñor; Lauro López-Mata; Enrique Martínez-Meyer; Enrique Ortiz)

A continuación se muestran los resultados obtenidos de las diferentes técnicas y diferentes modelos obtenidos, todos los pasos con mayor detalle se pueden ver en el anexo.

Tabla 2. Modelos estimados para la selección de variables

Modelo	Observaciones	Tasa de fallos	AUC	Nº variables seleccionadas
Regresión logística	Sin partición de datos	0.09264	0.8867	12
Regresión logística	Training 80%, test 20%	0.09407	0.8870	11
Regresión logística	Training 80%, test 20%	0.09264	0.88721	12
Regresión logística	Training 80%, test 20%	0.08793	0.88522	13
Regresión logística	Training 80%, test 20%	0.08725	0.89365	14
Regresión logística	Training 80%, test 20%	0.08634	0.88661	14
Regresión logística	Training 80%, test 20%	0.08723	0.88645	14
Regresión logística	Training 80%, test 20%	0.08734	0.88665	13
Regresión logística	Training 80%, test 20%	0.08992	0.88689	12
Árbol de decisión	- Nº observaciones por hoja 15. - BEST. - Criterio: probchisq	0.91786		6
Árbol de decisión	- Nº observaciones por hoja 25. - BEST. - Criterio: probchisq	0.08689		6
Árbol de decisión	- Nº observaciones por hoja 10. - LARGEST. - Criterio: probchisq	0.08689		10

Árbol de decisión	- Nº observaciones por hoja 10. - nleaves=30. - Criterio: Entropía	0.08070	12
Red neuronal	- 9 nodos - QUANEW - Tangente hiperbólica	0.06227	12

Fuente: Elaboración propia

Dado los resultados, el modelo con menor tasa de fallos es la red neuronal, con un valor de 0.0622, por lo tanto los modelos para obtener la distribución de los Aphodius en España utilizarán las 12 variables asociadas a esta red, bio1_15 *(Temperatura media anual),* bio2_15 *(Rango de temperatura diurno medio (Temp. Máxima – Temp. Mínima)),* bio3_15 *(Isotermalidad),* bio5_15 *(Temperatura máxima del mes más caliente),* bio8_15 *(Temperatura media del trimestre más húmedo),* bio9_15 *(Temperatura media del trimestre más seco),* bio10_15 *(Temperatura media del trimestre >más caliente),* bio11_15 *(Temperatura media del trimestre más frío),* bio12_15 *(Precipitación total anual),* bio14_15 *(Precipitación del mes más seco)* bio16_15 *(Precipitación del trimestre más húmedo),* bio18_15 (Precipitación del trimestre más caliente).

Estimación de modelos

Para la estimación de la distribución potencial se utilizaran varios algoritmos que se pueden clasificar como “profile”, “regresión” y “machine learning” (Robert J. Hijmans and Jane Elith, 2016).

- Los métodos “**profile**” únicamente tienen en cuenta datos de presencia, no necesitan datos de ausencias o “background”.
    - Domain
    - Bioclim
    - Mahalanobis
- Los métodos de **regresión** y de **machine learning** utilizan tanto la presencia y ausencia de datos o de “background”.
    - Modelos de Regresión
        - GLM
        - GAM

- Machine Learning
    - Gradient Boosting
    - Random Forest
    - SVM

Una distinción entre los modelos de regresión y machine learning es la forma de clasificar los modelos.

Además de estos modelos se va a realizar ensamblado, para mejorar la estimación obtenida y comprobar la idoneidad de cada uno de ellos.

- Una clase completamente diferente a los modelos anteriores consiste en los modelos que sólo, o principalmente, utilizan la ubicación geográfica de los Aphodius, y no se basan en los valores de las variables predictoras (variables bioclimáticas) en estos lugares, a estos modelos se les conocen como "**modelos geográficos**".
    - Distancia geográfica
    - Convex hulls
    - Circles
    - Presencia/ausencia
- Otra metodología que no es muy utilizada en modelos de distribución y que se va a ver en este trabajo son las redes bayesianas.
    - Naïves Bayes

SITUACIÓN DE INICIO

Como se ha dicho en la introducción, el estudio se centra en los datos de España que existen en la base de datos GBIF sobre las especies del género Aphodius.

Tabla 3. Especies españolas incluidas en la bbdd GBIF

Especie	Género	Familia	Registros
Aphodius alpinus	Aphodius	Scarabaeoidea	14
Aphodius ater	Aphodius	Scarabaeoidea	6
Aphodius brevis	Aphodius	Scarabaeoidea	1
Aphodius coniugatus	Aphodius	Scarabaeoidea	209
Aphodius elevatus	Aphodius	Scarabaeoidea	6
Aphodius fimetarius	Aphodius	Scarabaeoidea	1205

Aphodius foetidus	Aphodius	Scarabaeoidea	831
Aphodius lividus	Aphodius	Scarabaeoidea	5
Aphodius marini	Aphodius	Scarabaeoidea	7
Aphodius meyeri	Aphodius	Scarabaeoidea	1
Aphodius obscurus	Aphodius	Scarabaeoidea	4
Aphodius satellitius	Aphodius	Scarabaeoidea	6
Aphodius scrutator	Aphodius	Scarabaeoidea	7

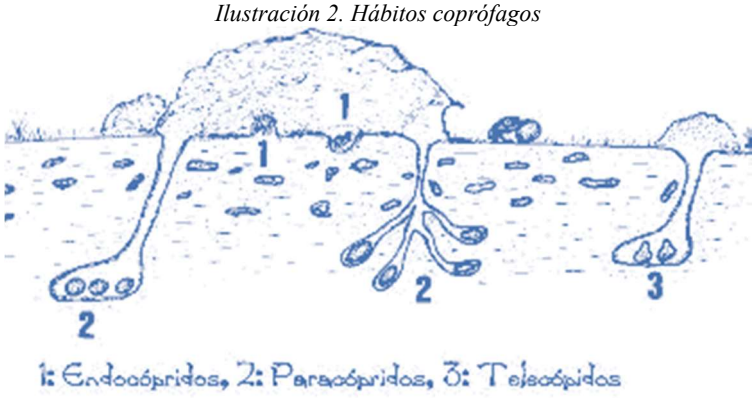
Fuente: Elaboración propia con datos de GBIF

Hasta la fecha hay descritas 102 especies de Aphodius en la península ibérica, 197 si se tienen en cuentas las distintas variedades dentro de una misma especie. Tan sólo dos especies son endémicas para España.

Las especies de la base de datos representan el 12,7% de las especies españolas, no encontrándose ningún endemismo.

Como la mayoría de Aphodiinae los Aphodius son de hábitos coprófagos y generalmente asociados a excrementos de vertebrados y mostrando un comportamiento habitualmente endocóprido, el cual se caracteriza por ser este lugar su alimento y su habitáculo para la reproducción (Halffter & Edmonds 1982, Cambefort 1991).

Ilustración 2. Hábitos coprófagos



Fuente: http://www.oocities.org

Según las claves descritas por Luis Baguena en 1967 son especies con pronoto sin impresiones, a lo sumo con una ligera fosita antemedia en algunos machos, nunca con puntuación foveolar, élitros normalmente estriados, sin quillas longitudinales, cara externa de las tibias posteriores con dos quillas transversales enteras.

A continuación se describe de manera breve datos morfológicos de cada una de las especies incluidas en la base de datos[1].

[1] Fotografías extraídas de la web http://www.colpolon.biol.uni.wroc.pl/

Aphodius alpinus (Scopoli, 1763) Ancho y corto, generalmente negro, con el ápic y los lados de los élitros ligeramente castaños (forma típica), o con ellos enteramente rojizos. Epistoma apenas sinuado por delante y con puntuación superficial gruesa, la del pronoto doble y bastante densa, más la fina que la gruesa, estrias poco pronunciadas.

Aphodius ater (De Geer, 1774) Especie convexa, enteramente negro, los élitros mates (forma típica) o más o menos brillantes (variedades), interestrias con puntuación fina pero manifiesta, la parasutural no más hundida en la región preapical que delante ni detrás.

Aphodius brevis (Erichson, 1848) Especie muy pequeña, de 4 a 5 mm. El epistoma inerme por delante, los lados de su escotadura redondeados, el pronoto sin borde anterior, la puntuación de las estrías muy gruesa y nada separada.

Aphodius coniugatus (Panzer, 1795) Cabeza y pronoto negros, éste con los ángulos anteriores rojizos, los élitros de color pajizo cruzados por una faja irregular que suele extenderse hasta el ápice por la interestria negruzca que suele extenderse hasta el ápice por la interestria parasutural, aunque hay veces que se interrumpe.

Aphodius elevatus (Olivier, 1789): Se trata de una especie en oligotróficos prefieren los pastos expuestos. Es una especie coprófagos, que se encuentran principalmente en el estiércol de vaca (siempre extremadamente secado), sino también en ovejas, caballos y excremento humano o bajo grueso de pellets de Lagomorfos. Los

adultos están activos a finales de primavera, verano y otoño, pero sobre todo en otoño. Se siente atraída por la luz. Se encuentra difundido desde el nivel del mar de hasta 2.000 m snm (Dellacasa y Dellacasa 2006).

Aphodius fimetarius (Linnaeus, 1758) Pronoto negro con los ángulos anteriores rojizos, los élitros rojizos sin manchas y con la cuarta interestria corta, y el abdomen negro. Los machos a diferencia de las hembras presentan una ligera foseta en el pronoto.

Aphodius foetidus (Herbst, 1783) Pronoto negro ligeramente punteado, los élitros de color pardo con interestrias densamente punteadas y la parasatural de color oscuro sin manchas, patas y tarsos de color pardo.

Aphodius lividus (Olivier, 1798) Muy brillante, de coloracion algo variable, en la forma típica son castaños el disco de la cabeza y el del pronoto, la interestria parasatural y una mancha vaga discal en cada élitro, el resto de éstos pajizo.

Aphodius marini (Baguena, 1930) Negro, brillante, el epistoma muy convexo y con puntuación muy densa y fuerte, el borde anterior cuadridentado la escotadura media mucho más ancha que las laterales, la frente también densamente punteada, el reborde basal del pronoto indistinto, la puntuaciñon de éste moderadamente fina y bastante densa, más hacia la periferia, el dendícuro humeral de los élitros fuerte, los puntos de las estrías separados y no muy fuertes.

Aphodius meyeri (Heer, 1847) Especie de tamaño pequeño de entre 3,5 y 4 mm. Especie de distribución paleártica encontrándose con mayor frecuencia en el centro de Europa. Especie poco conocida.

Aphodius obscurus (Fabricius, 1792) Especie grande de entre 7 y 7,5 mm, de color negro mate y lampiño (forma típica). Epistoma con puntuación densa y fuerte, la del pronoto, doble, muy densa y uniformemente repartida en todo el disco, los puntos finos y escasos, estrias normales, las interestrias planas con la puntuación fina y superficial.

Aphodius satellitius (Herbst, 1789) Cabeza y pronoto negros, los élitros rojizos, generalmente con una mancha oscura y alargada común sutural (forma típica) y algunas veces sin ellas (variedades). Epistoma apenas sinuado por delante, casi semicircular, la puntuación del pronoto fina, poco densa, uniformemente en el disco y mezclada con puntos más gruesos hacia los lados, estriación moderadamente fuerte, más ancha y profunda hacia el ápice, las interestrias planas en el disco y convexas hacia tras con puntuaciñin fina y poco densa en el disco y más fuerte hacia los lados y atrás.

Aphodius scrutator (Herbst, 1789) Bicolor, el pronoto negro con los lados anchamente rojizos (forma típica), élitros, abdomen y tarsos rojizos.

DISTRIBUCIÓN Y PREPARACIÓN DE LOS DATOS

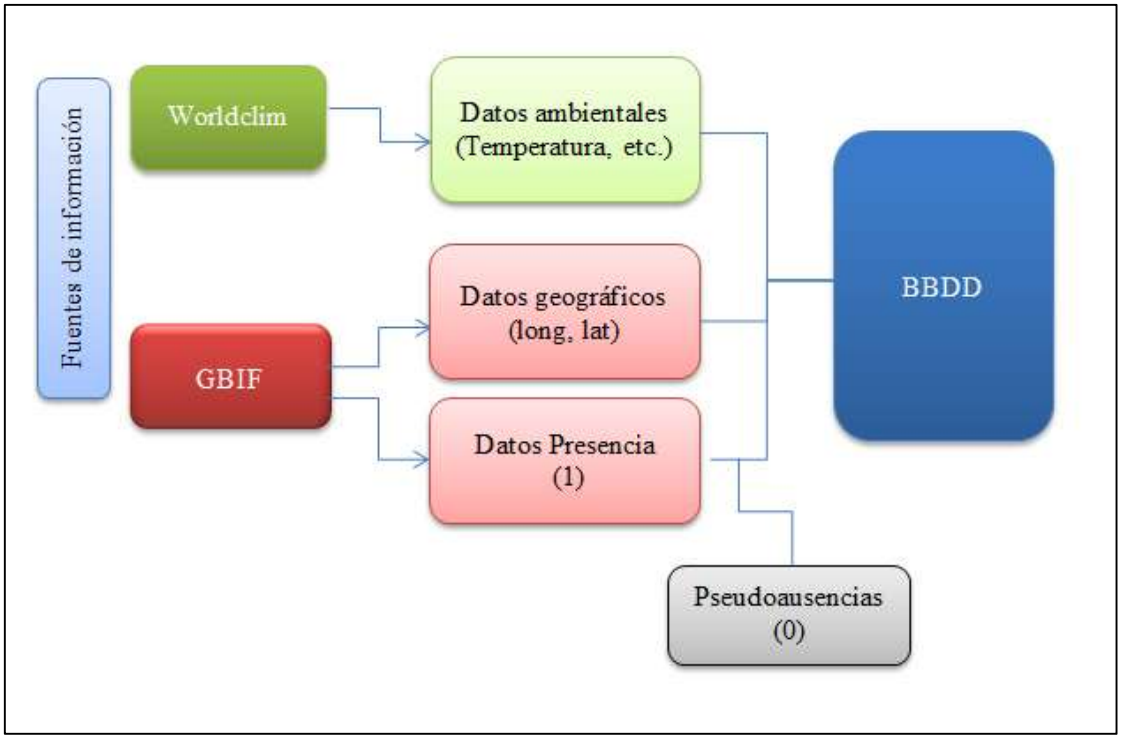
Construcción de la Base de datos

Antes de comenzar a explotar la información, como en todo trabajo o como en la mayoría hay que diseñar una base de datos de la que cuelgue todo la información necesaria.

La base de datos que se ha utilizado contiene información de las distintas fuentes consultadas, Worldclim y GBIF, de estas fuentes como se ha comentado en la introducción se ha descargado la información necesaria para la determinación de la distribución de los Aphodius en España, información tanto ambiental como información

geográfica entendiendo esta como la localización datada de la especie e información presencial, esta última es información geográfica sólo de datos de presencia. En el siguiente diagrama se puede ver representado el flujo de trabajo hasta obtener la base de datos definitiva. Para tener una imagen de la base datos final, en la tabla 5 se muestra un ejemplo de la estructura 'física' de la base de datos.

Tabla 4. Diagrama de fflujo de la construcción de la BBDD



Fuente: Elaboración propia

Tabla 5.Estructura de la BBDD

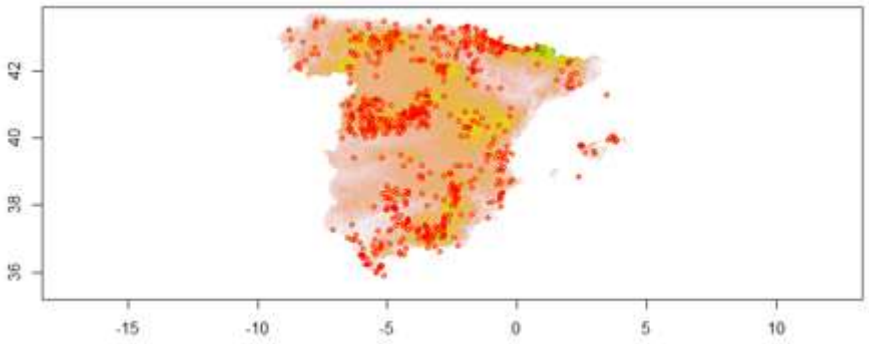
lon	lat	species	bio4_15	bio7_15	bio1_15	bio2_15	...
-3.1532	37.37121	Aphodius fimetarius	641.7	30.8	14.5	11.6	...
-3.1532	37.37121	Aphodius foetidus	641.7	30.8	14.5	11.6	...
-2.44431	37.02846	Aphodius fimetarius	525.9	24.2	16.8	9	...
-3.13546	37.36177	Aphodius foetidus	642.5	30.8	14.3	11.6	...
-3.13546	37.36177	Aphodius foetidus	642.5	30.8	14.3	11.6	...
-3.13603	37.36844	Aphodius foetidus	645.4	30.9	14.4	11.6	...
-5.77	40.39	Aphodius foetidus	631.6	29.5	11.9	11.2	...
-5.78	40.57	Aphodius foetidus	627.1	29.3	11.4	11.2	...
-5.3	40.58	Aphodius fimetarius	618.5	29.4	9.8	11.2	...
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Fuente: Elaboración propia

Representación de los datos (Aphodius)

Si representamos las observaciones sobre un mapa obtenemos una perspectiva de la distribución de todas las especies del estudio. Se observa como hay puntos incoherentes, posiblemente ocasionados por el sesgo en la extrapolación de los datos.

Gráfico 1. Representación de Aphodius sobre mapa de España

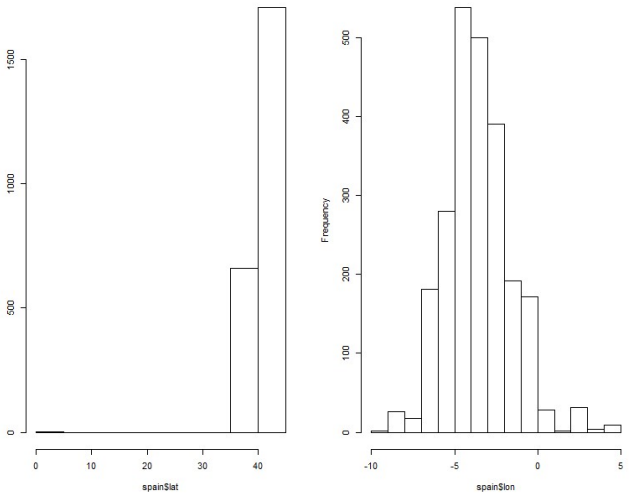


Fuente: Elaboración propia con datos de la base de datos GBIF

Concentración de especies

La mayoría de Aphodius de la base de datos han sido datados en la latitud 40º mayoritariamente y en torno a longitud -5º y 0º, siendo mayoritaria la ocurrencia en la longitud 6º, por lo que (lat 40º long 6º) corresponde a especies del entorno de zonas como Plasencia, Cáceres o Salamanca, de hecho en el gráfico anterior se puede ver como hay una mayor concentración de puntos sobre esta zona.

Gráfico 2. Histograma de latitud y longitud



Fuente: Elaboración propia con datos de la base de datos GBIF

Georeferenciar

Antes de continuar con el análisis, una de las partes de la comprensión de la base de datos o del tratamiento de los datos, es dar una georeferencia a observaciones que tengan una localidad descrita, pero no tengan coordenadas. En este apartado se georeferencian estos casos.

Se ha comprobado que de las 3256 observaciones 841 no tienen coordenadas. Teniendo en cuenta esto y si eliminamos las observaciones que comparten la misma georeferencia la base de datos final cuenta con 2808 observaciones.

Background o Pseudoausencias

Como se ha dicho anteriormente, algunos modelos no precisan de la necesidad de contar con datos de ausencia de las especies, sin embargo otros modelos si hacen de esto una necesidad para poder modelizar una distribución potencial.

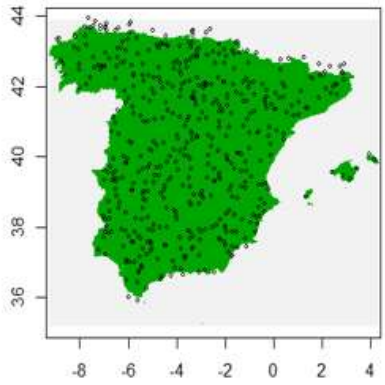
La base de datos con la que se cuenta para el trabajo sólo tiene datos presenciales, por lo que es necesario estimar datos de ausencia, en este sentido es necesario aleatorizar datos background.

Los datos background (e.g. Phillips *et al.* 2009) no pretenden predecir la localización de ausencia, más bien pretenden caracterizar ambientes en la zona de estudio. En este sentido, background es el mismo, independientemente del lugar donde se haya encontrado la especie. Los datos background establecen el dominio del medio ambiente del estudio, mientras que los datos de presencia deben establecer en qué condiciones una especie es más probable que se presente en torno a la media. Un concepto estrechamente relacionado pero diferente, es el de "pseudo-ausencias", también se utiliza para la generación de la clase no presencia de modelos logísticos. En este caso, a veces se trata de predecir dónde podrían producirse ausencias, se puede utilizar toda la zona de estudio excepto en los lugares donde haya presencia, o pueden utilizarse en lugares donde la presencia de la especie sea poco probable.

En este trabajo se utilizará el concepto de background ya que requiere menos suposiciones, además tiene algunos métodos estadísticos coherentes para hacer frente al "solapamiento" entre los puntos de presencia y de fondo (e.g. Phillips y Elith, 2011).

Para la estimación de datos background se utilizará del paquete dismo de R una función para tomar muestras aleatorias (datos background) de una zona de estudio, en este sentido se generaran 500 puntos repartidos por toda España.

Gráfico 3. Datos background estimados con el paquete Dismo de R



Fuente: Elaboración propia

Una vez obtenidos los background se generó una matriz de datos que contiene 500 filas de ausencias más las filas que corresponden al número de registros de presencia. El número de columnas es de 20 que incluye las 19 variables bioclimáticas más una de presencia (1) o ausencia (0). El utilizar base de datos con información de presencia y ausencias es recomendable según Elith et al. (2011), ya que se proporciona mejor información sobre la prevalencia (es decir, las zonas donde la especie está presente) que solo con los datos de presencia.

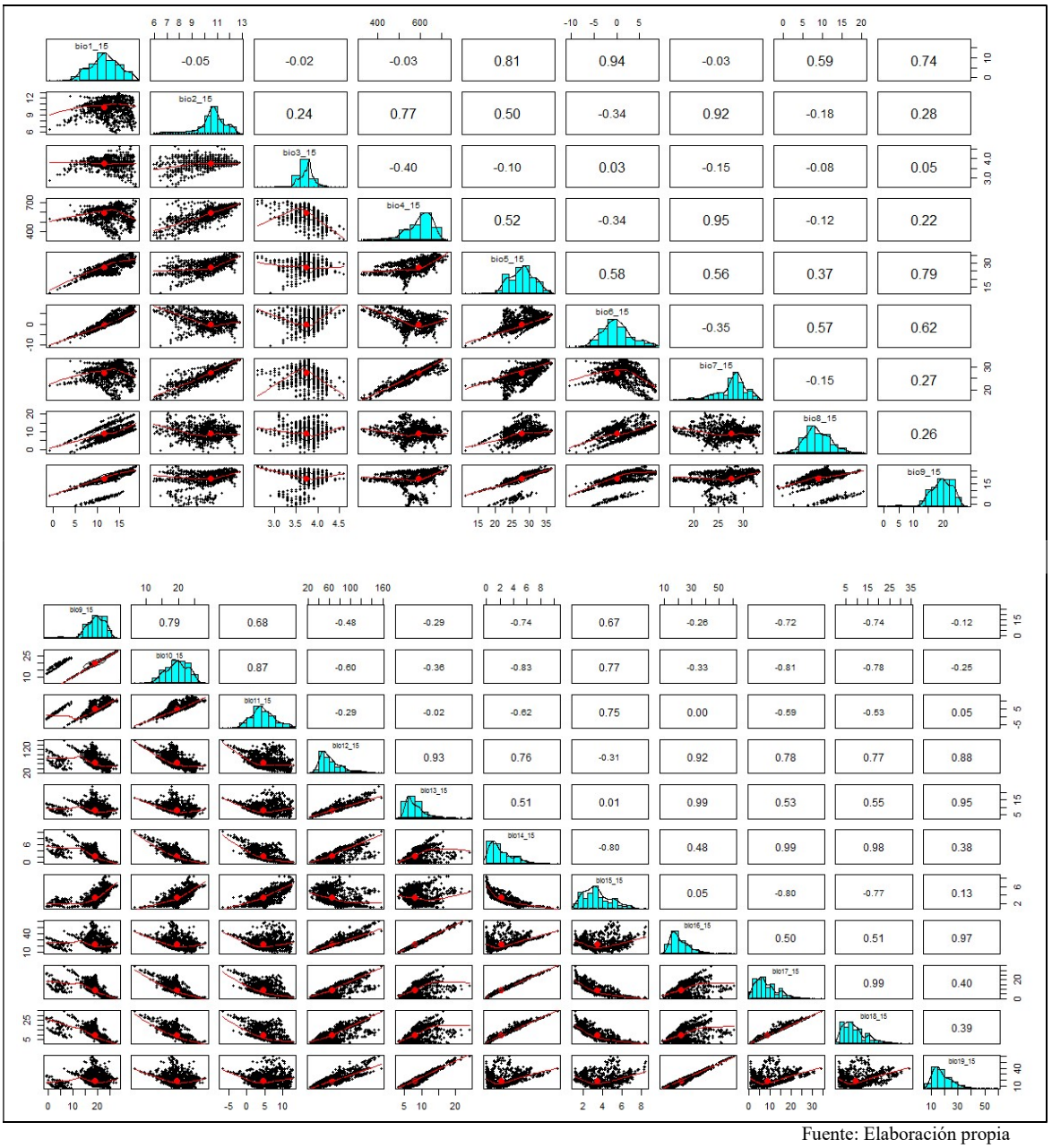
CORRELACIÓN ENTRE LA PRESENCIA/AUSENCIA Y LAS VARIABLES BIOCLIMÁTICAS

Al analizar la correlación entre los datos de presencia y las pseudoausencias estimadas se aprecia una correlación negativa muy significativa (p_valor < 0,01) con las variables Bio1_15, Bio3_15, Bio5_15, Bio6_15, Bio8_15, Bio9_15, Bio10_15, Bio11_15, Bio15_15, Bio16_15, Bio19_15, y una correlación positiva con Bio4_15 y Bio7_15. La tabla con todas las correlaciones se puede ver en el Anexo del documento.

En el siguiente gráfico se puede ver la correlación entre las variables bioclimáticas. Las correlaciones significativas son aquellas que tienen de mayor tamaño el dato enmarcado. Con los histogramas podemos hacernos una idea de la distribución de cada

variable, por ejemplo, variables como Bio1_15, Bio6_15 o Bio8_15 tienen una distribución centrada que en principio parece que podrían distribuirse normalmente.

Gráfico 4. Correlaciones entre variables bioclimáticas

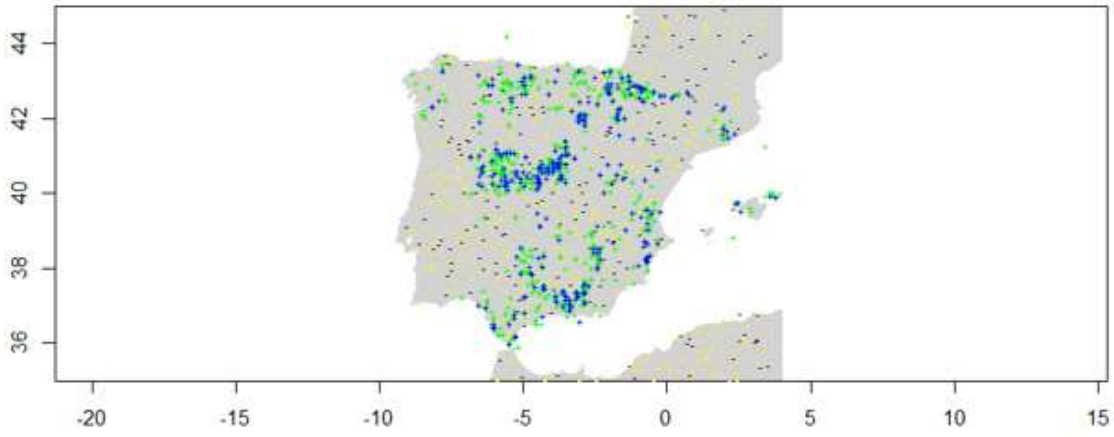


Fuente: Elaboración propia

DESARROLLO DEL TRABAJO Y PRINCIPALES RESULTADOS

Para obtener unos resultados validados, los modelos se estiman con datos training y datos test, utilinzado los datos training para la estimación de las predicciones y los daos test para comprobar la eficiencia del modelo. Para esto se ha decidido separar la base de datos en cinco grupos aleatoriamente seleccionados. Los datos test son aquellos incluidos en el grupo uno y el resto forman los datos training.

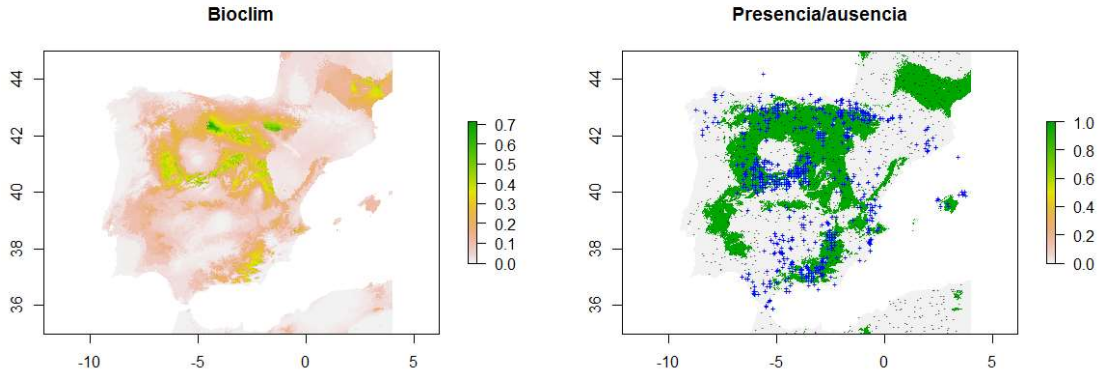
Gráfico 5. Representación de puntos de presencia y background



Fuente: Elaboración propia

Con el fin de realizar comparaciones, se estimaran los modelos definidos anteriormente con todas las *variables bioclimáticas* y con las *variables preseleccionadas* de esta manera podremos comprobar si la selección de variables es idónea o por lo contrario es innnecesaria.

El resultado esperado en cada algoritmo son dos mapas, en uno de ellos se dibujan las probabilidades estimadas de presencia/ausencia en función de las variables consideradas, a esta representación se la conoce como la proyección de las predicciones en el espacio. Y en otro mapa se grafica las probabilidad estimadas de presencia/ausencia teniendo en cuenta un umbral o punto de corte en el que la suma de sensitividad y especificidad sea lo mayor posible, sobre este último mapa se representan los puntos reales de presencia y los background para ver el comportamiento de los datos, ejemplo:



En el mapa de la izquierda se aprecia la presencia estimada[2] por el algoritmo BIOCLIM y en el mapa de la derecha (distribución potencial) se aprecia la estimación de presencia de Aphodius según el umbral que hace mayor la sensitividad y especificidad, sobre este mapa como se ha dicho anteriormente se representan los datos reales para estudiar posibles comportamientos de las especies.

Modelos Profile

DOMAIN (Carpenter et al., 1993)

Este algoritmo ha sido y está siendo muy utilizado para el modelado de la distribución de especies. No suele funcionar muy bien en un modelo de comparación (Elith et al., 2006) y muy mal cuando se utiliza como variables los efectos del cambio climático (Hijmans y Graham, 2006). Se basa en la distancia Gower entre las variables ambientales de cualquier zona y las que coinciden con las zonas donde está presente la especie (training sites). Esta distancia se calcula como la diferencia absoluta entre los valores de la variable climática de cualquier zona dividido por el rango de la variable a través de todos los puntos de presencia conocidos (es decir, la distancia es escalado por la gama de observaciones). Por cada variable la distancia mínima entre un sitio y otro cualquiera de los training sites. La distancia Gower es entonces la media de estas distancias sobre todas las variables ambientales. El algoritmo asigna a una zona la distancia del punto de ocurrencia más cercano. Para integrar más variables ambientales, se utiliza la distancia de cualquiera de las variables. A esta distancia se resta uno, y los

[2] Escala de color: de menor probalidad de presencia equivale a tonos claros; mayor probabilidad de presencia tonos oscuros.

valores por debajo de cero son truncados para que las puntuaciones queden entre 0 (puntuación baja) y 1 (puntuación alta).

Resultados con las variables preseleccionadas

Matriz de confusión

La matriz de confusión que se obtiene tras la clasificación que realiza el algoritmo muestra una alta clasificación para las ausencias, background en este caso.

Tabla 6. Matriz de confusión DOMAIN con las variables preseleccionadas

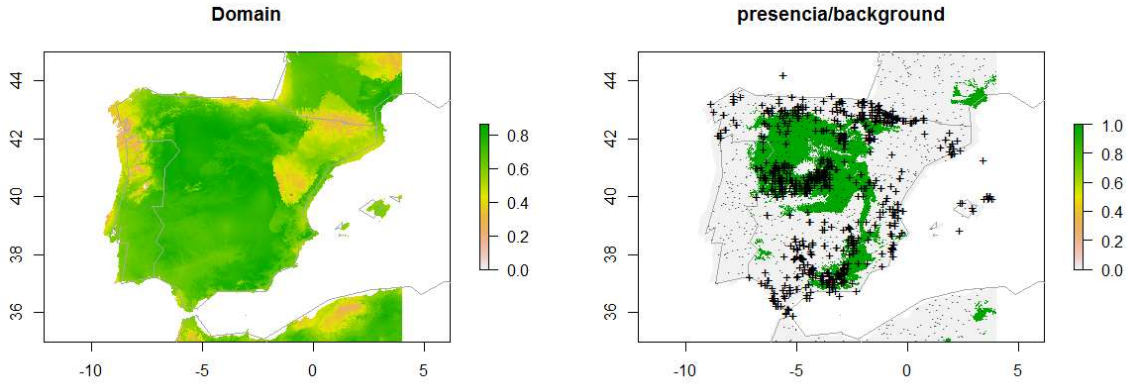
	Presencia	Ausencia
Presencia	2	1
Ausencia	18	182

Fuente: Elaboración propia con datos de GBIF

Evaluación del modelo

El AUC obtenido con este algoritmo no es muy alto, 0.685.

Representación gráfica de la distribución estimada



Resultados con todas las variables bioclimáticas

En esta ocasión como se puede ver el algoritmo clasifica en mayor media los verdaderos positivos.

Tabla 7. Matriz de confusión DOMAIN con todas las variables

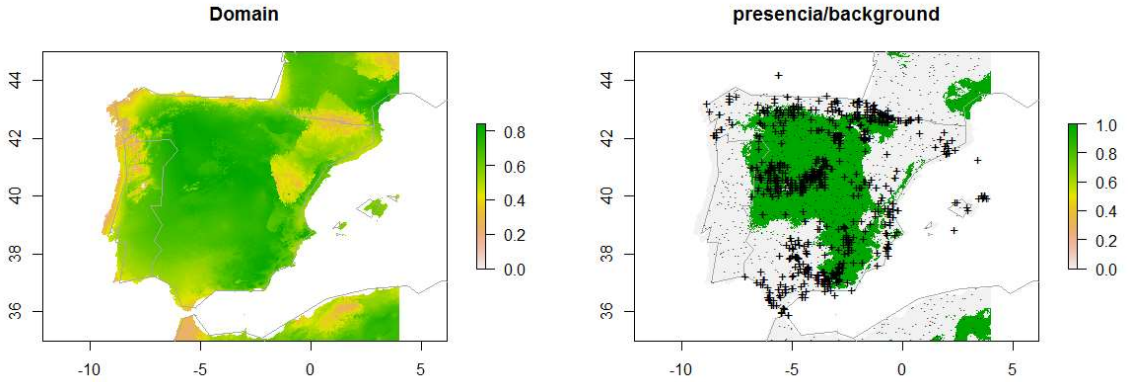
	Presencia	Ausencia
Presencia	281	183
Ausencia	93	107

Fuente: Elaboración propia con datos de GBIF

Evaluación del modelo

El AUC obtenido con este algoritmo no es muy alto, 0.653.

Representación gráfica de la distribución estimada



Comparación de modelos

Se aprecia en los gráficos anteriores como los resultados obtenidos con las variables pre seleccionadas las áreas predichas son más restrictivas que las estimadas con todas las variables bioclimáticas. El AUC obtenido con las variables preseleccionadas es algo mayor que del modelo que incluye todas las variables.

BIOCLIM (Bushby, 1991; Nix, 1986)

BIOCLIM también, denominado "Envoltura bioclimática", estima el Rango climático/topográfico de las zonas de presencia para cada variable, y calcula la distribución potencial de dicha especie en lugares con rangos climáticos y/o topográficos similares utilizando los percentiles de los valores más probables.

Es un algoritmo muy utilizado para el modelado de la distribución de especies, aunque por lo general no realiza muy buenos resultados como algunos otros métodos de modelado (Elith et al., 2006) todavía se utiliza.

Calcula la similitud de una ubicación mediante la comparación de los valores de las variables ambientales de cualquier zona a un percentil de los valores de las variables ambientales que coincide con zonas de presencia de las especies (training sites). Cuanto

más cerca del percentil 50 (la mediana), más óptima es la ubicación. Las colas de la distribución no se distinguen, es decir, el percentil 10 se trata como equivalente al percentil 90. Este valor se le resta uno y se multiplica por dos para que los resultados queden entre 0 y 1. Lo habitual es no encontrar un valor 1, ya que esto querrá decir que una ubicación tiene el valor de la mediana de los datos del training sites para todas las variables consideradas. El valor 0 si es más común que se obtenga, ya que se asigna a todas las celdas del raster que tengan un valor una variable ambiental fuera de la distribución percentil al menos en una de las variables del training sites.

Resultados con las variables preseleccionadas

Matriz de confusión

Tabla 8. Matriz de confusión BIOCLIM con las variables preseleccionadas

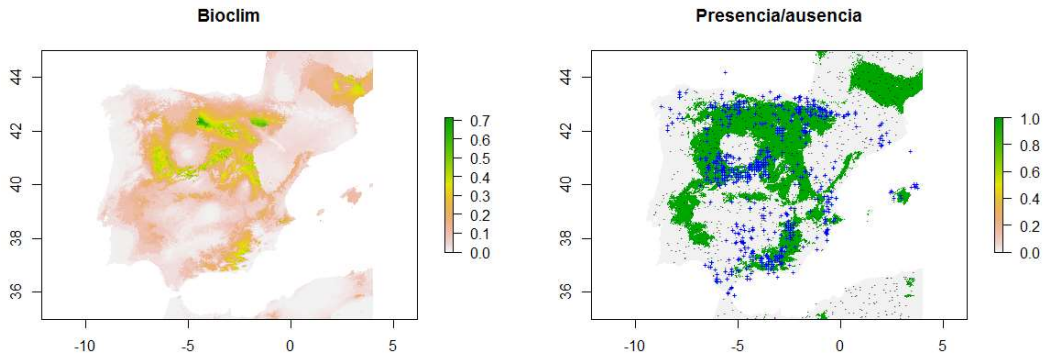
	Presencia	Ausencia
Presencia	1	2
Ausencia	29	171

Fuente: Elaboración propia con datos de GBIF

Evaluación del modelo

El AUC obtenido con este algoritmo es alto, 0.691.

Representación gráfica de la distribución estimada



Resultados con todas las variables bioclimáticas

Matriz de confusión

Tabla 9. Matriz de confusión BIOCLIM con todas las variables

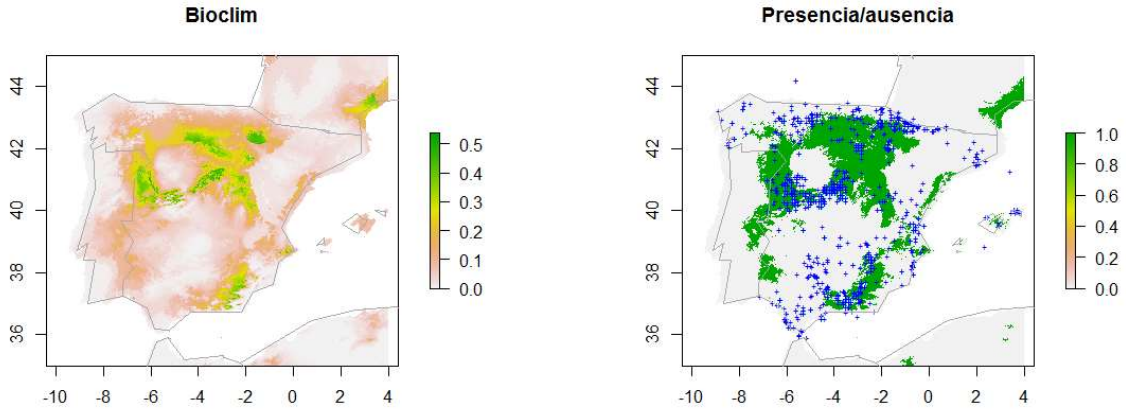
	Presencia	Ausencia
Presencia	243	221
Ausencia	62	138

Fuente: Elaboración propia con datos de GBIF

Evaluación del modelo

El AUC obtenido con este algoritmo es alto, 0.713.

Representación gráfica de la distribución estimada



Comparación de modelos

Con este algoritmo, el AUC como se ha podido comprobar aumenta cuando se incluyen todas las variables bioclimáticas, cierto es que no de una manera desproporcionada, por lo que siguiendo el criterio de parsimonia si hubiera que seleccionar alguno de estos dos enfoques seleccionaríamos el modelo que incluye las variables preseleccionadas.
Gráficamente se aprecia que no hay mucha diferencia ente la distribución predicha de un modelo a otro.

MAHALANOBIS

Este algoritmo se basa en predecir la distribución de las especies utilizando la distancia de Mahalanobis (Mahalanobis, 1936). Toma la distancia de Mahalanobis teniendo en cuenta las correlaciones de las variables bioclimáticas en los datos, y no depende de la escala de las medidas.

Resultados con las variables preseleccionadas

Matriz de confusión

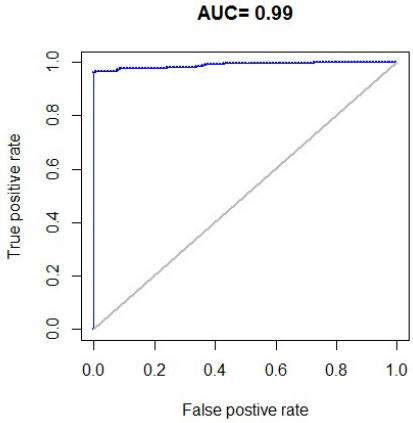
Tabla 10. Matriz de confusión Mahalanobis con las variables preseleccionadas

	Presencia	Ausencia
Presencia	3	0
Ausencia	98	102

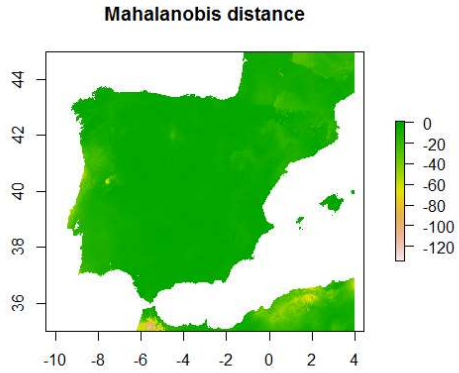
Fuente: Elaboración propia con datos de GBIF

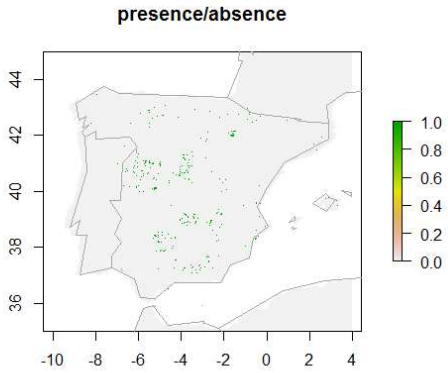
Evaluación del modelo

El AUC obtenido con este algoritmo es muy alto, casi hay un ajuste perfecto.



Representación gráfica de la distribución estimada





Resultados con todas las variables bioclimáticas

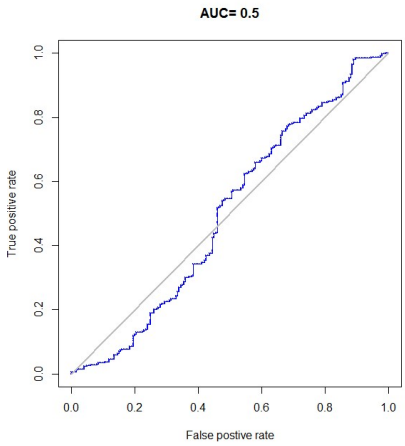
Matriz de confusión

Tabla 11. Matriz de confusión Mahalanobis con todas las variables

	Presencia	Ausencia
Presencia	233	231
Ausencia	100	97

Fuente: Elaboración propia con datos de GBIF

Evaluación del modelo



Comparación de modelos

La comparación entre las predicciones estimadas con este algoritmo la podemos considerar poco robusta. Se aprecia como los resultados del modelo que incluye las variables preseleccionadas ajustan "demasiado" bien la predicción, y por lo contrario como el modelo que incluye todas las variables bioclimáticas no es muy bueno el ajuste, posiblemente causado por la correlación espacial entre variables, como es el caso de las variables bio2_15 y bio7_15. En esta ocasión queda demostrado como la preselección de variables mejora notablemente las predicciones.

Modelos de Regresión

Modelos lineales generalizados (GLM)

Los modelos lineales generalizados (GLM) son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc) y varianzas no constantes (Luis Cayuela, 2010).

Los GLM permiten especificar distintos tipos de distribución de errores, en concreto para este trabajo lo más normal es tomar que los errores se distribuyan según una binomial.

Resultados con las variables preseleccionadas

Se estiman tres modelos, el primero de ellos siguiendo una distribución binomial, el segundo gaussiano y el tercero según una distribución de Poisson. A continuación se muestra el Criterio de información de Akaike para cada uno de los modelos.

Tabla 12. AIC de los modelos GLM para las variables preseleccionadas

Modelo GLM	AIC
Binomial	2643.13
Gauss	2821.04
Poisson	4912.04

Fuente: Elaboración propia

Tomando como referencia el valor AIC se considera el modelo binomial como el adecuado para obtener la distribución potencial de los Aphodius.

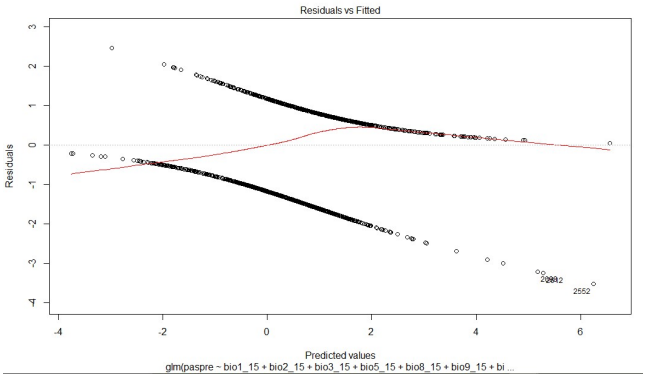
Los coeficientes estimados y la deviance residual del modelo seleccionado son:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5340  -0.8967   0.5267   0.7707   2.4611

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.30313    3.79378   2.189  0.02862 *
bio1_15     -3.71350    0.40076  -9.266  < 2e-16 ***
bio2_15     -0.09043    0.30040  -0.301  0.76339
bio3_15     -0.35200    0.95631  -0.368  0.71281
bio5_15     -0.07224    0.18099  -0.399  0.68977
bio8_15      0.05877    0.02336   2.516  0.01186 *
bio9_15      0.12130    0.01803   6.729 1.71e-11 ***
bio10_15     1.50333    0.31606   4.756 1.97e-06 ***
bio11_15     1.92342    0.22917   8.393  < 2e-16 ***
```

```
bio12_15     0.09963    0.02230   4.468 7.88e-06 ***
bio14_15     0.57282    0.20670   2.771  0.00558 **
bio16_15    -0.23521    0.04871  -4.828 1.38e-06 ***
bio18_15    -0.33237    0.05639  -5.894 3.77e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hay variables no significativas, bio2_15, bio3_15 y bio5_15, sin embargo para lo que nos ocupa en este trabajo vamos a tomar todas las variables dentro del modelo.

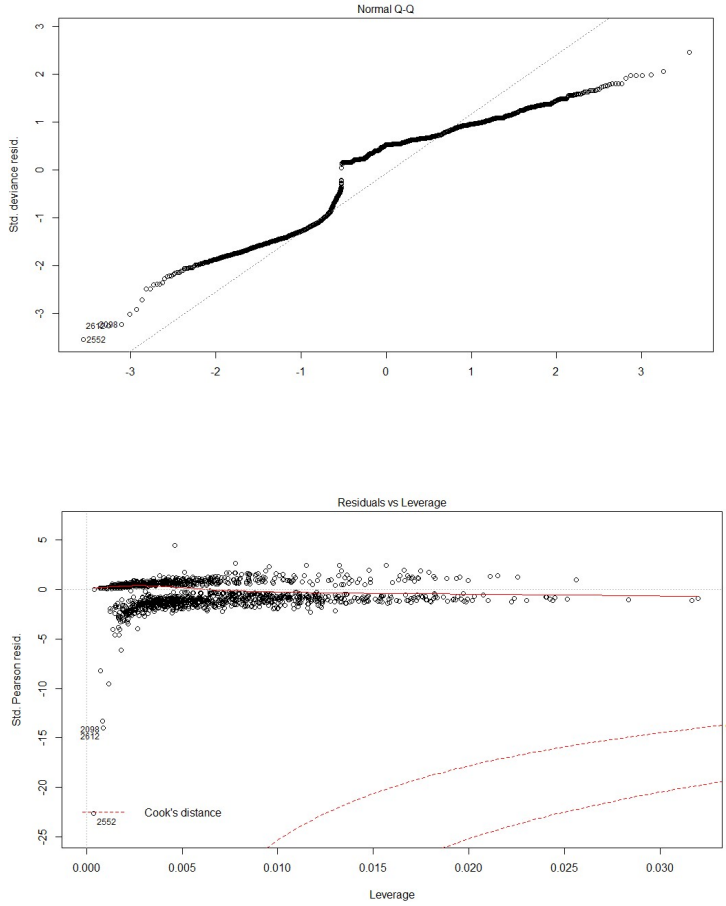
Análisis de los residuos



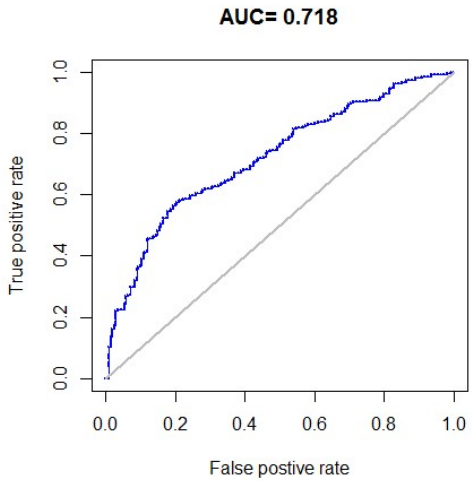
Se aprecia que los residuos no se distribuyen sobre una horizontal, por lo que se demuestra que el modelo es no lineal.

Para evaluar la normalidad de los residuos nos fijarnos en el siguiente gráfico, q-q plot, en el cual podemos ver como los residuos estandarizados no siguen la diagonal adecuadamente para afirmar normalidad, además hay observaciones identificadas por la cola que pueden ser posibles outliers.

Además si nos fijamos en el gráfico de los residuos frente al leverage podemos ver como se encuentran fuera de la línea discontinua que identifica la distancia de Cook, indicando la no linealidad de los residuos y por tanto del modelo, además se puede ver como la observación 2552 tiene una influencia importante en la tendencia de los residuos por lo que se podría tratar de eliminar.

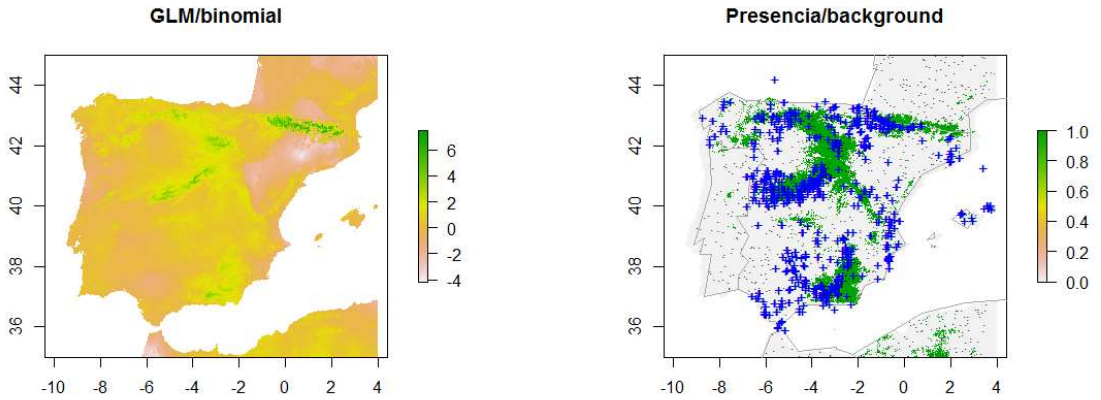


Evaluación del modelo



El AUC obtenido con este modelo no es muy elevado, 0.72 a pesar de este resultado podemos considerar el modelo como adecuado.

Representación gráfica de la distribución estimada



Resultados con todas las variables bioclimáticas

Al igual que en el apartado anterior se realizan tres modelos, binomial, gaussiano y según una distribución poisson, la diferencia con los modelos anteriores es que en esta ocasión se incluyen como variables predictoras todas las variables bioclimáticas.

Para seleccionar el modelo adecuado nos fijamos en el AIC asociado más bajo, en esta ocasión al igual que en el caso anterior se corresponde al modelo según una binomial.

Tabla 13. AIC de los modelos GLM para todas las variables bioclimáticas

Modelo GLM	AIC
Binomial	2416.7
Gaussiano	2575.4
Poisson	4842.7

Fuente: Elaboración propia

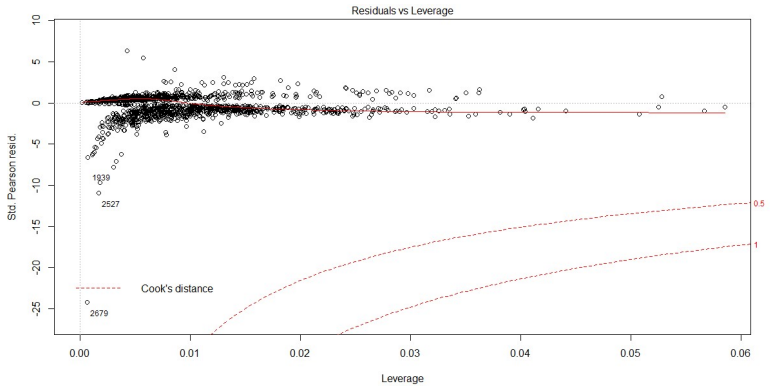
Los coeficientes estimados y la deviance residual del modelo binomial son:

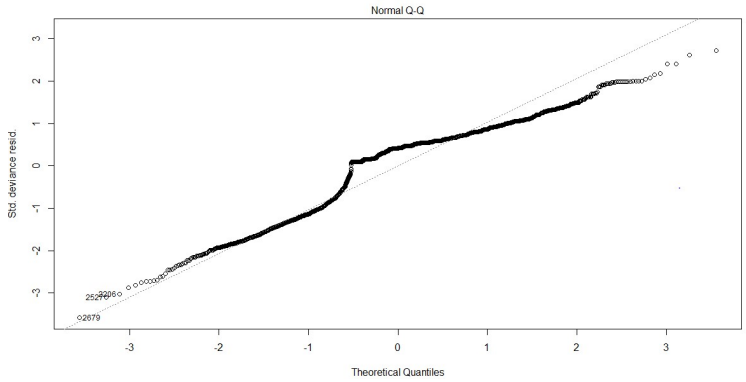
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5713  -0.6883   0.4166   0.6998   2.7199

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 22.63700    5.12249   4.419 9.91e-06 ***
bio1_15     -2.69368    0.58068  -4.639 3.50e-06 ***
bio2_15      2.62099    0.55216   4.747 2.07e-06 ***
bio3_15     -4.48922    1.29531  -3.466 0.000529 ***
bio4_15     -0.01966    0.02066  -0.952 0.341165
bio5_15     -1.13960    0.28334  -4.022 5.77e-05 ***
bio6_15      2.62539    0.32925   7.974 1.54e-15 ***
bio8_15     -0.03518    0.02978  -1.181 0.237464
bio9_15      0.22487    0.02726   8.249  < 2e-16 ***
bio10_15     2.21248    0.99632   2.221 0.026375 *
bio11_15    -1.40771    0.70891  -1.986 0.047062 *
bio12_15     0.23271    0.03701   6.288 3.21e-10 ***
bio13_15     1.19973    0.16392   7.319 2.50e-13 ***
bio14_15     2.87180    0.33873   8.478  < 2e-16 ***
bio15_15     0.44877    0.18402   2.439 0.014741 *
bio16_15    -0.60379    0.09152  -6.598 4.18e-11 ***
```

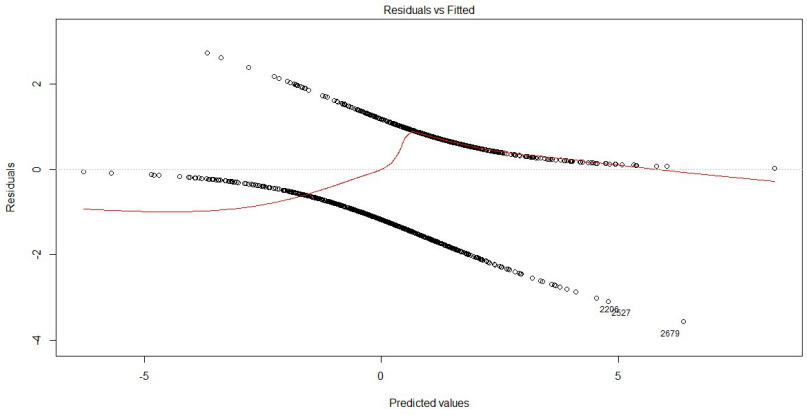
```
bio17_15    -0.95934    0.13346  -7.188 6.55e-13 ***
bio18_15    -0.26496    0.09200  -2.880 0.003977 **
bio19_15    -0.35781    0.06423  -5.571 2.53e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Se aprecian algunas variables no significativas estadísticamente hablando, sin embargo no se descaran del análisis que nos ocupa en este trabajo.

Análisis de los residuos







Analizando lo gráficos anteriores, se aprecia que los residuos no se distribuyen sobre una horizontal, por lo que se demuestra que el modelo es no lineal.

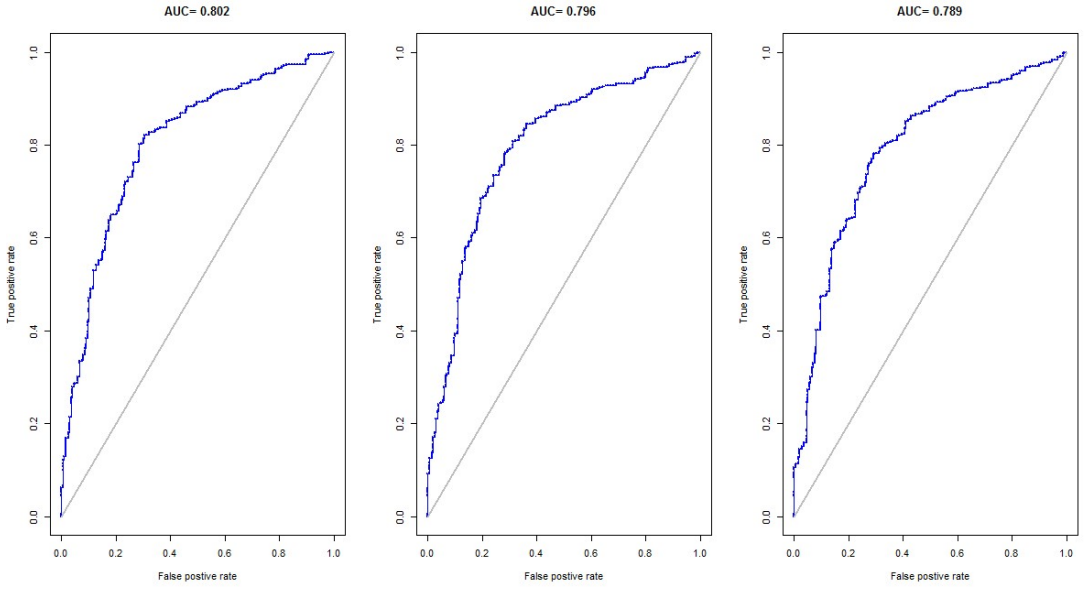
El gráfico q-q plot, indica que los residuos no siguen la diagonal adecuadamente para afirmar normalidad, además hay observaciones identificadas por la cola que pueden ser posibles outliers.

Además si nos fijamos en el gráfico de los residuos frente al leverage podemos ver como se encuentran fuera de la línea discontinua que identifica la distancia de Cook, indicando la no linealidad de los residuos y por tanto del modelo, además se puede ver que hay varias observaciones identificadas en el gráfico las cuales tienen una influencia importante en la tendencia de los residuos por lo que se podrían tratar de eliminar.

Evaluación del modelo

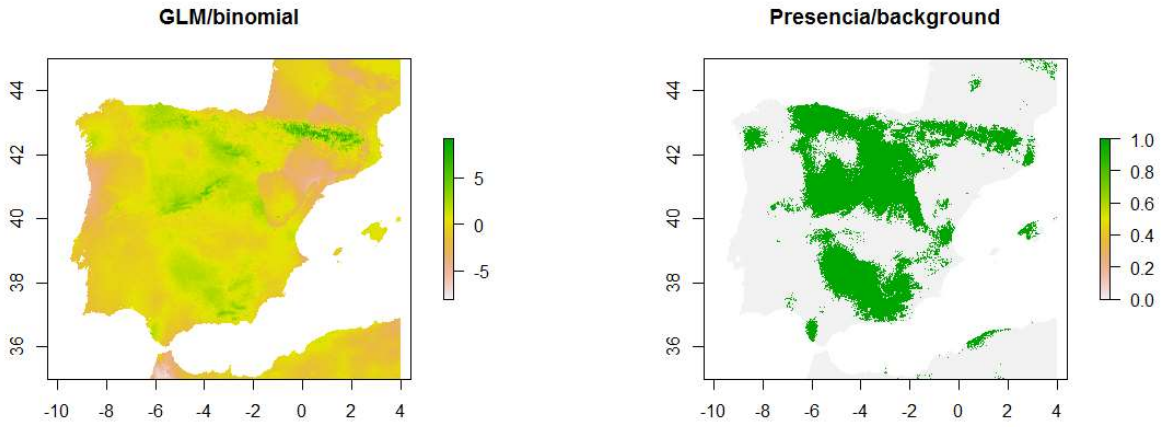
Como método de evaluación comparamos el AUC obtenido con los tres modelos, podemos ver como el área bajo la curva del modelo binomial es mayor, lo que nos hace indicar que hemos seleccionado el mejor modelo para predecir la distribución de los Aphodius.

Gráfico 6. AUC asociado a modelos GLM con todas las variables bioclimáticas



Fuente: Elaboración propia con datos de GBIF

Representación gráfica de la distribución estimada



Los resultados obtenidos con las variables preseleccionadas no son malas, sin embargo son mejores los obtenidos con el modelo que incluye todas las variables bioclimáticas. Puede que se esté sobre estimando.

Comparación de modelos

Los resultados obtenidos por ambos modelos son en general buenos, si es cierto que el modelo con las variables preseleccionadas tiene un AUC mayor que el modelo con todas las variables bioclimáticas.

Las predicciones, como se aprecia en las representaciones gráficas de sendos modelos difieren bastantes de uno y de otro, el modelo con toda las variables quizás tenga una distribución predicha menos dispersa que el modelo con las variables preselecciondas, este último muestra una predicción más amplia.
En el caso de tener que elegir uno de los dos modelos, se seleccionaría el modelo con las variables preseleccionadas por tener mejores resultados en los estadísticos de bondad de ajuste.

Modelos GAM

Este tipo modelo fue presentado por Hastie y Tibshirani en 1990, se consideran como una extensión de los modelos de regresión lineal, incorporando no linealidad y regresión no paramétrica. El modelo está construido por la suma de funciones suaves (splines) de

las variables predictoras, pudiendo ser estas variables continuas, variables categóricas, número de casos y series de datos. A diferencia de los modelos de regresión lineal (Modelo aditivo generalizado GAM: Regresión no lineal y no paramétrica, Isabel Quintas, 2012) donde se deben determinar los coeficientes correspondientes a cada uno de las variables independientes (xi), el modelo sustituye $\sum \beta i\ xi$ por una suma de funciones no necesariamente lineales $\sum ai\ fi\ (xi)$, donde cada una de las "fi" es estimada de manera muy flexible, pudiendo estas mostrar el efecto no lineal de esa relación.

Los modelos que se ejecutan a continuación realizan las estimaciones mediante mínimos cuadrados.

Resultados con las variables preseleccionadas

Se estiman tres modelos, el primero de ellos siguiendo una distribución binomial, el segundo gaussiano y el tercero según una distribución de Poisson. A continuación se muestra el Criterio de información de Akaike para cada uno de los modelos.

Tabla 14. AIC de los modelos GAM para las variables preseleccionadas

Modelo GAM	AIC
Binomial	2661.90
Gauss	2841.35
Poisson	4987.05

Fuente: Elaboración propia

Viendo los valores del estadístico AIC asociado a cada uno de los modelos, se considera el modelo binomial como el adecuado para obtener la distribución potencial de los Aphodius al tener el valor más bajo.

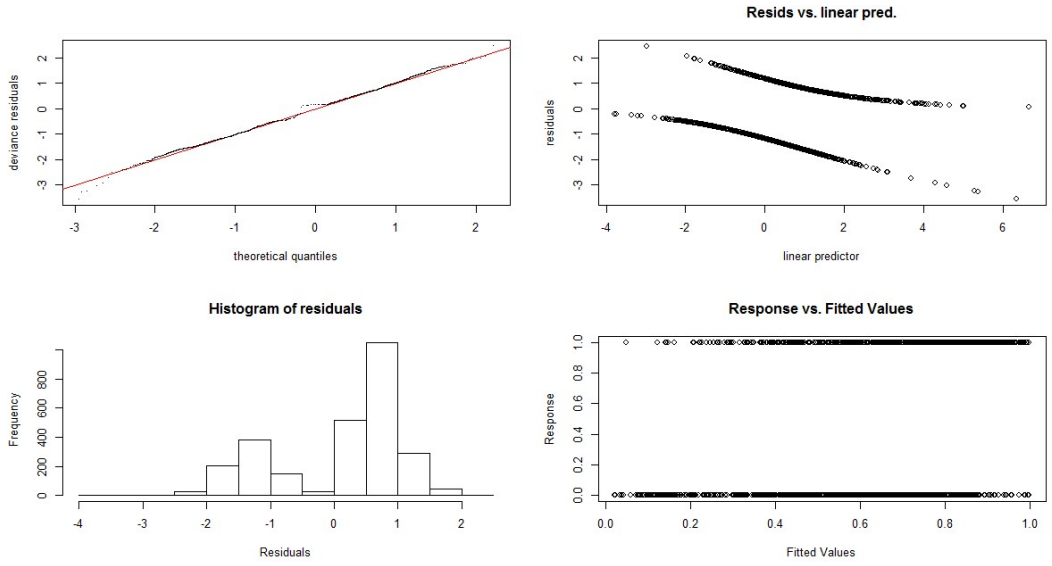
Los coeficientes estimados y la deviance residual del modelo seleccionado son:

```
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-3.5618 -0.8876  0.5180  0.7489  2.4636

(Dispersion Parameter for binomial family taken to be 1)

    Null Deviance: 3280.138 on 2697 degrees of freedom
Residual Deviance: 2635.909 on 2685 degrees of freedom
AIC: 2661.909

Number of Local Scoring Iterations: 5

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.41224    3.80384   2.212   0.0270 *
bio1_15     -3.76885    0.40263  -9.361  < 2e-16 ***
bio2_15     -0.08688    0.30140  -0.288   0.7732
bio3_15     -0.36206    0.95916  -0.377   0.7058
```

```
bio5_15      -0.07583    0.18154  -0.418   0.6762
bio8_15       0.06010    0.02348   2.560   0.0105 *
bio9_15       0.12161    0.01804   6.741 1.57e-11 ***
bio10_15      1.52928    0.31723   4.821 1.43e-06 ***
bio11_15      1.95287    0.23024   8.482  < 2e-16 ***
bio12_15      0.10072    0.02237   4.503 6.70e-06 ***
bio14_15      0.58239    0.20748   2.807   0.0050 **
bio16_15     -0.23744    0.04887  -4.859 1.18e-06 ***
bio18_15     -0.33673    0.05658  -5.952 2.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.219   Deviance explained = 19.6%
UBRE = -0.013377  Scale est. = 1         n = 2698
```

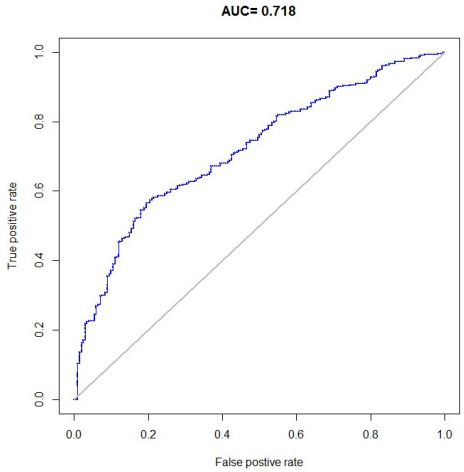
Tras los resultados vemos que variables como bio2_15, bio3_15 y bio5_15 no son nada significativas para la presencia o ausencia de Aphodius según este modelo. El R^2 estimado no es muy alto, tan solo el 22% de la presencia o ausencia depende de las variables seleccionadas. Para intentar mejorar estos resultados se podría analizar la posibilidad de eliminar las variables nada significativas del modelo, sin embargo, para lo que nos ocupa en este trabajo no nos centraremos en esto.

Análisis de los residuos

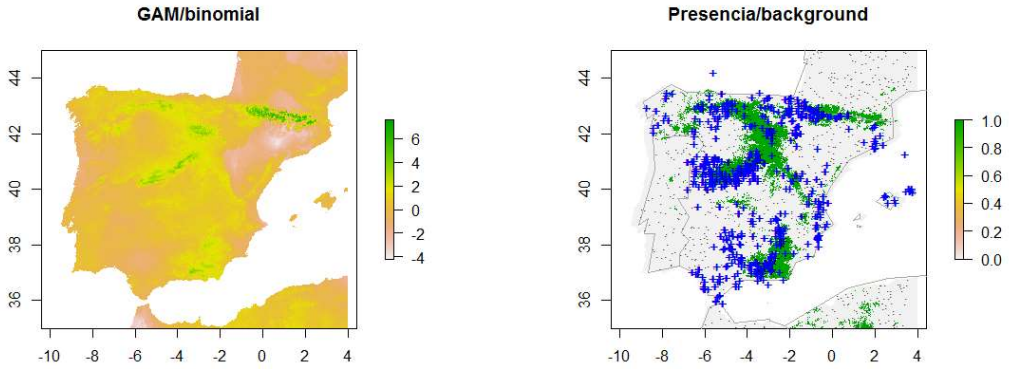


Evaluación del modelo

Para evaluar el modelo se representa la curva ROC y el valor del área bajo la curva (0.718), junto con el AIC confirmamos que el modelo según una binomial es el óptimo para la representación de las distribuciones predichas.



Representación gráfica de la distribución estimada



Resultados con todas las variables bioclimáticas

Al igual que anteriormente se han comparado tres modelos según una distribución binomial, gaussiana y una distribución de Poisson, incluyendo todas las variables bioclimáticas como predictoras, obteniéndose los siguientes resultados con respecto al Criterio de información de Akaike.

Tabla 15. AIC de los modelos GAM para las variables preseleccionadas

Modelo GLM	AIC
Binomial	2435.65
Gauss	2769.42
Poisson	4968.81

Fuente: Elaboración propia

Como se aprecia en la tabla, el valor AIC más bajo corresponde al modelo según una binomial, por lo que tomaremos este modelo como el idóneo para obtener y representar las predicciones de los Aphodius.

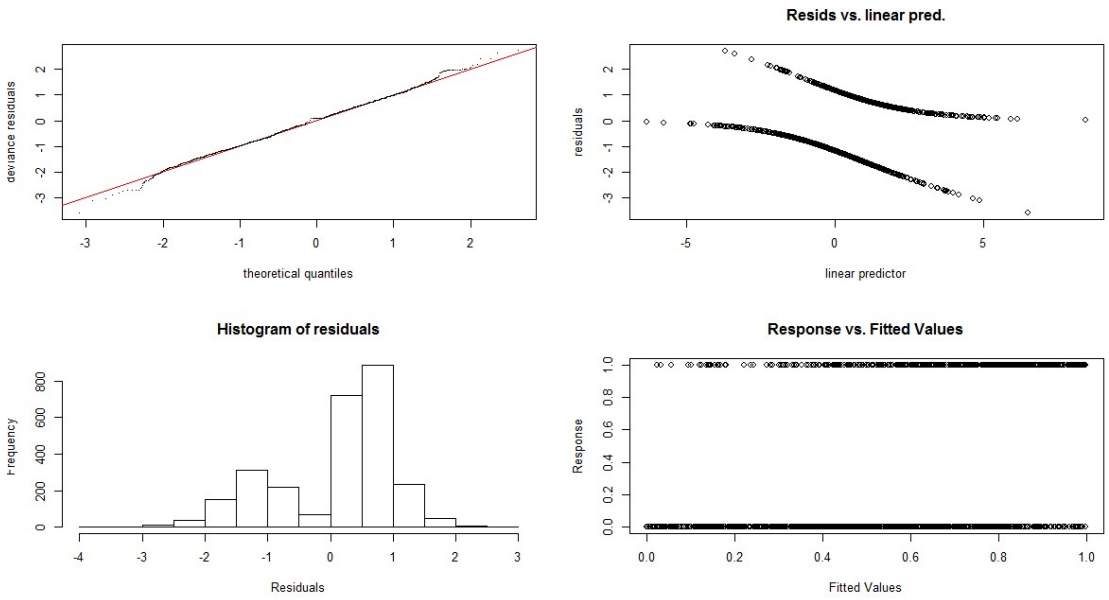
Los coeficientes estimados y la deviance residual del modelo seleccionado son:

```
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-3.6112 -0.6682  0.4172  0.6919  2.7277

(Dispersion Parameter for binomial family taken to be 1)

    Null Deviance: 3280.138 on 2697 degrees of freedom
Residual Deviance: 2397.651 on 2679 degrees of freedom
AIC: 2435.651

Number of Local Scoring Iterations: 5

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 22.89211    5.14070   4.453 8.46e-06 ***
bio1_15     -2.74297    0.58275  -4.707 2.51e-06 ***
bio2_15      2.65131    0.55447   4.782 1.74e-06 ***
bio3_15     -4.53517    1.29992  -3.489 0.000485 ***
bio4_15     -0.01846    0.02074  -0.890 0.373490
bio5_15     -1.15420    0.28469  -4.054 5.03e-05 ***
bio6_15      2.65573    0.33077   8.029 9.83e-16 ***
bio8_15     -0.03339    0.02998  -1.114 0.265401
bio9_15      0.22663    0.02735   8.287  < 2e-16 ***
bio10_15     2.19148    1.00028   2.191 0.028461 *
bio11_15    -1.36002    0.71214  -1.910 0.056162 .
bio12_15     0.23476    0.03719   6.312 2.76e-10 ***
bio13_15     1.21020    0.16466   7.349 1.99e-13 ***
bio14_15     2.89964    0.34028   8.521  < 2e-16 ***
bio15_15     0.46101    0.18517   2.490 0.012786 *
bio16_15    -0.61114    0.09196  -6.646 3.01e-11 ***
bio17_15    -0.96479    0.13399  -7.200 6.01e-13 ***
bio18_15    -0.26906    0.09241  -2.912 0.003595 **
bio19_15    -0.35935    0.06450  -5.571 2.53e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.313   Deviance explained = 26.9%
```

En este caso para la presencia o ausencia de Aphodius no es nada significativo las variables bio11_15 y bio18_15.

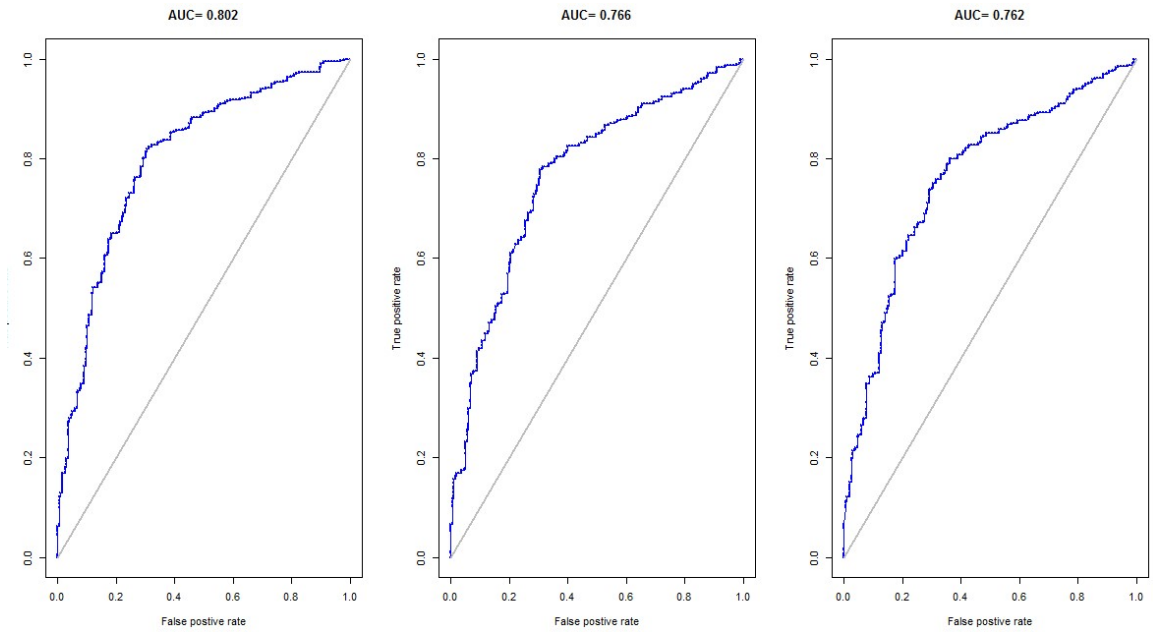
Análisis de los residuos

En este apartado se pueden ver la distribución de los residuos según la familia del modelo, en este caso según una binomial, no se aprecia heterocedasticidad y del gráfico izquierdo inferior podemos ver como la mayoría de residuos se concentran en torno a los valores 0, 1 al ser un modelo dicotómico. No se aprecian anomalías exageradas por lo que podemos considerar el modelo adecuado.

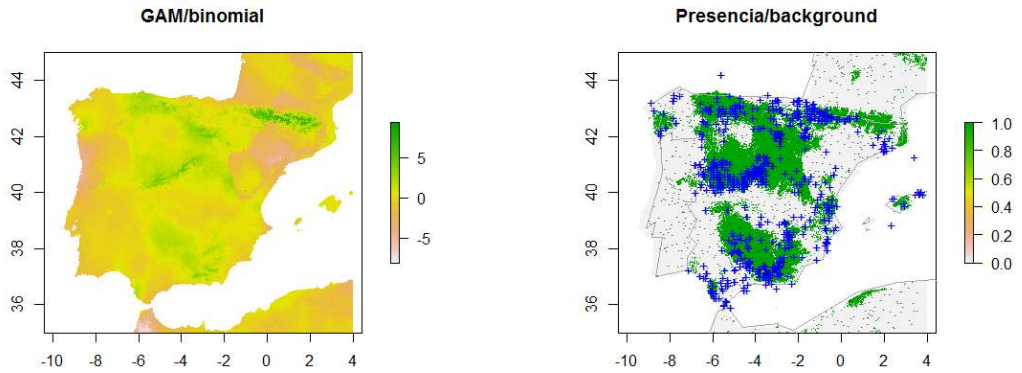


Evaluación del modelo

En este caso comparamos las curvas ROC de los tres modelos planteados, pudiendo ver en el gráfico de abajo como la primera curva correspondiente al modelo binomial es el que mayor AUC tiene, por lo que con este resultado además sabiendo que el modelo era el que menor AIC presentaba, podemos decir que es el más adecuado para representar la distribución predicha de los Aphodius.



Representación gráfica de la distribución estimada



Comparación de modelos

Tras los resultados obtenidos, el modelo que incluye todas las variables bioclimáticas como independientes tiene mejores resultados que el modelo con las variables preseleccionadas, en concreto el AUC de este último es de 0.718, en cambio el modelo con todas las variables es muy superior, 0.802.

En el caso de tener que elegir uno de los dos modelos para representar la distribución potencial de los Aphodius sería con el modelo que tiene todas las variables bioclimáticas.

Machine Learning

Gradient Boosting

Este algoritmo es un método iterativo, va alternando las constantes de regularización, las predicciones las va estimando poco a poco en la dirección de decrecimiento dada por el negativo del gradiente.

Resultados con las variables preseleccionadas

Utilizando las variables seleccionadas previamente se estiman varios modelos con el fin de elegir con validación cruzada el modelo que tenga menor tasa de error medio y finalmente rasterizar la distribución predicha obtenida por el modelo elegido.

Modelo	Selección[3]	Programa	Parámetros	Error medio
GB1	Vc	SAS	- 300 iteraciones - Shrinkage 0.02 - 2 divisiones en cada nodo - 5 hojas finales	0.0666
GB2	Vc	SAS	- 500 iteraciones - Shrinkage 0.02 - 2 divisiones en cada nodo - 5 hojas finales	0.0632
GB3	Vc	SAS	- 1000 iteraciones - Shrinkage 0.01 - 10 divisiones en cada nodo - 10 hojas finales	0.0606061
GB4	Vc	SAS	- 1000 iteraciones - Shrinkage 0.05 - 10 divisiones en cada nodo - 10 hojas finales	0.0659
GB5	Vc	R	- >1000 iteraciones - Shrinkage 0.01 - 10 divisiones en cada nodo - 2 hojas finales - Ntrees 10000	0.0274

De los árboles estimados el que menor **error medio** presenta es el calculado con R, el programa realiza inicialmente 10 modelos de 50 árboles estratificando por la prevalencia, continua haciendo lotes de 10 modelos aumentando el número de árboles de 50 en 50 y promediando la deviance residual, hasta que finalmente obtiene un modelo con la deviance residual más baja. En este caso y como se ve en la tabla anterior el programa encuentra un modelo con 10.000 árboles, y una desviance residual de 0.521, como se puede ver en la tabla siguiente.

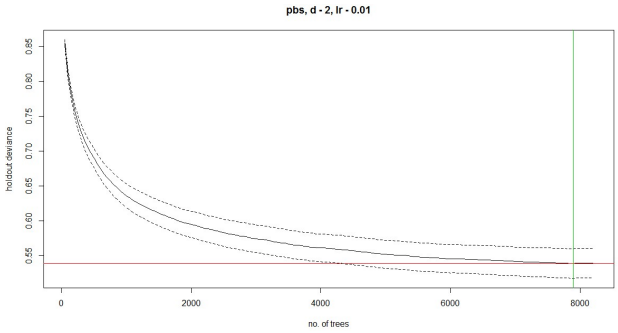
[3] El modelo final ha sido seleccionado con Validación cruzada

Tabla 16. Deviance residual del modelo GB elegido con las var. preseleccionadas

Media deviance total	**0.**934
Media deviance residual	0.284
Deviance residual obtenida con VC	0.521

Fuente: Elaboración propia con datos de GBIF

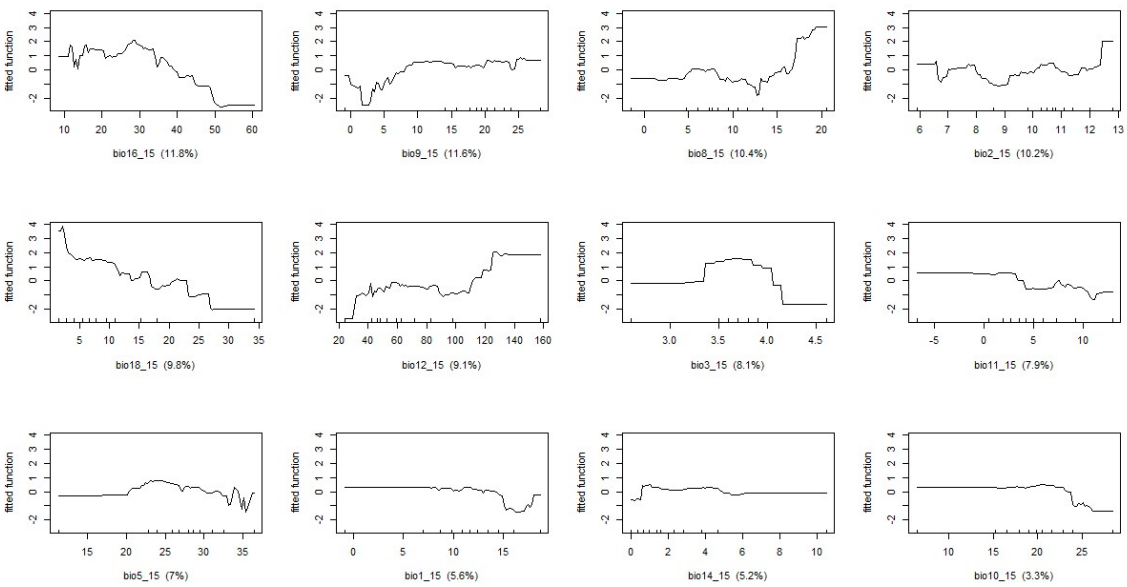
Gráficamente podemos ver la convergencia de los árboles hasta obtener el resultado óptimo. La deviance calculada para cada árbol se puede ver en el Anexo del documento.

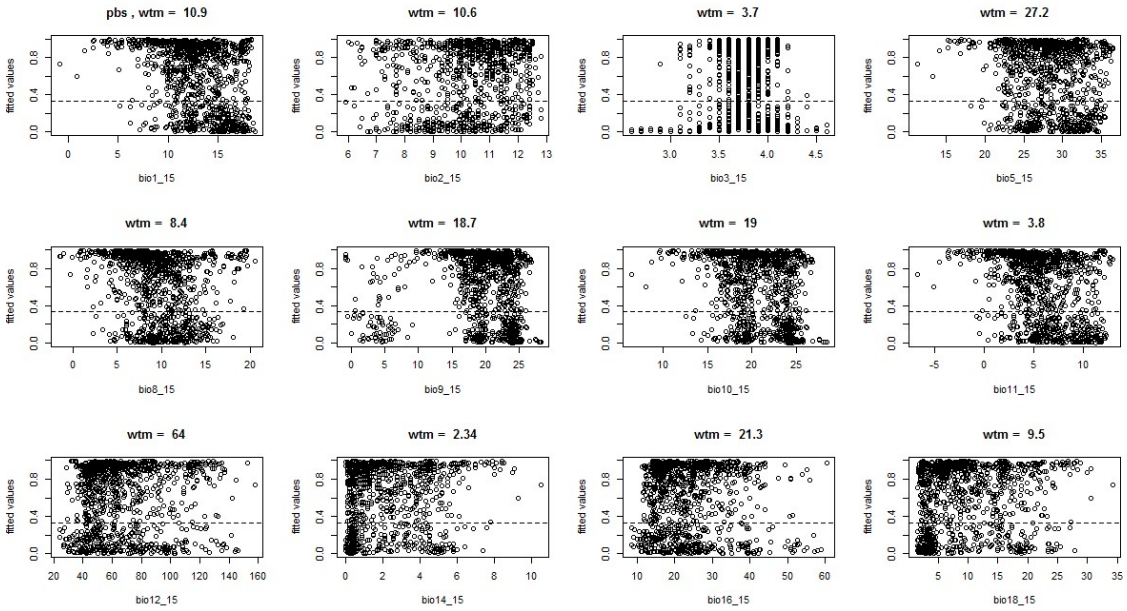


Plot de las funciones y valores predichos para cada variable bioclimática

En este apartado se puede ver la distribución de las funciones ajustadas para cada variable predictora.

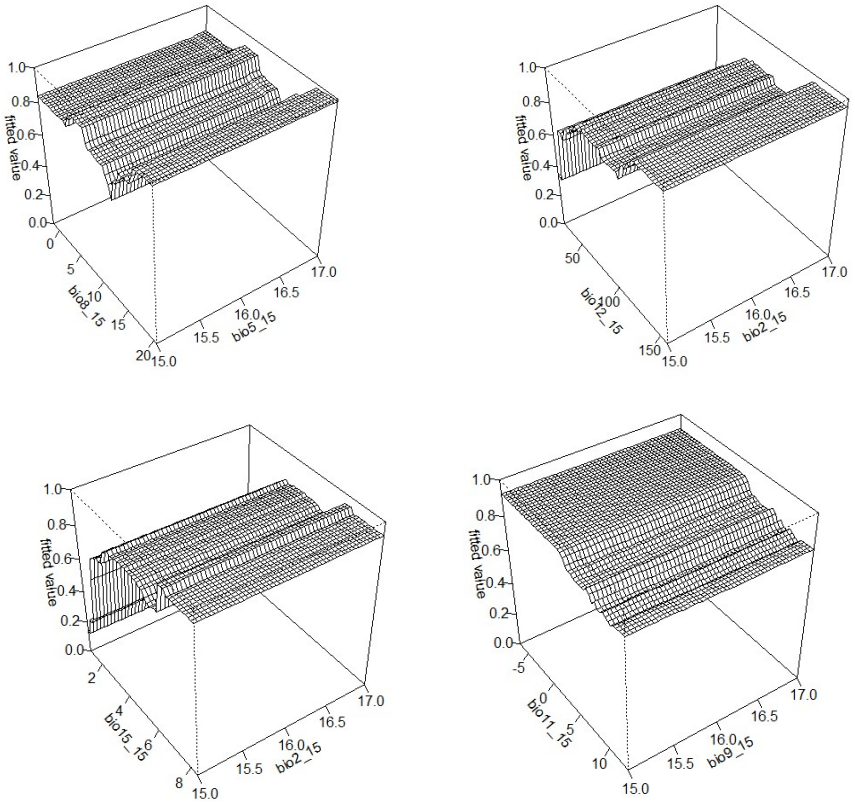
Dependiendo de la distribución de las observaciones dentro del espacio ambiental, las funciones ajustadas pueden dar una indicación engañosa sobre la distribución de los valores ajustados en relación con cada predictor. En este modelo predictivo parece que todas las observaciones ajustadas se distribuyen según la función estimada.





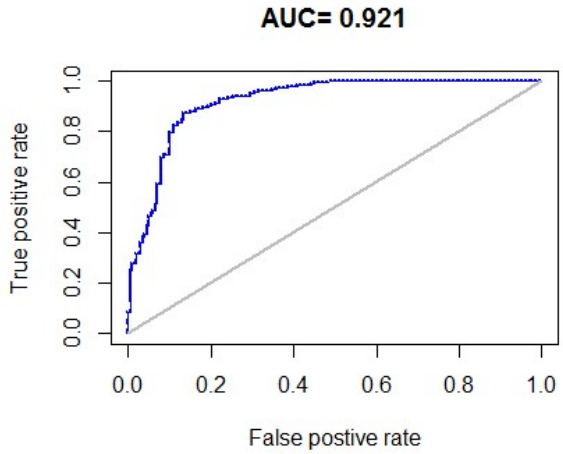
Las interacciones entre variables y el valor ajustado son interesantes para ver a que valores ajustados le corresponden un valor alto o escaso en función de un par de variables ambientales, a continuación se representa la interacción de las ocho variables con mayor influencia relativa.

Variable	Influencia relativa
Bio1_15	765.1034
Bio2_15	1401.9633
Bio3_15	1115.7572
Bio5_15	966.9879
Bio8_15	1425.4057
Bio9_15	1585.5685
Bio10_15	451.8386
Bio11_15	1080.1105
Bio12_15	1255.1667
Bio14_15	707.8071
Bio16_15	1617.7189
Bio18_15	1350.5055

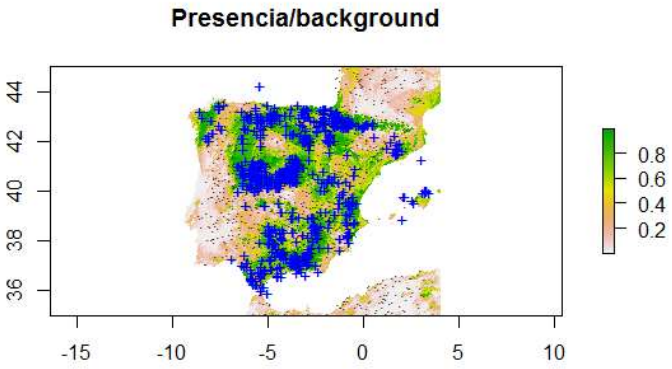


Evaluación del modelo

Como en los modelos anteriores se evalúa el modelo obteniendo el AUC, en este caso se puede ver como el valor obtenido es bastante bueno (0.92), por lo que podemos decir que el modelo es un buen predictor de la distribución de los Aphodius.



Representación gráfica de la distribución estimada



Resultados con todas las variables bioclimáticas

En este apartado se exponen los resultados obtenidos tomando como variables predictoras todas las variables bioclimáticas. Al igual que en el apartado anterior se han comparado los resultados de los mismos modelos anteriores.

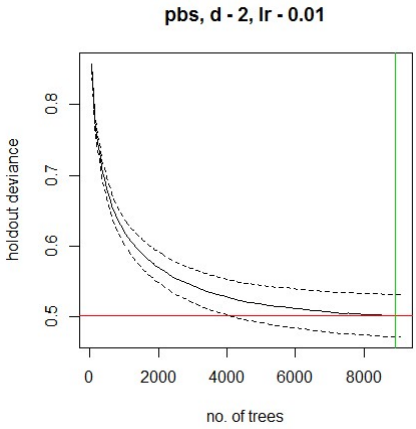
El modelo que menor error medio presenta es de nuevo el calculado con R, en este caso con 8.900 árboles, una desviance residual de 0.5018, y un error medio de 0.029.

Tabla 17. Deviance residual de modelo GB elegido con todas las variables

Media deviance total	**0.**934
Media deviance residual	0.284
Deviance residual obtenida con VC	0.502

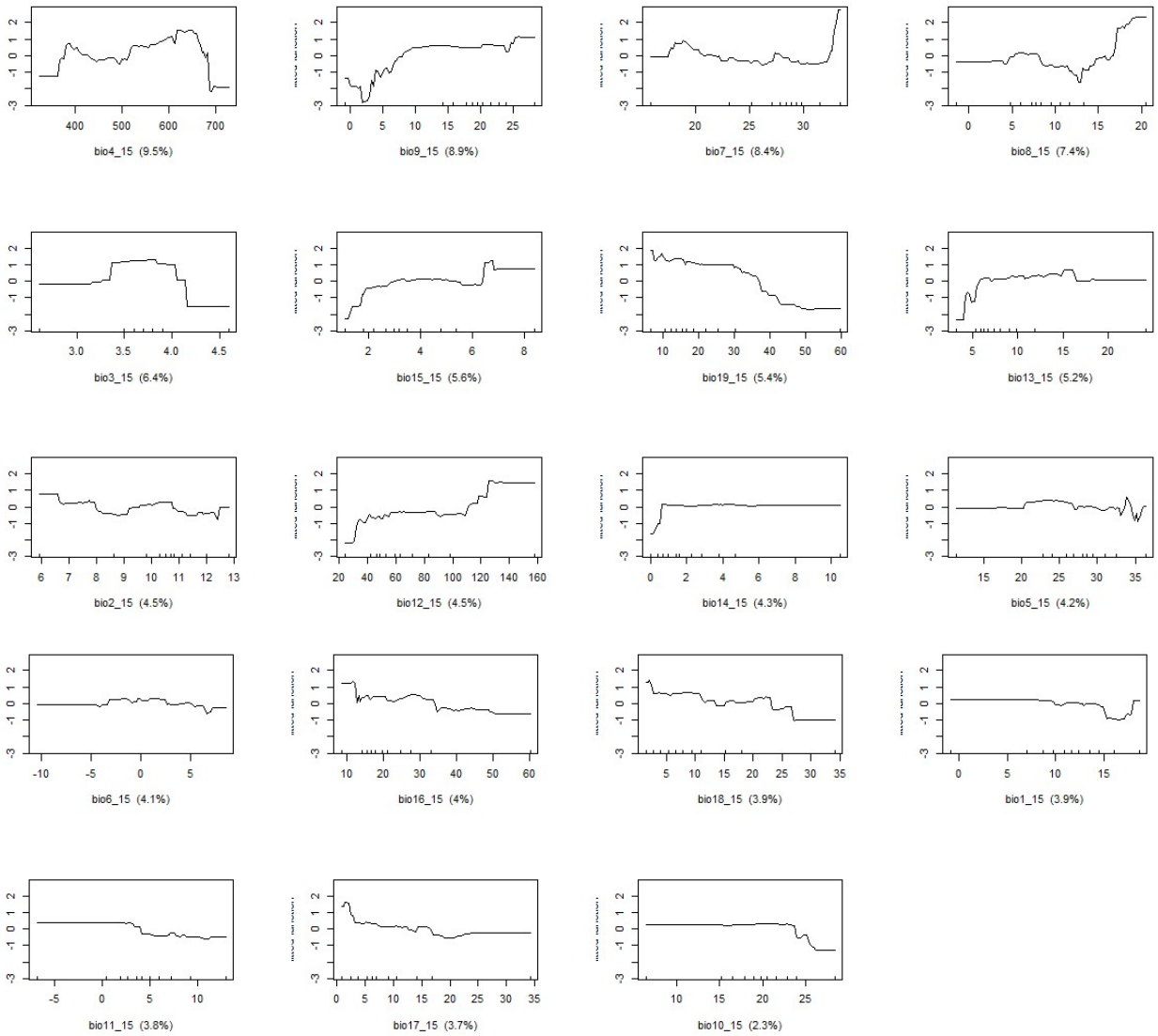
Fuente: Elaboración propia con datos de GBIF

Gráficamente podemos ver la convergencia de los árboles hasta obtener el resultado óptimo.



Plot de las funciones y valores ajustados por variable bioclimática

Las funciones ajustadas dibujan la distribución de las predicciones para cada variable predictora.



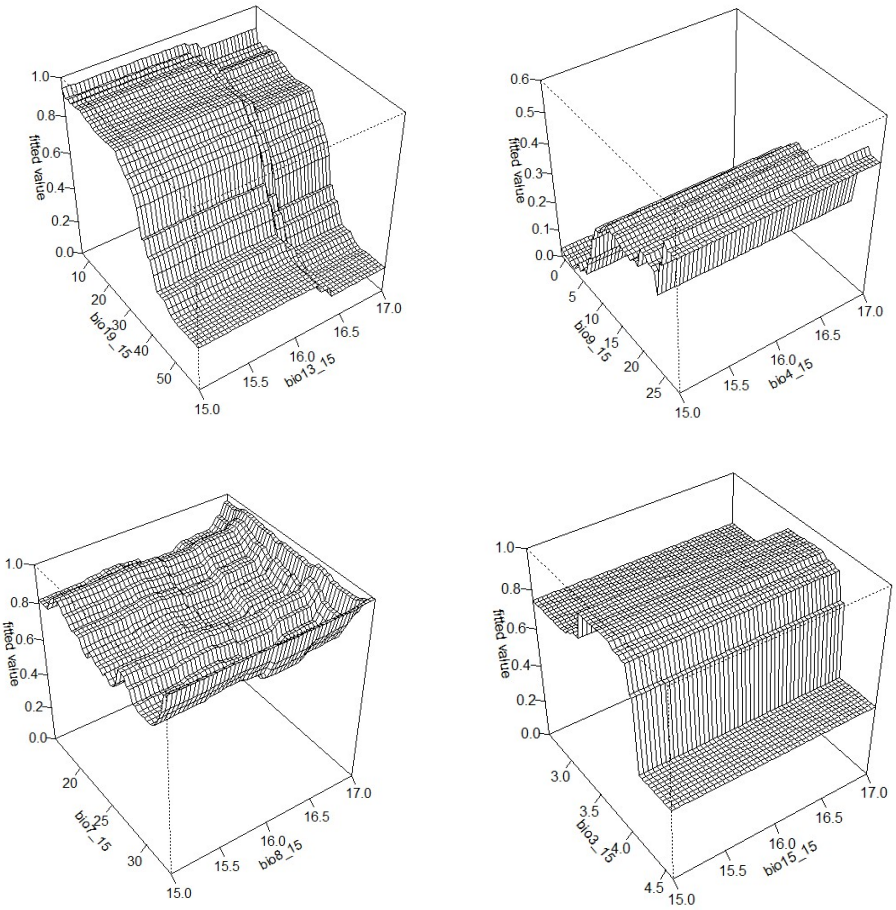
Influencia de las variables bioclimáticas

En la siguiente tabla se muestran las variables ordenadas de mayor a menor influencia relativa sobre la variable dependiente.

Variable	Influencia relativa
Bio4_15	9.478084
Bio9_15	8.929383
Bio7_15	8.354955
Bio8_15	7.355038
Bio3_15	6.427911

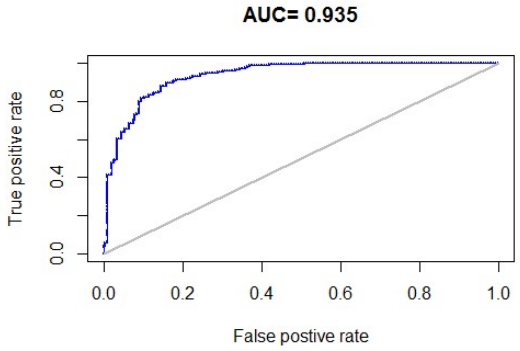
Bio15_15	5.635258
Bio19_15	5.364134
Bio13_15	5.161557
Bio2_15	4.498714
Bio12_15	4.454505
Bio14_15	4.282852
Bio5_15	4.190228
Bio6_15	4.136639
Bio16_15	4.034783
Bio18_15	3.945002
Bio1_15	3.805559
Bio11_15	3.748016
Bio17_15	2.310264
Bio10_15	9.478084

A continuación se muestran las interacciones de las ocho variables de mayor importancia para el GB.

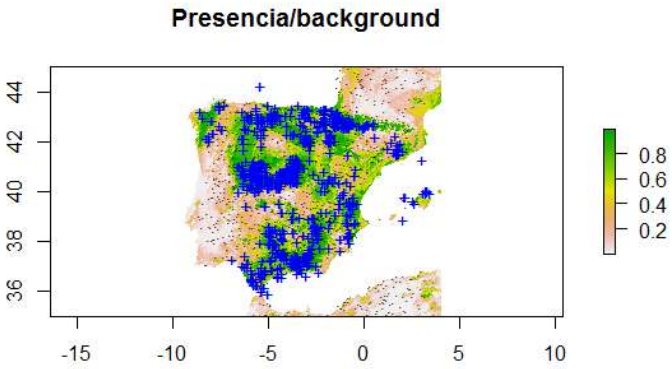


Evaluación del modelo

Como en los modelos anteriores se evalua el modelo obteniendo el AUC, en este caso se puede ver como el valor es bastante alto, por lo que podemos decir que el modelo es un buen predictor de la distribución de los Aphodius.

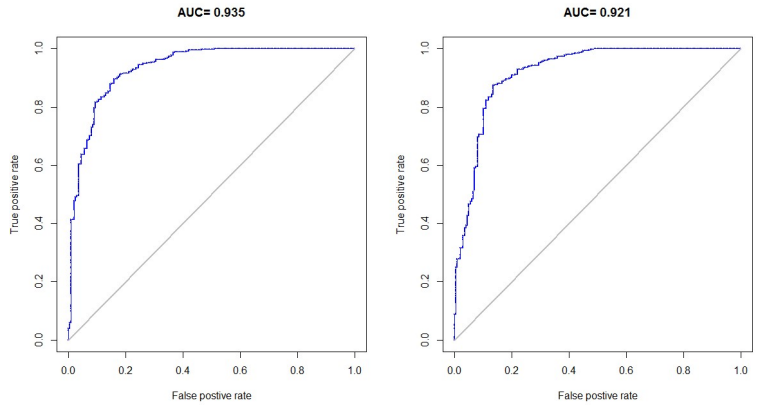


Representación gráfica de la distribución estimada



Comparación de modelos

El modelo ajustado con las variables previamente seleccionadas se obtiene un AUC algo menor (0.921) que el obtenido en el caso en el que se incluyen todas las variables (0.935)



Si vemos los gráficos de las distribuciones estimadas podemos ver que no hay diferencias muy significativas entre el modelo con todas las variables ambientales y el modelo con las variables preseleccionadas.

Siguiendo el criterio de parsimonia a la hora de seleccionar un modelo para representar la distribución potencia de los Aphodius nos quedaríamos con el modelo que incluye las variables preseleccionadas.

Random Forest

Este algoritmo por definirlo de alguna manera sencilla es una combinación de árboles de predicción tal que cada árbol depende de los valores de una componente aleatoria independiente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging.

En forma resumida sigue este proceso:

1- Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes set de datos.

2- Crea un árbol de decisión con cada set de datos, obteniendo diferentes árboles, ya que cada set contiene diferentes individuos y diferentes variables.

3- Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad.

4- Predice los nuevos datos usando el "voto mayoritario", donde clasificará como "positivo" si la mayoría de los arboles predicen la observación como positiva.

Resultados con las variables preseleccionadas

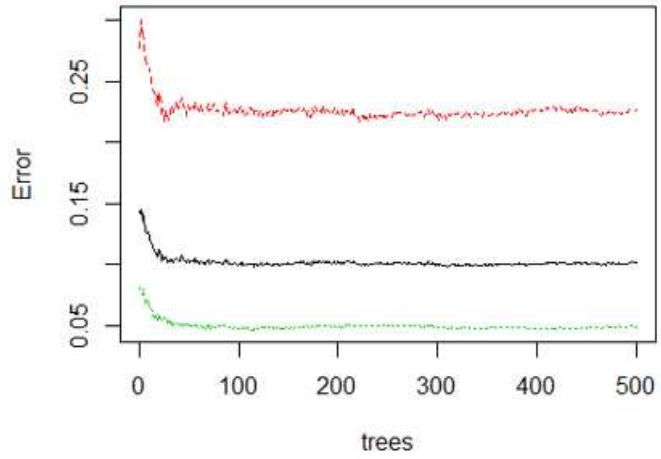
Se realizaron varios modelos de RF modificando parámetros tales como el nº máximo de árboles, cuantas divisiones por nodo, hojas finales, variables a muestrear en cada nodo, porcenbag, toma de muestras con o sin reemplazamiento, etc. Los resultados obtenidos se encuentran en el Anexo.

Modelo	Selección[4]	Parámetros	Error (Test)
Random Forest	Vc	- 30 árboles como máximo - 2 divisiones máximas en cada nodo - 5 hojas finales del árbol - Variables a muestrear en cada nodo 3 - Porcenbag 0.5 - Con reemplazamiento	0.0695
Random Forest	Vc	- 100 árboles como máximo - 10 divisiones máximas en cada nodo - 5 hojas finales del árbol - Variables a muestrear en cada nodo 3 - Porcenbag 0.01 - Con reemplazamiento	0.18716
Random Forest	Vc	- 500 árboles como máximo - 2 divisiones máximas en cada nodo - 3 hojas finales del árbol - Variables a muestrear en cada nodo 4 - Porcenbag 0.8 - Con reemplazamiento	0.0694

Después de ver los resultados obtenidos con los diferentes modelos, se elige como modelo óptimo el último de la tabla anterior. Este modelo tiene un erro medio de clasificación inferior al resto, además este error para los datos presenciales de Aphodius es muy bajo como se puede ver en el gráfico siguiente, indicando que su ajuste para estas observaciones es bueno. (Línea roja: background, línea negra: OOB, línea verde: presencia)

[4] El modelo final ha sido seleccionado con Validación cruzada

Gráfico 7. Error de clasificación para RF variables preseleccionadas



Fuente: Elaboración propia con datos de GBIF

Matriz de confusión

Con este modelo se obtiene una tasa de error de 10.12% para las predicciones de las observaciones que quedan fuera para la creación de los árboles (OOB, out of bag). La matriz de confusión que se obtiene con las observaciones que sí entran en la creación de los árboles es la siguiente.

Tabla 18. Matriz de confusión RF con las variables preseleccionadas

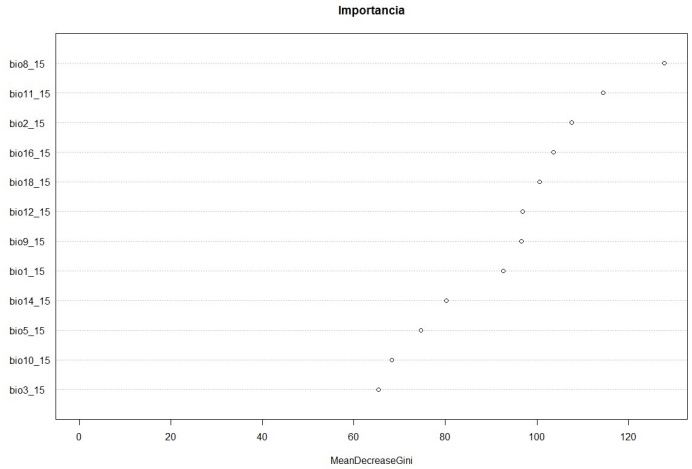
	Presencia	Ausencia	Error
Presencia	619	181	0.2262
Ausencia	92	1806	0.0484

Fuente: Elaboración propia con datos de GBIF

Se aprecia como el modelo predice mejor las observaciones de presencia de Aphodius, parece que esta sobreajustando dicha observaciones, sin embargo para el interés del trabajo es algo que se va persiguiendo, modelos que sean buenos predictores de los datos de presencia de especies, es decir obtener pocos falsos positivos .

Importancia de las variables bioclimáticas

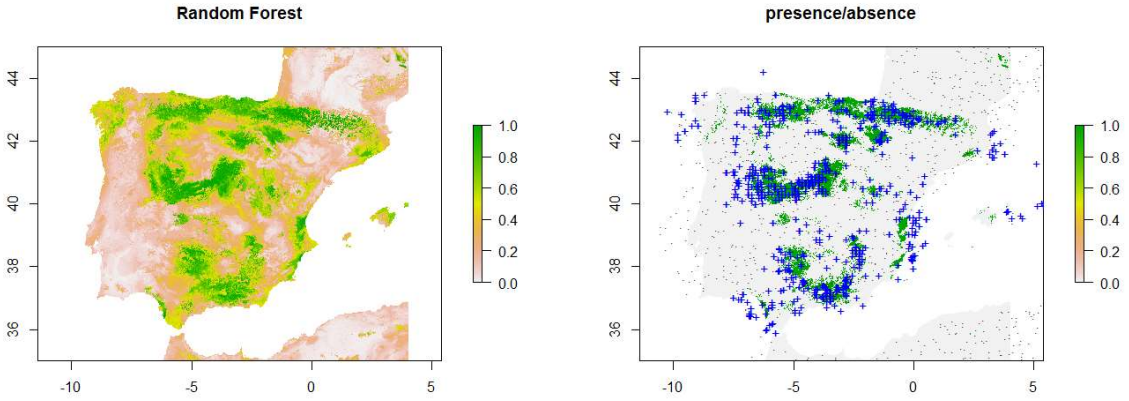
Según este algoritmo las variables con mayor importancia sobre la presencia o ausencia de Aphodius son bio8_15, bio11_15, bio2_15 y la de menor importancia bio3_15.



Evaluación del modelo

Como en los modelos anteriores se evalúa el modelo obteniendo el AUC, en este caso el valor obtenido es de 0.836, un resultado bastante alto, por lo que podemos decir que el modelo es un buen predictor de la distribución de los Aphodius.

Representación de la distribución estimada



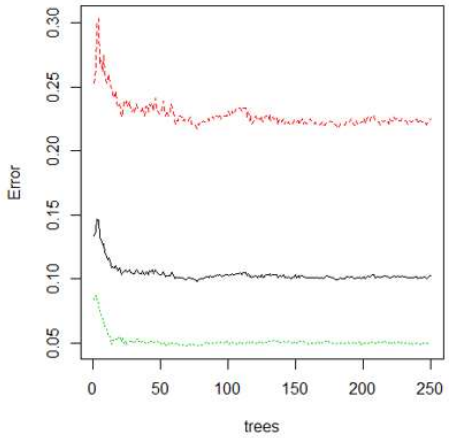
Resultados con todas las variables bioclimáticas

En este apartado se han comparado los mismos modelos del apartado anterior salvo que en esta ocasión se han tomado como variables predictoras todas las variables bioclimáticas. Los errores obtenidos son:

Modelo	Selección[5]	Parámetros	Error (Test)
Random Forest	Vc	- 30 árboles como máximo - 2 divisiones máximas en cada nodo - 5 hojas finales del árbol - Variables a muestrear en cada nodo 3 - Porcenbag 0.5 - Con reemplazamiento	0.0819
Random Forest	Vc	- 100 árboles como máximo - 10 divisiones máximas en cada nodo - 5 hojas finales del árbol - Variables a muestrear en cada nodo 3 - Porcenbag 0.01 - Con reemplazamiento	0.18716
Random Forest	Vc	- 500 árboles como máximo - 2 divisiones máximas en cada nodo - 3 hojas finales del árbol - Variables a muestrear en cada nodo 3 - Porcenbag 0.8 - Con reemplazamiento	0.0748

Después de ver los resultados obtenidos con los diferentes modelos, se elige como modelo óptimo el último de la tabla anterior. Este modelo tiene un error medio de clasificación inferior al resto, además este error para los datos presenciales de Aphodius es muy bajo como se puede ver en el gráfico siguiente, indicando que su ajuste para estas observaciones es bueno. (Línea roja: background, línea negra: OOB, línea verde: presencia).

Gráfico 8. Error de clasificación para RF todas las variables bioclimáticas



Fuente: Elaboración propia con datos de GBIF

[5] El modelo final ha sido seleccionado con Validación cruzada

Matriz de confusión

Al igual que ocurre con las variables preseleccionadas, este modelo sobreajusta las predicciones de las Presencia de Aphodius.

Tabla 19. Matriz de confusión RF con todas las variables

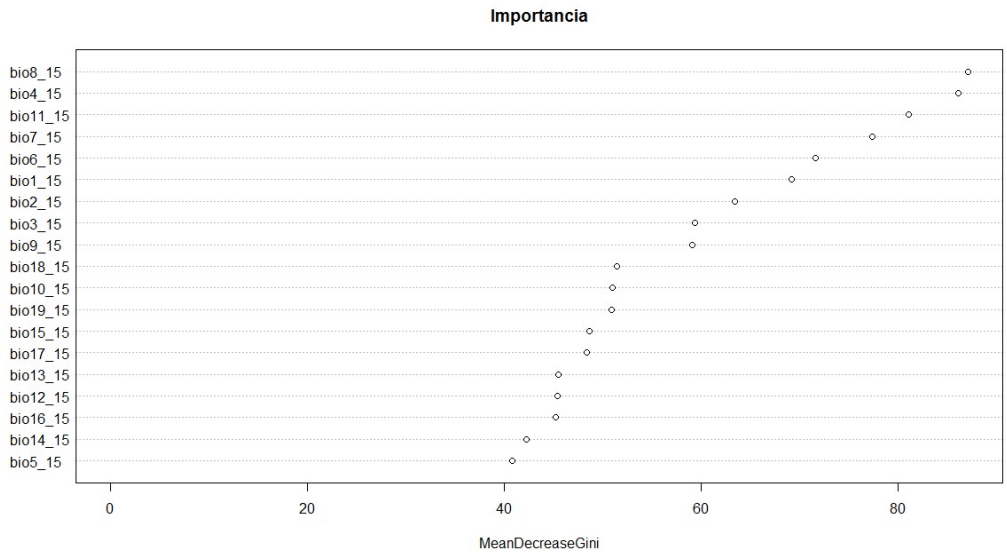
	Presencia	Ausencia	Error
Presencia	627	173	0.21625
Ausencia	88	1810	0.04634

Fuente: Elaboración propia

Con este modelo se obtiene una tasa de error de 9.67% para las predicciones de las observaciones que quedan fuera para la creación de los árboles (OOB, out of bag).

Importancia de las variables bioclimáticas

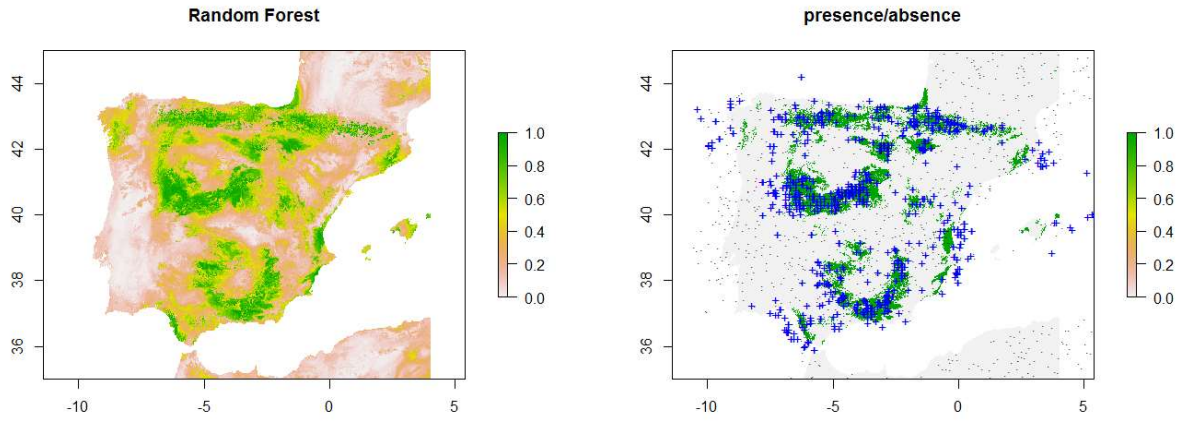
En el siguiente gráfico se puede ver la importancia que este algoritmo da a las variables utilizadas. Esto podría ser de utilidad en el caso de seleccionar variables para otras predicciones. Las variables bioclimáticas más importantes según este algoritmo son *bio8_15* seguida *bio4_15* y la de importancia menor es bio5_15 muy seguida de bio14_15.



Evaluación del modelo

Con este modelo hemos conseguido aumentar levemente el AUC, eso sí, complicando el modelo con siete variables más, en este caso el valor del área bajo la curva ROC es de 0.8409, un resultado bastante alto.

Representación de la distribución estimada



Comparación de los modelos

Con ambos modelos se obtienen resultados muy similares, ambos predicen bien tanto las presencias de Aphodius como los backgrounds, esta predicción se refleja en la representación de la distribución estimada de cada modelo, siendo muy similar la presencia real a la estimada.

En el caso de tener que elegir uno de los dos modelos, siguiendo el criterio de parsimonia seleccionaríamos el modelo con las variables preseleccionadas.

Suppor Vector Machine

A modo de resumen se puede decir que este algoritmo trata de plantear el problema de separación lineal de clases con métodos algebraicos (buscar el hiperplano de separación).

Resultados con las variables preseleccionadas

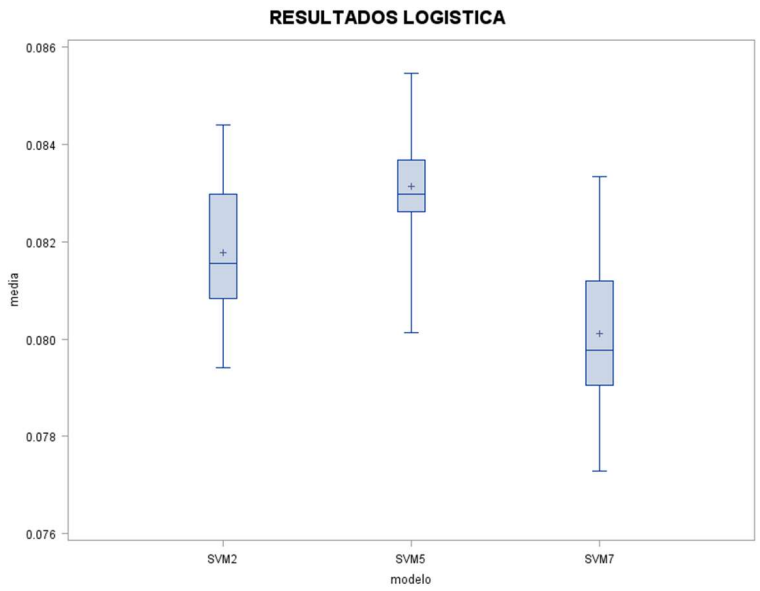
Al igual que en los algoritmos anteriores, previamente al cálculo de las predicciones, se busca el modelo que menor tasa error medio usando validación cruzada.

Modelo	Parámetro regularización	Kernel	Tasa error medio	Std
SVM1	C=10	Lineal	0.4524	0.1496448
SVM2	C=10	Polinómica grado 2	0.0817	0.0012763

SVM3	C=10	RBF	0.0840	-
SVM4	C=5	Lineal	0.3809	0.1334999
SVM5	C=5	Polinómica grado 2	0.0831	0.0011007
SVM6	C=5	RBF	0.0840	-
SVM7	C=20	Polinómica grado 2	0.0801	0.0013815

Fuente: Elaboración propia

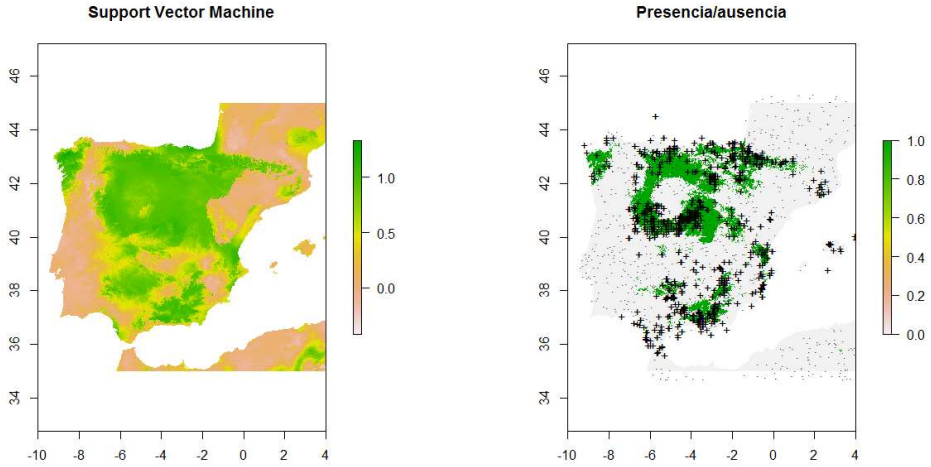
Los mejores modelos corresponden a una función polinómica de grado 2, como cuando hablamos de distribución potencial interesa reducir la variabilidad de los valores predichos, se selecciona como adecuado el modelo SVM2 que aunque no sea el de menor error si es el que tiene una variabilidad muy reducida que es lo que nos interesa, esto lo podemos ver gráficamente.



Evaluación del modelo

El modelo obtenido con las variables preseleccionadas es bastante bueno, el AUC asociado tiene un valor de 0.746.

Representación gráfica de la distribución estimada



Resultados con todas las variables bioclimáticas

Al igual que en el apartado anterior se estiman varios modelos de SVM, en este caso se incluyeron todas las variables bioclimáticas como predictoras.

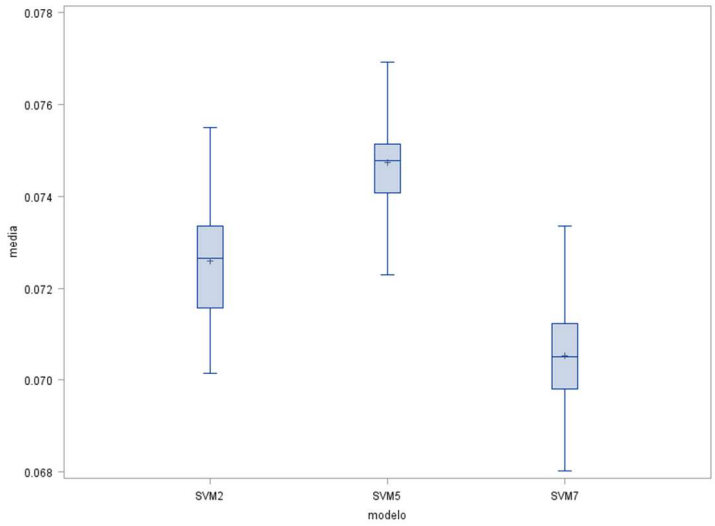
Tabla 20. Tasa de erro de los modelos estimados según SVM

Modelo	Parámetro regularización	Kernel	Tasa error medio	Std
SVM1	C=10	Lineal	0.4739768	0.1693055
SVM2	C=10	Polinómica grado 2	0.0725975	0.0011719
SVM3	C=10	RBF	0.0712251	-
SVM4	C=5	Lineal	0.3902960	0.1512500
SVM5	C=5	Polinómica grado 2	0.0747429	0.000919035
SVM6	C=5	RBF	0.0769231	-
SVM7	C=20	Polinómica grado 2	0.0705215	0.0011984

Fuente: Elaboración propia

Se observa que los mejores modelos corresponden a una función polinómica de grado 2, como cuando hablamos de distribución potencial interesa reducir la variabilidad de los valores predichos, se selecciona como adecuado el modelo SVM5 que aunque no sea el de menor error si es el que tiene una variabilidad muy reducida que es lo que nos interesa, esto lo podemos ver gráficamente.

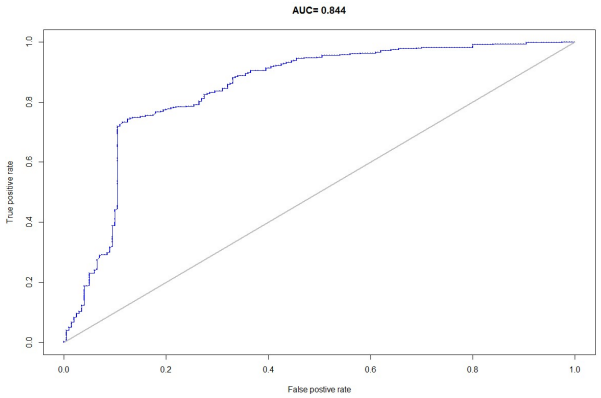
Gráfico 9. Box-plot del error medio - SVM



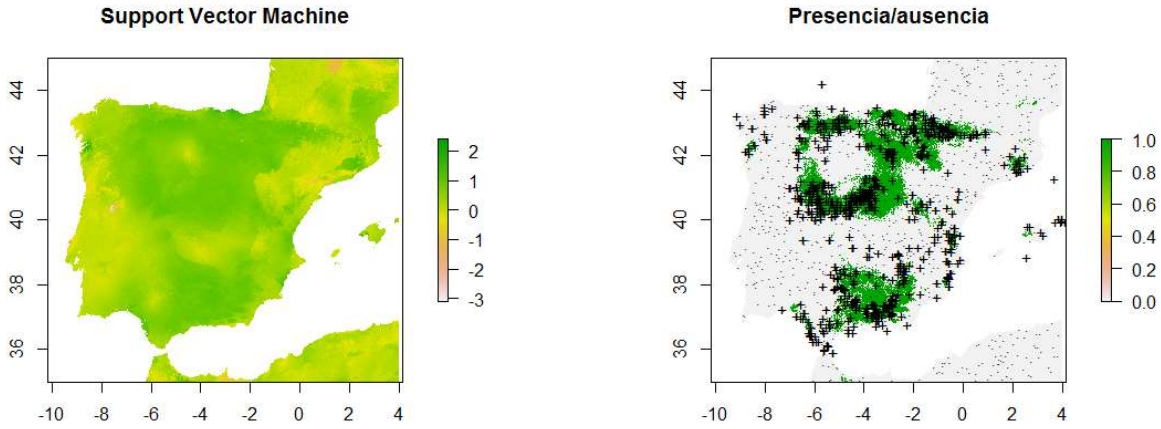
Fuente: Elaboración propia

Evaluación del modelo

En esta ocasión el modelo es bueno, el AUC obtenido es superior al del apartado anterior.



Representación gráfica de la distribución estimada



Comparación de modelos

Aunque a priori, si nos fijamos en las predicciones rasterizadas de ambos modelos, los resultados no son iguales. El modelo que incluye todas las variables es más restrictivo o lo que es lo mismo da sensación de ser más probable sus resultados, además el AUC obtenido es más elevado que el del modelo con variables preseleccionadas.
En el caso de tener que elegir uno de los dos modelos, nos decantaríamos utilizando este algoritmo por el modelo con todas las variables bioclimáticas.

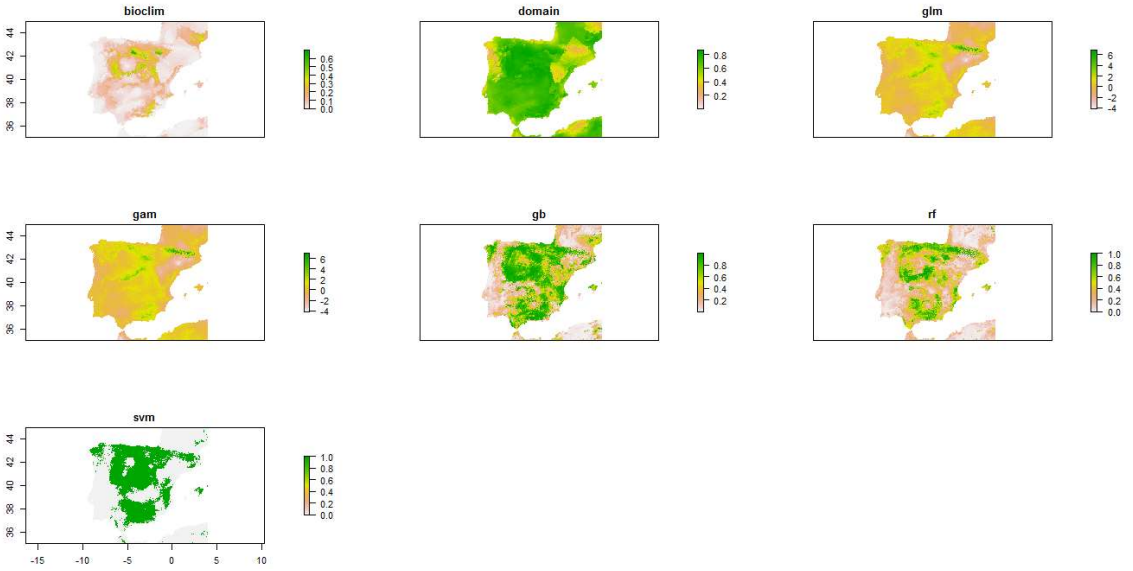
Ensamblado - Stacking

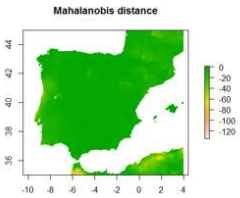
Una vez visto varios modelos se pretende en esta apartado construir un ensamblado de estos con el fin de intentar reducir la varianza de las predicciones, para ello se realiza Stacking, promediando las predicciones, en este caso al ser modelos de clasificación, se obtiene el promedio de las probabilidades.

Los modelos incluidos en el ensamblado son: Bioclim, Domain, Mahalanobis, GLM, GAM, Gradient Boosting, Random Forest y SVM.

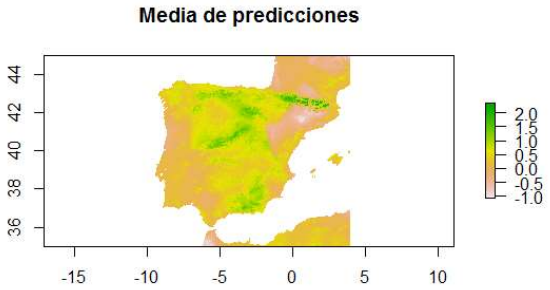
Resultados con las variables preseleccionadas

Los siguientes mapas se corresponden a los modelos seleccionados de cada algoritmo incluyendo todas las variables bioclimáticas como predictoras.



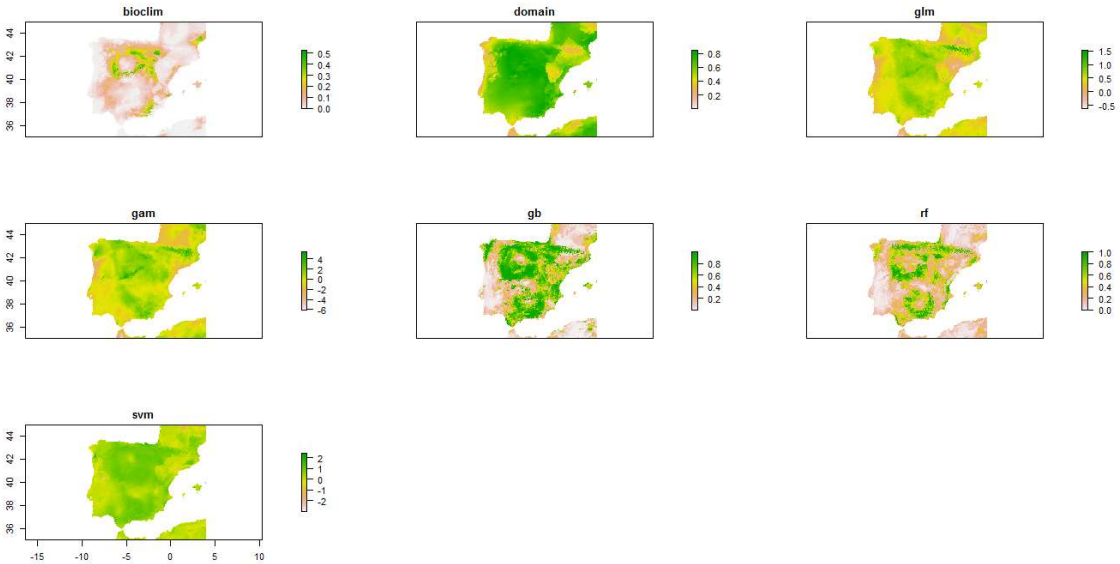


Representación gráfica del promedio de todas las predicciones

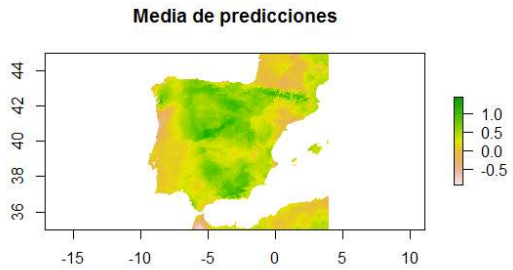


Resultados con todas las variables bioclimáticas

A continuación al igual que antes que muestran todos los de los modelos seleccionados de cada algoritmo incluyendo sólo las variables bioclimáticas preseleccionadas.



Representación gráfica del promedio de todas las predicciones



Tras la combinación de modelos se puede ver como las zonas con un puntuación mayor en la predicción se concentran cerca de los datos reales y sobre todo en la zona norte del país, demostrando la posible habitabilidad en zonas con temperaturas más fríos por los Aphodius.

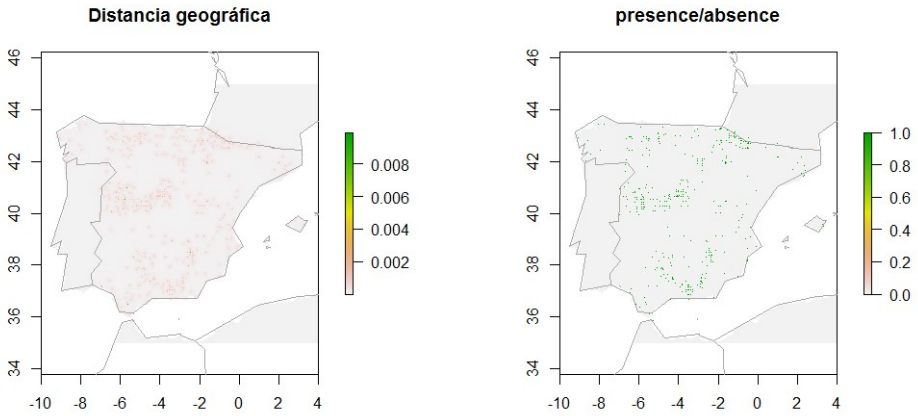
Modelos geográficos

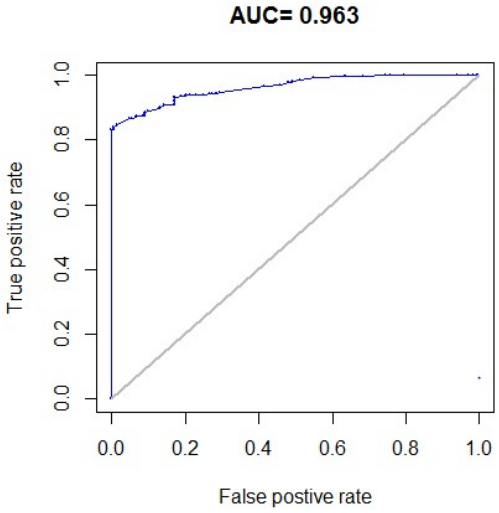
Estos modelos utilizan la localización geográfica de las ocurrencias conocidas, no se basan en las variables bioclimáticas de esas zonas. A continuación veremos unos cuantos algoritmos geográficos.

Distancia geográfica

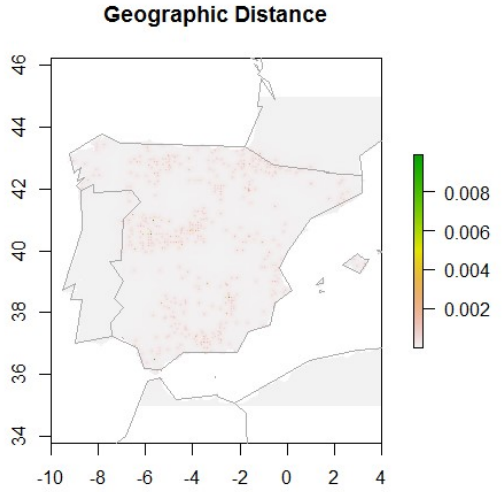
Modelo sencillo que presupone que cuanto más cerca de un punto de presencia conocido, es más probable encontrar la especie. La distancia geográfica comúnmente se mide con base en la distancia euclidiana —distancia lineal entre 2 puntos—, aunque existen también múltiples algoritmos diferentes que pueden aplicarse.

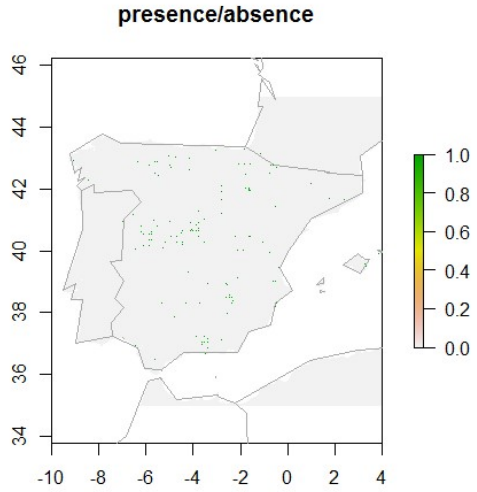
Resultados con las variables preseleccionadas

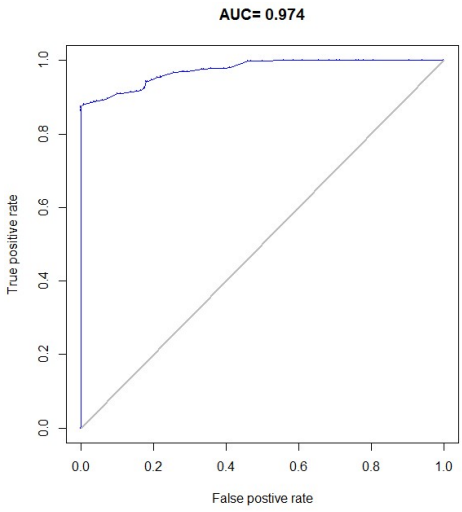




Resultados con todas las variables bioclimáticas



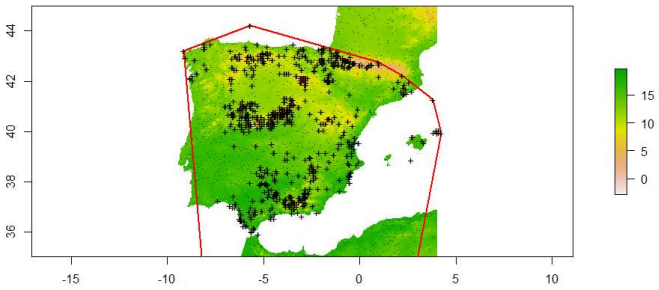




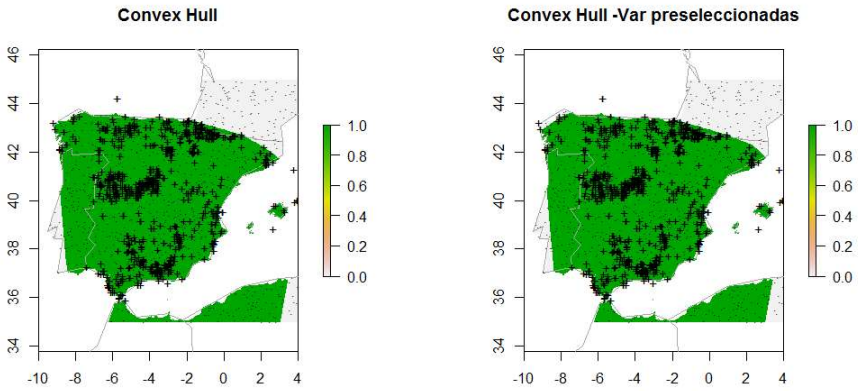
Rango convexo (Convex hull)

Este modelo dibuja un rango o celda convexa alrededor de todos los puntos de presencia de Aphodius.

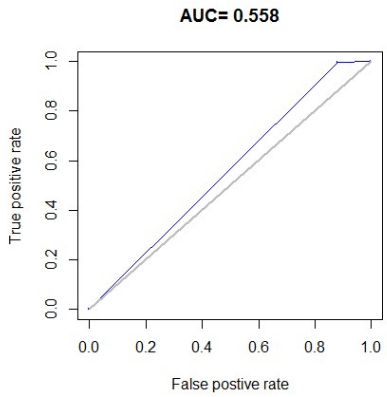
Aquellos puntos fuera de la celda convexa se consideran ausencias.



En función de la celda covexa se obtiene las predicciones, a continuación se puede ver el gráfico de la distribución predicha con este algoritmo.



Para este tipo de modelos, como se puede ver, no es de mucha utilidad hacer una preselección de variables predictoras previamente ya que dibuja una región convexa alrededor de todos los puntos de presencia sin tener en cuenta las variables bioclimáticas.



Presencia/ausencia

IDW - Inverse Distance Weight

La herramienta IDW (Ponderación de distancia inversa) utiliza un método de interpolación que estima los valores de las celdas asignando pesos a los datos del entorno en función inversa de la distancia que los separa. Cuanto más cerca está un punto del centro de la celda que se está estimando, más influencia o peso tendrá en el proceso de cálculo del promedio.

La fórmula general es:

$$\hat{z}_j = \sum_{i=1}^{n} k_{ij} \cdot z_i$$

Donde:

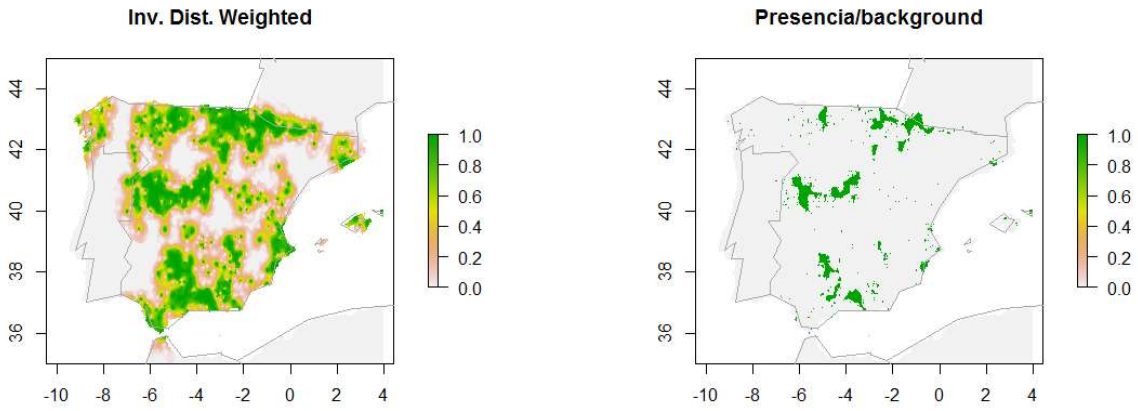
$\hat{z}_j$ es el valor estimado para el punto j ;

n es el número de puntos usados en la interpolación;

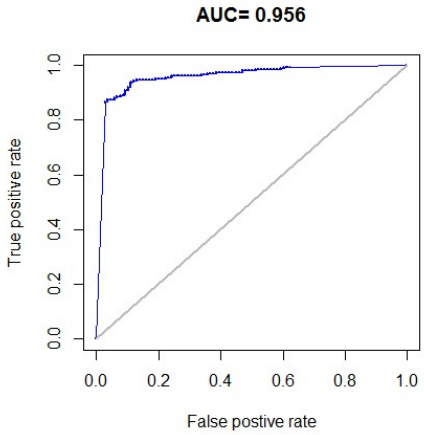
zi el valor en el punto i-ésimo; y

kij el peso asociado al dato i en el cálculo del nodo j. Los pesos varían entre 0 y 1 para cada dato y la suma total de ellos es la unidad.

Resultados con las variables preseleccionadas

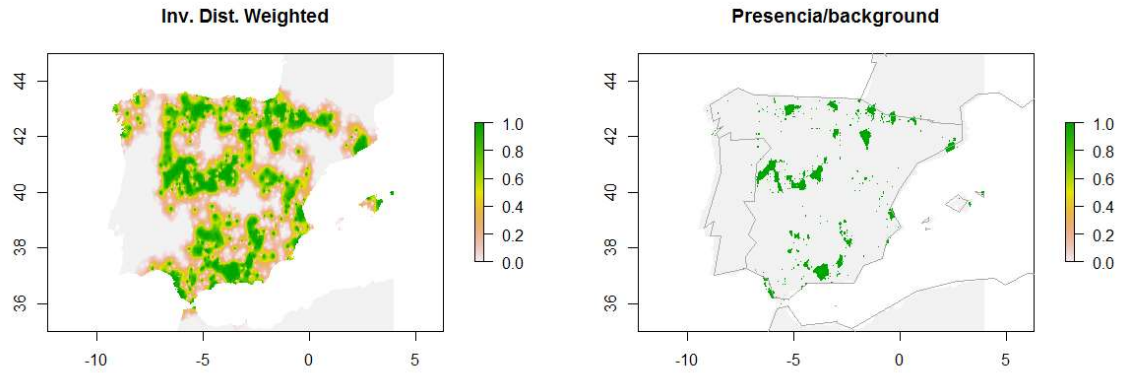


Evaluación del modelo

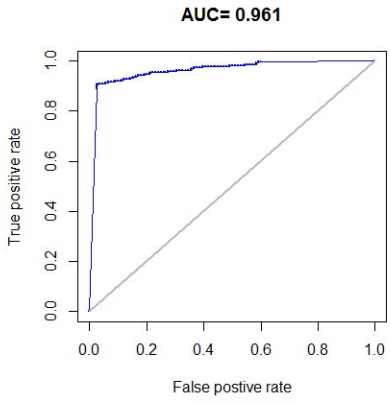


Como se aprecia el AUC obtenido con este algoritmo geográfico es muy elevado.

Resultados con todas las variables bioclimáticas



Evaluación del modelo



Comparación de modelos

Los dos planteamientos obtienen valores similares, por lo que siguiendo el criterio de parsimonia seleccionaríamos en el caso de tener que elegir entre uno de estos dos modelos por el de menor complejidad.

Modelos Bayesianos

Naïves Bayes

Este algoritmo asume que la presencia o ausencia de una especie no está relacionada con la presencia o ausencia de otra especie. Predice la probabilidad de posibles resultados.

Una ventaja de utilizar Naïves bayes para clasificar es que solo requiere una pequeña cantidad de datos de entrenamiento para estimar los parámetros (las medias y las varianzas de las variables) necesarias para la clasificación. Asumiendo independencia de las variables, solo es necesario determinar las varianzas de las variables de cada clase y no de toda la matriz de covarianza.

Se basa en el teorema de Bayes, o probabilidad condicionada:

$$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}$$

Si generalizamos la ecuación a más de un caso, se obtiene la siguiente expresión:

$$P(A|b_1,b_2,\dots,b_n)=P(A)\cdot P(b_1,b_2,\dots,b_n|A)=P(A)\cdot\prod_{i=1}^{n}P(a_i|A)$$

Para asignar a una especie (datos test) la probabilidad de presencia o ausencia, se calcula tantas probabilidades condicionadas como variables ambientales (condición) haya, y se le asigna el estado de presencia o ausencia tomando la probabilidad mayor.

$$Solucion=arg\max_{i=1}^{n}P(c_i)\cdot\prod_{j=1}^{m}P(a_j|c_i)$$

Resultados con las variables preseleccionadas

Probabilidades condicionas

Los resultados que se muestrasn son la probabilidades a posteriori para la clasificación que hace R sobre los datos test de presencia y background basándose en las características bioclimáticas.

Tabla 21. Probabilidades a posteriori Naïves Bayes con las var. presleccionadas

	Bio1_15		Bio2_15		Bio3_15		Bio8_15	
	1	2	1	2	1	2	1	2
0	13,28	2,76	10,19	1,39	3,77	0,28	28,5	3,7
1	11,31	3,3	10,6	1,18	3,73	0,17	27,48	3,69

	Bio9_15		Bio10_15		Bio11_15		Bio12_15	
	1	2	1	2	1	2	1	2
0	19,16	6,01	20,9	3,02	6,28	2,91	66,46	26,12
1	18,91	4,12	19,33	3,36	4,19	3,27	63,87	22,66

	Bio14_15		Bio16_15		Bio18_15	
	1	2	1	2	1	2
0	2,11	1,91	23,42	9,79	9,42	6,63
1	2,27	1,76	21,34	7,45	9,39	5,72

Fuente: Elaboración propia

Resultados con todas las variables bioclimáticas

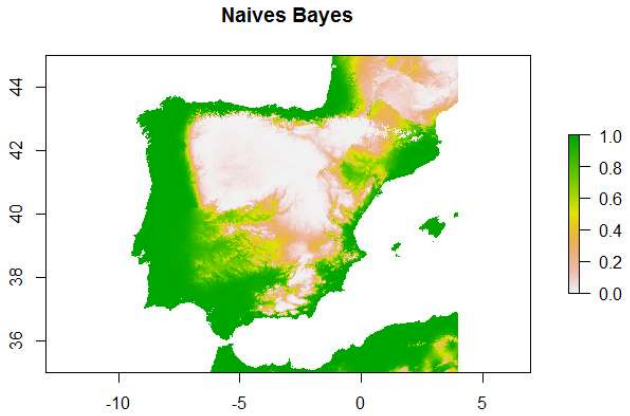
Probabilidades condicionas

Tabla 22. Probabilidades a posteriori Naïves Bayes con todas las variables

	Bio1_15		Bio2_15		Bio3_15		Bio4_15		Bio5_15	
	1	2	1	2	1	2	1	2	1	2
0	13,56	2,69	10,09	1,42	3,77	0,29	569,11	82,09	28,61	3,58
1	11,2	3,3	10,54	1,2	3,73	0,16	596,45	62,12	27,4	3,71
	Bio6_15		**Bio7_15**		**Bio8_15**		**Bio9_15**		**Bio10_15**	
	1	2	1	2	1	2	1	2	1	2
0	2,11	2,95	26,49	3,63	10,28	3,13	19,44	5,93	21,13	2,94
1	-0,44	3,24	17,84	3,13	8,61	3,62	18,83	4,08	19,23	3,37
	Bio11_15		**Bio12_15**		**Bio13_15**		**Bio14_15**		**Bio15_15**	
	1	2	1	2	1	2	1	2	1	2
0	6,22	2,88	66,63	24,89	8,72	3,41	1,96	1,82	3,93	1,66
1	4,08	3,26	64,37	22,3	8,02	2,65	2,31	1,76	3,38	1,32
	Bio16_15		**Bio17_15**		**Bio18_15**		**Bio19_15**			
	1	2	1	2	1	2	1	2		
0	23,88	9,59	8,27	5,96	9,12	6,42	21,15	10,47		
1	21,47	7,36	9,05	5,68	9,45	5,76	18,68	7,69		

Fuente: Elaboración propia

Representación gráfica de la distribución estimada



EVALUACIÓN DE LOS MODELOS

Para la evaluación de los modelos predictivos lo idóneo es realizar varios estadísticos y contratar sus resultados. En este caso se utilizarán cuatro índices utilizados por (Kazuya Naoki, M. Isabel Gómez, Ramiro P. López, Rosa I. Meneses & Julieta Vargas, 2006) para evaluar el desempeño de cada modelo de distribución: *sensitividad, especificidad, poder predictivo positivo (PPP), el índice Kappa*, además del *AUC* obtenido en cada modelo.

Estos cuatro índices fueron calculados tras obtener la matriz de confusión de cada modelo.

$$Sensitividad \quad \frac{a}{a+c}$$

$$Especificidad = \frac{d}{b+d}$$

$$PPP = \frac{a}{a+b}$$

$$Kappa = \frac{((a+b) - \frac{(a+c)(a+b)+(b+d)(c+d))}{N}}{N - ((a+c)(a+b)+(b+d)(c+d)}$$

Donde, *a* es el número de registros presentes correctamente predichos como presentes,
b es el número de registros ausentes incorrectamente predichos como presentes,
c es el número de registros presentes incorrectamente predichos como ausentes,
d es el número de registros ausentes correctamente predichos como ausentes, y

N es el número total de observaciones = $a + b + c + d$.

La *sensitividad* es la proporción de presencias correctamente predichas y un alto valor indica un bajo error (error tipo I[6]). La *especificidad* es la proporción de ausencias correctamente predichas y su valor alto indica un bajo error de comisión (error tipo II[7]). El *PPP* es la proporción de presencias correctamente predichas con relación a todas las localidades, donde la presencia de la especie fue predicha (Parra et al. 2004). *Kappa* es un índice utilizado para medir la precisión de la predicción en relación a la predicción al azar. Un valor alto de Kappa según (Fielding & Bell 1997) indica que la predicción tiene tanto el error tipo I y tipo II bajos.

Tabla 23. Escala valoración índide Kappa

kappa	grado de acuerdo
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

Estos cuatro índices oscilan entre 0 y 1, cuanto más cerca de 1 esté el valor, significa un mejor funcionamiento del modelo. Para calcular estos índices, se realizó un remuestreo de datos. Para eso, se escogieron aleatoriamente aproximadamente el 70% de las observaciones. Este 70% fue utilizado para generar una distribución con datos training, y el 30% restante considerado como datos test fue usado para comprobar la idoneidad/precisión de la predicción de cada modelo, por tanto los índices anteriores se realizan sobre los datos test.

Se comparan por tanto los modelos obtenidos con las ***variables preseleccionadas***, ya que los resultados son parecidos en alguna ocasión a los modelos que incluyen todas las variables bioclimáticas y en la mayoría de los casos son superiores su capacidad predictiva, además de ser modelos menos complejos.

[6] Error que se comete cuando se rechaza la hipótesis nula (H_o) siendo esta verdadera
[7] Error que se comete cuando se acepta la hipótesis nula (Ho) siendo esta falsa

Tabla 24. Índices para la evaluación de los modelos con var. preseleccionadas

Input	Modelo	Sensitividad	Especificidad	PPP	Kappa	AUC
Variables de presencia y bioclimáticas	**BIOCLIM**	0.526	0.671	0.787	0.161	0.691
	DOMAIN	0.621	0.555	0.762	0.159	0.686
	MAHALANOBIS	0.985	0.525	0.828	0.582	0.99
Variables presencia, ausencia y bioclimáticas	**GLM**	0.534	0.853	0.705	0.441	0.718
	GAM	0.534	0.853	0.705	0.441	0.718
	GRADIENT BOOSTING	0.782	0.954	0.901	0.760	0.921
	RANDOM FOREST	0.730	0.847	0.593	0.535	0.836
	SVM	0.425	0.957	0.809	0.442	0.746
	NAÏVES BAYES	0.615	0.689	0.462	0.278	
Variables geográficas (longitud, latitud)	**DISTANCIA GEOGRÁFICA**	0.978	0.175	0.737	0.198	0.963
	CONVEX HULL	0.746	0.310	0.719	0.058	0.558
	PRESENCIA/AUSENCIA	0.943	0.715	0.887	0.687	0.956

Fuente: Elaboración propia

Estos modelos dependen de un umbral (punto de corte), para determinar el mejor umbral para cada predicción, se escogió aquel umbral que fuese mayor que la sensitividad y de esta manera obtener el máximo valor de Kappa.

Los valores de los cuatro índices varían entre 0 y 1; por tanto, tienden a seguir la distribución binomial (Sokal & Rohlf 1995). Para normalizar los datos, todos los valores de los índices fueron transformados a arcoseno de raíz cuadrada antes de realizar los análisis estadísticos.

Para comparar los valores de índices, se aplicó una ANOVA bifactorial considerando los modelos de distribución y especies como los factores. Cuando se encontró una diferencia significativa en uno de estos factores, se aplicó una prueba a posteriorí de Tukey para realizar las comparaciones pareadas. Este análisis no se ha realizado para aquellas especies con muy pocas observaciones ya que las predicciones no son exactas o no se pueden calcular por ser matrices singulares o no invertibles. Por tanto, se compara los índices de las especies con mayor representatividad, *Aphodius alpinus, Aphodius coniugatus, Aphodius fimetarius y Aphodius foetidus.*

Tabla 25. Resultados del ANOVA entre modelos de distribución y especies

Factor	g.l.	F	P_valor	g.l.	F	P_valor
	Sensitividad			Especificidad		
Modelo	10	100,614	<0.001	10	202,594	<0.001
Especie	3	19,898	<0.001	3	30,000	<0.001
Modelo*Especie	20	6,050	<0.001	20	42,692	<0.001
	PPP			Kappa		
Modelo	10	104,133	<0.001	10	201,748	<0.001
Especie	2	864,829	<0.001	3	1270,165	<0.001
Modelo*Especie	20	48,721	<0.001	20	146,393	<0.001

Fuente: Elaboración propia

Se aprecian diferencias significativas en los cuatro índices entre los cuatro modelos de distribución (P_valor < 0.001) y entre las especies seleccionadas (P < 0.01), al igual que en las interacciones entre los modelos y las especies (P < 0.01).

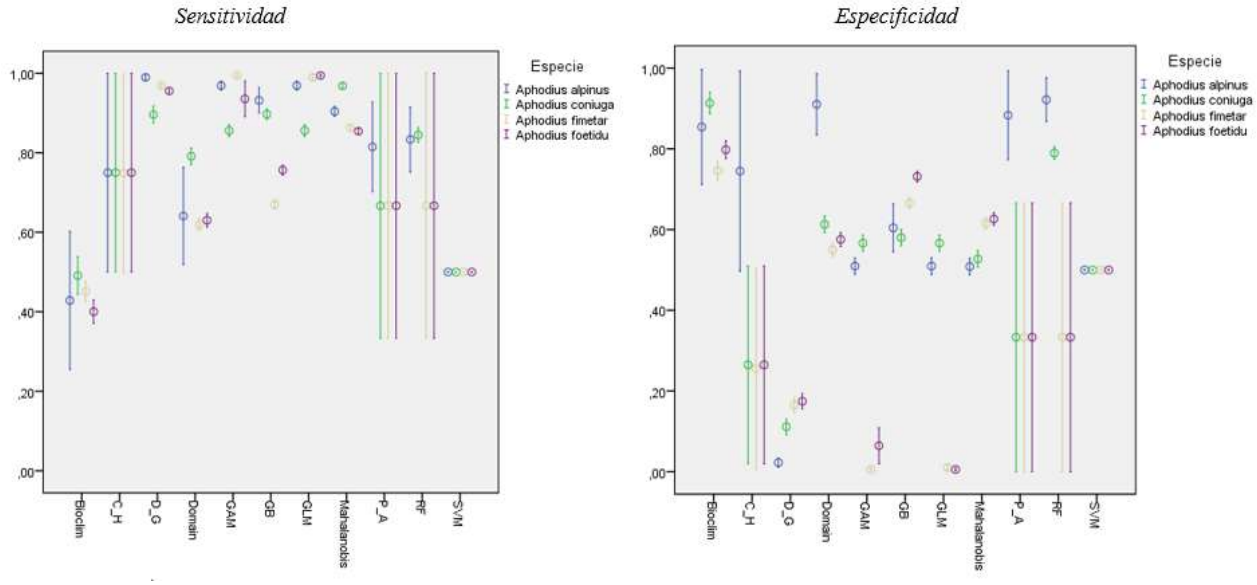
En relación a la *sensitividad*, BIOCLIM ha mostrado diferencias significativas con todos los demás modelos excepto con SVM (p_valor > 0.05). DOMAIN tiene diferencias significativas con todos los modelos salvo con Convent Hull (C_H), Presencia ausencia (P_A), Random Forest (RF) y Support vector machine (SVM). El algoritmo de Convent Hull tiene diferencias significativas con BIOCLIM (p_valor <0.05), GAM y GLM, con el resto de modelos el p_valor es mayor de 0.05. Distancia geográfica (D_G) es estadísticamente significativo con todos los modelos salvo con los modelos lineales, GLM y GAM (p_valor > 0.05). El algoritmo GAM es significativo con todos los modelos salvo con GLM y D_G (p_valor > 0.05). Gradiente Boosting (GB) es significativo con todos los modelos salvo con C_H, P_A, RF y SVM (p_valor > 0,05). El algoritmo GLM presenta significatividad con todos los modelos excepto con D_G y GAM (p_valor > 0,05). Mahalanobis no es significativo con los modelos C_H. P_A y RF (p_valor > 0,05) con el resto si tiene significatividad. Random Forest tan solo es significativo con BIOCLIM, D_G, GAM y GLM (p_valor < 0,05) con el resto de modelos las medias son diferentes. El algoritmo SVM es significativo con D_G, GAM, GLB y Mahalanobis (p_valor < 0,05). De todos estos modelos viendo el gráfico 10 y de carácter general se puede decir que BIOCLIM es el que peor rendimiento muestra, y los que mejor con respecto a la sensitividad son D_G, GAM, GLM y Mahalanobis, este último en menor medida.

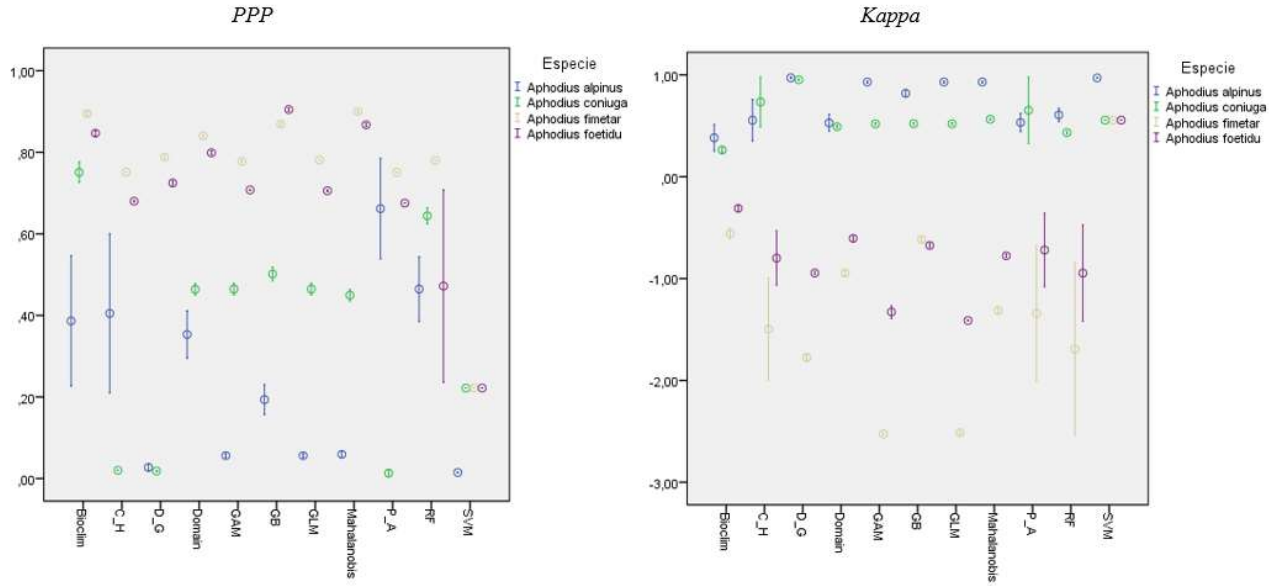
En cuanto a la *especificidad*, BIOCLIM es significativo con todos los modelos salvo con P_A, RF y SVM (p_valor > 0,05). DOMAIN es significativo con todos los modelos excepto con C_H, Mahalanobis, P_A y SVM (p_valor > 0,05). El algoritmo C_H no es significativo con los modelos DOMAIN, GAM, Mahalanobis, P_A y SVM (p_valor > 0,05) con el resto si guarda significatividad. El algoritmo D_G no es significativo con GAM, GLM y SCM (p_valor > 0,05), con el resto de modelos si hay significatividad. GAM no es significativo con los modelos C_H, D_G, GLM y SVM (p_valor > 0,05). El algoritmo GB es significativo con todos los modelos excepto con P_A, RF y SVM (p_valor > 0,05). GLM no es significativo con D_G, GAM y SVM (p_valor > 0,05), con el resto de modelos si presenta significatividad. Mahalanobis es significativo con todos los modelos salvo con C_H, DOMAIN, P_A y SVM (p_valor > 0,05). El algoritmo P_A solo tiene significatividad con los modelos D_G, GAM y GLM (p_valor < 0,05). Random Forest tiene significatividad con todos los modelos excepto con BIOCLIM, GB, P_A y SVM (p_valor > 0,05). SVM no es significativo con ningún modelo. Viendo el gráfico 10 por lo general se puede decir en términos de la especificidad que el modelo que peor rendimiento da son D_G, GAM y GLM, y el que mejor BIOCLIM junto con GB.

Con respecto al *PPP*, el algoritmo BIOCLIM es significativo con todos los modelos excepto con GB (p_valor > 0,05). C_H es significativo con BIOCLIM, DOMAIN, GB y SVM (p_valor < 0,05), con el resto no tiene significatividad. El algoritmo D_G muestra significatividad con todos los modelos excepto con los otros modelos geográficos, P_A y C_H (p_valor > 0,05). DOMAIN es significativo con todos los modelos (p_valor < 0,05). Los modelos GAM y GLM tienen significatividad con todos los modelos salvo con C_H y RF (p_valor > 0,05). El algoritmo GB es significativo con todos los modelos salvo con BIOCLIM. Mahalanobis es significativo con todos los modelos salvo con C_H y P_A (p_valor > 0,05). P_A es significativo con todos los modelos salvo con C_H, D_G, DOMAIN y Mahalanobis (p_valor > 0,05). RF es significativo con todos los modelos excepto con C_H, GAM y GLM (p_valor > 0,05). SVM es significativo con todos los modelos. En esta ocasión, viendo el gráfico 10 el modelo que mejor rendimiento tiene es Gradient Boosting.

Para el índice *Kappa,* BIOCLIM es significativo para todas las variables excepto para C_H y GB (p_valor > 0,05). El algoritmo C_H sólo es significativo con los modelos GAM, GLM, RF y SVM (p_valor < 0,05), con el resto de modelos no hay significatividad. D_G es significativo con todos los modelos excepto con C_H y DOMAIN (p_valor > 0,05). El algoritmo DOMAIN es significativo con todos los modelos salvo con D_G y C_H (p_valor > 0,05). GAM y GLM son significativos con todos los modelos. GB no es significativo con BIOCLIM, C_H y Mahalanobis (p_valor > 0,05). Mahalanobis y P_A son significativos con todos los modelos salvo con C_H (p_valor > 0,05). RF es significativo con todos los modelos salvo con SVM y viceversa (p_valor > 0,05). Para este índice los valores más altos como se pueden ver en el gráfico 10 se corresponden a modelos como RF y SVM. Es de destacar que para este índice los valores altos son para las especies *alpinus* y *coniugnatus*, mientras que *fimetarius* y *foetidus* tienen los valores más bajos.

Gráfico 10. IC-95% del error medio por índice y por especie





DISTRIBUCIÓN POTENCIAL EN EL AÑO 2070

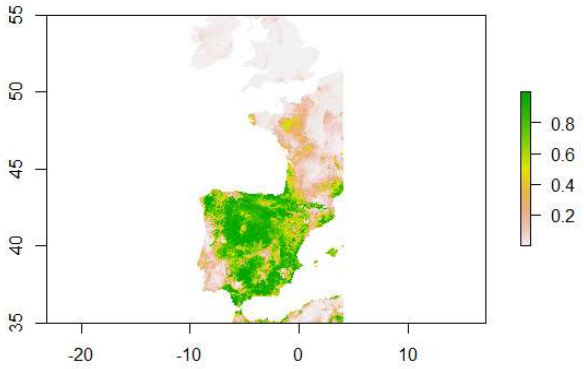
Como ya se comentó en introducción del documento, wordclim tiene en su base de datos, una predicción de las variables bioclimáticas para el año 2070, para generar estas capas/predicciones combinan información sobre cambio climático proveniente de modelos de circulación global (AOGCMs (atmosphere-ocean coupled general circulation models), modelos físicos sobre dinámica climática) y la capa para los datos del presente (proveniente de una interpolación).

A continuación se muestra el resultado obtenido aplicando el mejor modelo visto anteriormente (Gradient Boosting) utilizando los datos de las variables bioclimáticas estimadas para el año 2070 por worldclim.

Tabla 26. Matriz de confusión GB con datos del 2070

	Presencia	**Ausencia**
Presencia	355	119
Ausencia	54	145

Fuente: Elaboración propia



Sorprendentemente la distrbución potencial en el año 2070 es bastante amplia, pese a los cambios que seguramente sufra el país hasta ese año a consecuencia del calentamiento global/cambio climático. Este resultado se puede utilizar como un indicador de fortaleza de esta especie ante condiciones ambientales adversas. Sin embargo, estos resultados pueden ser engañosos siempre y cuando las especies de las cuales estan ligadas en la cadena trófica demuestren un fortaleza similar.

CONCLUSIONES

A través de este estudio se ha podido comprobar la eficacia de los modelos estadísticos para predecir la distribución potencial de los Aphodius, distribución que según los modelos puede basarse en datos geográficos, de presencia o datos ambientales. La utilización de unos u otros depende en gran medida de la toma de datos durante el muestreo, por eso para conocer la distribución geográfica real o potencial de un grupo taxonómico se ha comprobado que lo mejor es recopilar la información taxonómica, corológica y ambiental de todas las colecciones posibles, museos, así como de bibliografía, con el fin de tener una delimitación de las especies.

Ha quedado demostrado que una selección previa de las variables en la mayoría de los modelos estimados presentan mejores resultados que en el caso de incluir todas las variables bioclimáticas como predictoras. Teniendo esto encuenta, las conclusions se basan en los resultados obtenidos con los modelos de variables preseleccionadas.

A día de hoy no hay un algortimo que sea definitivo ni sea mejor que los demás, la elección de un modelo es complicado y depende mucho de la naturaleza de lo datos, hay modelos que funcionan mejor con pocos datos y otros que funcionan mejor con una

mayor cantidad de datos, teniendo en cuenta esto y por darle un significado a todo este trabajo seleccionaremos las distribuciones estimadas por Gradient Boosting o Presencia/ausencia como las mejores. Estos dos algoritmos muestran mejores resultados que el resto, un AUC bastante bueno y una sensitividad elevada hace de estos modelos unos buenos predictores de la posible presencia de Aphodius, uno basado en zonas con características ambientales idóneas para su localización o desplazamiento y otro basado en interpolación de pesos de datos presenciales.

A la hora de tener que elegir entre estos dos modelos elegiremos Gradient Boosting por ser un modelo que no está acotado por decirlo de una manera exclusivamente a la geolocalización de las especies, se basa en la utilización de otras predictoras a parte de la localización como otros modelos que hemos visto, pero con peores resultados.

Los peores resultados se han obtenido con BIOCLIM, sensitividad muy baja y un AUC no muy elevado corroboran lo dicho anteriormente, es un modelo muy utilizado en trabajos similares aunque los resultados no son muy buenos. Otro algoritmo en el que habría que profundizar es Mahalanobis, unos resultados muy buenos con las variables preseleccionadas y bastante “indecisos” si se incluyen todas las variables. Este modelo identifica los pixeles que pertenecen a una envoltura ambiental elíptica, dada la correlación entre variables, es posible que la autocorrelación entre variables de un resultado engañoso, como se ha dicho antes, habría que analizar esto con más detenimiento.

Gracias a las representaciones gráficas de la distribución potencial se puede ver como gran parte de las especies de Aphodius serían capaces de adaptarse a zonas del norte del país donde las condiciones ambientales son más adversas. Este tipo de conclusiones habría que contrastarla analizando para cada especie los modelos vistos y añadiendo variables tales como la altitud.

En cuanto a los principales factores bioclimáticos causales que han propiciado la actual distribución de los Aphodius se ha podido comprobar que en un alto grado la presencia esta correlacionada negativamente con casi todas las variables bioclimáticas y sólo tienen una correlación positiva con la estacionalidad de temperatura, y el rango de temperatura media, parece que hay evidencia de que la temperatura de la zona es el principal factor en la distribución.

Como resumen se muestran las representaciones obtenidas por Gradient Bossting para todas las variables bioclimáticas, las variables preseleccionadas y las variables del año 2070.

Tabla 27. GB con variables preseleccionadas

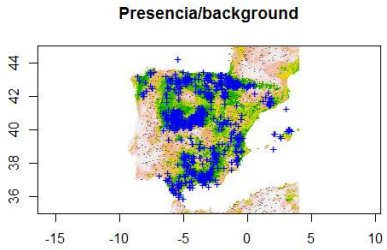


Tabla 28. GB con todas las variables bioclimáticas (presente)

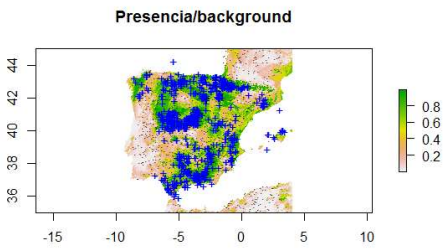
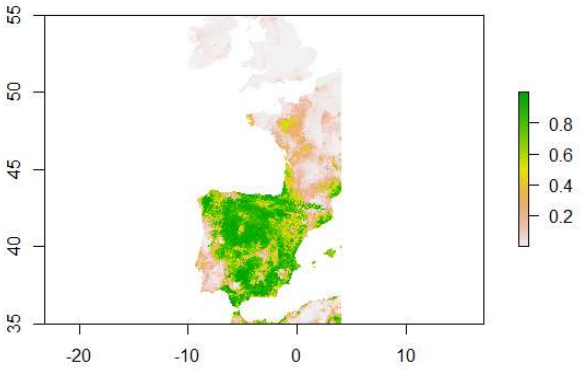


Tabla 29. GB con todas las variables bioclimáticas (2070)



Los resultados obtenidos en este estudio deben considerarse como una aportación más a los diversos trabajos existentes en este ámbito, por lo tanto es una aproximación que se debe validar y mejorar en acciones futuras.

ANEXO

Matriz de Correlaciones

		Presencia_ausencia
Presencia_ausencia	Correlación de Pearson	1
	Sig. (bilateral)	
	N	2808
Bio1	Correlación de Pearson	-,254**
	Sig. (bilateral)	,000
	N	2808
Bio2	Correlación de Pearson	-,011
	Sig. (bilateral)	,565
	N	2808
Bio3	Correlación de Pearson	-,259**
	Sig. (bilateral)	,000
	N	2808
Bio4	Correlación de Pearson	,152**
	Sig. (bilateral)	,000
	N	2808
Bio5	Correlación de Pearson	-,105**
	Sig. (bilateral)	,000
	N	2808
Bio6	Correlación de Pearson	-,247**
	Sig. (bilateral)	,000
	N	2808
Bio7	Correlación de Pearson	,082**
	Sig. (bilateral)	,000
	N	2808
Bio8	Correlación de Pearson	-,104**
	Sig. (bilateral)	,000
	N	2808
Bio9	Correlación de Pearson	-,058**
	Sig. (bilateral)	,002
	N	2808
Bio10	Correlación de Pearson	-,182**
	Sig. (bilateral)	,000
	N	2808
Bio11	Correlación de Pearson	-,287**
	Sig. (bilateral)	,000
	N	2808
Bio12	Correlación de Pearson	,006
	Sig. (bilateral)	,770
	N	2808
Bio13	Correlación de Pearson	-,008
	Sig. (bilateral)	,686
	N	2808
Bio14	Correlación de Pearson	-,018
	Sig. (bilateral)	,329
	N	2808
Bio15	Correlación de Pearson	-,157**
	Sig. (bilateral)	,000
	N	2808
Bio16	Correlación de Pearson	-,045*
	Sig. (bilateral)	,016
	N	2808
Bio17	Correlación de Pearson	,023

	Sig. (bilateral)	,228
	N	2808
Bio18	Correlación de Pearson	-,013
	Sig. (bilateral)	,481
	N	2808
Bio19	Correlación de Pearson	-,072**
	Sig. (bilateral)	,000
	N	2808

Modelos estimados para la selección de variables

Regresión logística - Stepwise

```
Proc logistic data=aphodius.aphodius;
class;
model pbs=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15/
selection=stepwise;
Run;
```

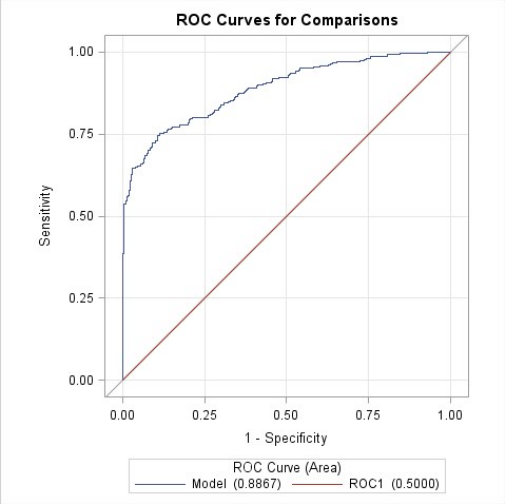
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.5253	1.1422	1.7833	0.1817
bio2_15	1	0.0331	0.0159	4.3257	0.0375
bio6_15	1	-0.1116	0.0265	17.6951	<.0001
bio9_15	1	-0.0196	0.00340	33.1413	<.0001
bio11_15	1	0.1665	0.0257	41.8980	<.0001
bio12_15	1	-0.0389	0.00434	80.4523	<.0001
bio13_15	1	-0.1849	0.0207	79.6175	<.0001
bio14_15	1	0.2235	0.0347	41.5353	<.0001
bio15_15	1	-0.1342	0.0206	42.3395	<.0001
bio16_15	1	0.0896	0.0114	61.8609	<.0001
bio17_15	1	-0.0880	0.0167	27.9119	<.0001
bio18_15	1	0.0708	0.0120	34.9741	<.0001
bio19_15	1	0.0675	0.00734	84.6149	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
bio2_15	1.034	1.002	1.066
bio6_15	0.894	0.849	0.942
bio9_15	0.981	0.974	0.987
bio11_15	1.181	1.123	1.242
bio12_15	0.962	0.954	0.970
bio13_15	0.831	0.798	0.866
bio14_15	1.250	1.168	1.338
bio15_15	0.874	0.840	0.910
bio16_15	1.094	1.070	1.118
bio17_15	0.916	0.886	0.946
bio18_15	1.073	1.048	1.099
bio19_15	1.070	1.055	1.085

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	88.6	**Somers' D**	0.774
Percent Discordant	11.2	**Gamma**	0.776
Percent Tied	0.3	**Tau-a**	0.227
Pairs	1154000	**c**	0.887



Las variables seleccionadas con este método son: bio2, bio6, bio9, bio11, bio12, bio13, bio14, bio15, bio16, bio17, bio18, bio19, estas variables como se aprecia en la tabla del análisis de máxima verosimilitud son bastante significativas, P_valor <0.001, y el AUC obtenido es bueno, sin embargo para no quedarnos con esta primera impresión, realizaremos una selección de variables ejecutando la macro randomselectlog con diversas semillas y criterios de selección para su posterior comparación con validación cruzada.

```
%randomselectlog(data=aphodius.aphodius,
listclass= ,
vardepen=pbs,
modelo=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15 bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15 ,
sinicio=12345,sfinal=12380,fracciontrain=0.8, directorio=L:);

bio6_15 bio9_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15 bio17_15 bio18_15
bio19_15

bio2_15 bio6_15 bio9_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15 bio17_15
bio18_15 bio19_15

bio1_15 bio2_15 bio4_15 bio6_15 bio9_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15 bio18_15 bio19_15
```

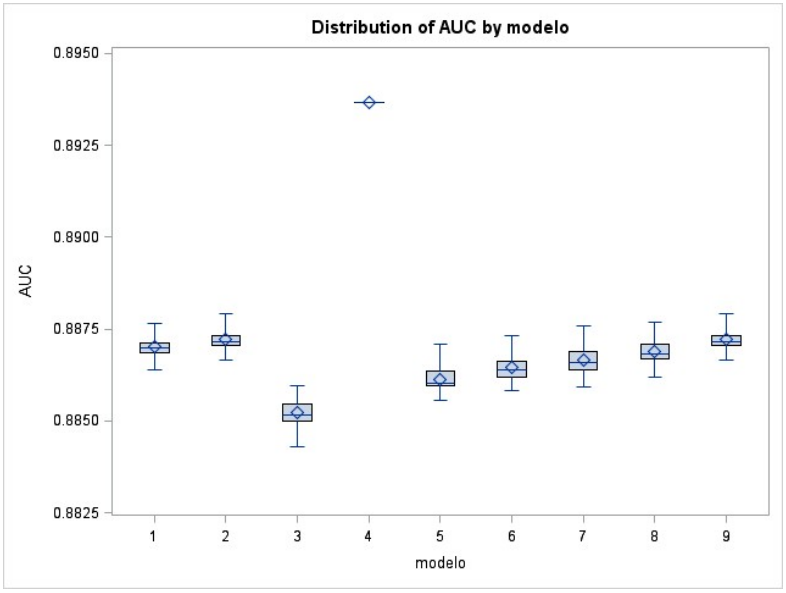
```
bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio9_15 bio12_15 bio13_15 bio14_15 bio15_15
bio16_15 bio17_15 bio18_15 bio19_15

bio1_15 bio2_15 bio3_15 bio5_15 bio9_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15
bio16_15 bio17_15 bio18_15 bio19_15

bio2_15 bio3_15 bio5_15 bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15
bio16_15 bio17_15 bio18_15 bio19_15

bio2_15 bio3_15 bio5_15 bio9_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15 bio18_15 bio19_15

bio4_15 bio6_15 bio9_15 bio10_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15 bio17_15
bio18_15 bio19_15

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio6_15 bio9_15 bio11_15 bio12_15 bio13_15 bio14_15
bio15_15 bio16_15 bio17_15 bio18_15 bio19_15,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final1;set final;modelo=1;

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio2_15 bio6_15 bio9_15 bio11_15 bio12_15
bio13_15 bio14_15 bio15_15 bio16_15 bio17_15 bio18_15 bio19_15,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final2;set final;modelo=2;

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio1_15 bio2_15 bio4_15 bio6_15 bio9_15 bio12_15 bio13_15
bio14_15 bio15_15 bio16_15 bio17_15 bio18_15 bio19_15,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final3;set final;modelo=3;

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio9_15 bio12_15
bio13_15 bio14_15 bio15_15 bio16_15 bio17_15 bio18_15 bio19_1,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final4;set final;modelo=4;

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio1_15 bio2_15 bio3_15 bio5_15 bio9_15 bio11_15 bio12_15
bio13_15 bio14_15 bio15_15 bio16_15 bio17_15 bio18_15 bio19_15,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final5;set final;modelo=5;

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio2_15 bio3_15 bio5_15 bio9_15 bio10_15 bio11_15 bio12_15
bio13_15 bio14_15 bio15_15 bio16_15 bio17_15 bio18_15 bio19_15,
```

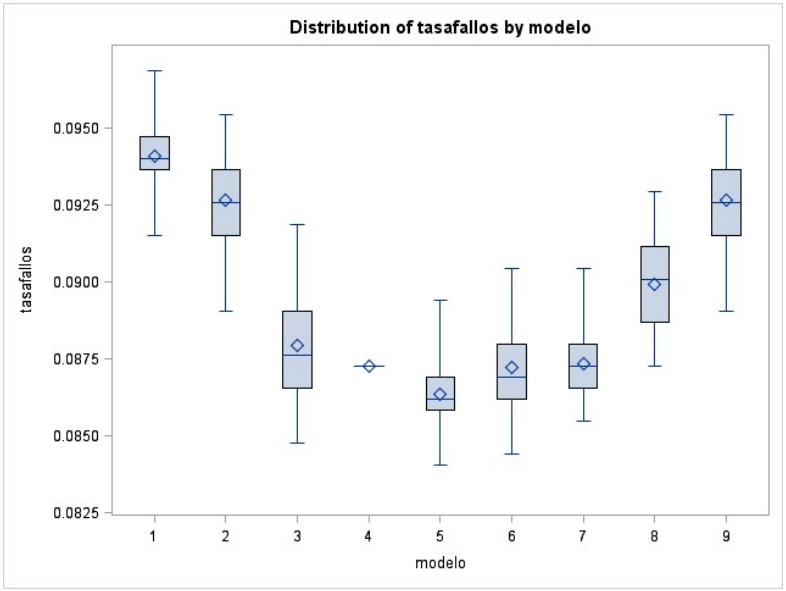
```
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final6;set final;modelo=6;

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio2_15 bio3_15 bio5_15 bio9_15 bio11_15 bio12_15 bio13_15
bio14_15 bio15_15 bio16_15 bio17_15 bio18_15 bio19_15,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final7;set final;modelo=7;

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio4_15 bio6_15 bio9_15 bio10_15 bio12_15 bio13_15 bio14_15
bio15_15 bio16_15 bio17_15 bio18_15 bio19_15,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final8;set final;modelo=8;

%cruzadalogistica
(archivo=aphodius.aphodius,vardepen=pbs,
conti=bio2_15 bio6_15 bio9_15 bio11_15 bio12_15 bio13_15
bio14_15 bio15_15 bio16_15 bio17_15 bio18_15 bio19_15,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final9;set final;modelo=9;


data union;set final1 final2 final3 final4 final5 final6 final7 final8
final9;
proc boxplot data=union;plot media*modelo;run;
```

En la tabla siguiente se puede ver el error medio obtenido en cada semilla. Esta media es la media del objetivo en todas las pruebas de validación cruzada (habitualmente tasa de fallos).

Obs	semilla	capturados	tasafallos	sensi	especif	Youden	AUC
1	12345	0.92500	0.090812	0.97618	0.59958	0.57576	0.88718
2	12346	0.90714	0.092593	0.97661	0.58873	0.56534	0.88714
3	12347	0.91250	0.095085	0.97312	0.58877	0.56189	0.88732
4	12348	0.90893	0.093661	0.97443	0.58906	0.56348	0.88687
5	12349	0.91607	0.094373	0.97531	0.58323	0.55854	0.88747
6	12350	0.94286	0.092236	0.97436	0.59987	0.57423	0.88717
7	12351	0.92857	0.093661	0.97530	0.58718	0.56248	0.88709
8	12352	0.91786	0.091168	0.97649	0.59471	0.57120	0.88705
9	12353	0.94286	0.094729	0.97488	0.58284	0.55773	0.88717
10	12354	0.91964	0.092949	0.97448	0.59366	0.56814	0.88676
11	12355	0.91786	0.092593	0.97529	0.59480	0.57008	0.88722
12	12356	0.91786	0.089031	0.97790	0.60231	0.58022	0.88779
13	12357	0.91607	0.091880	0.97538	0.59726	0.57265	0.88709
14	12358	0.93214	0.091168	0.97658	0.59557	0.57215	0.88732
15	12359	0.89643	0.092949	0.97486	0.59434	0.56920	0.88763
16	12360	0.90357	0.094729	0.97503	0.58339	0.55842	0.88687
17	12361	0.92500	0.094017	0.97449	0.59095	0.56544	0.88704
18	12362	0.90357	0.095442	0.97445	0.58401	0.55847	0.88666
19	12363	0.91964	0.091168	0.97575	0.59853	0.57428	0.88699
20	12364	0.91607	0.091880	0.97616	0.59184	0.56800	0.88687

21	12365	0.91250	0.094729	0.97497	0.58508	0.56005	0.88774
22	12366	0.91786	0.091880	0.97706	0.58944	0.56651	0.88710
23	12367	0.91250	0.093661	0.97481	0.58950	0.56431	0.88727
24	12368	0.91786	0.093305	0.97526	0.58814	0.56340	0.88713
25	12369	0.93750	0.090812	0.97700	0.59534	0.57234	0.88722
26	12370	0.92857	0.092593	0.97614	0.59022	0.56636	0.88760
27	12371	0.92679	0.091880	0.97578	0.59658	0.57236	0.88707
28	12372	0.90357	0.091880	0.97702	0.59124	0.56826	0.88722
29	12373	0.91429	0.090812	0.97528	0.60356	0.57885	0.88685
30	12374	0.92679	0.091524	0.97574	0.59920	0.57494	0.88790
31	12375	0.91786	0.092949	0.97529	0.58939	0.56467	0.88792





Árboles de decisión

```
proc arbor data=aphodius.aphodius criterion= probchisq;
```

```
input bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15 bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15 /level=interval;
input  /level=ordinal;
target pbs /level=binary;
interact largest;/* necesario para poder utilizar train */
train maxbranch=2 leafsize=15;/* maxbranch=máximas divisiones en cada
rama (2=arbol binario)
leafsize=numero de observaciones por hoja final, tamaño de la hoja
final */
assess measure=misc; /* misc para dependientes nominales o binarias,
ase para continuas */
subtree Best;/* número máximo de hojas finales del subárbol elegido:
BEST, LARGEST, nleaves=nhojas
(es bueno probar varios números de hojas finales)*/
score out=salprob;/* archivo de salida con predicciones y variables de
pertenencia a nodos */
describe file=print;/* describe la creación de nodos en la ventana
output */
save importance=importancia model=modelo;/* en el archivo importancia
viene la importancia de cada variable; el modelo se guarda
para futuros proc arbor si se desea */
run;
proc print data=importancia;run;
```

Obs	NAME	LABEL	NRULES	IMPORTANCE
1	bio2_15	bio2_15	2	1.00000
2	bio4_15	bio4_15	1	0.86482
3	bio5_15	bio5_15	3	0.70490
4	bio13_15	bio13_15	1	0.42215
5	bio16_15	bio16_15	1	0.35295
6	bio11_15	bio11_15	1	0.27728

```
proc arbor data=aphodius.aphodius criterion= probchisq;
input bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15 bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15 /level=interval;
input  /level=ordinal;
target pbs /level=binary;
interact largest;/* necesario para poder utilizar train */
train maxbranch=2 leafsize=25;/* maxbranch=máximas divisiones en cada
rama (2=arbol binario)
leafsize=numero de observaciones por hoja final, tamaño de la hoja
final */
assess measure=misc; /* misc para dependientes nominales o binarias,
ase para continuas */
subtree Best;/* número máximo de hojas finales del subárbol elegido:
BEST, LARGEST, nleaves=nhojas
(es bueno probar varios números de hojas finales)*/
score out=salprob;/* archivo de salida con predicciones y variables de
pertenencia a nodos */
describe file=print;/* describe la creación de nodos en la ventana
output */
save importance=importancia model=modelo;/* en el archivo importancia
viene la importancia de cada variable; el modelo se guarda
para futuros proc arbor si se desea */
run;
```

```
proc print data=importancia;run;
```

Obs	NAME	LABEL	NRULES	IMPORTANCE
1	bio2_15	bio2_15	2	1.00000
2	bio4_15	bio4_15	1	0.86482
3	bio5_15	bio5_15	2	0.63932
4	bio13_15	bio13_15	2	0.48142
5	bio7_15	bio7_15	1	0.33550
6	bio11_15	bio11_15	1	0.31036

```
proc arbor data=aphodius.aphodius criterion= probchisq;
input bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15 bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15 /level=interval;
input /level=ordinal;
target pbs /level=binary;
interact largest;/* necesario para poder utilizar train */
train maxbranch=2 leafsize=10;/* maxbranch=máximas divisiones en cada
rama (2=arbol binario)
leafsize=numero de observaciones por hoja final, tamaño de la hoja
final */
assess measure=misc; /* misc para dependientes nominales o binarias,
ase para continuas */
subtree Largest;/* número máximo de hojas finales del subárbol
elegido: BEST, LARGEST, nleaves=nhojas
(es bueno probar varios números de hojas finales)*/
score out=salprob;/* archivo de salida con predicciones y variables de
pertenencia a nodos */
describe file=print;/* describe la creación de nodos en la ventana
output */
save importance=importancia model=modelo;/* en el archivo importancia
viene la importancia de cada variable; el modelo se guarda
para futuros proc arbor si se desea */
run;
proc print data=importancia;run;
```

Obs	NAME	LABEL	NRULES	IMPORTANCE
1	bio2_15	bio2_15	2	1.00000
2	bio4_15	bio4_15	2	0.89125
3	bio5_15	bio5_15	3	0.71661
4	bio13_15	bio13_15	1	0.43131
5	bio16_15	bio16_15	1	0.36060
6	bio11_15	bio11_15	2	0.28519
7	bio17_15	bio17_15	1	0.21180
8	bio12_15	bio12_15	1	0.19947
9	bio3_15	bio3_15	1	0.15008
10	bio8_15	bio8_15	1	0.11400

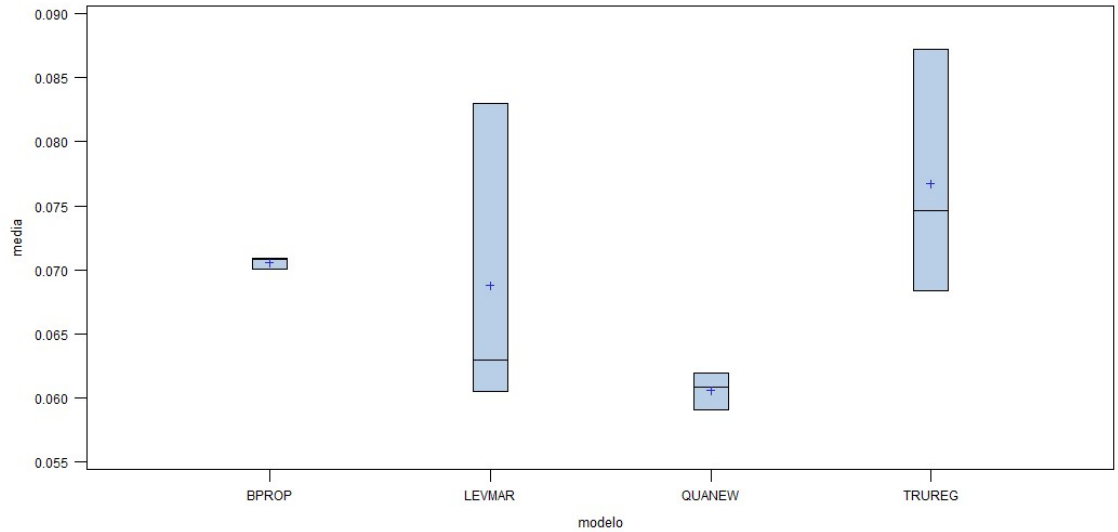
```
proc arbor data=aphodius.aphodius criterion=entropy;
input bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15 bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15 /level=interval;
input  /level=ordinal;
target pbs /level=binary;
interact largest;/* necesario para poder utilizar train */
train maxbranch=2 leafsize=5;/* maxbranch=máximas divisiones en cada
rama (2=arbol binario)
```

```
leafsize=numero de observaciones por hoja final, tamaño de la hoja
final */
assess measure=misc; /* misc para dependientes nominales o binarias,
ase para continuas */
subtree nleaves=30;/* número máximo de hojas finales del subárbol
elegido: BEST, LARGEST, nleaves=nhojas
(es bueno probar varios números de hojas finales)*/
score out=salprob;/* archivo de salida con predicciones y variables de
pertenencia a nodos */
describe file=print;/* describe la creación de nodos en la ventana
output */
save importance=importancia model=modelo;/* en el archivo importancia
viene la importancia de cada variable; el modelo se guarda
para futuros proc arbor si se desea */
run;
proc print data=importancia;run;
```

Obs	NAME	LABEL	NRULES	IMPORTANCE
1	bio2_15	bio2_15	4	1.00000
2	bio4_15	bio4_15	2	0.87595
3	bio5_15	bio5_15	3	0.70108
4	bio13_15	bio13_15	1	0.42881
5	bio16_15	bio16_15	1	0.32423
6	bio11_15	bio11_15	1	0.31736
7	bio17_15	bio17_15	1	0.26190
8	bio12_15	bio12_15	1	0.23786
9	bio3_15	bio3_15	2	0.16997
10	bio1_15	bio1_15	2	0.16347
11	bio8_15	bio8_15	1	0.15645
12	bio14_15	bio14_15	1	0.13647

Redes neuronales

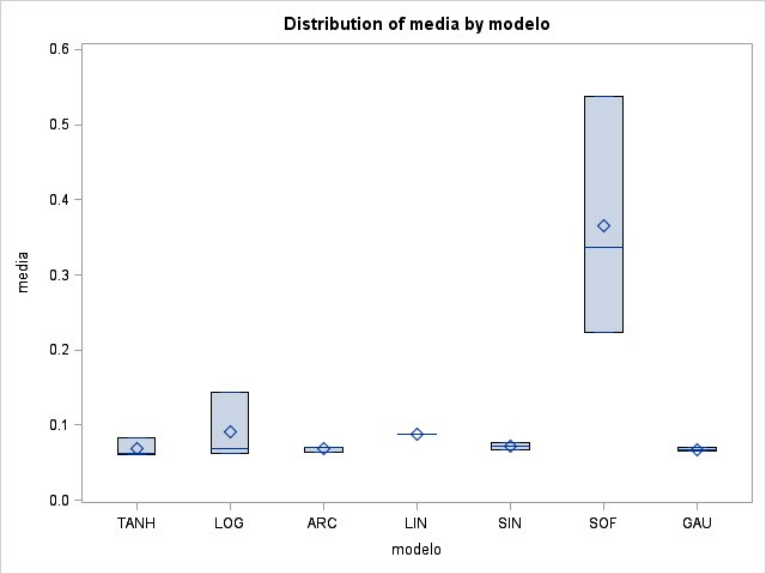
```
%macro algovalcruza;
%let lista='BPROP LEVMAR QUANEW TRUREG';
%let nume=4;
%do  i=1 %to &nume;
data _null_;meto=scanq(&lista,&i);call symput('meto',left(meto));run;
 %cruzadaneural(archivo=aphodius.aphodius,vardepen=pbs,conti=bio1_15
bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15 bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15
bio19_15,categor=,ngrupos=3,sinicio=12345,sfinal=12347,ocultos=10,meto
=&meto);
 data final&i;set final;modelo="&meto";put modelo=;run;
%end;
data union;set %do i=1 %to &nume; final&i %end;
%mend;

%algovalcruza;

proc print data=union;run;
proc boxplot data=union;plot media*modelo;run;
```



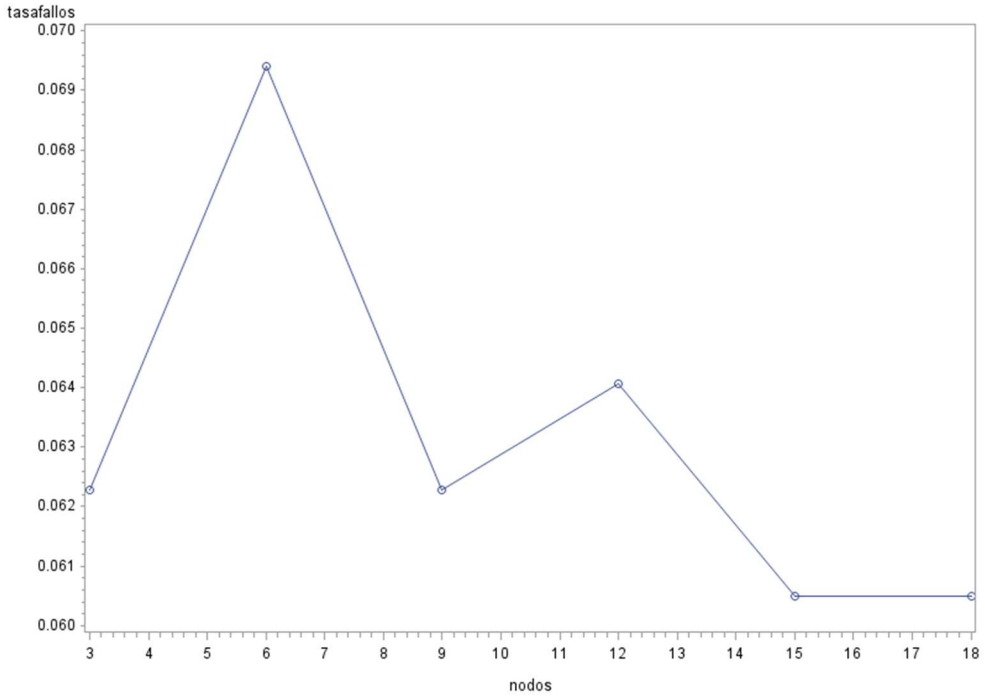
Según los resultados obtenidos con esta macro, en media el mejor resultado se obtienen con el método **Quasi Newton**.

Si nos centramos en la búsqueda de la función de activación con un error medio menor, en el siguiente gráfico podemos ver que en este caso lo es la **Arco Tangente**.

```
%macro activalcruza;
%let lista='TANH LOG ARC LIN SIN SOF GAU';
%let nume=7;
%do  i=1 %to &nume;
data _null_;activa=scanq(&lista,&i);call
symput('activa',left(activa));run;
 %cruzadaneural(archivo=aphodius.aphodius,vardepen=pbs,conti=bio1_15
bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15 bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15
bio19_15,categor=,ngrupos=3,sinicio=12345,sfinal=12347,ocultos=10,acti
=&activa);
 data final&i;set final;modelo="&activa";put modelo=;run;
%end;
data union;set %do i=1 %to &nume; final&i %end;
%mend;

%activalcruza;

proc print data=union;run;
proc boxplot data=union;plot media*modelo;run;
```




```
%macro numeronodos(inicionodos=,finalnodos=,increnodos=);
data union;run;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=aphodius.aphodius,
   listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15  ,

listclass=,vardep=pbs,nodos=&nodos,corte=50,semilla=12345,porcen=0.80)
;
   data estadisticos;set estadisticos;nodos=&nodos;run;
   data union;set union estadisticos;run;
%end;
data union;set union ;if _n_=1 then delete;run;
symbol v=circle i=join;
proc gplot data=union;plot (porcenVN porcenFN porcenVP porcenFP sensi
especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;
%numeronodos(inicionodos=3,finalnodos=20,increnodos=3);
```



Con estos resultados parece que con 15 nodos la tasa de fallos es menor.

Pues bien tenemos una red neuronla de 15 nodos, QNewton y ArcoTangente, pero con todas las variables.

Gradient Boosting

*Resultado con todas las variables

```
GBM STEP - version 2.9

Performing cross-validation optimisation of a boosted regression tree
model
for pbs and using a family of bernoulli
Using 2823 observations and 19 predictors
creating 10 initial models of 50 trees

 folds are stratified by prevalence
total mean deviance =  0.934
tolerance is fixed at  9e-04
ntrees resid. dev.
50    0.8534
now adding trees...
100   0.809
150   0.7784
200   0.7561
250   0.7389
300   0.7246
350   0.7111
400   0.6996
450   0.6899
500   0.6813
550   0.6721
600   0.6644
650   0.6573
700   0.6505
```

750	0.6444
800	0.6387
850	0.6337
900	0.6291
950	0.625
1000	0.6212
1050	0.6172
1100	0.6132
1150	0.6098
1200	0.6065
1250	0.6038
1300	0.6012
1350	0.5984
1400	0.5955
1450	0.5925
1500	0.5902
1550	0.5878
1600	0.5856
1650	0.5832
1700	0.5812
1750	0.5791
1800	0.5774
1850	0.5753
1900	0.5731
1950	0.5718
2000	0.5701
2050	0.5688
2100	0.5672
2150	0.5653
2200	0.5638
2250	0.5623
2300	0.5605
2350	0.5589
2400	0.5577
2450	0.5562
2500	0.5546
2550	0.5534
2600	0.5524
2650	0.5513
2700	0.5502
2750	0.5493
2800	0.5483
2850	0.5476
2900	0.5467
2950	0.5454
3000	0.5442
3050	0.543
3100	0.5422
3150	0.5411
3200	0.5401
3250	0.5391
3300	0.5381
3350	0.537
3400	0.5363
3450	0.5356
3500	0.5346
3550	0.5338
3600	0.5332
3650	0.5326
3700	0.5319
3750	0.5313
3800	0.5305
3850	0.5301
3900	0.5295
3950	0.5288
4000	0.5279
4050	0.5274

4100	0.5265
4150	0.526
4200	0.5255
4250	0.5247
4300	0.5241
4350	0.5234
4400	0.5231
4450	0.5226
4500	0.5219
4550	0.5212
4600	0.5207
4650	0.5204
4700	0.5205
4750	0.52
4800	0.5196
4850	0.5193
4900	0.5189
4950	0.5185
5000	0.5181
5050	0.5176
5100	0.517
5150	0.5168
5200	0.5164
5250	0.5159
5300	0.5155
5350	0.5152
5400	0.515
5450	0.5148
5500	0.5145
5550	0.5139
5600	0.5139
5650	0.5137
5700	0.5137
5750	0.5136
5800	0.5132
5850	0.5126
5900	0.5126
5950	0.5122
6000	0.5119
6050	0.5115
6100	0.5115
6150	0.5112
6200	0.5109
6250	0.5108
6300	0.5102
6350	0.5098
6400	0.5098
6450	0.5092
6500	0.5088
6550	0.5085
6600	0.5086
6650	0.5086
6700	0.5083
6750	0.5081
6800	0.5078
6850	0.5074
6900	0.507
6950	0.5069
7000	0.5066
7050	0.5063
7100	0.506
7150	0.5061
7200	0.5056
7250	0.5052
7300	0.5049
7350	0.5048
7400	0.5048

```
7450    0.505
7500    0.5049
7550    0.5051
7600    0.5048
7650    0.5047
7700    0.5044
7750    0.5042
7800    0.5042
7850    0.504
7900    0.5038
7950    0.5037
8000    0.5036
8050    0.5035
8100    0.5036
8150    0.5034
8200    0.5033
8250    0.5031
8300    0.503
8350    0.503
8400    0.5026
8450    0.5026
8500    0.5024
8550    0.5019
8600    0.5021
8650    0.5021
8700    0.502
8750    0.5019
8800    0.502
8850    0.502
8900    0.5018
8950    0.5019
9000    0.502
9050    0.5021
fitting final gbm model with a fixed number of 8900 trees for pbs

mean total deviance = 0.934
mean residual deviance = 0.284

estimated cv deviance = 0.502 ; se = 0.03

training data correlation = 0.873
cv correlation =  0.711 ; se = 0.022

training data AUC score = 0.982
cv AUC score = 0.914 ; se = 0.011
```

	bio1_15	bio2_15	bio3_15	bio4_15	bio5_15	bio6_15	bio7_15	bio8_15	bio9_15	bio10_15	bio11_15	bio12_15	bio13_15	bio14_15	bio15_15	bio16_15	bio17_15	bio18_15	bio19_15
bio1_15	0	0.14	0.47	0.23	0.79	0.40	0.04	10.10	0.50	1.20	1.55	0.08	0.06	0.01	0.63	0.06	0.39	0.43	0.76
bio2_15	0	0.00	0.90	10.47	0.26	0.09	12.36	7.11	24.09	0.47	0.01	1.27	0.20	0.02	8.21	0.63	0.29	40.09	4.40
bio3_15	0	0.00	0.00	0.82	0.23	0.16	6.73	6.11	6.54	1.35	2.01	2.69	6.16	0.07	0.75	1.01	0.77	0.15	8.86
bio4_15	0	0.00	0.00	0.00	5.05	1.88	104.50	16.91	0.71	0.54	0.18	1.45	3.55	0.09	7.98	2.58	0.82	0.17	7.27
bio5_15	0	0.00	0.00	0.00	0.00	0.49	0.40	1.78	18.43	0.24	0.40	0.55	0.98	0.04	2.61	0.35	1.38	2.58	0.72
bio6_15	0	0.00	0.00	0.00	0.00	0.00	0.40	149.82	6.93	0.12	2.53	0.05	1.83	0.05	2.74	0.26	0.24	1.81	0.35
bio7_15	0	0.00	0.00	0.00	0.00	0.00	0.00	4.36	10.23	0.27	0.17	0.20	1.88	0.26	3.30	1.14	0.13	0.12	16.88
bio8_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.18	10.10	6.41	0.21	0.28	1.07	3.83	1.04	1.15	0.82	6.82
bio9_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.78	1.43	0.13	0.36	0.86	2.76	0.30	0.08	0.92	0.10
bio10_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.61	0.00	0.20	0.30	0.34	0.56	0.05
bio11_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.00	15.80	0.04	0.01	0.10	0.07
bio12_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.04	0.48	0.20	9.70	11.63	4.39	2.58
bio13_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	1.43	14.80	9.99	1.54	1.10
bio14_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.70	0.06	3.08	1.19	0.03
bio15_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.83	0.07	0.31
bio16_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.14	1.22	0.67
bio17_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.54	0.65
bio18_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.24
bio19_15	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

*Resultados con las variables preseleccionadas

```
GBM STEP - version 2.9
```

```
Performing cross-validation optimisation of a boosted regression tree
model
for pbs and using a family of bernoulli
Using 2823 observations and 12 predictors
creating 10 initial models of 50 trees

 folds are stratified by prevalence
total mean deviance =  0.934
tolerance is fixed at  9e-04
ntrees resid. dev.
50    0.8554
now adding trees...
100   0.811
150   0.7822
200   0.7612
250   0.7457
300   0.7335
350   0.7223
400   0.7124
450   0.7038
500   0.696
550   0.689
600   0.6827
650   0.6767
700   0.6705
750   0.665
800   0.6599
850   0.6552
900   0.6511
950   0.6474
1000   0.6436
1050   0.6401
1100   0.6367
1150   0.634
1200   0.6318
1250   0.6284
1300   0.6258
1350   0.6235
1400   0.6214
1450   0.6192
1500   0.6167
1550   0.6147
1600   0.6125
1650   0.611
1700   0.6087
1750   0.6067
1800   0.6047
1850   0.6031
1900   0.6011
1950   0.5994
2000   0.5976
2050   0.5965
2100   0.5952
2150   0.594
2200   0.593
2250   0.5918
2300   0.5905
2350   0.5889
2400   0.5878
2450   0.5868
2500   0.5856
2550   0.5846
2600   0.5825
2650   0.5813
2700   0.58
2750   0.5786
2800   0.5775
```

2850	0.5766
2900	0.576
2950	0.5746
3000	0.5738
3050	0.5725
3100	0.5715
3150	0.5705
3200	0.5694
3250	0.5687
3300	0.5676
3350	0.5668
3400	0.566
3450	0.5652
3500	0.5647
3550	0.5638
3600	0.5632
3650	0.5622
3700	0.5612
3750	0.5604
3800	0.5593
3850	0.5586
3900	0.5577
3950	0.5571
4000	0.5564
4050	0.556
4100	0.5552
4150	0.5545
4200	0.5538
4250	0.5534
4300	0.5528
4350	0.5521
4400	0.5512
4450	0.5502
4500	0.5498
4550	0.5494
4600	0.549
4650	0.5484
4700	0.5478
4750	0.547
4800	0.5466
4850	0.5457
4900	0.5454
4950	0.5451
5000	0.5442
5050	0.5435
5100	0.5431
5150	0.5426
5200	0.5423
5250	0.5417
5300	0.5415
5350	0.5413
5400	0.5405
5450	0.5403
5500	0.5396
5550	0.5394
5600	0.5389
5650	0.5388
5700	0.5387
5750	0.5381
5800	0.538
5850	0.5375
5900	0.5375
5950	0.5373
6000	0.5366
6050	0.5361
6100	0.5358
6150	0.5354

6200	0.5346
6250	0.5342
6300	0.5345
6350	0.5338
6400	0.5333
6450	0.5331
6500	0.5329
6550	0.5322
6600	0.532
6650	0.5315
6700	0.5311
6750	0.5307
6800	0.5308
6850	0.5303
6900	0.5304
6950	0.53
7000	0.5295
7050	0.5296
7100	0.5297
7150	0.5295
7200	0.5292
7250	0.5288
7300	0.5284
7350	0.5281
7400	0.5278
7450	0.5276
7500	0.5276
7550	0.5274
7600	0.5276
7650	0.5274
7700	0.5273
7750	0.527
7800	0.5269
7850	0.5266
7900	0.5266
7950	0.5261
8000	0.5262
8050	0.5261
8100	0.526
8150	0.5256
8200	0.5255
8250	0.5256
8300	0.5254
8350	0.525
8400	0.5249
8450	0.5245
8500	0.5246
8550	0.5245
8600	0.5242
8650	0.5241
8700	0.5242
8750	0.5243
8800	0.5242
8850	0.5241
8900	0.5238
8950	0.5236
9000	0.5234
9050	0.5234
9100	0.5232
9150	0.5233
9200	0.5227
9250	0.5227
9300	0.5223
9350	0.5223
9400	0.5222
9450	0.522
9500	0.5217

```
9550   0.5214
9600   0.5215
9650   0.5212
9700   0.5213
9750   0.5214
9800   0.5208
9850   0.5204
9900   0.5203
9950   0.5202
10000   0.52
fitting final gbm model with a fixed number of 10000 trees for pbs

mean total deviance = 0.934
mean residual deviance = 0.307

estimated cv deviance = 0.52 ; se = 0.027

training data correlation = 0.858
cv correlation =  0.699 ; se = 0.017

training data AUC score = 0.977
cv AUC score = 0.907 ; se = 0.01
```

Random Forest

*Resultados con las variables preseleccionadas

```
%randomforestbin(archivo=aphodius.aphodius,
vardep=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio5_15 bio8_15 bio9_15 bio10_15
bio11_15 bio12_15 bio15_15 bio16_15 bio18_15,
listcategor=,
semilla1=12345,porcen1=0.80,

maxtrees=30,variables=3,porcenbag=0.5,maxbranch=2,tamhoja=5,maxdept
h=15,pvalor=0.2,compara=1);
```

- Iniciamos este algoritmo con los siguientes parámetros:
  - 30 árboles como máximo
  - Con 2 divisiones máximas en cada nodo
  - 5 hojas finales del árbol
  - Variables a muestrear en cada nodo 3
  - Porcenbag 0.5
  - Con reemplazamiento

RANDOM FOREST Iteraciones=30

The MEANS Procedure
Selection Indicator=0

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
561	0.0695187	0.2545611	0	1.0000000

Selection Indicator=1

Analysis Variable : error				
N	Mean	Std Dev	Minimum	Maximum
2247	0.0663106	0.2488797	0	1.0000000

RESULTADOS LOGISTICA

The MEANS Procedure
Selection Indicator=0

Analysis Variable : error				
N	Mean	Std Dev	Minimum	Maximum
561	0.0998217	0.3000297	0	1.0000000

Selection Indicator=1

Analysis Variable : error				
N	Mean	Std Dev	Minimum	Maximum
2247	0.1032488	0.3043513	0	1.0000000

*Resultados con todas las variables

- Iniciamos este algoritmo con los siguientes parámetros:
  - 30 árboles como máximo
  - Con 2 divisiones máximas en cada nodo
  - 5 hojas finales del árbol
  - Variables a muestrear en cada nodo 3
  - Porcenbag 0.5
  - Con reemplazamiento

```
%randomforestbin(archivo=aphodius.aphodius,
vardep=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15
bio16_15 bio17_15
bio18_15 bio19_15,
listcategor=,
semilla1=12345,porcen1=0.80,

maxtrees=30,variables=3,porcenbag=0.5,maxbranch=2,tamhoja=5,maxdept
h=15,pvalor=0.2,compara=1);
```

RANDOM FOREST Iteraciones=30

The MEANS Procedure
Selection Indicator=0

Analysis Variable : error				
N	Mean	Std Dev	Minimum	Maximum
561	0.0819964	0.2746041	0	1.0000000

Selection Indicator=1

Analysis Variable : error				
N	Mean	Std Dev	Minimum	Maximum
2247	0.0672007	0.2504250	0	1.0000000

RESULTADOS LOGISTICA

The MEANS Procedure
Selection Indicator=0
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
561	0.0802139	0.2718665	0	1.0000000

Selection Indicator=1
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
2247	0.0827770	0.2756062	0	1.0000000

- Modificamos el algoritmo con los siguientes parámetros:

```
%randomforestbin(archivo=aphodius.aphodius,
vardep=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15
bio16_15 bio17_15
bio18_15 bio19_15,
listcategor=,
semilla1=12345,porcen1=0.80,
maxtrees=100,variables=3,porcenbag=0.01,maxbranch=10,tamhoja=5,maxd
epth=8,pvalor=0.2,compara=1);
```

RANDOM FOREST Iteraciones=100

The MEANS Procedure
Selection Indicator=0
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
561	0.1871658	0.3903926	0	1.0000000

Selection Indicator=1
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
2247	0.1757899	0.3807261	0	1.0000000

RESULTADOS LOGISTICA

The MEANS Procedure
Selection Indicator=0
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
561	0.0802139	0.2718665	0	1.0000000

Selection Indicator=1
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
2247	0.0827770	0.2756062	0	1.0000000

- Modificamos el algoritmo con los siguientes parámetros:

```
%randomforestbin(archivo=aphodius.aphodius,
vardep=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15
bio16_15 bio17_15
bio18_15 bio19_15,
listcategor=,
semilla1=12345,porcen1=0.80,
maxtrees=30,variables=5,porcenbag=0.8,maxbranch=3,tamhoja=3,maxdept
h=8,pvalor=0.2,compara=1);
```

RANDOM FOREST Iteraciones=500

The MEANS Procedure
Selection Indicator=0
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
561	0.0748663	0.2634104	0	1.0000000

Selection Indicator=1
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
2247	0.0591900	0.2360326	0	1.0000000

RESULTADOS LOGISTICA

The MEANS Procedure
Selection Indicator=0
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
561	0.0802139	0.2718665	0	1.0000000

Selection Indicator=1
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
2247	0.0827770	0.2756062	0	1.0000000

Suppor Vector Machine

Modelo SVM1

```
%cruzadaSVMbin
(archivo=aphodius.aphodius,
vardepen=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15,
listclass=,
```

```
ngrupos=4,seminicio=12345,semifinal=12385,kernel=linear,c=10);
data final1;set final;modelo='SVM1';


%cruzadaSVMbin
(archivo=aphodius.aphodius,
vardepen=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15,
listclass=,
ngrupos=4,seminicio=12345,semifinal=12385,kernel=polynom
k_par=2,c=10);
data final2;set final;modelo='SVM2';

%cruzadaSVMbin
(archivo=aphodius.aphodius,
vardepen=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15,
listclass=,
ngrupos=4,seminicio=12345,semifinal=12385,kernel=RBF
k_par=gamma,c=10);
data final3;set final;modelo='SVM3';


/* c=5 */

%cruzadaSVMbin
(archivo=aphodius.aphodius,
vardepen=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15,
listclass=,
ngrupos=4,seminicio=12345,semifinal=12385,kernel=linear,c=5);
data final4;set final;modelo='SVM4';


%cruzadaSVMbin
(archivo=aphodius.aphodius,
vardepen=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15,
listclass=,
ngrupos=4,seminicio=12345,semifinal=12385,kernel=polynom k_par=2,c=5);
data final5;set final;modelo='SVM5';

%cruzadaSVMbin
```

```
(archivo=aphodius.aphodius,
vardepen=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15,
listclass=,
ngrupos=4,seminicio=12345,semifinal=12385,kernel=RBF k_par=gamma,c=5);
data final6;set final;modelo='SVM6';

%cruzadaSVMbin
(archivo=aphodius.aphodius,
vardepen=pbs,
listconti=bio1_15 bio2_15 bio3_15 bio4_15 bio5_15 bio6_15 bio7_15
bio8_15
bio9_15 bio10_15 bio11_15 bio12_15 bio13_15 bio14_15 bio15_15 bio16_15
bio17_15
bio18_15 bio19_15,
listclass=,
ngrupos=4,seminicio=12345,semifinal=12385,kernel=polynom
k_par=2,c=20);
data final7;set final;modelo='SVM7';



data union;set final1 final2 final3 final4 final5 final6 final7;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;

data union;set final2 final5 final7;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;

Proc print data=final7;
Run;

Proc means data=final1;
Run;
Proc means data=final2;
Run;
Proc means data=final3;
Run;
Proc means data=final4;
Run;
Proc means data=final5;
Run;
Proc means data=final6;
Run;
Proc means data=final7;
Run;
```

Tasa de error por modelo

Obs	media	semilla	modelo
1	0.65670	12345	SVM1
2	0.50356	12346	SVM1
3	0.66809	12347	SVM1
4	0.49430	12348	SVM1
5	0.50142	12349	SVM1
6	0.66311	12350	SVM1
7	0.17806	12351	SVM1
8	0.17806	12352	SVM1
9	0.34544	12353	SVM1
10	0.51140	12354	SVM1
11	0.34829	12355	SVM1
12	0.66667	12356	SVM1
13	0.34402	12357	SVM1
14	0.33903	12358	SVM1
15	0.33405	12359	SVM1
16	0.67023	12360	SVM1
17	0.49074	12361	SVM1
18	0.17806	12362	SVM1
19	0.66168	12363	SVM1
20	0.82194	12364	SVM1
21	0.66738	12365	SVM1
22	0.50071	12366	SVM1
23	0.50499	12367	SVM1
24	0.50427	12368	SVM1
25	0.50214	12369	SVM1
26	0.33974	12370	SVM1
27	0.17806	12371	SVM1
28	0.66168	12372	SVM1
29	0.33832	12373	SVM1
30	0.66453	12374	SVM1
31	0.66168	12375	SVM1
32	0.49003	12376	SVM1
33	0.33333	12377	SVM1
34	0.34473	12378	SVM1
35	0.49217	12379	SVM1
36	0.17806	12380	SVM1
37	0.34473	12381	SVM1
38	0.50142	12382	SVM1
39	0.50641	12383	SVM1
40	0.67236	12384	SVM1
41	0.49145	12385	SVM1

Obs	media	semilla	modelo
1	0.073362	12345	SVM2
2	0.071581	12346	SVM2
3	0.070869	12347	SVM2
4	0.074074	12348	SVM2
5	0.072650	12349	SVM2
6	0.074074	12350	SVM2
7	0.074074	12351	SVM2
8	0.074074	12352	SVM2
9	0.073362	12353	SVM2
10	0.073362	12354	SVM2
11	0.071225	12355	SVM2
12	0.073362	12356	SVM2
13	0.071225	12357	SVM2
14	0.073362	12358	SVM2
15	0.072293	12359	SVM2
16	0.071937	12360	SVM2
17	0.070157	12361	SVM2
18	0.072650	12362	SVM2
19	0.071937	12363	SVM2
20	0.071581	12364	SVM2
21	0.072650	12365	SVM2
22	0.073718	12366	SVM2
23	0.073362	12367	SVM2
24	0.073006	12368	SVM2
25	0.072650	12369	SVM2
26	0.073362	12370	SVM2
27	0.073718	12371	SVM2
28	0.071225	12372	SVM2
29	0.073362	12373	SVM2
30	0.071937	12374	SVM2
31	0.074074	12375	SVM2
32	0.072293	12376	SVM2
33	0.072650	12377	SVM2
34	0.071581	12378	SVM2
35	0.072293	12379	SVM2
36	0.070869	12380	SVM2
37	0.071581	12381	SVM2
38	0.075499	12382	SVM2
39	0.071937	12383	SVM2
40	0.073006	12384	SVM2
41	0.070513	12385	SVM2

Obs	media	semilla	modelo
1	0.071225	12345	SVM3
2	0.071225	12346	SVM3
3	0.071225	12347	SVM3
4	0.071225	12348	SVM3
5	0.071225	12349	SVM3
6	0.071225	12350	SVM3
7	0.071225	12351	SVM3
8	0.071225	12352	SVM3
9	0.071225	12353	SVM3
10	0.071225	12354	SVM3
11	0.071225	12355	SVM3
12	0.071225	12356	SVM3
13	0.071225	12357	SVM3
14	0.071225	12358	SVM3
15	0.071225	12359	SVM3
16	0.071225	12360	SVM3
17	0.071225	12361	SVM3
18	0.071225	12362	SVM3
19	0.071225	12363	SVM3
20	0.071225	12364	SVM3
21	0.071225	12365	SVM3
22	0.071225	12366	SVM3
23	0.071225	12367	SVM3
24	0.071225	12368	SVM3
25	0.071225	12369	SVM3
26	0.071225	12370	SVM3
27	0.071225	12371	SVM3
28	0.071225	12372	SVM3
29	0.071225	12373	SVM3
30	0.071225	12374	SVM3
31	0.071225	12375	SVM3
32	0.071225	12376	SVM3
33	0.071225	12377	SVM3
34	0.071225	12378	SVM3
35	0.071225	12379	SVM3
36	0.071225	12380	SVM3
37	0.071225	12381	SVM3
38	0.071225	12382	SVM3
39	0.071225	12383	SVM3
40	0.071225	12384	SVM3
41	0.071225	12385	SVM3

Obs	media	semilla	modelo
1	0.66239	12345	SVM4
2	0.33618	12346	SVM4
3	0.17806	12347	SVM4
4	0.33903	12348	SVM4
5	0.33547	12349	SVM4
6	0.32835	12350	SVM4
7	0.33974	12351	SVM4
8	0.17806	12352	SVM4
9	0.50427	12353	SVM4
10	0.33832	12354	SVM4
11	0.17806	12355	SVM4
12	0.34188	12356	SVM4
13	0.82194	12357	SVM4
14	0.33689	12358	SVM4
15	0.50142	12359	SVM4
16	0.33476	12360	SVM4
17	0.66097	12361	SVM4
18	0.33689	12362	SVM4
19	0.33689	12363	SVM4
20	0.51282	12364	SVM4
21	0.33262	12365	SVM4
22	0.17806	12366	SVM4
23	0.49501	12367	SVM4
24	0.33689	12368	SVM4
25	0.34473	12369	SVM4
26	0.34046	12370	SVM4
27	0.50214	12371	SVM4
28	0.65741	12372	SVM4
29	0.66809	12373	SVM4
30	0.49929	12374	SVM4
31	0.33832	12375	SVM4
32	0.17806	12376	SVM4
33	0.34259	12377	SVM4
34	0.33191	12378	SVM4
35	0.34615	12379	SVM4
36	0.17806	12380	SVM4
37	0.50214	12381	SVM4
38	0.34330	12382	SVM4
39	0.33191	12383	SVM4
40	0.34402	12384	SVM4
41	0.50855	12385	SVM4

Obs	media	semilla	modelo
1	0.075142	12345	SVM5
2	0.074074	12346	SVM5
3	0.072293	12347	SVM5
4	0.075142	12348	SVM5
5	0.074786	12349	SVM5
6	0.075142	12350	SVM5
7	0.075499	12351	SVM5
8	0.075142	12352	SVM5
9	0.075142	12353	SVM5
10	0.074786	12354	SVM5
11	0.074786	12355	SVM5
12	0.075499	12356	SVM5
13	0.073362	12357	SVM5
14	0.074786	12358	SVM5
15	0.074786	12359	SVM5
16	0.074430	12360	SVM5
17	0.074430	12361	SVM5
18	0.073006	12362	SVM5
19	0.074074	12363	SVM5
20	0.073718	12364	SVM5
21	0.074074	12365	SVM5
22	0.075142	12366	SVM5
23	0.076567	12367	SVM5
24	0.074786	12368	SVM5
25	0.075142	12369	SVM5
26	0.074786	12370	SVM5
27	0.076211	12371	SVM5
28	0.074074	12372	SVM5
29	0.074786	12373	SVM5
30	0.075855	12374	SVM5
31	0.075499	12375	SVM5
32	0.074786	12376	SVM5
33	0.075855	12377	SVM5
34	0.073362	12378	SVM5
35	0.074074	12379	SVM5
36	0.074430	12380	SVM5
37	0.073718	12381	SVM5
38	0.076923	12382	SVM5
39	0.074430	12383	SVM5
40	0.075499	12384	SVM5
41	0.074430	12385	SVM5

Obs	media	semilla	modelo
1	0.076923	12345	SVM6
2	0.076923	12346	SVM6
3	0.076923	12347	SVM6
4	0.076923	12348	SVM6
5	0.076923	12349	SVM6
6	0.076923	12350	SVM6
7	0.076923	12351	SVM6
8	0.076923	12352	SVM6
9	0.076923	12353	SVM6
10	0.076923	12354	SVM6
11	0.076923	12355	SVM6
12	0.076923	12356	SVM6
13	0.076923	12357	SVM6
14	0.076923	12358	SVM6
15	0.076923	12359	SVM6
16	0.076923	12360	SVM6
17	0.076923	12361	SVM6
18	0.076923	12362	SVM6
19	0.076923	12363	SVM6
20	0.076923	12364	SVM6
21	0.076923	12365	SVM6
22	0.076923	12366	SVM6
23	0.076923	12367	SVM6
24	0.076923	12368	SVM6
25	0.076923	12369	SVM6
26	0.076923	12370	SVM6
27	0.076923	12371	SVM6
28	0.076923	12372	SVM6
29	0.076923	12373	SVM6
30	0.076923	12374	SVM6
31	0.076923	12375	SVM6
32	0.076923	12376	SVM6
33	0.076923	12377	SVM6
34	0.076923	12378	SVM6
35	0.076923	12379	SVM6
36	0.076923	12380	SVM6
37	0.076923	12381	SVM6
38	0.076923	12382	SVM6
39	0.076923	12383	SVM6
40	0.076923	12384	SVM6
41	0.076923	12385	SVM6

Obs	media	semilla	modelo
1	0.072293	12345	SVM7
2	0.069444	12346	SVM7
3	0.069801	12347	SVM7
4	0.070513	12348	SVM7
5	0.070157	12349	SVM7
6	0.072650	12350	SVM7
7	0.071581	12351	SVM7
8	0.071225	12352	SVM7
9	0.071581	12353	SVM7
10	0.070157	12354	SVM7
11	0.069444	12355	SVM7
12	0.071581	12356	SVM7
13	0.069801	12357	SVM7
14	0.070513	12358	SVM7
15	0.070869	12359	SVM7

16	0.071225	12360	SVM7
17	0.070157	12361	SVM7
18	0.070513	12362	SVM7
19	0.070513	12363	SVM7
20	0.070157	12364	SVM7
21	0.071225	12365	SVM7
22	0.069444	12366	SVM7
23	0.071581	12367	SVM7
24	0.070157	12368	SVM7
25	0.070157	12369	SVM7
26	0.072650	12370	SVM7
27	0.070869	12371	SVM7
28	0.068732	12372	SVM7
29	0.070869	12373	SVM7
30	0.069444	12374	SVM7
31	0.073006	12375	SVM7
32	0.070513	12376	SVM7
33	0.069801	12377	SVM7
34	0.068732	12378	SVM7
35	0.069801	12379	SVM7
36	0.069444	12380	SVM7
37	0.070869	12381	SVM7
38	0.073362	12382	SVM7
39	0.069088	12383	SVM7
40	0.069444	12384	SVM7
41	0.068020	12385	SVM7

Comparación de modelos – Tuckey

	(I) Modelo	(J) Modelo	Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95%	
						Límite inferior	Límite superior
HSD Tukey	Bioclim	C_H	-,3127*	0,05935	0,00	-0,5038	-0,1216
		D_G	-,5310*	0,01417	0,00	-0,5766	-0,4853
		Domain	-,2221*	0,01402	0,00	-0,2673	-0,177
		GAM	-,5454*	0,01474	0,00	-0,5929	-0,498
		GB	-,2872*	0,01378	0,00	-0,3316	-0,2428
		GLM	-,5493*	0,01422	0,00	-0,595	-0,5035
		Mahalanobis	-,4499*	0,01353	0,00	-0,4935	-0,4063
		P_A	-,3186*	0,05336	0,00	-0,4904	-0,1468
		RF	-,3547*	0,04284	0,00	-0,4926	-0,2167
		SVM	-0,063	0,10135	1,00	-0,3894	0,2633
	C_H	Bioclim	,3127*	0,05935	0,00	0,1216	0,5038
		D_G	-,2183*	0,05854	0,01	-0,4068	-0,0298
		Domain	0,0905	0,0585	0,90	-0,0978	0,2789
		GAM	-,2328*	0,05868	0,00	-0,4217	-0,0438
		GB	0,0255	0,05844	1,00	-0,1627	0,2137
		GLM	-,2366*	0,05855	0,00	-0,4251	-0,048
		Mahalanobis	-0,1372	0,05839	0,40	-0,3252	0,0508
		P_A	-0,0059	0,07793	1,00	-0,2569	0,245
		RF	-0,042	0,07114	1,00	-0,2711	0,1871
		SVM	0,2497	0,11618	0,54	-0,1244	0,6238
	D_G	Bioclim	,5310*	0,01417	0,00	0,4853	0,5766
		C_H	,2183*	0,05854	0,01	0,0298	0,4068
		Domain	,3088*	0,01004	0,00	0,2765	0,3412
		GAM	-0,0145	0,01102	0,97	-0,0499	0,021
		GB	,2438*	0,00969	0,00	0,2126	0,275
		GLM	-0,0183	0,01032	0,80	-0,0515	0,0149
		Mahalanobis	,0811*	0,00934	0,00	0,051	0,1112
		P_A	,2124*	0,05246	0,00	0,0434	0,3813
		RF	,1763*	0,04171	0,00	0,042	0,3106
		SVM	,4680*	0,10087	0,00	0,1431	0,7928
	Domain	Bioclim	,2221*	0,01402	0,00	0,177	0,2673
		C_H	-0,0905	0,0585	0,90	-0,2789	0,0978
		D_G	-,3088*	0,01004	0,00	-0,3412	-0,2765
		GAM	-,3233*	0,01082	0,00	-0,3582	-0,2885
		GB	-,0651*	0,00947	0,00	-0,0956	-0,0346
		GLM	-,3271*	0,01011	0,00	-0,3597	-0,2946
		Mahalanobis	-,2278*	0,00911	0,00	-0,2571	-0,1984
		P_A	-0,0965	0,05242	0,76	-0,2653	0,0723
		RF	-0,1325	0,04166	0,06	-0,2667	0,0016
		SVM	0,1591	0,10085	0,89	-0,1656	0,4839
	GAM	Bioclim	,5454*	0,01474	0,00	0,498	0,5929
		C_H	,2328*	0,05868	0,00	0,0438	0,4217
		D_G	0,0145	0,01102	0,97	-0,021	0,0499
		Domain	,3233*	0,01082	0,00	0,2885	0,3582
		GB	,2582*	0,0105	0,00	0,2244	0,2921
		GLM	-0,0038	0,01108	1,00	-0,0395	0,0319
		Mahalanobis	,0955*	0,01018	0,00	0,0628	0,1283
		P_A	,2268*	0,05261	0,00	0,0574	0,3962
		RF	,1908*	0,0419	0,00	0,0558	0,3257
		SVM	,4824*	0,10095	0,00	0,1574	0,8075
	GB	Bioclim	,2872*	0,01378	0,00	0,2428	0,3316
		C_H	-0,0255	0,05844	1,00	-0,2137	0,1627
		D_G	-,2438*	0,00969	0,00	-0,275	-0,2126
		Domain	,0651*	0,00947	0,00	0,0346	0,0956
		GAM	-,2582*	0,0105	0,00	-0,2921	-0,2244
		GLM	-,2621*	0,00977	0,00	-0,2935	-0,2306
		Mahalanobis	-,1627*	0,00873	0,00	-0,1908	-0,1346
		P_A	-0,0314	0,05235	1,00	-0,2	0,1372
		RF	-0,0675	0,04158	0,87	-0,2014	0,0664
		SVM	0,2242	0,10082	0,49	-0,1005	0,5488
	GLM	Bioclim	,5493*	0,01422	0,00	0,5035	0,595
		C_H	,2366*	0,05855	0,00	0,048	0,4251
		D_G	0,0183	0,01032	0,80	-0,0149	0,0515
		Domain	,3271*	0,01011	0,00	0,2946	0,3597
		GAM	0,0038	0,01108	1,00	-0,0319	0,0395
		GB	,2621*	0,00977	0,00	0,2306	0,2935
		Mahalanobis	,0994*	0,00942	0,00	0,069	0,1297
		P_A	,2306*	0,05247	0,00	0,0617	0,3996
		RF	,1946*	0,04173	0,00	0,0602	0,3289
		SVM	,4862*	0,10088	0,00	0,1614	0,8111
	Mahalanobis	Bioclim	,4499*	0,01353	0,00	0,4063	0,4935
		C_H	0,1372	0,05839	0,40	-0,0508	0,3252
		D_G	-,0811*	0,00934	0,00	-0,1112	-0,051
		Domain	,2278*	0,00911	0,00	0,1984	0,2571
		GAM	-,0955*	0,01018	0,00	-0,1283	-0,0628
		GB	,1627*	0,00873	0,00	0,1346	0,1908
		GLM	-,0994*	0,00942	0,00	-0,1297	-0,069
		P_A	0,1313	0,05229	0,30	-0,0371	0,2997
		RF	0,0952	0,0415	0,44	-0,0384	0,2288
		SVM	,3869*	0,10078	0,01	0,0624	0,7114
	P_A	Bioclim	,3186*	0,05336	0,00	0,1468	0,4904
		C_H	0,0059	0,07793	1,00	-0,245	0,2569
		D_G	-,2124*	0,05246	0,00	-0,3813	-0,0434
		Domain	0,0965	0,05242	0,76	-0,0723	0,2653
		GAM	-,2268*	0,05261	0,00	-0,3962	-0,0574
		GB	0,0314	0,05235	1,00	-0,1372	0,2
		GLM	-,2306*	0,05247	0,00	-0,3996	-0,0617
		Mahalanobis	-0,1313	0,05229	0,30	-0,2997	0,0371
		RF	-0,0361	0,06623	1,00	-0,2493	0,1772
		SVM	0,2556	0,11323	0,46	-0,109	0,6202
	RF	Bioclim	,3547*	0,04284	0,00	0,2167	0,4926
		C_H	0,042	0,07114	1,00	-0,1871	0,2711
		D_G	-,1763*	0,04171	0,00	-0,3106	-0,042

Código R para los diferentes modelo

A fin de resumir se incluye solo el código de los algoritmos que incluyen las variables preseleccionadas.

```
### 1- Descargar todos los Aphodius de Gbif###
bandapho <- gbif("Aphodius", "*", geo=FALSE)

###Seleccionar solo España###
spain<-subset(bandapho,country=='Spain')

###Cogemos solo los datos que tengan completas las coordenadas###
acgeos<-subset(spain, !is.na(lon) & !is.na(lat))
dim(acgeos)
acgeos[1:4, c(1:5,7:10)]
summary(acgeos)

###Dibujamos los puntos con wrld_simpl###
data(wrld_simpl)
plot(wrld_simpl, xlim=c(-40,20), ylim=c(30,40), axes=TRUE, col="light yellow")
# restore the box around the map
box()
points(acgeos$lon, acgeos$lat, col='orange', pch=20, cex=0.75)
# plot points again to add a border, for better visibility
points(acgeos$lon, acgeos$lat, col='red', cex=0.75)

###Depurando la bbdd Spain. Vemos si ha datos con Longitud 0###
lonzeros=subset(acgeos,lon==0)
lonzeros[, 1:30]
dups <- duplicated(spain)
lonzeros  <-  lonzeros[dups, ]

###Eliminamos las observaciones duplicadas en función de especie y coordenadas###

dupss <- duplicated(acgeos[, c('species', 'lon', 'lat')])
# ignoring (sub) species and other naming variation
dupss <- duplicated(acgeos[, c('lon', 'lat')])

####Parte 2- Georeferencias####
georefs<-subset(spain, (is.na(lon)| is.na(lat)) & ! is.na(locality))
dim(georefs)
bs<- try( geocode(georefs$cloc[1:841]))
bs
bss<-subset(bs,!is.na(longitude)|!is.na(latitude))

spain2<-merge(spain,bss)

####Pseudo-ausencias######
filess <- raster(file.choose())
set.seed(1963)
bgs<-randomPoints(filess, 500)

par(mfrow=c(1,2))
plot(!is.na(filess), legend=FALSE)
points(bgs, cex=0.5)
# now we repeat the sampling, but limit
# the area of sampling using a spatial extent
es <- extent(-10, 10, 35, 45)
bg2s <- randomPoints(filess, 500, ext=es)
plot(!is.na(filess), legend=FALSE)
```

```
plot(es, add=TRUE, col='red')
points(bg2s, cex=0.5)

####Datos ambientales####
prec<-getData('worldclim', var='prec',res=0.5,lon=-4,lat=37.2)
plot(prec)
bio<-getData('worldclim', var='bio',res=0.5,lon=-10,lat=45) #left - Mapa izquierdo en el que esta España sin Cataluña y Baleares
bior<-getData('worldclim', var='bio',res=0.5,lon=10,lat=45) #right - Mapa derecho en el que estan Cataluña y Baleares
ext<-extent(-10,4,35,55)
l.c <- crop(bio,ext) # Asignar la extensión que buscamos
r.c <- crop(bior,ext)
m.c <- merge(l.c,r.c)# Unir los dos raster
plot(m.c) #Plot de la unión de raster
#reseting names to original
names(m.c)<-names(bio)
#setting correct decimal points
m.c <- m.c*0.1
names(m.c)

sr<-
writeRaster(m.c,"C:/Users/jose/Documents/pruebabio",format="ascii",bylayer=TRUE,suffix="names",overwrite=TRUE)
m.cstack<-stack(m.c) # Compactar en un RasterStack ya que es un Brink#
m.cstack
writeRaster(m.c, filename="C:/Users/jose/Documents/pruebabio/bio.grs", bylayer=T, overwrite=TRUE)# makes one .grd file for all variables.


m.cstack_pre <- dropLayer(m.cstack, c(4,6,7,13,15,17,19))



######CREAMOS DATOS TRAIN - TEST######

extraspain<-cbind(acgeos$lon,acgeos$lat)

colnames(extraspain)=c('lon','lat')

presvalss<-extract(m.cstack,extraspain)

set.seed(0)

groupspre<-kfold(extraspain,5)

pres_trainspre<-extraspain[groupspre !=1, ]

pres_testspre<-extraspain[groupspre ==1, ]

ext=extent(-10, 4, 35, 55)

backgspre<-randomPoints(m.cstack_pre, n=1000, ext=ext, extf=1.25)

colnames(backgspre)=c('lon', 'lat')

grouppre<-kfold(backgspre,5)

backg_trainspre<-backgspre[grouppre !=1, ]

backg_testspre<-backgspre[grouppre ==1, ]


rspre=raster(m.cstack_pre, 1)
```

```
plot(!is.na(rspre), col=c('white', 'light grey'), legend=FALSE)
plot(ext, add=TRUE, col='red', lwd=2)
points(backg_trainspre, pch='-', cex=0.5, col='yellow')
points(backg_testspre, pch='-', cex=0.5, col='black')
points(pres_trainspre, pch='+', cex=0.5, col='green')
points(pres_testspre, pch='+', cex=0.5, col='blue')

######  DOMAIN  #######

dmspre<-domain(m.cstack_pre, pres_trainspre)
e<-evaluate(pres_testspre, backg_testspre, dmspre, m.cstack_pre)
e
pdspre=predict(m.cstack_pre, dmspre, ext=ext, pogress='')
par(mfrow=c(1,2))
plot(pdspre, main='Domain')
plot(wrld_simpl, add=TRUE, border='dark grey')
trspre<-threshold(e, 'spec_sens')
plot(pdspre > trspre, main='presencia/background')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_trainspre, pch='+')
points(backg_trainspre, pch='-', cex=0.25)

str(pdspre)

#####  BIOCLIM  #######
bcspre<-bioclim(m.cstack_pre, pres_trainspre)
plot(bcspre, a=3, b=4, p=0.85)
e <- evaluate(pres_testspre, backg_testspre, bcspre, m.cstack_pre)
e
trspre <- threshold(e, 'spec_sens')
trspre
pbspre <- predict(m.cstack_pre, bcspre, ext=ext, progress='')
pbspre
```

```
par(mfrow=c(1,2))
plot(pbspre, main='Bioclim')
plot(wrld_simpl, add=TRUE, border='dark grey')
plot(pbspre > trspre, main='Presencia/ausencia')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_trainspre, pch='+',cex=0.5, col='blue')
points(backg_trainspre, pch='-', cex=0.25)

#### MAHALANOBIS  ####
pres_testsprelimpio<-na.omit(pres_tests)
pres_trainsprelimpio<-na.omit(pres_trainspre)
backg_testsprelimpio<-na.omit(backg_testspre)
mmpre <- mahal(m.cstack_pre, pres_trainsprelimpio)
e <- evaluate(pres_testsprelimpio, backg_testsprelimpio, mmpre, m.cstack_pre)
e
pmpre = predict(m.cstack_pre, mmpre, ext=ext, progress='')
par(mfrow=c(1,2))
pm[pmpre < -10] <- -10
plot(pmpre, main='Mahalanobis distance')
plot(wrld_simpl, add=TRUE, border='dark grey')
trmpre <- threshold(e, 'spec_sens')
plot(pmpre > trmpre, main='presence/absence')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_train, pch='+')
plot(e, 'ROC', col='blue', cex=0.1)
#### MODELOS DE REGRESIÓN #######

trainspre <- rbind(pres_trainspre, backg_trainspre)
pb_trainspre <- c(rep(1, nrow(pres_trainspre)), rep(0, nrow(backg_trainspre)))
envtrainspre <- extract(m.cstack_pre, trainspre)
envtrainspre <- data.frame( cbind(paspre=pb_trainspre, envtrainspre) )
head(envtrainspre)
envtrainspre
```

```
testpresspre<- data.frame( extract(m.cstack_pre, pres_testspre) )
testbackgspre <- data.frame( extract(m.cstack_pre, backg_testspre) )

### GLM   ###

gm4spre <- glm(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
            bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
            bio16_15 + bio18_15,
          family = binomial(link = "logit"), data=envtrainspre)
gausspre <- glm(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
              bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
              bio16_15 + bio18_15,
            family = gaussian(link = "identity"), data=envtrainspre)
gausspre <- glm(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
              bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
              bio16_15 + bio18_15,
            family = gaussian(link = "identity"), data=envtrainspre)
poisspre <- glm(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
              bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
              bio16_15 + bio18_15,
            family = poisson(link = "log"), data=envtrainspre)

plot(gm4spre)
ge4spre<-evaluate(testpresspre, testbackgspre, gm4spre)
ge4spre
pgs4pre <- predict(m.cstack_pre, gm4spre, ext=ext)
trsglmpre <- threshold(ge4spre, 'spec_sens')
par(mfrow=c(1,2))
plot(pgs4pre, main='GLM/binomial')
plot(pgs4pre > trsglmpre, main='Presencia/background')
plot(ge4spre, 'ROC', col='blue', cex=0.1)
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_trainspre, pch='+', col='blue')
```

```
points(backg_trainspre, pch='-', cex=0.25)
summary(gm4spre)

###PARA OBTENER LA MATRIZ DE CNFUSION#####
m5 <- glm(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
          bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
          bio16_15 + bio18_15,
        family=binomial(link="logit"), data=envtrainsprelimpio)
#Instale el paquete gmodels y llame la librería
install.packages("gmodels")
library(gmodels)
#Para determinar un punto de corte utilice tentativamente la media de los valores ajustado
threshold<-mean(fitted(m5))
threshold
#Ejecute la tabla cruzada o tabla de clasificación usando el valor real y el valor ajustado (pronosticado)
CrossTable(envtrainsprelimpio$paspre, fitted(m5) > threshold,expected=FALSE, prop.r=TRUE, prop.c=TRUE,
         prop.t=F, prop.chisq=F, chisq = FALSE, fisher=FALSE)

## sobre datos test ###
m5 <- glm(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
          bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
          bio16_15 + bio18_15,
        family=binomial(link="logit"), data=envtestsprelimpio)
#Instale el paquete gmodels y llame la librería
install.packages("gmodels")
library(gmodels)
#Para determinar un punto de corte utilice tentativamente la media de los valores ajustado
threshold<-mean(fitted(m5))
threshold
#Ejecute la tabla cruzada o tabla de clasificación usando el valor real y el valor ajustado (pronosticado)
CrossTable(envtestsprelimpio$paspre, fitted(m5) > threshold,expected=FALSE, prop.r=TRUE, prop.c=TRUE,
         prop.t=F, prop.chisq=F, chisq = FALSE, fisher=FALSE)
```

```
###### Modelos GAM  #####
gamspre<-gam(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
          bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
          bio16_15 + bio18_15, family = binomial, na=na.gam.replace,data=envtrainspre,
         trace=TRUE)
gemspre<-evaluate(testpresspre, testbackgspre, gamspre)
gemspre
pgamspre <- predict(m.cstack_pre, gamspre, ext=ext)
summary(gamspre)
pgamspre
gamspre
par(mfrow=c(1,2))
plot(pgamspre, main='GAM/binomial')
plot(wrld_simpl, add=TRUE, border='dark grey')
trsgampre <- threshold(gemspre, 'spec_sens')
plot(pgamspre > trsgampre, main='Presencia/background')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_trainspre, pch='+', col='blue')
points(backg_trainspre, pch='-', cex=0.25)
gamspre$aic

gamspre2<-gam(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
           bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
           bio16_15 + bio18_15, family = gaussian(link = "identity"), na=na.gam.replace,data=envtrainspre,
          trace=TRUE)
gamspre2
gamspre2$aic
gemspre2<-evaluate(testpresspre, testbackgspre, gamspre2)
gemspre2
gamspre3<-gam(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
            bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
            bio16_15 + bio18_15, family = poisson(link = "log"), na=na.gam.replace,data=envtrainspre,
           trace=TRUE)
```

```
gemspre3<-evaluate(testpresspre, testbackgspre, gamspre3)
gemspre3
gamspre3$aic

gam.check(gamspre)
par(mfrow=c(1,2))
plot(gemspre, 'ROC', col='blue', cex=0.1)

###Para obtener la MATRIZ DE CONFUSION####
envtrainsprelimpio<-na.omit(envtrainspre)
gamspre2 <- gam(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
                bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
                bio16_15 + bio18_15,
               family=binomial(link="logit"), data=envtrainsprelimpio)
#Instale el paquete gmodels y llame la librería
install.packages("gmodels")
library(gmodels)
#Para determinar un punto de corte utilice tentativamente la media de los valores ajustado
threshold<-mean(fitted(gamspre2))
threshold
#Ejecute la tabla cruzada o tabla de clasificación usando el valor real y el valor ajustado (pronosticado)
CrossTable(envtrainsprelimpio$paspre, fitted(gamspre2) > threshold,expected=FALSE, prop.r=TRUE,
prop.c=TRUE,
          prop.t=F, prop.chisq=F, chisq = FALSE, fisher=FALSE)

### sobre datos test ###

envtestsprelimpio<-na.omit(envtestspre)
attach(envtestsprelimpio)
#Estime el modelo LOGIT usando la función GLM
gamspre2 <- gam(paspre ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
                bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 +
                bio16_15 + bio18_15,
```

```
            family=binomial(link="logit"), data=envtestsprelimpio)
#Instale el paquete gmodels y llame la librería
install.packages("gmodels")
library(gmodels)
#Para determinar un punto de corte utilice tentativamente la media de los valores ajustado
threshold<-mean(fitted(gamspre2))
threshold
#Ejecute la tabla cruzada o tabla de clasificación usando el valor real y el valor ajustado (pronosticado)
CrossTable(envtestsprelimpio$paspre, fitted(gamspre2) > threshold,expected=FALSE, prop.r=TRUE, prop.c=TRUE,
        prop.t=F, prop.chisq=F, chisq = FALSE, fisher=FALSE)

#### Gradient Boosting#######

aa<-subset(sdmdatas, !is.na(bio1_15))
aphodiusGMB_PRE <- gbm.step(data=aa, gbm.x = c(2,3,4,6,9,10,11,12,13,15,17,19),
                        gbm.y = 1, family = "bernoulli",
                        tree.complexity = 2,learning.rate = 0.01,
                        bag.fraction = 0.5)

gbm.plot(aphodiusGMB_PRE, n.plots=2, write.title = FALSE) #No run

gbm.plot.fits(aphodiusGMB_PRE)# No run
find.int <- gbm.interactions(aphodiusGMB_PRE)
find.int$interactions
find.int$rank.list
gbm.perspec(aphodiusGMB, 12, 2, y.range=c(15,17), z.range=c(0,1.0))

testframepre<-aa[,-1]

Method <- factor('bio1_15', levels = levels(aa$Method))
add <- data.frame(Method)
p_pre <- predict(m.c, aphodiusGMB_PRE,
          n.trees=aphodiusGMB_PRE$gbm.call$best.trees, type="response")
```

```
p_pre <- mask(p_pre, raster(m.c, 1))
gbmevpre<-evaluate(testpresspre, testbackgspre,
aphodiusGMB_PRE,n.trees=aphodiusGMB_PRE$gbm.call$best.trees)
gbmevpre
plot(gbmevpre, 'ROC', col='blue', cex=0.1)
par(mfrow=c(1,2))
plot(p_pre, main='Presencia/background')
plot(wrld_simpl, add=TRUE, border='dark grey')
trsgbmpre <- threshold(gbmevpre, 'spec_sens')
plot(p_pre > trsgbmpre, main='Presencia/background')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_trains, pch='+', col='blue')
points(backg_trains, pch='-', cex=0.25)

gbm.plot(aphodiusGMB_PRE, n.plots=12, write.title = FALSE)
gbm.plot.fits(aphodiusGMB_PRE)
str(aphodiusGMB_PRE)
relative.influence(aphodiusGMB_PRE, n.trees=aphodiusGMB_PRE$gbm.call$best.trees)
permutation.test.gbm(object, n.trees)
gbm.loss(y,f,w,offset,dist,baseline, group, max.rank)

#### Matriz de confusion ###
aatest<-envtestsprelimpio

aphodiusGMB_PRE <- gbm.step(data=aatest, gbm.x = c(2:13),
                            gbm.y = 1, family = "bernoulli",
                            tree.complexity = 2,learning.rate = 0.01,
                            bag.fraction = 0.5)
threshold<-mean(fitted(aphodiusGMB_PRE))
threshold
#Ejecute la tabla cruzada o tabla de clasificación usando el valor real y el valor ajustado (pronosticado)
CrossTable(envtestsprelimpio$paspre, fitted(aphodiusGMB_PRE) > threshold,expected=FALSE, prop.r=TRUE,
prop.c=TRUE,
```

```
        prop.t=F, prop.chisq=F, chisq = FALSE, fisher=FALSE)

###### Random Forest############

bio.imputedspre <- rfImpute(paspre~., envtrainspre)



rfstrain<-subset(envtrains, !is.na(bio1_15))

modelpre<-paspre~bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
  bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 + bio16_15 + bio18_15

rfspre<-randomForest(modelpre,importance=TRUE,proximity=TRUE, data=bio.imputedspre)
rfspre<-randomForest(bio.imputedspre[,2:13],factor(pb_trainspre), nodes=TRUE)
plot(rfspre, main="Error")



erfspre<-evaluate(testpresspre, testbackgspre, rfspre)
erfspre
prspre<-predict(m.cstack_pre,rfspre,ext=ext)
par(mfrow=c(1,2))
plot(prspre, main='Random Forest')
plot(wrld_simpl, add=TRUE, border='dark grey')
trspre <- threshold(erfspre, 'spec_sens')
plot(prspre > trspre, main='presence/absence')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_trainspre, pch='+',col='blue')
points(backg_trainspre, pch='-', cex=0.25)

hist(treesize(rfspre)) #Histograma de nº arboles
varImpPlot(rfspre, main="Importancia") # Plot de la importancia de las variables
importance(rfspre)
importance(rfspre, type=2)
```

```
rfspre.res <- tuneRF(rfspre[,-300], rfspre[,], stepFactor=1.5)
plot(erfspre, 'ROC', col='blue', cex=0.1)
str(rfspre)

rfspre$votes
print(rfspre)

importance(modelpre, type=1, class="paspre")
rfspre$err.rate
confusionMatrix(rfspre, data$Group)
######### SVM  ############

library(kernlab)
svmpre <- ksvm(factor(paspre) ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
            bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 + bio16_15 + bio18_15, data=envtrainsprelimpio,
kernel='polydot',kpar=list(degree=2),C=10,prob.model=TRUE)
plot(envtrainspre$bio3_15, envtrainspre$bio4_15,
    col = as.integer(envtrainspre[, 5]), pch = c("o","+")[1:150 %in% svmpre@SVindex+1],
    cex = 2, xlab = "Bio3", ylab = "Bio4")

esvspre <- evaluate(testpresspre, testbackgspre, svmpre)
esvspre
plot(esvspre, 'ROC', col='blue', cex=0.1)
str(svmpre)
svmpre@error
ext<-extent(-10,6,35,55)
psspre <- predict(m.cstack_pre, svmpre, ext=ext)
par(mfrow=c(1,2))
plot(psspre, main='Support Vector Machine')
plot(wrld_simpl, add=TRUE, border='dark grey')
trsvmpre <- threshold(esvspre, 'spec_sens')
plot(psspre > trsvmpre, main='Presencia/ausencia')
plot(wrld_simpl, add=TRUE, border='dark grey')
```

```
points(pres_trainspre, pch='+')
points(backg_trainspre, pch='-', cex=0.25)

### Matriz de confusión datos test ##
svmpre <- svm(factor(paspre) ~ bio1_15 + bio2_15 + bio3_15 + bio5_15 + bio8_15 +
              bio9_15 + bio10_15 + bio11_15 + bio12_15 + bio14_15 + bio16_15 + bio18_15, data=envtrainsprelimpio,
kernel='polynomial',kpar=list(degree=2),C=10,prob.model=TRUE)
pred <- predict(svmpre, envtestsprelimpio,type="class")
# Matriz de confusión
mc <- table(pred,envtestsprelimpio[,1], dnn = c("Predicho","Observado"))
# Ordenar tabla alfabéticamente
# Por algún motivo a veces sale desordenada
mc <- mc[order(rownames(mc)),order(colnames(mc))]
cat("** SVM\n")
print(mc)
# Aciertos en %
aciertos3 <- sum(diag(mc)) / sum(mc) * 100
cat("\nCorrectamente clasificados:",round(aciertos3,2),"%\n\n\n")



##### Ensamblado ###########


modelsspre <- stack(pbspre, pdspre, mmpre, pgs4pre, pgamspre, p_pre,prspre, psspre)
names(modelsspre) <- c("bioclim","domain","mahal","glm","gam","gb","rf","svm")
modelsspre <- stack(pbspre, pdspre,pgs4pre, pgamspre, p_pre,prspre, psspre)
names(modelsspre) <- c('bioclim','domain','glm','gam','gb','rf','svm')
plot(modelsspre)


mspre <- mean(modelsspre)
plot(mspre, main='Media de predicciones')


############ Modelos geográficos #############
###Distancia geográfica#####
```

```
#first create a mask to predict to, and to use as a mask
# to only predict to land areas
seamaskpre <- crop(m.cstack_pre[[1]], ext)
distmpre <- geoDist(pres_trainspre, lonlat=TRUE)
dspre <- predict(seamaskpre, distmpre, mask=TRUE)
epre <- evaluate(distmpre, p=pres_testspre, a=backg_testspre)
epre
par(mfrow=c(1,2))
plot(dspre, main='Geographic Distance')

plot(wrld_simpl, add=TRUE, border='dark grey')
trpre <- threshold(epre, 'spec_sens')
plot(dspre > trpre, main='presence/absence')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_train, pch='+')
points(backg_train, pch='-', cex=0.25)

#### Convex hull ##############

hullspre <- convHull(pres_trainspre, lonlat=TRUE)
eshpre <- evaluate(hullspre, p=pres_testspre, a=backg_testspre)
eshpre
hspre <- predict(seamaskpre, hullspre, mask=TRUE)
plot(hspre, main='Convex Hull -Var preseleccionadas')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_trains, pch='+')
points(backg_trains, pch='-', cex=0.25)

#### Presencia/ausencia#####
pres_trainsprelimpio<-na.omit(pres_trainspre)
backg_trainsprelimpio<-na.omit(backg_trainspre)
pres_testsprelimpio<-na.omit(pres_testspre)
```

```
backg_testsprelimpio<-na.omit(backg_testspre)


idwmspre <- geoIDW(p=pres_trainsprelimpio, a=backg_trainsprelimpio)

espre <- evaluate(idwmspre, p=pres_testsprelimpio, a=backg_testsprelimpio)

espre

iwspre <- predict(seamaskpre, idwmspre, mask=TRUE)

par(mfrow=c(1,2))

plot(iwspre, main='Inv. Dist. Weighted')

plot(wrld_simpl, add=TRUE, border='dark grey')

trwspre <- threshold(espre, 'spec_sens')

pawspre <- mask(iwspre > trwspre, seamaskpre)

plot(pawspre, main='Presencia/background')

plot(wrld_simpl, add=TRUE, border='dark grey')

points(pres_train, pch='+')

points(backg_train, pch='-', cex=0.25)
```

Código R para estimar datos del año 2070

Previamente hay que descargar los datos de Wordclim

```
datos_2070<-stack(r1_2070,r2_2070,r3_2070,r4_2070,r5_2070,r6_2070,r7_2070,r8_2070,
                  r9_2070,r10_2070,r11_2070,r12_2070,r13_2070,r14_2070,r15_2070,
                  r16_2070,r17_2070,r18_2070,r19_2070)
plot(datos_2070)

ext<-extent(-10, 6, 35, 55)
#bio<-getData('worldclim', var='bio',res=0.5,lon=-10,lat=45) #left - Mapa izquierdo en el que esta España sin
Cataluña y Baleares
#bior<-getData('worldclim', var='bio',res=0.5,lon=10,lat=45) #right - Mapa derecho en el que estan Cataluña y
Baleares
ext<-extent(-10,4,35,55)
l.2070 <- crop(datos_2070,ext) # Asignar la extensión que buscamos
r.2070 <- crop(datos_2070,ext)
m.c2070 <- merge(l.2070,r.2070)# Unir los dos raster
plot(m.c2070) #Plot de la unión de raster
#reseting names to original
names(m.c2070)<-names(datos_2070)
#setting correct decimal points
m.c2070 <- m.c2070*0.1
names(m.c2070)

sr2070<-
writeRaster(m.c2070,"C:/Users/jose/Documents/pruebabio",format="ascii",bylayer=TRUE,suffix="names",overwrite
=TRUE)
m.cstack2070<-stack(m.c2070) # Compactar en un RasterStack ya que es un Brink#
m.cstack2070
writeRaster(m.c2070, filename="C:/Users/jose/Documents/pruebabio/bio2070.grs", bylayer=T, overwrite=TRUE)#
makes one .grd file for all variables.
```

```
m.cstackpre2070 <- dropLayer(m.cstack2070, c(4,6,7,13,15,17,19))
m.cstackpre2070

####Extraer valores de los Rasters######
extraspain<-cbind(acgeos$lon,acgeos$lat)
colnames(extraspain)=c('lon','lat')
presvalss2070<-extract(m.cstackpre2070,extraspain)
presvalss2070<-na.omit(presvalss2070)
set.seed(0)
backgrs2070<-randomPoints(m.cstackpre2070, 500)
absvalss2070<-extract(m.cstackpre2070, backgrs2070)
pbs2070<- c(rep(1, nrow(presvalss2070)), rep(0, nrow(absvalss2070)))
sdmdatas2070<-data.frame(cbind(pbs2070,rbind(presvalss2070,absvalss2070)))
head(sdmdatas2070)
tail(sdmdatas)
pairs(sdmdatas2070[,2:5], cex=0.1, fig=T)

######CREAMOS DATOS TRAIN - TEST######

groupspre2070<-kfold(extraspain,5)
pres_trainspre2070<-extraspain[groupspre2070 !=1, ]
pres_testspre2070<-extraspain[groupspre2070 ==1, ]
ext=extent(-10, 4, 35, 55)
backgspre2070<-randomPoints(m.cstackpre2070, n=1000, ext=ext, extf=1.25)
colnames(backgspre2070)=c('lon', 'lat')
grouppre2070<-kfold(backgspre2070,5)
backg_trainspre2070<-backgspre2070[grouppre2070 !=1, ]
backg_testspre2070<-backgspre2070[grouppre2070 ==1, ]

rspre2070=raster(m.cstackpre2070, 1)

#### UTILES PARA LOS MODELOS #######

trainspre2070 <- rbind(pres_trainspre, backg_trainspre)
pb_trainspre <- c(rep(1, nrow(pres_trainspre)), rep(0, nrow(backg_trainspre)))
envtrainspre <- extract(m.cstack_pre, trainspre)
envtrainspre <- data.frame( cbind(paspre=pb_trainspre, envtrainspre) )
head(envtrainspre)
envtrainspre
testpresspre2070<- data.frame( extract(m.cstackpre2070, pres_testspre2070) )
testbackgspre2070 <- data.frame( extract(m.cstackpre2070, backg_testspre2070) )



#### Gradient Boosting#######

aa2070<-subset(sdmdatas2070, !is.na(ac45bi501))
aphodiusGMB_PRE2070 <- gbm.step(data=aa2070, gbm.x = c(2:13),
                        gbm.y = 1, family = "bernoulli",
                        tree.complexity = 2,learning.rate = 0.01,
                        bag.fraction = 0.5)

gbm.plot(aphodiusGMB_PRE2070, n.plots=2, write.title = FALSE)

gbm.plot.fits(aphodiusGMB_PRE)
find.int <- gbm.interactions(aphodiusGMB_PRE2070)
find.int$interactions
find.int$rank.list
gbm.perspec(aphodiusGMB_PRE2070, 12, 2, y.range=c(15,17), z.range=c(0,1.0))


p_pre2070 <- predict(m.cstackpre2070, aphodiusGMB_PRE2070,
             n.trees=aphodiusGMB_PRE2070$gbm.call$best.trees, type="response")
```

```
p_pre <- mask(p_pre, raster(m.c, 1))
gbmevpre2070<-evaluate(testpresspre2070, testbackgspre2070,
aphodiusGMB_PRE2070,n.trees=aphodiusGMB_PRE2070$gbm.call$best.trees)
gbmevpre2070
plot(gbmevpre2070, 'ROC', col='blue', cex=0.1)
par(mfrow=c(1,2))
trsgbmpre2070 <- threshold(gbmevpre2070, 'spec_sens')
plot(p_pre2070 > trsgbmpre2070, main='Presencia/background')
plot(wrld_simpl, add=TRUE, border='dark grey')
points(pres_trainspre2070, pch='-', col='blue')
points(backg_trainspre2070, pch='-', cex=0.25)
```

Bibliografía

Álvarez-Uría T. P; Zamorano C. C. 2001--. La biodiversidad en España. *Ambienta*. Madrid: Secretaría General Técnica Ministerio de Agricultura, Alimentación y Medio Ambiente. ISSN: 1577-9491

Báguena Corella, L., 1967. Scarabaeoidea de la fauna íbero-balear y pirenaica. Instituto Español de Entomología, CSIC (ed). Madrid. 576pp.

Cabrero-Sañudo F. J; Lobo J. M. 2003. Reconocimiento de los factores determinantes de la riqueza de especies: el caso de los Aphodiinae (Coleoptera, Scarabaeoidea, Aphodiidae) en la Península Ibérica. *Graellsia*, Vol 59, (2-3): 155-177 doi: 10.3989/graellsia.2003.v59.i2-3.240

Cambefort, Y. & Hanski, I. 1991. Dung beetle population biology, p. 36– 50. In: Hanski, I. & Cambefort, Y. (eds.). Dung beetle ecology. Princeton, Princeton University Press, 481 p.

Cassiini, M.H. 2011. Ecological principles of species distribution models: the habitat matching rule. Journal of Biogeography 38: 2057-2065.

Cayuela L. 2010. Análisis de datos ecológicos en R. [Blog]. [España]: luiscayuela.com. [Consulta: junio 2016]. Disponible en: http://luiscayuela.blogspot.com.es/

Charles Elton, 1927. Chapter. V: 'The Animal Community'. En: Julian S. Huxley (ed) *Animal Ecology*. [En línea]. London. [Consulta 15 de Julio de 2016]. Texto escaneado. Disponible en: https://archive.org/stream/animalecology00elto#page/n9/mode/2up

Cruz-Cárdenas Gustavo; Villaseñor J. Luis; López-Mata Lauro; Martínez-Meyer Enrique; Ortiz Enrique. 2004. Selección de predictores ambientales para el modelado de la distribución de especies en Maxent. *Revista Chapingo Serie Horticultura*, Universidad Autónoma Chapingo. Volumen XX, nº 2: mayo-agosto 2014, p. 188-201 ISSN electrónico: 2007-4034

Dellacasa M. & Dellacasa G., 2006. Scarabaeidae: Aphodiinae. New nomenclatorial and taxonomic acts, and comments (p. 31). In Löbl i. & Smetana a. (Eds.): Catalogue of Palaearctic Coleoptera. - (Apollo Books), Stenstrup, 3: 1-690

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. and Yates, C. J. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions, 17: 43–57. doi:10.1111/j.1472-4642.2010.00725.x

Global Biodiversity Information Facility, 2001. Global Biodiversity Information Facility. [Sitio Web]. Madrid: GBIF España. Unidad de Coordinación. [Consulta: 4 abril de 2016]. Disponible en: http://datos.gbif.es

Grinnell, J. 1924. Geography and evolution. Ecology 5: 225-229 Reprinted 1943. In Joseph Cgrinnell's Philosophy of Nature. Berkley: University of California Press. 151-157ppp.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. [En línea]. USA: Chapman & Hall, Laboratorios A T &T. [Consulta: agosto 2016]. Disponible en: https://books.google.es/books?id=qa29r1Ze1coC&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Halffter, G. & Edmonds, W.D. 1982. The nesting behavior of dung beetles (Scarabaeinae): An ecological and evolutive approach. México D.F., Man and the Biosphere Program UNESCO, 177 p.

Hortal J; M. Lobo J; del Rey L, 2006. Distribución y patrones de diversidad de los Afódidos en la Comunidad de Madrid (coleoptera, scarabaeoidea, aphodiidae, aphodiinae y psammodiinae). *Graellsia*, 62(número extraordinario): 439-460ppp

Hutchinson, G. E., 1957. Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology, 22, 145-159ppp.

Hortal J; M. Lobo J; del Rey L, 2006. Distribución y patrones de diversidad de los Afódidos en la Comunidad de Madrid (coleoptera, scarabaeoidea, aphodiidae, aphodiinae y psammodiinae). *Graellsia*, 62(número extraordinario): 439-460ppp. doi:10.3989/graellsia.2006.v62.iExtra

Hortal J; M. Lobo J; Martín Piera, F., 2003. Una estrategia para obtener regionalizaciones bióticas fiables a partir de datos incompletos: el caso de los escarabeidos (coleoptera, scarabaeinae) ibérico-baleares. *Graellsia*, 59(2-3): 331-344ppp. doi:10.3989/graellsia.2003.v59.i2-3

J. Hijmans Robert ; Elith Jane, 2016. Species distribution modeling with R. [Sitio Web]. dismo: Species Distribution Modeling. [Consulta: 10 de Julio de 2016]. Disponible en: https://cran.r-project.org/web/packages/dismo

Lobo, J. M. 2000, Hacia un proyecto CYTED para el inventario y estimación de la diversidad entomológica en Iberoamérica: PrIBES. [en línea]. Aragón: m3m-*Monografías del Tercer Milenio*, Sociedad Entomológica Aragonesa. [Consulta: 20 de abril de 2016]. Texto en HTML. Disponible en: http://entomologia.rediris.es/pribes/Lobo/Subproyecto3.html

Lobo, J. M; Martín Piera, F., 1991. La creación de un Banco de Datos zoológico sobre los Scarabaeidae (Coleoptera, Scarabaeoidea) ibero-baleares: Una experiencia piloto. ELYTRON. Journal European Association Coleopterology, 5: 31-37.

Lobo, J. M; Martín Piera, F., 2002. Estableciendo las bases de un proyecto Iberoamericano para la estimación e inventario de la diversidad entomológica. Zaragoza: m3m: Monografías Tercer Milenio. Vol 2: 321–327ppp. ISBN: 84–922495–8–7.

Naoki, Kazuya et al. Comparación de modelos de distribución de especies para predecir la distribución potencial de vida silvestre en Bolivia. Ecología en Bolivia [online]. 2006, vol.41, n.1 [Consultado: septiembre 2016], pp. 65-78. Disponible en: <http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1605-25282006000700005&lng=es&nrm=iso>. ISSN 2075-5023.

Phillips, S. J.; Anderson, R.P.; Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190: 231-259.

P Vaquez Diego, 2005. Reconsiderando el nicho hutchinsoniano. *Ecología Austral* [En línea]. 15:149-158. [Consulta 15 de Julio de 2016]. ISSN 1667-782X. Disponible en: Asociación Argentina de Ecología. http://ojs.ecologiaaustral.com.ar/index.php/index/search/search

Quintas I., 2000. Modelo aditivo generalizado GAM: Regresión no lineal y no paramétrica. ISBN: 978-607-28-0154-7

Slack G, Nancy, 2010. G. Evelyn Hutchinson and the invention of modern ecology. [En línea]. Yale: University Press. [Consulta 15 de Julio de 2016]. Texto plano. Disponible en: https://books.google.es/books?id=OaW_rFxTm4MC&printsec=frontcover&dq=Evelyn+Hutchinson&hl=es&sa=X&ved=0ahUKEwjupqWe5ZjPAhUEPxoKHebdDNUQ6AEIHzAA#v=onepage&q=Evelyn%20Hutchinson&f=false

Steven J, Phillips; Elith J, 2013. On estimating probability of presence from use–availability or presence–background data. *Ecology*, 94(6), 1409–1419ppp.

Stockwell, DRB. and Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science, vol. 13, no. 2, p. 143-158.

Verdú, J. R. and Galante, E. 2002. Climatic stress, food availability and human activity as determinants of endemism patterns in the Mediterranean region: the case of dung beetles (Coleoptera, Scarabaeoidea) in the Iberian Peninsula. *Diversity and Distributions*, 8: 259–274. [Consulta: 10 de junio de 2016]. doi:10.1046/j.1472-4642.2002.00151.x. Disponible en: http://onlinelibrary.wiley.com/doi/10.1046/j.1472-4642.2002.00151.x/abstract

WorldClim. 2012. Global Climate Data: Free climate data for ecological modeling and GIS. [Sitio Web]. Museum of Vertebrate Zoology, University of California, Berkeley. [Consullta: 4 de abril de 2016] Disponible en: http://www.worldclim.org/current. Fecha último acceso: 3 octubre 2012.