UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS AGRÓNOMOS

BASES METODOLÓGICAS PARA UN SISTEMA DE APOYO A LA SELECCIÓN DE ESPECIES EN LA RESTAURACIÓN DE LA VEGETACIÓN

TESIS DOCTORAL

AITOR GASTÓN GONZÁLEZ

LICENCIADO EN CIENCIAS AMBIENTALES

2011

DEPARTAMENTO DE PRODUCCIÓN VEGETAL: BOTÁNICA Y PROTECCIÓN VEGETAL
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS AGRÓNOMOS
UNIVERSIDAD POLITÉCNICA DE MADRID

BASES METODOLÓGICAS PARA UN SISTEMA DE APOYO A LA SELECCIÓN DE ESPECIES EN LA RESTAURACIÓN DE LA VEGETACIÓN

TESIS DOCTORAL

AUTOR: AITOR GASTÓN GONZÁLEZ
LICENCIADO EN CIENCIAS AMBIENTALES

DIRECTOR: JUAN IGNACIO GARCÍA VIÑAS
DOCTOR INGENIERO DE MONTES

Madrid, octubre de 2011

Tribunal nombrado por el Magfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid, el día …… de …….. de 200….

Presidente: ……..

Vocal: ……..

Vocal: ……..

Vocal: ……..

Secretario: ……..

Suplente: ……..

Suplente: ……..

Realizado el acto de defensa y lectura de la Tesis el día …… de …….. de 200…. en la ETSI/Facultad ……..

Calificación: …….

EL PRESIDENTE LOS VOCALES

EL SECRETARIO

A María

ÍNDICE

AGRADECIMIENTOS

Resulta difícil escribir los agradecimientos de un trabajo que se ha dilatado tanto en el tiempo, es fácil pasar por alto a alguien, así que me disculpo por anticipado.

Al profesor Juan Ignacio García Viñas le debo el primer lugar en estos agradecimientos por haberme ayudado a salir del oscuro callejón en el que me encontraba y haberme guiado de forma excelente por el camino recorrido desde aquel reinicio de la Tesis hasta llegar a buen puerto.

Mi gratitud a los compañeros de la EUIT Forestal de Madrid, que han aguantado demasiadas veces dosis inapropiadas de "modelización predictiva". En especial a Pepa Aroca, Alejandro Vivar y Enrique Sadornil que me han dedicado tiempo y paciencia para aclarar mis dudas estadísticas. También debo una mención especial a los compañeros de la Unidad Docente de Botánica, Bárbara Herrero, César López Leiva, Mar Génova, Pablo Galán, Carlos Ropero, Patricia Barberá y Judit Maroto por su apoyo incondicional en los momentos tan difíciles que todos hemos pasado. Gracias también a los compañeros del grupo de investigación ECOGESFOR, Sonia Roig, Alfredo Bravo, Leticia Carrero, Sergio González, Celia García Feced, Valentín Gómez y todos los demás, que con sus consejos, sugerencias y palabras de ánimo han impulsado esta Tesis hasta donde ha llegado.

Gracias a Roberto Vallejo y a Leopoldo Medina por resolver dudas relacionadas con los datos del Inventario Forestal Nacional y del Proyecto Anthos respectivamente.

A mis amigos, Patricio, Adrián, Olga, Nacho, Guille, Marta, Carolina, Agus, Jerónimo, Paloma, Txoni, Joseán y todos los demás, les debo el reconocimiento de su delicadeza al preguntar por la Tesis, combinando palabras de ánimo e interés con la frecuencia justa para no agobiar al doctorando que no cumple los plazos que él mismo se impone.

A mis padres y al resto de mi familia les debo haber puesto los cimientos del edificio que me ha permitido llegar hasta aquí. Por último, pero no menos importante, gracias a Marta por su apoyo incondicional, por sus consejos juiciosos y llenos de cariño.

Mil gracias a todos.

RESUMEN

El apoyo a la selección de especies a la restauración de la vegetación en España en los últimos 40 años se ha basado fundamentalmente en modelos de distribución de especies, también llamados modelos de nicho ecológico, que estiman la probabilidad de presencia de las especies en función de las condiciones del medio físico (clima, suelo, etc.). Con esta tesis se ha contribuido a la mejora de la capacidad predictiva de los modelos introduciendo algunas propuestas metodológicas adaptadas a los datos disponibles actualmente en España y enfocadas al uso de los modelos en la selección de especies.

No siempre se dispone de datos a una resolución espacial adecuada para la escala de los proyectos de restauración de la vegetación. Sin embrago es habitual contar con datos de baja resolución espacial para casi todas las especies vegetales presentes en España. Se propone un método de recalibración que actualiza un modelo de regresión logística de baja resolución espacial con una nueva muestra de alta resolución espacial. El método permite obtener predicciones de calidad aceptable con muestras relativamente pequeñas (25 presencias de la especie) frente a las muestras mucho mayores (más de 100 presencias) que requería una estrategia de modelización convencional que no usara el modelo previo.

La selección del método estadístico puede influir decisivamente en la capacidad predictiva de los modelos y por esa razón la comparación de métodos ha recibido mucha atención en la última década. Los estudios previos consideraban a la regresión logística como un método inferior a técnicas más modernas como las de máxima entropía. Los resultados de la tesis demuestran que esa diferencia observada se debe a que los modelos de máxima entropía incluyen técnicas de regularización y la versión de la regresión logística usada en las comparaciones no. Una vez incorporada la regularización a la regresión logística usando penalización, las diferencias en cuanto a capacidad predictiva desaparecen. La regresión logística penalizada es, por tanto, una alternativa más para el ajuste de modelos de distribución de especies y está a la altura

de los métodos modernos con mejor capacidad predictiva como los de máxima entropía.

A menudo, los modelos de distribución de especies no incluyen variables relativas al suelo debido a que no es habitual que se disponga de mediciones directas de sus propiedades físicas o químicas. La incorporación de datos de baja resolución espacial proveniente de mapas de suelo nacionales o continentales podría ser una alternativa. Los resultados de esta tesis sugieren que los modelos de distribución de especies de alta resolución espacial mejoran de forma ligera pero estadísticamente significativa su capacidad predictiva cuando se incorporan variables relativas al suelo procedente de mapas de baja resolución espacial.

La validación es una de las etapas fundamentales del desarrollo de cualquier modelo empírico como los modelos de distribución de especies. Lo habitual es validar los modelos evaluando su capacidad predictiva especie a especie, es decir, comparando en un conjunto de localidades la presencia o ausencia observada de la especie con las predicciones del modelo. Este tipo de evaluación no responde a una cuestión clave en la restauración de la vegetación ¿cuales son las *n* especies más idóneas para el lugar a restaurar? Se ha propuesto un método de evaluación de modelos adaptado a esta cuestión que consiste en estimar la capacidad de un conjunto de modelos para discriminar entre las especies presentes y ausentes de un lugar concreto. El método se ha aplicado con éxito a la validación de 188 modelos de distribución de especies leñosas orientados a la selección de especies para la restauración de la vegetación en España.

Las mejoras metodológicas propuestas permiten incrementar la capacidad predictiva de los modelos de distribución de especies aplicados a la selección de especies en la restauración de la vegetación y también permiten ampliar el número de especies para las que se puede contar con un modelo que apoye la toma de decisiones.

SUMMARY

During the last 40 years, decision support tools for plant species selection in ecological restoration in Spain have been based on species distribution models (also called ecological niche models), that estimate the probability of occurrence of the species as a function of environmental predictors (e.g., climate, soil). In this Thesis some methodological improvements are proposed to contribute to a better predictive performance of such models, given the current data available in Spain and focusing in the application of the models to selection of species for ecological restoration.

Fine grained species distribution data are required to train models to be used at the scale of the ecological restoration projects, but this kind of data are not always available for every species. On the other hand, coarse grained data are available for almost every species in Spain. A recalibration method is proposed that updates a coarse grained logistic regression model using a new fine grained updating sample. The method allows obtaining acceptable predictive performance with reasonably small updating sample (25 occurrences of the species), in contrast with the much larger samples (more than 100 occurrences) required for a conventional modeling approach that discards the coarse grained data.

The choice of the statistical method may have a dramatic effect on model performance, therefore comparisons of methods have received much interest in the last decade. Previous studies have shown a poorer performance of the logistic regression compared to novel methods like maximum entropy models. The results of this Thesis show that the observed difference is caused by the fact that maximum entropy models include regularization techniques and the versions of logistic regression compared do not. Once regularization has been added to the logistic regression using a penalization procedure, the differences in model performance disappear. Therefore, penalized logistic regression may be considered one of the best performing methods to model species distributions.

Usually, species distribution models do not consider soil related predictors because direct measurements of the chemical or physical properties are often lacking. The inclusion of coarse grained soil data from national or continental soil maps could be a reasonable alternative. The results of this Thesis suggest that the performance of the models slightly increase after including soil predictors form coarse grained soil maps.

Model validation is a key stage of the development of empirical models, such as species distribution models. The usual way of validating is based on the evaluation of model performance for each species separately, i.e., comparing observed species presences or absence to predicted probabilities in a set of sites. This kind of evaluation is not informative for a common question in ecological restoration projects: which n species are the most suitable for the environment of the site to be restored? A method has been proposed to address this question that estimates the ability of a set of models to discriminate among present and absent species in a evaluation site. The method has been successfully applied to the validation of 188 species distribution models used to support decisions on species selection for ecological restoration in Spain.

The proposed methodological approaches improve the predictive performance of the predictive models applied to species selection in ecological restoration and increase the number of species for which a model that supports decisions can be fitted.

CAPÍTULO 1

INTRODUCCIÓN

La restauración ecológica es una actividad humana que pretende iniciar o acelerar la recuperación de un ecosistema con respecto a su salud, integridad o sostenibilidad tras un proceso de degradación o destrucción (SER, 2004). Los gobiernos se han comprometido a promover la restauración de hábitats degradados como parte de las políticas de gestión y conservación de los recursos naturales (p.ej. la Directiva 92/43/CEE relativa a la conservación de los Hábitats Naturales y de la Fauna y La Flora Silvestres o el Convenio sobre la Diversidad Biológica de la Organización de Naciones Unidas) y en muchos países se ha establecido la obligación de corregir los impactos ambientales derivados de las actividades humanas (p.ej. en España, Real Decreto Legislativo 1/2008 de Evaluación de Impacto Ambiental de Proyectos)

La restauración de la vegetación no siempre implica la siembra o plantación de especies vegetales (Balaguer *et al.*, 2011), pero es una medida frecuentemente prevista en los proyectos de restauración ecológica (Clewell *et al.*, 2005) y por lo tanto el diseño de las actividades a menudo incluye la selección de las especies a utilizar en la restauración de la vegetación. Aunque los criterios para seleccionar especies pueden ser variados y dependerán en gran medida de los objetivos del proyecto, el proceso de selección de especies debe considerar en todo caso la idoneidad de las especies candidatas respecto a las características del medio físico (Serrada, 2000). Este trabajo trata los aspectos metodológicos de la estimación de la idoneidad de las especies vegetales y no incluye otros criterios para la selección de especies en el marco de la restauración ecológica.

1.1. ANTECEDENTES

El método más directo para la identificación de especies compatibles con las características del medio físico en un proyecto de restauración de la vegetación es acudir al conocimiento disponible sobre la flora y la vegetación local. Esta recomendación es una de las más habituales en las guías prácticas de selección de especies (p.ej. Ruiz de la Torre *et al.*, 1990). Este método minimiza la probabilidad de introducir especies no adaptadas a las condiciones locales y es el más deseable en muchos casos, pero tiene algunos inconvenientes: no siempre existe información

previa suficiente y a menudo no se dispone de recursos para contratar a un experto en la flora local, introduce cierto grado de subjetividad y además pueden no quedar restos de vegetación natural en el entorno. Con el objetivo de facilitar la selección de especies a personas no especialistas en la flora local, se han desarrollado diversos sistemas de apoyo a las decisiones.

Un método muy habitual para facilitar la selección de especies a personas no especialistas es recopilar la información disponible sobre los rangos de tolerancia de las especies frente a variables climáticas, edáficas y/o fisiográficas (Ruiz de la Torre *et al.*, 1990; Webb *et al.*, 1980; García Salmerón, 1980; De la Rosa *et al.*, 1992; Ruiz de la Torre *et al.*, 1996; CAB International, 2000; Ellis *et al.*, 2005; White & Dominy, 2005; Bravo & Montero, 2008; McVicar *et al.*, 2010). Los rangos de tolerancia suelen estar basados en estudios corológicos previos y en la opinión de expertos. Este método permite identificar las especies compatibles con las características del medio físico pero no permite ordenar dichas especies en función de su idoneidad.

Una alternativa a los rangos de tolerancia es el uso de clasificaciones ecológicas (Elena Rosselló, 1997; Pyatt & Suárez, 1997) en combinación con indicadores de asociación de las especies con las clases ecológicas (Castejón *et al.*, 1998; Ray *et al.*, 1998). Estos métodos se basan en muestreos y mapas temáticos, tanto para generar la clasificación ecológica como para estimar el grado de idoneidad de cada especie a cada clase ecológica. El procedimiento consiste en la identificación de la clase ecológica que corresponde a las características del medio físico del lugar a restaurar y la extracción de los valores de idoneidad de las especies para esa clase.

Algunos autores proponen usar los mapas de vegetación potencial como base para la selección de especies (Valle Tendero & Lorite Moreno, 2004; Carque *et al.*, 2008). El procedimiento consiste en ubicar la localidad del proyecto en un mapa de series de vegetación (p.ej. Rivas Martínez, 1987) y una vez identificada la serie a la que corresponde, consultar en la memoria del mapa las especies "indicadoras" de las etapas de sucesión para usarlas como lista de especies idóneas para la restauración. Estas series de vegetación no proceden del análisis de sucesiones reales observadas,

sino de reconstrucción a partir de observaciones e indicios (Terradas, 2001). La asignación o no de una especie a una serie es en gran medida fruto de la interpretación por parte del autor de los datos disponibles y esto le confiere al sistema un alto grado de subjetividad.

Un avance importante en la identificación de las especies idóneas para las condiciones del medio físico del lugar a restaurar lo constituye la introducción de modelos de distribución de especies (a menudo denominados modelos de nicho ecológico). Se trata de modelos empíricos que relacionan la distribución de las especies con la distribución de los factores del medio físico (Guisan & Thuiller, 2005). Dado que están basados en muestreos y en análisis numéricos, los modelos distribución de especies disminuyen la subjetividad en la identificación de las especies idóneas para la restauración. Es habitual que este tipo de modelos no consideren otros factores que determinan la presencia de las especies, como los procesos de fuente-sumidero, la heterogeneidad ambiental de grano fino o la ausencia de equilibrio entre la distribución actual y los factores del medio físico (Montoya *et al.*, 2009). Los modelos que incorporan características fisiológicas de las especies e interacciones interespecíficas (p.ej. Zavala & Bravo de la Parra, 2005) pueden ser más útiles que los modelos de distribución de especies, pero requieren información no disponible para muchas especies. A pesar de las simplificaciones que implican, los modelos de distribución de especies son actualmente la mejor alternativa disponible para la mayoría de los casos, cuando se pretende identificar especies idóneas asociadas a las condiciones del medio físico del lugar a restaurar.

Las primeras aplicaciones a la selección de especies en la restauración de la vegetación de estos modelos se pueden enmarcar en las denominadas envolventes ambientales que consisten en identificar la mínima envolvente que encierra las localidades de las especies en el espacio multidimensional formado por las variables ecológicas consideradas (Allué, 1990; Gandullo & Sánchez Palomares, 1994; García López & Allúe Camacho, 2004; Alonso Ponce *et al.*, 2010b). La disponibilidad de datos de presencia y ausencia de las especies (p.ej. el Mapa Forestal de España, Ruiz de la

Torre, 1990) permite el uso de técnicas estadísticas más potentes como la regresión logística (Morote *et al.*, 2001; Felicísimo, 2003).

Referencia bibliográfica	Tipo de método	Especies ibéricas	Área de estudio
García Salmerón, 1980	Rangos de tolerancia	6 arbóreas	España peninsular
Allué, 1990	Envolvente ambiental	12	España peninsular
Ruiz de la Torre *et al.*, 1990	Rangos de tolerancia	47 arbóreas, 65 arbustivas, 26 herbáceas	España peninsular
De la Rosa *et al.*, 1992	Rangos de tolerancia	17 arbóreas	España (Región Mediterránea)
Gandullo & Sánchez Palomares, 1994; Gandullo *et al.*, 2004b; Gandullo *et al.*, 2004a; Sánchez Palomares *et al.*, 2007; Sánchez Palomares *et al.*, 2008; Alonso Ponce *et al.*, 2010a	Envolvente ambiental	11 arbóreas	España peninsular
Ruiz de la Torre *et al.*, 1996	Rangos de tolerancia	29 arbóreas, 37 arbustivas, 10 herbáceas	Andalucía oriental
Castejón *et al.*, 1998	Clasificación ecológica	15 arbóreas	España peninsular
García López & Allúe Camacho, 2004	Envolvente ambiental	16 arbóreas	España peninsular
Morote *et al.*, 2001	Regresión logística	3 arbóreas	La Mancha
Felicísimo, 2003	Regresión logística	1 arbórea	Extremadura
Heredia, 2007	Red neuronal	39 arbustivas	España (Región Mediterránea)
Bravo & Montero, 2008	Rangos de tolerancia	25 arbóreas	España

Tabla 1.1. Resumen de los sistemas de apoyo a la selección de especies para la restauración de la vegetación en España

La tabla 1.1 recoge los sistemas de apoyo a la selección de especies para la restauración de la vegetación disponibles en España. A la vista de estos datos se puede

afirmar que las especies arbóreas dominantes de primer orden están bien estudiadas, pero el resto de especies arbóreas, las arbustivas y herbáceas no están tan bien representadas. Si nos centramos en los sistemas basados en análisis cuantitativos de datos obtenidos por muestreo, el sesgo en favor de las especies arbóreas dominantes es aún más claro, únicamente el modelo Sierra2 (Heredia, 2007) considera especies no arbóreas, pero no cubre toda la España peninsular.

1.2. JUSTIFICACIÓN DEL TRABAJO

Una metodología que pudiera generar modelos de distribución de especies a partir de datos disponibles para un numeroso grupo de especies vegetales y usarlos como base para un sistema de apoyo a las decisiones supondría un avance significativo en la selección de especies para la restauración de la vegetación en España. Esta tesis pretende identificar aspectos metodológicos de los modelos de distribución de especies susceptibles de ser mejorados con el objetivo de realizar propuestas metodológicas que aumenten la capacidad predictiva desde el punto de vista de la selección de especies en la restauración de la vegetación. Para ello, se han considerado los tres componentes principales que conforman los modelos de distribución de especies: un modelo ecológico, un modelo de datos y un modelo estadístico (Austin, 2002).

En una situación ideal, el desarrollo del modelo de datos se centraría en cómo realizar el muestreo para obtener los datos y cómo medir las variables de interés. Dado que es improbable que se realice un esfuerzo de muestreo a escala nacional con el objetivo de desarrollar modelos optimizados para la selección de especies, la identificación de aspectos susceptibles de mejora en cuanto al modelo de datos tiene que restringirse a cómo usar los datos disponibles y qué resultados se pueden esperar de los modelos ajustados con dichos datos.

La primera limitación para desarrollar modelos de distribución de especies válidos para toda España es la escasez de datos de buena calidad. La situación ideal es tener registros de presencia y ausencia de las especies en parcelas de pequeño tamaño

elegidas utilizando alguna técnica de muestreo. Este requisito solo lo cumplen las especies consideradas en el Inventario Forestal Nacional (IFN) en el que se registra el número y tamaño de los individuos de un centenar de especies leñosas (la mayoría árboles) en aproximadamente 90.000 parcelas de 0,2 ha. El Mapa Forestal de España (MFE) a escala 1:200.000 (Ruiz de la Torre, 1990) aporta datos de distribución para 267 de taxones leñosos y herbáceos a nivel de especie en aproximadamente 100.000 teselas forestales de tamaño variable y promedio de 250 ha. Aunque los inventarios de vegetación del Mapa Forestal no se pueden considerar exhaustivos, durante su elaboración se visitaron todas las teselas y eso lo convierte en una fuente de datos mucho más completa que las habituales de los estudios corológicos (Gastón & Soriano, 2006). Otra posible fuente de datos de distribución de especies la constituyen las bases de datos que recopilan datos florísticos e inventarios de vegetación como el Proyecto Anthos (www.anthos.es), el Sistema de Información de la Vegetación Ibérica y Macaronésica (www.sivim.es) o el portal español del Global Bodiversity Information Facility (ww.gbif.es). La exhaustividad y resolución espacial de los datos depende de las especie considerada, pero, en general, la fiabilidad de las ausencias es baja y la resolución espacial es de hasta 100 km^2. Esta resolución espacial tan grosera queda muy lejos de la escala de trabajo de un proyecto de restauración de la vegetación, por lo tanto el método de modelización debería permitir transferir modelos ajustados a baja resolución espacial a situaciones de mayor resolución si se pretende modelizar la distribución de especies para las que solamente se cuenta con las recopilaciones corológicas. Ninguno de los métodos citados anteriormente prevé la transferencia de las predicciones de los modelos entre escalas, de manera que es necesaria una nueva metodología de modelización, al menos para las especies que no cuentan con datos de alta resolución espacial.

Uno de los aspectos fundamentales en el establecimiento de un modelo ecológico es la selección de variables independientes. Pero la selección de variables no solo depende del modelo ecológico adoptado, también es decisiva la disponibilidad de información respecto a los factores ecológicos que se pretendan usar como variables independientes de los modelos, es decir, también depende del modelo de datos.

Así como no se disponen de los datos de distribución de especies óptimos, tampoco se cuenta con datos adecuados de todos los factores del medio físico que condicionan la distribución de las especies. Obtener estimaciones de los parámetros climáticos y fisiográficos es relativamente sencillo gracias a los modelos de estimación de parámetros climáticos (Sánchez Palomares *et al.*, 1999) y a los modelos digitales de elevaciones (Farr *et al.*, 2007). Sin embargo es raro contar con parámetros edafológicos medidos en campo y a menudo no son considerados en los modelos de distribución de especies a pesar de su clara influencia en la distribución de las especies vegetales (Coudun *et al.*, 2006). Una manera de incorporar datos edafológicos a los modelos es usar mapas de suelos. No existen mapas de suelo que actualmente cubran toda España de escalas superiores a 1:1.000.000 (Gómez-Miguel, 2007; Van Liedekerke *et al.*, 2006) y esto podría ser un obstáculo a la hora de utilizar su información como variables independientes de los modelos. Es necesario evaluar si la incorporación de datos del suelo provenientes de mapas de baja resolución espacial mejora significativamente los modelos de distribución de especies vegetales ibéricas.

El modelo estadístico es una parte fundamental de los modelos de distribución de especies, ya que determina, entre otras cuestiones, el método matemático usado para construirlos. Cuando solamente se disponen de datos de presencia de la especie y no hay posibilidad de generar pseudo-ausencias (localidades elegidas al azar en las que se supone ausencia de la especie y para las que se conocen los valores de los factores ecológicos considerados) las envolventes ambientales son la única opción. Este es el caso de los trabajos de autoecología paramétrica de especies forestales que utilizan variables del suelo obtenidas de muestras tomadas en campo y que no están disponibles en un mapa del que poder tomar pseudo-ausencias (Gandullo & Sánchez Palomares, 1994; Gandullo *et al.*, 2004b; Gandullo *et al.*, 2004a; Sánchez Palomares *et al.*, 2007; Sánchez Palomares *et al.*, 2008; Alonso Ponce *et al.*, 2010a). Si además de datos de presencia de las especies, se dispone de datos de ausencia (como en el IFN) o pseudo-ausencia (como en el MFE) y las variables independientes se pueden extraer de la cartografía temática, los modelos de regresión logística superan a las envolventes ambientales en cuanto a capacidad predictiva (Elith *et al.*, 2006). La regresión logística ha sido el método más usado en los modelos de distribución de especies de las dos

últimas décadas del siglo XX (Pearce & Ferrier, 2000), pero los resultados del estudio comparativo de métodos más exhaustivo hasta la fecha indicaban que algunos métodos más modernos como los de máxima entropía (Phillips *et al.*, 2006) superan en capacidad predictiva a la regresión logística (Elith *et al.*, 2006). Estos métodos más modernos tienen en común que incluyen técnicas de regularización que combaten el riesgo de sobreajuste cuando hay demasiadas variables independientes en el modelo y la muestra es demasiado pequeña. A pesar de que la regularización se puede aplicar a la regresión logística (Harrell, 2001), los modelos de regresión se han ajustado usando el método de máxima verosimilitud estándar cuando se los ha comparado a los de máxima entropía y otros métodos novedosos (Elith *et al.*, 2006; Gibson *et al.*, 2007; Elith & Graham, 2009; Roura-Pascual *et al.*, 2009; Tognelli *et al.*, 2009; Marini *et al.*, 2010). Si las diferencias observadas en cuanto a capacidad predictiva se debieran a la regularización, la incorporación de la regularización a la regresión logística podría situarla a la altura de los métodos más modernos y se justificaría su uso para la selección de especies en la restauración de la vegetación. Es necesario realizar un estudio comparativo que evalúe la influencia de la incorporación de la regularización a la regresión logística en la capacidad predictiva de los modelos de distribución de especies.

El otro aspecto fundamental del modelo estadístico es la validación. Como cualquier modelo empírico con vocación predictiva, los modelos de distribución de especies deben ser validados con respecto a su capacidad predictiva antes de ser usados en la práctica. Existen numerosos estadísticos para evaluar la capacidad predictiva que se pueden clasificar en dos grandes grupos, los que evalúan la discriminación y los que evalúan la calibración (Pearce & Ferrier, 2000). La discriminación estudia la capacidad de las predicciones de un modelo para discriminar entre presencias y ausencias observadas. Uno de los estadísticos de discriminación más populares es el área bajo la curva característica operativa del receptor (AUC por sus siglas en inglés, Fielding & Bell, 1997), que estima la probabilidad de que una observación de presencia de la especie tomada aleatoriamente obtenga una predicción de idoneidad mayor que una observación de ausencia tomada al azar. La discriminación es importante cuando se necesita que el modelo identifique las mejores

zonas para el desarrollo de una especie, por ejemplo cuando se usan los modelos para localizar nuevas poblaciones de plantas raras en áreas poco estudiadas. En el caso de la selección de especies para la restauración de la vegetación es más importante la calibración, que evalúa la verosimilitud de las probabilidades pronosticadas por el modelo (Harrell, 2001), es decir, si visitamos las *n* localidades en las que un modelo pronostica una probabilidad *p*, esperamos encontrar la especie estudiada en $n \cdot p$ localidades. Con un modelo bien calibrado obtenemos buenas estimaciones de la probabilidad de que una especie sea idónea para las características ecológicas del lugar que se pretende restaurar. La calibración se ha usado mucho menos que la discriminación en la evaluación de modelos de distribución de especies (Vaugham & Ormerod, 2005) y esta circunstancia podría comprometer algunas de las recomendaciones habituales sobre la estrategia de modelización cuando el objetivo sea la selección de especies. Es necesario incorporar los estadísticos de calibración a la evaluación de la capacidad predictiva de los modelos de distribución de especies orientados a la selección de especies para la restauración de la vegetación.

Todas las consideraciones sobre la validación de modelos hechas hasta ahora suponen que las predicciones de un modelo para una especie se están evaluando frente a la distribución conocida de la especie. Este tipo de validación interesa en el marco de la selección de especies cuando se plantean preguntas como ¿cuál es el mejor lugar para reintroducir una especie desaparecida? o ¿cuál es la probabilidad la especie elegida sobreviva en el lugar donde se va a realizar la restauración? Pero es habitual que la pregunta principal a responder en el proyecto de restauración sea más bien ¿cuáles son las tres especies más idóneas para las condiciones ecológicas del lugar a restaurar? Esta pregunta asume que existen *n* modelos correspondientes a *n* especies que ofrecen estimaciones de idoneidad para las características del medio físico en el lugar a restaurar y por lo tanto se puede generar una lista de especies ordenada por idoneidad. Lo que el usuario de los modelos espera es que las especies presentes en lugares similares al que se van a restaurar estén entre las primeras de la lista, o dicho de otra manera, que el conjunto de *n* modelos aplicados a un lugar del que se conoce la flora discriminen bien entre especies presentes y ausentes. Ninguna metodología de evaluación de modelos de distribución de especies responde a esta

pregunta y por lo tanto es necesario desarrollar un método que evalué este aspecto específico de la selección de especies para la restauración de la vegetación.

1.3. OBJETIVOS

El objetivo general es mejorar la capacidad predictiva de los modelos de distribución de especies aplicados a la selección de especies en la restauración de la vegetación, y que éstos sirvan de base para un sistema de apoyo a la toma de decisiones que sea aplicable al mayor número de especies vegetales de la España peninsular con los datos disponibles en la actualidad.

Se pretende avanzar en la consecución del objetivo general a través de los siguientes objetivos específicos que responden a las necesidades expuestas en el apartado de justificación del trabajo:

1. Desarrollar una metodología que permita transferir modelos de distribución de especies entre escalas geográficas para obtener predicciones de alta resolución espacial a partir de modelos de baja resolución espacial.
2. Evaluar el efecto de la incorporación de la regularización en la capacidad predictiva de los modelos de distribución de especies ajustados usando regresión logística y compararlos con los modelos de máxima entropía.
3. Evaluar el efecto de la incorporación de variables de suelo obtenidas de mapas de baja resolución espacial en la capacidad predictiva de los modelos de distribución de especies.
4. Desarrollar un método de validación de modelos de distribución de especies adaptado a la selección de especies para la restauración de la vegetación.

1.4. ESTRUCTURA DEL TRABAJO

Cada uno de los objetivos específicos enunciados en el apartado anterior corresponde a uno de los cuatro capítulos siguientes. Cada capítulo se ha escrito para que pueda ser leído independientemente de los otros y cuenta con apartados dedicados a la metodología, resultados y discusión. Por último, el sexto capítulo consiste en una discusión general de los resultados obtenidos y exposición de las conclusiones finales. Las referencias bibliográficas completas de todos los capítulos se presentan a continuación de la discusión general.

CAPÍTULO 2

TRANSFERENCIA DE MODELOS DE DISTRIBUCIÓN DE ESPECIES ENTRE ESCALAS GEOGRÁFICAS: UN MÉTODO DE ACTUALIZACIÓN DE MODELOS DE BAJA RESOLUCIÓN ESPACIAL USANDO MUESTRAS PEQUEÑAS DE ALTA RESOLUCIÓN ESPACIAL

Los resultados de este capítulo han sido publicados en: GASTÓN A., GARCÍA-VIÑAS J.I., 2010. Updating coarse-scale species distribution models using small fine-scale samples. Ecol. Model. 221 (21): 2576-2581.

2.1. INTRODUCCIÓN

La capacidad predictiva de los modelos es uno de los aspectos que más atención recibe por parte de los expertos en modelización. Diversos factores pueden causar una baja capacidad predictiva: baja calidad de los datos, omisión de variables independientes importantes, problemas con la integración de datos de diferentes escalas y tamaños de muestra pequeños. Este capítulo trata sobre las dos últimas causas de baja capacidad predictiva citadas. A menor tamaño de muestra, menor es la capacidad predictiva de los modelos de distribución de especies y el número de registros de presencia disponibles para cada especie es a menudo limitado (Wisz *et al.*, 2008). Los atlas corológicos de baja resolución espacial son cada día más abundantes, pero las aplicaciones de los modelos a la gestión de los recursos naturales suelen requerir predicciones de alta resolución espacial y los intentos de trasladar predicciones entre escalas han sido raros y los resultados no muy alentadores (McPherson *et al.*, 2006; Barbosa *et al.*, 2010). La situación ideal es contar con muestras de gran tamaño con datos de alta resolución espacial para construir modelos aplicados a la gestión de recursos naturales, pero no se dispone de ese tipo de muestras para la mayoría de las especies y el coste de este tipo de muestreo es generalmente inasequible. Si se dispone de grandes muestras de baja resolución, una solución intermedia podría basarse en modelos de baja resolución que incorporen información de alta resolución obtenible con un esfuerzo de muestreo limitado (McPherson *et al.*, 2006).

La regresión logística se ha usado frecuentemente en modelos de distribución de especies (véase Guisan *et al.*, 2002 para una revisión) y, en general, cuando se necesita predecir una variable binaria, p.ej. la mortalidad a corto plazo en modelos predictivos clínicos (Harrell, 2001). En modelización clínica se han usado métodos de actualización de modelos de regresión logística para ajustar modelos previos a circunstancias locales y/o contemporáneas cuando se dispone de nuevas muestras (Steyerberg *et al.*, 2004). Si se dispone de una muestra de actualización pequeña, los métodos que simplemente recalibran (actualizar los coeficientes y/o la constante del predictor lineal) pueden ser una estrategia razonable para obtener un modelo válido

para las nuevas circunstancias y predicen mejor que mantener los coeficientes en los valores originales o que la reestimación del modelo (Steyerberg *et al.*, 2004).

La recalibración actúa sobre la calibración del modelo (el nivel de coincidencia entre las probabilidades pronosticadas y las frecuencias observadas), pero no se espera que afecte a la discriminación (la habilidad del modelos para ordenar los lugares en función de la idoneidad de la especie), porque no se altera el orden de las predicciones. Si, como en el caso de la selección de especies para la restauración de la vegetación, se necesitan estimaciones fiables de la probabilidad de presencia de las especies, la calibración debería ser considerada antes que otras medidas de la capacidad predictiva de los modelos. Se propone usar la estrategia de actualización de modelos para incorporar información de alta resolución espacial a modelos de distribución de especies de baja resolución espacial.

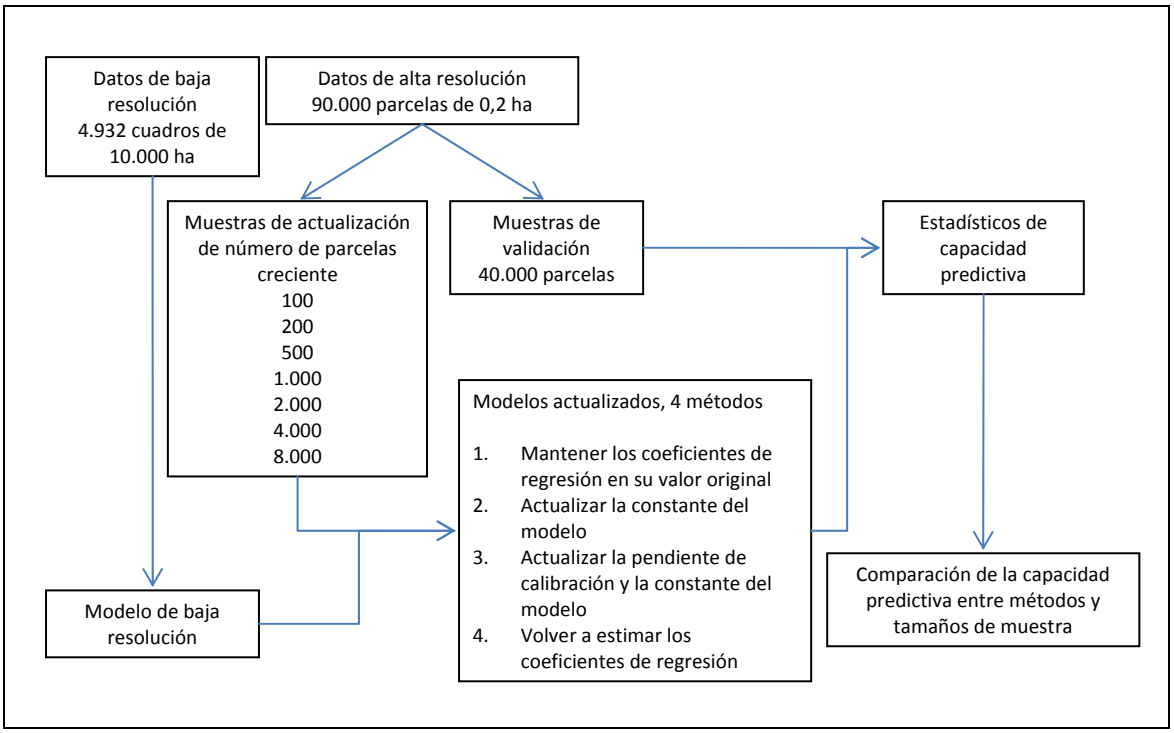
¿Pueden usarse esos métodos de recalibración para actualizar modelos de distribución de especies de baja resolución espacial usando muestras de alta resolución? ¿Cuál es el tamaño mínimo de muestra requerido para este tipo de estrategia de modelización? Se pretende responder a estas preguntas evaluando la capacidad predictiva de modelos de baja resolución recalibrados. También se ha estudiado el efecto del tamaño de la muestra de actualización en la capacidad predictiva de los modelos, ensayando con diferentes tamaños de muestra tomados aleatoriamente de una base de datos con gran número de registros de alta resolución espacial.

2.2. MATERIAL Y MÉTODOS

2.2.1. Marco experimental

Se ajustaron modelos para cinco especies arbóreas españolas usando datos de 10 km de resolución (4.932 cuadros UTM). Se generaron modelos actualizados con siete diferentes tamaños de muestra (entre 100 y 8.000 parcelas tomadas al azar de una base de datos de alta resolución con más de 90.000 parcelas de 0,2 ha). Se usaron

cuatro métodos para generar predicciones a alta resolución: (1) mantener los coeficientes de regresión en su valor original, (2) actualizar la constante del modelo, (3) actualizar la pendiente de calibración y la constante del modelo, (4) volver a estimar los coeficientes de regresión. Los modelos actualizados se evaluaron posteriormente en muestras independientes de 40.000 parcelas de alta resolución tomadas al azar, considerando tanto la calibración como la discriminación (véase cuadro 2.1). El procedimiento se repitió 120 veces para cada una de las 5 especies, generando un total de 600 evaluaciones para cada combinación de método y tamaño de muestra.



Cuadro 2.1. Esquema del diseño experimental. Los procedimientos del esquema se repitieron 120 veces para cada una de las 5 especies consideradas.

2.2.2. Datos de distribución de especies

Se consideraron cinco especies con prevalencias a alta resolución cercanas al 5% y prevalencias a baja resolución variables (véase el porcentaje tras el nombre de las especies): *Castanea sativa* Mill. (18%), *Fagus sylvatica* L. (11%), *Quercus humilis* Mill. (=*Q. pubescens* Willd.) (5%), *Q. suber* L. (20%) and *Pinus pinea* L. (26%).

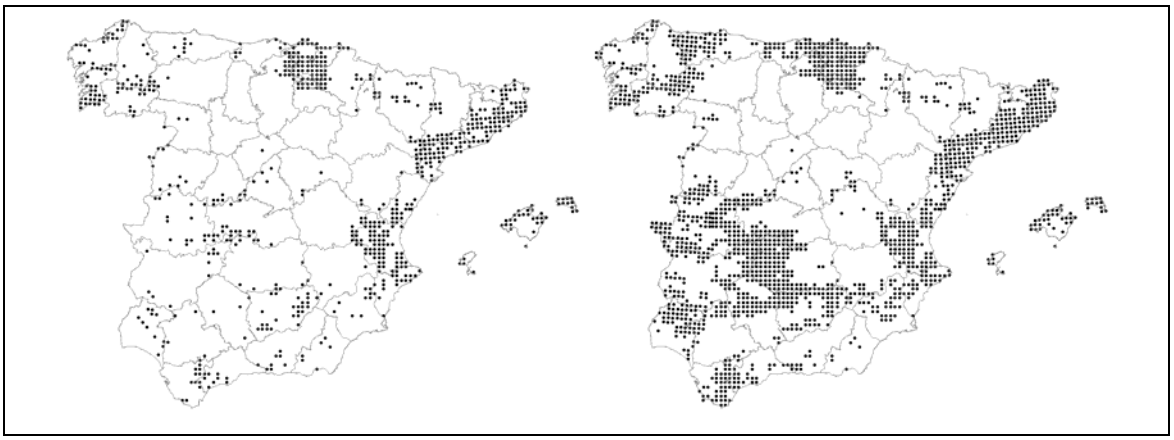


Figura 2.1. Aportación del Mapa Forestal de España (Ruiz de la Torre, 1990) a la corología de *Arbutus unedo*. A la izquierda el mapa de distribución basado en bibliografía botánica y pliegos de herbario, a la derecha, la distribución una vez incorporadas las citas el Mapa Forestal (Gastón & Soriano, 2006).

Los datos de baja resolución (10 km) se obtuvieron del Sistema de Información sobre las Plantas de España (www.anthos.es), que recopila citas de especies de plantas vasculares tomadas de la bibliografía botánica y las colecciones de herbario. Como en la mayoría de los atlas corológicos, los datos recopilados en el proyecto Anthos no responden a una estrategia de muestreo prediseñada y solo se registra la presencia, consecuentemente se desconoce la fiabilidad de las ausencias. Dado que la fiabilidad de las ausencias usadas en los modelos debería ser tan buena como la de las ausencias (Lobo, 2008), los datos de los atlas corológicos pueden ser insuficientes para una modelización precisa. Para evitar este problema, se añadieron los inventarios de plantas leñosas del Mapa Forestal de España (Ruiz de la Torre, 1990) a los datos del proyecto Anthos. Como resultado de un trabajo de campo exhaustivo, la aportación del Mapa Forestal a la corología de plantas leñosas puede doblar su área de distribución conocida (Gastón & Soriano, 2006) y por lo tanto incrementar la fiabilidad de las ausencias (véase figura 2.1). Los datos de presencia-ausencia de alta resolución se tomaron del Tercer Inventario Forestal Nacional (IFN), una malla sistemática con más de 90.000 parcelas de 0,2 hectáreas.

2.2.3. Variables ambientales

Se generaron capas de información climática aplicando un modelo de regresión múltiple basado en datos de estaciones meteorológicas (Sánchez Palomares *et al.*,

1999) a los datos de elevación de 3 arcosegundos (≈ 90 m) de resolución del STRM (Farr *et al.*, 2007). Inicialmente se consideraron 17 variables climáticas habitualmente usadas en estudios autoecológicos de especies arbóreas en España (p.ej., Alonso Ponce *et al.*, 2010b): Precipitaciones medias estacionales (4), precipitación media anual, temperaturas medias estacionales (4), temperatura media anual, temperatura media de las máximas del mes más cálido, temperatura media de la mínimas del mes más frío, duración del periodo árido, intensidad de la aridez, evapotranspiración potencial anual media, superávit hídrico anual medio, déficit hídrico anual medio. La información climática se integró dentro de cada celda de baja resolución (10 x 10 km) simplemente calculando el valor medio (McPherson *et al.*, 2006).

La distribución de los sustratos calcáreos es una variable útil para modelos predictivos de plantas en el área de estudio (véase capítulo 4). La información litológica de 1 km de resolución de la Base de Datos de Suelos Europeos (Van Liedekerke *et al.*, 2006) se integró dentro de cada celda de baja resolución simplemente calculando la proporción de área ocupada por materiales calcáreos.

2.2.4. Modelos de baja resolución

Se usó regresión logística para construir los modelos de baja resolución siguiendo las estrategias recomendadas por Harrell (2001) y Steyerberg (2009) en cuanto a selección de variables candidatas, reducción de número de variables, relaciones no lineales, validación interna y reducción de los coeficientes de regresión. Se usaron las funciones de la librería *Design* (Harrell, 2009) para R (R Development Core Team, 2009).

Se espera que la relación entre la presencia de las especies y las variables ambientales no sea lineal, es más, una proporción significativa de las respuestas puede ser asimétrica (Oksanen & Minchin, 2002). Por lo tanto, la complejidad de la variables explicativas se fijó a priori como splines cúbicas restringidas de 4 nodos (k=4) que permiten respuestas unimodales asimétricas. Dado que se estiman 3 parámetros (k-1) por variable y se necesitan un mínimo de 10 registros de presencia por parámetro

estimado para evitar el sobreajuste (o de ausencia, si son menos frecuentes que las presencias), se requiere una muestra muy grande para modelos con 18 variables independientes y sus correspondientes términos no lineales e interacciones. Para evitar el sobreajuste se usaron técnicas de análisis de conglomerados como estrategia para reducir el número de variables. Se construyeron grupos de variables con análisis jerarquizado de conglomerados aplicados a una matriz de similitud entre variables (coeficientes de correlación de Spearman al cuadrado). Una vez se habían definido los grupos de variables, se calculó el primer Componente Principal como representante de cada grupo.

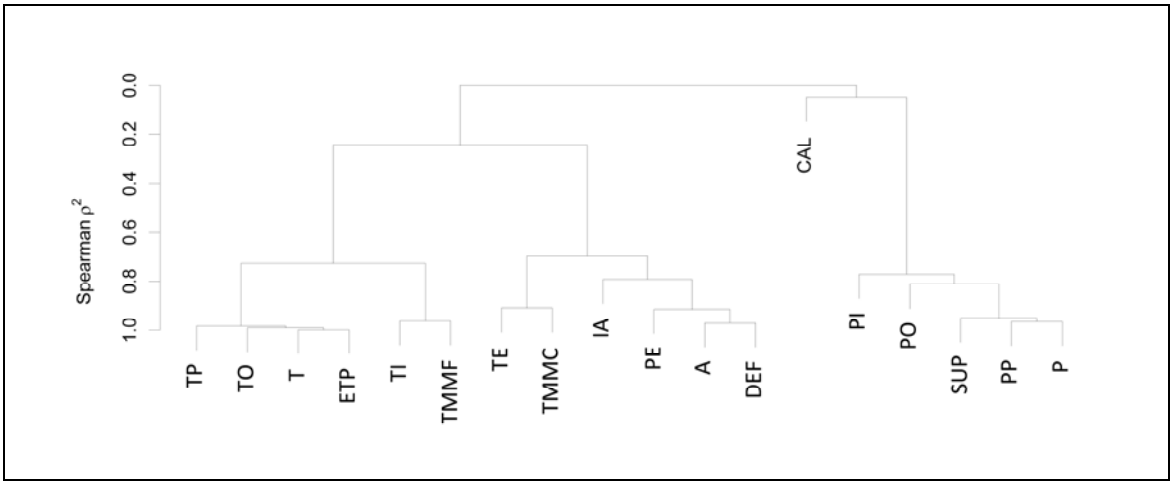


Figura 2.2. Resultado del análisis jerarquizado de conglomerados aplicado a una matriz de similitud entre variables. Abreviaturas: T, temperatura media anual; TI, temperatura media invernal; TP, temperatura media primaveral; TE, temperatura media estival; TO, temperatura media otoñal; P, precipitación anual media; PI, precipitación invernal media; PP, precipitación primaveral media; PE, precipitación estival media; PO, precipitación otoñal media; A, duración media del periodo árido; IA, intensidad media del periodo árido; ETP, evapotraspiración potencial anual media; DEF, déficit hídrico anual medio; SUP, superávit hídrico anual medio; CAL, proporción de superficie ocupada por sustratos calcáreos.

El procedimiento de reducción de variables generó seis grupos (figura 2.2): variables relacionadas con las condiciones térmicas medias (*T, TP, TO, ETP*), con las condiciones térmicas estivales (*TE, TMMC*), con las condiciones térmicas invernales (*TI, TMMF*), con la disponibilidad hídrica durante el periodo árido (*A, IA, PE, DEF*), con las disponibilidad hídrica media (*PI, PO, PP, P, SUP*) y con la proporción superficie ocupada por materiales calcáreos (*CAL*).

Los modelos preliminares, ajustados con el método estándar de máxima verosimilitud, se validaron internamente siguiendo un procedimiento de remuestreo (*bootstrap*). Se tomaron 200 muestras con reemplazo del mismo tamaño que la original. Se ajustó un nuevo modelo con cada una de las 200 muestras siguiendo el procedimiento original. Se estimó el optimismo en los estadísticos que evalúan la capacidad predictiva del modelo como la diferencia media entre los estadísticos calculados con la muestra de *bootsptrap* y los calculados con la original. El optimismo se restó a la capacidad predictiva aparente del modelo original. Los estadísticos usados para evaluar la capacidad predictiva de los modelos fueron: área bajo la curva característica operativa del receptor (AUC), pendiente de calibración e índice U (véase apartado 2.2.7).

Los modelos finales se obtuvieron aplicando un procedimiento de reducción a los coeficientes de regresión de los modelos preliminares. Se calculó un factor de reducción uniforme (s) igual a la pendiente de calibración del modelo una vez descontado el optimismo. Los coeficientes reducidos se calcularon como $s \cdot \beta_i$, donde β_i son los coeficientes de regresión originales.

2.2.5. Muestras de actualización y evaluación

Se tomaron 40.000 parcelas del IFN aleatoriamente para ser usadas como muestra independiente para la evaluación de los modelos actualizados. Se generaron muestras de actualización de 100, 200, 500, 1.000, 2.000, 4.000 y 8.000 parcelas aplicando muestreo estratificado con afijación proporcional al conjunto de parcelas no usadas para evaluación. Los estratos se definieron siguiendo una clasificación biogeoclimática disponible para España (Elena Rosselló, 1997). El mismo procedimiento de muestreo se repitió 120 veces.

2.2.6. Métodos de actualización

La actualización de modelos más allá de una simple recalibración requiere grandes muestras (Steyerberg *et al.*, 2004). Dado que nuestro interés reside en minimizar el tamaño de muestra, descartamos la extensión de modelos (añadir más variables independientes) y nos centramos en métodos de recalibración y reestimación (los cuatro primeros métodos descritos por Steyerberg *et al.*, 2004). La recalibración consiste en actualizar la constante y/o la pendiente de calibración del modelo. Actualizar la constante pretende corregir la diferencia entre la probabilidad media pronosticada y la prevalencia observada. La actualización de la pendiente de calibración intenta forzar a que la pendiente entre las probabilidades pronosticadas y las frecuencias observadas sea igual a la unidad. Se compararon cuatro métodos:

Método 1: mantener los coeficientes de regresión en sus valores originales. El predictor linear Z para el método 1 (Z_1) se calcula como $Z_1 = \alpha_{baja} + \Sigma\, \beta_{i\ baja} \cdot x_i$ donde α_{baja} y $\beta_{i\ baja}$ son la constante y los coeficientes de regresión estimados previamente con los datos de baja resolución y x_i los valores de la variables independientes en la muestra de alta resolución. Este método ofrece una referencia que debería ser mejorada por los métodos de actualización.

Método 2: actualizar la constante ajustando un modelo de regresión logística en la muestra de actualización con la constante α como el único parámetro libre y el predictor lineal del modelo de baja resolución como variable con coeficiente fijo igual a la unidad: $Z_2 = \alpha + Z_1$

Método 3: actualización de la constante α y la pendiente de calibración β_{global} ajustando un modelo de regresión logística en la muestra de actualización con el predictor lineal del modelo de baja resolución como única covariable: $Z_3 = \alpha + \beta_{global} \cdot Z_1$

Método 4: ajuste de un nuevo modelo siguiendo el mismo procedimiento que en el modelo de baja resolución pero usando la muestra de actualización para entrenar

los modelos: $Z_4 = \alpha + \Sigma\ \beta'_i \cdot x_i$, donde β'_i son los coeficientes estimados nuevamente para las covariables consideradas en el modelo de baja resolución. La recalibración (métodos 2 y 3) debería ser descartada si se obtiene mejor capacidad predictiva volviendo a ajustar los modelos (método 4).

2.2.7. Evaluación de la capacidad predictiva

Tanto la discriminación como la calibración pueden ser consideradas en la evaluación de la capacidad predictiva de los modelos (Pearce & Ferrier, 2000). La capacidad discriminativa es lo principal si el objetivo fundamental es ordenar los lugares de acuerdo con la idoneidad para las especies (p.ej. diseño de espacios protegidos o cartografía de lugares adecuados para reintroducir especies). Si lo que se necesita son estimaciones fiables de la probabilidad de presencia (p.ej. selección de especies para la restauración de la vegetación), la calibración (nivel de concordancia entre las probabilidades pronosticadas y observadas) debe considerarse antes de examinar la discriminación (Harrell, 2001). Se usó el área bajo la curva característica operativa del receptor (AUC) para evaluar la capacidad discriminativa (Fielding & Bell, 1997). El AUC varía entre 0,5 para un modelo que no discrimina mejor que el azar y 1 para una discriminación perfecta. Se usaron la pendiente de calibración y el índice U (Harrell *et al.*, 1984) para cuantificar la calibración. La pendiente de calibración es la pendiente de una regresión logística de las observaciones de presencia y ausencia en función de las probabilidades pronosticadas). El índice U es la diferencia en los valores de máxima verosimilitud entre el modelo usado para calcular la pendiente de calibración y un modelo ideal con constante igual a cero y pendiente igual a 1. Los modelos bien calibrados obtendrán pendientes de calibración cercanas a 1 e índices U cercanos a 0.

2.3. RESULTADOS

La capacidad predictiva media de los modelos de baja resolución, una vez corregido el optimismo, fue buena, tanto considerando la discriminación como la calibración (AUC = 0,909, pendiente de calibración = 0,968 y U = 0,0004).

Consecuentemente, se necesitó una ligera reducción de los coeficientes de regresión (0,968 de media).

Dado que el tamaño de muestra efectivo en la regresión logística está determinado por el número de presencias (o ausencias sin son más escasas que las presencias), se estudiaron los números medios de presencias en cada tamaño de muestra de actualización (tabla 2.1). Como se esperaba, el número medio de presencias resultó cercano al 5% de las parcelas muestreadas debido a la prevalencia de las especies estudiadas.

Tamaño de muestra	100	200	500	1000	2000	4000	8000
Número medio de presencias	5	10	25	49	97	193	386

Tabla 2.1. Número medio de presencias para cada tamaño de muestra.

Los resultados medios de los métodos de actualización se muestran en las figuras 2.3, 2.4 y 2.5 para los cuatro métodos y los siete tamaños de muestra. Como era de esperar, el método 1 generó modelos mal calibrados (baja pendiente de calibración y alto índice U, figuras 2.3 y 2.4). La pendiente de calibración no mejoró al actualizar la constante (método 2), pero la fiabilidad del modelo aumento considerablemente (menor índice U). La calibración de los modelos actualizados con el método 3 fue buena para muestras de 200 parcelas o más (10 o más presencias), por el contrario, el método 4 necesitó muestras más grandes (2.000 parcelas, cerca de 100 presencias) para alcanzar una fiabilidad similar.

Los métodos 1, 2 y 3 obtuvieron valores medios de AUC ligeramente superiores a 0,8 (figura 2.5). Las predicciones de los modelos reestimados (método 4) discriminaron peor que las del método de referencia, excepto para la muestras mayores (8.000 parcelas).

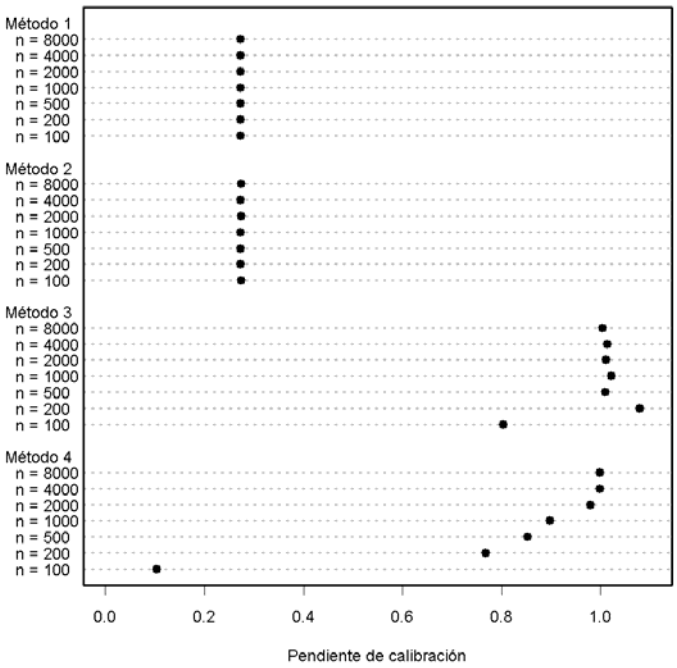


Figura 2.3. Promedio de la pendiente de calibración para los cuatro métodos de actualización y los siete tamaños de muestra considerados.

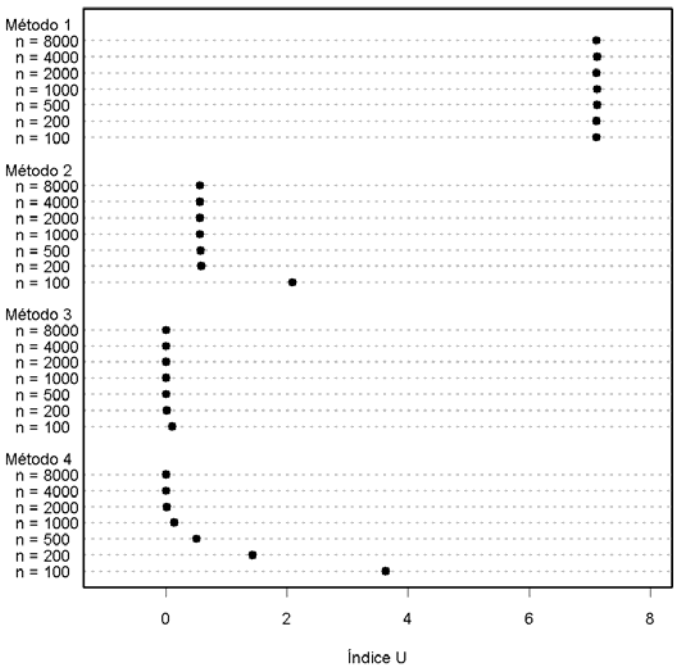


Figura 2.4. Promedio del índice U para los cuatro métodos de actualización y los siete tamaños de muestra considerados.

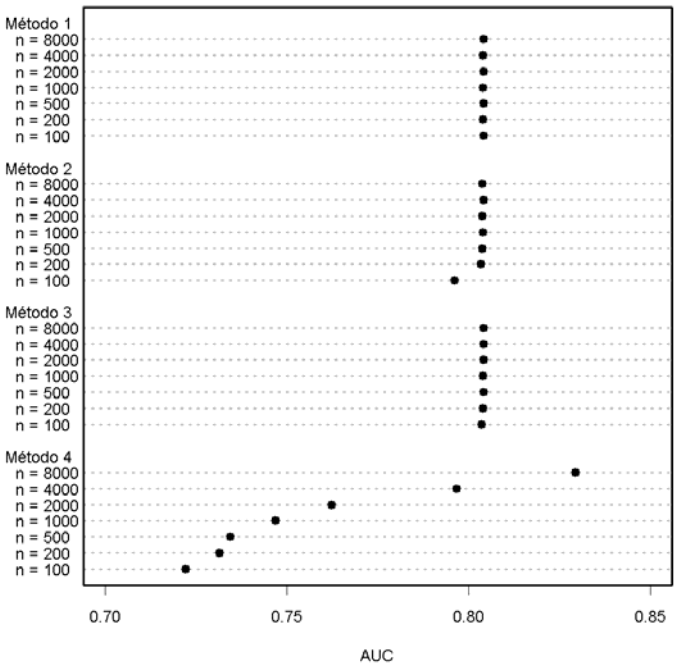


Figura 2.5. Promedio del área bajo la curva característica operativa del receptor para los cuatro métodos de actualización y los siete tamaños de muestra considerados.

2.4. DISCUSIÓN

Se obtuvieron predicciones poco fiables cuando se aplicaron directamente los modelos de baja resolución a datos de alta resolución (pendiente de calibración muy baja e índice U muy alto). La diferencia de resolución entre las muestras de entrenamiento y validación (10.000 y 0,2 hectáreas, respectivamente) puede afectar a la prevalencia y a la importancia relativa de las variables. Estas diferencias causan falta de calibración de los modelos (Pearce & Ferrier, 2000) e impiden aplicar directamente los modelos de baja resolución a datos de alta resolución si la calibración es el objetivo principal. No conocemos ningún estudio sobre la calibración de predicciones entre escalas para modelos de distribución de especies. Más aún, la calibración rara vez ha sido estudiada en modelos de distribución de especies (Vaugham & Ormerod, 2005; pero véase Carroll *et al.*, 1999; Pearce & Ferrier, 2000; Ferrier *et al.*, 2002; Reineking & Schröder, 2006).

La actualización de la constante (método 2) generó modelos con índices U menores que los modelos no actualizados (método 1), pero con pendiente de calibración similar. Este método corrige la calibración general, es decir, hace que la probabilidad pronosticada media sea igual a la prevalencia observada (Steyerberg *et al.*, 2004), pero no afecta a la pendiente de calibración.

Se obtuvieron modelos razonablemente calibrados actualizando la constante y la pendiente de calibración (método 3) con muestras que contienen 10 o más presencias. Por el contrario, para evitar una mala calibración se requirieron mayores muestras (100 o más presencias) al descartar la información derivada de los modelos de baja resolución y ajustar nuevos modelos (método 4). No fue posible mejorar los resultados de los métodos de recalibración con muestras pequeñas, como ya habían mostrado otros estudios (Steyerberg *et al.*, 2004). No conocemos ningún estudio sobre el efecto del tamaño de muestra sobre la calibración en el marco de los modelos de distribución de especies, pero se recomienda un mínimo de 10 presencias por parámetro estimado (sin contar la constante) para la regresión logística en modelos clínicos (Harrell, 2001). Actualizar la constante y la pendiente de calibración únicamente requiere estimar un parámetro. Por lo tanto, los modelos recalibrados son más robustos a la reducción del tamaño de muestra que los modelos reajustados, que habitualmente estiman más parámetros (16 en nuestro caso).

Los modelos de baja resolución discriminaron razonablemente la distribución de las especies a alta resolución (valores de AUC ligeramente superiores a 0,8, Swets, 1988). La transferencia entre escalas de predicciones para el desmán ibérico y la nutria en la Península Ibérica obtuvieron valores similares de discriminación (Barbosa *et al.*, 2010) Por el contario, modelos ajustados a baja resolución discriminaron mal la distribución de grano fino de especies de aves en Uganda (McPherson *et al.*, 2006). Otros intentos de predicciones entre escalas (Lloyd & Palmer, 1998; Collingham *et al.*, 2000; Barbosa *et al.*, 2003; Araújo *et al.*, 2005) usaron estadísticos que dependen de umbrales de probabilidad que clasifican las predicciones en presencias o ausencias, y los resultados no son comparables a los de este estudio.

Como habían mostrado previamente Steyerberg *et al.* (2004), la capacidad discriminativa no se vio afectada por los métodos de recalibración 2 y 3, independientemente del tamaño de muestra. Este resultado es el esperado, ya que la recalibración no altera el orden de las predicciones. La discriminación fue peor para los modelos reajustados (método 4) que para los de referencia (método 1, ni recalibrar, ni reajustar) si la muestra contenía menos de 200 presencias. Como era de esperar, los valores de AUC de los modelos reajustados crecieron con el tamaño de muestra según disminuye el riesgo de sobreajuste (Harrell, 2001). Estudios previos han encontrado tendencias similares en los valores de AUC respecto al tamaño de muestra (Wisz *et al.*, 2008; Cumming, 2000; Reese *et al.*, 2005; Hernandez *et al.*, 2006). Otros estudios del efecto del tamaño de muestra en la capacidad predictiva de los modelos usaron estadísticos que dependen de umbrales de probabilidad (Stockwell, 2002; Kadmon *et al.*, 2003; Pearson *et al.*, 2007) y no son directamente comparable a este estudio.

2.5. CONCLUSIONES

Se han actualizado con éxito modelos de baja resolución espacial con datos de alta resolución usando un método simple de recalibración (actualizar la constante y la pendiente de calibración de modelos de regresión logística). Los modelos actualizados tienen una capacidad predictiva a alta resolución razonable y superan a los modelos reajustados en el caso de muestras pequeñas (10 - 100 presencias). Si se dispone de un modelo de baja resolución (o puede ser generado fácilmente) y se requieren predicciones a alta resolución con un esfuerzo de muestreo limitado, la actualización puede ser mejor opción que ajustar un nuevo modelo.

La actualización de modelos de baja resolución espacial podría ser un enfoque adecuado para especies comunes con distribución bien conocida a baja resolución, pero no suficientemente representadas en los inventarios de alta resolución disponibles (p.ej. la mayoría de las plantas leñosas en España). En estos casos, un esfuerzo de muestreo limitado (p.ej. 200 parcelas para especies con prevalencias superiores a 5%) debería ser suficiente para desarrollar modelos con calibración y discriminación razonable.

Dado que la fiabilidad de las ausencias es alta en los datos de baja resolución usados en este estudio, tal vez no se obtengan los mismos resultados si se usan datos de presencia únicamente. Es muy probable que los modelos ajustados con datos de baja calidad tengan baja capacidad predictiva (Lobo, 2008) y la transferencia de las predicciones a otra escala generará más reducción de la discriminación. Los métodos de recalibración no mejoran la discriminación y se necesitan muestras muy grandes para una actualización más extensiva (Steyerberg *et al.*, 2004). En cuanto a la calibración, es necesario investigar la eficacia de los métodos de actualización para el caso de que los modelos de baja resolución se hayan ajustado con datos de presencia únicamente.

Los buenos resultados obtenidos alientan la investigación de la aplicación de las técnicas de actualización en otras situaciones en las que los modelos de distribución de especies tienen que usarse en condiciones diferentes a las que se usaron para ajustarlos (p.ej., diferentes periodos de tiempo o diferentes regiones).

CAPÍTULO 3

MODELOS DE DISTRIBUCIÓN DE ESPECIES USANDO REGRESIÓN LOGÍSTICA REGULARIZADA: UNA COMPARACIÓN CON MODELOS DE MÁXIMA ENTROPÍA

Los resultados de este capítulo han sido publicados en: GASTÓN A., GARCÍA-VIÑAS J.I., 2011. Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. Ecol.Model., 222(13), 2037-2041.

3.1. INTRODUCCIÓN

Tres componentes principales conforman los modelos de distribución de especies: un modelo ecológico, un modelo de datos y un modelo estadístico (Austin, 2002). Un aspecto importante del modelo estadístico es la elección del método matemático de modelización, ya que una elección subóptima puede causar baja capacidad predictiva. Los expertos en modelización ecológica han mostrado un interés significativo en el efecto del método matemático en la capacidad predictiva de los modelos de distribución de especies (p.ej. Muñoz & Felicísimo, 2004; Segurado & Araújo, 2004). Un grupo de trabajo del National Center for Ecological Analysis and Synthesis (NCEAS) de la Universidad de California ha desarrollado el estudio comparativo de técnicas de modelización más exhaustivo hasta la fecha (Elith *et al.*, 2006). El estudio evaluó la capacidad predictiva de 16 métodos usando datos procedentes de 6 regiones y correspondientes a 226 especies. Los modelos se ajustaron usando datos de presencia/pseudoausencia o solamente presencia y la capacidad predictiva se evaluó con datos de presencia/ausencia. Los resultados demostraron que métodos novedosos como los modelos de máxima entropía (Maxent) ofrecen mejor capacidad predictiva que otros métodos más arraigados como la regresión logística (ajustada tanto con modelos lineares generalizados, GLM, como con modelos aditivos generalizados, GAM). Estudios posteriores también han obtenido mejor capacidad predictiva para Maxent que para la regresión logística (Gibson *et al.*, 2007; Elith & Graham, 2009; Roura-Pascual *et al.*, 2009; Tognelli *et al.*, 2009; Marini *et al.*, 2010).

Investigaciones subsiguientes con los datos del estudio del NCEAS demostraron que la diferencia en capacidad predictiva entre Maxent y regresión logística se reduce con el aumento del tamaño de muestra (Wisz *et al.*, 2008). Estos resultados sugieren que Maxent es menos sensible al sobreajuste y en consecuencia supera a la regresión logística en situaciones de tamaño de muestra reducido. El grupo de trabajo del NCEAS usó una media de 7,5 registros de presencia por cada parámetro estimado en los modelos de regresión logística, un número de registros inferior al mínimo recomendado de 10 presencias por parámetro (Harrell, 2001). Otros estudios

comparativos que incluían Maxent y regresión logística usaron ratios de presencias por parámetro por debajo de 10 (Roura-Pascual *et al.*, 2009; Marini *et al.*, 2010) o ligeramente por encima (Gibson *et al.*, 2007; Elith & Graham, 2009). En estos casos de tamaño muestral reducido, las técnicas de regularización pueden ayudar a evitar los problemas de capacidad predictiva causados por el sobreajuste (Steyerberg *et al.*, 2000). El equipo del NCEAS no usó técnicas de regularización en los modelos de regresión logística y sí lo hizo en Maxent. Esta diferencia podría ser la causa de la diferencia observada entre la capacidad predictiva de esos dos métodos.

Una forma de aplicar la regularización a los modelos de regresión logística es la estimación de máxima verosimilitud penalizada (Harrell, 2001). La regresión penalizada superó a otra técnica de regularización llamado *Lasso* (Tibshirani, 1994) cuando los modelos se ajustaron con muestras pequeñas en una comparación de métodos de regularización aplicados a modelos de distribución de especies (Reineking & Schröder, 2006).

En la regresión logística penalizada se maximiza una versión penalizada de la función de verosimilitud (PML):

$$\text{PML}=\ln L - 0.5\,\lambda \sum (s_i\,\beta_i)^2$$

donde L es la función de verosimilitud convencional, λ es un factor de penalización, β_i son los coeficientes de regresión y es un factor de escala para que $s_i\,\beta_i$ sea adimensional. Este procedimiento de estimación reduce el valor absoluto de los coeficientes de regresión acercándolos a cero, esto causa predicciones sesgadas para la muestra con la que se ajusta el modelo, pero mejora la precisión de las predicciones con datos nuevos. Aunque la penalización no elimina variables independientes del modelo, reduce el número efectivo de parámetros estimados y, por lo tanto, ayuda a evitar los problemas de capacidad predictiva causados por el sobreajuste (Harrell, 2001).

El objetivo de este capítulo es comparar la regresión logística penalizada con Maxent (uno de los mejores métodos según la comparación del NCEAS) y analizar las causas que pudieran explicar las diferencias en la capacidad predictiva entre estos dos

métodos. La hipótesis es que Maxent obtuvo mejores resultados que la regresión logística en la comparación del NCEAS porque el primer método incluía técnicas de regularización y el segundo no. Si la hipótesis es cierta, la regresión penalizada y Maxent debería obtener resultados similares en cuanto a capacidad predictiva. Una hipótesis alternativa podría ser que los métodos generativos (como Maxent) supuestamente obtienen mejores resultados que los métodos discriminativos (como la regresión logística) cuando el tamaño de muestra es pequeño (Phillips & Dudík, 2008). Si la hipótesis alternativa es cierta, Maxent debería obtener mejores resultados que la regresión logística penalizada. Pretendemos contrastar las hipótesis comparando la capacidad predictiva de modelos ajustados con Maxent y regresión logística penalizada para especies arbóreas en España y usando muestras con diferente número de presencias.

3.2. MATERIALES Y MÉTODOS

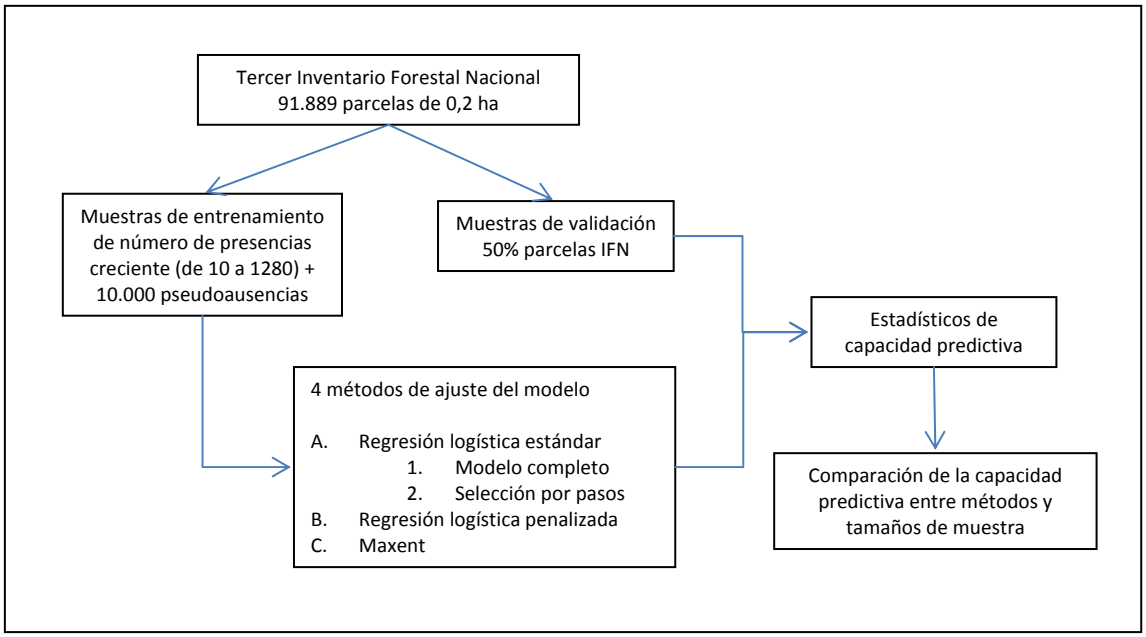
3.2.1 Diseño experimental

Se ajustaron modelos de distribución para 13 especies arbóreas usando datos de datos de presencia/pseudoausencia y evaluando la capacidad predictiva con datos de presencia/ausencia. Se tomaron muestras de ajuste de los modelos de tamaño variable (de 10 a 1280 presencias) para evaluar el efecto del tamaño de muestra en la capacidad predictiva (véase cuadro 3.1). Se evaluaron cuatro estrategias de modelización (véanse los detalles en la sección 3.2.4): dos variantes de la regresión logística estándar (modelos completos y selección de variables por pasos), regresión logística penalizada y Maxent con la configuración por defecto.

3.2.2 Datos de distribución de especies

Las muestras de entrenamiento y evaluación de los modelos se generaron a partir del Tercer Inventario Forestal Nacional de España (IFN). El IFN está formado por una malla regular de parcelas de 0,2 ha, de la que se tomaron las 91.889 peninsulares. Se usaron dos para métodos separar las muestras de entrenamiento y evaluación. El

primer método consistió en la separación aleatoria de las parcelas del IFN en dos muestras del mismo tamaño. Para reducir el efecto de la autocorrelación espacial entre las muestras de entrenamiento y evaluación se usó un segundo método de separación espacial. Para cada especie, se dividió la malla de parcelas a lo largo del meridiano que deja la mitad de las parcelas con presencia de la especie a cada lado (Este u Oeste). La mitad con mayor número de parcelas se eligió como muestra de entrenamiento.



Cuadro 3.1. Esquema del diseño experimental. Los procedimientos del esquema se repitieron para las 13 especies consideradas y aplicando dos métodos de separación de la muestra de validación (véase sección 3.2.2).

En ambos métodos de separación de la muestras se varió el número de presencias de las especies entre 10 y 1280 parcelas, considerando submuestras no anidadas de la muestra de entrenamiento. Se seleccionaron aleatoriamente 10.000 parcelas de la muestra de entrenamiento (incluyendo parcelas con registros de presencia) para ser usadas como pseudoausencias. Las pseudoausencias fueron las mismas para todos los modelos en el caso de la separación aleatoria y variaron entre especies en el caso de separación espacial.

Se ajustaron modelos para las especies arbóreas ibéricas de las familias *Pinaceae* y *Fagaceae*, excluyendo las especies con menos de 1300 registros de presencia en la muestra de entrenamiento para que los modelos de todas las especies tuvieran el mismo rango de tamaños de muestra. Se usaron un total de 13 especies: *Castanea sativa* Miller*, Fagus sylvatica* L.*, Pinus halepensis* Miller*, Pinus nigra* Arnold*, Pinus pinea* L.*, Pinus pinaster* Aiton*, Pinus sylvestris* L.*, Quercus faginea* Lam.*, Quercus ilex* L.*, Quercus humilis* Miller (=*Q. pubescens* Willd.)*, Quercus pyrenaica* Willd.*, Quercus robur* L.*,* y *Quercus suber* L.

3.2.3. Variables ambientales

La configuración por defecto de Maxent está optimizada para modelos con entre 11 y 13 variables independientes (Phillips & Dudík, 2008), por lo que se seleccionaron 11 variables ambientales para la comparación con la regresión logística.

Los datos climáticos se obtuvieron aplicando modelos de estimaciones climáticas (Sánchez Palomares *et al.*, 1999) a un modelo digital de elevaciones de 90 m de resolución (Farr *et al.*, 2007). Los modelos de estimaciones climáticas interpolan datos climáticos mensuales de estaciones meteorológicas usando la latitud, la longitud y la altitud como variables independientes. Se usaron 10 variables climáticas que habitualmente se usan en autoecología de especies arbóreas en España (p.ej. Alonso Ponce *et al.*, 2010b): precipitación estival media, precipitación anual media, temperatura media estival, temperatura media anual, temperatura media de las máximas del mes más cálido, temperatura media de las mínimas del mes más frío, duración del periodo de aridez, evapotranspiración potencial anual media, superávit hídrico anual medio y déficit hídrico anual medio.

Algunas variables climáticas están muy correlacionadas entre sí (coeficiente de correlación de Pearson mayor de 0,8 para 12 de las 45 pares de variables) y en estas circunstancias de colinealidad podría ser recomendable una reducción de variables previa al ajuste de los modelos. Aún así, se mantuvieron todas las variables para reproducir las condiciones para las que se ha optimizado la configuración por defecto

de Maxent, es decir, entre 11 y 13 variables independientes y alta colinealidad (Elith *et al.*, 2006, Tabla 3; véase Phillips & Dudík, 2008).La colinealidad puede causar la sobreestimación de los errores típicos de los coeficientes de regresión, pero no afecta a las predicciones hechas sobre datos que tengan el mismo grado de colinealidad que la muestra de entrenamiento, siempre que no intente una extrapolación extrema (Harrell, 2001). Este estudio se centra en las predicciones de los modelos y no en contrastar hipótesis sobre el efecto de las variables independientes y el grado de colinealidad de las muestras de entrenamiento y evaluación es prácticamente el mismo (los valores de la matriz de correlación no variaron más de 0,019 entre las muestras de evaluación y entrenamiento).

La distribución de los sustratos calcáreos es una variable útil para modelos predictivos de plantas en el área de estudio (véase capítulo 4). Se usó la Base de Datos de Suelos Europeos (Van Liedekerke *et al.*, 2006) para asignar cada parcela a un tipo de sustrato (calcáreo o silíceo).

3.2.4. Estrategias de modelización

Los modelos de máxima entropía se ajustaron usando la versión 3.3.2 de Maxent (Phillips *et al.*, 2006) con la configuración por defecto. La configuración por defecto de Maxent consiste en un conjunto valores de los parámetros del modelo obtenidos a partir de un proceso empírico de puesta a punto usando los datos del estudio del NCEAS (Elith *et al.*, 2006). El procedimiento de puesta punto consiste en la elección de los valores del parámetro de regularización y del tipo de respuestas (*feature classes*) en función del número de registros de presencias disponibles en la muestra (Phillips & Dudík, 2008). Maxent permite la modelización de respuestas complejas de las especies a los factores ambientales combinando diferentes tipos de respuestas (*feature classes*). Por defecto se usan todos los tipos de respuesta (incluyendo interacciones entre variables) cuando hay al menos 80 presencias en la muestra; entre 15 y 79 presencias se usan las respuestas lineales, cuadráticas y de bisagra; entre 10 y 14 presencias las lineales y cuadráticas, y únicamente las lineales por debajo de 10 presencias.

Se usaron tres tipos de regresión logística para evaluar el efecto de la regularización (véase tabla 3.1). La regresión logística estándar sin selección de variables se incluyó como referencia sobre la que la penalización debería ofrecer alguna mejora. Se consideró además la regresión logística estándar con selección de variables por pasos para reproducir los modelos lineales generalizados (GLM) usados en la comparación del NCEAS. Los modelos de regresión logística penalizada (Harrell, 2001) se ajustaron sin interacciones para muestras de menos de 80 presencias y con interacciones para el resto, es decir, usando el criterio de Maxent para incluir las interacciones.

Estrategia de modelización	**Regularización**	**Términos no lineales**	**Selección de variables**	**Interacciones entre variables**
Regresión logística estándar (modelo completo)	No	Splines cúbicos restringidos (4 nodos)	No	No
Regresión logística estándar (selección de variables por pasos)	No	Polinomios cúbicos	Por pasos	No
Regresión logística penalizada	Si	Splines cúbicos restringidos (4 nodos)	No	Si, en el caso de 80 o más presencias
Maxent	Si	Respuestas tipo umbral, bisagra, lineal y cuadrática	No	Si, en el caso de 80 o más presencias

Tabla 3.1. Características de las estrategias de modelización evaluadas.

Se espera que la relación entre la presencia de las especies y las variables ambientales no sea lineal, es más, una proporción significativa de las respuestas puede ser asimétrica (Oksanen & Minchin, 2002). Por lo tanto, la complejidad de la variables independientes se fijó a priori usando splines cúbicas restringidas (Harrell, 2001) de cuatro nodos, excepto para la regresión con selección de variables, para la que se añadieron los términos no lineales usando polinomios cúbicos para reproducir los modelos GLM de la comparación el NCEAS. Ambas formas de añadir términos no

lineales permiten respuestas desde lineales hasta unimodales asimétricas y requieren la estimación de tres parámetros.

Los modelos con selección de variables por pasos se ajustaron comenzando por el modelo completo, buscando en ambas direcciones y usando el índice de información de Akaike (AIC) como criterio de finalización de la búsqueda. Los pasos permitidos para variables continuas fueron los siguientes (en orden): polinomios cúbicos, funciones lineales y omitidas.

Los modelos penalizados se ajustaron usando la desviación típica de cada variable como factor de escala y usando una versión modificada del AIC para seleccionar el factor de penalización óptimo (Harrell, 2001). El factor de penalización óptimo (λ) es el que maximiza el valor del AIC modificado, elegido de entre una serie de valores de 0 a 10 con paso de 0,25. Los valores óptimos de λ variaron entre 0 y 9,25 y fueron menores de 2,75 en el 90% de los modelos penalizados. Se usaron las funciones *lrm* y *pentrace* de la librería *Design* (Harrell, 2009) del entorno de cálculo estadístico R (R Development Core Team, 2009) para ajustar los modelos de regresión logística, excepto los modelos con selección de variables para los que se usaron las funciones *glm* y *step* de la librería *stats* de R.

3.2.5. Evaluación de la capacidad predictiva

Tanto la discriminación como la calibración pueden ser consideradas en la evaluación de la capacidad predictiva de los modelos (Pearce & Ferrier, 2000). La capacidad discriminatoria es lo principal si el objetivo fundamental es ordenar los lugares de acuerdo con la idoneidad para las especies (p.ej. diseño de espacios protegidos o cartografía de lugares adecuados para reintroducir especies). Si lo que se necesita son estimaciones fiables de la probabilidad de presencia (p.ej. selección de especies para la restauración de la vegetación), la calibración (nivel de concordancia entre las probabilidades pronosticadas y observadas) debe considerarse antes de examinar la discriminación (Harrell, 2001). Se usó el área bajo la curva característica operativa del receptor (AUC) para evaluar la capacidad discriminatoria (Fielding & Bell,

1997). El AUC varía entre 0,5 para un modelo que no discrimina mejor que el azar y 1 para una discriminación perfecta. Se han usado la pendiente de calibración para cuantificar la calibración. La pendiente de calibración es la pendiente de una regresión logística de las observaciones de presencia y ausencia en función de las probabilidades pronosticadas. Los modelos bien calibrados obtendrán pendientes de calibración cercanas a 1.

Los valores de AUC y pendiente de calibración obtenidos en el proceso de evaluación de los modelos se analizaron usando modelos mixtos lineales, usando el estadístico de capacidad predictiva como variable dependiente, la estrategia de modelización , el tamaño de muestra y su interacción como factores fijos, y la especie como factor aleatorio. La estrategia de modelización es una variable cualitativa y el tamaño de muestra es una variable cuantitativa, de modo que el modelo estima una constante y una pendiente de regresión entre la capacidad predictiva y el tamaño de muestra para cada estrategia de modelización. El tamaño de muestra medido como el número de presencias de la especie en la muestra de entrenamiento se transformó usando el logaritmo en base 10 para inducir linealidad. Los análisis se realizaron usando la librería *nlme* de R (Pinheiro *et al.*, 2011).

3.3. RESULTADOS

Se presentan por separado los resultados de los dos métodos de separación de muestras (aleatorio y espacial, véase sección 3.2.2).

3.3.1. Separación aleatoria

Los efectos de la estrategia de modelización, el tamaño de muestra y su interacción en la capacidad discriminativa fueron significativos (véase tabla 3.2). Los modelos de Maxent superaron a los de regresión logística estándar (con o sin selección de variables) y la diferencia fue menor cuanto mayor el tamaño de muestra (véase figura 3.1). El modelo mixto confirmo la interpretación gráfica: la constante de Maxent fue mayor que las regresiones no regularizadas (0,83 frente a 0,80, $p < 0,001$) y se

observó lo contario para las pendientes (0,03 frente a 0,04, p<0,001, véase tabla 3.3). Los modelos de Maxent y los de regresión logística penalizada obtuvieron valores similares de AUC (véase figura 3.1). No se observaron diferencias significativas entre las constantes de los modelos de Maxent y los de regresión logística penalizada (p=0,60) y tampoco entre las pendientes (p=0,30). La selección de variables por pasos no supuso una mejora de la discriminación en comparación con los modelos completos (véase figura 3.1). Las diferencias entre estas dos estrategias no fueron significativas según el modelo mixto (p=0,73 para las constantes y p=0,74 para las pendientes).

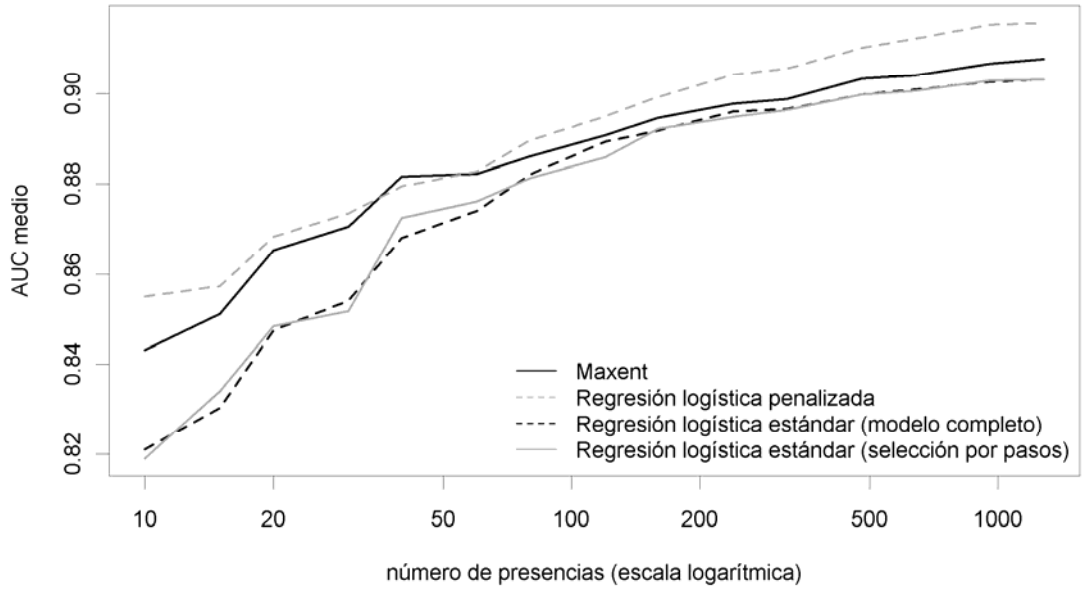


Figura 3.1. Resumen de los resultados de la evaluación de la capacidad discriminativa usando la separación aleatoria de muestras. Valores medios de AUC para cada combinación de estrategia de modelización y tamaño de muestra medido como el número de presencias de la especie en la muestra.

Los efectos de la estrategia de modelización, el tamaño de muestra y su interacción en la pendiente de calibración fueron significativos (véase tabla 3.2). La comparación de la calibración de las diferentes estrategias de modelización ofrecieron resultados similares a los ya descritos para la discriminación (véase tabla 3.3). No se observaron diferencias significativas entre la calibración de los modelos de Maxent y los de regresión logística penalizada. Los métodos regularizados superaron a los no regularizados y la diferencia fue menor cuanto mayor el tamaño de muestra. La mayor

diferencia respecto al análisis del AUC fue que los modelos de regresión logística con selección de variables por pasos obtuvieron mejores pendientes de calibración que los modelos completos.

			Valores de F			
			Separación aleatoria		Separación espacial	
	g.l. num.[1]	g.l. den.[2]	Discriminación (AUC)[3]	Pendiente calibración	Discriminación (AUC)[3]	Pendiente calibración
Constante	1	579	2754,73 ***	1135,15 ***	1068,09 ***	212,93 ***
Estrategia de model.	3	579	82,86 ***	247,90 ***	22,09 ***	195,38 ***
Tamaño muestra	1	181	578,75 ***	301,69 ***	207,73 ***	209,51 ***
Interacción	3	579	18,22 ***	194,04 ***	5,86 ***	143,87 ***

Tabla 3.2. Análisis de desviación de los modelos mixtos para ambas medidas de capacidad predictiva y métodos de separación de muestras (una columna para uno de los cuatro modelos mixtos). Códigos de significación de los valores de p: *** <0,001, ** 0,001-0,01, * 0,01-0,05. [1] Grados de libertad del numerador. [2] Grados de libertad del denominador. [3] AUC: área bajo la curva característica operativa del receptor.

3.3.2. Separación espacial

La separación espacial de las muestras generó valores de discriminación peores que los de la separación aleatoria (una diferencia media en el AUC de 0,051±0,006) y peor calibración (una separación más clara respecto a la pendiente de calibración ideal). Los resultados del análisis con el modelo mixto fueron similares a los de la separación aleatoria (véanse tablas 3.2 y 3.3). Los efectos de la estrategia de modelización, el tamaño de muestra y su interacción fueron significativos, tanto para la discriminación como para la calibración. Los métodos regularizados superaron a los no regularizados y la diferencia fue menor cuanto mayor el tamaño de muestra. La selección de variables por pasos en la regresión no mejoró la discriminación frente a los modelos completos, pero si lo hizo para la calibración.

		Separación aleatoria		Separación espacial	
		Discriminación (AUC)[1]	Pendiente calibración	Discriminación (AUC)[1]	Pendiente calibración
Constantes	Maxent	0,8268 ***	1,0770 ***	0,7839 ***	0,7900 ***
	Maxent - Reg. Log. Pen.	0,0018	-0,0640	0,0060	0,0033
	Maxent - Reg. Log. Std.	-0,0279 ***	-1,6038 ***	-0,0237 ***	-1,2466 ***
	Maxent - Reg. Log. Pasos	-0,0267 ***	-0,9500 ***	-0,0135 *	-0,6603 ***
Pendientes	Maxent	0,0285 ***	0,0635 **	0,0257 ***	0,0755 ***
	Maxent - Reg. Log. Pen.	0,0017	0,0118	-0,0030	-0,0148
	Maxent - Reg. Log. Std.	0,0094 ***	0,5657 ***	0,0070 **	0,4364 ***
	Maxent - Reg. Log. Pasos	0,0089 ***	0,3401 ***	0,0019	0,2213 ***

Tabla 3.3. Resumen de los efectos fijos de los modelos mixtos. Se presentan los coeficientes de ambas medidas de capacidad predictiva y métodos de separación de muestras (una columna para uno de los cuatro modelos mixtos). Se usó el contraste por defecto en R: se calcula el coeficiente del primer nivel del factor, Maxent en este caso, después cada nivel se compara con el primero y se contrasta si las diferencias son diferentes de cero. Códigos de significación de los valores de p: *** <0,001, ** 0,001-0,01, * 0,01-0,05. [1] AUC: área bajo la curva característica operativa del receptor.

3.4. DISCUSIÓN

Los resultados de este estudio concuerdan con otros trabajos previos que informan de una mayor capacidad discriminativa de los modelos de Maxent comparada a la de los modelos de regresión logística ajustados con máxima verosimilitud estándar (Elith *et al.*, 2006; Gibson *et al.*, 2007; Elith & Graham, 2009; Roura-Pascual *et al.*, 2009; Tognelli *et al.*, 2009; Marini *et al.*, 2010) y de una reducción de la diferencia con el aumento del tamaño de muestra (Wisz *et al.*, 2008). Este comportamiento apunta al sobreajuste como posible causa de la peor capacidad

discriminativa de los modelos de regresión logística estándar y por lo tanto la diferencia decrece cuando hay más registros de presencia disponibles para el ajuste de los modelos. Es más, la diferencia entre Maxent y regresión logística desapareció al incorporar la regularización a través de la estimación penalizada de la máxima verosimilitud, todo ello considerando tanto la calibración como la discriminación y usando diferentes estrategias de muestreo.

La selección de variables por pasos no mejoró la capacidad predictiva respecto a los modelos completos, pero obtuvo mejor calibración. Esto es consistente con estudios otros estudios previos que comparaban la capacidad predictiva de la selección de variables por pasos frente a los modelos completos en el contexto de la regresión logística (Steyerberg *et al.*, 2000). Esos estudios mostraron que la selección de variables por pasos no mejora la capacidad discriminativa respecto a los modelos completos a no ser que la muestra tenga más de 50 presencias por parámetro estimado (hay que tener en cuenta que se deben contabilizar el número de variables candidatas y no solo las del modelo final, Harrell, 2001). En nuestro caso el número de presencias por parámetro varió entre 0,3 y 41,3, por lo tanto es razonable que no se encontraran diferencias significativas entre las dos estrategias de modelización.

Los resultados apoyan la hipótesis de que Maxent superó a la regresión logística en la comparación del NCEAS (Elith *et al.*, 2006) debido a que el primer método incluyó regularización y el segundo no. En el presente estudio la regresión logística penalizada generó predicciones tan precisas como las de Maxent. Este resultado no apoya la hipótesis de que los métodos generativos (como Maxent) obtienen mejores resultados que los métodos discriminativos (como la regresión logística) cuando el tamaño de muestra es pequeño (Phillips & Dudík, 2008).

3.5. CONCLUSIONES

La regresión logística penalizada puede considerarse uno de los mejores métodos para desarrollar modelos de distribución de especies usando datos de

presencia/pseudoausencia, comparable a Maxent (uno de los mejores métodos según el estudio comparativo del NCEAS).

Los resultados alientan a un uso más frecuente de la regresión logística penalizada para la modelización de la distribución de las especies, especialmente en aquellos casos en los se requiere ajustar un modelo complejo con un tamaño de muestra limitado.

CAPÍTULO 4

EVALUACIÓN DEL EFECTO DE LA INCORPORACIÓN DE VARIABLES DE SUELO OBTENIDAS DE MAPAS DE BAJA RESOLUCIÓN ESPACIAL EN LA CAPACIDAD PREDICTIVA DE LOS MODELOS DE DISTRIBUCIÓN DE ESPECIES

.

4.1. INTRODUCCIÓN

Se asume que el clima es el factor ecológico más determinante en la distribución geográfica de las especies vegetales y que la importancia del clima aumenta según desciende la resolución espacial de los datos (Thuiller *et al.*, 2004). En consecuencia, con frecuencia los modelos de distribución de especies solamente consideran factores climáticos (Coudun *et al.*, 2006).

Algunos estudios recientes han detectado mejoras de la capacidad predictiva tras incorporar variables relacionadas con el suelo a modelos de distribución de especies de plantas (Coudun *et al.*, 2006; Coudun & Gégout, 2007), comunidades vegetales (Marage & Gégout, 2009) e insectos (Titeux *et al.*, 2009). Lo recomendable es incorporar la información sobre el suelo a través de mediciones directas de sus propiedades físicas y químicas (Austin, 2002). Desafortunadamente, no es habitual que las bases de datos de distribución de especies cuenten con datos de mediciones directas asociadas debido al alto coste de la toma de datos en campo y el posterior análisis en el laboratorio. Una alternativa a las mediciones directas es la información contenida en los mapas de suelos. Los mapas de suelo que cubren grandes extensiones tienen baja resolución espacial (p.ej. 1:1.000.000 para España, Gómez-Miguel, 2007; o Eurasia, Van Liedekerke *et al.*, 2006) y se han usado con éxito para modelos ajustados con datos de distribución de especies de baja resolución espacial (Titeux *et al.*, 2009; Gastón *et al.*, 2009). No conocemos ninguna evaluación previa del efecto de la incorporación de variables provenientes de mapas de suelo de baja resolución a modelos ajustados usando datos de distribución de mayor resolución. Si la diferencia de escalas introduce demasiado error y la capacidad predictiva de los modelos de alta resolución no va a mejorar, no es recomendable introducir las variables de suelo en el modelo, ya que al aumentar el número de parámetros a estimar aumenta el riesgo de sobreajuste, sobre todo cuando el tamaño de muestra es limitado. Este estudio pretende evaluar la aportación de un mapa de suelos de baja resolución espacial (1:1.000.000, Van Liedekerke *et al.*, 2006) a modelos de distribución de especies de vegetales ajustados con datos de distribución de especies de mayor resolución (1:200.000, Ruiz de la Torre, 1990).

4.2 MATERIAL Y MÉTODOS

4.2.1. Datos de distribución de especies

Se ajustaron modelos para 188 especies arbóreas, arbustivas y herbáceas a partir de los datos de distribución de dichas especies en las 120.938 teselas peninsulares del Mapa Forestal de España a 1:200.000 (Ruiz de la Torre, 1990). Cuando la variable dependiente es binaria el tamaño de muestra efectivo no es el número total de observaciones en la muestra sino el número de casos positivos o negativos, el que sea menor. El número de presencias por especie osciló entre 15 y 42.720, con un 90% de la especies entre 28 y 8.735 presencias.

4.2.2. Variables ambientales

Se generaron capas de información climática aplicando un modelo de regresión múltiple basado en datos de estaciones meteorológicas (Sánchez Palomares *et al.*, 1999) a los datos de elevación de 3 arcosegundos (≈ 90 m) de resolución del STRM (Farr *et al.*, 2007). Inicialmente se consideraron 17 variables climáticas habitualmente usadas en estudios autoecológicos de especies arbóreas en España (p.ej. Alonso Ponce *et al.*, 2010b): Precipitaciones medias estacionales (4), precipitación media anual, temperaturas medias estacionales (4), temperatura media anual, temperatura media de las máximas del mes más cálido, temperatura media de la mínimas del mes más frío, duración del periodo árido, intensidad de la aridez, evapotranspiración potencial anual media, superávit hídrico anual medio, déficit hídrico anual medio. La información climática se asoció a cada tesela tomado el valor correspondiente al centroide del polígono.

Se usaron técnicas de análisis de conglomerados como estrategia para reducir el número de variables (Harrell, 2001). Se construyeron grupos de variables con análisis jerarquizado de conglomerados aplicados a una matriz de similitud entre variables (coeficientes de correlación de Spearman al cuadrado). Una vez se habían

definido los grupos de variables, se calculó el primer Componente Principal como representante de cada grupo. El procedimiento de reducción de variables generó seis grupos (véase la figura 2.1 del capítulo 2): variables relacionadas con las condiciones térmicas medias (*T, TP, TO, ETP*), con las condiciones térmicas estivales (*TE, TMMC*), con las condiciones térmicas invernales (*TI, TMMF*), con la disponibilidad hídrica durante el periodo árido (*A, IA, PE, DEF*) y con las disponibilidad hídrica media (*PI, PO, PP, P, SUP*).

Se usó la Base de Datos de Suelos Europeos (ESDB) que incluye una mapa a escala 1:1.000.000 (Van Liedekerke *et al.*, 2006) como fuente de datos de suelo. Se extrajeron tres variables del mapa de suelo: la naturaleza calcárea del sustrato (variable binaria), la presencia de yeso (variable binaria) y el grupo de suelo según la leyenda del mapa de suelos de la FAO (variable cualitativa con 14 niveles en el área de estudio).

Se ajustaron tres modelos de complejidad creciente para cada especie (véase cuadro 4.1): el primero con variables independientes exclusivamente climáticas (15 parámetros, véase sección 4.2.3), el segundo añadiendo las dos variables litológicas (17 parámetros) y el tercero añadiendo al segundo el grupo de suelo (30 parámetros).

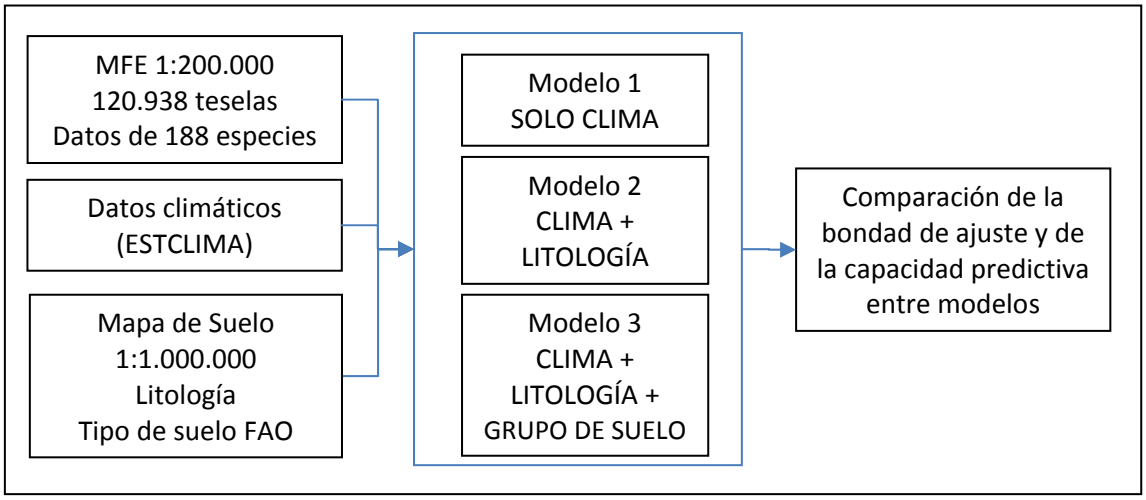
4.2.3. Estrategia de modelización

Se optó por una técnica de regularización aplicada a la regresión logística como alternativa a la selección por pasos de variables. Una forma de aplicar regularización a la regresión logística es usar la estimación de parámetros por máxima verosimilitud penalizada (Harrell, 2001; véanse aplicaciones a los modelos de distribución de especies en Reineking & Schröder, 2006 y en el capítulo 3). En la regresión logística penalizada se maximiza una versión penalizada de la función de verosimilitud (PML):

$$\mathrm{PML}=\ln L - 0.5\,\lambda \sum (s_i\,\beta_i)^2$$

donde L es la función de verosimilitud convencional, λ es un factor de penalización, β_i son los coeficientes de regresión y es un factor de escala para que $s_i\,\beta_i$ sea adimensional. Este procedimiento de estimación reduce el valor absoluto de los

coeficientes de regresión acercándolos a cero, esto causa predicciones sesgadas para la muestra con la que se ajusta el modelo, pero mejora la precisión de las predicciones con datos nuevos. Aunque la penalización no elimina variables independientes del modelo, reduce el número efectivo de parámetros estimados y, por lo tanto, ayuda a evitar los problemas de capacidad predictiva causados por el sobreajuste



Cuadro 4.1. Esquema del diseño experimental. Los procedimientos del esquema se aplicaron a cada una de las 188 especies consideradas.

Los modelos penalizados se ajustaron usando la desviación típica de cada variable como factor de escala y usando una versión modificada del índice de información de Akaike (AIC) para seleccionar el factor de penalización óptimo (Harrell, 2001). El factor de penalización óptimo (λ) es el que maximiza el valor del AIC modificado, elegido de entre una serie de valores de 0 a 10 con paso de 0,25. Se usaron las funciones *lrm* y *pentrace* de la librería *Design* (Harrell, 2009) del entorno de cálculo estadístico R (R Development Core Team, 2009).

Se espera que la relación entre la presencia de las especies y las variables ambientales no sea lineal, es más, una proporción significativa de las respuestas puede ser asimétrica (Oksanen & Minchin, 2002). Por lo tanto, la complejidad de la variables independientes cuantitativas se fijó a priori usando splines cúbicas restringidas (Harrell, 2001) de cuatro nodos. Esta forma de añadir términos no lineales permite

respuestas desde lineales hasta unimodales asimétricas y requieren la estimación de tres parámetros por variable.

4.2.4. Identificación del mejor modelo

La identificación del mejor modelo se acometió de dos formas diferentes. En primer lugar se comparó la bondad de ajuste de los modelos ajustados con los diferentes conjuntos de variables (solo clima; clima y litología; clima, litología y tipo de suelo). En segundo lugar se comparó la capacidad predictiva.

La bondad de ajuste se evaluó usando el índice de información de Akaike (AIC) que es adecuado para comparar modelos de diferente complejidad previamente especificados (Harrell, 2001). Cualquier modelo ganará en bondad de ajuste al añadir más variables debido al sobreajuste. El AIC calcula la bondad de ajuste penalizando la complejidad del modelo, de manera que para que un modelo más complejo obtenga un AIC mejor que otro más sencillo, el aumento en la bondad de ajuste tiene que compensar el aumento en complejidad.

En el caso de la evaluación de la capacidad predictiva también existe riesgo de extraer conclusiones inadecuadas si no se tiene en cuenta el sobreajuste. Un modelo más complejo obtendrá mejores evaluaciones de la capacidad predictiva que uno más sencillo si se ajustan y evalúan usando la misma muestra. Una forma de evitar esto es evaluar los modelos con muestras independientes. Si, como en este caso, no se dispone de muestras independientes, una alternativa es la evaluación interna de los modelos usando una técnica de remuestreo llamada *bootstrap* (Steyerberg *et al.*, 2001). El procedimiento de validación interna consiste en tomar una muestra de *n* parcelas con reemplazo sobre la muestra original, siendo *n* el número total de observaciones en la muestra original. Se ajustó un modelo usando la nueva muestra y el mismo método que en el modelo original. El modelo resultante se usó para obtener predicciones de la probabilidad de presencia tanto en la nueva muestra como en la original. Se calcularon los estadísticos de capacidad predictiva para ambas muestras y se obtuvo la diferencia entre ambos. Se denomina optimismo a esa diferencia, ya que los valores son mayores para la nueva muestra porque el modelo se ajustó con ella y el

sobreajuste hace que al aplicarlo a otra muestra la capacidad predictiva empeore. Se repitió el procedimiento 150 veces y se obtuvo el optimismo medio. Por último, se ajustó el modelo con la muestra original, se calcularon los estadísticos con la misma muestra y se le restaron los optimismos correspondientes para obtener una versión corregida de los mismos. Esta estrategia ha resultado más eficaz para la regresión logística que otras opciones de validación interna (Steyerberg *et al.*, 2001).

Tanto la discriminación como la calibración pueden ser consideradas en la evaluación de la capacidad predictiva de los modelos (Pearce & Ferrier, 2000). La capacidad discriminativa es lo principal si el objetivo fundamental es ordenar los lugares de acuerdo con la idoneidad para las especies (p.ej. diseño de espacios protegidos o cartografía de lugares adecuados para reintroducir especies). Si lo que se necesita son estimaciones fiables de la probabilidad de presencia (p.ej. selección de especies para la restauración de la vegetación), la calibración (nivel de concordancia entre las probabilidades pronosticadas y observadas) debe considerarse antes de examinar la discriminación (Harrell, 2001). Se usó el área bajo la curva característica operativa del receptor (AUC) para evaluar la capacidad discriminativa (Fielding & Bell, 1997). El AUC varía entre 0,5 para un modelo que no discrimina mejor que el azar y 1 para una discriminación perfecta. Se han usado la pendiente de calibración para cuantificar la calibración. La pendiente de calibración es la pendiente de una regresión logística de las observaciones de presencia y ausencia en función de las probabilidades pronosticadas. Los modelos bien calibrados obtendrán pendientes de calibración cercanas a 1.

Los resultados de capacidad predictiva se analizaron usando análisis de varianza de medidas repetidas basado en modelos mixtos. Los análisis se realizaron usando la librería *nlme* de R (Pinheiro *et al.*, 2011).

4.3. RESULTADOS

La incorporación de las variables relacionadas con el suelo tuvo un efecto significativo en la bondad de ajuste (AIC) y en la discriminación (AUC) de los modelos, pero no generó mejoras significativas en la pendiente de calibración (tabla 4.1).

			Valores de F		
	g.l. num. [1]	g.l. den. [2]	Bondad de ajuste (AIC) [3]	Discriminación (AUC) [4]	Pendiente calibración
Constante	1	374	96,55***	39.494,81***	15.847,14***
Conj. de variables independientes	2	374	50,92***	149,31***	0,48

Tabla 4.1. Análisis de la varianza de los modelos mixtos para las medidas de bondad de ajuste y capacidad predictiva (una columna para uno de los tres modelos mixtos). Códigos de significación de los valores de p: *** <0,001, ** 0,001-0,01, * 0,01-0,05. [1] Grados de libertad del numerador. [2] Grados de libertad del denominador. [3] AIC: índice de información de Akaike. [4] AUC: área bajo la curva característica operativa del receptor.

Diferencia	Bondad de ajuste (AIC) [1]	Discriminación (AUC) [2]	Pendiente de calibración
Clima & Litología - Sólo clima	353,81 ***	0,0073 ***	-0,0050
Clima & Litología & Tipo suelo - Sólo clima	545,78 ***	0,0141 ***	0,0077
Clima & Litología & Tipo suelo - Clima & Litología	191,96 **	0,0068 ***	0,0127

Tabla 4.2. Pruebas de comparación de medias de Tukey correspondientes a los modelos mixtos. Se presentan las diferencias de las tres medidas de bondad de ajuste y capacidad predictiva (una columna para uno de los tres modelos mixtos). Códigos de significación de los valores de p: *** <0,001, ** 0,001-0,01, * 0,01-0,05. [1] AIC: índice de información de Akaike. [2] AUC: área bajo la curva característica operativa del receptor.

Los modelos que incluyen la litología obtuvieron mayores AIC en el 85,6% de las especies consideradas y ese porcentaje subió al 93,6 al añadir el grupo de suelo. Únicamente en 8 de las 188 especies consideradas el modelo con variables climáticas superó en el valor de AIC a ambos modelos con información del suelo.

La mejora media del AUC respecto a los modelos climáticos fue de 0,0073 ($p<0,001$) para los modelos con litología y de 0,0141 ($p<0,001$) para los que además

incluyeron el grupo de suelo (tabla 4.2). La mejora media del AUC de los modelos más complejos (clima, litología y grupo se suelo) respecto a los de complejidad intermedia (clima y litología) también fue significativa (0,0068, p<0,001).

La diferencia media de la pendiente de calibración respecto a los modelos climáticos no difirió significativamente de 0 ni para los modelos con litología (p=0,881), ni para los que además incluyeron el grupo de suelo (p=0,745, tabla 4.2). La diferencia media de la pendiente de calibración de los modelos más complejos (clima, litología y grupo se suelo) respecto a los de complejidad intermedia (clima y litología) tampoco fue significativa (p=0,47).

4.4. DISCUSIÓN

4.4.1. Aspectos metodológicos

Los estudios previos sobre el efecto de la incorporación de variables de suelo en la capacidad predictiva de los modelos de distribución de especies se centraron en la evaluación de la discriminación, es decir, en la capacidad de las predicciones para separar las observaciones de presencia de las de ausencia (Coudun *et al.*, 2006; Coudun & Gégout, 2007; Marage & Gégout, 2009; Titeux *et al.*, 2009; Gastón *et al.*, 2009). No se consideró la calibración, es decir, la correspondencia entre las probabilidades pronosticadas y las frecuencias observadas. La calibración debería ser considerada antes que la discriminación si se requieren estimaciones fiables de la probabilidad de presencia (Harrell, 2001). A pesar de ello, la calibración rara vez ha sido estudiada en modelos de distribución de especies (Vaugham & Ormerod, 2005; pero véase Carroll *et al.*, 1999; Pearce & Ferrier, 2000; Ferrier *et al.*, 2002; Reineking & Schröder, 2006, capítulo 3 de esta tesis). Este estudio incluyó la calibración en la evaluación de la calidad del modelo y eso contribuye a mejorar el conocimiento sobre el efecto de la incorporación de variables del suelo a los modelos de distribución de especies.

Además, los estudios previos usaron modelos de regresión con selección de variables por pasos (Coudun *et al.*, 2006; Coudun & Gégout, 2007; Marage & Gégout, 2009; Titeux *et al.*, 2009; Gastón *et al.*, 2009). Algunos estudios de simulación han puesto de manifiesto que este procedimiento puede resultar en el mantenimiento de variables que son solo ruido aleatorio y en la eliminación de variables verdaderamente importantes (Derksen & Keselman, 1992). Esta característica de la selección de variables por pasos la convierte en un método subóptimo para comprobar si un conjunto de variables es preferible a otro. En este trabajo se ha usado un procedimiento más fiable que consiste en usar el conocimiento previo para elegir un pequeño número de conjuntos de variables, para luego comparar su bondad de ajuste y capacidad predictiva usando métodos que corrigen las diferencias de complejidad entre los modelos comparados (AIC y validación interna usando *bootstrap*, Harrell, 2001).

4.4.2. Efecto de la incorporación de las variables relacionadas con el suelo en la bondad de ajuste y capacidad predictiva de los modelos

Los resultados de este estudio confirman los resultados de otros trabajos previos que informaban de mejoras en la discriminación de los modelos que incorporan variables relacionadas con el suelo en comparación con lo que únicamente usan variables climáticas (Coudun *et al.*, 2006; Coudun & Gégout, 2007; Marage & Gégout, 2009; Titeux *et al.*, 2009; Gastón *et al.*, 2009).

La bondad de ajuste medida usando el AIC mejoró en un 87,5 % de las especies de insectos al incorporar el tipo de suelo, en el único estudio que consideró la bondad de ajuste además de la capacidad predictiva (Titeux *et al.*, 2009). Este porcentaje es similar al encontrado en nuestro estudio (93,6%).

La incorporación de las variables relacionadas con el suelo no consiguió la mejora de la calibración de los modelos exclusivamente climáticos. La calibración informa de la fiabilidad de las probabilidades pronosticadas para cada especie y es un aspecto decisivo de los modelos si se van a usar para tomar decisiones que se basen en

la probabilidad. Una posible aplicación de este tipo de modelos a la restauración de ecosistemas es ayudar a decidir si una especie debería ser plantada en un lugar en base a su probabilidad de supervivencia. Para este tipo de aplicación es fundamental que la calibración de los modelos sea buena y los resultados de este estudio cuestionarían la idoneidad de introducir variables relacionadas con el suelo para un buen número de especies. En todo caso, la decisión se podría tomar caso a caso usando el procedimiento de validación interna usado en este estudio. Pero en los proyectos de restauración de la vegetación es mucho más habitual enfrentarse a otro tipo de situación en la que ya se ha decidido que se va a instalar algún tipo de cubierta vegetal y lo que resta es decidir cuales son las *n* especies más idóneas para el lugar a restaurar. Este tipo pregunta requiere un conjunto de *N* modelos correspondientes a *N* especies disponibles en el mercado, siendo $N > n$. Además se requiere un método para evaluar conjuntamente la calidad de las predicciones que los *N* modelos ofrecen para un lugar concreto y no solamente la calidad de los modelos de cada especie evaluados en todo los lugares que forman el área de estudio. En el capítulo 5 se propone un método para llevar a cabo este tipo de evaluación y se evalúa el efecto de la incorporación de las variables relacionadas con el suelo a los resultados de esa evaluación conjunta.

4.4.3. Aspectos relacionados con la escala

Los estudios previos evaluaron el efecto de la incorporación de variables relacionadas con el suelo en los modelos usando datos de suelo de resolución espacial igual o mayor que la de los datos de distribución de especies, tanto con datos de baja (Titeux *et al.*, 2009; Gastón *et al.*, 2009) como de alta resolución espacial (Coudun *et al.*, 2006; Coudun & Gégout, 2007; Marage & Gégout, 2009). En este estudio se han incorporado a los modelos datos de suelo de menor resolución espacial que los datos de distribución de especies. La mejoría observada en la bondad de ajuste y en la discriminación, sin sacrificar la calibración, sugieren que los mapas de suelo de baja resolución espacial son una fuente válida para la modelización de la distribución de plantas con datos de mayor resolución espacial, a pesar de los errores derivados de la mezcla de datos a diferentes escalas.

4.5. CONCLUSIONES

Los resultados de este estudio permiten recomendar la incorporación de la información de los mapas de suelo de baja resolución a modelos de distribución de especies vegetales ajustados con datos de mayor resolución, si el objetivo es obtener un modelo que discrimine mejor entre presencias y ausencias de la especie. Si se pretende optimizar la calibración del modelo habrá que decidir en cada caso si se incorporan las variables de suelo de baja resolución espacial, ya que la incorporación no ha sido efectiva para un número significativo de especies. En todo caso, la metodología presentada permite tomar esa decisión usando una técnica de validación interna basada en *bootstrap*.

CAPÍTULO 5

EVALUACIÓN DE LA CAPACIDAD PREDICTIVA DE MODELOS DE DISTRIBUCIÓN DE ESPECIES APLICADOS A LA SELECCIÓN DE ESPECIES PARA LA RESTAURACIÓN DE LA VEGETACIÓN

Los resultados de este capítulo han sido publicados en: GASTÓN A., GARCÍA-VIÑAS J.I., 2013. Evaluating the predictive performance of stacked species distribution models applied to plant species selection in ecological restoration. Ecol.Model., 263(0), 103-108.

5.1. INTRODUCCIÓN

Como cualquier modelo empírico con vocación predictiva, los modelos de distribución de especies deben ser validados con respecto a su capacidad predictiva antes de ser usados en la práctica. Existen numerosos estadísticos para evaluar la capacidad predictiva que se pueden clasificar en dos grandes grupos, los que evalúan la discriminación y los que evalúan la calibración (Pearce & Ferrier, 2000). La discriminación es la capacidad de las predicciones del modelo para separar ausencias de presencias de la especie. Un modelo que discrimina bien es capaz de ordenar los lugares del área de estudio en función de la idoneidad para la especie. La calibración evalúa la capacidad del modelo para generar predicciones de probabilidad fiables, es decir, que se parezcan a las frecuencias observadas en la realidad. Un modelo puede discriminar bien pero estar mal calibrado porque la discriminación únicamente evalúa cómo se ordenan las observaciones en función de las predicciones de idoneidad y no controla si esos valores son fiables o proporcionados. Por ejemplo, supongamos que un modelo (llamémoslo A) pronostica una probabilidad para *Pinus uncinata* Ramond ex DC. en el Cabo de Gata de 0,01 y de 0,99 en el Valle de Benasque y otro modelo (llamémoslo B) pronostica 0,98 y 0,99 respectivamente. Dado que la especie está ausente en el Cabo de Gata y presente en el Valle de Benasque, y que ambos modelos pronostican una probabilidad mayor para el enclave pirenaico, la capacidad discriminativa del modelo A no es mejor que la del modelo B. Es obvio que el modelo A se ajusta mucho mejor a la realidad porque está mejor calibrado, es decir, si visitáramos puntos al azar en ambas zonas anotando la presencia o ausencia de la especie, las frecuencias que encontraríamos serían mucho más parecidas a las predicciones del modelo A que a las del modelo B.

Darle prioridad a la discriminación o a la calibración dependerá de la aplicación práctica para la que se destine el modelo de distribución de especies. Las aplicaciones en las que se requiera identificar las zonas donde las características del medio físico son más idóneas para una especie requerirán modelos que discriminen bien (p.ej. la identificación de las áreas donde es más probable que sobreviva una especie amenazada). En los casos en que haya que comparar probabilidades en términos

cuantitativos y no solamente ordenarlas, la calibración es mucho más importante (p.ej. cuantificar las perdidas económicas debidas al cambio climático y el consiguiente aumento de la incidencia de un patógeno en la producción de especies con interés comercial).

En el ámbito de la selección de especies en la restauración de la vegetación la fiabilidad de las probabilidades pronosticadas es crucial si el modelo se va a usar para descartar o no la plantación o siembra de una especie concreta. Pero en los proyectos de restauración de la vegetación es mucho más habitual enfrentarse a otro tipo de situación en la que ya se ha decidido que se va a instalar algún tipo de cubierta vegetal y lo que resta es decidir cuales son las *n* especies más idóneas para el lugar a restaurar. Este tipo de pregunta requiere un conjunto de *N* modelos correspondientes a *N* especies disponibles en el mercado, siendo $N > n$. En este caso, lo que interesa no es comparar las predicciones de una especie en todas las localidades del área de estudio con las observaciones de dicha especie, sino comparar las predicciones para las *N* especies en una localidad concreta con la lista de especies presentes y ausentes de dicha localidad. Se trata de comprobar si las predicciones del conjunto de modelos discriminan adecuadamente entre especies presentes y ausentes en una localidad.

La validación simultánea de un conjunto de modelos de distribución de especies es habitual en los estudios que intentan pronosticar la composición específica de una comunidad a partir de modelos de especies individuales, para aplicaciones varias en el ámbito de la conservación de los hábitats y la gestión de los recursos naturales. En algunos casos se ha usado la evaluación de la bondad de ajuste entre predicciones y observaciones usando el logaritmo de la verosimilitud (Oberdorff *et al.*, 2001; Clarke *et al.*, 2003). También se han usado medidas de disimilitud como la de Bray-Curtis entre las probabilidades pronosticadas y la composición específica observada en cada localidad (Van Sickle, 2008). El enfoque más habitual es convertir las probabilidades en pronósticos de ausencia o presencia y aplicar alguno de los estadísticos de discriminación derivados de la matriz de confusión (Fielding & Bell, 1997): la fracción de falsos positivos y la de falsos negativos (p.ej. Avery & Van Riper, 1990; Block, 1994; Feria A. & Peterson, 2002), la fracción de verdaderos positivos y la

de verdaderos negativos (p.ej. Feria A. & Peterson, 2002; Kattwinkel *et al.*, 2009), la fracción de casos correctamente clasificados (p.ej. Gabriels *et al.*, 2007; Kattwinkel *et al.*, 2009) o el índice de Kappa (p.ej. Gabriels *et al.*, 2007; Kattwinkel *et al.*, 2009;).

No se ha localizado ninguna evaluación de la capacidad predictiva de un conjunto de modelos aplicados a la selección de especies para la restauración de la vegetación. Lo más adecuado sería usar un estadístico de discriminación que informara si las predicciones de los modelos para un lugar permiten separar adecuadamente las especies presentes de las ausentes. Los estadísticos de discriminación usados en los estudios citados anteriormente requieren que la probabilidad de presencia de cada especie sea convertida en un pronóstico de presencia o ausencia. Para realizar la conversión hay que fijar un umbral de probabilidad por encima del cual se considera que el pronóstico es de presencia. La determinación de este umbral no es una tarea sencilla como ponen de manifiesto las numerosas formas de hacerlo que se han ensayado (Liu *et al.*, 2005). Además, este tipo de conversión elimina parte de la información que ofrecen los modelos, igualando especies con probabilidades muy distintas (p.ej. 0,2 y 0,9 serían lo mismo si el umbral es 0,15). Una forma de evitar esta conversión es usar medidas de la discriminación independientes del umbral de probabilidad, como el área bajo la curva característica operativa del receptor (AUC por sus siglas en inglés, Fielding & Bell, 1997). En este caso, el AUC estimaría la probabilidad de que una especie tomada al azar entre las que están presentes en una localidad obtenga una probabilidad pronosticada mayor que una especie tomada al azar entre las que están ausentes.

Se propone un método para la evaluación de un conjunto de modelos de distribución de especies en el marco de la selección de especies para la restauración de la vegetación, usando el AUC como medida de la capacidad del conjunto de modelos para discriminar entre las especies presentes y ausentes en cada localidad de la muestra de validación. Se presenta una aplicación del método a modelos de distribución de especies de plantas leñosas en España, en la que se evalúa si la incorporación de información de mapas de suelo de baja resolución mejora la

capacidad del conjunto de modelos para discriminar entre especies presentes y ausentes en cada localidad.

5.2 MATERIAL Y MÉTODOS

5.2.1. Datos de distribución de especies

Se ajustaron modelos para 188 especies arbóreas, arbustivas y herbáceas a partir de los datos de distribución de dichas especies en las 120.938 teselas peninsulares del Mapa Forestal de España a 1:200.000 (Ruiz de la Torre, 1990). Cuando la variable dependiente es binaria el tamaño de muestra efectivo no es el número total de observaciones en la muestra sino el número de casos positivos o negativos, el que sea menor. El número de presencias por especie osciló entre 15 y 42.720, con un 90% de la especies entre 28 y 8.735 presencias.

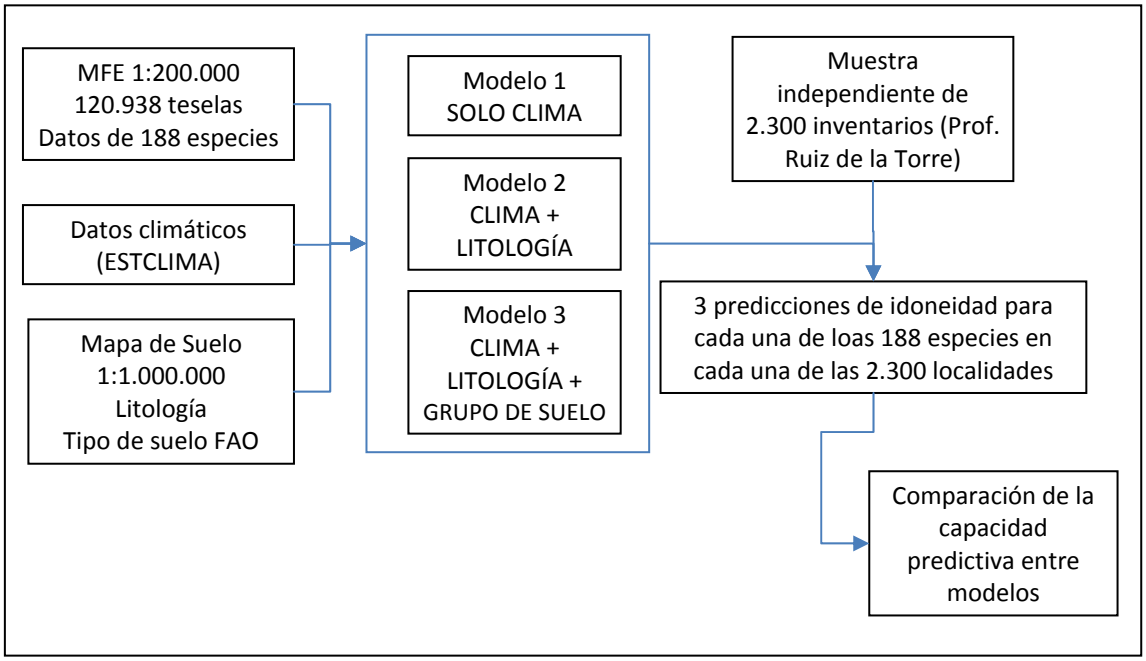
5.2.2. Variables ambientales

Se generaron capas de información climática aplicando un modelo de regresión múltiple basado en datos de estaciones meteorológicas (Sánchez Palomares *et al.*, 1999) a los datos de elevación de 3 arcosegundos (≈ 90 m) de resolución del STRM (Farr *et al.*, 2007). Inicialmente se consideraron 17 variables climáticas habitualmente usadas en estudios autoecológicos de especies arbóreas en España (p.ej. Alonso Ponce *et al.*, 2010b): Precipitaciones medias estacionales (4), precipitación media anual, temperaturas medias estacionales (4), temperatura media anual, temperatura media de las máximas del mes más cálido, temperatura media de la mínimas del mes más frío, duración del periodo árido, intensidad de la aridez, evapotranspiración potencial anual media, superávit hídrico anual medio, déficit hídrico anual medio. La información climática se asoció a cada tesela tomando el valor correspondiente al centroide del polígono.

Se usaron técnicas de análisis de conglomerados como estrategia para reducir el número de variables (Harrell, 2001). Se construyeron grupos de variables con

análisis jerarquizado de conglomerados aplicados a una matriz de similitud entre variables (coeficientes de correlación de Spearman al cuadrado). Una vez se habían definido los grupos de variables, se calculó el primer Componente Principal como representante de cada grupo. El procedimiento de reducción de variables generó seis grupos (véase la figura 2.1 del capítulo 2): variables relacionadas con las condiciones térmicas medias (*T, TP, TO, ETP*), con las condiciones térmicas estivales (*TE, TMMC*), con las condiciones térmicas invernales (*TI, TMMF*), con la disponibilidad hídrica durante el periodo árido (*A, IA, PE, DEF*) y con las disponibilidad hídrica media (*PI, PO, PP, P, SUP*).

Se usó la Base de Datos de Suelos Europeos (ESDB) que incluye una mapa a escala 1:1.000.000 (Van Liedekerke *et al.*, 2006) como fuente de datos de suelo. Se extrajeron tres variables del mapa de suelo: la naturaleza calcárea del sustrato (variable binaria), la presencia de yeso (variable binaria) y el grupo de suelo según la leyenda del mapa de suelos de la FAO (variable cualitativa con 14 niveles en el área de estudio).



Cuadro 5.1. Esquema del diseño experimental.

Se ajustaron tres modelos de complejidad creciente para cada especie (véase cuadro 5.1): el primero con variables independientes exclusivamente climáticas (15 parámetros, véase sección 5.2.3), el segundo añadiendo las dos variables litológicas (17 parámetros) y el tercero añadiendo al segundo el tipo de suelo (30 parámetros).

5.2.3. Estrategia de modelización

Se optó por una técnica de regularización aplicada a la regresión logística como alternativa a la selección por pasos de variables. Una forma de aplicar regularización a la regresión logística es usar la estimación de parámetros por máxima verosimilitud penalizada (Harrell, 2001; véanse aplicaciones a los modelos de distribución de especies en Reineking & Schröder, 2006 y en el capítulo 3). En la regresión logística penalizada se maximiza una versión penalizada de la función de verosimilitud (PML):

$$\mathrm{PML}=\ln L - 0.5\,\lambda \sum (s_i\,\beta_i)^2$$

donde L es la función de verosimilitud convencional, λ es un factor de penalización, β_i son los coeficientes de regresión y es un factor de escala para que $s_i\,\beta_i$ sea adimensional. Este procedimiento de estimación reduce el valor absoluto de los coeficientes de regresión acercándolos a cero, esto causa predicciones sesgadas para la muestra con la que se ajusta el modelo, pero mejora la precisión de las predicciones con datos nuevos. Aunque la penalización no elimina variables independientes del modelo, reduce el número efectivo de parámetros estimados y, por lo tanto, ayuda a evitar los problemas de capacidad predictiva causados por el sobreajuste (Harrell, 2001).

Los modelos penalizados se ajustaron usando la desviación típica de cada variable como factor de escala y usando una versión modificada del índice de información de Akaike (AIC) para seleccionar el factor de penalización óptimo (Harrell, 2001). El factor de penalización óptimo (λ) es el que maximiza el valor del AIC modificado, elegido de entre una serie de valores de 0 a 10 con paso de 0,25. Se usaron las funciones *lrm* y *pentrace* de la librería *Design* (Harrell, 2009) del entorno de cálculo estadístico R (R Development Core Team, 2009).

Se espera que la relación entre la presencia de las especies y las variables ambientales no sea lineal, es más, una proporción significativa de las respuestas puede ser asimétrica (Oksanen & Minchin, 2002). Por lo tanto, la complejidad de la variables independientes cuantitativas se fijó a priori usando splines cúbicas restringidas (Harrell, 2001) de cuatro nodos. Esta forma de añadir términos no lineales permite respuestas desde lineales hasta unimodales asimétricas y requieren la estimación de tres parámetros por variable.

5.2.4. Evaluación de la capacidad predictiva

Una evaluación de la capacidad predictiva de los modelo basada en la misma muestra con la que se ajustaron podría ofrecer resultados demasiado optimistas (Harrell, 2001), de modo que se opto por validar los modelos usando una muestra independiente de 2.300 inventarios de vegetación realizados por el profesor Juan Ruiz de la Torre. Se trata de un conjunto de inventarios distribuidos por toda la Península Ibérica, desde el nivel del mar hasta 2880 m de altitud, que abarcan agrupaciones arbóreas, arbustivas y herbáceas, usando parcelas rectangulares grandes (la mayoría de más de 400 m^2), con una media de 34 especies por parcela y que se encuentran disponibles a través del proyecto HispaVeg (http:/www.hispaveg.org).

La capacidad predictiva se evalúo considerando el conjunto de especies simultáneamente para poder estudiar si el conjunto de probabilidades pronosticadas discrimina bien entre las presencias y ausencias en un inventario concreto. Se eligió el AUC como estadístico de discriminación porque no requiere la conversión de las probabilidades en una variable binaria. En este caso, el AUC estima la probabilidad de que una especie tomada al azar entre las que están presentes en un inventario obtenga una probabilidad pronosticada mayor que una especie tomada al azar entre las que están ausentes.

El AUC se calculó siguiendo el procedimiento de cálculo del índice de concordancia (C) de Harrell que es idéntico al AUC en el caso de variables binarias (Harrell *et al.*, 1982). A partir de los 188 modelos de distribución de especies, se

estimaron las probabilidades de presencia en cada localidad de inventario para las 188 especies consideradas. Se emparejó cada especie presente en el inventario con cada especie ausente y se determinó la proporción de parejas en las que la especie presente tenía una probabilidad estimada mayor (véase cuadro 5.2). El procedimiento se repitió para los 2.300 inventarios y las tres combinaciones de variables independientes (véase sección 5.2.2).

Especie	Probabilidad pronosticada por el modelo	Presencia de la especie
Quercus ilex	**0,61**	**Si**
Quercus faginea	**0,34**	**Si**
Pinus nigra	**0,28**	**Si**
Pinus pinaster	0,20	No
Pinus halepensis	**0,11**	**Si**
Quercus pyrenaica	0,04	No
Pinus sylvestris	0,01	No
Pinus pinea	0,01	No
Castanea sativa	0,00	No
Quercus humilis	0,00	No
Pinus uncinata	0,00	No
Quercus suber	0,00	No
Quercus robur	0,00	No
Quercus petraea	0,00	No
Quercus canariensis	0,00	No
Fagus sylvatica	0,00	No
Abies alba	0,00	No

CÁLCULOS

Cálculo del número total de parejas compuestas por una especie presente y otra ausente:
→ 17 especies: 4 presentes, 13 ausentes
→ 4 x 13 = 52 parejas

Cálculo del número de parejas en las que la especie presente obtienen una probabilidad mayor que la ausente:
→ 51 parejas (todas menos *Pinus pinaster – P. halepensis*)

Cálculo del AUC:
→ AUC = 51/52 = 0,98

Cuadro 5.2. Ejemplo de cálculo simplificado del AUC para una localidad de inventario siguiendo el procedimiento de cálculo del índice de concordancia (C) de Harrell que es idéntico al AUC en el caso de variables binarias (Harrell *et al.*, 1982). Solamente se han considerado las especies arbóreas dominantes de primer orden para ilustrar el cálculo de forma sencilla.

Los valores de AUC correspondientes a las tres combinaciones de variables independiente se compararon usando análisis de varianza de medidas repetidas basado en modelos mixtos. Los análisis se realizaron usando la librería *nlme* de R (Pinheiro *et al.*, 2011).

5.3. RESULTADOS

La capacidad del conjunto de modelos para discriminar entre las especies presentes y ausentes en los inventarios de vegetación que forman la muestra de

validación fue, en promedio, aceptable para los modelos climáticos (AUC media: 0,883, error típico: 0,0015).

La incorporación de variables relacionadas con el suelo tuvo un efecto significativo (véase tabla 5.1). La litología mejoró la discriminación entre especies presentes y ausentes en los inventarios de validación respecto a los modelos climáticos y la adición del grupo de suelo también supuso mejoría, tanto respecto a los modelos climáticos como a los que ya incluían la litología. Las diferencias observadas fueron significativas según el análisis de la varianza de medidas repetidas usando modelos mixtos (véase tabla 5.2).

	Grados libertad numerador	Grados libertad denominador	Valor de F	Prob(>|F|)
Constante	1	4596	357.954,5	< 0,0001
Conj. variables independientes	2	4596	33,3	< 0,0001

Tabla 5.1. Análisis de la varianza del modelo mixto de medidas repetidas.

Diferencia	Estima	Error típico	Valor de z	Prob(>|z|)
Clima & Litología - Sólo clima	0,0027	0,0006	4,415	0,0001
Clima & Litología & Grupo suelo - Sólo clima	0,0050	0,0006	8,157	< 0,0001
Clima & Litología & Grupo suelo - Clima & Litología	0,0023	0,0006	3,742	0,0005

Tabla 5.2. Prueba de comparación de medias de Tukey correspondiente al modelo mixto de medidas repetidas.

5.4. DISCUSIÓN

El método propuesto ofrece una nueva forma de evaluar la calidad de las predicciones de los modelos de distribución de especies aplicados a la selección de especies en la restauración de la vegetación. Se adapta a una de las cuestiones que más frecuentemente hay que responder en los proyectos de restauración de la vegetación: ¿cuales son las n especies más idóneas para las características del medio físico del lugar a restaurar? La respuesta se obtiene midiendo la capacidad de las

predicciones del conjunto de especies para discriminar entre especies presentes y ausentes. A diferencia de los estudios en otros ámbitos de la modelización ecológica que han evaluado la capacidad discriminativa del conjunto de modelos usando estadísticos derivados de la matriz de confusión (p.ej. Avery & Van Riper, 1990; Block, 1994; Feria A. & Peterson, 2002; Kattwinkel *et al.*, 2009; Gabriels *et al.*, 2007), el método propuesto utiliza el área bajo la curva característica operativa del receptor o AUC, que ofrece una medición de la discriminación sin el inconveniente de tener que fijar un umbral para convertir las probabilidades en una variable binaria (Fielding & Bell, 1997).

La utilidad del método de evaluación se ha puesto de manifiesto en el estudio de caso presentado. Los resultados contradicen en parte a las evaluaciones de modelos realizadas especie a especie con la misma muestra y estrategia de modelización (véase capítulo 4). Cuando las predicciones de los modelos se evaluaron especie a especie, la incorporación de variables relativas al suelo no consiguió una mejora significativa de la calibración, es decir, no aumento la fiabilidad de las probabilidades pronosticadas (véase tabla 4.2). Dado que una estimación fiable de las probabilidades de presencia de las especies parece clave en la selección de especies para la restauración de la vegetación, la incorporación de variables relativas al suelo parecía cuestionable si los modelos se van a aplicar a la selección de especies. Al evaluar las predicciones localidad a localidad en vez de especie a especie, se ha obtenido un resultado diferente: la incorporación de las variables relativas al suelo mejora la capacidad de los modelos para discriminar entre especies presentes y ausentes en los inventarios de la muestra de validación.

5.5. CONCLUSIONES

Sería recomendable incorporar el método de evaluación de la capacidad predictiva propuesto a los estudios de modelización de la distribución de especies aplicados a la selección de especies en la restauración de la vegetación.

La metodología propuesta también podría ser aplicada a otros ámbitos de la modelización ecológica en los que se requiera evaluar la concordancia entre las

probabilidades de presencia provenientes de modelos de distribución ajustados para un conjunto de especies y las observaciones presencia/ausencia de dichas especies en una localidad.

CAPÍTULO 6

DISCUSIÓN Y CONCLUSIONES FINALES

El apoyo a la selección de especies a la restauración de la vegetación en España en los últimos 40 años se ha apoyado fundamentalmente en modelos de distribución de especies (p.ej. Gandullo & Sánchez Palomares, 1994; García López & Allúe Camacho, 2004; Morote *et al.*, 2001; Alonso Ponce *et al.*, 2010b). Con esta tesis se ha intentado contribuir a la mejora de la capacidad predictiva de los modelos introduciendo algunas propuestas metodológicas. A continuación se discuten las aportaciones de esta tesis a cada uno los tres principales componentes que conforman los modelos de distribución de especies: un modelo ecológico, un modelo de datos y un modelo estadístico (Austin, 2002). Todo ello en el marco de la selección de especies para la restauración de la vegetación.

6.1. MODELO ECOLÓGICO

El modelo ecológico incluye las hipótesis de partida respecto a qué factores determinan la distribución de las especies o qué forma tienen las respuestas de las especies. En general, se han asumido las hipótesis sobre el modelo ecológico más frecuentes en la bibliografía, es decir, el clima es el factor ecológico más determinante en la distribución geográfica de las especies vegetales (p.ej. Thuiller *et al.*, 2004) y la relación entre la presencia de las especies y las variables ambientales no tiene porque ser únicamente lineal o unimodal simétrica (Oksanen & Minchin, 2002).

La incorporación de la información relativa al suelo en los modelos es deseable como han mostrado algunos estudios previos (Coudun *et al.*, 2006; Coudun & Gégout, 2007), pero es habitual que no se disponga de mediciones directas de las propiedades físicas y químicas del suelo, de manera que a los criterios ecológicos se unen, en este caso, los criterios relativos al modelo de datos. Una alternativa a las mediciones directas es la información contenida en los mapas de suelos. Los mapas de suelo que cubren grandes extensiones tienen baja resolución espacial (p.ej. 1:1.000.000 para España, Gómez-Miguel, 2007; o Eurasia, Van Liedekerke *et al.*, 2006) y se han usado con éxito para modelos ajustados con datos de distribución de especies de baja resolución espacial (Titeux *et al.*, 2009; Gastón *et al.*, 2009). Los resultados del capítulo 4 permiten recomendar la incorporación de la información de los mapas de suelo de

baja resolución a modelos de distribución de especies ajustados con datos de mayor resolución, si el objetivo es obtener un modelo que discrimine mejor entre presencias y ausencias de la especie. Si se pretende optimizar la calibración (la concordancia entre probabilidades pronosticadas y las frecuencias observadas) del modelo habrá que decidir en cada caso si se incorporan las variables de suelo de baja resolución espacial, ya que la incorporación no ha sido efectiva para un número significativo de especies. En todo caso, la metodología presentada permite tomar esa decisión usando una técnica de validación interna basada en *bootstrap* (Steyerberg *et al.*, 2001).

6.2. MODELO DE DATOS

En una situación ideal, la discusión sobre el modelo de datos se centraría en cómo realizar el muestreo para obtener los datos y cómo medir las variables de interés. Dado que es improbable que se realice un esfuerzo de muestreo a escala nacional con el objetivo de desarrollar modelos optimizados para la selección de especies, la discusión sobre el modelo de datos tiene que restringirse a cómo usar los datos disponibles y qué resultados se pueden esperar de los modelos ajustados con dichos datos. Para la mayoría de las especies de plantas, las fuentes de datos corológicos disponibles las constituyen bases de datos que recopilan datos florísticos e inventarios de vegetación como el Proyecto Anthos (www.anthos.es), el Sistema de Información de la Vegetación Ibérica y Macaronésica (www.sivim.es) o el portal español del Global Bodiversity Information Facility (ww.gbif.es). La exhaustividad y resolución espacial de los datos depende de las especie considerada, pero, en general, la fiabilidad de las ausencias es baja y la resolución espacial es de hasta 100 km^2. Esta resolución espacial tan grosera queda muy lejos de la escala de trabajo de un proyecto de restauración de la vegetación.

Los resultados expuestos en el capítulo 2 confirman que la aplicación directa de los modelos obtenidos con datos de baja resolución espacial genera predicciones de alta resolución poco fiables. La diferencia de resolución espacial afecta a la prevalencia y a la importancia relativa de las variables y causa falta de calibración de los modelos (Pearce & Ferrier, 2000), es decir, baja concordancia entre las probabilidades

pronosticadas y las frecuencias observadas. El método de recalibración propuesto para trasladar las predicciones entre escalas consigue mejorar la fiabilidad de las probabilidades usando muestras pequeñas de alta resolución (véase capítulo 2). Además se trata de la primera evaluación de la calibración de predicciones entre escalas para modelos de distribución de especies, ya que otros estudios de modelización entre escalas usaron medidas de discriminación para evaluar la capacidad predictiva de los modelos (Lloyd & Palmer, 1998; Collingham *et al.*, 2000; Barbosa *et al.*, 2003; Araújo *et al.*, 2005; Barbosa *et al.*, 2010). El método propuesto permite usar los datos de baja resolución disponibles para casi cualquier especie vegetal en España en el ajuste de un modelo de baja resolución y a continuación tomar una muestra pequeña de datos de alta resolución espacial con la que recalibrar el primer modelo. Esta estrategia permite obtener modelos de distribución de especies aplicables a alta resolución espacial con un coste mucho menor que el derivado de un muestreo clásico. Para las especies con datos provenientes de un muestreo de alta resolución espacial, como el Inventario Forestal Nacional, no será necesario acudir al método de trasladar predicciones ente escalas y se podrán aplicar métodos de modelización en una única escala.

6.3. MODELO ESTADÍSTICO

El modelo estadístico incluye la selección del método estadístico para estimar los coeficientes, el método de estimación de los errores o el método de validación de las predicciones.

6.3.1. Selección del método estadístico

La regresión logística, un caso particular de los modelos lineales generalizados, ha sido el método más usado en los modelos de distribución de especies de las dos últimas décadas del siglo XX (Pearce & Ferrier, 2000), pero los resultados del estudio comparativo de métodos más exhaustivo hasta la fecha indicaban que algunos métodos más modernos como los de máxima entropía (Phillips *et al.*, 2006) superan en capacidad predictiva a la regresión logística (Elith *et al.*, 2006). Estos métodos más

modernos tienen en común que incluyen técnicas de regularización que combaten el riesgo de sobreajuste cuando hay demasiadas variables independientes en el modelo y la muestra es demasiado pequeña.

Los resultados del capítulo 3 han demostrado que la incorporación de técnicas de regularización a los modelos de regresión logística mejoran su capacidad predictiva, situándolos a la altura de los métodos más modernos como los de máxima entropía. Los resultados alientan a un uso más frecuente de la regresión logística penalizada para la modelización de la distribución de las especies, especialmente en aquellos casos en los se requiere ajustar un modelo complejo con un tamaño de muestra limitado.

6.3.2. Validación de las predicciones

La validación de un modelo predictivo se basa en la evaluación de la capacidad predictiva de sus predicciones. Dicha evaluación se lleva a cabo comparando las predicciones del modelo en un conjunto de observaciones que forman la muestra de validación. Lo más habitual es evaluar las predicciones para una especie comparándolas con observaciones de presencia y ausencia de dicha especie (p.ej. Elith *et al.*, 2006). Esta forma de evaluar informa sobre la capacidad del modelo para identificar los lugares idóneos para una especie concreta, pero en los proyectos de restauración de la vegetación el problema a resolver suele consistir en identificar las *n* especies más idóneas para un lugar concreto. Este tipo pregunta requiere un conjunto de *N* modelos correspondientes a *N* especies disponibles en el mercado, siendo $N > n$. En este caso lo que interesa no es comparar las predicciones de una especie en todas las localidades del área de estudio con las observaciones de dicha especie, sino comparar las predicciones para las *N* especies en una localidad concreta con la lista de especies presentes y ausentes de dicha localidad. Se trata de comprobar si las predicciones del conjunto de modelos discriminan adecuadamente entre especies presentes y ausentes en una localidad. A diferencia de los estudios en otros ámbitos de la modelización ecológica que han evaluado la capacidad discriminativa del conjunto de modelos usando estadísticos derivados de la matriz de confusión (p.ej. Avery & Van

Riper, 1990; Block, 1994; Feria A. & Peterson, 2002; Gabriels *et al.*, 2007; Kattwinkel *et al.*, 2009), el método propuesto en el capítulo 5 utiliza el área bajo la curva característica operativa del receptor o AUC, que ofrece una medición de la discriminación sin el inconveniente de tener que fijar un umbral para convertir las probabilidades en una variable binaria (Fielding & Bell, 1997).

La utilidad del método de evaluación propuesto se ha puesto de manifiesto en el estudio de caso presentado en el capítulo 5. Los resultados contradicen en parte a las evaluaciones de modelos realizadas especie a especie con la misma muestra y estrategia de modelización (véase capítulo 4). Cuando las predicciones de los modelos se evaluaron especie a especie, la incorporación de variables relativas al suelo no consiguió una mejora significativa de la calibración, es decir, no aumento la fiabilidad de las probabilidades pronosticadas (véase tabla 4.2). Dado que una estimación fiable de las probabilidades de presencia de las especies parece clave en la selección de especies para la restauración de la vegetación, la incorporación de variables relativas al suelo parecía cuestionable si los modelos se van a aplicar a la selección de especies. Al evaluar las predicciones localidad a localidad en vez de especie a especie, se ha obtenido un resultado diferente: la incorporación de las variables relativas al suelo mejora la capacidad de los modelos para discriminar entre especies presentes y ausentes en los inventarios de la muestra de validación.

6.4. CONCLUSIONES FINALES

Los resultados de este trabajo han permitido mejorar algunos aspectos clave de los modelos de distribución de especies aplicados a la selección de especies en la restauración de la vegetación, las mejoras se pueden resumir en las siguientes conclusiones finales:

1. La aplicación de técnicas de regularización a la regresión logística ha permitido obtener modelos con capacidad predictiva equiparable a los métodos más modernos como los de máxima entropía. La regresión logística penalizada puede considerarse como una de las mejores opciones metodológicas para los

modelos de distribución de especies aplicados a la selección de especies en la restauración de la vegetación.

2. La incorporación de información relativa al suelo proveniente de mapas de baja resolución espacial a modelos de alta resolución espacial ha mejorado la capacidad predictiva de forma ligera pero significativa. Cuando no se dispone de mediciones directas de las propiedades del suelo, es recomendable incorporar la información proveniente de mapas de suelo de baja resolución a los modelos de distribución de especies aplicados a la selección de especies en la restauración de la vegetación.
3. Se ha desarrollado un método de evaluación de la capacidad predictiva que informa de la capacidad de los modelos para identificar las especies más idóneas para un lugar concreto y esto lo hace más adecuado para validar los modelos de distribución de especies aplicados a la selección de especies en la restauración de la vegetación.
4. La estrategia de modelización propuesta puede extenderse a la mayoría de las especies de plantas en España con un reducido esfuerzo de muestreo. Partiendo de un modelo de baja resolución espacial ajustado con datos ya disponibles en las bases de datos públicas, el método de recalibración propuesto permite obtener predicciones probabilidad de presencia de las especies fiables y de alta resolución espacial, usando una muestra pequeña de nuevos datos de alta resolución espacial.

CAPÍTULO 7

REFERENCIAS BIBLIOGRÁFICAS

ALLUÉ J.L., 1990. Atlas fitoclimático de España. Instituto de Nacional de Investigaciones Agrarias, Madrid (España). 221 pp.

ALONSO PONCE R., SÁNCHEZ PALOMARES O., ROIG S., LÓPEZ SENESPLEDA E., GANDULLO J.M., 2010a. Las estaciones ecológicas actuales y potenciales de los sabinares albares españoles. Monografías INIA. Serie Forestal nº 19, Madrid. 188 pp.

ALONSO PONCE R., LÓPEZ SENESPLEDA E., SÁNCHEZ PALOMARES O., 2010b. A novel application of the ecological field theory to the definition of physiographic and climatic potential areas of forest species. European Journal of Forest Research, 129(1), 119-131.

ARAÚJO M.B., THUILLER W., WILLIAMS P.H., REGINSTER I., 2005. Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. Global Ecol.Biogeogr., 14(1), 17-30.

AUSTIN M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol.Model., 157(2-3), 101-118.

AVERY M.L., VAN RIPER C., 1990. Evaluation of wildlife-habitat relationships data base for predicting bird community composition in central California chaparral and blue oak woodlands. Calif.Fish Game, 76(2), 103-117.

BALAGUER L., VALLADARES F., ESCUDERO A., MOLA I., ALFAYA V., 2011. Restauración ecológica e infraestructuras de transporte: perspectivas y recomendaciones. In: Restauración ecológica de áreas afectadas por infraestructuras de transporte. Bases científicas para soluciones técnicas.(Valladares F., Balaguer L., Mola I., Escudero A., Alfaya V. , ed.). Fundación Biodiversidad, Madrid, España. pp. 304-309.

BARBOSA A.M., REAL R., VARGAS J.M., 2010. Use of coarse-resolution models of species' distributions to guide local conservation inferences. Conserv.Biol., 24(5), 1378-1387.

BARBOSA A.M., REAL R., OLIVERO J., MARIO VARGAS J., 2003. Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. Biol.Conserv., 114(3), 377-387.

BLOCK W.M., 1994. Assessing wildlife-habitat-relationships models: a case study with California oak woodlands. Wildl.Soc.Bull., 22(4), 549.

BRAVO A., MONTERO G., 2008. Descripción de los caracteres culturales de las principales especies forestales de España. In: Compendio de Selvicultura Aplicada en España (Serrada R., Montero G., Reque J.A. , ed.). INIA, Madrid. pp. 1037-1114.

CAB International, 2000. Forestry compendium, global module. CAB International. Wallingford, United Kingdom.

CARQUE E., MARRERO M.V., NARANJO A., 2008. Identificación y selección de especies adecuadas para la recuperación de hábitats afectados por la desertificación en Canarias. Gobierno de Canarias, 58 pp.

CARROLL C., ZIELINSKI W.J., NOSS R.F., 1999. Using Presence-Absence Data to Build and Test Spatial Habitat Models for the Fisher in the Klamath Region, U.S.A. Conserv.Biol., 13(6), 1344-1359.

CASTEJÓN M., SÁNCHEZ F., ELENA-ROSSELLÓ R., 1998. SIGREFOR: Sistema de Información Geográfica para la Reforestación. Fundación Conde del Valle de Salazar, Madrid (España). 18 pp.

CLARKE R.T., WRIGHT J.F., FURSE M.T., 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecol.Model., 160(3), 219-233.

CLEWELL A., RIEGER J., MUNRO J., 2005. Guidelines for developing and managing ecological restoration projects. Society for Ecological Restoration, Tucson (USA). 16 pp.

COLLINGHAM Y.C., WADSWORTH R.A., HUNTLEY B., HULME P.E., 2000. Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. J.Appl.Ecol., 37, 13-27.

COUDUN C., GÉGOUT J.C., 2007. Quantitative prediction of the distribution and abundance of *Vaccinium myrtillus* with climatic and edaphic factors. Journal of Vegetation Science, 18(4), 517-524.

COUDUN C., GÉGOUT J.C., PIEDALLU C., RAMEAU J.C., 2006. Soil nutritional factors improve models of plant species distribution: an illustration with *Acer campestre* (L.) in France. J.Biogeogr., 33(10), 1750-1763.

CUMMING G.S., 2000. Using between-model comparisons to fine-tune linear models of species ranges. J.Biogeogr., 27(2), 441-455.

DE LA ROSA D., MORENO J.A., GARCIA L.V., ALMORZA J., 1992. MicroLEIS: A microcomputer-based Mediterranean land evaluation information system. Soil use Manage., 8(2), 89-96.

DERKSEN S., KESELMAN H.J., 1992. Backward, forward and stepwise automated subset selection algorithms : frequency of obtaining authentic and noise variables. Brit J Math Stat Psy, 45(2), 265-282.

ELENA ROSSELLÓ R., 1997. Clasificación biogeoclimática de España peninsular y balear. Ministerio de Agricultura, Pesca y Alimentación, Madrid (España). 446 pp.

ELITH J., GRAHAM C.H., 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography, 32(1), 66-77.

ELITH J., H. GRAHAM C., P. ANDERSON R., DUDÍK M., FERRIER S., GUISAN A., J. HIJMANS R., HUETTMANN F., R. LEATHWICK J., LEHMANN A., LI J., G. LOHMANN L., A. LOISELLE B., MANION G., MORITZ C., NAKAMURA M., NAKAZAWA Y., MCC. M. OVERTON J., TOWNSEND PETERSON A., J. PHILLIPS S., RICHARDSON K., SCACHETTI-

PEREIRA R., E. SCHAPIRE R., SOBERÓN J., WILLIAMS S., S. WISZ M., E. ZIMMERMANN N., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29(2), 129-151.

ELLIS E.A., NAIR P.K.R., JESWANI S.D., 2005. Development of a web-based application for agroforestry planning and tree selection. Comput.Electron.Agric., 49(1), 129-141.

FARR T.G., ROSEN P.A., CARO E., CRIPPEN R., DUREN R., HENSLEY S., KOBRICK M., PALLER M., RODRIGUEZ E., ROTH L., SEAL D., SHAFFER S., SHIMADA J., UMLAND J., WERNER M., OSKIN M., BURBANK D., ALSDORF D., 2007. The Shuttle Radar Topography Mission. Rev.Geophys., 45(2), RG2004.

FELICÍSIMO A.M., 2003. Uses of spatial predictive models in forested areas territorial planning. CIOT 2003. IV International Conference on Spatial Planning. Zaragoza (Spain), April 2-3.

FERIA A. T.P., PETERSON A.T., 2002. Prediction of bird community composition based on point-occurrence data and inferential algorithms: a valuable tool in biodiversity assessments. Divers.Distrib., 8(2), 49-56.

FERRIER S., WATSON G., PEARCE J., DRIELSMA M., 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. Biodivers.Conserv., 11(12), 2275-2307.

FIELDING A.H., BELL J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ.Conserv., 24(01), 38.

GABRIELS W., GOETHALS P., DEDECKER A., LEK S., DE PAUW N., 2007. Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks. Aquat.Ecol., 41(3), 427-441.

GANDULLO J.M., BLANCO A., SÁNCHEZ PALOMARES O., RUBIO A., ELENA ROSSELLÓ R., GÓMEZ SANZ V., 2004a. Las estaciones ecológicas de los hayedos españoles.

Monografías INIA. Serie Forestal nº 8. Ministerio de Educación y Ciencia, Madrid. 299 pp.

GANDULLO J.M., BLANCO A., SÁNCHEZ PALOMARES O., RUBIO A., ELENA-ROSSELLÓ R., GÓMEZ SANZ V., 2004b. Las estaciones ecológicas de los castañares españoles. Monografías INIA. Serie Foretsal nº 7. Ministerio de Educación y Ciencia, Madrid. 224 pp.

GANDULLO J.M., SÁNCHEZ PALOMARES O., 1994. Estaciones ecológicas de los pinares españoles. ICONA, Madrid (España). 188 pp.

GARCÍA LÓPEZ J.M., ALLÚE CAMACHO C., 2004. Ensayo de un sistema fitoclimático de carácter autoecológico para especies arbóreas forestales en la Península Ibérica y su aplicación en labores de repoblación forestal. Actas IV Congreso Forestal Español. Zaragoza (España), 26-30 sept.

GARCÍA SALMERÓN J., 1980. Los diagramas bioclimáticos y su utilización forestal. Foret Mediterranéenne I, 22, 105-133.

GASTÓN A., SORIANO C., 2006. Contribution of the Forest Map of Spain to the chorology of woody plant species. Investigación Agraria: Sistemas y Recursos Forestales, Fuera de serie, 9-13.

GASTÓN A., SORIANO C., GÓMEZ-MIGUEL V., 2009. Lithologic data improve plant species distribution models based on coarse-grained occurrence data. Forest Systems, 18(1), 42-49.

GIBSON L., BARRETT B., BURBIDGE A., 2007. Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. Divers.Distrib., 13(6), 704-713.

GÓMEZ-MIGUEL V., 2007. Geología, Geomorfología y Edafología. Monografía del Instituto Geográfico Nacional, Madrid (España). 196 pp.

GUISAN A., EDWARDS T.C., HASTIE T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol.Model., 157(2-3), 89-100.

GUISAN A., THUILLER W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol.Lett., 8(9), 993-1009.

HARRELL F.E., 2001. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. Springer, New York. 568 pp.

HARRELL F.E., 2009. Design Package. R package version 2.3-0. http://CRAN.R-project.org/package=Design,

HARRELL F.E., CALIFF R.M., PRYOR D.B., LEE K.L., ROSATI R.A., 1982. Evaluating the Yield of Medical Tests. JAMA: The Journal of the American Medical Association, 247(18), 2543-2546.

HARRELL F.E., LEE K.L., CALIFF R.M., PRYOR D.B., ROSATI R.A., 1984. Regression modelling strategies for improved prognostic prediction. Stat.Med., 3(2), 143-152.

HEREDIA, N. (2007). Desarrollo de un modelo de evaluación de tierras en red neuronal (Sierra2) para la selección de especies arbustivas en la reforestación de zonas mediterráneas, un nuevo componente del sistema MicroLEIS. Unpublished CIHEAM-IAMZ, Zaragoza (España).

HERNANDEZ P.A., GRAHAM C.H., MASTER L.L., ALBERT D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography, 29(5), 773-785.

KADMON R., FARBER O., DANIN A., 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecol.Appl., 13(3), 853-867.

KATTWINKEL M., STRAUSS B., BIEDERMANN R., KLEYER M., 2009. Modelling multi-species response to landscape dynamics: mosaic cycles support urban biodiversity. Landscape Ecol., 24(7), 929-941.

LIU C., BERRY P.M., DAWSON T.P., PEARSON R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography, 28(3), 385-393.

LLOYD P., PALMER A.R., 1998. Abiotic factors as predictors of distribution in southern African Bulbuls. The Auk, 115(2), 404-411.

LOBO J.M., 2008. More complex distribution models or more representative data? Biodiversity Informatics, 5, 14-19.

MARAGE D., GÉGOUT J.C., 2009. Importance of soil nutrients in the distribution of forest communities on a large geographical scale. Global Ecol.Biogeogr., 18(1), 88-97.

MARINI M., BARBET-MASSIN M., LOPES L., JIGUET F., 2010. Predicting the occurrence of rare Brazilian birds with species distribution models. Journal of Ornithology, , 1-10.

MCPHERSON J.M., JETZ W., ROGERS D.J., 2006. Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions—possibilities and limitations. Ecol.Model., 192(3-4), 499-522.

MCVICAR T.R., VAN NIEL T.G., LI L., WEN Z., YANG Q., LI R., JIAO F., 2010. Parsimoniously modelling perennial vegetation suitability and identifying priority areas to support China's re-vegetation program in the Loess Plateau: Matching model complexity to data availability. For.Ecol.Manage., 259(7), 1277-1290.

MONTOYA D., PURVES D.W., URBIETA I.R., ZAVALA M.A., 2009. Do species distribution models explain spatial structure within tree species ranges? Global Ecol.Biogeogr., 18(6), 662-673.

MOROTE A., OROZCO E., LÓPEZ F., DEL CERRO A., ANDRÉS M., SELVA M., BRIONGOS J., NAVARRO R., 2001. Aplicación de un sistema de información geográfica para la elección de especie en la forestación de terrenos agrícolas de la Mancha. III Congreso Forestal Español. Granada (España), 3-5 septiembre. pp. 62-68.

MUÑOZ J., FELICÍSIMO Á.M., 2004. Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science, 15(2), 285-292.

OBERDORFF T., PONT D., HUGUENY B., CHESSEL D., 2001. A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. Freshwat.Biol., 46(3), 399-415.

OKSANEN J., MINCHIN P.R., 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecol.Model., 157(2-3), 119-129.

PEARCE J., FERRIER S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol.Model., 133(3), 225-245.

PEARSON R.G., RAXWORTHY C.J., NAKAMURA M., TOWNSEND PETERSON A., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J.Biogeogr., 34(1), 102-117.

PHILLIPS S.J., ANDERSON R.P., SCHAPIRE R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol.Model., 190(3-4), 231-259.

PHILLIPS S.J., DUDÍK M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31(2), 161-175.

PINHEIRO J., BATES D., DEBROY S., SARKAR D., TEAM R.D.C., 2011. nlme: Linear and Nonlinear Mixed Effects Models.

PYATT D.G., SUÁREZ J.C., 1997. An ecological site classification for forestry in Great Britain with special reference to Grampian, Scotland. Technical Paper-Forestry Commission (United Kingdom),

R DEVELOPMENT CORE TEAM., 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, http://www.R-project.org, Vienna, Austria.

RAY D., REYNOLDS K., SLADE J., HODGE S., 1998. A spatial solution to ecological site classification for British forestry using Ecosystem Management Decision Support. Proceedings of Third International Conference on GeoComputation Conference. Bristol, UK, 17–19 September.

REESE G.C., WILSON K.R., HOETING J.A., FLATHER C.H., 2005. Factors affecting species distribution predictions: a simulation modeling experiment. Ecol.Appl., 15(2), 554-564.

REINEKING B., SCHRÖDER B., 2006. Constrain to perform: Regularization of habitat models. Ecol.Model., 193(3-4), 675-690.

RIVAS MARTÍNEZ S., 1987. Memoria del mapa de series de vegetación de España: 1: 400.000. ICONA, Madrid (España). 268 pp.

ROURA-PASCUAL N., BROTONS L., PETERSON A., THUILLER W., 2009. Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. Biol.Invasions, 11(4), 1017-1031.

RUIZ DE LA TORRE J., 1990. Mapa forestal de España. Memoria general. ICONA, Madrid (España). 191 pp.

RUIZ DE LA TORRE J., CARRERAS C., GARCÍA VIÑAS J.I., ORTI M., 1996. Manual de la flora para la restauración de áreas críticas y diversificación en masas forestales. Consejería de Medio Ambiente de la, Sevilla (España). 208 pp.

RUIZ DE LA TORRE J., GIL P., GARCÍA-VIÑAS J.I., GONZÁLEZ-ADRADOS J.R., GIL F., 1990. Catálogo de especies vegetales a utilizar en plantaciones de carreteras. Ministerio de Obras Públicas y Urbanismo, Madrid (España). 497 pp.

SÁNCHEZ PALOMARES O., ROIG S., RÍO GAZTELURRUTIA M., RUBIO A., GANDULLO J.M., 2008. Las estaciones ecológicas actuales y potenciales de los rebollares españoles. Monografías INIA. Serie Forestal nº 17. Ministerio de Educación y Ciencia, Madrid. 343 pp.

SÁNCHEZ PALOMARES O., SÁNCHEZ SERRANO F., CARRETERO M.P., 1999. Modelos y cartografía de estimaciones climáticas termopluviométricas para la España peninsular. Instituto Nacional de Investigacines Agrarias, Madrid, Spain. 192 pp.

SÁNCHEZ PALOMARES O., JOVELLAR L.C., SARMIENTO L.A., RUBIO A., GANDULLO J.M., 2007. Las estaciones ecológicas de los alcornocales españoles. Monografías INIA. Serie Forestal nº 14. Ministerio de Educación y Ciencia, Madrid. 230 pp.

SEGURADO P., ARAÚJO M.B., 2004. An evaluation of methods for modelling species distributions. J.Biogeogr., 31(10), 1555-1568.

SER, 2004. The SER International Primer on Ecological Restoration. Society for Ecological Restoration International, Tucson (USA). 13 pp.

SERRADA R., 2000. Apuntes de repoblaciones forestales. Fundación Conde del Valle de Salazar, Madrid (España). 435 pp.

STEYERBERG E.W., EIJKEMANS M.J., HARRELL F.E.,JR, HABBEMA J.D., 2000. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat.Med., 19(8), 1059-1079.

STEYERBERG E.W., HARRELL F.E.,JR, BORSBOOM G.J.J.M., EIJKEMANS M.J.C., VERGOUWE Y., HABBEMA J.D., 2001. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J.Clin.Epidemiol., 54(8), 774-781.

STEYERBERG E.W., 2009. Clinical prediction models : a practical approach to development, validation, and updating. Springer, New York ; London. 497 pp.

STEYERBERG E.W., BORSBOOM G.J.J.M., VAN HOUWELINGEN H.C., EIJKEMANS M.J.C., HABBEMA J.D.F., 2004. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat.Med., 23(16), 2567-2586.

STOCKWELL D., 2002. Effects of sample size on accuracy of species distribution models Ecol.Model., 148(1), 1 - 13.

SWETS J., 1988. Measuring the accuracy of diagnostic systems. Science, 240(4857), 1285-1293.

TERRADAS J., 2001. Ecología de la vegetación: de la ecofisiología de las plantas a la dinámica de comunidades y paisajes. Editorial Omega, Barcelona (España). 703 pp.

THUILLER W., ARAÚJO M.B., LAVOREL S., 2004. Do we need land-cover data to model species distributions in Europe? J.Biogeogr., 31(3), 353-361.

TIBSHIRANI R., 1994. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statisical Society, Series B, 58, 267-288.

TITEUX N., MAES D., MARMION M., LUOTO M., HEIKKINEN R.K., 2009. Inclusion of soil data improves the performance of bioclimatic envelope models for insect species distributions in temperate Europe. J.Biogeogr., 36(8), 1459-1473.

TOGNELLI M.F., ROIG-JUNENT S.A., MARVALDI A.E., FLORES G.E., LOBO J.M., 2009. An evaluation of methods for modelling distribution of Patagonian insects. Revista Chilena de Historia Natural, 82(3), 347-360.

VALLE TENDERO F., LORITE MORENO J., 2004. Datos botánicos aplicados a la Gestión del Medio Natural Andaluz III. Junta de Andalucía. Consejería de Medio Ambiente, Sevilla. 512 pp.

VAN LIEDEKERKE M., JONES A., PANAGOS P., 2006. ESDBv2 Raster Library, a set of rasters derived from the European Soil Database distribution v2.0. European Commission and the European Soil Bureau Network, CDROM, EUR 19945 EN,

VAN SICKLE J., 2008. An index of compositional dissimilarity between observed and expected assemblages. J.N.Am.Benthol.Soc., 27(2), 227.

VAUGHAM I.P., ORMEROD S.J., 2005. The continuing challenges of testing species distribution models. J.Appl.Ecol., 42(4), 720-730.

WEBB D.B., WOOD P.J., SMITH J., 1980. A guide to species selection for tropical and sub-tropical plantations. Tropical Forestry Papers 15, University of Oxford, Oxford. 256 pp.

WHITE T.J.R., DOMINY S.W.J., 2005. Review of Best Practices for Tree Planting on Marginal Agriculture Lands in Ontario. Canadian Forest Service, Great Lakes Forestry Centre, Ontario, Canada. 98 pp.

WISZ M.S., HIJMANS R.J., LI J., PETERSON A.T., GRAHAM C.H., GUISAN A., NCEAS PREDICTING SPECIES DISTRIBUTIONS WORKING GROUP., 2008. Effects of sample size on the performance of species distribution models. Divers.Distrib., 14(5), 763-773.

ZAVALA M.A., BRAVO DE LA PARRA R., 2005. A mechanistic model of tree competition and facilitation for Mediterranean forests: Scaling from leaf physiology to stand dynamics. Ecol.Model., 188(1), 76-92.