Universitat Oberta de Catalunya

uoc.edu

Modelización espacial de la distribución de los anfibios ibéricos a partir de variables ambientales

Elena Herrero Esteban
Máster en Bioestadística y Bioinformática

Área 2, Subárea 11: Análisis de Datos y Técnicas de Clustering

Nombre Consultor/a: Daniel Fernández Martínez

Nombre tutor externo: Pablo Alberto Refoyo Román *(Facultad de Ciencias Biológicas, Universidad Complutense de Madrid)*

Nombre Profesor/a responsable de la asignatura: Carles Ventura Royo

Fecha Entrega 2/06/2022

FICHA DEL TRABAJO FINAL

Título del trabajo	Modelización espacial de la distribución de los anfibios ibéricos a partir de variables ambientales
Nombre del autor	Elena Herrero Esteban
Nombre del consultor/a	Daniel Fernández Martínez
Nombre del tutor externo	Pablo Alberto Refoyo Román
Nombre del PRA	Carles Ventura Royo
Fecha de entrega	06/2022
Titulación	Máster en Bioestadística y Bioinformática
Área del Trabajo Final	M0.178 TFM - Estadística y Bioinformática 2 aula 1
Idioma del trabajo	Castellano
Número de créditos	15
Palabras clave	*Riqueza, anfibios, variables climáticas, random forest, supervised Learning, modelización.*

Resumen del Trabajo

Los anfibios están experimentando un gran declive poblacional a nivel mundial, sufriendo el mayor porcentaje de extinciones por año en el último siglo, lo que les convierte en el grupo más amenazado del reino animal según la Unión Internacional para la Conservación de la Naturaleza. España es un *hotspot* de biodiversidad donde se encuentran representadas el 35% de todas las especies de anfibios europeos, además de poseer varios endemismos ibéricos. Evaluar la importancia de las variables bióticas, abióticas y antrópicas en la riqueza de anfibios en España peninsular y Baleares es clave para orientar los planes de Gestión y Conservación en el escenario de cambio global. En el presente trabajo se emplean Análisis de Componentes Principales y Modelos Lineales Generalizados para generar un modelo predictivo que mediante herramientas de geoprocesamiento espacial se representa en un Mapa de Idoneidad de la riqueza de anfibios en el territorio. Random Forest es la técnica de Machine Learning más apropiada para analizar la relación de las variables con la riqueza de especies. Las variables climáticas resultan relevantes para la riqueza de anfibios, siendo las de precipitaciones las más significativas (BIO15, BIO19, BIO16, BIO13). Las zonas del norte y occidente peninsular son las más idóneas para la riqueza de anfibios. La riqueza de anfibios puede emplearse como un parámetro bioindicador fiable del cambio climático. La combinación de

herramientas estadísticas, bioinformáticas y de geoprocesamiento representan una poderosa herramienta para los estudios de distribución de especies y su distribución futura en respuesta a cambios ambientales.

Abstract

Amphibians are experimenting a large population decline worldwide, having the highest percentage of extinctions per year in the last century, which makes them the most threatened group in the animal kingdom according to the International Union for Conservation of Nature. Spain is a biodiversity hotspot, where 35% of amphibian european species are represented, in addition to having several Iberian endemisms. Evaluating the importance of biotic, abiotic and anthropic variables in the richness of amphibians in Spain and the Balearic Islands is key to being able to guide Management and Conservation plans in the current global change scenario. In this paper, Principal Component Analysis and Generalized Linear Models are used to generate a predictive model that, through spatial geoprocessing tools, is represented in a Suitability Map of the richness of amphibians in the territory. Random Forest is the most appropriate Machine Learning technique to analyze the relationship of variables with species richness. The climatic variables are relevant for the richness of amphibians, with precipitation being the most significant (BIO15, BIO19, BIO16, BIO13). The northern and western areas within the Iberian Peninsula are the most suitable areas for the richness of amphibians. Amphibian richness can be used as a reliable bioindicator parameter in the context of climate change. The combination of statistics, bioinformatics and geoprocessing tools achieves a complete toolbox for species distribution studies and their future distribution due to environmental changes.

Índice

Lista de figuras

Lista de tablas

1 Resumen

Los anfibios están experimentando un gran declive poblacional a nivel mundial, sufriendo el mayor porcentaje de extinciones por año en el último siglo, lo que les convierte en el grupo más amenazado del reino animal según la Unión Internacional para la Conservación de la Naturaleza. España es un *hotspot* de biodiversidad donde se encuentran representadas el 35% de todas las especies de anfibios europeos, además de poseer varios endemismos ibéricos.

Evaluar la importancia de las variables bióticas, abióticas y antrópicas en la riqueza de anfibios en España peninsular y Baleares es clave para orientar los planes de Gestión y Conservación en el escenario de cambio global.

En el presente trabajo se emplean Análisis de Componentes Principales y Modelos Lineales Generalizados para generar un modelo predictivo que mediante herramientas de geoprocesamiento espacial se representa en un Mapa de Idoneidad de la riqueza de anfibios en el territorio. Random Forest es la técnica de Machine Learning más apropiada para analizar la relación de las variables con la riqueza de especies.

Las variables climáticas resultan relevantes para la riqueza de anfibios, siendo las de precipitaciones las más significativas (BIO15, BIO19, BIO16, BIO13). Las zonas del norte y occidente peninsular son las más idóneas para la riqueza de anfibios. La riqueza de anfibios puede emplearse como un parámetro bioindicador fiable del cambio climático. La combinación de herramientas estadísticas, bioinformáticas y de geoprocesamiento representan una poderosa herramienta para los estudios de distribución de especies y su distribución futura en respuesta a cambios ambientales.

2 Introducción

2.1 Contexto y justificación del trabajo

Los anfibios fueron los primeros vertebrados que cambiaron el medio acuático por el terrestre hace más de 350 millones de años. Han colonizado prácticamente todos los ecosistemas terrestres, excepto zonas con climatologías extremas, frías y secas [1]. *Amphibia* es una clase compuesta a su vez por tres órdenes, *Gymnophiona* (excavadores vermiformes, sin extremidades, como las cecilias), *Caudata* (poseen cola y cuatro patas, como las salamandras y tritones) y *Anura* (sin cola, como los sapos y las ranas). En España peninsular y Baleares únicamente encontramos anfibios del orden *caudata* y *anura*, ya que el orden *Gymnophiona* habita selvas tropicales húmedas de América, África, la India e Indochina.

ORDEN *ANURA*

Los anuros son un grupo de anfibios que incluye las ranas y los sapos. Se estima que en este orden existen más de 5.000 especies, repartidas en 48 familias [4]. Las características básicas de dicho grupo son que tienen un cuerpo corto, ensanchado y con las patas posteriores modificadas para el salto terminando en cinco dedos palmeados, mientras que las patas anteriores tienen cuatro dedos [2]. La mayoría de los anfibios tienen un ciclo de vida bifásico, con una etapa larvaria que se desarrolla en medios acuáticos, y una fase post-metamórfica que tiene lugar principalmente en el medio terrestre [3]. El orden *anura* tras la etapa larvaria pierde la cola, lo que los diferencia del orden *caudata*. Tienen su hábitat cerca de ríos, lagos y estanques, donde tiene lugar la reproducción y su fase larvaria. Su respiración es cutánea por lo que son muy sensibles a la abrasión y a la deshidratación. La mayor parte son nocturnos y para controlar el equilibrio hídrico corporal están más activos cuando hay humedad o pasan mucho tiempo dentro del agua.

Los sapos tienen la piel seca, rugosa y habitan en suelos húmedos excavando galerías. Las ranas son mucho más dependientes de medios con humedad, tienen la piel lisa, siempre humectada porque su respiración es cutánea, y poseen locomoción trepadora o acuática [2] [4]. Las figuras 1, 2 y 3 muestran especies de anfibios del orden *anura*, los dos primeros ampliamente distribuidos en la península ibérica y catalogados como Preocupación Menor en la lista roja de la UICN para España, mientras que la figura 3 está catalogada como Casi Amenazada tanto a nivel global como en España.

Figura 2. ***Discoglossus galganoi***
Sapillo pintojo ibérico

Figura 1. ***Epidalea calamita***
Sapo corredor

Figura 3. ***Pelobates cultripes***
Sapo de espuelas

ORDEN *CAUDATA*

Los *caudata* son un grupo de anfibios que incluye salamandras, tritones y gallipatos. En este orden se estiman más de 580 especies, repartidas en 9 familias. La estructura del cuerpo es cabeza, tronco y cola, con patas del mismo tamaño, como podemos apreciar en las figuras 4 y 5. Tienen su hábitat en bosques húmedos y cerca de ríos y lagos. Algunas especies viven exclusivamente en el agua, otras habitan la tierra pero necesitan zonas de constante humedad, tanto para su reproducción, como para evitar la desecación de la piel. Tienen distintos tipos de respiración; las larvas y algunos adultos acuáticos emplean branquias, mientras que las especies más adaptadas al medio terrestre tienen respiración pulmonar y cutánea [5] [6].

Figura 4. ***Triturus marmoratus***
Tritón jaspeado

Figura 5. ***Salamandra salamandra***
Salamandra común

El papel de los anfibios en los ecosistemas es fundamental: participan en los ciclos de nutrientes y en la dispersión de semillas, son importantes bioturbadores (modifican la distribución de los sedimentos) y tienen un papel fundamental en los flujos de energía a través de cadenas tróficas (como depredador y presa). Proveen de servicios ecosistémicos tan importantes como el control biológico de plagas que atacan a la agricultura además de depredar sobre insectos que diseminan graves enfermedades humanas (mosquitos del género *Aedes* transmisores de los virus zika y dengue o mosquitos del género *Anopheles* transmisores de la malaria). Por otro lado, son presas de mamíferos, aves, reptiles y peces, por lo que su declive poblacional afecta a toda la cadena trófica [7].

También son bioindicadores de la salud ambiental de los ecosistemas puesto que su piel es extremadamente permeable; esto los hace muy vulnerables a la absorción de sustancias tóxicas que les causan la muerte, por lo que su presencia en los ecosistemas es indicadora de salud ambiental [8] [9].

En el escenario actual de cambio climático los anfibios son uno de los grupos de vertebrados que más afectado se ve, por sus necesidades hídricas, su dependencia de las variables ambientales y su sensibilidad a los contaminantes.

Existen evidencias de que los cambios a corto plazo en el clima pueden ocasionar el declive de las poblaciones de anfibios [10]. Una de las consecuencias de estos fenómenos climatológicos son los cambios en la distribución de las poblaciones de anfibios [11]. Los cambios bruscos de temperatura y las nevadas tardías son las principales causas de mortalidad de anfibios durante el periodo reproductor [12]. Los huevos de los anfibios no tienen pared protectora por lo que deben depositarse en agua para que no se sequen [13], situación que se complica por los periodos de sequía cada vez más frecuentes. El aumento de las temperaturas con la consiguiente desecación de masas de agua, tanto estacionales como permanentes, provoca que las larvas no puedan completar su metamorfosis [14]. El aumento de la

temperatura del agua adelanta el desarrollo larvario, disminuyendo el tamaño de los ejemplares en plena metamorfosis y afectando así a su supervivencia [12].

La conservación de anfibios es una necesidad. Según la Unión Internacional para la Conservación de la Naturaleza (UICN), casi un tercio de las especies conocidas de anfibios están catalogadas bajo alguna categoría de amenaza, lo que les convierte en el grupo más amenazado del reino animal. Hay que tener en cuenta que la evaluación no está completa, además no se conocen todos los animales, plantas y hongos que existen en el planeta. Los datos son muy alarmantes porque de las 7.261 especies de anfibios que se estiman en el mundo, el 41% está en declive [15] como podemos observar en la figura 6. La figura 7 muestra que en el último siglo los anfibios son el grupo de vertebrados que han sufrido un mayor porcentaje de extinciones por año.

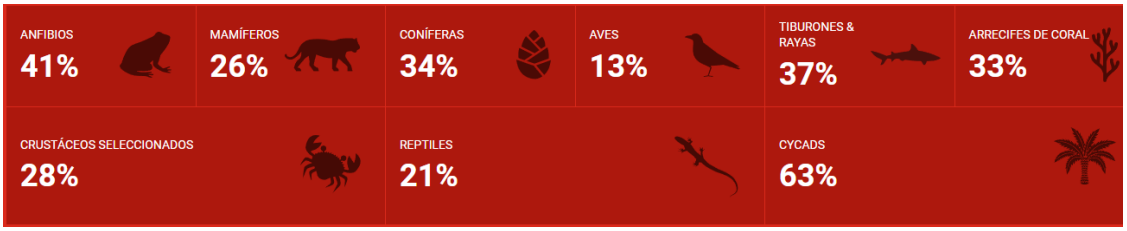


Figura 6. Especies evaluadas por la Lista Roja de la UICN a nivel mundial
[Fuente: Lista Roja de las Especies Amenazadas de la UICN] [15]

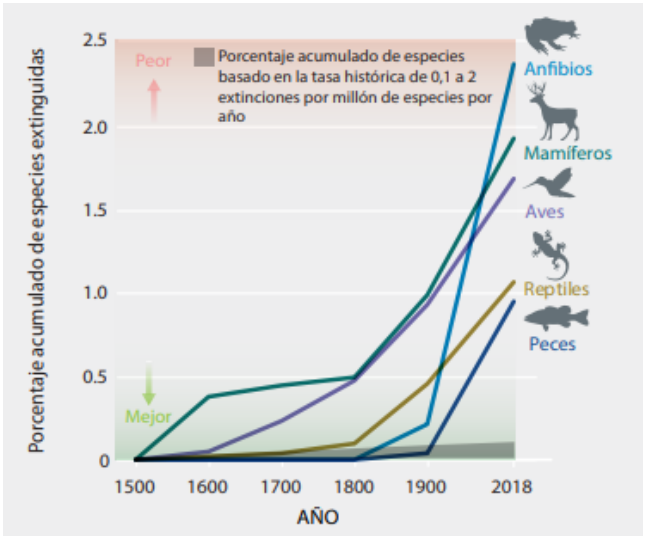


Figura 7. Extinciones desde 1500 para grupos de vertebrados
[Fuente: IPBES, Informe de la Evaluación Mundial sobre la diversidad biológica y los servicios de los ecosistemas] [16]

España es uno de los países más ricos en biodiversidad de Europa, por su posición geográfica entre dos continentes y la insularidad de algunos territorios. En España se encuentra representado el 35% de todas las especies de anfibios de Europa. Según el Atlas y Libro Rojo de Anfibios y Reptiles en España de 2002, un tercio de las especies españolas de anfibios están amenazadas. La península ibérica es un área importante para la conservación

de anfibios por la existencia de varias especies endémicas, algunas de las cuales se encuentran en Peligro Crítico de Extinción como por ejemplo el tritón del Montseny (*Calotriton arnoldi),* En Peligro como la *Rana pyrenaica,* y Vulnerable como el sapo partero bético (*Alytes dickhilleni)*, el sapillo balear (*Alytes muletensis)* y la salamandra rabilarga (*Chioglossa lusitanica)*. Estos datos están anticuados y es necesaria una reevaluación a nivel europeo del estado de conservación de los anfibios [18] como muestran las estadísticas de las figuras 8 y 9.

Total Especies	Evaluadas UICN (nº/% del total)	Desactualizadas (nº / % respecto a evaluadas)
39	33/85%	24/62%

Figura 8. Tabla resumen del estado de actualización de las especies de anfibios presentes en España incluidas en la lista roja de la UICN a nivel europeo.
[Fuente: Red List UICN septiembre 2019] [17]

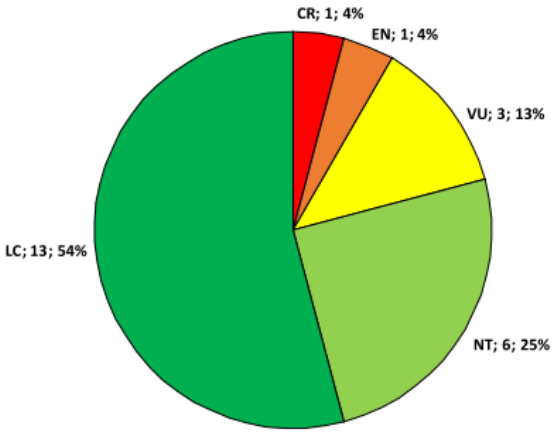


Figura 9. Número y porcentaje de anfibios de España que necesitan una reevaluación según su categoría de amenaza.
Las siglas corresponden a: EX=extinto, EW=extinto en estado silvestre, CR=en peligro crítico, EN=en peligro, VU=vulnerable, NT=casi amenazado, DD=datos insuficientes, LC=preocupación menor.
[Fuente: Red List UICN septiembre 2019] [17]

Las principales amenazas de las poblaciones de anfibios de España son la destrucción de hábitat por ocupación del suelo e intensificación de usos, que además conlleva fragmentación de hábitats; presencia de contaminantes químicos en el medio; introducción de especies exóticas invasoras (alimentación o mascotismo); exceso de radiación ultravioleta; atropellos en carreteras, y enfermedades y parásitos como la quitridiomicosis [19].

La mayoría de los estudios coinciden en que los gradientes ambientales explican la distribución de los anfibios [20] [21] [22] y por tanto es interesante estudiar si la variación de estos es la principal causante de su declive poblacional.

Sin embargo, y a pesar de su aparente relevancia en los ecosistemas como indicadores de cambio climático, son pocos los estudios de analizar el peso que cada tipo de variables (antrópicas, climáticas, ambientales, etc.) presentan en la distribución de las especies en el ámbito ibérico. Sillero et al., [23] establece los patrones de distribución para todo el grupo y la incidencia de las variables, pero no indica el valor relativo de cada una de ellas.

El objetivo de este trabajo es conocer las variables ambientales y antrópicas que más influyen en la distribución de la riqueza de anfibios de España peninsular y Baleares, ya que este conocimiento permitirá optimizar los esfuerzos de muestreo para los planes de gestión y conservación de las poblaciones.

2.2 Objetivos del trabajo

Objetivos generales

Los dos objetivos generales del estudio son:

1. Creación de un algoritmo ad-hoc basado en los datos de la relevancia de cada variable en la distribución de los anfibios para generar un mapa de distribución de idoneidad considerando el peso relativo que cada variable tiene en la distribución del grupo.

2. Evaluar la importancia de las variables bióticas, abióticas y antrópicas en la riqueza de anfibios en España peninsular y Baleares.

Objetivos específicos

- Crear una base de datos a partir de datos georreferenciados de presencia de las especies de anfibios registrados en la base de datos del Ministerio de Transición Ecológica, 19 variables climáticas de *WorldClim* (temperaturas y precipitaciones), el mapa de altitud del Instituto Geográfico Nacional, la incidencia antrópica (*Global footprint network*) y presencia de masas de agua.

- Preprocesar la base de datos y estudiar las variables con un Exploratory Data Analysis (EDA).

- Determinar mediante Análisis Estadístico y Machine Learning qué variables bióticas, abióticas y antrópicas tienen más peso en la riqueza de especies.

- Modelizar la idoneidad del territorio en la península ibérica y Baleares con el fin de detectar áreas de presencia no conocidas.

2.3 Enfoque y método seguido

Para el primer objetivo, tras un extenso estudio bibliográfico se decidió agrupar las presencias/ausencias de las diferentes especies en la variable riqueza. Dicha variable explica mejor la pérdida de biodiversidad que cualquier otra variable ya que engloba la diversidad existente de los grupos biológicos, sin embargo, su estudio no está muy extendido por la dificultad que conlleva analizar la relevancia de las variables en la distribución de grupos completos en lugar de las específicas para especies concretas. Se decidió elaborar un modelo predictivo que pudiese ser extrapolado a un Sistema de Información Geográfica para representar en forma de mapa las posibles zonas de presencia de anfibios por adecuación a las variables ambientales del territorio.

Para el segundo objetivo, se decidió emplear Random Forest. Este algoritmo de machine learning presenta numerosas ventajas, maneja cientos de variables sin excluir ninguna, es muy útil cuando se poseen datos ruidosos o redundantes, como es nuestro caso puesto que las 19 variables climáticas empleadas de precipitación y temperatura están muy correlacionadas entre ellas, y estima qué variables son las más importantes para el modelo.

2.4 Planificación del trabajo

Se ha estipulado que la duración de cada tarea sea de una semana natural. Esta planificación es meramente orientativa ya que si se consigue avanzar más una semana se seguirá adelantando trabajo de las siguientes, al igual que si hay problemas técnicos y se produce un retraso, se retrasarán el resto de las actividades.

2.4.1 Tareas

T1. Revisión bibliográfica de Análisis Estadísticos y técnicas de Machine Learning

7/03/2022 – 13/03/2022

Esta primera semana se empleará para terminar de decidir y perfilar los análisis que se llevarán a cabo en consenso con ambos tutores.

T2. Creación base de datos

14/03/2022 – 20/03/2022

En base a lo decidido la anterior semana se creará una base de datos que se modificará en función de los requerimientos de los algoritmos de Machine Learning elegidos.

T3. Análisis Estadísticos

21/03/2022 – 3/04/2022

- EDA (Exploratory Data Analysis)
- Análisis de Correlaciones
- ACP (Análisis de Componentes Principales)
- Modelo Lineal Generalizado con distribución de Poisson
- Creación de un algoritmo predictivo empleando los coeficientes obtenidos en el modelo lineal generalizado de Poisson

T4. Elaboración Informe del Desarrollo del Trabajo. Fase I

4/04/2022 -10/04/2022

Esta semana se escribirá el informe de la PEC2.

T5. Algoritmos Machine Learning

11/04/2022-8/05/2022

- Preparación base de datos para los algoritmos de Machine Learning.
- División data *train* y data *test*.
- Entrenamiento del modelo.
- Predicción y evaluación.

T6. Elaboración Informe del Desarrollo del Trabajo. Fase II

9/05/2022 – 15/05/2022

Esta semana se escribirá el informe de la PEC3.

T7. Elaboración mapa de idoneidad de la distribución de anfibios

16/05/2022-22/05/2022

Con la fórmula creada a partir de los coeficientes obtenidos en el modelo lineal generalizado con distribución de Poisson (*T3.Análisis Estadísticos*) se generará un mapa con el programa ArcMap que detectará las áreas para establecer las zonas más idóneas para la presencia de una mayor diversidad de especies (riqueza de especies).

T8. Cierre de la memoria

23/05/2022 –2/06/2022

En estas últimas semanas se terminará de redactar la memoria.

T9. Elaboración de la presentación

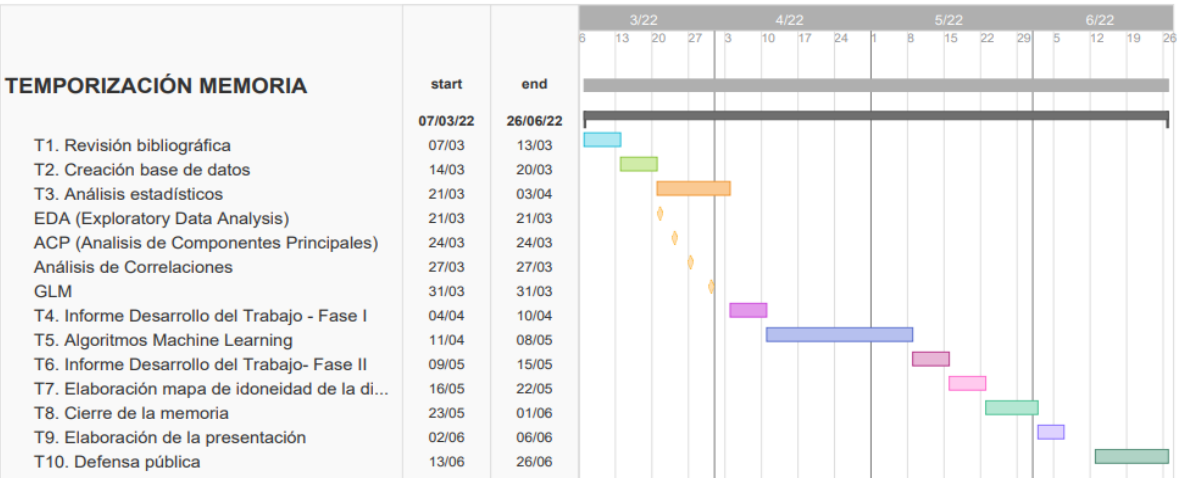
2/06/2022 - 6/06/2022

T10. Defensa pública

13/06/2022 –23/06/2022

2.4.2 Calendario

El siguiente Diagrama de Gantt, elaborado en la plataforma *team gantt*, recoge la temporización de tareas establecida:



TEMPORIZACIÓN MEMORIA	start	end
	07/03/22	26/06/22
T1. Revisión bibliográfica	07/03	13/03
T2. Creación base de datos	14/03	20/03
T3. Análisis estadísticos	21/03	03/04
EDA (Exploratory Data Analysis)	21/03	21/03
ACP (Analisis de Componentes Principales)	24/03	24/03
Análisis de Correlaciones	27/03	27/03
GLM	31/03	31/03
T4. Informe Desarrollo del Trabajo - Fase I	04/04	10/04
T5. Algoritmos Machine Learning	11/04	08/05
T6. Informe Desarrollo del Trabajo- Fase II	09/05	15/05
T7. Elaboración mapa de idoneidad de la di...	16/05	22/05
T8. Cierre de la memoria	23/05	01/06
T9. Elaboración de la presentación	02/06	06/06
T10. Defensa pública	13/06	26/06

2.4.3 Hitos

- ✓ Creación base de datos.
- ✓ Análisis estadísticos (EDA, Análisis de Correlación, ACP y GLM con distribución de Poisson).
- ✓ Entrega Informe Desarrollo del Trabajo Fase I.
- ✓ Aplicación Algoritmos Machine Learning.
- ✓ Entrega Informe Desarrollo del Trabajo Fase II.
- ✓ Representación gráfica final: elaboración mapa idoneidad distribución anfibios España peninsular y Baleares.

2.4.4 Análisis de riesgos

- Limitación temporal por la gran cantidad de análisis que se quieren llevar a cabo. Puede que los objetivos planteados sean demasiado optimistas en cuanto al tiempo del que se dispone y finalmente ciertos análisis no den tiempo de completarlos.
- Poca capacidad de cómputo para realizar los análisis. Existen limitaciones técnicas porque el ordenador del que dispongo no tiene una gran capacidad de procesamiento de grandes bases de datos, lo que ralentizará la ejecución de los análisis.
- Necesidad de crear otra base de datos teniendo en cuenta más variables (como por ejemplo, georreferenciar las zonas de presencia-ausencia) para poder emplear las técnicas de Machine Learning.

2.5 Breve sumario de contribuciones y productos obtenidos

Al finalizar este trabajo se espera obtener:

1. Base de datos.
2. Plan de trabajo.
3. Memoria.
4. Presentación virtual.
5. Publicación de un artículo en una revista como *Amphibia-Reptilia* [https://brill.com/view/journals/amre/amre-overview.xml]

2.6 Breve descripción de los otros capítulos de la memoria

En el capítulo del Estado del arte se describe la hipótesis de partida de este trabajo y como se plantea resolverla, teniendo en cuenta los métodos estadísticos y de machine learning empleados por otros grupos de investigación para modelizar la distribución de distintos grupos de especies y así seleccionar el método más apropiado con el cual conseguir los objetivos de nuestro trabajo.

En el apartado de Metodología se relata la creación de la base de datos y se explican teóricamente los análisis que se van a llevar a cabo en el apartado de resultados; Exploratory Data Analysis, ACP, GLM de Poisson y Random Forest.

En Resultados se expone el procesamiento de la base de datos; el análisis individual de las variables y el *test* Shapiro-Wilks para comprobar que siguen una distribución normal; análisis bivariado para ver cómo interactúan entre ellas; Análisis de Correlaciones de Spearman para comprobar si están asociadas; los análisis estadísticos (ACP, GLM) llevados a cabo para crear un algoritmo final que permita la representación de la predicción de la riqueza en un mapa, y la selección de las variables más influyentes en la riqueza de anfibios mediante técnicas de machine learning (Random Forest).

En la Discusión se analizan los resultados obtenidos en base a los conocimientos previos, a lo aportado por otros autores y se responde a las hipótesis de la investigación planteada y dando una interpretación biológica de estos resultados comparándolos con la bibliografía existente.

Las Conclusiones sintetizan los resultados obtenidos en dicho trabajo, las líneas de futuro que se podrían seguir para mejorarlo y ampliarlo y el seguimiento de la planificación, con los contratiempos que han ido surgiendo y como ha variado la temporización inicial. También se plantea si se debería haber utilizado otra metodología distinta.

3 Estado del arte

El clima es la fuerza abiótica con más peso en las distribuciones de especies a amplias escalas espaciales [24]. Las distribuciones pueden verse limitadas por los requisitos de temperatura o humedad [25]. El aumento de temperaturas en la península ha provocado una rápida respuesta al cambio climático para las especies de anfibios, que han pasado a ocupar zonas de alta montaña, con mayor altitud y temperaturas más frescas [26].

Los anfibios corren riesgo de extinguirse silenciosamente por tener datos insuficientes (categoría DD de la UICN) sobre su situación poblacional [27]. El mayor problema para detectar el declive de especies y poblaciones de anfibios en el territorio español es que no se poseen series históricas de datos que permitan detectar cambios en su abundancia y distribución. El último inventario que se hizo fue en 2002 para la creación del Atlas y Libro Rojo de los Anfibios y Reptiles de España [28]. Sin una estrategia común de muestreo, se obtienen representaciones sesgadas de la distribución de la riqueza de anfibios.

En este trabajo queremos conocer la dependencia de la distribución de la riqueza de anfibios de las principales variables climáticas (temperatura y precipitación), antrópicas y topográficas (orilla y altitud).

Actualmente existe un Programa para el Seguimiento a largo plazo de las poblaciones de anfibios y reptiles en el territorio español (Proyecto SARE) [29]. Conocer el mapa actual de idoneidad del territorio (*suitability model)* permitirá focalizar el esfuerzo de muestreo en los puntos donde se identifique que potencialmente podría haber más riqueza de especies, para tener datos reales de la situación actual de las especies de anfibios en España. Esto, junto con el descubrimiento de qué variables climáticas tienen más peso en esta distribución permitirá orientar los planes de gestión y conservación de anfibios en el escenario actual de cambio climático y declive poblacional.

Se han desarrollado diferentes estrategias estadísticas e informáticas que permiten predecir y explicar la distribución de la biodiversidad: técnicas de teledetección, análisis multivariante y análisis espacial con Sistemas de Información Geográfica [30] son las más empleadas en la actualidad.

Machine Learning es una técnica puntera para este tipo de investigaciones. Las técnicas más utilizadas suelen ser las Redes Neuronales Artificiales [31] y Random Forest [32] [33]. El aprendizaje automático permite modelizar escenarios climáticos pasados, presentes y futuros. En el trabajo de Sousa-Guedes trabajaron con 21 especies de anfibios, seleccionaron aleatoriamente el 70% de los registros de presencia de cada especie como datos de entrenamiento y el 30% como datos de prueba; el rendimiento del modelo se evaluó mediante AUC, que permite discriminar el modelo de una especie de un

modelo aleatorio. De esta forma predijeron mapas de idoneidad del territorio en diferentes escenarios climáticos para cada especie de anfibios [34]. Campos [35] también trabajó a nivel de especie, pero descartando variables climáticas Worldclime que se encontraban muy correlacionadas antes de crear sus modelos predictivos. Sillero [22] estableció los patrones de distribución para todo el grupo de anfibios ibéricos y la incidencia de las variables, pero no indicó el valor relativo de cada una de ellas.

Mientras que la mayoría de los estudios analizan los patrones de distribución a nivel de especie, en este trabajo lo que planteamos es estudiarlos a nivel de diversidad, usando como aproximación la riqueza (número de especies).

Las técnicas de aprendizaje supervisado permiten abordar de maneras muy diferentes el problema de la modelización espacial de la distribución de diferentes organismos, pero todas estas simulaciones y modelizaciones solo son precisas y útiles para la conservación si se poseen datos actualizados de los censos poblacionales y la biología de las especies de estudio.

4 Metodología

4.1 Creación base de datos

Para la creación de la base de datos se han empleado Sistemas de Información Geográfica. Los Sistemas de Información Geográfica (SIG) permiten relacionar cualquier tipo de dato con una localización geográfica. Las capas geográficas de este estudio se descargaron de diferentes páginas de acceso libre. Empleamos el software de ArcMap (*ArcGIS® software by Esri*), creado por la empresa ESRI; es un programa de procesamiento geoespacial que se utiliza principalmente para ver, editar, crear y analizar datos geoespaciales. Todas las capas fueron recortadas para España peninsular y Baleares, utilizando como base la capa vectorial de España (IGN), georreferenciada en coordenadas geográficas WGS84. [https://centrodedescargas.cnig.es/CentroDescargas/index.jsp]

A continuación se detalla la obtención de las diferentes variables y las fuentes desde las cuales se descargaron.

- **Riqueza de especies de anfibios**

La capa se descargó del Banco de Datos de la Naturaleza del Ministerio para la transición ecológica y el reto demográfico.

[https://www.miteco.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/acceso-rapido-datos.aspx]

En ArcMap se exportó esa capa a un Excel, que recogía la siguiente información: una columna cuyas filas eran las etiquetas de cada coordenada geográfica UTM 10x10km y varias columnas con todas las especies de anfibios citados en España, cuyas filas tenían un 0 o 1 en función de la ausencia-presencia de la especie en dicha coordenada.

En España (entendida como los territorios de la península ibérica, Ceuta, Melilla, islas y peñones norteafricanos, Islas Baleares e Islas Canarias) existen 39 especies de anfibios, con 14 endemismos y 2 especies introducidas. En este estudio se han empleado 26 especies, las presentes en la Península Ibérica e Islas Baleares (tabla 1), descartándose las siguientes:

- Sapo común (*Bufo spinosus*) y sapo corredor (*Epidalea calamita*), dada la plasticidad de estas especies y su escasa dependencia de zonas húmedas (solo requieren cursos de agua para los procesos reproductores y pueden ser temporales). La inclusión de estas especies distorsionaría los reales requerimientos de humedad del resto de especies dada su elevada presencia en la Península.

- Tritón del Montseny (*Calotriton arnoldi*), sapo partero almogávar (*Alytes almogavarii*), sapillo moteado occidental (*Pelodytes atlanticus*) y sapillo moteado mediterráneo (*Pelodytes hespericus*), por tener una distribución muy restringida y específica.

- Especies introducidas.

Dichas especies se eliminaron con una tabla dinámica de Excel, deseleccionando las descartadas. Trabajamos con datos de riqueza, entendida como el número de especies detectadas en cada cuadrícula UTM 10x10km, por lo que mediante una tabla dinámica se sumaron todas las presencias de anfibios creando esa nueva variable. Esta nueva variable será la variable dependiente de los posteriores análisis estadísticos y representaciones geográficas.

Orden *Caudata*		
1	*Calotriton asper*	Tritón pirenaico
2	*Chioglossa lusitanica*	Salamandra rabilarga
3	*Ichthyosaura alpestris*	Tritón alpino
4	*Lissotriton boscai*	Tritón ibérico
5	*Lissotriton helveticus*	Tritón palmeado
6	*Pleurodeles waltl*	Gallipato
7	*Salamandra salamandra*	Salamandra común
8	*Triturus marmoratus*	Tritón jaspeado
9	*Triturus pygameus*	Tritón pigmeo
Orden *Anura*		
10	*Alytes cisternasii*	Sapo partero ibérico
11	*Alytes dickhilleni*	Sapo partero bético
12	*Alytes muletensis*	Sapillo balear
13	*Alytes obstetricans*	Sapo partero común
14	*Discoglossus galganoi*	Sapillo pintojo ibérico
15	*Discoglossus jeanneae*	Sapillo pintojo meridional
16	*Discoglossus pictus*	Sapillo pintojo mediterráneo
17	*Hyla meridionalis*	Ranita meridional
18	*Hyla molleri*	Ranita de San Antonio
19	*Pelobates cultripes*	Sapo de espuelas
20	*Pelodytes ibericus*	Sapillo moteado ibérico
21	*Pelodytes punctatus*	Sapillo moteado septentrional
22	*Phelophylax perezi*	Rana común
23	*Rana dalmatina*	Rana ágil
24	*Rana ibérica*	Rana patilarga
25	*Rana pyrenaica*	Rana pirenaica
26	*Rana temporaria*	Rana bermeja

Tabla 1. Especies de anfibios de estudio

Las figuras 10, 11 y 13 muestran tres especies de anuros y la figura 12 una especie de urodelo, todos ellos incluidos en el estudio (tabla 1).

Figura 10. ***Pelodytes punctatus***
Sapillo moteado meridional

Figura 11. ***Hyla molleri***
Ranita de San Antón

Figura 12. ***Pleurodeles waltl***
Gallipato

Figura 13. ***Alytes obstetricans***
Sapo partero común

- **UTM 10km x 10 Km**

Capa de malla obtenida del Ministerio de Transición Ecológica y Retos Demográficos. Esta variable es la "unidad de medida", será la base sobre la que se analizarán y obtendrán todas las demás variables.

[https://www.miteco.gob.es/en/biodiversidad/servicios/banco-datos-naturaleza/informacion-disponible/bdn-cart-aux-descargas-ccaa.asp]

- **Variables climáticas**

Las variables climáticas son variables ampliamente utilizadas a la hora de analizar la dependencia de la distribución de especie frente a los requerimientos ambientales. Concretamente las variables Wordclime están especialmente aceptadas en los estudios científicos por su adecuada precisión y variedad (representan tendencias anuales, por ejemplo temperatura media anual y precipitación anual; estacionalidad, por ejemplo rango anual de temperatura y precipitación, y factores ambientales extremos o limitantes, por ejemplo temperatura del mes más frío y cálido, o precipitación de los meses húmedo y cálido). Emplearemos en total 19 de estas variables.

[https://www.worldclim.org/]

- **Altitud**

Esta variable, aunque representa cierta relación con las variables climáticas se considera adecuada ya que integra de forma sinérgica las variables de precipitación y temperatura, además de variables como la presión o la concentración de oxígeno, lo que caracteriza bastante bien las limitaciones ecológicas de muchas especies de fauna y flora. Esta capa se obtuvo a partir del Centro de Descargas IGN.

[https://centrodedescargas.cnig.es/CentroDescargas/index.jsp]

- ***Footprint***

Esta variable sintetiza mediante una fórmula la presión que diferentes contaminantes y variables antrópicas producen sobre el entorno, está relacionada con la huella ecológica y se basa en datos de contaminación lumínica, sonora, influencia antrópica, etc. sobre el territorio. Esta capa es de elaboración propia a partir de la capa obtenida en Sanderson et al., 2003.

[https://www.nature.com/articles/sdata201667]

- **Orilla**

Esta variable pretende reflejar de forma rápida el hábitat adecuado para la presencia de anfibios. Si bien existen muchas representaciones digitales de hábitat terrestre (Land Cover Corine, Mapas de Vegetación…) la dependencia de los anfibios a los entornos húmedos hace que estas variables no reflejen adecuadamente el hábitat disponible, mientras que la superficie disponible de orilla representa la disponibilidad de zonas húmedas en cada cuadrícula. Esta capa es de elaboración propia, calculada a partir de una capa digital del Centro de descargas del Ministerio de Agricultura, Alimentación y Pesca. Es una intersección entre dos capas vectoriales: cuadriculas UTM (polígonos) y ríos y zonas húmedas (líneas). Una vez realizada la intersección, se calculó la longitud de las líneas en cada cuadrícula con la herramienta de ArcMap de Calculadora de tablas. De esta forma se calculó la superficie de orilla medida en metros de orillas de cursos y masas de agua para pixeles de 1km^2.

[https://www.mapama.gob.es/ide/metadatos/srv/spa/metadata.show?uuid=1d7418d0-39fe-4206-9b8a-a971a49d65a2]

Una vez descargadas todas las capas, se cargaron en ArcMap. En primer lugar, se unió la capa UTM10x10km al Excel creado anteriormente con la riqueza de anfibios. Esto se consigue porque ambas capas tienen como nexo de unión la columna con la etiqueta de las coordenadas UTM.

Después se realizó un muestreo, seleccionando 10 puntos aleatorios para cada cuadrícula con presencia de especies. Tras el muestreo, se cargaron el resto de capas (variables climáticas de *BIO1* a *BIO19*, *altitud*, *orilla* y *footprint*). Mediante la herramienta *spatial analysis* de ArcMap extraemos valores múltiples a puntos, creando una tabla de atributos con todas las variables que podemos exportar fácilmente a Excel.

Tras este procesamiento, se generó una base de datos muy completa y compleja ya que contiene los valores de cada variable en cada cuadricula, además de la riqueza presente en las misma. Concretamente la base de datos estaba constituida por 47.170 filas y 31 variables: *riqueza, footprint, orilla, altitud, BIO1* a *BIO19, UTM10, FID1, FID2, CID, puntos, y, cuadricula* y *super.*

Aunque hay variables que presentan el mismo valor para toda la cuadrícula (riqueza de anfibios) no es así para otras variables que muestran diversidad de valores. Al obtener 10 valores aleatorios y promediarlos obtenemos un valor medio de dicha variable para dicha cuadricula. Esto se realizó mediante tablas dinámicas en Excel, obteniéndose así una única fila por cada coordenada UTM.

Eliminamos en Excel las variables *FID1, CID, puntos, y, cuadricula* y *super* ya que solo se necesitan para generar la base de datos georreferenciada pero no para los análisis estadísticos. Además la variable *super* no aporta información porque todas las cuadrículas poseen la misma superficie (100 km^2). *FID2* (número de etiqueta asociado a cada coordenada UTM) es una variable importante porque, tras crear nuestro modelo predictivo, nos permitirá reasignar las coordenadas geográficas para la representación en el mapa sin necesidad de mantener en la base de datos todas las variables geográficas mientras se hacen los análisis estadísticos. La base de datos resultante tiene 4717 observaciones y 23 variables, especificadas en la tabla 2.

Etiqueta	Tipo variable	Descripción variable
Riqueza	Cuantitativa discreta	Suma de las especies de anfibios que aparecen en una cuadrícula UTM
Footprint	Cuantitativa continua	Huella ecológica, presiones humanas directas e indirectas sobre el medio ambiente a nivel mundial
Orilla	Cuantitativa continua	Superficie de orilla medida en metros de orillas de cursos y masas de agua para píxeles de 1km^2
Altitud	Cuantitativa continua	medida en m
BIO1	Cuantitativa continua	Temperatura Media Anual
BIO2	Cuantitativa continua	Rango Diurno Medio [media mensual (temperatura máxima - temperatura mínima)]
BIO3	Cuantitativa continua	Isotermalidad = (BIO2/BIO7) * 100
BIO4	Cuantitativa continua	Estacionalidad de la Temperatura (desviación estándar * 100)
BIO5	Cuantitativa continua	Temperatura Máxima del mes más cálido
BIO6	Cuantitativa continua	Temperatura Mínima del mes más frío
BIO7	Cuantitativa continua	Rango Anual de Temperatura (BIO5 – BIO6)
BIO8	Cuantitativa continua	Temperatura Media del trimestre más húmedo
BIO9	Cuantitativa continua	Temperatura Media del trimestre más seco
BIO10	Cuantitativa continua	Temperatura Media del trimestre más cálido
BIO11	Cuantitativa continua	Temperatura Media del trimestre más frío
BIO12	Cuantitativa continua	Precipitación Anual
BIO13	Cuantitativa continua	Precipitación del mes más lluvioso
BIO14	Cuantitativa continua	Precipitación del mes más seco
BIO15	Cuantitativa continua	Estacionalidad de la Precipitación (Coeficiente de Variación)
BIO16	Cuantitativa continua	Precipitación del trimestre más húmedo
BIO17	Cuantitativa continua	Precipitación del trimestre más seco
BIO18	Cuantitativa continua	Precipitación del trimestre más cálido
BIO19	Cuantitativa continua	Precipitación del trimestre más frío

Tabla 2. Caracterización de las variables climáticas utilizadas para el análisis

**Las variables wordclime (BIO1 a BIO 11) relacionadas con temperatura están en ºC multiplicados por 10*

Una vez creada la base de datos se cargó en el software de código libre R *[R Core Team (2021). R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, *Austria. https://www.R-project.org/*] para realizar los análisis estadísticos.

4.2 Exploratory Data Analysis (EDA)

El Análisis Exploratorio de Datos fue desarrollado en los años 70 por el matemático John Tukey. Este análisis permite mediante estadísticas descriptivas y herramientas gráficas tener un primer contacto con la base de datos a estudiar, para obtener un resumen de las principales características, comprender las variables y su distribución, hallar anomalías en los datos como valores atípicos, eliminar valores nulos para trabajar con bases de datos completas, revelar patrones entre variables y generar nuevas hipótesis que contrastar más adelante con otros métodos estadísticos más elaborados. Es muy útil porque también permite determinar si las técnicas estadísticas que se van a emplear en el análisis de datos son apropiadas. En este apartado se realizaron los siguientes análisis:

1. **Preprocesado de la base de datos**

   Se eliminaron los valores nulos (NA) para trabajar con filas completas.

2. **Análisis univariado**

El análisis individual de las variables se hizo con **estadísticas descriptivas** (media aritmética, desviación estándar, máximo, mínimo, media del error estándar IQR intercuartílico (Q3-Q1)), sesgo, curtosis y varios percentiles) y representándolas gráficamente mediante **histogramas** y **boxplots**.

Otro análisis importante en este apartado es comprobar si las variables siguen una distribución normal con un nivel de significación de 0,05 mediante la prueba de **Shapiro-Wilks**. En dicha prueba la hipótesis nula (H_0) sostiene que la distribución es normal frente a la hipótesis alternativa (H_1) de que la distribución no es normal:

$$H_0: X \sim N(\mu, \sigma^2)$$
$$H_1: X \nsim N(\mu, \sigma^2)$$

3. **Análisis bivariado**

Para observar las relaciones entre la variable respuesta riqueza y el resto de variables se representaron gráficamente con **diagramas de puntos** y **diagramas de cajas y bigotes**.

4. Análisis de Correlaciones

El estudio de la correlación entre dos variables se refiere a un conjunto de relaciones estadísticas que involucran una dependencia entre ellas. El **Análisis** de **Correlaciones** por el método de **Spearman** es un método estadístico no paramétrico que pretende examinar la intensidad de asociación entre dos variables cuantitativas, y que esa relación no sea debida al azar, sino que sea estadísticamente significativa.

El **estadístico r de Spearman** se calcula de la siguiente forma:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2-1)}$$

- *d* es la diferencia entre los correspondientes valores de *x – y*
- *n* es el número de parejas.

El coeficiente de correlación oscila entre -1 y +1. El valor 0 indica que no existe asociación lineal entre las dos variables en estudio (puede que exista otro tipo de correlación, pero no lineal); valores próximos a 1 indican una correlación fuerte y positiva; valores próximos a -1 indican una correlación fuerte y negativa.

4.3 Análisis de Componentes Principales (ACP)

El Análisis de Componentes Principales es una técnica de machine learning de aprendizaje no supervisado que sintetiza la información, reduciendo la dimensionalidad de los datos con una gran cantidad de medidas interrelacionadas. El objetivo es reducir el número de variables perdiendo la menor cantidad de información posible. Las variables originales se convierten en un nuevo conjunto, que se conoce como componentes principales o factores principales, que son una combinación lineal de las variables originales, e independientes entre sí.

Matemáticamente, dadas *n* observaciones de *p* variables, se buscan m<p variables que sean combinaciones lineales de las *p* originales y que no estén correlacionadas, recogiendo la mayor parte de la información o variabilidad de los datos. Se consideran una serie de variables $(x_1,...x_p)$ sobre un grupo de objetos o individuos y se calcula a partir de ellas un nuevo conjunto de variables $(y_1,...y_p)$ no correlacionadas entre sí, cuyas varianzas decrecen progresivamente.

Cada y_j (donde $j=1,...,\ p$) es una combinación lineal de las $x_1,\ x_2,...x_p$ originales, es decir:

$$y_j = a_{j1}x_1 + a_j2x_2 + \ldots + a_jPx_p = a'_jxy$$

donde $a'_j = (a_{j1}, a_{j2}, \ldots, a_{jP})$ es un vector de constantes y $x = \begin{bmatrix} x_1 \\ x_2 \\ \ldots x_p \end{bmatrix}$

Para realizar el ACP las variables deben estar estandarizadas para que sean comparables entre ellas. Se escalan para que tengan desviación estándar uno y media cero.

Resumiendo, las **fases** de un Análisis de Componentes Principales son las siguientes:

1. Análisis de la matriz de correlaciones

Un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, esto demuestra que existe información redundante y, por tanto, pocos factores explicarán gran parte de la variabilidad total.

2. Selección de los factores

La elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente. Los valores propios o *eigenvalues*, que miden la cantidad de variación retenida por cada componente principal, se pueden utilizar para determinar el número de componentes principales a retener [36]. Un *eigenvalue* mayor que 1 indica que las componentes principales explican más varianza que la explicada por una de las variables originales en los datos estandarizados. Esto se usa como valor umbral o punto de corte para el cual se retienen las componentes.

3. Análisis de la matriz factorial

Una vez seleccionados los componentes principales, se representan en forma de matriz. Cada elemento de ésta representa los coeficientes factoriales de las variables (las correlaciones entre las variables y los componentes principales). La matriz tendrá tantas columnas como componentes principales y tantas filas como variables.

4. Interpretación de los factores

Este es un aspecto muy importante en la ACP porque debe deducirse tras observar la relación de los factores con las variables iniciales y hacer esto de manera correcta depende del conocimiento que se tenga sobre el tema de la investigación.

Para que un factor sea fácilmente interpretable debería tener las siguientes características, que son difíciles de conseguir:

- Los coeficientes factoriales deben ser próximos a 1.
- Una variable debe tener coeficientes elevados sólo con un factor.
- No deben existir factores con coeficientes similares.

5. Cálculo de las puntuaciones factoriales

Son las puntuaciones que tienen los componentes principales para cada caso, que nos permitirán su representación gráfica. Se calculan mediante la expresión:

$$X_{ij} = a_{i1}Z_{j1} + \cdots + a_{jk}Z_{kj} = \sum_{s=1}^{k} a_{is}\, z_{sk}$$

donde *a* son los coeficientes y *Z* son los valores estandarizados que tienen las variables en cada uno de los sujetos de la muestra.

La contribución total de una determinada variable, al explicar las variaciones retenidas por dos componentes principales se calcula como:

$$Contribución = \frac{(C1 * Eig1) + (C2 * Eig2)}{Eig1 + Eig2}$$

donde C1 y C2 son las contribuciones de la variable sobre la componente 1 y la componente 2, y Eig1 y Eig2 son los valores propios respectivamente [37] [38].

4.4 Modelo Lineal Generalizado: modelo log-lineal de Poisson

El **Modelo de Regresión Lineal Múltiple** presenta la siguiente forma:

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

donde:

- *y*: variable respuesta
- $X = (X_1,\ldots,X_p)$: vector de *p* variables predictoras
- *α:* intersección
- $\beta = \{ \beta_1 \ldots \beta_p \}$: vector de *p* coeficientes de regresión
- *ε*: error de medida (varianza no explicada por el modelo)

En un modelo lineal se asumen tres supuestos:

1. Los errores (ε_i) siguen una distribución normal.

$$Y_i \sim N(\mu, \sigma^2)$$

2. La varianza es constante.
3. La variable dependiente se relaciona linealmente con las variables independientes.

Estas asunciones hacen que la regresión lineal presente dificultades al trabajar con variables ecológicas.

Los Modelos Lineales Generalizados (GLM) son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores y varianzas no constantes. La combinación de variables predictoras está relacionada con la media de la variable respuesta mediante una función de conexión (*link function).* En función del tipo de datos se puede elegir entre diferentes familias de distribuciones (Gaussiana, Poisson, Binomial, Gamma) con sus respectivas funciones de conexión (identidad, logarítmica, *logit* e inversa).

El objetivo de los Modelos Lineales Generalizados es encontrar una función con capacidad predictiva y estadísticamente significativa, mediante procesos *forward, backward* o mediante criterios de información (AIC o Akaike y BIC o Bayesian Criterion Information). Estos últimos buscan el modelo más óptimo penalizando la capacidad de ajuste por su complejidad (número de variables). Esta técnica de selección de las variables no soluciona el problema de la colinealidad; las variables que seleccionen podrían estar incluidas por su correlación con otras y no porque tengan relaciones de causalidad con la variable respuesta. Es por esto que para generar nuestro modelo trabajaremos con las variables climáticas agrupadas en los factores generados en la ACP.

Con los datos de los que disponemos en este estudio, el modelo más apropiado es una regresión puesto que la que la variable respuesta (riqueza) es de conteo y las variables explicativas son continuas, empleando la *link function* logarítmica y la distribución de los errores de Poisson (modelo log-lineal).

La fórmula matemática es la siguiente:

$$\log(y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

equivalente a

$$y = e^{\alpha + \beta X}$$

Las asunciones de este modelo son que la **varianza** debe ser **igual** a la **media**.

- La varianza mide la dispersión de los datos, es el promedio de las diferencias al cuadrado de la media. La varianza es igual a 0 si todos los valores son idénticos. Cuanto mayor sea la diferencia entre los valores, mayor será la varianza.
- La media (μ) es el promedio de valores de un conjunto de datos.

$$E(X) = \mu$$

Para la regresión de Poisson, la media y la varianza se relacionan como:

$$var(X) = \sigma^2 E(X)$$

siendo σ^2 el parámetro de dispersión.

La evaluación de la calidad del modelo se hace en base a la **Devianza (D^2).** La devianza es una medida de bondad de ajuste que compara el modelo saturado o modelo nulo (aquel en el cual se asume que hay tantos parámetros como observaciones) con el modelo de interés.

$$Devianza\ (D^2) = \frac{devianza\ modelo\ nulo - devianza\ residual}{devianza\ modelo\ nulo} * 100$$

Para valorar el ajuste de los modelos lineales generalizados se puede emplear el **estadístico X^2,** que se define como el doble de la diferencia entre el máximo del logaritmo de la verosimilitud que se podría conseguir con la mínima o máxima parametrización y el valor del máximo del logaritmo de la verosimilitud que se consigue con el modelo que se quiere evaluar.

$$X^2 = \sum_i \frac{(y_i - \mu'_i)^2}{\mu_i} = \sum_i \frac{(observado - ajustado\ )^2}{ajustado}$$

Es importante analizar los **residuos** del modelo:

- Normalidad, usando los residuos de devianza (no en la escala original de la respuesta).

- Homocedasticidad de los residuos a través de las predicciones del modelo (aplicando la transformación de la link function).

- Existencia de datos influyentes y perdidos con la distancia de Cook y Leverage.

4.5 Machine Learning: Random Forest

Machine Learning es una rama del campo de la inteligencia artificial que a través de algoritmos permite a las máquinas realizar tareas específicas autónomamente, sin necesidad de programarlos, identificando patrones de datos masivos y elaborando predicciones mediante el uso de distintos algoritmos. Los algoritmos de machine learning se pueden dividir en tres categorías, aprendizaje supervisado, no supervisado y de refuerzo.

Los **Random Forest** (o bosques aleatorios), creados por Leo Breiman y Adele Cutlerson, son técnicas de aprendizaje automático supervisado muy usadas porque tienen una capacidad de generalización muy alta. Las limitaciones de los **Árboles de Decisión** es que tienen la tendencia de sobre-ajustar (*overfit*); aprenden muy bien los datos de entrenamiento pero su generalización no es tan buena, además si el tamaño de muestra es pequeño y el p-valor grande, ignoran algunas variables. Sin embargo, los Random Forest son un tipo de método de partición recursivo muy adecuado para problemas de tamaño de muestra pequeño y p-valor grande, porque aunque pierden interpretabilidad, aumentan el rendimiento del modelo final.

Random Forest es un conjunto (*ensemble*) de Árboles de Decisión combinados con *bagging* (figura 14). Cada árbol se entrena con distintas muestras de datos para un mismo problema, ningún árbol ve todos los datos de entrenamiento, sino que los distintos árboles ven distintas porciones de datos. Al combinar los resultados y promediarlos, unos errores se compensan con otros y se obtiene una predicción que generaliza mejor.

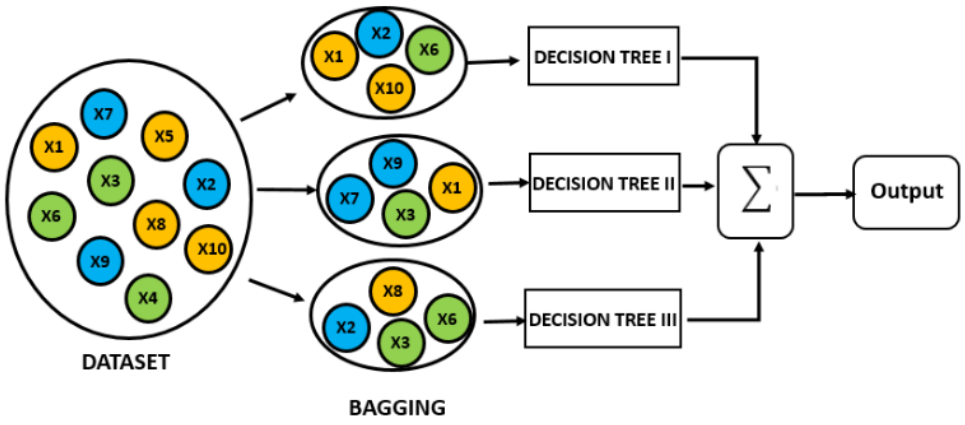


Figura 14. Figura explicativa proceso de *bagging*

[Fuente: https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html]

En la tabla 3 se muestra un resumen de las fortalezas y debilidades de dicho algoritmo:

Fortalezas	Debilidades
• Modelo muy generalizable, funciona bien en la mayoría de los problemas. • Tiene muy pocas suposiciones adjuntas por lo que la preparación de los datos es mínima. • Se puede utilizar en bases de datos masivas produciendo clasificadores muy certeros. • Maneja cientos de variables de entrada, sin excluir ninguna, identificando las más significativas. Método de reducción de dimensionalidad. • Posee un método eficaz para estimar datos ruidosos o faltantes y mantiene la exactitud cuando una gran proporción de los datos está perdida. • Maneja tanto variables categóricas como continuas. • Una de las salidas del modelo es la importancia de variables.	• A diferencia de un árbol de decisiones, el modelo no es fácilmente interpretable • Poco control en lo que hace el modelo (modelo caja negra para modeladores estadísticos) • Los datos que incluyen variables categóricas con diferente número de niveles, el random forest se parcializa a favor de esos atributos con más niveles. La posición que marca la variable no es fiable para este tipo de datos. • Si los datos contienen grupos de atributos correlacionados con similar relevancia para el rendimiento, entonces los grupos más pequeños están favorecidos sobre los grupos más grandes.

Tabla 3. Fortalezas y debilidades de Random Forest

Los pasos a seguir para construir un Random Forest son los siguientes:

1. Teniendo un conjunto ***N*** observaciones diferentes, se elige una muestra ***N*** aleatoria con reemplazamiento (***bootstrapping***), para introducir aleatoriedad al algoritmo. Cada árbol se entrena con aproximadamente 2/3 del total de datos de entrenamiento.
2. Dadas las ***M*** variables de entrada, en cada nodo se seleccionan de forma aleatoria ***p<<M*** variables. ***p*** es constante en todo el proceso de formación del árbol e introduce el segundo elemento de aleatoriedad en el algoritmo.
3. Se deja crecer el árbol hasta la máxima extensión posible sin poda.
4. Una vez construido el bosque, se realiza la predicción.

La aleatoriedad introducida reduce la correlación entre árboles porque en la formación de los mismos cada uno parte de una muestra ligeramente distinta y en cada nodo la selección de variables es diferentes.

Los **hiperparámetros** más importantes en un Random Forest son:

- ***Ntree***: número de árboles que forma el bosque.

- ***Mtree*** (o *mtry* en las fórmulas de R): número de variables ***p*** que se seleccionan en cada nodo.

La tasa de error de los Random Forest está relacionada con estos parámetros. Al reducir *p* (número de variables) se reduce la correlación entre los árboles porque cada nodo tiene menos posibilidades entre las que elegir, pero también se reduce la precisión del árbol. Por defecto, el valor de *mtree* (*mtry*) es la raíz cuadrada del número total de todos los predictores (*p*) para la clasificación.

El número de árboles (*ntree*) influye en la precisión de la predicción, a mayor número de árboles mejor es la predicción porque el número de datos a promediar es mayor. Existe un valor umbral a partir del cual el error ya no disminuye y se estanca.

OOB (Out of Bag Error) es una medida que calcula la tasa de clasificación errónea y se aplica a modelos que utilizan la técnica del bootstrapping. Usar este conjunto de *test* (OOB) es tan preciso como si se usara un conjunto de *test* del mismo tamaño que el de entrenamiento. Al elegir los ***N*** datos con reemplazamiento, alrededor de 1/3 de los datos nunca son seleccionados, por tanto OOB representa el error de predicción cometido por el Random Forest cuando se tienen en cuenta el conjunto de variables que han quedado "fuera de bolsa". OOB disminuye cuando aumentan el número de árboles hasta que llega un punto en el que se estanca y no disminuye más al aumentar *ntree**,*** pudiendo producir problemas de sobreajuste u *overfitting*.

Se seleccionó este algoritmo para el presente trabajo porque se pretende descubrir la **importancia de las variables**, es decir, ver cómo afecta a la salida del modelo cuando se producen cambios en las variables de entrada. Las variables de entrada que más variabilidad produzcan en la salida son aquellas que más influencia tienen y mejor explican el modelo.

Hay dos formas de medir esta importancia:

1. Índice de Gini

Mide la reducción de la impuridad nodal media, es decir, haciendo la media sobre todos los árboles de todas las reducciones de todas las variables, se extrae la media de ese valor de impuridad. La variable que más reduzca la impuridad del Random Forest es la más importante.

2. Mean Squared Error (MSE)

Se calcula el MSE o Error Cuadrático Medio con las variables del OOB que son las que se han quedado fuera, se hace una permutación aleatoria en una de las variables de entrada lo que produce un cambio en las *y* de salida que produce un cambio en el MSE de las variables que han quedado fuera. Se

realiza ese proceso con todas las variables de entrada y en todos los árboles permutando aleatoriamente y se compara el valor del MSE para cada variable permutada, antes y después de permutar. Esa diferencia se suma para cada árbol, se normaliza y promedia. Cuanto mayor sea ese valor, más importante es la variable permutada.

La ecuación del MSE es la siguiente:

$$MSE(X_j) = \frac{1}{n}\sum_{i=1}^{n}(y - y'(X_j))^2$$

Para medir el **rendimiento del modelo** se necesitan métricas que cuantifiquen la calidad del clasificador, permitiendo la comparación y selección. La mayoría de las medidas están basadas en la matriz de confusión. No se va a entrar en explicar cada una de ellas en detalle porque en el presente estudio se va a emplear únicamente el ***Accuracy*** que mide el porcentaje de casos que el modelo ha acertado.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

donde TP=tasa positivos, TN=tasa negativos, FP=falsos positivos, FN=falsos negativos [39] [40] [41] [42] [43].

5 Resultados

5.1 Exploratory Data Analysis (EDA)

5.1.1 Preprocesado base de datos: eliminación valores nulos

La base de datos final está constituida por 23 variables (Tabla 2). El **Análisis Exploratorio de los Datos** determinó que el 99'28% de las filas están completas, encontrando un 0'54% de observaciones faltantes o valores nulos (NA). De las 4.171 filas de la base de datos están completas 4.683. Se eliminaron 609 valores nulos para trabajar con las filas completas.

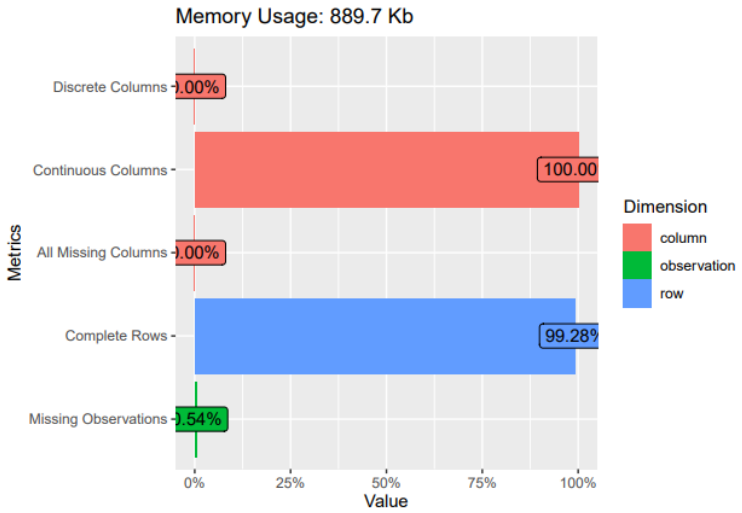


Figura 15. Resultados del Exploratory Data Analysis

5.1.2 Análisis Univariado

La variable riqueza parece seguir una distribución de Poisson como se aprecia en la siguiente gráfica; los datos son conteos de eventos (enteros no negativos, sin límite superior) y todos los eventos son independientes.

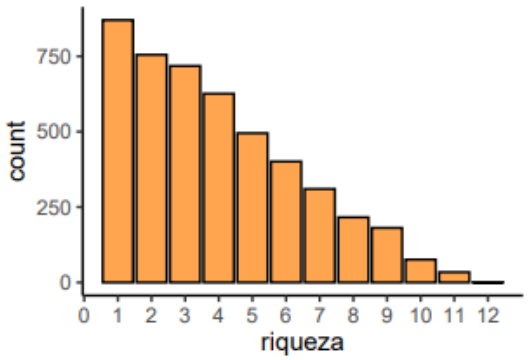


Figura 16. Histograma de la variable respuesta riqueza de especies

riqueza	n
1	870
2	755
3	718
4	626
5	495
6	401
7	310
8	216
9	181
10	76
11	34
12	1

Tabla 4. Recuento riqueza de especies por cuadrícula UTM

El resto de variables también se representaron mediante **histogramas** (figuras 17, 18 y 19). Se observan grandes diferencias en la forma de los histogramas, por lo que se realizó la prueba de **Shapiro-Wilks** para comprobar si las variables siguen una distribución normal, obteniéndose en todas ellas un p-valor inferior a 0'05 (ver anexo) y por tanto aceptando la hipótesis de normalidad para continuar con los análisis.

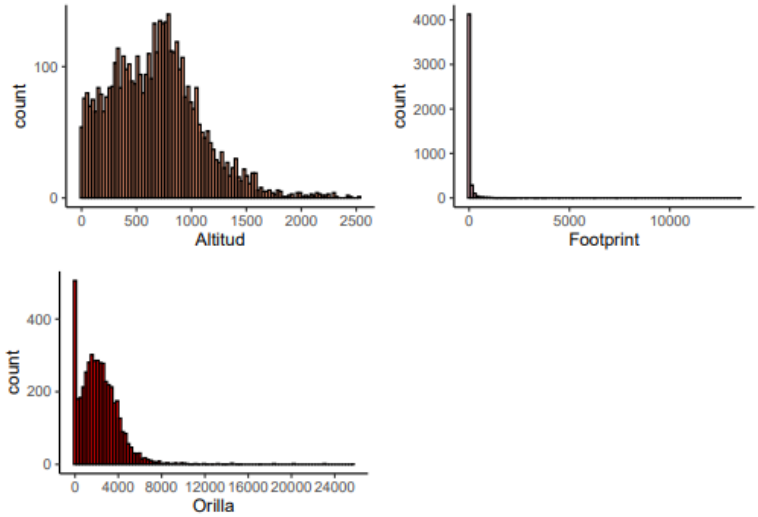


Figura 17. Histogramas de las variables altitud, *footprint* y orilla

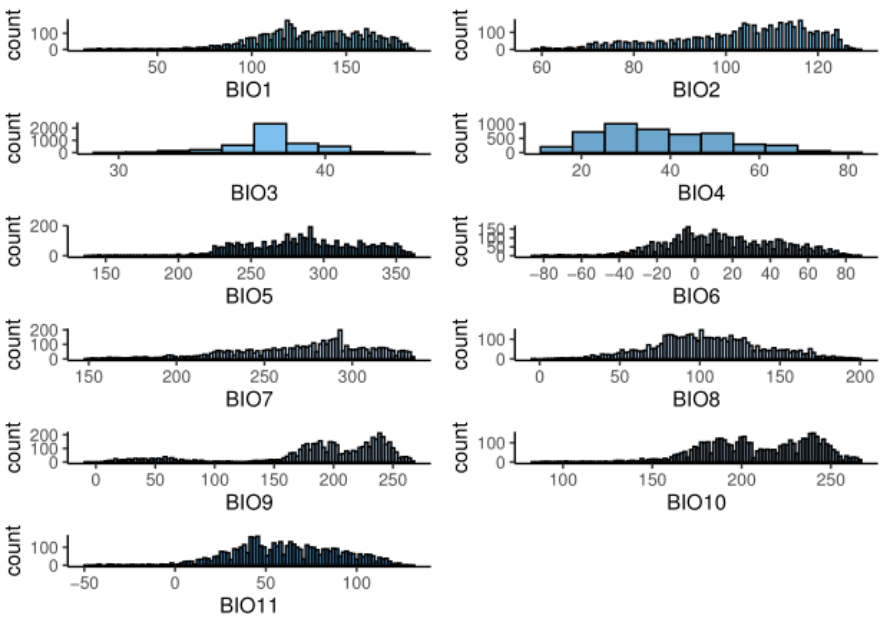


Figura 18. Histogramas de las distintas variables de temperatura

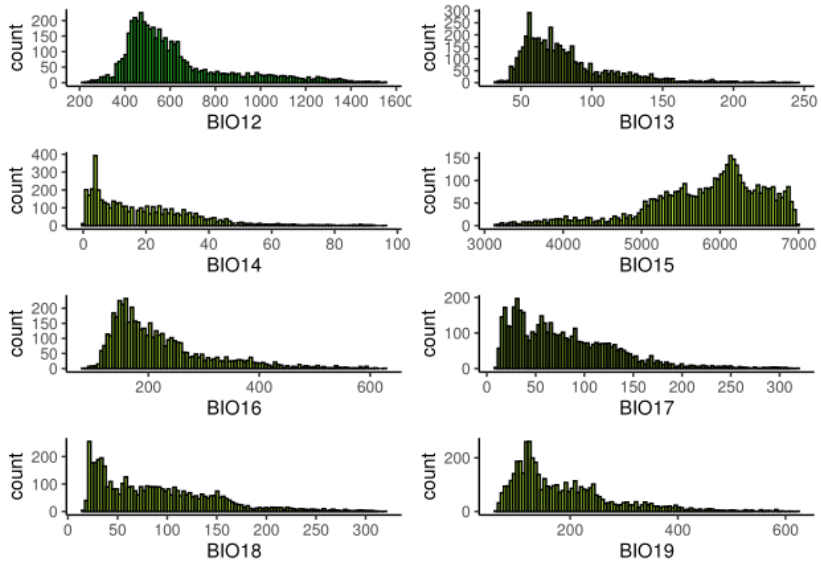


Figura 19. Histogramas de las distintas variables de precipitación

En el anexo se incluyen **boxplots** de cada una de las variables en los que se pueden ver el valor máximo, la media, el valor mínimo, quartiles y los valores atípicos de cada una de ellas. También se incluyen las **estadísticas descriptivas** de las variables: media aritmética, desviación estándar, máximo, mínimo, media del error estándar IQR (rango intercuartílico (Q3-Q1)), sesgo, curtosis y varios percentiles, que aparecen resumidos en la tabla 5 y ampliados en el anexo.

described_variables	n	na	mean	sd	se_mean	IQR	skewness	kurtosis
orilla	4683	0	2301.374760	1934.866881	28.2741145	2235.5	2.7245548	19.2396524
altitud	4683	0	674.062140	405.854168	5.9307270	550.5	0.6666254	0.7700328
BIO1	4683	0	131.370489	28.542070	0.4170839	41.0	-0.4903404	0.3839091
BIO2	4683	0	103.240658	15.101068	0.2206712	21.0	-0.7145191	-0.2465008
BIO3	4683	0	37.560752	1.883371	0.0275216	2.0	-0.6711639	1.7762874
BIO4	4683	0	37.928678	13.655039	0.1995404	21.0	0.4513261	-0.5524128
BIO5	4683	0	286.497971	39.534700	0.5777186	59.0	-0.2850198	-0.1293357
BIO6	4683	0	15.078796	29.249125	0.4274160	42.0	0.0330775	-0.3089551
BIO7	4683	0	271.427504	38.270386	0.5592433	52.0	-0.6825092	0.0379752
BIO8	4683	0	105.091394	33.999582	0.4968342	45.0	0.0966871	-0.1706994
BIO9	4683	0	190.036729	60.739149	0.8875782	58.0	-1.4461357	1.3444485
BIO10	4683	0	208.875080	31.098643	0.4544429	50.0	-0.4902453	0.0619801
BIO11	4683	0	60.997651	29.145817	0.4259064	42.0	-0.1643041	-0.0866916
BIO12	4683	0	626.724322	241.474126	3.5286495	235.0	1.4185152	1.5036445
BIO13	4683	0	81.670083	32.262404	0.4714489	34.0	1.6172116	3.1031387
BIO14	4683	0	19.281444	16.373987	0.2392723	24.0	1.2980104	2.0540342
BIO15	4683	0	5805.468290	768.659122	11.2323779	955.5	-0.9827269	0.8140972
BIO16	4683	0	220.585522	89.519595	1.3081454	99.0	1.4887190	2.2728733
BIO17	4683	0	80.085842	53.473028	0.7813987	76.0	1.1858533	1.6395892
BIO18	4683	0	89.171044	58.568565	0.8558596	87.0	1.0287085	0.8371726
BIO19	4683	0	186.976938	95.551835	1.3962943	115.0	1.4057756	2.0346571
footprint	4683	0	63.632287	407.963803	5.9615550	15.0	19.7774412	498.4956537
riqueza	4683	0	3.970105	2.484459	0.0363053	4.0	0.6890932	-0.3430557

Tabla 5. Estadísticas descriptivas de las variables

5.1.3 Análisis Bivariado

Como el volumen de datos es elevado, los diagramas de puntos enfrentando riqueza al resto de variables son bastante confusos (ver anexo). Mediante **diagramas de cajas y bigotes** se consigue una mejor representación. Si la media de una de las categorías cae fuera de la desviación estándar de otra caja, existen diferencias significativas. En la figura 20 observamos que, a priori no se observan diferencias significativas en orilla y *footprint* entre los distintos valores de riqueza (1-12), mientras que en altitud si se observan ligeras diferencias. Sin embargo, tanto en la figura 21, que representa distintas mediciones de la variable climática temperatura, como en la figura 22, que corresponde a distintas mediciones de las precipitaciones, hay grandes diferencias entre las medias de los diferentes grupos de riqueza; por tanto, parece que existen diferencias significativas que se analizarán posteriormente con otros análisis estadísticos.

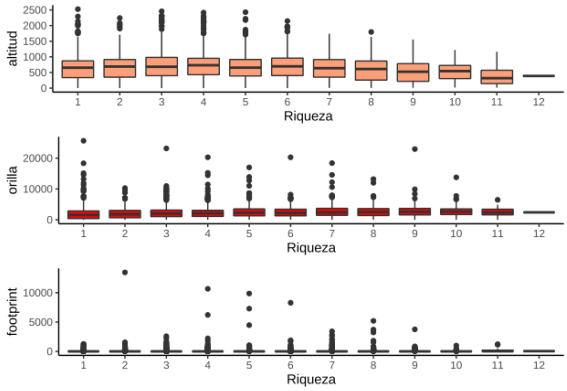


Figura 20. Boxplots bivariados riqueza vs altitud, orilla y *footprint*

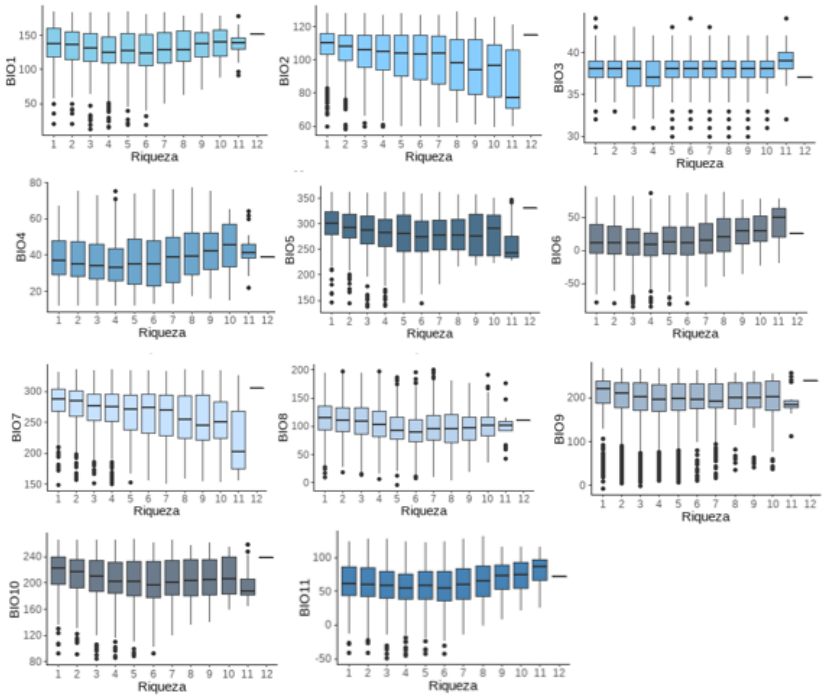


Figura 21. Boxplots bivariados riqueza vs distintas variables de temperatura (BIO1 a BIO11)

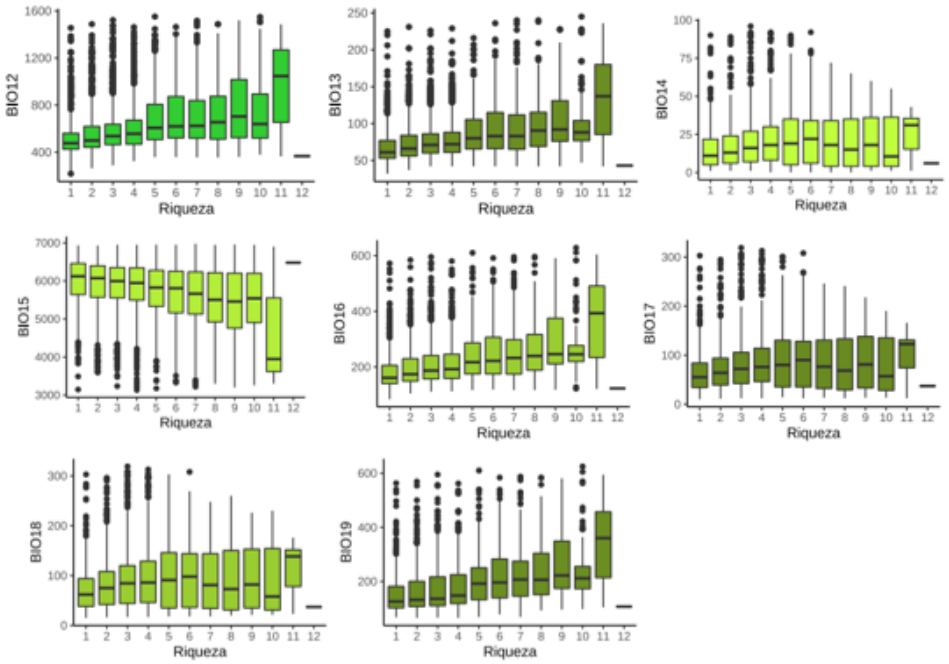


Figura 22. Boxplots bivariados riqueza vs distintas variables de precipitación (BIO12 a BIO19)

5.1.4 Análisis de Correlaciones

La figura 23 muestra las correlaciones positivas entre las variables en azul, y las negativas en rojo. Por tanto, los primeros resultados obtenidos en el **Análisis de Correlaciones** por el método de **Spearman** arrojan que:

- Las variables BIO12, BIO13, BIO14, BIO16, BIO17, BIO18, BIO19, correspondientes a las variables climáticas de precipitación, orilla y BIO6 muestran correlaciones positivas con la riqueza. La interpretación de este resultado es que a mayores valores de precipitación, cuánto mayor es la superficie de orilla o cuanto mayor es la temperatura mínima del mes más frío, mayor es la riqueza de anfibios.

- Las variables BIO2, BIO5, BIO7, BIO8, BIO10, BIO15, (correspondientes a las variables climáticas de temperaturas) y altitud, muestran correlaciones negativas con la riqueza. Cuanto mayor es la temperatura o más altitud hay en la cuadrícula UTM, la riqueza de especies disminuye.

- Las variables BIO1, BIO3, BIO4, BIO9, BIO11 y *footprint*, a priori, no son significativas.

- Como era de esperar, todas las variables climáticas de temperatura (BIO1 a BIO11) y precipitaciones (BIO12 a BIO19) están muy correlacionadas entre sí.

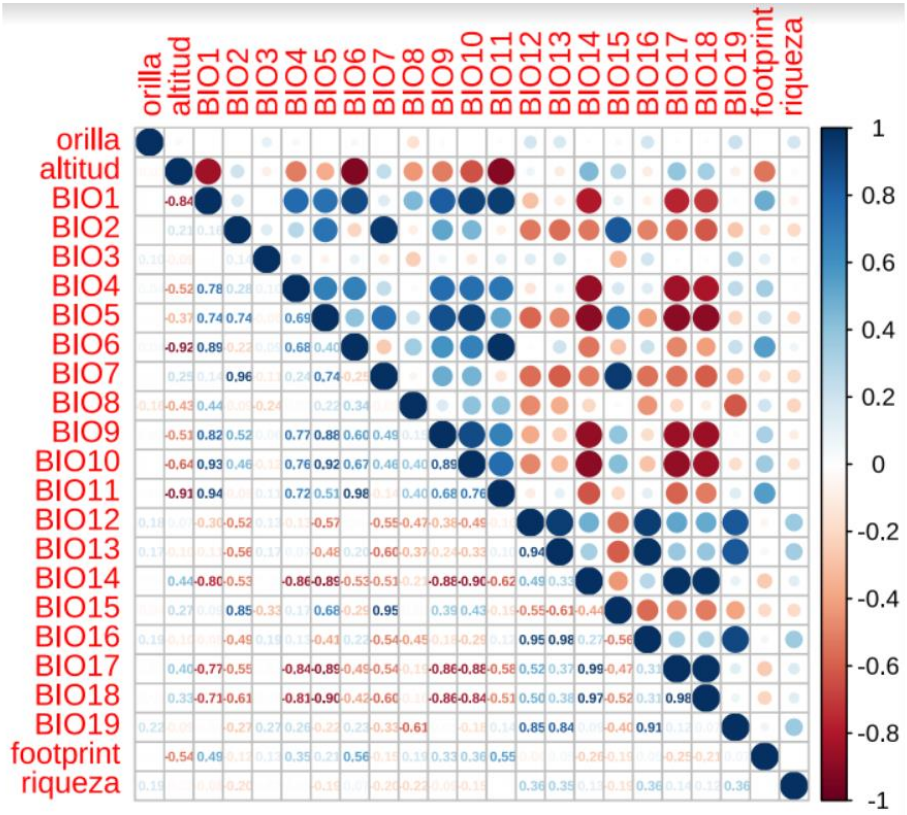


Figura 23. Representación correlaciones entre variables

5.2 Análisis de Componentes Principales (ACP)

El resultado del análisis de correlaciones es que existe multicolinealidad en las 19 variables climáticas (BIO1-BIO19), por eso se utilizó el Análisis de Componentes Principales, para obtener factores agrupados y emplearlos como nuevas variables en el modelo final que incluye orilla, altitud y *footprint.* De esta forma se evita la multicolinealidad efectuando una clasificación automática que tiene en cuenta la información esencial, es decir, conserva solo los primeros factores. Se usaron la librería *FactoMineR* para el análisis y *factoextra* para la visualización basada en ggplot2. Para realizar la ACP hay que estandarizar las variables para que sean comparables entre ellas. Se escalan para que tengan desviación estándar uno y media cero.

Los *eigenvalues* o valores propios miden la cantidad de variación retenida por cada componente principal. Las primeras componentes principales corresponden a las direcciones con la máxima cantidad de variación en el conjunto de datos. En la tabla 6 tenemos en la primera columna los eigenvalues, en la segunda columna la proporción de variación explicada por cada valor propio y en la tercera el porcentaje acumulado explicado, que se obtiene sumando las sucesivas proporciones de variación explicadas para obtener el total acumulado.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	9.4762400	49.8749475	49.87495
Dim.2	5.0497701	26.5777376	76.45269
Dim.3	2.2785556	11.9923978	88.44508
Dim.4	1.0117720	5.3251156	93.77020
Dim.5	0.4619234	2.4311756	96.20137
Dim.6	0.4055892	2.1346801	98.33605
Dim.7	0.1910426	1.0054874	99.34154
Dim.8	0.0600695	0.3161551	99.65770
Dim.9	0.0242672	0.1277223	99.78542
Dim.10	0.0139995	0.0736816	99.85910
Dim.11	0.0098612	0.0519012	99.91100
Dim.12	0.0054742	0.0288114	99.93981
Dim.13	0.0040077	0.0210930	99.96091
Dim.14	0.0028783	0.0151487	99.97605
Dim.15	0.0022293	0.0117333	99.98779
Dim.16	0.0013519	0.0071151	99.99490
Dim.17	0.0006949	0.0036576	99.99856
Dim.18	0.0002064	0.0010861	99.99965
Dim.19	0.0000671	0.0003531	100.00000

Tabla 6. Valores propios obtenidos en el ACP

Los valores propios o *eigenvalues* se utilizan para determinar el número de componentes principales a retener, valores mayores de 1 indican que los componentes principales explican más varianza que la explicada por una de las variables originales en los datos estandarizados, por tanto, observando la tabla 6 concluimos en que debemos quedarnos con las cuatro primeras componentes. En la figura 24 observamos que alrededor del 76'45% de la variación se explica por los dos primeros valores propios juntos, y el 93'8% de la información contenida en los datos es retenida por los cuatro primeros componentes principales.

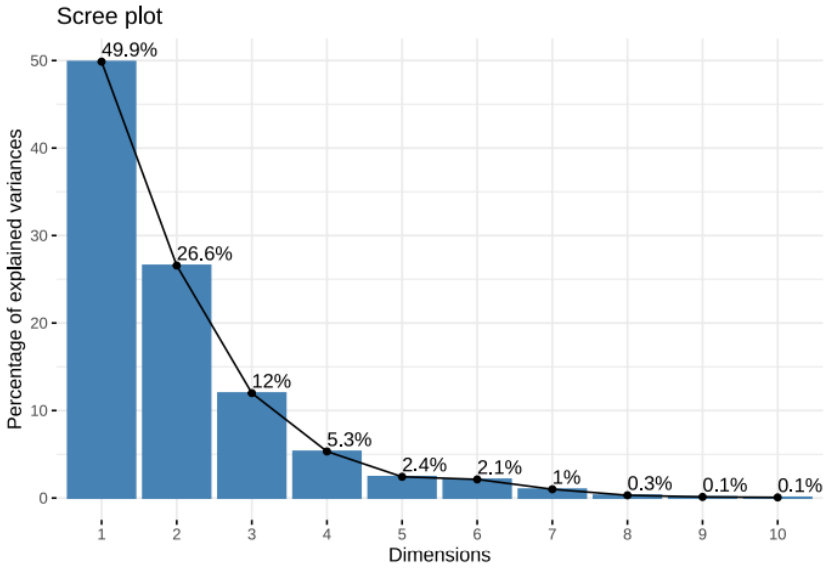


Figura 24. Gráfico de sedimentación

El ACP se puede representar con un gráfico de correlación de variables (figura 25) que muestra las relaciones entre todas las variables. Las variables

correlacionadas positivamente se agrupan, mientras que las variables correlacionadas negativamente se colocan en cuadrantes opuestos del origen del gráfico.

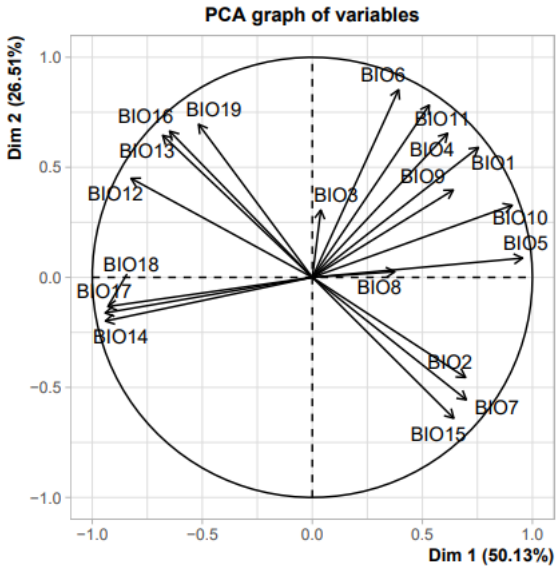


Figura 25. Gráfico de correlación de variables de las dos primeras componentes

La calidad de representación (cos^2) de las variables se mide por la distancia entre las variables y el origen; las variables que están alejadas del origen están bien representadas en el mapa de factores. Por tanto, observamos en la figura 26 que las variables posicionadas cerca de la circunferencia del círculo de correlación tienen un cos^2 alto (coloreadas en rojo) y por tanto una buena representación de la variable en el componente principal. Las variables situadas cerca del centro del círculo tienen valores de cos^2 bajos (coloreadas en azul cian), lo que indica que la variable no está perfectamente representada por las componentes principales. Las variables con valores medios de cos^2 se colorean en amarillo.

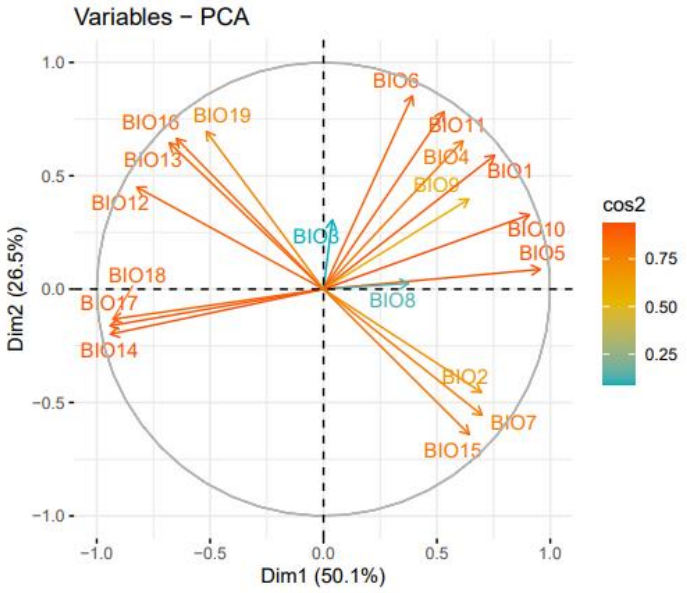


Figura 26. Calidad de representación de las variables en el mapa de factores

Las **contribuciones de las variables** para explicar la variabilidad en un componente principal dado se expresan en porcentaje (tabla 7). La línea discontinua roja de la figura 27 indica la contribución promedio esperada.

	Dim.1	Dim.2	Dim.3	Dim.4
BIO1	6.0190306	6.9837652	2.0727062	0.0645716
BIO2	4.8205966	4.5718047	11.0136310	1.0545310
BIO3	0.0152292	1.4448899	8.8370205	67.6857465
BIO4	4.0237492	8.3401251	2.6646963	3.4103375
BIO5	9.6134091	0.1147448	0.9777419	1.6331787
BIO6	1.6536763	14.5176075	3.9239628	0.0277513
BIO7	4.9299377	6.5611747	6.4298897	2.0976882
BIO8	1.5193717	0.0572568	29.8186906	0.7314443
BIO9	4.3561352	3.0489796	5.0762722	1.4549497
BIO10	8.7154304	2.1544261	0.6099212	1.6961369
BIO11	3.0043533	12.2423680	3.3810527	0.3922177
BIO12	7.1915029	3.8717742	3.0380297	2.6807190
BIO13	4.8647890	8.2249642	2.5655466	2.3876658
BIO14	9.3104934	0.7934852	0.3905980	0.0314138
BIO15	4.2125735	8.4240260	2.1276919	9.6168485
BIO16	4.4397493	8.6228212	4.2245794	2.6070776
BIO17	9.3610386	0.5081155	0.4864080	0.0102159
BIO18	9.1125024	0.2925984	2.4966435	0.0433115
BIO19	2.8364316	9.2250728	9.8649177	2.3741943

Tabla 7. Contribución de las variables a los componentes pirncipales

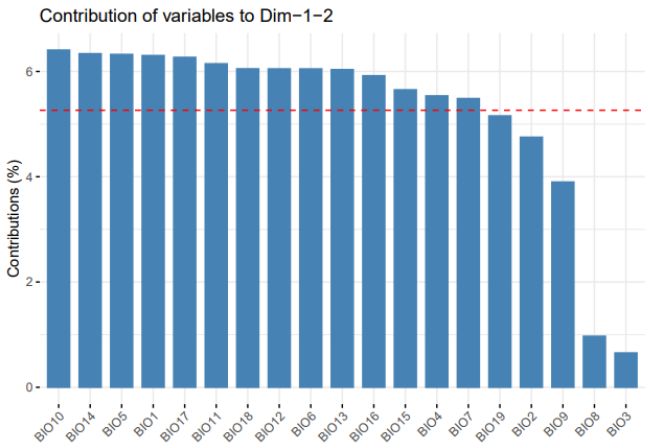


Figura 27. Gráfica de la contribución de las variables a los componentes principales

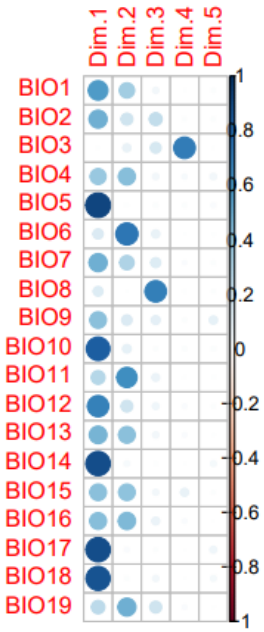


Figura 28. Contribución de las variables a cada dimensión

La figura 28 muestra las variables que más contribuyen a cada dimensión.

- El factor 1 está representado por BIO5, BIO10, BIO14, BIO17 y BIO18 esencialmente.
- El factor 2 por BIO6 y BIO11.
- El factor 3 por BIO8.
- El factor 4 por BIO3.

Mediante el ACP agrupamos las 19 variables climáticas de temperatura y precipitación en cuatro factores. Nos interesa quedarnos con el valor de cada factor para cada caso, es decir, los valores de las cuatro primeras componentes principales para las 4.683 observaciones de nuestra base de datos. Los guardamos como un nuevo *data.frame* en R*,* al que posteriormente unimos las demás variables de riqueza, orilla, altitud y *footprint* que se emplean en el siguiente apartado.

5.3 Modelo Lineal Generalizado: modelo log-lineal de Poisson

En este trabajo se ha decidido representar la biodiversidad mediante el parámetro de la riqueza, entendida como número de especies detectadas en cada cuadrícula UTM 10x10km. La modelización espacial de la biodiversidad presenta bastantes dificultades. Los procesos de modelización espacial de la riqueza asumen un equilibrio entre el medio ambiente y el patrón de biodiversidad observado, es decir, la información con la que se trabaja es estática por la escasez de muestreos realizados en diferentes momentos para valorar cómo influyen las variables ambientales en la riqueza. Otro aspecto a tener en cuenta es que las variables predictoras utilizadas pueden no ser las variables directas que influyen en el patrón espacial, las variables más influyentes pueden no haberse considerado por la dificultad de su medición (por ejemplo, capacidad de dispersión de los anfibios).

Las asunciones del modelo de Poisson es que la media es igual a la varianza, por eso antes de ajustar el modelo se calcularon estos parámetros obteniéndose que μ=3’970105 y σ^2=6’172536. Que la varianza sea superior a la media nos sugiere que tendremos una sobredispersión en el modelo.

A continuación ajustamos el modelo log-lineal de Poisson, obteniendo que las siete variables son significativas: Dim.1, Dim.2, Dim.3, Dim.4, *footprint,* orilla y altitud, con p-valores de $<2e^{-16}$, $<2e^{-16}$, 1’10e^{-16}, $<2e^{-16}$, 0’0463, $<2e^{-16}$ y 6.37e^{-6} respectivamente.

La dispersión del modelo se calculó con un test overdispersion, obteniéndose un valor de 1’3. Hay una pequeña sobredispersión, que al no ser demasiado elevada se podría obviar puesto que las estimaciones son correctas, a pesar de que la desviación estándar sea incorrecta y el modelo no la tenga en cuenta.

La evaluación de la calidad del modelo se calcula con la Devianza (D^2), donde se obtuvo que el modelo explica el 14’42% de la variabilidad. El *test* X^2, salió significativo (p-valor: $<2'2e^{-16}$) por lo que el ajuste del modelo es bueno.

Cuando el número de observaciones es lo suficiente grande, como es nuestro caso, los estadísticos de Devianza (D^2) y X^2 son equivalentes.

Se analizaron los residuos del modelo, sin obtener ningún resultado que no valide las asunciones (figura 29).

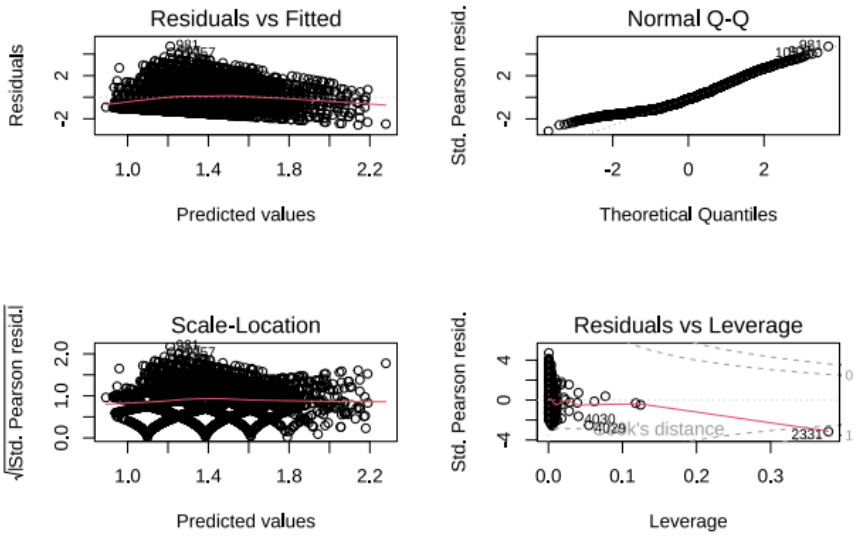


Figura 29. Residuos del modelo

El test de Shapiro-Wilk, usando los residuos de devianza (no en la escala original de la respuesta) tiene un p-valor significativo (<2’2e-16), aceptando por tanto la hipótesis de normalidad aunque en el gráfico de residuos ya se apreciaba que se cumplía.

Para tener un error estándar más correcto, podemos usar un modelo cuasi-poisson y comparar los resultados de ambos modelos (tabla 8).

	coef1	se.coef1	coef2	se.coef2	exponent
(Intercept)	1.4134070	0.0332046	1.4134070	0.0383346	4.1099341
Dim.1	-0.0509341	0.0033757	-0.0509341	0.0038972	0.9503413
Dim.2	0.0314192	0.0064477	0.0314192	0.0074439	1.0319180
Dim.3	0.0575691	0.0066554	0.0575691	0.0076836	1.0592585
Dim.4	0.0909593	0.0074981	0.0909593	0.0086565	1.0952244
footprint	0.0000296	0.0000148	0.0000296	0.0000171	1.0000296
orilla	0.0000320	0.0000036	0.0000320	0.0000042	1.0000320
altitud	-0.0002025	0.0000449	-0.0002025	0.0000518	0.9997975

Tabla 8. Comparación GLM poisson y GLM cuasi-poisson

Los coeficientes son los mismos, el error estándar es el que varía. La devianza del modelo de quasi-poisson es 1’323846, no varía mucho en comparación con la devianza del GLM Poisson.

5.4 Modelización espacial de la riqueza de anfibios

Los coeficientes obtenidos con los que elaborar un modelo extrapolable a una representación geográfica con ArcMap son los siguientes:

	x
(Intercept)	1.4134070
Dim.1	-0.0509341
Dim.2	0.0314192
Dim.3	0.0575691
Dim.4	0.0909593
footprint	0.0000296
orilla	0.0000320
altitud	-0.0002025

Tabla 9. Coeficientes modelo de Poisson

Por tanto, el **modelo de predicción de Poisson** obtenido fue el siguiente:

$$\mathbf{y} = e^{(1.413407D - \mathbf{Dim1} * 5.093408e^{-2} + \mathbf{Dim2} * 3.141923e^{-2} + \mathbf{Dim3} * 5.756913e^{-2} + \mathbf{Dim4} * 9.095929e^{-2} + \mathbf{footprint} * 2.959291e^{-2} + \mathbf{orilla} * 3.198961e^{-5} - \mathbf{altitud} * 2.024981e^{-4})}$$

Se creó una nueva base de datos con las variables de orilla, altitud, *footprint* y los cuatro primeros factores de la ACP, además de una nueva variable de idoneidad para riqueza donde se metió la fórmula del modelo predictivo de Poisson en Excel con los coeficientes de la tabla 9.

La figura 30 representa la riqueza de especies por cuadrícula UTM con los datos extraídos del Banco de Datos de la Naturaleza del Ministerio para la transición ecológica y el reto demográfico. Observamos que la mayor riqueza de especies se sitúa en ciertas zonas del norte, como en Galicia y Asturias, Cataluña (especialmente en su zona norte), el oeste de la península, algunas zonas del centro peninsular relacionado con cadenas montañosas y zonas húmedas de Andalucía.

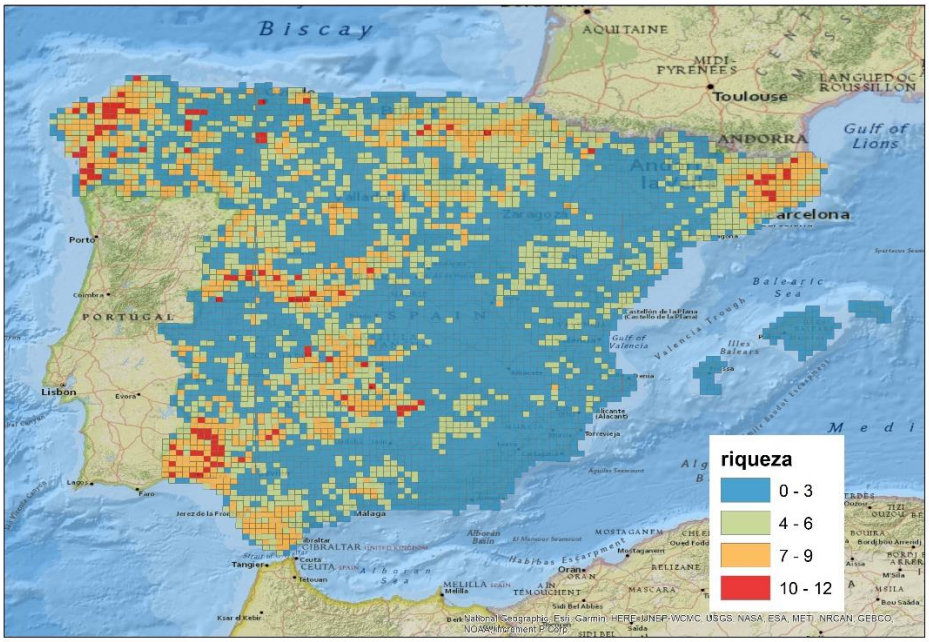


Figura 30. Representación gráfica de la riqueza de especies por cuadrícula UTM

El mapa de la figura 31 se creó mediante el algoritmo obtenido en la modelización de Poisson. Este mapa representa los lugares más adecuados para la existencia de una mayor diversidad de especies de anfibios en España peninsular y Baleares. La Idoneidad está categorizada en cuatro niveles: riqueza nula (0), riqueza baja (0-3´83577), riqueza media (3´83577-5´32386) y riqueza alta (5´32386-9´77132).

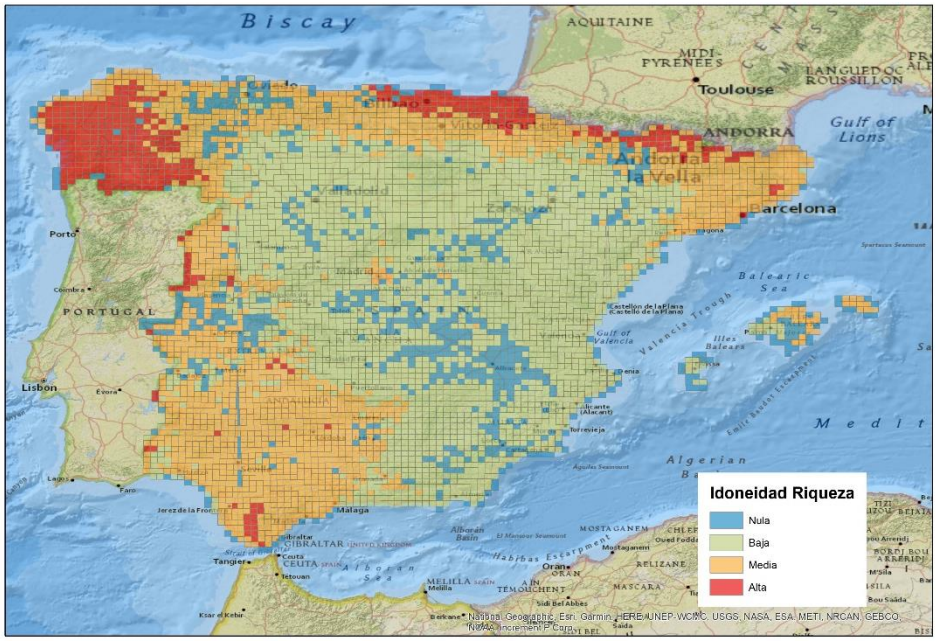


Figura 31. Mapa de idoneidad del territorio para la riqueza de especies de anfibios en la Península Ibérica y Baleares

El mapa de idoneidad (figura 31) muestra diferencias con el mapa de presencias conocidas (figura 30), pero se aprecia que las zonas con más idoneidad del territorio vuelven a ser la zona norte (especialmente Galicia) y

toda la España Eurosiberiana, además del norte de Cataluña y Aragón, junto con el oeste peninsular, zonas húmedas de Andalucía y otros puntos dispersos con valores elevados de idoneidad. En las Islas Baleares es donde se nota más diferencia, teníamos registrada una riqueza de 0 a 3 especies en esas cuadrículas, pero el mapa de idoneidad refleja que hay zonas en las que la idoneidad es media, por lo que las condiciones ambientales permitirían la presencia de un número de especies mayor al que actualmente se conocen.

5.5 Selección de las variables más influyentes con Random Forest

Se decidió categorizar la variable riqueza puesto que no nos interesa que el modelo prediga la riqueza exacta, lo que se quiere es averiguar qué variables tienen más importancia en la riqueza de anfibios. Se hicieron pruebas con **distintas categorizaciones de la variable riqueza** en Random Forest**.**

- **Categorización en tres clases**

Se establecieron tres clases: riqueza baja (0-3 especies por cuadricula UTM), media (4-6) y alta (7-12). El valor obtenido de OOB fue de 29’62% y el *Accuracy* fue de 0’716.

- **Categorización en dos clases**

Se establecieron dos clases, la clase 1 (baja riqueza) de 0 a 6 y la clase 2 (alta riqueza) de 6 a 12 especies por cuadrícula UTM. En este modelo el valor de OOB fue de 13’48% y el de *Accuracy* 0’8612.

- **Categorización en dos clases separando por la mediana**

En la anterior categorización en dos clases los datos no estaban balanceados, por lo que se probó a establecer dos clases a partir de la mediana. La mediana es una medida de localización central que se define como aquel valor que divide un conjunto de observaciones, ordenadas de menor a mayor, en dos partes con el mismo número de observaciones o como aquel valor que divide los datos en dos partes de igual probabilidad. La mediana tenía un valor de 3, por tanto la clase 1 de baja riqueza abarcaba de 1 a 3 especies por cuadrícula UTM, y la clase 2, de alta riqueza, de 4 a 12 especies por cuadrícula UTM. Los valores obtenidos de OOB y de *Accuracy* fueron 27’09% y 0’7381 respectivamente. Partiendo los datos por la mediana se obtenían peores resultados que en la categorización de clase 1 (0-6) y clase 2 (7-12), por lo que en posteriores análisis se empleó la anterior partición de datos.

También se emplearon los algoritmos de k-vecinos más cercanos y Árboles de Decisión, obteniéndose valores de *Accuracy* más bajos que los conseguidos

al emplear Random Forest. Se pueden consultar en el anexo estos resultados. Random Forest es una mejora del modelo de los árboles de decisión por lo que era esperable obtener un mejor valor de *Accuracy*.

Por tanto, el Random Forest elegido fue con la categorización en dos clases, la primera de 0 a 6, y la segunda de 7 a 12. Se seleccionó el 70% de los datos de la base original para el conjunto de entrenamiento o *train* y el 30% de los datos restantes para el conjunto *test*. La base de datos original tiene 4.638 registros; el conjunto *train* estaba constituido por 3.278 y el conjunto *test* por 1.405 observaciones.

En el **entrenamiento** del modelo se obtuvo la matriz de confusión de la tabla 10. En la figura 32 se aprecia como el error de predicción de las clases y de OOB disminuye y se estabiliza en un punto al aumentar el número de árboles.

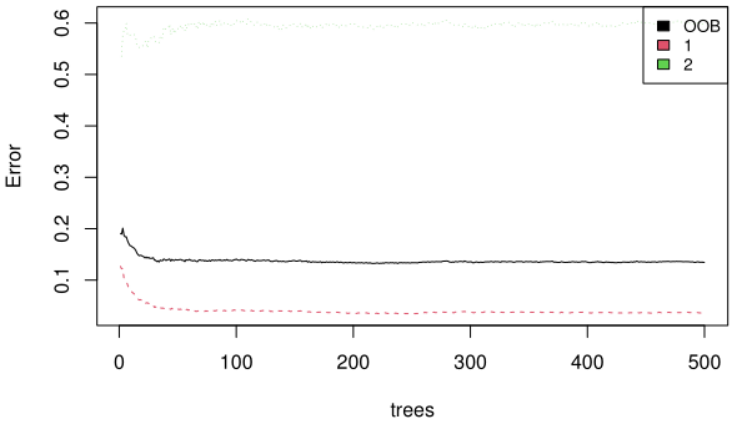


Figura 32. Gráficas trees vs error

1	2	class.error
2605	98	0.0362560
344	231	0.5982609

Tabla 10. Matriz de confusión con el conjunto *train*

Al realizar las **predicciones** con el conjunto *test* se obtuvo la matriz de confusión de la tabla 11. Como se ha dicho anteriormente, el valor de Accuracy es de 0’8612, un muy buen valor que indica que el modelo predice bien el 86’12%.

1	2
1116	46
149	94

Tabla 11. Matriz de confusión con el conjunto *test*

Una vez calculado el modelo, miramos cuáles son las variables que ha considerado como más importantes mediante el **Índice de Gini** que mide la reducción de la impuridad nodal media. En la tabla 12 y la figura 33 observamos los valores de cada variable; la que más reduce la impuridad del Random Forest es la más importante.

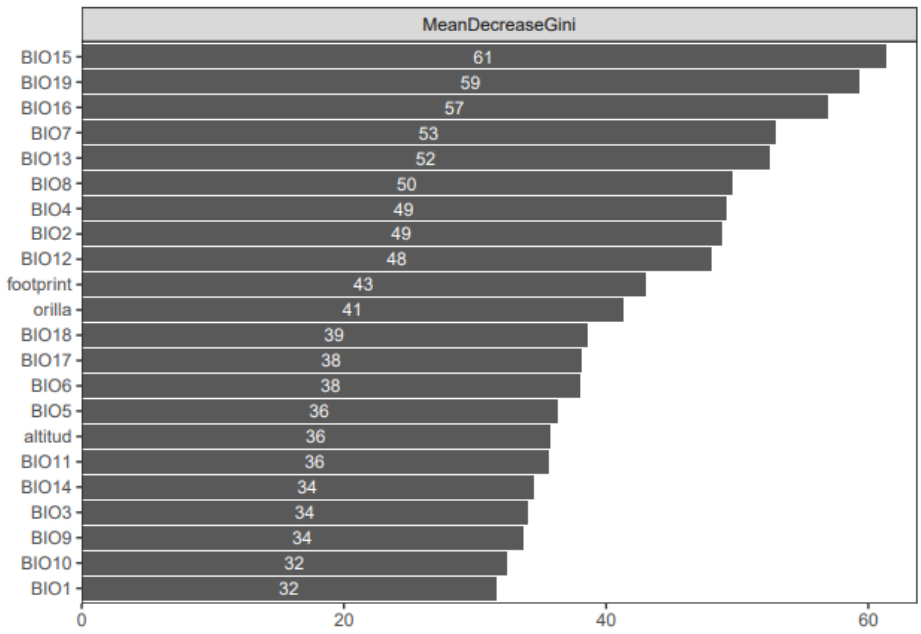


Figura 33. Gráfica de importancia de las variables

	MeanDecreaseGini
orilla	41.28433
altitud	35.68719
BIO1	31.62867
BIO2	48.75822
BIO3	34.01936
BIO4	49.14804
BIO5	36.21526
BIO6	37.97462
BIO7	52.93403
BIO8	49.62447
BIO9	33.65471
BIO10	32.41072
BIO11	35.53966
BIO12	48.04492
BIO13	52.45040
BIO14	34.41342
BIO15	61.27157
BIO16	56.91187
BIO17	38.03711
BIO18	38.54708
BIO19	59.31052
footprint	42.99971

Tabla 12. Valores del Indice de Gini de las variables

Establecemos 50 como valor umbral (figura 34) y nos quedamos con seis variables que son las que más influyen en la distribución de la riqueza de anfibios en España peninsular y Baleares. Son las siguientes, ordenadas de mayor a menor importancia:

- BIO15: estacionalidad de la precipitación
- BIO19: precipitación del trimestre más frío
- BIO16: precipitación del trimestre más húmedo
- BIO7: rango anual de temperatura (temperatura del mes más cálido – temperatura del mes más frío)
- BIO13: precipitación del mes más lluvioso
- BIO8: temperatura media del trimestre más húmedo

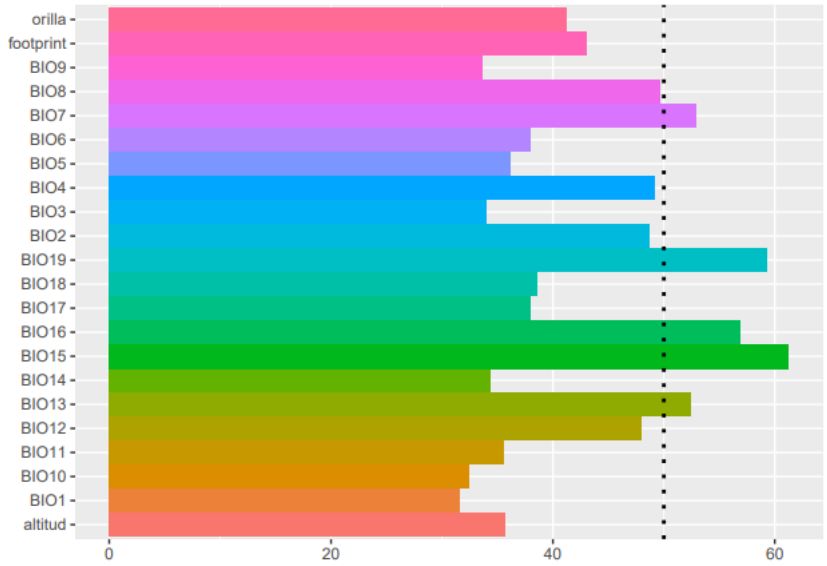


Figura 34. Gráfica de la selección de las variables más influyentes (punto de corte en 50)

En el anexo se presenta el código empleado para optimizar los hiperparámetros. Se obtuvo que los valores óptimos para mejorar la predicción eran *mtry* = 17 y *ntree* = 5.000. El coste computacional de emplear 5.000 árboles es muy alto por lo que finalmente nos quedamos con el anterior modelo cuyos valores son buenos y fiables para la selección de variables, además de poseer sentido biológico, como se explica en la discusión.

6 Discusión

La importancia de las variables ambientales para la distribución, supervivencia y etología de los anfibios ha sido ampliamente relatada en la bibliografía [10] [11] [12] [14] [20] [21] [22] [23] [25] [26]. Por ejemplo, nuestros resultados muestran que las variables de temperatura se correlacionan negativamente con la riqueza de anfibios; a mayor temperatura, menor riqueza, tal y como sugieren también Montori [12] y Pérez-Castillo [46]. Por otro lado, las altas temperaturas generan las condiciones adecuadas para el surgimiento del hongo quitridio (*Batrachochytrium dendrobatidis)* que causa la enfermedad de la quitridiomicosis, grave causante del declive poblacional de los anfibios [44] [19]. Además, los machos de diversas especies de anuros emplean reclamos acústicos durante la época reproductora para atraer a las hembras, esta condición etológica está estrechamente ligada con la temperatura por lo que el aumento de temperaturas podría afectar a dicha actividad [45].

Otras variables climáticas relevantes son aquellas relativas a las precipitaciones, tal y como indica Guerrero [48] y Antutúnez [49]. Los anfibios dependen de masas de agua para llevar a cabo la reproducción; el amplexo tiene lugar en el agua, los huevos carecen de pared protectora y si no se depositan en el agua se secan [13], por otro lado, los renacuajos son plenamente acuáticos [14], es decir, los requerimientos hídricos de los anfibios son elevados. Esto queda reflejado en la correlación positiva que muestran nuestros resultados en relación con las variables de precipitación; a mayor precipitación, mayor riqueza. De forma indirecta, un aumento de las precipitaciones provoca una mayor superficie de orilla lo que amplía la disponibilidad de zonas reproductoras para los anfibios (charcas, ríos, lagos, lagunas, arroyos, pozas…).

Es decir, tal y como indican los autores Werner [20], Enriqueza-Urzelai [21] y Sillero [22] y muestran nuestros resultados, las variables ambientales tienen un gran peso en la distribución de los anfibios, específicamente las referidas a las precipitaciones. Al realizar el Random Forest nos sale que las variables con más influencia en la distribución de la riqueza de anfibios son BIO15 (estacionalidad de la precipitación), BIO19 (precipitación del trimestre más frío), BIO16 (precipitación del trimestre más húmedo), BIO7 (rango anual de temperatura), BIO13 (precipitación del mes más lluvioso) y BIO8 (temperatura media del trimestre más húmedo), mientras que factores topográficos como altitud y orilla, y antrópicos, como *footprint*, tenían mucha menos relevancia. Esta importancia de las variables de precipitación se refleja fielmente en nuestra representación gráfica, ya que muestra que los mayores valores de riqueza se encuentran en el norte y zonas de costa del noreste de la península, donde las precipitaciones son más abundantes.

Campos [35] también establece la variable BIO19 como relevante, sin embargo, no considera el resto de variables, posiblemente la localización

reducida de su trabajo (limitada a la Reserva de la Biosfera Transfronteriza Meseta Ibérica) pueda influir en dicho resultado ya que nosotros trabajamos a una escala espacial mayor correspondiente a toda España peninsular y Baleares.

La representación geográfica de la idoneidad del territorio para la riqueza de especies muestra que las zonas más adecuadas coinciden con áreas asignadas a climas atlánticos o con influencia de estos, como son las zonas más occidentales del país, mientras que las áreas mediterráneas y continentales (Sistema Central, Meseta Castellana, zonas de Ciudad Real, Toledo y Extremadura…) presentan valores de baja idoneidad, siendo estas las regiones que más se van a ver afectadas por el cambio climático y por tanto mayor afección a la riqueza de anfibios [11]. Esta afección es mayor cuando se producen cambios climáticos a corto plazo provocando el declive de las poblaciones [10] o modificando su área de distribución [11].

Esta dependencia en las variables climáticas (junto a su sensibilidad a los contaminantes) hace que podamos considerar a los anfibios como potenciales bioindicadores de la salud ambiental de los ecosistemas [8] [9] o del cambio climático, tal y como ocurre con otros grupos faunísticos (lepidópteros) [50] [51].

La varianza explicada por el modelo es del 14’42%, por tanto no se puede establecer una relación causa-efecto porque hay muchas variables que deberían tenerse en cuenta y no pueden considerarse. Los estudios basados en la relación directa de variables predictoras y la distribución de especies faunísticas se encuentran con el problema de la imposibilidad de valorar la importancia de variables no mesurables; poder cuantificar la competencia por los recursos, las interacciones ecológicas entre los distintos niveles de las cadenas tróficas, o la biogeografía histórica en entornos abiertos y no controlables como es la naturaleza, resultan difícilmente valorables. Por ello, debemos considerar nuestros resultados como de carácter asociativo y no causal, pero al ser descriptivos, permiten establecer la base para futuros trabajos.

7 Conclusiones

7.1 Conclusiones

• A pesar de la dificultad que representa la utilización de una variable combinatoria como es la riqueza de anfibios, que implica la selección de variables generalistas predictoras para todas las especies involucradas, la combinación de distintas herramientas bioinformáticas y de representación geográfica resultan especialmente útiles para valorar la biodiversidad de taxones concretos en el territorio peninsular.

• Las variables ambientales han resultado ser elementos fundamentales para establecer la distribución de la riqueza de anfibios a nivel peninsular e islas Baleares.

• La combinación de variables correlacionadas entre ellas es útil siempre que se puedan agrupar en un ACP que reduce la dimensionalidad sin perder información, sin embargo, esto imposibilita establecer la importancia de cada una de las variables en una situación climática futura, por lo cual se hace necesaria la utilización de Random Forest para establecer el peso de cada una de las variables.

• Random Forest es la técnica más adecuada para este tipo de análisis por su capacidad para tratar con datos ruidosos, resulta especialmente útil para la discriminación de nuestras variables climáticas predictoras que están bastante correlacionadas.

• El conocimiento de las variables predictoras más importantes nos permite pronosticar el declive de la biodiversidad de anfibios a partir de los diferentes escenarios climáticos propuestos y establecer medidas de gestión adecuadas para la conservación.

• El mapa de idoneidad obtenido ayudará en la selección de zonas de muestreo futuras en los distintos programas de seguimiento que se realizan, como por ejemplo el proyecto SARE.

• La riqueza de anfibios puede emplearse como un parámetro bioindicador fiable en el contexto de cambio climático.

7.2 Líneas de futuro

- Comprobación empírica de la coincidencia de la riqueza de especies con el mapa de idoneidad establecido mediante muestreos de campo.

- Sistematizar y generalizar la toma de datos en todas las unidades de muestreo de España peninsular y Baleares para obtener datos fiables de cada cuadrícula UTM y aumentar así la precisión de los análisis. El proyecto SARE (Seguimiento de Anfibios y Reptiles de España) podría ayudar a resolver este problema. Dicho programa de voluntariado consiste en el seguimiento a largo plazo de las poblaciones de anfibios y reptiles para obtener series largas que permitan determinar la evolución de las poblaciones.

- Intentar inferir, aunque sea de forma indirecta, las variables no mesurables comentadas, por ejemplo, capacidad de dispersión de los anfibios, competencia por los recursos con otras especies, interacciones ecológicas entre los distintos niveles de las cadenas tróficas (presencia de depredadores), razones históricas como movimientos tectónicos o glaciaciones que influyen en la distribución biogeográfica de las poblaciones, razones ambientales como la respuesta de las especies a la energía disponible, estructura del paisaje y vegetación…

- Realizar un análisis sectorial por regiones bioclimáticas (mediterránea, continental y eurosiberiana) analizando el peso de cada variable climática en cada uno de ellos.

- Repetir los análisis con niveles taxonómicos inferiores, por ejemplo a nivel de familia; de esta forma podríamos observar si, por ejemplo, a las especies de sapillos les afectan variables distintas a las de ranas, tritones o gallipatos.

7.3 Seguimiento de la planificación

Inicialmente se hicieron regresiones lineales múltiples con enfoque descendente, es decir, ir eliminando las variables que no saliesen significativas en el modelo hasta quedarnos con las más importantes con el mejor valor de AIC. Como se ha visto a lo largo del trabajo, las variables climáticas (BIO1 a BIO19) estaban muy correlacionadas por lo que no era correcto incluir todas las variables en el análisis sin hacer una selección previa. Para ello se emplearon análisis de correlaciones y factores de inflación de la varianza, pero puesto que la hipótesis del trabajo era que todas son importantes e influyen en mayor o menor medida, y emplear todas garantizaba una mejor representación geográfica, se decidió agruparlas en factores con un Análisis de Componentes Principales. Además, las regresiones lineales múltiples estaban mal

planteadas, porque la variable riqueza sigue una distribución de Poisson, por eso se realizaron Modelos Lineales Generalizados.

En la realización del trabajo surgieron varios contratiempos personales y técnicos (fallos en el ordenador) que retrasaron la planificación inicial por semanas. No se pudieron probar todos los algoritmos de Machine Learning que se plantearon en un principio, porque además, tras realizar el *Exploratory Data Analysis* y tener una buena visión sobre la base de datos a la que nos enfrentamos, quedó patente que muchos de ellos no iban a ser útiles para resolver nuestras hipótesis. El estudio bibliográfico reveló que para nuestro problema planteado lo más adecuado era utilizar Random Forest.

A pesar de los retrasos en la planificación, la metodología final planteada fue la correcta para la consecución de nuestros objetivos.

8 Glosario

- **UICN**: Unión Internacional para la Conservación de la Naturaleza
- **ACP**: Análisis de Componentes Principales
- **GLM**: Modelo Lineal Generalizado
- **Amplexo**: modo de apareamiento propio de los anfibios anuros y urodelos en el que hembras y machos se reúnen en el agua y se acoplan; el macho, más pequeño que la hembra, se abraza a ella sujetándola por debajo de sus extremidades anteriores o por encima de las posteriores. La hembra vierte al agua los huevos sin fecundar y el macho libera los espermatozoides a la vez, produciéndose inmediatamente la fecundación.
- ***Hotspot* de biodiversidad:** término acuñado por Norman Myers refiriéndose a las regiones biogeográficas con una alta biodiversidad amenazada.

9 Bibliografía

1. Beltrán, N., Martínez, R., Perales.J (2016). Guía de Anfibios de los Parques Nacionales españoles. Organismo Autónomo Parques Nacionales Ministerio de Agricultura, Alimentación y Medio Ambiente. ISBN 978-84-8014-900-6

2. https://www.fundacionaquae.org/wiki-explora/45_anfibios/index.html [2/05/2022]

3. Rittenhouse, T.A.G. & Semlitsch, R.D (2007). Distribution of amphibians in terrestrial habitat surrounding wetlands. *Wetlands* 27:153-161.

4. https://www.asturnatura.com/orden/anura.html [2/05/2022]

5. https://www.asturnatura.com/orden/caudata.html [2/05/2022]

6. https://anfibios.animalesbiologia.com/informacion/orden-caudata-caudados [2/05/2022]

7. Cortéz-Gómez, AMM., Ruiz-Agudelo, CA., Valencia-Aguilar, A., & Ladle, RJ (2015). Ecological functions of neotropical amphibians and reptiles: a review. *Universitas Scientiarum,* 20(2): 229-245

8. Lambert, M.R.K. (1997). Environmental effects of heavy spillage from a destroyed pesticide store near Hargeisa (Somaliland) assessed during the dry season, using reptiles and amphibians as bioindicators. *Archives of Environmental Contamination and Toxicology*, *32*(1): 80-93.

9. Saber, S., Tito, W., Said, R., Mengistou, S., & Alqahtani, A. (2017). Amphibians as bioindicators of the health of some wetlands in Ethiopia. *The Egyptian Journal of Hospital Medicine*, 66(1), 66-73.

10. Blaustein, R.A., Belden, L.K., Olson, D.H., Green, D.M., Root, T.L. & Kiesecker, J.M. (2001). Amphibian breeding and climate change. *Biology*, 15: 1804-1809.

11. Araújo, M.B., Thuiller, W., & Pearson, R. G. (2006). Climate warming and the decline of amphibians and reptiles in Europe. *Journal of Biogeography*, 33(10): 1712-1728.

12. Montori, A., Giner, G., Béjar, X., & Álvarez, D. (2011). Descenso brusco de temperaturas y nevadas tardías como causas de mortalidad de anfibios durante el período reproductor. *Boletín de la Asociación Herpetológica Española* 22: 72-74.

13. Duellman, W. E., & Trueb, L. (1994). Biology of amphibians. *Baltimore*: Johns Hopkins University Press.

14. Aguillón-Gutiérrez, D. R. (2018). Anomalías macroscópicas en larvas de anfibios anuros. *Revista Latinoamericana de Herpetología*, 1(1), 8-21.

15. https://www.UICNredlist.org/resources/summary-statistics [3/05/2022]

16. https://www.UICNredlist.org/es/ [3/05/2022]

17. IPBES (2019): Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. *IPBES secretariat*, Bonn, Germany. 56 pages.

18. Comité Español de la UICN y Fundación Naturaleza y Hombre (2019). Análisis de las especies de la Lista Roja de la UICN en España: una llamada urgente a la acción. Málaga-Santander (España). [https://www.uicn.es/web/pdf/Analisis_L_Roja_Spain2019.pdf]

19. WorldWildlifeFund (2013). Prioridades para la conservación de anfibios en España. Madrid: WWF, Programa de especies WWF en España [https://www.wwf.es/?28125/WWF-alerta-de-que-son-necesarias-medidas-urgentes-para-la-proteccin-de-los-anfibios[

20. Werner, E., Skelly, D., Relyea, R., & Yurewicz, K (2007). Amphibian species richness across environmental gradients. *Oikos* 116: 1697-1712.

21. Enriquez-Urzelai, U., Bernardo, N., Moreno-Rueda, G., Montori, A., & Llorente, G. (2019). Are amphibians tracking their climatic niches in response to climate warming? A test with Iberian amphibians. C*limatic Change* 154(1): 289-301.

22. Sillero, Neftalí. (2021). Climate change in action: local elevational shifts on Iberian amphibians and reptiles. *Regional Environmental Change*. 21. 10.1007/s10113-021-0183

23. Sillero, N., Brito, J.C., Skidmore, A.K. & Toxopeus, A.G. (2009). Biogeographical patterns derived from remote sensing variables: the amphibians and reptiles of the Iberian Peninsula. *Amphibia-Reptilia* 30: 185-206.

24. Hutchinson, A. H. (1918). Limiting factors in relation to specific ranges of tolerance of forest trees. *Bot. Gaz.* 96: 465–493.

25. Cunningham, Heather & Rissler, Leslie & Buckley, Lauren & Urban, Mark. (2015). Abiotic and biotic constraints across reptile and amphibian ranges. *Ecography*. 39(1).

26. Enriquez-Urzelai, U., Bernardo, N., Moreno-Rueda, G., Montori, A. & Llorente G.A.(2019). Are amphibians tracking their climatic niches in response to climate warming? A test with Iberian amphibians. *Climatic Change*.

27. Howard, Sam & Bickford, David. (2014). Amphibians over the edge: Silent extinction risk of Data Deficient species. *Diversity and Distributions*, 20 (1)

28. Pleguezuelos, J.M., Márquez, R. & Lizana, M. (2002). Atlas y libro rojo de los anfibios y reptiles de España. Dirección general de Conservación de la Naturaleza (Ministerio de Medio Ambiente) - Asociación Herpetológica Española. Madrid.

29. https://herpetologica.es/category/programas/programa-sare/ [10/05/2022]

30. Nogues, D.B. 2003. Estudio de la distribución espacial de la biodiversidad: conceptos y métodos. *Cuadernos de Investigación Geográfica* 29:67-82.

31. Habib, N. S., Abu Maghasib, O. K., & Al-Ghazali, A. R. (2021). Predicting the Presence of Amphibian Species Using Features Attained from GIS and Satellite Images. *International Journal of Academic and Applied Research (IJAAR),* 5(4), 56-65.

32. Girardello, M., Griggio, M., Whittingham, M. J., & Rushton, S. P. (2010). Models of climate associations and distributions of amphibians in Italy. Ecological research, 25(1), 103-111.

33. Martín, B., González-Arias, J., Vicente-Vírseda, J. A. (2021). Machine learning as a successful approach for predicting complex spatio–temporal patterns in animal species abundance. *Animal Biodiversity and Conservation*, 44: 289-301.

34. Sousa-Guedes, D.; Arenas-Castro, S.; Sillero, N. (2020) Ecological Niche Models Reveal Climate Change Effect on Biogeographical Regions: The Iberian Peninsula as a Case Study. *Climate* 8(3):42.

35. Campos, João & Rodrigues, Sara & Freitas, Teresa Raquel & Santos, João & Honrado, João & Regos, Adrián. (2021). Climatic variables and ecological modelling data for birds, amphibians and reptiles in the Transboundary Biosphere Reserve of Meseta Ibérica (Portugal-Spain). *Biodiversity Data Journal*. 9. 10.3897/BDJ.9.e66509.

36. Kaiser, Henry F. A Note on Guttman's Lower Bound for the Number of Common Factors. British Journal of Statistical Psychology 14: 1-2 (1961)

37. https://www.fuenterrebollo.com/Economicas/ECONOMETRIA/MULTIVARIANTE/ACP/ACP.pdf [5/04/2022]

38. https://www.xlstat.com/es/soluciones/funciones/analisis-de-componentes-principales-acp [5/04/2022]

39. https://www.r-bloggers.com/2017/09/how-random-forests-improve-simple-regression-trees/ [20/05/2022]

40. https://www.iartificial.net/random-forest-bosque-aleatorio/ [20/05/2022]

41. https://www.listendata.com/2014/11/random-forest-with-r.html#id-0c1e65 [21/05/2022]

42. https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting#Random_Forest [22/05/2022]

43. https://rpubs.com/Joaquin_AR/255596 [23/05/2022]

44. Andrés Fernández Loras. Quitridiomicosis en anfibios: Inmunidad, tratamiento y mitigación en el medio natural (2021). *Tesis Doctoral,* Universidad Complutense de Madrid.

45. Llusia, D., Márquez, R., Beltrán, J. F., Benítez, M. Do Amaral, J. (2013). Calling behaviour under climate change: geographical and seasonal variation of calling temperatures in ectotherms. *Global Change Biology* 19, 2655-2674.

46. Pérez-Castillo, C.D, Fabián Medina-Rangel, G.F (2018). Influence of some environmental variables on relative abundance and bodyly characteristics of the rain frog *Pristimantis renjiforum* (Lynch, 2000) in Cundinamarca, Colombia. *Biodivers. Neotrop.* 8 (3): 157-67

47. Mendoza Miranda, D. P. (2018). Impacto de la precipitación en la distribución potencial de anfibios en el Parque Nacional Carrasco. Chapare, Bolivia (*Doctoral dissertation*).

48. Guerrero, J. C., Real, & Vargas, J. M. (1999). Asociaciones interespecíficas de los anfibios en los gradientes ambientales del sur de España. *Rev. Esp. Herp*, *13*, 49-59.

49. Antúnez, A., Real, R., & Vargas, J. M. (1988). Análisis biogeográfico de los anfibios de la vertiente sur de la Cordillera Bética. *Miscelania Zoologica*, 261-272.

50. España, Red de Parques Nacionales (2018). Seguimiento de lepidópteros en la red de parques nacionales de España.

51. Saldivar Solano, D. J., & Rigby Omier, K. K. (2020). Inventario de mariposas diurnas asociada a los agroecosistemas del Centro de Transferencia Agroforestal (CeTAF) como bioindicadores de la calidad

ambiental (*Doctoral dissertation*, Bluefields Indian & Caribbean University).