BIOGEOGRAFÍA
SCIENTIA BIODIVERSITATIS

Editores:
Raimundo Real y Ana Luz Márquez

Edición:

Raimundo Real

Ana Luz Márquez

Autores: Los propios autores responsables de cada comunicación.
Diseño portada: Marcelo van Rompaey.

Maquetación: Ana Luz Márquez y Raimundo Real

Colaboran:

UNIVERSIDAD DE MÁLAGA

GOBIERNO DE ESPAÑA MISTERIO DE CIENCIA E INOVACIÓN

JUNTA DE ANDALUCIA

Scientia Biodiversitatis
Biogeografía, Diversidad y Conservación

fqrma

Edita: Raimundo Real
Ana Luz Márquez

Depósito Legal: MA-2.756-09

ISBN: 978-84-692-5169-0

Una estimación de la capacidad predictiva de los modelos de distribución

Aranda, S. C. [1, 2] y Lobo, J. M. [1]

[1]Dep. Biodiversidad y biología Evolutiva. Museo Nacional de Ciencias Naturales (CSIC). C/ José Gutiérrez Abascal, 2. 28006, Madrid, Spain.
[2]Universidade dos Açores. Dep. de Ciências Agrárias, CITA-A (Azorean Biodiversity Group). Terra-Chã, 9700-851 Angra do Heroísmo, Terceira, Açores, Portugal.
*Autor responsable: mcnsc850@mncn.csic.es

Resumen: Puesto que la sobrepredicción es uno de los principales problemas en los modelos de distribución de especies, en este trabajo examinamos si superponer muchos modelos individuales permite representar correctamente la riqueza y composición reales. Utilizamos los datos de distribución de las plantas de Tenerife (Islas Canarias) para modelizar individualmente cada una de las 841 especies nativas reconocidas en la isla (base de datos BIOTA-Canarias). Para la modelización empleamos MaxEnt, una de las técnicas que, supuestamente, ofrece mejores resultados cuando sólo se dispone de información sobre las presencias de las especies. Después de superponer todas las predicciones de los modelos individuales, evaluamos los resultados comparando la riqueza de especies y la composición florística así obtenidas para cada cuadrícula de 500 x 500 m previamente definidas como bien muestreadas. Nuestros resultados demuestran que este tipo de modelos generan una gran sobrepredicción, independientemente del punto de corte aplicado para transformar los valores continuos de favorabilidad en valores de presencia/ausencia. Sin embargo, se pueden mejorar parcialmente los resultados eligiendo un punto de corte que compense los errores de riqueza y composición y examinando las representaciones obtenidas a una escala regional.
Palabras clave: Errores de predicción, espermatofitos, modelos de distribución, sobrepredicción, Tenerife.

Abstract: Overestimation is considered one of the main problems in the accomplishment of species distribution models. Here, we examine whether the overlay of many individual species models is able to adequately represent the species richness and compositional patterns. For this purpose, we used a database on the distribution of seed plants in an intensively surveyed, species-rich island (Tenerife, Canary Islands) to individually model all of the 841 native plant species. We used MaxEnt, one of the highest-ranked, presence-only modelling techniques. All prediction results from individual models were overlapped, and we obtained and compared species richness and species composition values against those previously determined well-surveyed grid-cells. Our results show high levels of overprediction, independently of the suitability threshold applied to transform continuous values into presences/absences. However, it is possible to improve the results by choosing a threshold that compensates errors in both species richness and species composition, as well as by examining the obtained geographical representation at a regional scale.
Keywords: Distribution models, over-prediction, prediction errors, spermatophytes, Tenerife Island.

1. Introducción

Nuestro conocimiento sobre la distribución de los organismos es incompleto y sesgado, sobre todo en grupos hiperdiversos en los que el propio conocimiento taxonómico es insuficiente. Los modelos predictivos de distribución buscan paliar estas deficiencias proporcionando representaciones geográficas fiables a partir de un conjunto parcial de datos (Guisan y Zimmermann 2000, Araújo y Guisan 2006). Sin embargo, antes de realizar cualquier modelo predictivo de distribución es conveniente conocer: i) la técnica de modelización más apropiada según se busque estimar la distribución real o potencial de las especies (Soberón y Peterson 2005, Peterson 2006, Soberón 2007), y ii) tanto la calidad de los datos biológicos disponibles como la capacidad predictiva de las variables que pueden utilizarse. De este modo, obtener modelos destinados a proveer representaciones cercanas a la distribución real o la distribución potencial de las especies requiere utilizar datos, predictores y técnicas diferentes (Jiménez-Valverde *et al.* 2008).

Cuando se pretende obtener simulaciones sobre la distribución real de los organismos es imprescindible que la información de partida esté homogéneamente repartida a lo largo de todo el gradiente geográfico-ambiental del territorio analizado (Hortal *et al.* 2008, Barbosa *et al.* 2009), de manera que el modelo interpole los nuevos valores a partir de los disponibles. Si hay regiones ambientales o geográficas ocupadas por la especie pero no detectadas, o con datos insuficientes, será probable que los modelos extrapolen erróneamente sus predicciones más allá del rango de condiciones reales en los que han sido realizados (Kadmon *et al.* 2004, Hortal *et al.* 2007). Predecir la distribución real de un especie requiere, además, poseer información fiable tanto de las localidades en las que la especie está presente, como de aquellas otras donde está ausente. En realidad, conocer las ausencias es muy importante pues sólo ellas nos ayudan a investigar los factores demográficos, históricos, bióticos o geográficos que han limitado la colonización de algunos territorios ambientalmente favorables (Ricklefts y Schluter 1993, Hanski 1998, Pulliam 1998, 2000). Si estos requisitos se cumplen (tener un conjunto de datos representativo e información fiable sobre las ausencias), la utilización de técnicas de parametrización complejas que representan la estructura espacial completa de los datos puede ser recomendable (Bahn y McGill 2007).

Desafortunadamente, siempre es difícil asegurar la ausencia de una especie en una localidad (Mackenzie *et al.* 2004) y solamente el estudio pormenorizado de las bases de datos biológicas permite estimar si la información de partida constituye una muestra representativa de la "población" que pretende interpolarse (Hortal *et al.* 2008). Se han recomendado diferentes técnicas de modelización para obtener estimas de la distribución real de los organismos (Elith *et al.* 2006, Tsoar *et al.* 2007), sin considerar la calidad de los datos usados ni la necesidad de utilizar ausencias fiables. En este trabajo pretendemos evaluar la capacidad de predicción de una de estas técnicas recomendadas, utilizando como ejemplo las especies de espermatófitos de la isla de Tenerife. Para ello, primero hemos modelizado cada una de las especies y después hemos obtenido representaciones de la variación en la riqueza de especies y la composición superponiendo todos los modelos individuales. Por último, evaluamos las predicciones así obtenidas comparándolas con algunas localidades cuyos datos de riqueza y composición se consideraron fiables de acuerdo al esfuerzo de colecta realizado en ellas. Este procedimiento pretende: i) ofrecer una medida de incertidumbre para las predicciones realizadas con estos modelos, ii) estudiar la distribución espacial y ambiental de los errores existentes, y iii) proponer un método simple capaz de corregir parcialmente la sobrepredicción.

2. ¿Cómo hacemos los modelos?

A continuación explicamos de forma sencilla los tres pasos que hemos seguido para obtener las representaciones geográficas de la riqueza de especies y la composición de los espermatófitos de Tenerife.

Paso 1. Elección de los datos biológicos y ambientales.

La información biológica que hemos empleado procede de BIOTA-Canarias (<http://www.gobiernodecanarias.org/cmayot/medioambiente/medionatural/biodiversidad/bancodatos/biotaespecies.html >), probablemente la mejor base de datos que hay para el archipiélago. Concretamente, utilizamos todas las plantas nativas de espermatófitos reconocidas en Tenerife a una resolución de cuadrículas UTM de 500 x 500 metros. Elegimos esta isla como área de estudio ya que, debido a su gran tradición botánica, reúne un número impresionante de registros (casi un millón de datos de presencia para 841 especies) (Izquierdo *et al.* 2005).

Como información ambiental, hemos elegido los datos de temperatura y precipitación mensual y anual procedentes del Instituto Nacional de Meteorología, así como cuatro variables topográficas (altitud, rango de altitudes, pendiente y diversidad de orientaciones) obtenidas a partir de un Modelo Digital del Terreno. Con estas variables continuas se realizó un análisis de componentes principales obteniendo cinco componentes ortogonales con autovalores mayores que la unidad, los cuales explicaron el 91% de la variabilidad en los predictores elegidos. Elegimos las cinco variables con mayor peso para cada factor, y aquellas que no estaban significativamente correlacionadas con ninguno de estos cinco factores, con el objeto de

representar la mayor variabilidad climática del territorio analizado (temperatura media de octubre, precipitación media anual, rango de altitudes, precipitación de junio, diversidad de orientaciones, precipitación de diciembre y diversidad de pendientes). Se incluyeron también 16 variables categóricas que representan el sustrato geológico, el tipo de suelo y la orientación del terreno, facilitadas por la Consejería de Política Territorial y Medio Ambiente del Gobierno de Canarias.

Paso 2. Realización de los modelos de distribución

Hemos utilizado la técnica de Máxima Entropía o MaxEnt (Phillips *et al.* 2004, 2006) por dos razones principales: i) porque se supone que es una de las técnicas que produce predicciones más fiables cuando sólo se dispone de datos de presencia (Elith *et al.* 2006), y ii) porque nos permite hacer todos los modelos para las 841 especies al mismo tiempo y de una manera relativamente rápida. Además, esta técnica está implementada en un software gratuito, libre y de fácil manejo (Maximum Entropy Species Distribution Modelling v.2.3, disponible en <http://www.cs.princeton.edu/~schapire/maxent/ >). Así, sólo hay que incluir las variables en el formato adecuado y seleccionar el tipo de respuesta para las variables explicativas en relación a los datos biológicos. Nosotros elegimos las respuestas linear y cuadrática. Para los otros parámetros del modelo mantuvimos los valores recomendados por defecto. El resultado final son unos mapas de favorabilidad que varían entre 0 y 100 para cada una de las 841 especies.

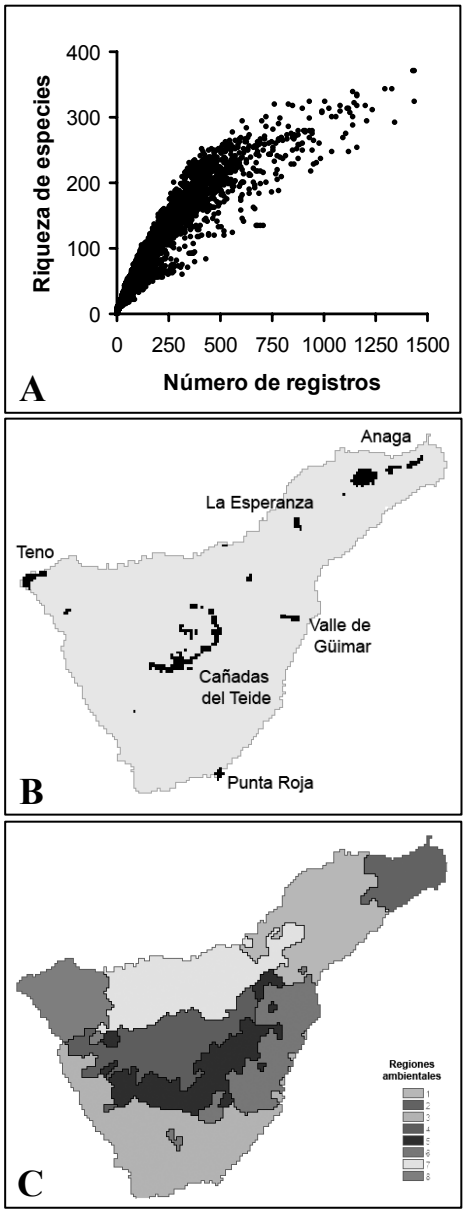

Figura 1. Relación entre el número de especies y el esfuerzo de muestreo (número de registros) realizado en cada uno de los sitios de Tenerife (A), y selección de las cuadrículas de 500 m (n = 264) que presentan inventarios fiables (B). También se muestra la regionalización ambiental de Tenerife (C) descrita por Hortal *et al.* (2007). *Relationship between the number of species and the sampling effort (number of records) in all sites of Tenerife (A), and selection of 500 m grid-cells (n = 264) with reliable inventories. The environmental regionalization of Tenerife described by Hortal et al. (2007) is also shown (C).*

Paso 3. Superponiendo las distribuciones individuales

Para hallar la riqueza de especies bastaría con sumar los resultados de cada uno de los modelos individuales. Sin embargo, MaxEnt genera una variable continua y es necesario elegir un punto de corte para transformar las favorabilidades en una variable binaria (0/1). Para ello, elegimos 21 umbrales diferentes asumiendo que cada una de las especies se encontraba presente cuando los valores de favorabilidad fuesen iguales o mayores que 1, 5, 10, 15, …, 100. Además, también utilizamos como punto de corte el valor específico de cada especie que garantizaba la predicción correcta de todos sus puntos de presencia (el valor mínimo de favorabilidad existente en una localidad de presencia o VMF).

Posteriormente, superpusimos estos mapas a fin obtener una representación de la riqueza de especies y la composición florística existente en cada cuadrícula de 500 x 500 m. En total, obtuvimos 22 mapas de riqueza y composición correspondientes a los 22 puntos de corte.

3. Evaluando las predicciones

Para poder evaluar la capacidad predictiva de los modelos de distribución necesitamos un conjunto de sitios "reales" donde se conozcan las presencias y las ausencias de las especies. Cuando sólo se poseen datos de presencia, una posibilidad para estimar las localidades con inventarios fiables es analizar el esfuerzo de colecta realizado en cada localidad (Hortal *et al.* 2007). En nuestro caso, hemos asumido que aquellas celdas donde hay un gran número de registros (más de 50) y, además, 3 o más registros por especie, tienen mayor probabilidad de ofrecer inventarios fiables (Figura 1A). En total, 264 cuadrículas cumplen estos criterios (Figura 1B), así que seleccionamos dichos sitios para comprobar los patrones generales de riqueza y composición que predicen los modelos y para cuantificar el tipo y la magnitud de los errores que se producen en cada caso.

3.1. ¿Se predicen bien los patrones generales de riqueza y composición?

Hemos calculado las correlaciones entre los valores observados y predichos en las celdas con inventarios fiables, tanto para la riqueza de especies, aplicando el test de Spearman (r_s), como para la composición florística, aplicando el test de Mantel (R). Estos análisis nos han permitido conocer cuál es el punto de corte en los valores de favorabilidad que genera las predicciones relativas de riqueza y composición más parecidas a las reales. Así, las correlaciones no paramétricas de Spearman nos indican si la riqueza de especies observada y predicha muestran un patrón general de variación espacial similar, mientras que el test de Mantel nos ofrece una visión global de cómo varían las predicciones de especies entre los diferentes sitios en comparación a las variaciones en la estructura de la composición observada. Para calcular las correlaciones de Mantel primero hay que hallar dos matrices de similitud de especies entre sitios, una para los datos predichos y otra para los observados, y después testar la correlación entre ambas. Como medida de similitud utilizamos el índice de diversidad β-Simpson que, independientemente de la riqueza, da cuenta del recambio de especies o *turnover* entre sitios. Hicimos el test de Mantel con el software gratuito PAST v. 1.68 (Hammer *et al.* 2001).

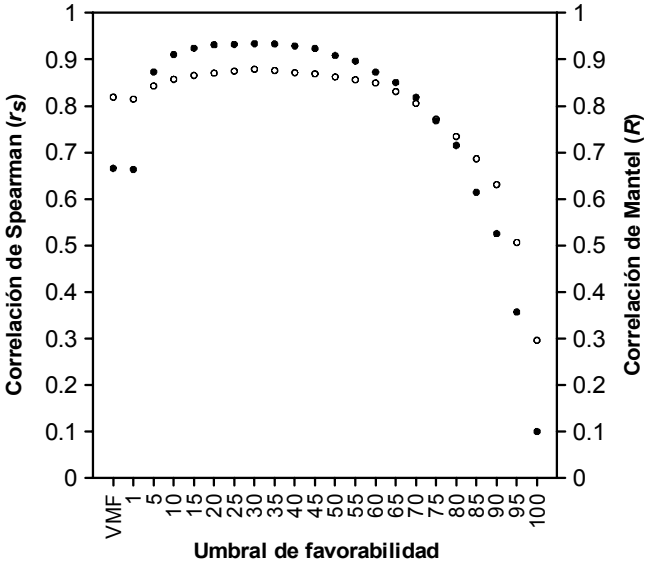

Figura 2. Correlaciones de Spearman entre la riqueza predicha y observada (círculos negros) y correlaciones de Mantel entre la composición predicha y observada (círculos blancos) en los diferentes umbrales utilizados para cortar los valores continuos de favorabilidad en valores binarios. El VMF es el valor mínimo de favorabilidad existente en una localidad de presencia. *Spearman correlations between observed and predicted species richness (dark circles) and Mantel correlations between observed and predicted species composition (open circles) at different thresholds used to cut continuous suitability values into binary ones. VMF is the lowest suitability value associated with an observed presence record.*

Según el punto de corte que se elija, las correlaciones entre los valores de riqueza predichos y observados oscilan entre 0,30 y 0,88, mientras que los valores de correlación para la composición de especies varían entre 0,10 y 0,93 (Figura 2). Utilizar umbrales de favorabilidad extremos, tanto en el límite inferior como en el superior, produjo patrones de riqueza y composición que varían ostensiblemente respecto a los "reales". Las predicciones más ajustadas para ambos casos se logran con puntos de corte situados entre valores de favorabilidad de 30 y 35 (r_s = 0,88; p < 0,0001 y R = 0,93; p < 0,0001). Elegir el VMF a fin de garantizar la predicción

correcta de todas presencias, parece producir resultados menos fiables (r_s = 0,82 y R = 0,67; n = 264, p < 0,0001).

En realidad, utilizar el VMF como punto de corte produce claras sobrepredicciones (Fig 3A). El intercepto de la relación lineal entre los valores observados y esperados de riqueza puede considerarse una medida de la sobrepredicción (153 especies), mientras que una pendiente significativamente mayor que la unidad (1,51 ± 0,06, t = 23,89, p < 0,0001) sugiere que estas sobrepredicciones son mayores cuanto mayor es la riqueza de las celdas. Aplicar el punto de corte que genera las distribuciones de riqueza y composición de especies más parecidas a la realidad (35; ver Figura 2) produce, en cambio, mejores resultados. En este caso, el intercepto de la relación lineal no difiere significativamente de cero (2,17 ± 5,24; t = 0,41, p = 0,7), mientras que la pendiente es sólo ligeramente inferior a la unidad (0.95 ± 0,03; t = 32,02, p < 0,0001), de modo que las celdas mas ricas en especies son levemente infrapredichas (Figura 3B).

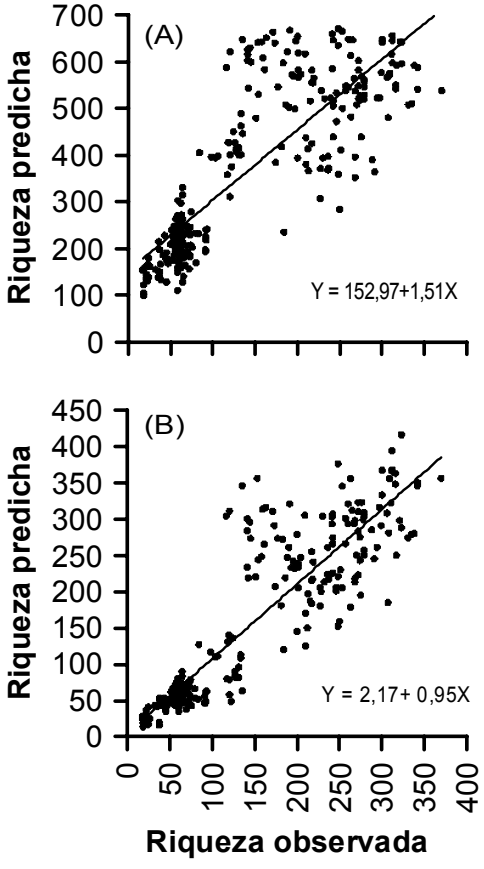

Figura. 3. Relación entre la riqueza predicha y observada usando como punto de corte el valor mínimo de favorabilidad existente en una localidad de presencia (VMF) (A), o usando el punto de corte que genera mejores correlaciones tanto para la riqueza como para la composición (B). *Relationship between predicted and observed richness using the lowest predicted value associated with an observed presence record (VMF) as threshold criteria to cut continuous suitability values (A), or using as threshold the suitability value that generates higher species richness and compositional correlations (B).*

3.2. ¿Cuál es la magnitud de los errores? ¿Qué nos desvela su distribución espacio-ambiental?

Una vez que hemos determinado el umbral de favorabilidad con el cual podemos predecir patrones coherentes de riqueza y composición, ahora queremos saber más específicamente qué errores hay detrás de estos patrones. Para ello, hemos calculado los errores de riqueza hallando la diferencia entre los valores predichos y los reales, y los errores de composición sumando los falsos negativos (presencias que se predicen incorrectamente como ausencias) y los falsos positivos (ausencias que se predicen erróneamente como presencias). Después, hemos dividido ambos cómputos por el número de especies observadas en cada sitio para reflejar la magnitud real del error en relación a las especies presentes.

Nuestros resultados demuestran que, aun habiendo elegido el modelo que ofrece mejores predicciones en los patrones generales de riqueza y composición, se siguen prediciendo más especies de las que hay realmente (el sitio con más sobrepredicción alcanza el 138%, y el mínimo error no es inferior al 67%). Así pues, utilizar solamente los datos de presencia para modelizar la distribución de las especies siempre parece sobreestimar la realidad. No obstante, en promedio, se pueden lograr predicciones aceptables de riqueza (4,5%, DE = 32). A pesar de ello, cuando examinamos específicamente estos errores, observamos que un 59,7% (DE = 21,9) de las especies no se predicen bien, y el error mínimo es del 29,6%. En definitiva, nuestros resultados muestran que los errores en composición son inevitables, a pesar de utilizar un gran conjunto de variables ambientales como predictores. Utilizar solamente datos de presencia para modelizar la distribución real de las especies generará predicciones globales de biodiversidad poco fiables, al menos, para la resolución y escala de trabajo aquí empleados.

Si los modelos de distribución fallan al predecir la riqueza y composición, ¿dónde ocurren dichos errores?, ¿suceden al azar o, por el contrario, se repiten sistemáticamente en ciertas regiones geográficas y/o en determinados ambientes? Para abordar esta cuestión, hemos regresado las "variables de error" (de riqueza y composición) separadamente frente a las variables ambientales y geográficas, utilizando Modelos Lineares Generalizados (MLG). Para construir los MLG (ambiental y espacial), hemos elegido la distribución de Poisson para las variables dependientes, y la función logarítmica para relacionarlas con las variables explicativas en cada caso. Seleccionamos iterativamente las variables independientes en orden creciente según la varianza que explican, y eliminamos aquellas que no contribuyen significativamente cuando se hace el modelo completo. Realizamos estos análisis en el software STATISTICA v.7 (StatSoft. Inc. 2004).

Los resultados de los MLG demuestran que los errores producidos en los modelos de distribución poseen una estructura ambiental y espacial significativa. El MLG ambiental explica un 20,2% de la varianza en el error de riqueza ($F_{4, 259} = 17{,}64$; $p < 0{,}0001$), y un 28,9% en el error de composición ($F_{7, 256} = 16{,}27$; $p < 0{,}0001$). Las variables más importantes en cada caso indican que las zonas que son geográficamente más accidentadas, o bien, aquellas donde apenas varía la pendiente, tienden a acumular mayores errores de sobrepredicción, y que en las zonas típicamente de inceptisoles (suelos con poco desarrollo horizontal, característicos de pendientes muy pronunciadas) se predice peor la composición de especies. El MLG espacial también explica un alto porcentaje de la varianza en los errores de riqueza (33,2%; $F_{5, 258} = 27{,}15$; $p < 0{,}0001$), debido principalmente a que el número de especies se sobreestima en las zonas más orientales. En el caso de la composición, las variables espaciales explican un 14,6% de la variabilidad ($F_{2, 261} = 23{,}41$; $p < 0{,}0001$), mostrando mayores errores en la parte nororiental de la isla.

4. Un método sencillo y rápido, ¿pero fiable?

Contar con un conjunto representativo de información fiable sobre las presencias y las ausencias de las especies es imprescindible para poder, al menos, validar las estimaciones de distribución real que obtenemos con los modelos (Fielding y Bell 1997, Vaughan y Ormerod 2003). Cuando tenemos muchos datos y son de buena calidad (presencias y ausencias) se pueden utilizar simplemente técnicas de validación cruzada que suponen reservar una parte de la información para hacer el modelo y la otra para validarlo. Sin embargo, cuando los datos son peores (sólo presencias), o cuando tenemos poca información, sacrificar parte de los datos para validar el modelo no es aconsejable. Aquí hemos explicado cómo poder estimar los sitios que tienen inventarios fiables, demostrando que no es suficiente con tener una gran cantidad de registros (como en nuestro caso, casi un millón de datos) sino que, además, se debe poseer una proporción significativa de datos para cada especie bien distribuida en el espectro ambiental y espacial del territorio.

Debido a que las presencias/ausencias reales de las especies no son directamente comparables con las predicciones continuas de favorabilidad obtenidas en MaxEnt, hemos de "cortar" necesariamente dichos valores en unos y ceros para poder validar los modelos. En general, se han recomendado diferentes puntos de corte o umbrales como, por ejemplo, utilizar la prevalencia de las especies (Liu *et al.* 2005). En nuestro caso, esto carece de sentido ya que sólo conocemos las presencias. También, entre los resultados que genera el propio software donde se elaboran los modelos con MaxEnt se proponen varios *thresholds* más o menos típicos, aunque algunos resultan poco comprensibles (ver *"balance" threshold;* también en el foro de discusión de MaxEnt <http://groups.google.com/group/Maxent>). Hemos demostrado que utilizar el VMF tampoco genera buenos resultados, tan sólo garantiza predecir bien las presencias observadas pero a costa de sobreestimar bastante el número real de especies observadas. Nuestro propósito con el procedimiento que hemos utilizado aquí ha sido, no tanto encontrar el mejor umbral de corte para cada especie en el sentido funcional, sino que (sólo a efectos metodológicos) hemos tratado de determinar aquel que proporciona mejores predicciones probando el mayor rango posible de puntos de corte.

Desafortunadamente, un relativo acierto en la predicción de la riqueza no implica estimar correctamente la composición. Cuando analizamos detalladamente los errores de predicción, nuestros resultados ponen en duda la fiabilidad y utilidad de este tipo de modelos. A raíz de estos resultados, nos hemos planteado la posibilidad de si estas predicciones, al menos, mejoran a escala regional. Asumimos, en este caso, que predecir incorrectamente una especie en las celdas contiguas o próximas a donde realmente está puede ser menos grave que hacerlo en zonas más alejadas. Para ello, hemos examinado los errores de predicción en las diferentes regiones ambientales de la isla previamente determinadas (Figura 1C, Hortal *et al.* 2007). Los resultados obtenidos muestran que en todas las regiones ambientales se llegan a predecir correctamente entre el 80-90% de las especies, y que los errores se deben fundamentalmente a la sobrepredicción (errores de comisión). Además, comprobamos que dichos errores son poco frecuentes (ocurren puntualmente en determinadas cuadrículas), de modo que el número medio de cuadrículas que presentan dichos errores de comisión oscila desde 5,5 ± 9.7 (±1DE), en la zona 4 (alrededor del 1% de su superficie total), hasta 31,9 ± 49,3 (±1DE), en la zona 2 (alrededor del 7% del área).

El hecho de que los errores de los modelos muestren una estructura en el espacio ambiental y geográfico pueden estar sugiriendo que: i) los sesgos que hay en la información biológica a nivel de especie influyen y se propagan cuando hacemos modelos a nivel de comunidad, o bien que ii) los errores debidos a factores que son más complejos de modelizar para una única especie (como elementos históricos, geográficos o bióticos), cobran un mayor peso cuando afectan a la comunidad entera. El mapa con la distribución de las localidades que poseen buenos inventarios demuestra claramente que los sitios mejor muestreados están sesgados hacia las zonas de interés natural o paisajístico, como el área central de las Cañadas del Teide o los macizos de Anaga y Teno, que aún poseen laurisilva relícta (Figura 1B). Al haber más información biológica en estas zonas, los modelos tenderán a sobreestimar el área de distribución de las especies allí presentes incluyendo zonas potencialmente favorables. Por otra parte, lograr predicciones mucho mejores cuando hacemos los modelos considerando regiones mayores (tanto de riqueza como de composición de especies) pone de manifiesto que hay otras "barreras" físicas o históricas, incluso bióticas, que no se pueden tener en cuenta localmente utilizando sólo los datos de presencia. Además, aparte de estos argumentos, no conviene olvidar que siempre existe el inconveniente metodológico de tener que elegir el mismo umbral de favorabilidad para todas las especies que, evidentemente, generará sobre- o subestimaciones en el número local de especies, según sea el caso para cada grupo de plantas. Es probable que estas diferencias se minimicen a escala regional debido a que las plantas también forman comunidades más o menos homogéneas para ambientes que también lo son.

La dificultad de conocer qué factores influyen en la distribución geográfica de cada especie es nuestro mayor inconveniente para realizar modelos predictivos fiables que representen la distribución real de los organismos. Este problema se acentúa cuando desconocemos la información de las ausencias que, por lo menos, nos ayuda a discriminar entre los factores que "favorecen" (fundamentalmente ambientales) y las fuerzas que "restringen" la distribución de las especies (como la capacidad dispersiva, los elementos históricos, factores demográficos, etc). Por ello, cuando se abordan aproximaciones como la presentada en este trabajo, donde se consideran, no una, sino muchas especies, cada una con sus requerimientos específicos, hay que adoptar un compromiso práctico para lograr predicciones fiables. En nuestro caso, tal compromiso fue asumir que podíamos utilizar una de las mejores técnicas que se basan en datos de "sólo presencia" para predecir la distribución real de todas las especies de espermatófitos en Tenerife. Siguiendo estas recomendaciones, hemos demostrado que hacerlo así tampoco solventa las limitaciones en la información biológica de partida y que debemos esforzarnos por conseguir datos mejores (más representativos), y por incluir variables que reflejen los factores que han limitado la colonización de determinadas áreas. En este estudio proponemos un protocolo sencillo para estimar la incertidumbre asociada a estos modelos analizando primero las predicciones de los patrones generales de distribución, y calculando posteriormente los errores que se producen específicamente bajo dichos patrones. Visualizar la estructura espacial de los errores a diferentes escalas demuestra que existen otros factores además de los puramente

ambientales que afectan localmente a la distribución de las especies. Finalmente, también sugerimos un método simple para reducir las sobrepredicciones en los modelos predictivos de distribución corrigiendo por el intercepto y la pendiente de la recta cuando se regresan los valores de riqueza predichos frente a los valores observados.

Agradecimientos

Agradecemos al proyecto BIOTA-Canarias la información aportada. Este artículo ha sido apoyado por el proyecto del MEC (CGL2004-04309) y el proyecto de la Fundación BBVA. SC trabaja con una beca de la DRCT (M311/I009A/2005).

Bibliografía

Araújo, M. B. y Guisan, A. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33: 1677-1688.

Bahn, V. y McGill, B.J. 2007. Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography,* 16: 733-742.

Barbosa, M., Real, R, y Vargas, J.M. 2009. Transferability of environmental favourability models in geographic space: The case of the Iberian desman (*Galemys pyrenaicus*) in Portugal and Spain. *Ecological Modelling* (in press).

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, McC. J., Peterson, T. A., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S. y Zimmermann, N. E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29: 129-151.

Fielding, A. H. y Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24: 38-49.

Guisan, A. y Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135: 147-186.

Hammer, O., Harper, D. A. T. y Ryan, P. D. 2001. PAST: Palaeontological Statistics software package for education and data analysis. *Palaeontologia Electronica*, 4: 1-9.

Hanski, I. 1998. Metapopulation dynamics. *Nature*, 396: 41-49.

Hortal, J., Lobo, J. M. y Jiménez-Valverde, A. 2007. Limitations of Biodiversity Databases: Case Study on Seed-Plant Diversity in Tenerife, Canary Islands. *Conservation Biology*, 21: 853-863.

Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, 117: 847-858.

Izquierdo, I., Martín, J. L., Zurita, N. y Arechavaleta, M. 2005. Lista de Especies Silvestres de Canarias (Hongos, Plantas y Animales Terrestres), 2nd ed. Consejería de Política Territorial y Medio Ambiente, Gobierno de Canarias.

Jiménez-Valverde, A., Lobo, J. M. y Hortal, J. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, 14: 885 – 890.

Kadmon, R., Farber, O. y Danin, A. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, 14: 401-413.

Liu, C., Berry, P. M., Dawson, T. P. y Pearson, R. G. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28: 385-393.

Mckenzie, D. I., Bailey, L. L. y Nichols, J. D. 2004. Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology*, 73: 546-555.

Peterson, T. A. 2006. Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*, 3: 59-72.

Phillips, S. J., Dudík, M. y Schapire, R. E. 2004. A maximum entropy approach to species distribution modeling. En: Proceedings of the 21st International Conference on Machine Learning ACM Press, pp. 665-662.

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

Pulliam, H. R. 1988. Sources, sinks and population regulation. *American Naturalist*, 132: 652-661.

Pulliam, H. R. 2000. On the relationship between niche and distribution. *Ecology Letters*, 3: 349-361.

Ricklefs, R. E. y Schluter, D. 1993. Species Diversity in Ecological Communities. Historical and Geographical Perspectives. University Chicago Press.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*,10: 1115-1123.

Soberón, J. y Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, 2: 1-10.

StatSoft. Inc. 2004. STATISTICA (data analysis software system), version 7. www.statsoft.com.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. y Kadmon, R. 2007. A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, 13: 397–405.

Vaughan, V. y Ormerod, S. J. 2005. The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, 42: 720-730.