Universidad Autónoma de Madrid
Departamento de Biología
Facultad de Ciencias

Consejo Superior de Investigaciones Científicas
Departamento de Biodiversidad y Biología Evolutiva
Museo Nacional de Ciencias Naturales

TESIS DOCTORAL

Modelos predictivos aplicados a la conservación de invertebrados protegidos ibero-baleares

Rosa María Chefaoui Díaz

Madrid, Mayo de 2010

Universidad Autónoma de Madrid

Departamento de Biología
Facultad de Ciencias

Consejo Superior de Investigaciones Científicas

Departamento de Biodiversidad y Biología Evolutiva
Museo Nacional de Ciencias Naturales

MODELOS PREDICTIVOS APLICADOS A LA CONSERVACIÓN DE INVERTEBRADOS PROTEGIDOS IBERO-BALEARES

Memoria presentada por ROSA MARÍA CHEFAOUI DÍAZ *para optar al Grado de Doctor en Ciencias Biológicas*

Rosa María Chefaoui Díaz

Vº Bº Director de Tesis	Vº Bº Director de Tesis	Vº Bº Tutor de Tesis
Dr. Jorge Miguel Lobo	Dr. Joaquín Hortal Muñoz	Dr. Enrique García-Barros

Madrid, Mayo de 2010

A mis abuelos

AGRADECIMIENTOS

En primer lugar, merecen toda mi gratitud mis directores, pues sin ellos esta Tesis no habría sido posible, me dieron toda la ayuda e impulso posible. Gracias a Jorge Lobo, pues a la vez de brindarme su confianza para trabajar con él y animarme desde el principio, me ha guiado durante este tiempo. No deja de sorprenderme su brillantez científica, por lo que espero seguir aprendiendo de él mucho más. Pero sobre todo, ha sido también un amigo y siempre grata su compañía. Con Joaquín Hortal aprendí el uso de los sistemas de información geográfica, pero eso es lo mínimo que aprendí de él, tanto en el terreno científico como personal. Sin su estimable asesoramiento científico y su disposición permanente a aclarar mis dudas esta Tesis no habría sido posible. Muchas gracias a ambos por todo el tiempo que me habéis prestado y por vuestra amistad, he tenido mucha suerte de haberos conocido.

Seguidamente quisiera agradecer la labor de mi tutor, Enrique García-Barros, siempre amable, atento y presto a colaborar. Gracias por tus aportaciones académicas, minuciosas correcciones, apoyo logístico en las labores administrativas y agradables conversaciones.

Gracias a todos los investigadores que he conocido en el Museo durante estos años, por su simpatía y ayuda. A Alberto que me ayudó con el R, Sara y Silvia que siempre tienen una sonrisa, Jesús y sus cenas, Tere, Belén, Marisa…

Debo dar las gracias a Alexander Hirzel, que me resolvió muy amablemente ciertas dudas sobre Biomapper.

También a los compañeros de la enseñanza que he tenido el placer de conocer durante estos años y que han sido mis otros colegas: Manolo, Begoña, Marga, Laura,

Inma, Damiana, Blanca, Josefina, Chari, Jose, Juan Diego, Isabel..., no puedo poner a todos, pues son muchos; he aprendido mucho de ellos y sus conversaciones han sido muy enriquecedoras. Por otra parte, reconozco que los alumnos me han dado cada día esa chispa de juventud necesaria para no perderme demasiado en el mundo de los adultos.

No puedo dejar de agradecer a mis amigos de bata blanca su estupenda labor profesional y humana. Ellos han sido también determinantes para la finalización de esta Tesis.

Gracias a todos mis amigos, a los de siempre, a los que veo más, a los que veo menos… todos habéis estado cerca en algún momento para escuchar, hacerme reír, enseñarme nuevos lugares y, en definitiva, hacer la vida más bonita. No busco hacer una relación pormenorizada de todos aquellos a los que he de agradecer su apoyo, su presencia en un momento difícil o, simplemente su sonrisa; pues sería imposible.

Esta Tesis está especialmente dedicada a mis abuelos, a los que sigo queriendo mucho y echo de menos. Por todo lo que me dieron. Y, por supuesto, al resto de la familia, que me ha apoyado en momentos difíciles y a los que quiero mucho, un abrazo a todos.

A Cora, mi compañera inseparable que tanto da a cambio de tan poco.

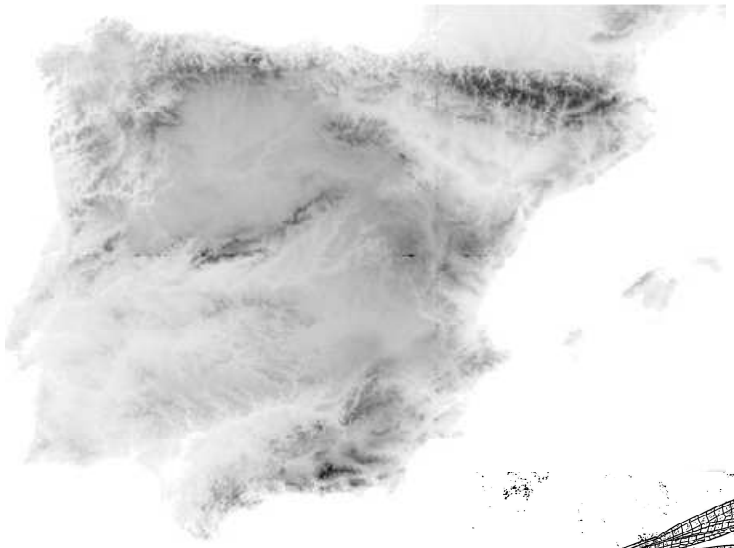
Y, por último, a Juanan, por su cariño, conversaciones y hacer que cada día merezca la pena. Muchas gracias por tu apoyo y paciencia.

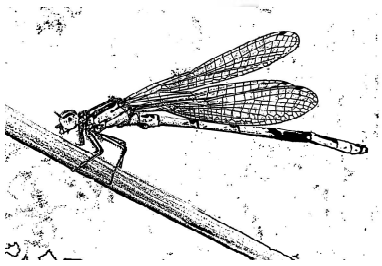
Rosa

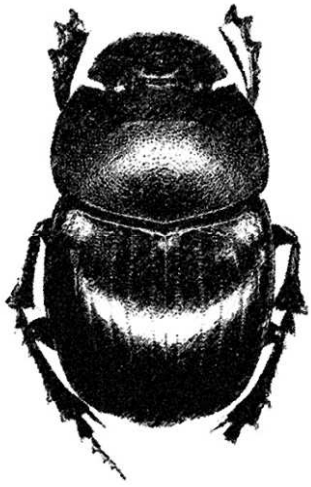
ÍNDICE

INTRODUCCIÓN Y ESTRUCTURA DE LA TESIS DOCTORAL

CAPÍTULOS







INTRODUCCIÓN

INTRODUCCIÓN

Ante la rápida desaparición de hábitats y especies, la necesidad de conservar la biodiversidad se enfrenta con la imposibilidad de inventariar y proteger todas las especies individualmente. Por un lado, nuestro conocimiento sobre la diversidad del planeta continua siendo insuficiente y son muchas las especies aún por describir ("Linnean Shortfall"; Brown & Lomolino, 1998) y asimismo, desconocemos en gran medida la distribución local, regional o global de numerosos taxones ("Wallacean Shortfall"; Lomolino, 2004). El sesgo en el desconocimiento aumenta según los organismos sean más pequeños y complejos (Medellín & Soberón, 1999; Ødegaard, *et al.*, 2000; Whittaker *et al.*, 2005) por lo que los taxones animales peor conocidos suelen ser grupos de invertebrados. Resulta por lo tanto esencial identificar las especies amenazadas y describir su distribución, mediante planteamientos de trabajo que no resulten inabarcables por falta de tiempo o presupuesto, problemas recurrentes con los que se enfrenta la planificación sistemática de la conservación.

Durante las últimas décadas, se ha intentado priorizar la selección de reservas mediante la identificación de "hotspots" o zonas de máxima riqueza. Sin embargo, frecuentemente éstas no coinciden para taxones diferentes, las especies raras no están presentes en ellas o la congruencia entre los distintos índices de biodiversidad es baja (Prendergast *et al.*, 1993; Orme *et al.*, 2005; Grenyer *et al.*, 2006). La utilización del criterio de complementariedad para la búsqueda de "huecos" ("gap analysis"; Scott *et al.*, 1993) en la red de espacios protegidos es mucho más efectiva (Williams *et al.*, 1996; Kati *et al.*, 2004) y su utilización en el caso de la Península Ibérica demuestra que se necesitan áreas protegidas adicionales a las existentes para la conservación efectiva de la diversidad de plantas y vertebrados (Araújo *et al.*, 2007). Aunque es necesario realizar

un ejercicio similar en el caso de los invertebrados, datos provisionales recientemente publicados sugieren que nuestra actual red de reservas sería poco efectiva a la hora de representar estas especies (Verdú & Galante, 2009). Desafortunadamente, aunque la mayoría de las especies son invertebrados, éstos son comúnmente olvidados pues su conservación cuenta con serias dificultades: i) es complejo elaborar listados e inventarios de especies a proteger debido a su elevada diversidad; ii) requieren mayor esfuerzo de muestreo que los vertebrados y su posterior identificación por expertos es muy laboriosa; iii) debido a su tamaño, la escala de estudio se refiere a menudo a microhábitats difíciles de detectar y iv) se desconocen los ciclos biológicos de la mayoría de las especies, así como su ecología y su distribución (ver New, 1998). Para conseguir una mejor protección de los invertebrados es necesario compensar estos obstáculos mediante estrategias que nos aporten la información indispensable acerca de las especies amenazadas (distribución, requerimientos ambientales, exigencia de recursos, etc.) en un plazo de tiempo aceptable y poder determinar la idoneidad de los hábitats a proteger.

Los Sistemas de Información Geográfica (SIG) han supuesto un avance significativo para la conservación de especies amenazadas pues permiten integrar información geográficamente referenciada, tanto de datos ambientales como biológicos, y pueden solventar muchas de las dificultades unidas al estudio de los invertebrados. Los modelos predictivos obtenidos mediante SIG nos permiten delimitar la distribución potencial de las especies (p. e., Dennis & Hardy, 1999; Reutter *et al.*, 2003; Hortal *et al.*, 2005); controlar sus poblaciones (p. e., Allen *et al.*, 2001; Davies *et al.*, 2005; Westerberg & Wennergren, 2003); analizar su nicho (p. e., Peterson *et al.*, 2002; Hirzel *et al.*, 2002; Cassinello *et al.*, 2006); diseñar redes de espacios protegidos (p. e., Cabeza

et al., 2004; Pearce & Boyce, 2006); realizar previsiones de futuro (p. e., Iverson & Prasad, 1998; Hill *et al.*, 2002), etc. Conjuntamente, las Bases de Datos tomadas de atlas, museos y herbarios se han revelado como una fuente de información muy valiosa para obtener datos de presencia de las especies (Soberón & Peterson, 2004; Gaubert *et al.*, 2006; Elith & Leathwick, 2007). Sin embargo, estos datos de origen heterogéneo, además de poder contener errores o proceder de muestreos sesgados (Hortal *et al.*, 2007, 2008; Newbold, 2010), no suelen aportar ausencias fiables, necesarias para poder realizar modelos predictivos coherentes (Anderson *et al.*, 2003; Loiselle *et al.*, 2003; Lobo *et al*., 2007), por lo que se han buscado alternativas generando modelos basados exclusivamente en presencias (Hirzel *et al.*, 2002; Pearce & Boyce, 2006) o en pseudo-ausencias obtenidas de diferentes maneras (Zaniewski *et al.*, 2002; Engler *et al.*, 2004; Lobo *et al.*, 2006, 2010). No obstante, elaborar modelos sin ausencias fiables tiene sus contrapartidas, pues se pierde información relevante acerca de los factores que limitan la distribución de las especies (Jiménez-Valverde *et al.*, 2008). Por tanto, resulta necesario estudiar las posibilidades que brindan estas técnicas para obtener aproximaciones a la distribución de las especies más cercanas a la real o la potencial en función del modo de obtención de las pseudo-ausencias. ¿Varían las predicciones de distribución según la ubicación de estos datos de ausencia?, ¿cuáles son la verdaderas posibilidades de estas pseudo-ausencias?

La Directiva 92/43/CEE (Directiva Hábitats) propone y regula la creación de una red de espacios naturales protegidos única para la Unión Europea, denominada Red Natura 2000. Según esta directiva, la Red Natura 2000 debe mantener un régimen especial de protección que asegure un estado de conservación favorable de los hábitats y las especies, especialmente las incluídas en sus Anexos II, III y IV. En concreto, su

Anexo II incluye un conjunto de especies de artrópodos considerados de interés comunitario para su conservación. Dentro del contexto de España, las Comunidades Autónomas son las encargadas de implementar la Red Natura 2000 dentro de sus respectivos territorios. Por tanto, cada una de ellas debería poner en práctica medidas de protección y, eventualmente, seguimiento de todas las especies incluídas en dicho Anexo II, siendo necesarios los estudios enfocados a una mejora del conocimiento sobre su distribución, requerimientos ambientales y estimas de sus poblaciones.

En esta tesis se evalua la utilidad de los modelos de distribución potencial de especies, elaborados mediante la combinación de datos ambientales en formato SIG y de datos de presencia obtenidos de museos, atlas y bases de datos, para la conservación de invertebrados amenazados en la Península Ibérica. Se hará uso de técnicas predictivas que utilizan exclusivamente presencias, como ENFA (Ecological Niche Factor Analysis) y MDE (modelo de envoltura ambiental), junto a otras que introducen en su cálculo presencias y ausencias (en este caso, pseudo-ausencias): GAM (Modelos Aditivos Generalizados), GLM (Modelos Lineales Generalizados) y NNET (Modelos de Redes Neuronales).

OBJETIVOS

El objetivo general de esta tesis es contestar al siguiente interrogante: ¿Es posible mejorar el estado de conservación de los invertebrados protegidos Ibero-Baleares mediante la aplicación de modelos predictivos? Para abordar este objetivo general se plantean los siguientes objetivos específicos:

1. ¿Podemos conocer de una manera fiable la distribución potencial de los invertebrados aunque no dispongamos de datos procedentes de muestreos exhaustivos?

El primer objetivo específico de la tesis consiste en:

Evaluar el comportamiento de distintas técnicas predictivas empleadas para la modelización de la distribución potencial de invertebrados usando datos de presencia disponibles en museos, atlas y bases de datos.

2. ¿Aportan los modelos predictivos información relevante sobre la ecología de las especies?

El siguiente objetivo será:

Determinar el nicho ambiental ocupado por una especie y las variables que afectan en mayor medida a su presencia mediante modelos predictivos.

3. ¿Pueden las características particulares de las especies afectar a los modelos?

Este interrogante determina el tercer objetivo específico:

Evaluar los efectos que tanto los datos como las características ecológicas de las especies pueden causar en la precisión de los modelos de distribución de invertebrados.

4. ¿Qué información biogeográfica pueden aportar los modelos de distribución?

Se plantea el siguiente objetivo:

Buscar explicaciones sobre la dinámica de la distribución de las poblaciones de determinadas especies coherentes con las predicciones obtenidas.

5. ¿Se encuentran suficientemente protegidas estas especies?

Esta tesis tiene también como objetivo:

Evaluar el estado de conservación de determinadas especies de invertebrados protegidos examinando la distribución de sus poblaciones y la eficacia de las reservas para preservarlas.

ESTRUCTURA DE LA TESIS

Esta tesis doctoral consta de cuatro capítulos:

I.- *Modelos de distribución potencial, caracterización del nicho y evaluación del estado de conservación usando herramientas SIG: el estudio de las especies ibéricas de Copris.*

En este capítulo se delimita la distribución potencial de las dos especies del género *Copris* que habitan la Península Ibérica (Coleoptera, Scarabaeidae) mediante ENFA (Ecological Niche Factor Análisis), un método que compara los valores ambientales de las localidades en las que ha sido observada una especie respecto a los valores ambientales del territorio analizado. Se explora el nicho ambiental ocupado por cada especie en una región pequeña (la Comunidad de Madrid), a fin de restringir el papel de los limitantes de dispersión discriminando las posibles áreas de co-ocurrencia e identificando las características ambientales específicas de cada especie. Se estudia también el grado de protección actual de las poblaciones clave de *C. hispanus* y *C. lunaris*, realizándose una propuesta para mejorar su conservación.

II.- *Evaluación de los efectos del uso de pseudo-ausencias en los modelos predictivos de distribución.*

Se estudia cómo influye el tipo de procedimiento para obtener pseudo-ausencias en los mapas predictivos generados. Se compararán las variaciones que sufren los modelos obtenidos para una especie enblemática, *Graellsia isabelae*, según la localización y selección de éstas: al azar dentro de la región de estudio o a diferentes distancias

ambientales por fuera de las regiones con condiciones, a priori, climáticamente favorables.

III.- *Evaluación de las variables ambientales más relevantes para explicar la distribución de* Graellsia isabelae *y delimitación de áreas importantes para su conservación.*

En este capítulo se realiza la modelización de la distribución potencial de una especie de insecto protegida mediante el empleo de pseudo-ausencias y Modelos Lineales Generalizados (GLM). Se estima la distribución potencial de *Graellsia isabelae* en la Península Ibérica y se identifican las variables que afectan en mayor medida a su distribución. Analizamos la posibilidad de conectividad y fragmentación de sus poblaciones así como el grado de protección de la especie respecto a los Lugares de Interés Comunitario (LICs).

IV.- *Efectos de las características ecológicas y de los datos en el comportamiento de los modelos de distribución de especies de invertebrados protegidos.*

En este capítulo se evalúan los efectos que ejercen las características ecológicas y biogeograficas de las especies, concretamente el número de observaciones y su dispersión en la región de estudio, sobre la precisión de los modelos, con el objeto de contribuir al conocimiento de la predicción de especies que, como los invertebrados, tienen características metodológicas y ecológicas muy heterogéneas.

REFERENCIAS BIBLIOGRÁFICAS

Allen, C. R., Pearlstine, L. G. & Kitchens, W. M. 2001. Modeling viable mammal populations in gap analyses. *Biological Conservation* **99**(2), 135-144.

Anderson, R. P., Lew, D. & Peterson, A. T. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling* **162**, 211-232.

Araújo, M. B., Lobo, J. M. & Moreno, J. C. 2007. The Effectiveness of Iberian Protected Areas in Conserving Terrestrial Biodiversity. *Conservation Biology* **21**(6), 1423-1432.

Brown, J. H. & Lomolino, M. V. 1998. *Biogeography*, 2nd edn. Sinauer Press, Sunderland, Massachusetts.

Cabeza, M., Araújo, M. B., Wilson, R. J., Thomas, C. D., Cowley, M. J. R. & Moilanen, A. 2004. Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology* **41**, 252-262.

Cassinello, J., Acevedo, P., & Hortal, J. 2006. Prospects for population expansion of the exotic aoudad (*Ammotragus lervia*; Bovidae) in the Iberian Peninsula: clues from habitat suitability modelling. *Diversity and Distributions* **12**, 666-678.

Davies, Z. G., Wilson, R. J., Brereton, T. M. & Thomas, C. D. 2005. The re-expansion and improving status of the silver-spotted skipper butterfly (*Hesperia comma*) in Britain: a metapopulation success story. *Biological Conservation* **124**(2), 189-198.

Dennis, R. L. H. & Hardy, P. B. 1999. Targeting squares for survey: predicting species richness and incidence of species for a butterfly atlas. *Global Ecology & Biogeography* **8**(6), 443-454.

Elith, J. & Leathwick, J. R. 2007. Predicting species' distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* **13**(3), 165-175.

Engler, R., Guisan, A. & Rechsteiner, L. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* **41**(2), 263-274.

Gaubert, P., Papes, M. & Peterson, A. T. 2006. Natural history collections and the conservation of poorly known taxa: Ecological niche modeling in central African rainforest genets (*Genetta* spp.). *Biological Conservation* **130**(1), 106-117.

Grenyer, R., Orme, C. D. L., Jackson, S. F., Thomas, G. H., Davies, R. G., Davies, T. J., Jones, K. E., Olson, V. A., Ridgely, R. S., Rasmussen, P. C., Ding, T. S., Bennett, P. M., Blackburn, T. M., Gaston, K. J., Gittleman, J. L., & Owens, I. P. F. 2006. Global distribution and conservation of rare and threatened vertebrates. *Nature* **444**, 93-96.

Hill, J. K., Thomas, C. D., Fox, R., Telfer, M. G., Willis, S. G., Asher, J. & Huntley, B. 2002. Responses of butterflies to twentieth century climate warming: implications for future ranges. *Proceedings of the Royal Society* **269**(1505), 2163-2171.

Hirzel, A., Hausser, J., Chessel, D. & Perrin., N. 2002. Ecological-Niche Factor Analysis: How to compute habitat-suitability maps without absence data? *Ecology* **83**(7), 2027-2036.

Hortal, J., Borges, P. A. V., Dinis, F., Jiménez-Valverde, A., Chefaoui, R. M., Lobo, J. M., Jarroca, S., Brito de Azevedo, E., Rodrigues, C., Madruga, J., Pinheiro, J., Gabriel, R., Cota Rodrigues, F. & Pereira, A. R. 2005. *Using ATLANTIS - Tierra 2.0 and GIS environmental information to predict the spatial distribution and habitat suitability of endemic species*. Direcção Regional de Ambiente and Universidade dos Açores, Horta, Angra do Heroísmo and Ponta Delgada, Horta, Faial.

Hortal, J., Lobo, J. M., & 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., & Baselga, A. 2008. Historical bias in biodiversity inventories affects the observed realized niche of the species. *Oikos* **117**, 847-858.

Iverson, L. R. & Prasad, A. M. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs* **68**(4), 465-485.

Jiménez-Valverde, A., Lobo, J. M. & Hortal, J. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions* **14**(6), 885-890.

Kati, V., Devillers, P., Dufrêne, M., Legakis, A., Vokou, D. y Lebrun, P. 2004. Hotspots, complementarity or representativeness? Designing optimal small-scale reserves for biodiversity conservation. *Biological Conservation* **120**, 471-480.

Lobo, J. M., Verdú, J. R. & Numa, C. 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions* **12**(2), 179-188.

Lobo, J. M., Baselga, A., Hortal, J., Jiménez-Valverde, A., & Gómez, J. F. 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Diversity and Distributions* **13**, 772-780.

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. 2010. The uncertain nature of absences and their importance in species distribution modelling. *Ecography* doi:10.1111/j.1600-0587.2009.06039.x.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G. & Williams, P. H. 2003. Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology* **17**(6), 1591-1600.

Lomolino, M. V. 2004. Conservation biogeography. *Frontiers of Biogeography: new directions in the geography of nature* (ed. by M. V. Lomolino and L. R. Heaney), pp. 293-296. Sinauer Associates, Sunderland, Massachusetts.

Medellín, R. A. & Soberón, J. 1999. Predictions of mammal diversity on four land masses. *Conservation Biology* **13**, 143-149.

New, T. R. 1998. *Invertebrate surveys for conservation*. Oxford University Press, New York.

Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography* **34**, 3-22.

Ødegaard, F. Diserud, O. H., Engen, S. & Aagaard, K. 2000. The magnitude of local host specificity for phytophagous insects and its implications for estimates of global species richness. *Conservation Biology* **14**, 1182-1186.

Orme, C. D. L., Davies, R. G., Burgess, M., Eigenbrod, F., Pickup, N., Olson, V. A., Webster, A. J., Ding, T-S., Rasmussen, P. C., Ridgely, R. S., Stattersfield, A. J., Bennett, P. M., Blackburn, T. M., Gaston, K. J. & Owens, I. P. F. 2005. Global hotspots of species richness are not congruent with endemism or threat. *Nature* **436**, 1016-1019.

Pearce, J. & Boyce, M. S. 2006. Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology* **43**(3), 405-412.

Peterson, A. T., Ball, L. G. & Cohoon, K. P. 2002. Predicting distributions of Mexican birds using ecological niche modelling methods. *Ibis* **144**, E27-E32.

Prendergast, J. R., Quinn, R. M., Lawton, J. H., Eversham, B. C. & Gibbons, D. W. 1993. Rare species, the coincidence of diversity hotspots and conservation strategies. *Nature* **365**, 335-337.

Reutter, B., Helfer, V., Hirzel, A. H. & Vogel, P. 2003. Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography* **30**(4), 581-590.

Scott, J. M., Davis, F. W., Csuti, B., Noss, R., Butterfield, B., Groves, C., Anderson, H., Caicco, S., D'Erchia, F., Edwards, T. C., Ulliman, J. & Wright, R. G. 1993. Gap analysis: a geographic approach to protection of biological diversity. *Wildlife Monographs* **123**, 1-41.

Soberón, J. & Peterson, A. T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London B* **359**, 689-698.

Verdú, J. R. & Galante, E., eds. 2009. *Atlas de los Invertebrados Amenazados de España (Especies En Peligro Crítico y En Peligro)*. Direccion General para la Biodiversidad, Ministerio de Medio Ambiente, Madrid, 340 pp.

Westerberg, L. & Wennergren, U. 2003. Predicting the spatial distribution of a population in a heterogeneous landscape. *Ecological Modelling* **166**(1-2), 53-65.

Whittaker, R. J., Araújo, M. B., Jepson, P., Ladle, R. J., Watson, J. E. M., & Willis, K. J. 2005. Conservation biogeography: assessment and prospect. *Diversity and Distributions* **11**, 3-23.

Williams, P., Gibbons, D., Margules, C., Rebelo, A., Humphries, C., Pressey, R., 1996. A comparison of richness hotspots, rarity hotspots, and complementary areas for conserving diversity of British birds. Conservation Biology **10**, 155-174.

Zaniewski, A. E., Lehmann, A. & Overton, J. M. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* **157**(2-3), 261-280.

CAPÍTULOS

Modelos de distribución potencial, caracterización del nicho y evaluación del estado de conservación mediante herramientas SIG: el estudio de las especies ibéricas de *Copris*.

RESUMEN

Los escarabajos coprófagos desempeñan una función ecológica importante mediante el reciclado de materia orgánica en ecosistemas de pastizal; sin embargo, sus poblaciones se encuentran en declive, por lo que es importante su conservación. Para modelizar los nichos ambientales de *Copris hispanus* (L.) y *Copris lunaris* (L.) (Coleoptera, Scarabaeidae) en la Comunidad de Madrid (CM), se utilizaron los datos de presencia disponibles para estas especies y BIOMAPPER, una herramienta basada en SIG. Las distribuciones potenciales obtenidas para ambas especies se usaron para ejemplificar la utilidad de este tipo de metodologías en la valoración del estado de conservación de las especies, así como su capacidad de describir la simpatría potencial entre dos o más especies. Ambas especies, distribuidas a lo largo de un gradiente que abarca desde condiciones ambientales secas-mediterráneas hasta húmedas-alpinas, coinciden en zonas de moderadas temperaturas y precipitaciones medias anuales del norte de la Comunidad de Madrid. Las especies de *Copris* se encuentran mal conservadas por la red de espacios protegidos existente, pero la protección proporcionada por los nuevos espacios incluídos en la futura Red Natura 2000 mejorará el estado de conservación general de estas especies en la Comunidad de Madrid.

Palabras clave: *Copris*; Península Ibérica; conservación de escarabajos coprófagos; modelización del nicho mediante SIG; distribución de especies.

Este capítulo ha sido publicado como:

Chefaoui, R. M., Hortal, J. & Lobo, J. M. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation* **122,** 327-338.

Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species.

ABSTRACT

Dung beetle populations, in decline, play a critical ecological role in extensive pasture ecosystems by recycling organic matter; thus the importance of their conservation status. Presence data available for *Copris hispanus* (L.) and *Copris lunaris* (L.) (Coleoptera, Scarabaeidae) in Comunidad de Madrid (CM), and BIOMAPPER, a GIS-based tool, was used to model their environmental niches.

The so derived potential distributions of both species were used to exemplify the utility of this kind of methodologies in conservation assessment, as well as its capacity to describe the potential sympatry between two or more species. Both species, distributed along a Dry-Mediterranean to Wet-Alpine environmental conditions gradient, overlap in areas of moderate temperatures and mean annual precipitations in the north of CM. *Copris* are poorly conserved in the existing protected sites network, but protection provided by new sites included in the future Natura 2000 Network will improve the general conservation status of these species in CM.

Keywords: *Copris*; Iberian Peninsula; Dung beetle conservation; GIS predictive niche-modelling; Species distribution.

INTRODUCTION

Research has become increasingly focused on the extrapolation of species distribution from incomplete data to obtain reliable distribution maps most efficiently (Mitchell, 1991; Pereira & Itami, 1991; Buckland & Elston, 1993; Iverson & Prasad, 1998; Manel *et al*., 1999a, b; Parker, 1999; Peterson *et al*., 1999; Pearce & Ferrier, 2000; Vayssières *et al*., 2000; Hirzel *et al*., 2001; Guisan *et al*., 2002; Hortal & Lobo, 2002; Ferrier *et al.*, 2002). By processing environmental information and presence/absence data, several statistical methods can provide estimates of the probability of occurrence of a given species (Guisan & Zimmermann, 2000). However, since the absence of a species from a locality is difficult to demonstrate, and since faunistic atlases do not usually cover localities where sampling has failed to produce a capture, false absences can decrease the reliability of predictive models. As an alternative, species distribution prediction based on presence-only data has been developed (Busby, 1991; Mitchell, 1991; Walker & Cocks, 1991; Carpenter *et al*., 1993; Scott *et al*., 1993; Stockwell & Peters, 1999; Peterson *et al*., 1999; Hirzel *et al*., 2001, 2002; Robertson *et al*., 2001). Generally, these alternative methods delimit the environmental niche of species within a geographical area and with a given resolution by comparing the environmental distribution of all the cells with that of cells where the species has been observed.

Pinpointing the areas where appropriate environmental conditions exist to sustain species is vital for biogeographical and conservation studies. It allows identifying environmentally suitable regions still not colonized, or where the species has become extinct; then the contribution of unique historical or geographical factors to the shaping of the current distribution of a species can be judged. With regard to conservation, potential distribution area identification can help locate sites suitable for

reintroduction programs, or faunistic corridors, favouring success in regional conservation planning. Following this line of thought, niche-based modelling of potential distributions has been used recently to examine different ecological and evolutionary aspects, such as competition between phylogenetically related species (Anderson *et al.*, 2002) or variation in species' niche requirements through evolutionary time (Peterson *et al.*, 1999; Peterson & Holt, 2003).

To exemplify the use of these techniques for conservation and ecological purposes, the Ecological-Niche Factor Analysis method (ENFA; Hirzel *et al.*, 2001, 2002) was used to delimit the potential distributional areas for two *Copris* species (Coleoptera, Sacarabaeoidea) in central Spain. This genus is made up of around 70 large-size dung beetle species, three of them present in the Western Palaearctic region (Baraud, 1992), being two of them, *Copris lunaris* and *Copris hispanus*, present in the Iberian Peninsula (Martín-Piera, 2000). *C. lunaris*, widely distributed throughout the Palaearctic region, inhabits mainly northern and temperate Iberian localities below 1000 m. in altitude, although it does reach 1700 m. in the south (Fig. 1.1). On the contrary, *C. hispanus*, a western Mediterranean species, frequents the southern half of the Iberian Peninsula at a slightly lower altitude. Both species occur together rarely, as they do in the Comunidad de Madrid (CM), in the centre of the Iberian Peninsula.

Dung beetles play a key role in Mediterranean cattlegrazed traditional landscapes. They are responsible for most organic matter recycling (Martín-Piera & Lobo, 1995) and control dipterous populations (Hanski, 1991). However, Western European dung-beetle assemblages present several conservation problems related to intensive management of agriculture and farming activities (Martín-Piera & Lobo, 1995; Barbero *et al.*, 1999; Hutton & Giller, 2003). Their sensitivity to landscape

transformation (e.g., Verdú *et al.*, 2000; Verdú & Galante, 2002) and several cattle antibiotic treatments (e.g., ivermectins; Lumaret *et al.*, 1993; Hutton & Giller, 2003), has leaded to the proposal of making use of them as indicators for conservation evaluation (Halffter, 1998; Davis *et al.*, 2001; Andresen, 2003). Among them, large-sized dung beetles, such as *Copris* species, seem to be the most affected. Their European populations are declining, and even becoming extinct (e.g., rollers, see Lobo,

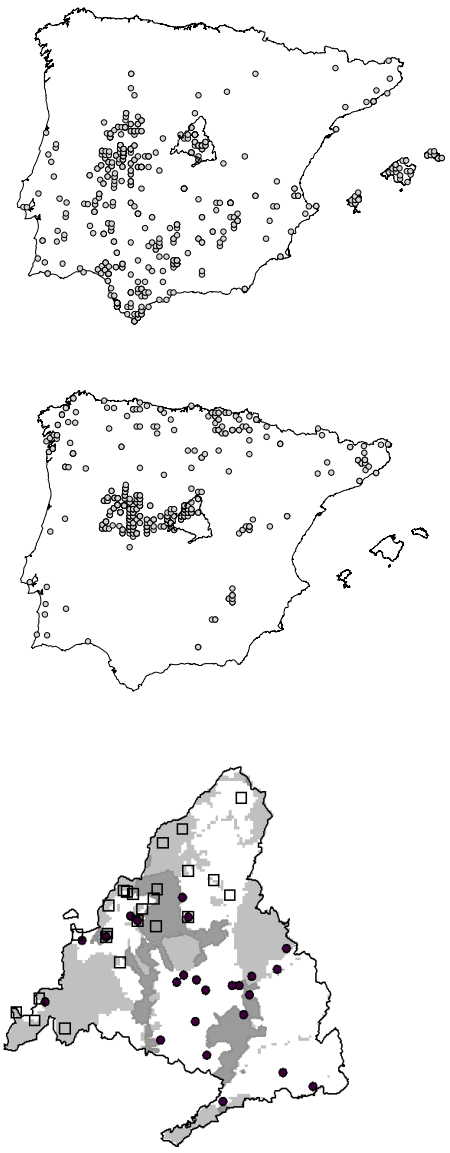


Fig. 1.1 - Presence of *Copris hispanus* (upper) and *C. lunaris* (centre) in the Iberian Peninsula and distribution of both species in the Comunidad de Madrid (lower). Squares represent *Copris lunaris* presences, and dark circles do for *Copris hispanus* ones. Dark shadow represents existing Protected Natural Sites, light shadow future Nature 2000 network sites.

2001). This is the case for *C. lunaris*, that has been declared critically endangered or extinct in the countries located at its northern range margins (see, e.g., Skidmore, 1991; Rassi et al., 1992; http://www2.dmu.dk/1_Om_DMU/2_Tvaerfunk/_3_fdc_bio/projekter/redlist/redlist_en.asp or http://www.daba.lu.lv/ldf/ CORINE/Insect.html).

Habitats Directive, the European Community initiative for a continental-scale network of protected areas (Natura 2000 Network), uses different kinds of habitats as conservation goals. In Spain, the selection and management of these areas has been led by the Autonomous Communities, which constitute administrative divisions with full environmental jurisdictional autonomy. In this paper, we explore how this habitat-based selection may be useful to preserve the populations of two sympatric dung beetle species in an area where their respective northern and southern range margins overlap. We have chosen these species as indicators of the conservation status of traditionally managed landscapes, one of the targets of Natura 2000 Network. In addition, we explore how these two species respond to a strong environmental gradient at the edge of their respective distributions.

Using all distributional information available for *C. hispanus* and *C. lunaris*, the environmental niche occupied by each species in CM was modelled and used to extrapolate their respective potential distributions. Environmental requirements of both species are reviewed to identify differences and similarities. From the maps so-obtained, most probable areas of joint occurrence are identified. Specific environmental conditions and habitat heterogeneity are considered as possible causes for co-occurrence, while taking into account the probability that competitive interactions may play a significant role in shaping local species distribution in these areas. The efficacy of existing protected natural sites (PNS) of Madrid, and also that of the complete set of

sites included in the Spanish proposal for Natura 2000 Network in the preservation of populations of both species is assessed. Finally, we identify key *Copris* population sites in Madrid.

MATERIALS AND METHODS

Study site

CM, an autonomous Spanish region with full jurisdiction over local environmental policy, also complies with Spanish and European policy (see Fig. 1.1). Its northern peak, Somosierra (latitude 41º 8' N) is 140 km from the most southernmost point (the Tajo valley, latitude 39º 52' N). Although its mean altitude is around 800 m, CM climate and topography vary, along with elevations, from 434 m in the Alberche valley, to the 2430 m of the Peñalara peak, in the Central System mountain range. Its geologic history, also very eventful, gave rise to considerable lithologic diversity, with acidic rocks in the mountains; alluvial deposits on mountain slopes, terraces and valleys; and calcareous rocks and clays, and even gypsum soils, in the southeast. Its diversity, together with strategic positioning in the centre of the Iberian Peninsula, has made of CM a region of transition between Mediterranean and Eurosiberian faunas (see Fernández-Galiano & Ramos-Fernández, 1987), an ideal region for small-scale pilot studies, as it is home to a synthesis of all inland Iberia.

Data sources

Biological data came from BANDASCA database, a compilation of all the information available in the bibliography and collections of natural history on the 53 Iberian species of the Scarabaeidae family (see structure in Lobo & Martín-Piera, 1991), as well as from

a number of standardized surveys. The most recent of these sampling campaigns was environmentally and spatially designed explicitly to account for the spatial patterns of biodiversity variations in the region (see Hortal, 2004 and Hortal & Lobo, 2005). After its results, it can be assumed that current presence records cover the main environmental and spatial patterns in both *Copris* species' distributions (for a detailed assessment of sampling effort and success see Hortal, 2004).

In this kind of geographically explicit analyses, the spatial resolution (grain size) constitutes a key decision for the accuracy and reliability of the obtained results. If cell size is larger than the area required to support a population, then the model will have very poor resolution. On the other hand, if it is too much small, then the model would present a high false prediction rate. Western Palaearctic temperate dung beetle populations have been estimated to have an approximate size of 1 km^2 (Roslin, 2000, 2001a, b; Roslin & Koivunen, 2001). A small scale study carried out in a semi-arid area of Central Spain gives support to a similar population size for Mediterranean species (Lobo et al., 2006). Thus, we have chosen 1 km^2 as the most appropriate spatial scale to carry out our analysis.

From the 72 database records available for *C. hispanus* and 111 for *C. lunaris* in the CM, only 24 reliable presence points could be obtained for each of the two species (see Fig. 1.1). The spatial resolution of most BANDASCA records, referred to the UTM 1 x 1 km grid, was also 1 km^2, so most of this biological information extracted from the database was directly used for the analyses. However, ten presence records for *C. lunaris* and seven for *C. hispanus* were limited to 10 x 10 km squares. We explored database information for each of these records (coming from museum specimen labels or from the literature). We thus assigned, where possible, their geographical position to

the 1 km^2 pixel placed nearest to the centroid of the 10 x 10 km grid square, that complies with the altitude and/or geographical information in the database. If, e.g., a record was referred to have an altitude of 750 m.a.s.l, and to pertain to the San Lorenzo de El Escorial territory in the 30TVK09 UTM 10 km cell, we located a presence in one of the pixels that comply with both characteristics, using the rule of thumb of selecting the closest to the centroid of the UTM cell. We assume that the error thus introduced is negligible, since both species are excellent flyers, as are almost all dung beetles.

Environmental data comes from CM-SIG, an environmental GIS database of CM (J. Hortal, unpublished; see Hortal, 2004), which contains the information of several variables relevant to the distribution of Scarabaeidae species. The richness and variation of Scarabaeidae assemblages in Western Europe has been formerly related with topography (Lobo *et al.*, 2002; Hortal *et al.*, 2003), climate (Lobo & Martín-Piera, 2002; Hortal *et al.*, 2001, 2003; Lobo *et al.*, 2002; Verdú & Galante, 2002), and soil composition (Hortal *et al.*, 2001, 2003). Thus, we have selected five variables to account for these factors, on the assumption that they constitute the most important environmental determinants of the distribution of *Copris* species in the studied region. Landscape structure variables are known to affect the microdistribution of dung beetles (i.e., at spatial scales smaller than 1 km^2), but were not considered for the environmental niche modelling procedure because such present-day land use variables are not adequate to model presence records from a large temporal resolution. However, land use information has been used to characterize the habitat heterogeneity of the areas potentially adequate for both species (see below). A Digital Elevation Model (DEM; map of elevations) was extracted from a global DEM with 1 km spatial resolution (Clark Labs, 2000a). Mean annual precipitation and mean annual temperature scores for 41

stations of Central Iberia (30-year monthly data) were obtained from an agroclimatic atlas (Ministerio de Agricultura, Pesca y Alimentación, 1986). We interpolated data from these points onto 1-km-spatial-resolution maps using a moving-average procedure (using a six-point search radius; see Clark Labs, 2000b). Maps of solar radiation and lithology (11-categories) were digitized from a CM Atlas (ITGE, 1988). Categories in the lithology map were reclassified into areas with stony acidic soils; with calcareous soils or deposits; and with acidic deposits. As ENFA does not work with multinomial data, we derived three maps of the proportion of each kind of soil category in the 5 x 5 km^2 window surrounding each pixel, using IDRISI32 Pattern module (Clark Labs, 2000b). Only the first two lithological variables were used, as the information from the third was redundant.

To ascertain if areas highly suitable for both species were more heterogeneous than the rest of the region, we also extracted five heterogeneity variables, three of them to take into account habitat heterogeneity. As steeper slopes are correlated with higher environmental variability, a slope map was calculated from the DEM using the GIS (Clark Labs, 2000b). A nine-categories aspect map was also derived from the DEM, and a land use map of 14-categories, obtained by reclassifying and enlarging the 250 m European Land Use/Land Cover map provided by the CORINE programme (European Environment Agency, 1996). For both maps, the Shannon diversity index of the cells within a 5 x 5 km window was calculated to obtain an aspect-diversity map and a land-use diversity map (Clark Labs, 2000b). Annual variation (i.e. temporal heterogeneity) in monthly precipitation and temperature were also calculated for each climatic station as the mean of the differences between monthly extreme values within each year, and then interpolated using the procedure referred to above.

Finally, we obtained additional vectorial cartography of CM administrative limits from the digital version of the CM 1:200.000 map (Servicio Cartográfico de la Comunidad de Madrid, 1996). Protected natural sites (PNS) and future Natura 2000 network sites were obtained from the ‘‘Banco de Datos de la Naturaleza’’ of the Spanish ‘‘Dirección General de Conservación de la Naturaleza’’ (see http://www.mma.es/bd_nat/menu.htm).

Data analysis

Potential Distribution Maps - Ecological-niche factor analysis (ENFA) was done using BIOMAPPER 2.1 software (Hirzel *et al.*, 2000; see http://www.unil.ch/biomapper). ENFA uses diverse environmental information to characterize the ecological distribution of the species. It computes a group of uncorrelated factors, summarizing the main environmental gradients in the region considered, similarly to common ordination techniques such as Principal Component Analysis. However, ENFA derives these factors using data only from known species presences (and absences, when available), thus providing factors with biological meaning. The first axis (marginality factor) is chosen to describe the marginality of the niche with respect to the regional environmental conditions, by maximizing the difference between the environmental mean value of the species' presences, and the global mean environmental value of all the studied region. The following axes (specialization factors), sorted according to their decreasing amounts of explained variance, are used to represent the species' degree of specialization in the rest of the (orthogonal) environmental gradients identified in the study area. Habitat suitability is modelled using the so-selected factors by estimating the ecogeographic degree of similarity between each grid square and the environmental

preferences of the species, that is, the probability that a given grid square belongs to the environmental domain of the presence observations. Thus, starting from a species presence map a potential distribution map takes on the form of a habitat suitability map (HSM) of values that vary from 0 (minimum habitat quality) to 100 (maximum). The distribution models obtained were validated by a Jackknife procedure, whereby each HSM was computed 24 times (the number of presence points of each species), leaving out one point of presence with each iteration. By this procedure one independent habitat suitability score for each presence point was obtained and the observed and estimated scores compared. For a more extensive explanation of the method see Hirzel *et al.* (2002).

Environmental and spatial characterization of the realized niche - The HSM maps obtained were reclassified as of: very low habitat suitability (0–25); low habitat suitability (26–50); high habitat suitability (51–75) or very high habitat suitability (76–100). These new maps were cross-tabulated in the GIS environment to pinpoint zones of spatial coincidence (very high/very high and very low/very low habitat suitability) and also of difference (very high for one species and very low for the other) for both species. By means of a Mann-Whitney U test (StatSoft, 1999), we extracted those environmental variables that characterize each of these four zones, because of being significantly different to the conditions in the rest of CM. In order to compare the environmental variability between the cells with a very high suitability value for both species and the remaining cells, another Mann-Whitney U test was carried out, taking into account the five heterogeneity variables described above.

Conservation status - The degree of protection of *C. hispanus* and *C. lunaris*, achieved by existing PNS, and to be achieved by future Natura 2000 network sites in CM, was evaluated by extracting minimum, maximum, mean and standard deviation of the suitability values for both species in each protected site. The area of the zones with very high suitability values for each species (HS > 75) and for both together was also located. To assess conservation status of each species we have used two different criteria: the mean suitability scores, and the area with very high suitability scores (HS > 75) per PNS.

RESULTS

Potential distributional maps

The six environmental variables considered were reduced to two factors for each species that explained a similar percentage of the variance: 96.7% for *C. hispanus* and 95.9% for *C. lunaris*, respectively. The first selected axis, which maximizes the absolute difference between global environmental mean and the species mean (the marginality factor), explains 74% of the specialization for *C. hispanus* and 72% for *C. lunaris* (see Hirzel *et al.*, 2002), (i.e. the ratio of the standard deviation between the global distribution and that of the species). These high percentages of specialization point out that the high importance of these first factors to explain both marginality and niche breadth of each one of the two species. The next factors (specialization factors) explain 19% and 18% respectively. Solar radiation and calcareous soils are the variables with higher marginality coefficients for *Copris hispanus*, showing that the scores of these variables in the presence cells differ from the mean values in the region (Table 1.1). As these coefficients are positive, this species is shown to prefer sunny areas and basic

soils. Mean annual precipitation has the higher coefficient of the specialization factor, showing that the distribution of *C. hispanus* in CM is specially restricted by this variable. In the case of *C. lunaris*, acid soils and mean annual precipitation are the variables related to the marginality factor, meaning a higher probability of presence in siliceous and rainy cells. The specialization of this species is mainly conditioned by the presence of calcareous soils, mid-to-high altitudes and high solar radiation scores. Marginality scores characterize how much each species' habitat differs from the conditions available in the study area (from 0, close to the mean, to 1, when it prefers habitats extreme in the region). Overall marginality value was higher than 0.65 for both species, evidencing a high separation of both species from the central part of the strong environmental gradient present in CM. *C. lunaris*, adapted to cold mountain environments (see below), which are more rare in the region, presented a very high marginality (0.91), whilst *C. hispanus*, more adapted to the intermediate environments of the marginal slopes of the mountains, presented lower values (0.68). On the other hand, the global tolerance values (the opposite of specialization ones) were 0.43 for *C. lunaris* and 0.17 for *C. hispanus*. The score for this species (close to 0) suggests that *C. hispanus* tends to live near mean regional conditions, and tolerates an smaller environmental range than does *C. lunaris*, which is adapted to conditions that are more extreme at CM.

Habitat suitability maps so obtained (Fig. 1.2) show a high probability of appearance for *C. lunaris* in north-western CM, while highest habitat suitability values for *C. hispanus*, distributed patchily across the region, are basically limited to the centre and southeast. Jackknife validation results for these HSMs indicate that the *C. hispanus* potential map is more reliable than the one for *C. lunaris*. A habitat suitability value

greater than 50 was found in 87.5% of 24 1 km^2 grid squares in the case of *C. hispanus* (SD=33.1 %), while in the case of *C. lunaris* such a suitability value was found in just 66.7% of the 24 1 km^2 cells with presence (SD=47.1 %).

Copris hispanus		***Copris lunaris***	
Marginality factor (74%)	*Specialization factor (19%)*	*Marginality factor (72%)*	*Specialization factor (18%)*
Solar radiation (0.80)	Precipitation (0.89)	Acid soil (0.69)	Calcareous soil (0.64)
Calcareous soil (0.45)	Acid soil (0.32)	Precipitation (0.51)	Altitude (0.50)
Acid soil (0.29)	Temperature (0.27)	Temperature (-0.34)	Solar radiation (0.50)
Altitude (-0.24)	Solar radiation (0.12)	Solar radiation (0.28)	Precipitation (0.22)
Temperature (0.11)	Altitude (0.11)	Altitude (0.26)	Temperature (0.15)
Precipitation (-0.06)	Calcareous soil (0.01)	Calcareous soil (-0.05)	Acid soil (0.12)

Table 1.1 - Specialisation explained by the two factors extracted by ENFA, and coefficient values of the six environmental variables used in the analysis. Positive values on the marginality factor mean that the species prefers localities with higher values regarding to the CM mean score. Variables with higher specialisation coefficients restrict more the distribution range of the species.

Environmental and spatial characterization of the realized niche

Reclassified and cross-tabulated habitat suitability maps show the areas of spatial coincidence and difference between both species (Fig. 1.3). The very highly suitable areas for both species are located in the north of CM, in the spurs of the ‘‘Sierra de Guadarrama’’ (Fig. 1.3a). These zones differ significantly from the rest of CM because of their higher altitudes (Mann-Whitney U test; $Z = 14.7$, $p < 0.001$), higher mean annual precipitations ($Z = 11.9$, $p < 0.001$), greater presence of stony acid soils ($Z = 13.83$, $p < 0.001$) and lower mean annual temperatures ($Z = 9.97$, $p < 0.001$). Three of the five variables considered as environmental heterogeneity surrogates also differ significantly between these coincidence areas and the rest of CM, which present higher

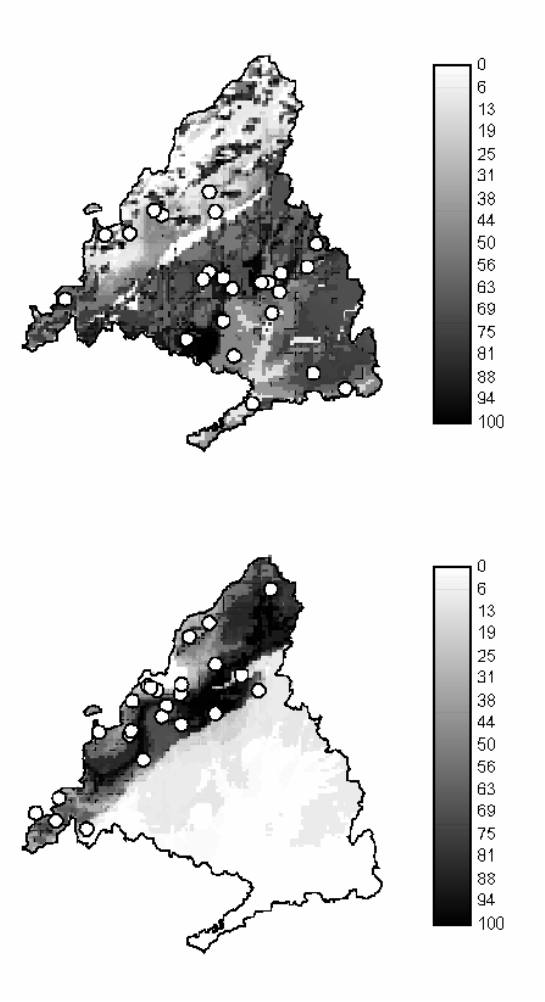


Fig. 1.2 - Habitat Suitability Maps for *C. hispanus* (upper) and *C. lunaris* (lower). The scale on the right shows the habitat suitability values (0= low suitability; 100= high suitability).

values of annual range of precipitation ($Z = 12.69$, $p < 0.01$) and slope ($Z = 12.56$, $p < 0.01$), and lower values of annual range of temperature ($Z = 12.37$, $p < 0.01$). On the contrary, landscape heterogeneity (aspect and land use diversity variables) was not significantly different between these areas and the rest of CM.

The zones with very poor suitability scores for both species are, on the one hand, Cotos, Navacerrada and ‘‘Sierra de Cuerda Larga’’, mountainous areas with altitudes higher than 1300 m; and on the other, low altitude quaternary terraces (around 600 m) of the rivers Jarama, Manzanares, Tajo and Tajuña; and also transition zones between stony acid soils of the sierra and acid deposits of the ‘‘ramp’’ (the southern slope of the

Guadarrama mountains; see Fig. 1.3b). Very low suitability areas also differ from the rest of CM in altitude ($Z = 6.9$, $p < 0.001$), the presence of stony acid soils ($Z = 4.9$, $p < 0.001$) and mean annual temperature ($Z = 3.6$, $p < 0.001$).

The areas in which the niches of both species do not coincide are markedly different. Areas of very high suitability for *C. lunaris* and very low for *C. hispanus* are found in the ‘‘Sierra de Guadarrama’’ (Fig. 1.3c), where all environmental variables considered are significantly different from the rest of CM ($p < 0.001$). Stony acid soil is predominant, altitude (1061.8 m ± 191.6 m) and mean annual precipitation (716.9 mm ± 50.5 mm) are higher than median values of CM, while solar radiation (4 kwh/m^2/día) and mean annual temperature (11.8 ºC ± 0.8 ºC) are lower. On the contrary, *C. hispanus* finds very-high suitability and *C. lunaris* very-poor suitability areas in the ‘‘ramp’’ of acid deposits, and also in a calcareous soil area between the rivers Tajo and Tajuña (Fig. 1.3d). In these areas, all the environmental variables considered differ significantly from the rest of CM ($p < 0.001$). Stony acid soil is less frequent in these areas, mean annual temperatures (13.9 ºC ± 0.5 ºC) and the solar radiation are higher, while altitude (636.5 m ± 87.6 m) and mean annual precipitation (459.5 mm ± 58.7 mm) are lower than in the rest of the CM.

Conservation status

At present, there are only two PNS where *C. hispanus* has considerable areas with habitat suitability scores higher than 75 (PNS 2 and 8). On the other hand, *C. lunaris* is well represented in just one site (PNS 2), the broadest park with mountainous territory, and the only one with sites highly suitable for both species (Table 1.2). The mean suitability scores for both species in the PNS are lower than 30%. Future Natura 2000

network sites will protect a more extensive area (Table 1.2), and consequently, will improve the protection of *Copris* species, facilitating greater interconnectivity among populations. The area with high suitability scores increases eight times for *C. hispanus* and three for *C. lunaris*.

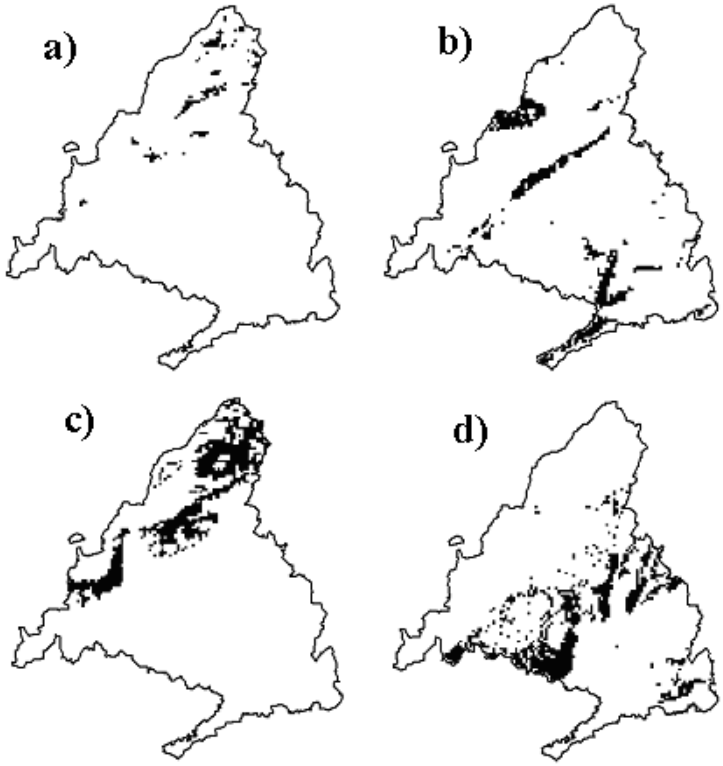


Fig. 1.3 - Maps of areas that are: a) very highly suitable for both species; b) very poor for both species; c) very poor suitability for *C. hispanus* and very high for *C. lunaris*; d) very highly suitable for *C. hispanus* and very poorly for *C. lunaris*.

		Copris hispanus					*Copris lunaris*				Co-occurrence
Protected natural sites (PNS)	Area (Km^2).	Min.	Max.	Mean	SD	HS>75 Area	Min.	Max.	Mean	SD	HS>75 Area	HS>75 Area
1. Peñalara	7	0	8	2.3	3.9	0	0	42	22.3	14.2	0	0
2. Cuenca Alta del Manzanares	458	0	100	34.7	27.6	32	0	100	50.9	38.7	183	15
3. Parque Regional del Sureste	315	0	98	40.5	20.2	7	0	8	0.2	1.3	0	0
4. Refugio de la Laguna de San Juan y Torcal de Valmayor	1	11	11	11.0	0.0	0	0	0	0.0	0.0	0	0
5. Sitio Natural de Interés Nacional del Hayedo de Montejo	3	0	7	2.3	4.0	0	61	77	71.3	9.0	2	0
6. Regajal y Mar de Ontígola	6	10	75	45.7	23.5	0	0	0	0.0	0.0	0	0
7. Paraje Pintoresco del Pinar de Abantos y zona de Herrería	17	0	98	41.1	33.4	3	33	61	43.2	8.9	0	0
8. Parque Regional del Curso Medio del Río Guadarrama	183	0	98	57.9	25.7	29	0	75	17.0	21.0	0	0
9. M. N. I. N. del Collado del Arcipreste de Hita	1	0	0	0.0	0.0	0	53	53	53.0	0.0	0	0
Total	991					61					185	15
New Natura 2000 network sites	3457	0	100	45.0	25.9	487	0	100	31.8	33.3	598	50

Table 1.2 - Habitat suitability values for *C. hispanus* and *C. lunaris,* and co-occurrence zones in each protected natural site and Natura 2000 network sites.
HS>75 Area: zones with a suitability value greater than 75, expressed in Km^2.

DISCUSSION

The niches of Iberian Copris species

The study area, representative of central Iberia, is a zone of confluence of Mediterranean and Eurosiberian-like climate regions. We find that both *Copris* species are distributed along a gradient from the Tajo basin (warmer, dryer, with strong annual weather variations) where *C. hispanus* is found, towards the mountain slopes of the Sistema Central (colder and rainier) where *C. lunaris* predominates. Interestingly, both species present nearly equal marginality factors (and also specialization factors), these axes being so highly correlated that they may be considered identical (Pearson correlation coefficients higher than 0.99). Thus, it can be assumed that *Copris* species are responding to the same main environmental variations in Madrid. However, as can be seen in Table 1.1, the factors driving each one's distribution seem to be opposite, evidencing different environmental responses with respect to the average conditions of the region. To ascertain the way they confront the environmental determinants described by these axes, we have represented the means and deviations of the habitat suitability values for each species along them (see Fig. 1.4). Both species seem to show opposite environmental adaptations: whilst the niche of *C. hispanus* is mainly restricted to calcareous bedrock areas with intense solar radiation, *C. lunaris* prefers sites with acid bedrock and more abundant precipitation. Thus, the principal environmental adaptations of both species respond to the same environmental variations in the studied area, but in a different way (see Fig. 1.4).

Copris dung beetles, tunnelling nesters, construct a tunnel under cattle droppings, burying several dung balls (up to 250 gr., unpublished data) where they lay their eggs. So, environmental factors that affect temperature extremes and water content

in the soil throughout the year are likely, in the main, to shape their distribution in Madrid. Probably *C. hispanus*' physiological adaptations to warm environments with long dry spells, and avoidance of freezing, allow it to nest in highly water-stressed soils, such as sunny calcareous ones. *C. lunaris*, on the contrary, may not be able to nest in such dry areas, but its tolerance of freezing allows it to nest in soils with greater water availability but lower temperatures. Hence, our data show that the environmental niche of both species is biased towards two extreme environments at each of the edges of the gradient: (1) Dry-Mediterranean, with high temperatures and intense solar radiation, calcareous soils and low altitudes and precipitation, and (2) Wet-Alpine, with high altitudes and precipitation, acid soils and low temperatures and weaker solar radiation (Fig. 1.4). Whilst *C. hispanus* does not find suitable areas near the semi-arid first extreme, *C. lunaris* is able to reach the Alpine limit of the gradient. Both species find suitable sites in between the extremes, due to their respective tolerance to medium-to-low temperatures with high-to-moderate precipitations and acid stony and sandy soils. In these sites, the greater water content of the soil and the infrequency of freezing temperatures throughout the year probably constitute highly suitable environmental conditions for nesting success of both *Copris* species.

Competition remarks

Both species seem to co-occur in some areas located in the mid-slopes of the sierra, where competition might take place as a result of their large-size and high capability of nutrient removal (a couple is able to bury up to 250 gr. of dung). However, the presence of competition between both species is possible just in the case they live in the same habitat, appearing at the same time and pasture. Environmental heterogeneity may allow

both species to coexist in the same area but in different habitats; or they may occur in the same locality but at different dates, due to seasonal weather variation; for that reason, competition may not exist. Present results do not clearly support the hypothesis of a higher heterogeneity in the co-occurrence areas, as only two of the five environmental diversity variables tested presented higher values in these zones. However, competitive interactions have been proven for large Afrotropical dung beetles (see Hanski & Cambefort, 1991, and references therein), but information is lacking about interspecific competition in Mediterranean Scarabaeidae (see Finn & Gittings, 2003). Although unpublished data (Veiga, 1982) suggest that specimens of the two species could inhabit the same dung pat, no extensive data is available, so no evaluation of competitive exclusion can yet be made. Further small-scale studies in the sympatric area are necessary to clarify how populations from both species coexist.

Conservation status assessment

Biodiversity conservation of insects, a challenge difficult to respond to due to the lack of information, requires predictive models, as both the most efficient way to obtain reliable maps of insect distributions, and also to evaluate the ability of proposed and existing sites to further conservation. Comunidad de Madrid, an autonomous region with complete jurisdictions over environmental policies, needs an evaluation of both the effectiveness of its PNS and of the potential gains from new ones.

As we commented before, populations of *Copris lunaris* and *Copris hispanus*, as well as those of other dung beetles, are in decline in the Iberian Peninsula, probably because of the use of ivermectines (Lumaret *et al.*, 1993) and the diminution of traditional cattle herding (Lobo, 2001; Roslin & Koivunen, 2001). These species play an

important role in extensive pasture ecosystems by recycling organic matter (Andrzejewska & Gyllenberg, 1980) that, otherwise could cause major damage through accumulation (as occurred in Australia; see Bornemissza, 1976). For this reason, it is important to control and reverse any decline in their populations.

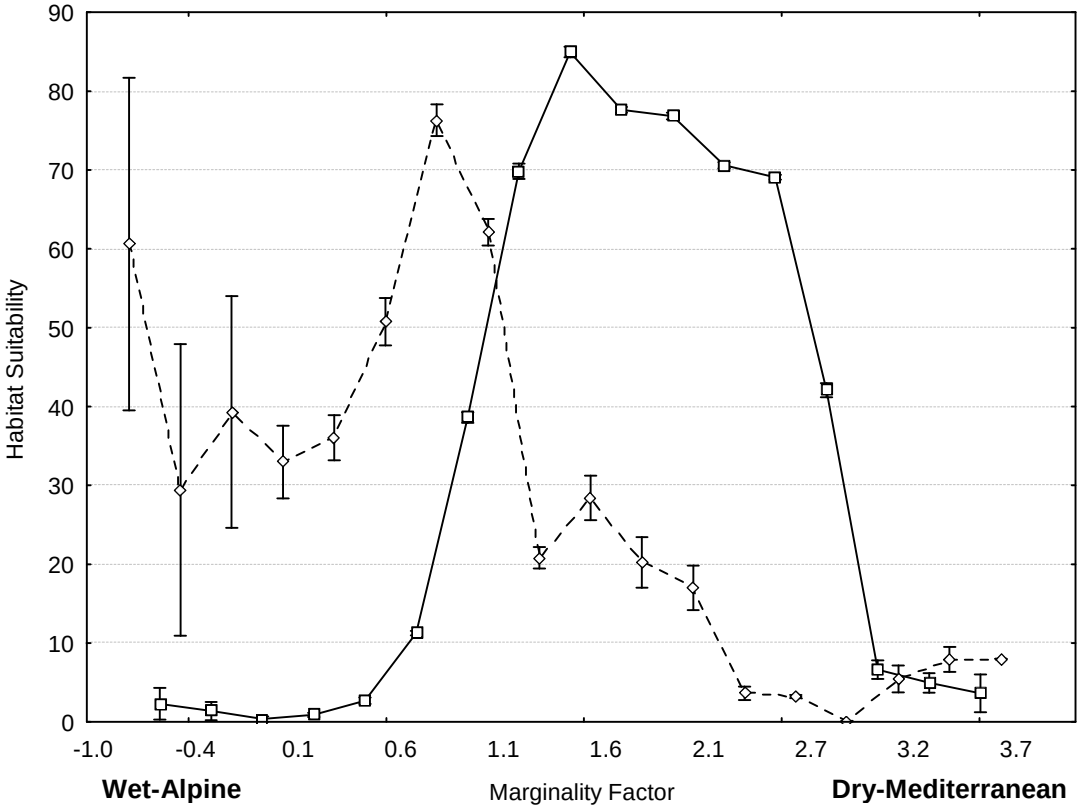


Fig. 1.4 - Variation of mean Habitat Suitability scores along the Marginality Factor (ranging from Wet-Alpine to Dry-Mediterranean environmental conditions).The Factor was divided into 20 intervals, and mean values are shown. *Copris hispanus* is represented by squares and solid lines, and *C. lunaris* by rhombus and broken lines. Vertical lines delimit 95% confidence interval. As Marginality Factors for both species were highly correlated, the one used for representation was that of *C. hispanus* (see text).

To evaluate the conservation status of *Copris* species, we have taken into account the size of protected sites as well as the values of habitat suitability in each PNS and future Nature 2000 network sites. Only one protected site (Hayedo de Montejo; PNS 5) presented an average habitat suitability higher than 70 for one of the species, *C.*

lunaris. However, this site is an ancient beech forest, only 3 km^2 in extent, and so not very effective in preserving populations of this species. Mean suitability values alone are not enough to guarantee protection for a species in protected areas. It is also necessary to take into account the size of the area highly suitable for the species in each protected site. Using the area with habitat suitability greater than 75 for this task, important differences between the two species appear. Whilst for *C. hispanus* only two PNS (2 and 8) measured around 30 km^2 (for a total area of 61 km^2), for *C. lunaris* a single area (PNS 2) measured 183 km^2.

The rarity in PNS of areas highly suitable for both species at the same time highlights two main deficiencies in the CM conservation network. One of them is the area of replacement between basin and mountain assemblages, a gradient zone called ‘‘ramp’’ (‘‘rampa’’ in Spanish), protected in part by PNS 2, that has been identified as an important dung beetle diversity hotspot (Martín-Piera, 2001); another is the Sierra of Guadarrama, scarcely protected by the already-mentioned PNS 5. Areas of faunistic replacement and range-margins are of great importance for the survival of most species (Spector, 2002) where important processes occur (Thomas *et al.*, 2001), specially when faced with climate change (Hill *et al.*, 2002). Using data from additional, extant groups, these areas should be identified, studied, and protected effectively. Connectivity is another weak point of CM protected sites (Sastre *et al.*, 2002). This may be of secondary importance for many dung beetles, such as *Copris* species, as they are presumably good fliers. But less vagile species would need dispersal corridors to be able to disperse as climate change occurs.

In the future, Nature 2000 network will improve the general conservation status of CM because the area and connectivity of protected sites will be increased

substantially. Nature 2000 Network annexes habitats for protection that favour *Copris* species presence such as ‘‘dehesas’’ (forests of Quercus sp. used for grazing), and natural pastures of *Festuca indigesta*. In Europe, protected sites’ agriculture and cattle uses are restricted to traditional ones. Intensive agriculture or monocultures are not let while traditional cattle herding is promoted and conserved so *Copris* species will be favoured by this new protection programme wherever traditional cattle herding is promoted and conserved (Barbero *et al.*, 1999; Verdú *et al.*, 2000; Lobo, 2001; Roslin & Koivunen, 2001). Transhumance, pasture conservation and avoidance of the use of cattle antibiotics such as ivermectins are vital conditions for the conservation of *Copris* populations.

It is important to remark that, although this study has been developed on a small working scale, the studied region has full jurisdiction over local environmental policy. Thus, it has a direct application to the conservation of these species. However, the same methodology can be applied to similar studies on different working scales. In our opinion, accurate estimates of the potential distribution of species are obtainable without recourse to exhaustive data. Habitat suitability maps elaborated with this or similar methods have proven to be quite reliable, insofar as they provide a reasonable approximation to the species niche, even without very many presence points. Together with GIS, habitat suitability maps delimit quite well areas highly suitable for each species and for both species (sympatric areas). A greater sampling effort in these areas would validate them as sympatric and would confirm their actual presences.

ACKNOWLEDGEMENTS

We are especially thankful to Alexandre Hirzel, who has given valuable assistance on Biomapper 2.1, and to James Cerne, who helped us with the English review. The comments from an anonymous referee improved a former version of the manuscript. We are also indebted to the Banco de Datos de la Naturaleza of the Spanish Ministerio de Medio Ambiente, which provided us with the Protected Natural Sites and Natura 2000 proposal GIS layers, as well as to the Servicio Cartográfico de la Comunidad de Madrid for providing the digital version of their CM 1:200.000 map. This paper was supported by the projects REN2001-1136/GLO (Spanish D.G.I.) and 07M/0080/2002 (Comunidad de Madrid). J.H. was supported by a Ph.D. Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid grant.

REFERENCES

Anderson, R. P., Peterson, A. T. & Gómez-Laverde, M. 2002. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos* **98**, 3-16.

Andresen, E. 2003. Effect of forest fragmentation on dung beetle communities and functional consequences for plant regeneration. *Ecography* **26**(1), 87-97.

Andrzejewska, L. & Gyllenberg, G. 1980. Small herbivore subsystem. In: *Grasslands Systems Analysis and Man* (edited by Breymeyer, A. I. & van Dyne, G. M.). Cambridge University Press, Cambridge, 201-267.

Baraud, J. 1992. *Coléoptères Scarabaeoidea d´Europe, Faune de France 78*. Société Linnéenne de Lyon, Lyon.

Barbero, E., Palestrini, C. & Rolando, A. 1999. Dung beetle conservation: effects of habitat and resource selection (Coleoptera: Scarabaeidae). *Journal of Insect Conservation* **3**(2), 75-84.

Bornemissza, G. F. 1976. The Australian dung beetles project 1965–1975. *Australian Meat Research Commission Review* **30**, 1-30.

Buckland, S. T. & Elston, D. A. 1993. Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology* **30**, 478-495.

Busby, J. R. 1991. BIOCLIM- a bioclimate analysis and prediction system. In: *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* (edited by Margules, C. R. & Austin, M. P.). CSIRO, Melbourne, 64-68.

Carpenter, G., Gillison, A. N. & Winter, J. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* **2**(6), 667-680.

Clark, L. 2000a. Global Change Data Archive Vol. 3. 1 km Global Elevation Model. Clark University.

Clark, L. 2000b. Idrisi 32.02. GIS software package. Clark University.

Davis, A. J., Holloway, J. D., Huijbregts, H., Krikken, J., Kirk-Spriggs, A. H. & Sutton, S. L. 2001. Dung beetles as indicators of change in the forests of northern Borneo. *Journal of Applied Ecology* **38**, 593-616.

European Environment Agency. 1996. Natural Resources CD-Rom. European Environment Agency.

Fernández-Galiano, E. & Ramos-Fernández, A. 1987. La Naturaleza de Madrid. In: *Comunidad de Madrid. Consejería de Agricultura y Ganadería*, Madrid.

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. *Biodiversity and Conservation* **11**(12), 2275-2307.

Finn, J. A. & Gittings, T. 2003. A review of competition in north temperate dung beetle communities. *Ecological Entomology* **28**(1), 1-13.

Guisan, A., Edwards Jr., T. C. & Hastie, J. T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* **157**, 89-100.

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

Halffter, G. 1998. A strategy for measuring landscape biodiversity. *Biology International* **36**, 3-17.

Hanski, I. 1991. The dung insect community. In: *Dung Beetle Ecology* (edited by Hanski, I. & Cambefort, Y.). Princeton University Press, New Jersey, 5-21.

Hanski, I. & Cambefort, Y. 1991. Competition in dung beetles. In: *Dung Beettle Ecology* (edited by Hanski, I. & Cambefort, Y.). Princeton University Press, New Jersey, 305-329.

Hill, J. K., Thomas, C. D., Fox, R., Telfer, M. G., Willis, S. G., Asher, J. & Huntley, B. 2002. Responses of butterflies to twentieth century climate warming: implications for future ranges. *Proceedings of the Royal Society B* **269**(1505), 2163-2171.

Hirzel, A., Hausser, J., Chessel, D. & Perrin., N. 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology* **83**(7), 2027-2036.

Hirzel, A. H., Hausser, J. & Perrin, N. 2000. Biomapper 2.0. Laboratory for conservation biology, University of Lausanne, Laussane.

Hirzel, A. H., Helfer, V. & Metral, F. 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling* **145**, 111-121.

Hortal, J. 2004. Selección y diseño de áreas prioritarias de conservación de la biodiversidad mediante sinecología. Inventario y modelización predictiva de la distribución de los escarabeidos coprófagos (Coleoptera, Scarabaeoidea) de Madrid. Unpublished PhD Thesis, Universidad Autónoma de Madrid, Madrid.

Hortal, J. & Lobo, J. M. 2002. Una metodología para predecir la distribución espacial de la diversidad biológica. *Ecología*(16), 405-432.

Hortal, J. & Lobo, J. M. 2005. An ED-based Protocol for Optimal Sampling of Biodiversity. *Biodiversity and Conservation* **14**(12), 2913-2947.

Hortal, J., Lobo, J. M. & Martín-Piera, F. 2001. Forecasting insect species richness scores in poorly surveyed territories: the case of the Portuguese dung beetles (Col. Scarabaeinae). *Biodiversity and Conservation* **10**(8), 1343-1367.

Hortal, J., Lobo, J. M. & Martín-Piera, F. 2003. Una estrategia para obtener regionalizaciones bióticas fiables a partir de datos incompletos: el caso de los Escarabeidos (Coleoptera) Ibérico-Baleares. *Graellsia* **59**(2-3), 331-344.

Hutton, S. A. & Giller, P. S. 2003. The effects of intensification of agriculture on northern temperate dung beetle communities. *Journal of Applied Ecology* **40**(6), 994-1007.

ITGE. 1988. Atlas Geocientífico y del Medio Natural de la Comunidad de Madrid. Serie: Medio Ambiente. Instituto Tecnológico GeoMinero de España, Madrid.

Iverson, L. R. & Prasad, A. M. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs* **68**(4), 465-485.

Lobo, J. M. 2001. Decline of roller dung beetle (Scarabaeinae) populations in the Iberian Peninsula during the 20th century. *Biological Conservation* **97**(1), 43-50.

Lobo, J. M., Hortal, J., Cabrero-Sañudo, F. J. 2006. Regional and local influence of grazing activity on the diversity of a semiarid dung beetle community. *Diversity and Distributions* **12**, 111-123.

Lobo, J. M., Lumaret, J. P. & Jay-Robert, P. 2002. Modelling the species richness of French dung beetles (Coleoptera, Scarabaeidae) and delimiting the predictive capacity of different groups of explanatory variables. *Global Ecology and Biogeography* **11**, 265-277.

Lobo, J. M. & Martín-Piera, F. 1991. La creación de un banco de datos zoológico sobre los Scarabaeidae (Coleoptera: Scarabaeoidea) íbero-baleares: una experiencia piloto. *Elytron* **5**, 31-38.

Lobo, J. M. & Martín-Piera, F. 2002. Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. *Conservation Biology* **16**(1), 158-173.

Lumaret, J. P., Galante, E., Lumbreras, C., Mena, J., Bertrand, M., Bernal, J. L., Cooper, J. F., Kadiri, N. & Crowe, D. 1993. Field effects of ivermectin residues on dung beetles. *Journal of Applied Ecology* **30**(3), 428-436.

Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. 1999. Alternative methods for predicting species distribution: an ilustration with Himalayan river birds. *Journal of Applied Ecology* **36**, 734-747.

Manel, S., Dias, J. M. & Ormerod, S. J. 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling* **120**(2-3), 337-347.

Martín-Piera, F. 2000. Familia Scarabaeidae. In: *Fauna Ibérica 14. Coleoptera, Scarabaeoidea I* (edited by Martín-Piera, F. & López-Colón, J. I.). Museo Nacional de Ciencias Naturales. Consejo Superior de Investigaciones Científicas, Madrid.

Martín-Piera, F. 2001. Area networks for conserving Iberian insects: a case study of dung beetles (col., Scarabaeoidea). *Journal of Insect Conservation* **5**(4), 233-252.

Martín-Piera, F. & Lobo, J. M. 1995. Diversity and ecological role of dung beetles in Iberian grassland biomes. In: *Farming on the Edge: the Nature of Traditional Farmland in Europe* (edited by McCracken, D. I., Bignal, E. M. & Wenlock, S. E.). Joint Nature Conservation Committee, 147-153.

Ministerio de Agricultura, Pesca y Alimentación. 1986. *Atlas Agroclimático Nacional de España*. Dirección General de la Producción Agraria, Madrid.

Mitchell, N. D. 1991. The derivation of climate surfaces for New Zealand, and their application to the bioclimatic analysis of the distribution of kauri (*Agathis australis*). *Journal of the Royal Society of New Zealand* **21**(1), 13-24.

Parker, V. 1999. The use of logistic regression in modelling the distributions of bird species in Swaziland. *South African Journal of Zoology* **34**, 39-47.

Pearce, J. & Ferrier, S. 2000. An evaluation of alternative algorithm for fitting species distribution models using logistic regression. *Ecological Modelling* **128**, 127-147.

Pereira, J. M. C. & Itami, R. M. 1991. GIS-based habitat modelling using logistic multiple regression: a study of the Mt. Graham Red Squirrel. *Photogrammetric Engineering and Remote Sensing* **57**(11), 1475-1486.

Peterson, A. T. & Holt, R. D. 2003. Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. *Ecology Letters* **6**, 774-782.

Peterson, A. T., Soberon, J. & Sánchez-Cordero, V. 1999. Conservatism of ecological niches in evolutionary time. *Science* **285**, 1265-1267.

Rassi, P., Kaipainen, H., Mannerkoski, I. & Ståhls, G. 1991. Report on the monitoring of threatened animals and plants in Finland. Committee Report 1991, Ministry of the Environment, Helsinki, Finland.

Robertson, M. P., Caithness, N. & Villet, M. H. 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions* **7**(1-2), 15-27.

Roslin, T. 2000. Dung beetle movement at two spatial scales. *Oikos* **91**, 323-335.

Roslin, T. 2001a. Large-scale spatial ecology of dung beetles. *Ecography* **24**, 511-524.

Roslin, T. 2001b. Spatial population structure in a patchily distributed beetle. *Molecular Ecology* **10**, 823-837.

Roslin, T. & Koivunen, A. 2001. Distribution and abundance of dung beetles in fragmented landscapes. *Oecología* **127**, 69-77.

Sastre, P., de Lucio, J. V. & Martínez, C. 2002. Modelos de conectividad del paisaje a distintas escalas. Ejemplos de aplicación en la Comunidad de Madrid. *Ecosistemas, 2002/2*. http://www.revistaecosistemas.net/revista_frame.asp?pagina=%2Farticulo.asp%3FId%3D287%26Id_Categoria%3D2%26tipo%3Dportada

Scott, J. M., Davis, F. W., Csuti, B., Noss, R., Butterfield, B., Groves, C., Anderson, H., Caicco, S., D´Erchia, F., Edwards, T.C Ulliman, J. & Wright, R.G. 1993. Gap analysis: a geographic approach to protection of biological diversity. *Wildlife Monographs* **123**, 1-41.

Servicio Cartográfico de la Comunidad de Madrid. 1996. Mapa de la Comunidad de Madrid. Escala 1:200.000. Consejería de Ordenación Territorial, Comunidad de Madrid, Madrid.

Skidmore, P. 1991. *Insects of the British Cow Dung Community*. Field Studies Council, Monfort Bridge.

Spector, S. 2002. Biogeographic crossroads as priority areas for conservation. *Conservation Biology* **16**(6), 1480-1487.

StatSoft, I. 2001. STATISTICA (data analysis software system), version 6.

Stockwell, D. & Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13**(2), 143-158.

Thomas, C. D., Bodsworth, E. J., Wilson, R. J., Simmons, A. D., Davies, Z. G., Musche, M. & Conradt, L. 2001. Ecological and evolutionary processes at expanding range margins. *Nature* **411**(6837), 577-581.

Vayssières, M. P., Plant, R. E. & Allen-Díaz, B. H. 2000. Classification trees: an alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science* **11**(5), 679-694.

Veiga, C. M. 1982. Los Scarabaeoidea (Col.) coprófagos de Colmenar Viejo (Madrid). Perfiles autoecológicos. Licenciature Thesis, Universidad Complutense de Madrid, Madrid.

Verdú, J. R., Crespo, M. B. & Galante, E. 2000. Conservation of a nature reserve in Mediterranean ecosystems: the effects of protection from grazing on biodiversity. *Biodiversity and Conservation* **9**(12), 1707-1721.

Verdú, J. R. & Galante, E. 2002. Climatic stress, food availability and human activity as determinants of endemism patterns in the Mediterranean region: the case of dung beetles (Coleoptera, Scarabeoidea) in the Iberian Peninsula. *Diversity and Distributions* **8**, 259-274.

Walker, P. A. & Cocks, P. A. 1991. HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters* **1**, 108-118.

Evaluación de los efectos del uso de pseudo-ausencias en los modelos predictivos de distribución.

RESUMEN

Predecir la distribución de especies con datos de presencia obtenidos de atlas, colecciones de museos y bases de datos es un desafío. En este capítulo, se comparan siete procedimientos para obtener pseudo-ausencias, que después se usarán para generar modelos de regresión logística (GLM). Las pseudo-ausencias son seleccionadas al azar o mediante métodos elaborados a partir de datos de presencia (ENFA y MDE) para obtener modelos de la distribución de *Graellsia isabelae*, una especie amenazada de polilla endémica de la Península Ibérica. Los resultados muestran que el método de selección de pseudo-ausencias influye en gran medida en el porcentaje de variabilidad explicada, en los resultados de las medidas de precisión y, lo más importante, en el nivel de restricción de la distribución estimada. Cuanto más alejada está la región de la que se extraen las pseudo-ausencias del óptimo ambiental definido por las presencias, los modelos generados obtienen mejor valor de precisión, y la sobrepredicción es mayor. Cuando otras variables distintas de las ambientales influyen en la distribución de las especies (p.ej. en estado de no equilibrio) y no existe información precisa sobre las ausencias, la selección de pseudo-ausencias al azar o de lugares con condiciones ambientales similares a las de las presencias genera los mapas de distribución potencial más restringidos, ya que las pseudo-ausencias pueden estar situadas dentro de zonas

ambientalmente adecuadas. Este estudio muestra que si no existen ausencias fiables, el método de elección de pseudo-ausencias condiciona el modelo obtenido, ya que se generarán diferentes predicciones a lo largo del gradiente entre la distribución real y la potencial.

Palabras clave: pseudo-ausencias, modelos de distribución, precisión del modelo, no equilibrio, *Graellsia isabelae*, Península Ibérica.

Este capítulo ha sido publicado como:

Chefaoui, R. M. & Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological modelling* **210**, 478-486.

Assessing the effects of pseudo-absences on predictive distribution model performance.

ABSTRACT

Modelling species distributions with presence data from atlases, museum collections and databases is challenging. In this paper, we compare seven procedures to generate pseudoabsence data, which in turn are used to generate GLM-logistic regressed models when reliable absence data are not available. We use pseudo-absences selected randomly or by means of presence-only methods (ENFA and MDE) to model the distribution of a threatened endemic Iberian moth species (*Graellsia isabelae*). The results show that the pseudo-absence selection method greatly influences the percentage of explained variability, the scores of the accuracy measures and, most importantly, the degree of constraint in the distribution estimated. As we extract pseudo-absences from environmental regions further from the optimum established by presence data, the models generated obtain better accuracy scores, and over-prediction increases. When variables other than environmental ones influence the distribution of the species (i.e., non-equilibrium state) and precise information on absences is non-existent, the random selection of pseudo-absences or their selection from environmental localities similar to those of species presence data generates the most constrained predictive distribution maps, because pseudo-absences can be located within environmentally suitable areas. This study shows that if we do not have reliable absence data, the method of pseudo-

absence selection strongly conditions the obtained model, generating different model predictions in the gradient between potential and realized distributions.

Keywords: Pseudo-absences; Distribution models; Model accuracy; Non-equilibrium; *Graellsia isabelae*; Iberian Peninsula

INTRODUCTION

Reliable species distribution information on various scales is needed for both biogeographic and conservation purposes. Taking advantage of computing developments such as databases and GIS, many different initiatives aim to compile massive amounts of taxonomic and distribution information (Bisby, 2000). Atlases, museum data and databases can provide information relevant to the development of prediction maps (Dennis & Hardy, 1999; Reutter *et al.*, 2003; Chefaoui *et al.*, 2005; Hortal *et al.*, 2005). Since these heterogeneous data sources do not indicate the locations where the species have not been found after a sufficiently intense collection effort, false absences can decrease the reliability of prediction models (see Anderson, 2003; Loiselle *et al.*, 2003). Group discrimination techniques that use presence-absence data (Guisan & Zimmermann, 2000) seem to predict species distributions more accurately than profile techniques, which only use presence data (Ferrier & Watson, 1997; Manel *et al.*, 1999; Hirzel *et al.*, 2001; Guisan *et al.*, 2002; Brotons *et al.*, 2004; Gu & Swihart, 2004; Segurado & Araújo, 2004). However, group discrimination techniques are appropriate only in the case where absence data indicate the entire area unsuitable for the species are available. Since a quick and feasible method to overcome this problem is needed, the following approaches have been suggested: (i) randomly choosing absence points across

all of the available territory (for example, Stockwell & Peters, 1999), (ii) selecting random absence points but weighting them in favour of areas known to contain true absences (Zaniewski *et al.*, 2002), and (iii) including absence points identified from a circular buffer area around each presence point (Hirzel *et al.*, 2001). Since all of these methods may produce false absences, even in areas that are environmentally favourable for the species, using a profile technique to calculate a habitat suitability map has been proposed as a way to select weighted absence points, which can subsequently be used with presence data in a logistic regression procedure (Engler *et al.*, 2004). Absences obtained with this method, "pseudo-absences", can be considered an intermediate methodological approach between presence-only and presence-absence distribution models, which are especially useful when accurate absence data are not available.

In this study, two profile techniques were used to select pseudo-absences progressively near to the environmental domain of the presences, while also selecting them at random. Presence-absence models derived from these pseudo-absences and occurrence data from *Graellsia isabelae* (Graells, 1849) (Lepidoptera: Saturniidae), a protected moth endemic to Spain (see Fernández-Vidal, 1992), are compared with the purpose of showing that it is possible to achieve differently forecasted distributions depending on the method and the threshold used to select these pseudo-absences. The variation in these predictions will be subsequently related to the ambivalent capacity of distribution predictive models to represent realized and potential species distributions (*sensu* Svenning & Skov, 2004).

METHODS

Study area and biological data

The area considered was mainland Spain and the Balearic islands. Since this species has an eastern Iberian distribution, we assume that our study area included most of the suitable habitat area for this species. The studied area comprised 498,150 km^2 divided into 5270 cells of 10 km × 10 km, to which biological and environmental data are referred.

G. isabelae, a sedentary and non-gregarious caterpillar, lives in pine forests and has five developmental stages. From June to August, the larvae feed before metamorphosizing into pupae. Since *G. isabelae* is a conspicuous and well-known species (adults are beautiful and exhibit marked sexual dimorphism), occurrence records can be considered reliable.

Species-presence data were mainly obtained from a distribution atlas (Galante & Verdú, 2000), as well as unpublished data from the Valencia region (Baixeras, 2001; J. Baixeras, personal communication, 2004) and other bibliographic references (Viejo, 1992; García-Barros & Herranz, 2001; López-Sebastián *et al.*, 2002). Because species data came from diverse sources and some references were old (since 1849), all data were checked by comparing their locations with the distribution of pinewoods to eliminate possible outliers. As a result, six presence data points were discarded. A total of 136 presence data points with a spatial resolution of 100 km^2 (UTM cells) were considered (see Fig. 2.1).

Predictor variables

The explanatory variables used in the preparation of distribution models (Table 2.1) come from different sources and have been set up with the aid of IDRISI Kilimanjaro software (Clark Labs, 2003). Topographic variables, elevation and slope were extracted from a global DEM with a 1 km spatial resolution (Clark Labs, 2000). Aspect diversity was calculated by means of the Shannon Index, which estimated the aspect variation in

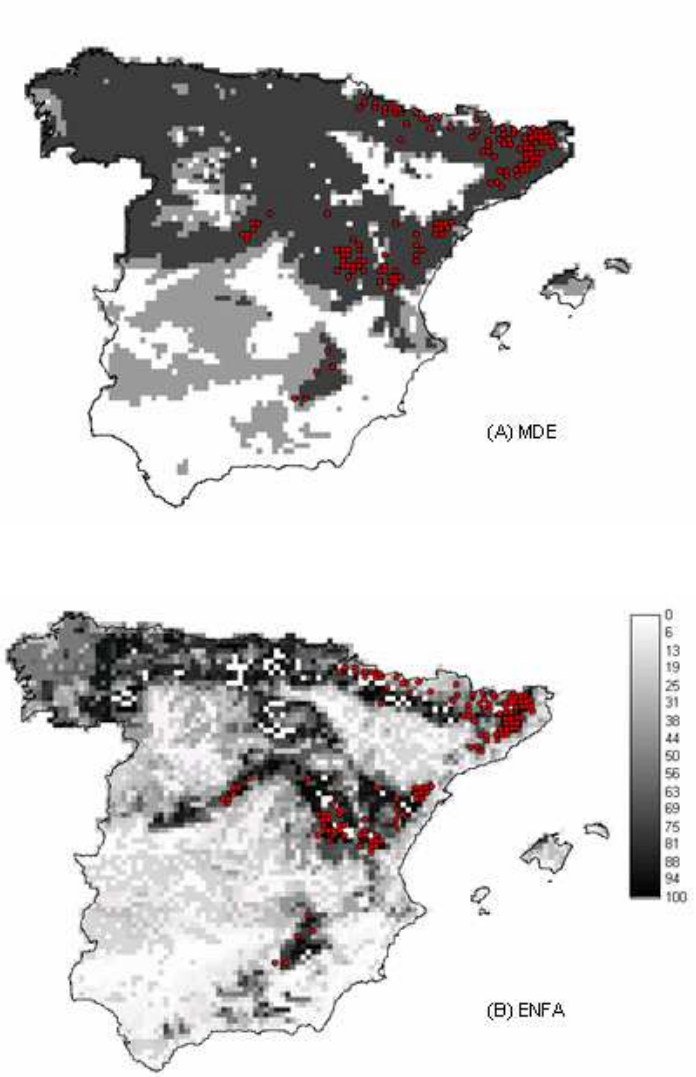


Fig. 2.1 - Habitat suitability maps representing the potential distribution obtained from presence-only models. (A) Dark grey indicates suitable area obtained from a multi-dimensional envelope model (MDE); light grey indicates potential area added by increasing maximum and minimum scores 10% for each environmental variable (Expanded-MDE). (B) Scale on the right shows different habitat suitability (HS) scores obtained from an ENFA model with the same environmental variables.

all 1-km pixels composing each 100 km^2 cell. Temperature and precipitation data were provided by the Spanish Instituto Nacional de Meteorología. Aridity was calculated as: $Ia = 1/((P/T) + 10)\times10^2$, where P is the mean annual precipitation and T is the mean annual temperature (see Verdú & Galante, 2002). In addition, four lithological variables were digitized from a lithological map (Instituto Geográfico and Nacional, 1995). The resulting polygon vector layers were rasterized at 1 km^2 resolution, and the areas of calcareous deposits, siliceous sediments, stony acidic soils and calcareous soils were subsequently calculated for each cell. These variables were included to incorporate the basic-acidic nature of the soils and their hardness, variables that can be relevant to explain plant species distribution. The third-degree polynomial of the central latitude (Lat) and longitude (Lon) of each grid cell (Trend Surface Analysis; see Legendre and Legendre, 1998) was included after the environmental variables in order to determine if it helped explain anything more about the species distribution (see Lobo *et al.*, 2006). All continuous independent variables were referenced to the same 10 km × 10 km UTM grid square as species data. Predictor environment variables were standardized to 0 mean and 1 standard deviation to eliminate the effect of varying measurement scales. Finally, latitude and longitude were standardized in the same way as the environmental variables.

Presence-only models

We used Multi-Dimensional Niche Envelope (MDE; Busby, 1991; Lobo *et al.*, 2006) and Ecological Niche Factor Analysis (ENFA; Hirzel *et al.*, 2002) to elaborate the presence-only models. These models were generated from the presence data (n = 136)

Predictor variables	Minimum - Maximum values
Environmental variables	
Mean elevation (m)	0 - 2722
Aspect diversity	0 - 16
Slope (º)	0 - 46
Summer precipitation (July, Aug. and Sept.) (mm)	6.6 - 472
Annual precipitation (mm)	178 - 2201
Aridity	0 - 1.64
Minimum annual temperature (ºC)	-3.6 - 14.3
Maximum annual temperature (ºC)	9.1 - 24.9
Area with acidic stony soils (km^2)	0 - 100
Area with calcareous stony soils (km^2)	0 - 100
Area with acidic sediments (km^2)	0 - 100
Area with calcareous sediments (km^2)	0 - 100
Spatial variables (in UTM coordinates)	
Latitude (Y)	3990000 - 4860000
Longitude (X)	-20000 - 1060058

Table 2.1 - Explanatory variables used to generate the distribution models. Spatial variables were used only with presence-absence GLM models.

and the information from 12 environmental predictor variables (Table 2.1). For the MDE model, maximum and minimum scores for all environmental variables from presence cells were used to select the suitable grid squares, with environment scores falling within that range. Thus, the generally appropriate environmental conditions for the species were established according to the environmental conditions of the observed presence points. In the Expanded-MDE, this range was expanded by 10% to guarantee that absences selected were environmentally distant from presence localities. MDE and Expanded-MDE were generated in EXCEL spreadsheets, while binary maps were elaborated with IDRISI Kilimanjaro.

ENFA was performed using BIOMAPPER 3.1 software (Hirzel *et al.*, 2004). The ENFA modelling technique (similar to Principle Component Analysis in that it generates orthogonal axes) computes a group of uncorrelated factors with ecological meaning (marginality and specialization), summarizing the main environmental gradients in the region considered. Habitat suitability (HS) is modelled using the selected factors to estimate the ecogeographic degree of similarity between each grid

square and the environmental preferences of the species; that is, this method estimates the probability that a given cell belongs to the environmental domain of the presence observations. The resulting habitat suitability map has scores (HS values) that vary from 0 (minimum habitat suitability) to 100 (maximum). The predictor variables were normalized through a Box-Cox transformation (Sokal & Rohlf, 1981), and a "distance geometric-mean" algorithm, which provides a good generalization of the niche (Hirzel & Arlettaz, 2003), was chosen to perform the analyses.

Pseudo-absences

Identifying unsuitable habitats by profile techniques enables us to produce reliable pseudo-absences for presence-absence modelling. Previous results (Jiménez-Valverde *et al.*, 2008) clearly demonstrate that it is necessary to use as much good absence data as possible, especially when dealing with small numbers of presences, to correctly classify the absence zone (see also Thuiller *et al.*, 2004). However, to avoid biases caused by the inclusion of an extremely high number of absences (King & Zeng, 2000; Dixon *et al.*, 2005), we selected 10 times more absences than presences (1360) from each model. To compare the effect of obtaining pseudo-absences with each method, we also randomly selected absences from all regions excepting occurrence cells. Seven groups of pseudo-absences were obtained: one at random, one from MDE, one from Expanded-MDE and four from ENFA. From the ENFA model, the sets were extracted according to four different habitat suitability (HS) thresholds: HS≤10 (ENFA-10), HS≤20 (ENFA-20), HS≤30 (ENFA-30) and HS≤40 (ENFA-40). The upper limit of the selected HS threshold was established as the mean HS score of presences (67) minus its standard deviation (26).

Presence-absence models

The 136 presence data points and each set of 1360 pseudoabsences were subsequently analyzed with the stepwise logistic regression method using Generalized Linear Models (GLM). GLM are an extension of classic linear regression models that allow for analysis of non-linear effects among variables and non-normal distributions of the independent variables (McCullagh & Nelder, 1989). The relationship between the dependent and the explanatory variables (the link function) is logit, and a binomial distribution of the dependent variable was assumed for this analysis.

The species presence-absence data for each of the 10 km × 10 km UTM cells were first compared to linear, quadratic and cubic functions of each environmental variable in order to account for possible curvilinear relationships (Austin, 1980). Next, a model using all environmental variables was built, adding the variables sequentially, in order of their estimated importance (i.e., in a forward-backward stepwise procedure). Lastly, the third-degree polynomial of the central latitude and longitude of each cell was included in the model (TSA) to account for spatial variation due to historic, biotic or environmental factors otherwise not directly considered by this analysis (Legendre & Legendre, 1998). Backward-stepwise regression, with 9 terms of the equation used as predictor variables, removed insignificant spatial terms. Significant terms ($p < 0.05$) were retained and included in the final environmental model. Including spatial variables after environmental ones partially prevented the model from accurately representing ecological niches but allowed us to increase the explanatory capacity of the model by incorporating unconsidered non-environmental factors. The STATISTICA 6.0 package (StatSoft Inc., 2001) was used for all statistical computations.

Validation and cut-off threshold

The Receiver Operating Characteristic (ROC; Zweig & Campbell, 1993; Schröder, 2004) was used to measure performance of the models. A ROC curve is a plot of sensitivity (ratio of correctly classified positives to the total number of positive cases) versus 1 − specificity (false positive rate) at all possible thresholds of presence-absence classification. The area under the ROC function (AUC), independent of the presence-absence threshold (Fielding, 2002), is widely used as a measure of model prediction accuracy. An AUC value of 0.5, from a possible range of 0-1, indicates that prediction of species presence-absence does not deviate from that of a random assignment, while an AUC score of 1 indicates perfect presence-absence prediction. Prediction maps were also compared by calculating sensitivity and specificity (percentages of correctly predicted presences and absences).

To compare observed and predicted maps, a cut-off point is needed to transform continuous probabilities obtained in GLM models to binary probabilities (i.e., presence-absence). The sensitivity-specificity difference minimizer (Liu *et al.*, 2005; Jiménez-Valverde & Lobo, 2006, 2007) was used to select this threshold due to its generally good performance. These three accuracy measures (AUC, sensitivity and specificity) were computed with the aid of a jackknifing procedure (see Olden *et al.*, 2002; Engler *et al.*, 2004). With a dataset of n observations, the model was recalculated n times, leaving out one observation in turn. Each one of the regression models based on the $n-1$ observations was then applied to the excluded observation, and these models derived predictions for all observations, which were used again to calculate new sensitivity, specificity and AUC jackknife-derived scores.

Since all the resulting models use pseudo-absences, both specificity and AUC scores estimate the degree of accuracy of the absence information used in the model training process. Thus, a high specificity score only implies that most of the data considered as absence data are correctly predicted and does not imply a high performance in the prediction of the unknown true absences.

RESULTS

Habitat suitability maps from the profile modelling techniques show remarkable differences (Fig. 2.1). The suitable area predicted by Expanded-MDE is 31% greater (356,700 km^2) than the area predicted by MDE (244,100 km^2). The predicted area generated by applying the four ENFA threshold-related models decreased with increases in the HS threshold; increasing the HS threshold from 10 to 40 produces a 53% reduction in the suitable area (Fig. 2.2 and Table 2.2). The mean ENFA habitat suitability score values for the 136 presence data points was 67.3 ± 25.6 (S.D.) with HS scores oscillating between 5 and 100; 52 presence points had very high suitability scores (HS > 75), 45 had high suitability scores (75 ≥ HS > 50), 33 had low suitability scores (50 ≥ HS > 25), and 6 had very low suitability scores (HS ≤ 25).

Five to seven predictor variables were selected in the seven final GLM logistic models ($p \leq 0.05$), highlighting the relevance of some explanatory variables in all models: mean elevation, summer precipitation and aridity (not shown). Spatial variables added after environmental ones only slightly improved the explanatory capacity of the models (around 1% of total variability), except the model in which pseudo-absences were selected at random (around 5% of added variability). Final GLM models in which pseudo-absences were selected by a profile technique accounted for a high percentage of

total explained deviance (from 87.6% to 97.6%, see Table 2.3), while GLM models with pseudo-absences selected at random had a lower explanatory capacity (around 68% of total deviance). In general, models using absence data further away environmentally from the presence data possessed a higher explanatory capacity (Table 2.3).

All the models that used pseudo-absences selected by profile techniques had impressive sensitivity, specificity and AUC scores (mean ± 95% confidence interval: 0.9792 ± 0.0123, 0.9843 ± 0.0126 and 0.9952 ± 0.0066, respectively), which were significantly higher than those obtained by selecting pseudoabsences at random (Table 2.3). Jackknife estimates of the three accuracy measures showed that the model results were highly stable: they differed by less than 2% of the estimates obtained using all the observations (Table 2.3). As with the explained percent deviance, models that used absence data that where environmentally further away from the presence data also had higher accuracy scores (Table 2.3).

After selecting the threshold value, continuous GLM probability maps were converted to binomial distributions (Fig. 2.2). The restrictive character of GLM versus profile techniques is evident; the suitable area generated by GLM models was smaller than that derived from profile-techniques (from 31% to 48% smaller; see Table 2.2). The GLM model performed using pseudo-absences derived from Expanded-MDE was the one that generated wider forecasted areas. Interestingly, in the case of ENFA-derived GLM models, the estimated species distribution area decreased with gradual increases in the habitat suitability threshold used to discern pseudo-absences. Moreover, when pseudo-absence data were randomly selected, the predicted species distribution area was almost 40% smaller than the most restricted distribution area estimated using pseudo-absences derived from profile techniques (Table 2.2 and Fig. 2.3). This

reduction in the predicted area always followed a well-defined geographical pattern; cells in the northwestern corner of the study area, where the species has never been collected, disappeared gradually from the potential distribution area.

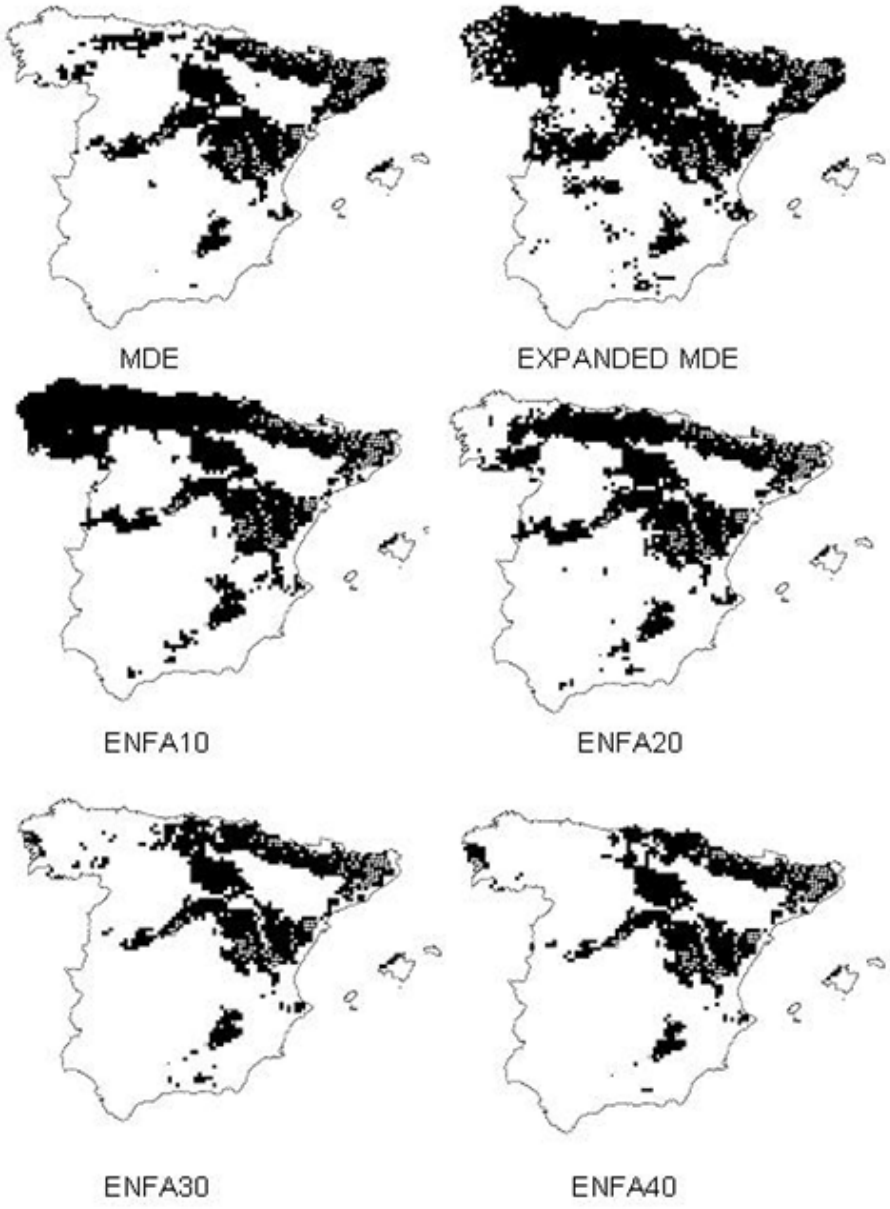


Fig. 2.2 - Obtained distribution maps from logistic GLM models using pseudo-absences derived from profile techniques, which vary in the threshold applied to select probable absence points (see methods section).

	MDE	Expanded-MDE	ENFA10	ENFA20	ENFA30	ENFA40	Random
Unsuitable area estimated	282,900	170,300	144,800	243,700	311,600	347,600	-
Suitable area estimated	244,100	356,700	382,200	283,300	215,400	179,400	-
Suitable area estimated	132,300	245,400	199,200	159,600	121,200	113,200	68,400

Table 2.2 - Predicted suitable and unsuitable areas of distribution for *Graellsia isabelae* in Spain according to the two profile techniques (ENFA and MDE) and suitable area predicted by the logistic GLM models from selected pseudo-absences at different thresholds from the profile techniques or randomly (see Methods). Values are expressed in km^2.

Method of pseudo-absence selection	Deviance	Explained Deviance (%)	Sensitivity	Specificity	AUC
MDE					
Environmental	69.61	94.48			
Environmental + TSA	64.43	94.89	0.9779 (0.9619)	0.9816 (0.9711)	0.9990 (0.9824)
Expanded-MDE					
Environmental	42.29	96.65			
Environmental + TSA	42.29	96.65	0.9708 (0.9632)	0.9956 (0.9956)	0.9831 (0.9794)
ENFA10					
Environmental	35.62	97.17			
Environmental + TSA	30.33	97.59	0.9926 (0.9779)	0.9941 (0.9765)	0.9998 (0.9910)
ENFA20					
Environmental	55.64	95.59			
Environmental + TSA	45.53	96.39	0.9926 (0.9779)	0.9926 (0.9779)	0.9988 (0.9909)
ENFA30					
Environmental	115.26	90.87			
Environmental + TSA	109.74	91.31	0.9779 (0.9632)	0.9764 (0.9633)	0.9961 (0.9896)
ENFA40					
Environmental	171.14	86.45			
Environmental + TSA	156.08	87.64	0.9632 (0.9559)	0.9654 (0.9573)	0.9942 (0.9851)
Random					
Environmental	467.22	63.01			
Environmental + TSA	406.86	67.79	0.8970 (0.8823)	0.8970 (0.8860)	0.9599 (0.9505)

Table 2.3 - Comparison of the GLM models obtained from pseudo-absences generated at random and from two profile techniques (ENFA and MDE) at different threshold scores (see Methods). GLM models were built by including environment and environment + spatial (TSA) variables. The percentage of correctly predicted presences (sensitivity) and absences (specificity), as well as the area under the ROC function (AUC) are measurements derived from the confusion matrix to estimate model prediction accuracy. The average scores of these accuracy measures are showed in brackets after accomplishing a Jackknifing procedure in which all the regression models based on the *n*-1 observations were calculated and the model applied to that excluded one.

DISCUSSION

In this study, selecting pseudo-absences appears to be a good strategy to make GLM modelling possible when true absence data are not available. Taking into account that an AUC value > 0.90 is qualified as outstanding (Hosmer & Lemeshow, 2000), our validation results for all the pseudoabsence selection approaches are excellent. Engler *et al.* (2004) show that GLM models using ENFA-weighted pseudoabsences provide significantly better results than those that use randomly chosen pseudo-absences or profile techniques such as ENFA alone, due mainly to the tendency of profile techniques to over-predict species distributions. In agreement with Engler *et al.* (2004), we find that this strategy provides a way to enhance the quality of GLM-based potential distribution maps. In our case, the GLM model derived from pseudo-absences extracted from cells with an ENFA habitat suitability score equal to or lower than 20 (ENFA-10 and ENFA-20) seems to be the most accurate, although Expanded-MDE pseudo-absence selection also provides rather good validation results. However, the profile method used and the environmental limits defined when selecting pseudo-absences greatly influences the percentage of explained variability, the scores of the accuracy measures and, most importantly, the degree of constraint on the distribution estimated.

In the case of *G. isabelae*, where presence data occur in the eastern territory, GLM spatial predictions exclude the Ebro Valley and other areas of low elevation. However, in the western area, where the species has not been observed, the lack of reliable absence data causes high variability in the predictions; the models tend to expand the suitable area for this species to the northwestern Iberian corner. Presence only methods always generate wider potential distributional areas than GLM models derived from pseudo-absences. Moreover, those strategies in which pseudo-absences

were selected from a smaller area environmentally distant from the optimum established by the presence data (Expanded-MDE and ENFA-10) generate final GLM models that explain a higher percentage of total variability, have higher accuracy scores and wider distributions. Conversely, the profile techniques that generate wider unsuitable areas, such as MDE, ENFA-30 and ENFA-40, produce functions with lower percentages of explained deviance and poorer accuracy scores, but more restricted predictive distribution maps, similar to the observed distribution. The random selection of pseudo-absences generates the most constrained predictive distribution map because all absence data are included, even those data located within environmentally favourable areas.

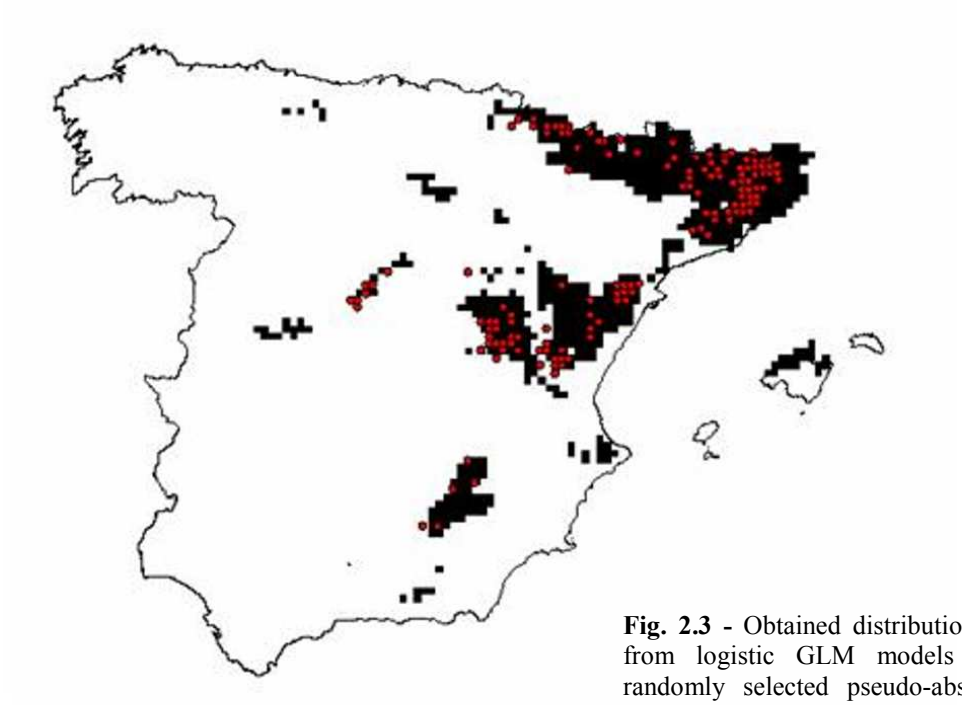


Fig. 2.3 - Obtained distribution map from logistic GLM models using randomly selected pseudo-absences. Red dots represent the observed distribution of *Graellsia isabelae*.

Only an appropriate selection of presence and absence locations can guarantee the reliability of distribution model predictions. First, however, we must determine

whether we would like to produce a distributional range closer to the potential or closer to the realized distribution. Species distributions should be considered abstractions of a dynamic reality. We can be interested in providing a distributional hypothesis able to reflect all the environmental suitable places in which a species can occur according to a group of environmental variables (the potential distribution). Profile techniques such as MDE and ENFA estimate the distribution of the species considering the environmental information of the localities in which the species has been observed, generating wide suitable distributions; this is because these techniques cannot incorporate the absence information on the climatically suitable localities in which the species does not occur. Many theoretical arguments and empirical studies show that it is possible to find reliable absence data in sites with environmentally favourable conditions (Ricklefs & Schluter, 1993; Hanski, 1998; Pulliam, 1988, 2000). Obviously, including such absence information in predictive modelling techniques should inevitably diminish the obtained range size until a distributional hypothesis nearer to the realized distribution is reached. That happens because some "a priori" favourable environmental localities are considered as absences. Hence, only the use of reliable presence and absence data and discrimination techniques such as GLMs allows the production of a reliable approach to model the "real" distribution of a species; a distribution in which contingent distribution restriction forces as historical factors, biotic interactions or dispersal limitations play an effective role.

The current distribution of *G. isabelae* is reasonably well known, due to its conspicuous nature. Thus, we are inclined to believe that the lack of presence data in suitable habitat areas in western Iberia indicates actual absence, and not a sampling artefact. Profile techniques indicate that the potential distributional range of *G. isabelae*

is wider than realized in the western region. Thus, reasons other than environmental characteristics may be the cause of this non-equilibrium state in which species do not occupy all suitable habitats (see Austin, 2002; Guisan & Thuiller, 2005). Under equilibrium conditions, good absence data should always come from locations with unfavourable environments (see, for example, Hirzel & Arlettaz, 2003). Contrarily, in a non-equilibrium scenario, the cells considered environmentally unfavourable and chosen as pseudo-absences can influence the obtained predictive functions and the difference between potential and realized distributions (see Svenning & Skov, 2004). The principal difficulty lies in obtaining predictive distributional models that closely approximate the realized distribution of species under non-equilibrium conditions; obtaining these models causes reductions in goodness-of-fit, similar to those caused by using MDE, ENFA-30 or ENFA-40. This response is due to the fact that both presence and absence data may be possible under similar environmental conditions (see also Collingham *et al.*, 2000). Hence, neither the coefficient of determination, sensitivity, specificity, nor AUC scores are appropriate measures of the performance of models if the objective is to obtain a model under non-equilibrium conditions. Selecting pseudo-absences environmentally distant from the presences unavoidably facilitates the production of models that over-predict presences, as well as the discrimination between presences and absences. The discrimination ability of distribution models must be evaluated according to the pursued purposes. Profile methods must be used if we want to generate a hypothesis on the potential distribution. Using discrimination methods and selecting pseudo-absences by Expanded-MDE and ENFA-10 methods also allows us to obtain models nearer to the potential distribution of the species because pseudo-absences are selected from environments dissimilar to those of species presence data.

On the contrary MDE, ENFA-30 and ENFA-40 are better models to represent the approximate range of the realized distribution. Paradoxically, the random selection of pseudo-absences can be a satisfactory alternative procedure to model the realized distribution of the species, provided good absence data are not available, because we include in the modelling process many absences near the environmental domain of the presences. Since there is no single way to build, evaluate and interpret distribution models, it is necessary to carefully consider the available distribution and biological information of each species in an individualized way (Zimmermann & Kienast, 1999; Rushton *et al.*, 2004; Soberón & Peterson, 2005). In conclusion, as the degree of prediction over-estimation varies with the method applied, the success that we can achieve using correlative static models and environmental predictors is determined by two factors: (1) the distribution equilibrium state of the species in the analyzed region and (2) the method used to select pseudo-absences.

How does one construct predictive models without using variables that describe the biotic or historical factors likely to influence the non-equilibrium, present-day distribution of the species? The major challenges of distribution modelling are accounting for the distributions of most species that are likely to be in non-equilibrium states. If it is not possible to assume that environmental factors are the unique determinants of species distributions (Davis *et al.*, 1998; Iverson *et al.*, 2004; Skov & Svenning, 2004; Thomas *et al.*, 2004; Soberón & Peterson, 2005), perhaps including spatial variables along with environmental ones would help account for variability due to non-environmental factors (Legendre & Legendre, 1998; Lobo *et al.*, 2004). In our case, the addition of spatial variables after environmental ones increases the explanatory capacity of the GLM models when pseudo-absences are randomly selected or when the

habitat suitability range is augmented to select pseudo-absences. Although the additional deviance explained by environmental + TSA models can be small, it is important to remember that all environmental variables are spatially structured, and that changes in the environmental variables included in the models can cause a better recovery of the spatial variability in the dependent variable. For example, the inclusion of significant environmental variables obtained from the GLM model built with pseudo-absences from ENFA-10 in the ENFA-20, ENFA-30 and ENFA-40 models does not noticeably reduce the explained deviance (from 95.0%, 88.8% and 82.8% to 93.4%, 84.1% and 80.5%), but the added percentage of variability explained by the spatial variables increases notably (7.3%, 10.0% and 13.3%, respectively). Another promising option is to consider some geographical variables as predictors indirectly related to the failure of a species to colonize the entire suitable territory (see Lobo *et al.*, 2006). That is, if one includes a measure of connectivity between areas, or a "distance cost" (Hortal *et al.*, 2005), one can quantify the dispersal effort necessary to inhabit areas farther from the area with well-known presences, directly integrating dispersal models and environmental data (Iverson *et al.*, 2004). When the biological, historical and physiological information necessary to describe the realized distribution of species is lacking, our predictions should continue to be based on correlative statistical models in which the role of non-environmental processes must be considered. Good models need good data. Thus, the elaboration of reliable simulations on the realized distribution of species unavoidably requires good absence data, as well as the inclusion of non-environmental processes in the model procedure. Our study shows that if we do not have reliable absence data the method of pseudo-absence selection strongly conditions

the obtained model, generating different model predictions in the gradient between potential and realized distributions.

ACKNOWLEDGEMENTS

Special thanks to Alberto Jiménez-Valverde and Joaquín Hortal for their valuable suggestions. This paper has been supported by a Fundación BBVA project (Diseño de una red de reservas para la protección de la Biodiversidad en América del sur austral utilizando modelos predictivos de distribución con taxones hiperdiversos) and a MEC Project (CGL2004-04309).

REFERENCES

Anderson, R. P. 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography* **30**(4), 591-605.

Austin, M. P. 1980. Searching for a model for use in vegetation analysis. *Plant ecology* **42**(1-3), 11-21.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecological Modelling* **157**(2-3), 101-118.

Baixeras, J. 2001. Seguimiento de poblaciones de *Graellsia isabelae* en zonas de actuación del proyecto Life-habitats. Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, Valencia.

Bisby, F. A. 2000. The quiet revolution: biodiversity informatics and the internet. *Science* **289**(5488), 2309-2312.

Brotons, L., Thuiller, W., Araújo, M. B. & Hirzel, A. H. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* **27**(4), 437-448.

Busby, J. R. 1991. BIOCLIM- a bioclimate analysis and prediction system. In: *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* (edited by Margules, C. R. & Austin, M. P.). CSIRO, Melbourne, 64-68.

Chefaoui, R. M., Hortal, J. & Lobo, J. M. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation* **122**(2), 327-338.

Clark, L. 2000. Global Change Data Archive Vol. 3. 1 km Global Elevation Model. Clark University.

Clark, L. 2003. Idrisi Kilimanjaro, Worcester, MA.

Collingham, Y. C., Wadsworth, R. A., Huntley, B. & Hulme, P. E. 2000. Predicting the spatial distribution of alien riparian species: issues of spatial scale and extent. *Journal of Applied Ecology* **37**, 13-27.

Davis, A. J., Jenkinson, L. S., Lawton, J. H., Shorrocks, B. & Wood, S. 1998. Making mistakes when predicting shifts in species range in response to global warming. *Nature* **391**(6669), 783-786.

Dennis, R. L. H. & Hardy, P. B. 1999. Targeting squares for survey: predicting species richness and incidence of species for a butterfly atlas. *Global Ecology & Biogeography* **8**(6), 443-454.

Dixon, P. M., Ellison, A. M. & Gotelli, J. 2005. Improving the precision of estimates of the frequency or rare events. *Ecology* **86**, 1114-1123.

Engler, R., Guisan, A. & Rechsteiner, L. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* **41**(2), 263-274.

Fernández-Vidal, E. H. 1992. Comentarios acerca de la distribución geográfica francesa y notas taxonómicas sobre *Graellsia isabelae* (Graells, 1849). *SHILAP Revista de Lepidopterología* **20**(77), 29-49.

Ferrier, S. & Watson, G. 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. NSW National Parks and Wildlife Service Department of Environment, Sport and Territories, Canberra, Australia.

Fielding, A. H. 2002. What are the appropiate characteristics of an accuracy measure? In: *Predicting Species Occurrences: Issues of Accuracy and Scale* (edited by J. M. Scott, P. J. Heglund, J. B. Haufler, M. L. Morrison, M. G. Raphael, W. A. Wall & F. B. Samson). Island Press, Washington, DC, 271-280.

Galante, E. & Verdú, J. R. 2000. *Los Artrópodos de la "Directiva Hábitat" en España*. Ministerio de Medio Ambiente, Madrid.

García-Barros, E. & Herranz, J. 2001. Nuevas localidades de *Proserpinus proserpina* (Pallas, 1772) y *Graellsia isabelae* (Graells, 1849) del centro peninsular. *SHILAP Revista de Lepidopterología* **29**(114), 183-184.

Gu, W. & Swihart, R. K. 2004. Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biological Conservation* **116**, 195-203.

Guisan, A., Edwards Jr., T. C. & Hastie, J. T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* **157**, 89-100.

Guisan, A. & Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters* **8**, 993-1009.

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

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

Hirzel, A., Hausser, J., Chessel, D. & Perrin, N. 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology* **83**(7), 2027-2036.

Hirzel, A. H. & Arlettaz, R. 2003. Modelling habitat suitability for complex species distributions by the environmental-distance geometric mean. *Environmental Management* **32**(5), 614-623.

Hirzel, A. H., Hausser, J. & Perrin, N. 2004. Biomapper 3.1. Lab. of Conservation Biology, Department of Ecology and Evolution, University of Lausanne, Laussane, Switzerland.

Hirzel, A. H., Helfer, V. & Metral, F. 2001. Assessing habitat-suitability models with a virtual species. *Ecological Modelling* **145**, 111-121.

Hortal, J., Nieto, M., Rodríguez, J. & Lobo, J. M. 2005. Evaluating the roles of connectivity and environment on faunal turnover: patterns in recent and fossil Iberian mammals. In: *Migration in Organisms-Climate, Geography, Ecology* (edited by Elewa, A. M. T.). Springer-Verlag, Berlin, Germany, 301-328.

Hosmer, D. W. J. & Lemeshow, S. 2000. *Applied Logistic Regression*. John Wiley & Sons, New York.

Instituto Geográfico Nacional. 1995. Atlas nacional de España, vol. 1-2. Centro Nacional de Información, Madrid, Spain.

Iverson, L. R., Schwartz, M. W. & Prasad, A. M. 2004. How fast and far might tree species migrate in the eastern United States due to climate change? *Global Ecology and Biogeography* **13**, 209-219.

Jiménez-Valverde, A. & Lobo, J. M. 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica* **31**, 361-369.

Jiménez-Valverde, A. & Lobo, J. M. 2006. The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions* **12**(5), 521-524.

Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions* **14**(6), 885-890.

King, G. & Zeng, L. 2000. Logistic regression in rare events data. The Global Burden of Disease 2000 in Aging Populations, Research Paper No. 2 (edited by Havard Burden of Disease Unit, C. f. P. a. D. S.). http://www.hsph.harvard.edu/burdenofdisease/publications/papers/Logistic%20Regression.pdf

Legendre, P. & Legendre, L. 1998. *Numerical Ecology*, Amsterdam, Holland.

Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. 2005. Selecting thresholds of ocurrence in the prediction of species distributions. *Ecography* **28**(3), 385-393.

Lobo, J. M., Jay-Robert, P. & Lumaret, J. P. 2004. Modelling the species richness for French Aphodiidae (Coleoptera, Scarabaeoidea). *Ecography* **27**, 145-156.

Lobo, J. M., Verdú, J. R. & Numa, C. 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions* **12**(2), 179-188.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G. & Williams, P. H. 2003. Avoiding Pitfalls of Using Species Distribution Models in Conservation Planning. *Conservation Biology* **17**(6), 1591-1600.

López-Sebastián, E., López, J. C., Juan, M. J. & Selfa, J. 2002. Primeras citas de mariposa isabelina en la Comunidad Valenciana. *Quercus* **193**, 10-13.

Manel, S., Dias, J. M. & Ormerod, S. J. 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling* **120**, 337-347.

McCullagh, P. & Nelder, J. A. 1989. *Generalized Linear Models*. Chapman & Hall, London, England.

Olden, J. D., Jackson, D. A. & Peres-Neto, P. 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society* **131**, 329-336.

Pulliam, H. R. 1988. Sources, Sinks, and Population Regulation. *The American Naturalist* **132**(5), 652-661.

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

Reutter, B., Helfer, V., Hirzel, A. H. & Vogel, P. 2003. Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography* **30**, 581-590.

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

Rushton, S. P., Ormerod, S. J. & Kerby, G. 2004. New paradigms for modelling species distributions? *Journal of Applied Ecology* **41**, 193-200.

Schröder, B. 2004. ROC Plotting and AUC Calculation Transferability Test. Institute for Geoecology, Potsdam University. Potsdam. http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html.

Segurado, P. & Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. *Journal of Biogeography* **31**(10), 1555-1568.

Skov, F. & Svenning, J. C. 2004. Potential impact of climate change on the distribution of forest herbs in Europe. *Ecography* **27**(3), 366-380.

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

Sokal, R. R. & Rohlf, F. J. 1981. *Biometry*. Freeman, New York.

StatSoft, I. 2001. STATISTICA (data analysis software system), version 6.

Stockwell, D. & Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13**(2), 143-158.

Svenning, J. C. & Skov, F. 2004. Limited filling of the potential range in European tree species. *Ecology Letters* **7**(7), 565-573.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C. & Erasmus, B. F. N., Ferreira de Siqueira, M., Grainger, A., Hannah, L., Hughes, L., Huntley, B., S. van Jaarsveld, A., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L., & Williams, S.E. 2004. Extinction risk from climate change. *Nature* **427**, 145-148.

Thuiller, W., Brotons, L., Araújo, M. B. & Lavorel, S. 2004. Effects of restricting environmental range of data to project current and future species distributions. *Ecography* **27**, 165-172.

Verdú, J. R. & Galante, E. 2002. Climatic stress, food availability and human activity as determinants of endemism patterns in the Mediterranean region: the case of dung beetles (Coleoptera, Scarabeoidea) in the Iberian Peninsula. *Diversity and Distributions* **8**, 259-274.

Viejo, J. L. 1992. Biografía de un naturalista y biología del lepidóptero por él descrito. Graells y la *Graellsia*. *Quercus*(74), 22-30.

Zaniewski, A. E., Lehmann, A. & Overton, J. M. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* **157**(2-3), 261-280.

Zimmermann, N. E. & Kienast, F. 1999. Predictive mapping of alpine grasslands in Switzerland: Species versus community approach. *Journal of Vegetation Science* **10**(4), 469-482.

Zweig, M. H. & Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**(4), 561-577.

Evaluación de las variables ambientales más relevantes para explicar la distribución de *Graellsia isabelae* y delimitación de áreas importantes para su conservación.

RESUMEN

Conocer la distribución de las especies amenazadas es muy importante para su conservación. En este trabajo se examina un método útil para modelizar la distribución de especies cuando se maneja información de atlas y museos, sin disponibilidad de datos fiables de ausencia. Creando un modelo de la distribución de *Graellsia isabelae*, una especie amenazada de polilla, evaluamos su estado de conservación actual e identificamos las variables explicativas más relevantes para su distribución mediante Sistemas de Información Geográfica y Modelos Lineales Generalizados. El modelo de distribución fue realizado a partir de 136 datos de presencia y 25 variables explicativas digitalizadas a una resolución de 10 x 10 km. Los modelos predictivos se elaboraron mediante la regresión logística de pseudo-ausencias identificadas a partir de un método que usa solamente presencias (Ecological-Niche Factor Analysis; ENFA), y se obtuvo un valor explicativo de la varianza total de 96,23%. Se encontró que las mejores variables predictivas fueron la precipitación durante el verano, la aridez y la altitud media. En relación a las plantas hospedadoras, la presencia de *G. isabelae* se asoció principalmente con el pino albar (*Pinus sylvestris*) y el pino salgareño (*P. nigra*). La existencia de 8 poblaciones, exclusivamente en la zona este de la Península, y una amplia zona adecuada para la especie y no ocupada en la parte oeste de la Península

Ibérica, indica que la especie probablemente no está en equilibrio con el ambiente debido a factores históricos. La protección de los Lugares de Interés Comunitario (LICs) no parece ser suficiente para mantener las poblaciones actuales, siendo necesaria la protección de hábitats adecuados para la especie en lugares limítrofes a ellos. Nuestra metodología es útil para manejar la conservación de especies para las que no hay datos de ausencia disponibles. Ha sido posible determinar aquellas variables que afectan en mayor medida a la distribución así como las áreas potencialmente adecuadas para la especie con el propósito de evaluar las áreas protegidas, la conectividad entre poblaciones y las posibles reintroducciones.

Palabras clave: modelos de distribución, *Graellsia isabelae*, Península Ibérica, no equilibrio, pseudo-ausencias, especies amenazadas.

Este capítulo ha sido publicado como:

Chefaoui, R. M. & Lobo, J. M. 2007. Assessing the conservation status of an Iberian moth using pseudo-absences. *The Journal of Wildlife Management* **71**(8), 2507-2516.

Assessing the Conservation Status of an Iberian Moth Using Pseudo-Absences.

ABSTRACT

Knowing the distribution of endangered species is of substantial importance for their conservation. Here we evaluate a useful approach for modeling species distribution when managing information from atlases and museums but when absence data is not available. By modeling the distribution for *Graellsia isabelae*, a threatened moth species, we assessed its current conservation status and identified its most relevant distribution explanatory variables using Geographic Information System and Generalized Linear Models. The distribution model was built from 136 occurrence records and 25 digitized explanatory variables at a 10 x 10 km resolution. Model predictions from logistic-regressed pseudo-absences, obtained from a presence-only method (Ecological-Niche Factor Analysis), explained 96.23% of the total deviance. We found that the best predictor variables were summer precipitation, aridity, and mean elevation. With respect to host plants, the presence of *G. isabelae* associated mainly with Scots pine (*Pinus sylvestris*) and Austrian pine (*P. nigra*). The finding of 8 areas, exclusively in the eastern Iberian territory, and a larger unoccupied habitat in the western Iberian Peninsula indicates that this species is probably not in equilibrium with its environment because of historical factors. Sites of Community Importance under protection do not seem sufficient to maintain current populations, necessitating the protection of suitable neighboring habitats. Our methodology is useful to assess the

conservation status of species for which reliable absence data is not available. It is possible to determine those variables that most affect the distribution of species as well as the potential suitable areas with the purpose of evaluating protected areas, connectivity among populations, and possible reintroductions.

Keywords: distribution modeling, *Graellsia isabelae*, Iberian Peninsula, nonequilibrium, pseudo-absences, threatened species.

INTRODUCTION

The prediction of species' distributions is relevant to diverse applications in evolution, ecology, and conservation science. Producing accurate predictions with available data is challenging due to the lack of information regarding the great majority of species. In order to solve the limitations in data, several statistical techniques and computer tools for data management have been combined for the purpose of obtaining information about the conservation status, geographic distribution, and habitat requirements of endangered species.

The conservation of biodiversity is a priority that has lead to the elaboration of multiple Red Lists in an effort to document the status of endangered species. Analyses of habitat requirements, distribution, and habitat suitability of threatened and endangered species can help to compensate for the lack of information on their ecology, a major obstacle to conservation, especially of invertebrates. Computer tools such as Geographic Information System (GIS) and statistical modeling techniques applied to information from atlases and museums can be used to draw up predictive maps of the

requirements and conservation status of such species (Dennis & Hardy, 1999; Reutter *et al.*, 2003; Chefaoui *et al.*, 2005). Because sampling and identification are laborious tasks, predictions for regions not yet exhaustively surveyed can be based on pseudo-absences (Zaniewski *et al.*, 2002; Engler *et al.*, 2004; Lobo *et al.*, 2006). Good absence data is fundamental to consistent models (see Anderson *et al.*, 2003 or Loiselle *et al.*, 2003). Unfortunately, our maps were not drawn up from reliable absence data. Thus, probable absence localities far from the environmental domain defined by presences may be selected for the modeling of species distribution to avoid false absences that can decrease model reliability. With the goal of obtaining a predictive model based on generalized linear models (GLM) without reliable absence data, we procured the pseudo-absences with a presence-only method. This strategy allows researchers to use presence data alone to obtain distribution models able to delimit the potential range of species (Svenning & Skov, 2004).

Graellsia isabelae (Lepidoptera: Saturniidae; Graëlls, 1849) is a host-limited species protected by the Habitats Directive (the European Community initiative for an ecological network of special protected areas, known as Natura 2000), Bern Convention, Red Book of Lepidoptera, and other regional catalogues. Over the last 30 years, the decline of European butterflies (Warren *et al.*, 2001; Wenzel *et al.*, 2006) has occurred mostly in specialist and sedentary species (Steffan-Dewenter & Tscharntke, 2000; Thomas, 2000), highlighting the need to protect species with characteristics similar to those of *G. isabelae*.

Graellsia isabelae has a sedentary and nongregarious caterpillar, develops in 5 stages, and dwells in pine forests. Larvae feed from June to August before pupating. It is a univoltine species that flies at dusk, from March to July (only 1 brood/yr; Masó &

Ylla, 1989). Because *G. Isabelae* is an emblematic and conspicuous species (the beautiful adults are markedly sexually dimorphic), occurrence records can be considered reliable (Ylla, 1997).

There is controversy about larval food plants as various authors (Agenjo, 1943; Gómez-Bustillo & Fernández-Rubio, 1974; Vuattoux, 1984; Masó & Ylla, 1989; Ylla, 1997) have cited different pine species: Scots pine, Austrian pine, dwarf mountain pine (*P. uncinata*), aleppo pine (*P. halepensis*), maritime pine (*P. pinaster*), and stone pine (*P. pinea*) as possible host plants based on captivity experiments.

There are 9 described subspecies of *G. isabelae* (Vives, 1994), but the real variability among populations is not clear. Among them, the autochthonous origin of *G. Isabelae galliaegloria* (Oberthür, 1923), present in the Jura and Alps mountains of France and Switzerland, is controversial (Fernández-Vidal, 1992; Ylla, 1997). With the exception of these populations, the distribution of the insect is eastern Iberian.

We aimed to estimate the potential distribution of *G. isabelae* on the Iberian Peninsula and also to identify the explanatory variables most relevant to its occurrence. In an effort to identify the current conservation status of *G. isabelae*, we examined suitable regions for fragmentation, degree of connection, and the area currently under protection.

METHODS

Study area

As *G. isabelae* had an eastern Iberian distribution, the study area was the Iberian Peninsula (excluding the Portuguese territory but including the Balearic Islands), which included the whole known range of the species. In total, the area comprised 498,150

km^2 divided into 5,270 cells of 10 x 10 km, for which corresponding biological and environmental data were described. We chose this resolution because the majority of biological data were originally referred to by that pixel size and the relation between grid size and study area was appropriate.

Biological Data

We obtained species presence data mainly from a distribution atlas (Galante & Verdú, 2000), additional unpublished data from the Valencia region (J. Baixeras, University of Valencia, personal communication), and other bibliographic references (Viejo, 1992; García-Barros & Herranz, 2001; López-Sebastián *et al.*, 2002). As species data came from diverse sources and some references were old, we checked all references by comparing species locations with pine woods distributions to eliminate possible outliers. We discarded 6 presence data points with probably erroneous Universal Transverse Mercator (UTM) coordinates. Finally, we considered 136 presence points at a UTM spatial resolution of 100 km^2.

Predictor Variables

We used IDRISI Kilimanjaro GIS software to set up the explanatory variables we introduced in the preparation of distribution models (Table 3.1) from different sources. We extracted topographic variables (elevation and slope) from a global digital elevation model with 1-km spatial resolution (Clark Labs, 2000), and we calculated aspect diversity using the Shannon index. Temperature and precipitation data at 1-km resolution were provided by the Spanish Instituto Nacional de Meteorología. We calculated aridity as $Ia = 1/(P/T + 10) \times 10^2$, where P is the mean annual precipitation

and *T* the mean annual temperature (see Verdú & Galante, 2002). We extracted from a forest map woods containing the host plants species cited above (Ruíz, 2002). We included in the analysis as a predictor variable the area of each forest patch present in each cell with respect to the different kinds of pine woods. In addition, we digitized 4 lithology variables from a lithology map (Instituto Geográfico Nacional, 1995) to calculate the area of calcareous deposits, siliceous sediments, stony acidic, and calcareous soils for each cell. Spatial variables were the central latitude (Lat) and longitude (Lon) of each UTM cell and we derived their polynomial transformations from Trend Surface Analyses. The inclusion of these variables can help to determine unaccounted-for variable influence on species distribution (see Legendre & Legendre, 1998). All continuous explanatory variables referred to the same 10 x 10 km UTM grid cells as those of species data using IDRISI Kilimanjaro's Resample and Contract modules. We standardized the predictor environment variables to zero means and one standard deviation to eliminate the effect of varying measurement scales. We also standardized latitude and longitude in the same way as environment variables.

Predictive Distribution Models

Because accurate absence data were not available, we used a presence-only modeling technique (Ecological-Niche Factor Analysis [ENFA]) to map habitat suitability, from which we selected pseudo-absences to be used with presence data in a logistic regression procedure (GLM; see Engler *et al.*, 2004). We applied ENFA to presence data (n = 136) and the 28 predictor variables (Table 3.1) by means of BIOMAPPER 3.1 software (Hirzel *et al.*, 2004), which was designed to build habitat suitability models

and maps for any species. The principle of ENFA is to compare the distributions of the predictor variables between the species distribution and the whole area. This modeling

Variables	df	Deviance	% explained deviance	Sign
Environmental variables				
f^2 summer precipitation	1,493	222.03	75.63	− +
f^2 max. annual temp	1,493	400.89	56.01	+ +
f^3 aridity	1,492	420.19	53.89	+ + −
f^2 min. annual temp	1,493	533.82	41.43	+ +
f^3 $\bar{x}$ elevation	1,492	552.77	39.35	− − +
f^2 annual precipitation	1,493	577.93	36.59	− +
f^3 slope	1,492	698.18	23.39	− + −
f^3 aspect diversity	1,492	857.84	5.88	+++
Vegetation				
f^2 *P. sylvestris*	1,493	568.27	37.65	− +
f^3 total area with any *Pinus* sp.	1,492	594.63	34.76	− + −
f^3 *P. sylvestris* and *P. nigra*	1,492	621.43	31.82	− + −
P. nigra	1,494	798.67	12.37	−
f^2 *P. nigra* and others	1,493	837.18	8.15	− +
P. sylvestris and *P. uncinata*	1,494	852.62	6.45	−
P. sylvestris and others	1,495	867.75	4.79	−
Groves of *P. sylvestris* and *P. nigra*	1,494	879.87	3.46	−
P. halepensis, *P. pinaster* and *P. pinea*	1,494	889.62	2.39	+
P. uncinata	1,494	906.74	0.51	−
Mixture of pines	1,494	910.96	0.05	
Lithology variables				
Calcareous stony soils	1,494	635.97	30.22	−
Acidic sediments	1,494	723.15	20.66	+
Acidic stony soils	1,494	902.12	1.02	+
Calcareous sediments	1,494	906.59	0.53	−
Spatial variables				
Longitude^2 x latitude	1,494	647.43	28.96	−
f^3 latitude	1,492	658.39	27.76	− + −
Longitude	1,494	671.94	26.27	−
Longitude x latitude^2	1,494	750.5	17.65	−
Longitude x latitude	1,494	859.85	5.66	−

Table 3.1 - Individual logistic regression of presence-absence of *Graellsia isabelae* against each one of the selected explanatory variables, indicating relationships as linear, quadratic (f^2), or cubic (f^3). Biological data were collected from Spain (since 1849). The sign column indicates the sign for each selected term of each function. We chose spatial variables after backward-stepwise elimination of nonsignificant terms from a third-degree polynomial of latitude and longitude.

technique (similar to Principal Component Analysis in that it generates orthogonal axes) computes a group of uncorrelated factors with ecological meaning, summarizing the main environmental gradients in the region considered. These factors are 1) the marginality factor, which describes how far the species optimum is from the mean habitat in the study area, and 2) the specialization factors that describe how specialized the species is. We used the selected factors to estimate the degree of ecogeographic similarity of each grid cell to the environmental preferences of the species, that is, the probability of a given cell belonging to the environmental domain of the presence observations. From this, we drew up a habitat suitability (HS) map with values that varied from zero (min. HS) to 100 (max.). We normalized predictor variables through a box-cox transformation (Sokal & Rohlf, 1981), and we chose a geometric-mean distance algorithm, which provides a good generalization of the niche (Hirzel & Arlettaz, 2003), to perform the analyses.

Unsuitable habitats determined by this profile technique help to identify reliable pseudo-absences for presence-absence modeling. To avoid bias due to the inclusion of a comparatively higher number of absences (King & Zeng, 2000), we randomly selected 10 times more absences (1,360) than presences from the model. We chose pseudo-absences from unsuitable habitats with HS < 10, a threshold value that has been shown to produce good validation results (Chefaoui & Lobo, 2008). We regressed the 136 presence data points and the 1,360 pseudo-absences selected from the presence-only model using logistic regressions in GLM. Generalized Linear Models are an extension of the classical linear regression models that allow for nonlinearity in the data as well as a range of independent variable distributions other than the normal (McCullagh & Nelder, 1989). The relationship between the dependent and the explanatory variables

(the link function) was logistic, and we assumed a binomial error distribution of the dependent variable.

To perform a statistical analysis of the variables, we first related the presence-absence data of the species for the 10 x 10 km UTM cells under consideration separately to each predictor variable. First, to consider possible curvilinear relationships, we selected for inclusion the linear, quadratic, or cubic function of the variable that accounts for the most important change in deviance with significant terms (Austin, 1980). With this procedure, we identified the most relevant explanatory variables. Next, we built 4 models from each of the variable sets to estimate the relative relevance of each group of explanatory variables: an environmental model (E), a vegetation model (V), a lithology model (L), and a spatial model (S). Subsequently, we accomplished different models considering all possible combinations among the 4 types of variables (E, V, L, and S). We used Akaike's Information Criterion (AIC), the measures associated with it (ΔAIC, Akaike wt, and Model likelihood; Hastie *et al.*, 2001; Burnham & Anderson, 2002), and the percentage of explained deviance values to choose between competing models. We used the STATISTICA 6.0 package (StatSoft Inc., Tulsa, OK) for all statistical computations.

We used the Receiver Operating Characteristic curve (ROC; Zweig & Campbell, 1993; Schröder, 2004) to estimate model accuracy. A ROC curve is a plot of sensitivity (the ratio of correctly predicted presences to the total no. of presences) versus 1-specificity (false positive rate) as the threshold changes, and the calculation of the area under this curve (AUC) provides a single number performance measure across all possible ranges of thresholds (Fielding, 2002). However, when a cut-off point was needed to transform continuous probabilities obtained in GLM models to binary ones

(presence-absence), we used the sensitivity-specificity difference minimizer (Liu *et al.*, 2005; Jiménez-Valverde & Lobo, 2006) to select this threshold, due to its generally good performance. Because the small sample size did not allow for the performance of ROC analysis with independent data, we obtained model validation scores by means of a jackknifing procedure (see Olden *et al.*, 2002; Engler *et al.*, 2004) developed with R v.2.2.0 (R Development Core Team, Vienna, Austria) just in those models with better results in AIC-derived measures and deviance scores. In the jackknife procedure, with a data set of *n* observations, the model was recalculated *n* times, leaving out one observation each turn. We then applied each of the regression models based on the *n* -1 observations to the excluded observation, obtaining a probability value for each of the observations. We subsequently used these jackknifing probabilities, together with the binary dependent scores, to calculate AUC, sensitivity, and specificity.

We used hierarchical partitioning to measure the relative importance of each type of explanatory variable (Birks, 1996; MacNally, 2000, 2002). First, we calculated the percentage of explained deviance for each type of variable, as well as the variability explained by all possible variable combinations. Subsequently, we calculated the average effect of inclusion of each type of variable in all models for which this type of variable was relevant. We took such averages as estimations of the independent contribution of each type of explanatory variable.

Conservation

Because *G. isabelae* is a host-limited moth, we used pine species significantly related with the distribution of this insect species to filter potential predictive model habitats as a means of identifying currently suitable regions. After assigning a buffer area of 10 km

around each suitable UTM cell, we identified groups of connected cells (habitat patches) that maintain a priori separate populations. We selected the buffer size in accordance with dispersal distance data cited by other authors (Montoya & Hernández, 1975; Baixeras, 2001).

Because habitat area has been shown to greatly influence the conservation of European specialist butterflies (Steffan-Dewenter & Tscharntke, 2000; Wahlberg *et al.*, 2002; Krauss *et al.*, 2003), we calculated the area, perimeter, and a compactness ratio (Clark Labs, 2003) for each habitat patch. The compactness ratio compares the patch area: perimeter ratio with that of a circle of the same perimeter. We characterized isolation of populations or patches by the maximum, minimum, and mean distance to the nearest occupied patch or region, computed as the distance to the closest edge of nearby patches.

To examine the possible distribution expansion through reforested pine woods, reflected in the reforestation relationship with species occurrence, we used a contingency table with Cramer's V coefficient, a measure of the strength of variable association (Ott *et al.*, 1983; Clark Labs, 2003). We made Gap Analysis of habitat patches and Natura 2000 protected Sites of Community Importance (SCIs) to evaluate the current conservation status of *G. isabelae*.

RESULTS

Relevant Explanatory Variables

Regression of each variable separately showed that environmental variables, mainly those related to precipitation and temperature, were most relevant to the prediction of *G. isabelae* distribution (Table 3.1). Vegetation variables also explained the occurrence of

G. isabelae, although they were less relevant. Mean value of percentage of explained deviance for environmental variables was 31.18% larger than for vegetation variables. The presence of Scots pine and Austrian pine woods seemed to be the most important vegetation variable, along with the highly correlated total area of *Pinus* species. With regard to lithology variables, species presence correlated negatively (sign −) with calcareous stony soils but positively (sign +) with acidic sediments; both variables explained >45% of the total variability in species presence (see lithology model; Table 3.2). Lastly, spatial variables also explained the occurrence of *G. isabelae*, confirming that the Iberian distribution of this species forms a spatially structured pattern.

The complete environmental model accounted for >95% of total variability, an astonishing percentage of variability that none of the other types of predictor variables could explain (Table 3.2). The addition of vegetation, lithology, and spatial variables slightly increased (<1%) the explanatory capacity of the environmental variables. The mean percentage of variation accounted for by vegetation, lithology, and spatial variables was 22.25%, 12.34%, and 11.09%, respectively.

Predictive Models

The analysis for model selection suggested that the model carried out with environmental, lithological, and spatial variables (E + L + S) and the model encompassing only the environmental and lithological variables (E + L) were those that had higher percentages of explained deviances, lower AIC values, and the best model likelihood (Table 3.2). However, jackknifing validation indicated that the E + L + S model had higher accuracy scores; its AUC, sensitivity, and specificity scores were 0.9841, 0.9705, and 0.9977, respectively, whereas the results for the E + L model were

Predictive models	df	Deviance	% explained deviance	AIC	ΔAIC	AIC wt	Model likelihood
E	1,488	43.67	95.20	59.67	5.99	0.0135	0.0500
V	1,486	187.24	79.45	207.24	153.56	0.0000	0.0000
L	1,491	500.58	45.08	510.57	456.89	0.0000	0.0000
S	1,490	539.85	40.77	551.85	498.17	0.0000	0.0000
E + V	1,487	39.38	95.67	57.38	3.70	0.0424	0.1572
E + L	1,487	35.68	96.08	53.68	0.00	0.2695	1.0000
E + S	1,488	49.26	95.59	65.26	11.58	0.0008	0.0031
V + L	1,483	121.40	86.68	147.36	93.68	0.0000	0.0000
V + S	1,484	229.41	86.89	143.41	89.73	0.0000	0.0000
L + S	1,490	176.40	80.64	188.44	134.76	0.0000	0.0000
E + V + L	1,487	35.68	96.08	53.68	0.00	0.2695	1.0000
E + V + S	1,487	42.09	95.67	60.09	6.41	0.0109	0.0406
E + L + S	1,486	34.31	96.23	54.31	0.63	0.1967	0.7298
V + L + S	1,486	82.92	90.90	102.92	49.24	0.0000	0.0000
E + V + L + S	1,486	34.31	96.23	54.31	0.63	0.1967	0.7298

Table 3.2 - Deviance, percentage of explained deviance, Akaike Information Criterion (AIC), ΔAIC, AIC weight, and model likelihood values for each one of the generalized linear model accomplished with each type of explanatory variables and with all possible variable combinations. To perform environmental (E), vegetation (V), lithology (L) and spatial (S) models we used *Graellsia isabelae* data from Spain since 1849. Variables that constitute models E + V + L and E + L are coincident because vegetation variables did not contribute to models when added, so these two models are equivalent (the same with E + V + L+ S and E + L + S).

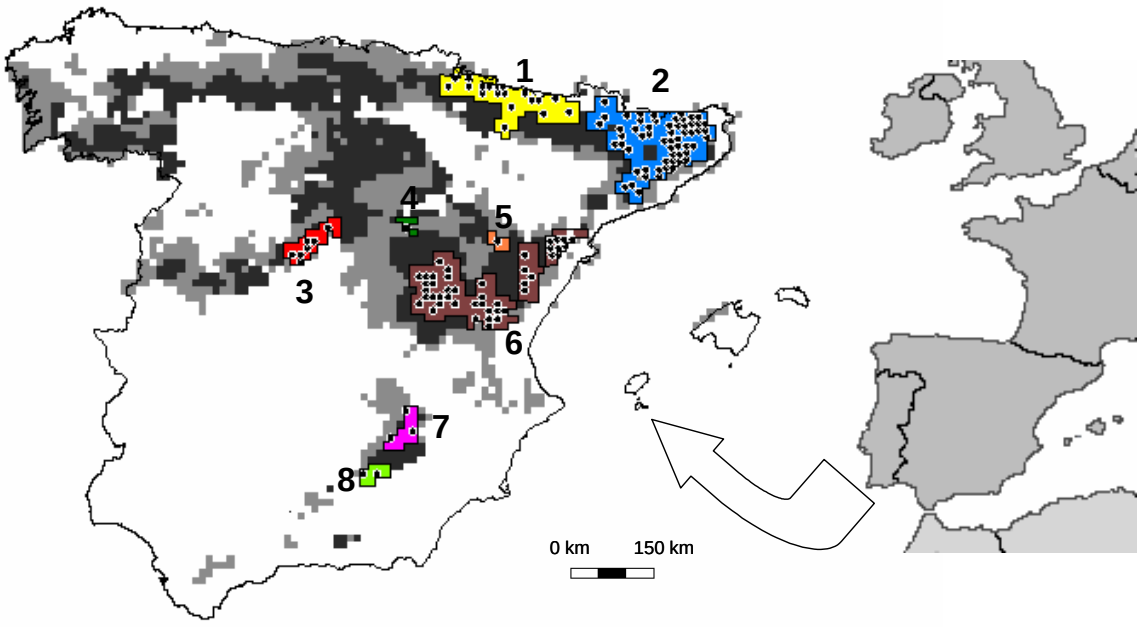


Fig. 3.1 - Distribution of occurrence data for *Graellsia isabelae* in Spain since 1849 (black dots), suitable area according to the final Generalized Linear Model (light grey), and suitable area with Scots pine and Austrian pine woods (dark grey). Habitat patches (in colours): 1) Pyrenees, 2) Catalonia, 3) Sierra of Guadarrama, 4) Anguita (Guadalajara), 5) Montalbán (Teruel), 6) Iberian System, 7) Sierras of Segura and Alcaraz, 8) Sierra of Cazorla.

0.9834, 0.9705, and 0.9963, respectively.

This E + L + S model explained >96% of the total variability when we considered all significant explanatory variables together in order of importance (Table 3.2). Variables retained in the final model were summer precipitation, aridity, mean elevation, slope, calcareous stony soils, calcareous sediments, and latitude (Table 3.3). After selecting the appropriate threshold value (0.09) for this final model, we converted the continuous GLM probability map values to binomial (Fig. 3.1). The total a priori suitable area, found to be about 203,100 km^2, was reduced to a current suitable area of 114,500 km^2 after filtering for the presence of the appropriate pine woods (Scots pine and Austrian pine).

The regions considered to be suitable exhibit very specific environmental conditions (Table 3.4), in which precipitation is higher than the average for the entire territory, and temperatures and aridity lower than the Iberian average. Elevation and slope scores were also higher than those observed for the whole territory, showing that the general environmental conditions for this species were those found in medium elevation mountain ranges. This final distribution model also revealed the existence of unoccupied regions, potentially suitable for *G. isabelae*, in the northwestern quadrant of the Iberian Peninsula (Cantabric mountains, Zamora and Galician mountains, the western area of the Iberian Central System, and the Iberian System), and in some southern mountains (Sierra Nevada).

Parameters	Wald	*P*-values	Coeff.	SE
E + L				
Summer precipitation	17.83	< 0.001	-13.03	3.08
Summer precipitation^2	14.67	< 0.001	2.68	0.70
Max. annual temp	8.76	0.003	2.27	0.76
Aridity	13.10	< 0.001	6.14	1.69
Min. annual temp	10.65	0.001	-10.12	3.10
$\bar{x}$ elevation	15.20	< 0.001	-12.04	3.08
Slope	10.59	0.001	2.10	0.64
Calcareous sediments	6.63	0.010	-1.70	0.66
E + L + S				
Summer precipitation	17.20	< 0.001	-10.29	2.48
Summer precipitation^2	18.74	< 0.001	2.69	0.62
Aridity	13.80	< 0.001	6.17	1.66
$\bar{x}$ elevation	8.85	0.003	-2.39	0.80
Slope	6.00	0.014	1.53	0.62
Calcareous stony soils	11.07	< 0.001	-2.62	0.78
Calcareous sediments	5.92	0.010	-1.72	0.70
Latitude	5.49	0.020	4.20	1.79
Latitude^2	7.33	0.007	2.77	1.02

Table 3.3 - Parameter estimates from environmental + lithology (E + L) and environmental + lithology + spatial (E + L + S) final generalized linear models (±SE) performed for Spanish *Graellsia isabelae* data since 1849. Wald statistic scores test the significance of regression coefficients.

Environmental variables	**Study area**		**Suitable habitat**			
	Min.-Max.	$\bar{x}$	**Min.**	**Max.**	$\bar{x}$	**SD.**
Summer precipitation (mm)	0-472.4	90.4	22.9	472.4	130.9	61.4
Max. annual temp (º C)	9.1-24.9	19.3	9.6	23.1	17.3	1.9
Aridity	0-1.64	0.41	0.07	0.6	0.3	0.1
Min. annual temp (º C)	-3.5-14.3	7.34	-3.5	12.7	5.2	2.2
$\bar{x}$ elevation (m)	0-2,722	665	2	2,632	907	374
Annual precipitation (mm)	0-2,200	698.9	365.4	2,165.2	844.9	386.6
Slope (º)	0-46.0	3.4	0	39	4.3	3.4

Table 3.4 - Environmental conditions of suitable areas for *Graellsia isabelae* in Spanish territory (with data since 1849). The suitable areas were defined by generalized linear model performed with environmental, lithology and spatial variables (E + L + S model not filtered with pine woods) compared with those of the entire study area.

Connectivity and Conservation Status

The presence of *G. isabelae* was observed in eight unconnected patches, comprising the buffered occurrence cells of the currently suitable area (see Fig. 3.1). The total area was 41,600 km^2, although the individual patch sizes were quite dissimilar; 67% of this area consisted solely of the two largest patches (Table 3.5). The mean distance to the nearest occupied patch was 47.2 km, being 10 km the minimum distance between patches (pairs 1-2 and 5-6) and 120 km the maximum (between patches 6 and 7). These habitat patches exhibited slightly differing compactness ratios (patches 1, 2, 3, and 6 with a larger edge effect; Table 3.5). The patches could be grouped in four regions connected by suitable habitat between them (see Fig. 3.1): 1-2 (Catalan-Pyrenean), 3 (Guadarrama Mountains), 4-5-6 (Iberian System and associated mountains), and 7-8 (Sierras of Cazorla, Segura and associated mountains).

Around 39% of the suitable area is currently protected by the Natura 2000 proposal (SCIs), with protected areas differing greatly from region to region. Included in this reserve proposal are >80% of regions 7 and 8, whereas <20% of regions 2 and 5 would be protected (Table 3.5). Finally, the presence of *G. isabelae* does not correlate significantly with reforested pine woods (Cramer's V score is 0.02); of the 136 presences, just 8 fell within reforested woods.

Region	Area (km^2)	Perimeter (km)	Compactness ratio	Area included in SCIs (%)	Suitable habitat bordering the region (%)	Oldest reference available (yr)
1 Pyrenees	7,700	660	0.47	35.0	51.5	1943
2 Catalonia	13,800	880	0.47	19.8	36.3	1920
3 Sierra of Guadarrama	2,400	320	0.54	65.7	25.0	1849
4 Anguita (Guadalajara)	500	140	0.56	47.2	35.7	1993
5 Montalbán (Teruel)	600	120	0.72	10.8	50.0	1974
6 Iberian System	14,100	1120	0.37	46.8	60.7	1920
7 Sierras of Segura and Alcaraz	1,600	220	0.64	83.5	40.9	1943
8 Sierra of Cazorla	900	140	0.75	99.3	28.5	1943

Table 3.5 - Main characteristics of *Graellsia isabelae* Iberian patches of occurrence data (since 1849). Numbers of each region are those shown in Fig. 1. SCIs are Sites of Community Importance protected by the Habitats Directive.

DISCUSSION

The Distribution Model

Suitable area identification was highly reliable; 97% of presences and 99% of environmentally derived pseudoabsences were correctly predicted. Pseudo-absence selection by a modeling method that requires only presence data and the inclusion of such absences in presence-absence modeling seems a promising distribution prediction procedure (see also Zaniewski *et al.*, 2002; Engler *et al.*, 2004; Lobo *et al.*, 2006).

Evidently, true distribution absences can never be distinguished with certainty from false ones, due to the lack of information on the species. In our case, the size and showiness of *G. isabelae* should have led to a reasonably well-known distribution. A nonequilibrium distribution pattern is supposed if a species does not occupy its entire suitable habitat. Because the occupied area is less than the potential derived in our final predictive model, we deduce that the species is not in equilibrium with the current climate. Given the nonequilibrium state, reliable absence information should be sought exclusively in environmentally favorable areas by standardized sampling to confirm

species absence there. In addition, the ability of *G. isabelae* to colonize these areas should be examined in the near future. However, the inclusion of reliable absence data from regions with environments similar to those with presences implies a reduction in the goodness-of-fit of models obtained (see Collingham *et al.*, 2000). Predictive distribution modeling assumes that the distribution of species is in an equilibrium or pseudo-equilibrium state (Guisan & Theurillat, 2000; Guisan & Zimmermann, 2000; Austin, 2002; Pearson & Dawson, 2003; Guisan & Thuiller, 2005). Because nonequilibrium with environmental variables will be common among some groups and in some regions (White *et al.*, 2001; Pearson *et al.*, 2002; Skov & Svenning, 2004; Araújo & Pearson, 2005), success in forecasting actual species distribution could depend on the inclusion of variables representing geographic, demographic, or historical factors that inhibit species distribution across all environmentally favorable locations.

The Most Relevant Variables

We showed that the main explanatory variables accounting for potential distribution may be identified by available species presence information alone. Our results demonstrate that *G. isabelae* does not need environmental conditions marginal to those of the Iberian Peninsula. We found that there are suitable habitats in a wide range of environments, with a preference for midrange mountainous conditions. Keeping in mind that changes in resolution and extent can alter the relevance of explanatory variables, we still found climate variables (summer precipitation, aridity, and mean elevation) to be the best predictors. The curvilinear relationship between summer precipitation and the presence-absence of *G. isabelae* is especially important (Fig. 3.2 and Table 3.1); precipitation from 1,250 mm to 3,250 mm makes species presence highly probable.

Among all possible host plants examined, species presence was shown to be related with Scots pine and Austrian pine, the pine woods most frequently cited in the literature. Other pine species occasionally cited as food in captivity studies (aleppo pine, maritime pine, stone pine, and dwarf mountain pine; see Agenjo, 1943; Gómez-Bustillo & Fernández-Rubio, 1974; Vuattoux, 1984; Masó & Ylla, 1989; Ylla, 1997) were only marginally related with *G. isabelae* presence.

Because *G. isabelae* feeds only on plants, we did not expect the best model predictions of its distribution to be independent of vegetation variables; although not altogether irrelevant, their inclusion, after environmental variables, did not increase model prediction accuracy. Similar findings presented by Warren *et al.* (2001) showed that the range limits of 46 British butterflies could be described by three bioclimatic variables. Such a result may be due to environmental collinearity between climatic and vegetation variables. The major part (82%) of preferred Scots pine and Austrian pine woods were located within the suitable environmental area, where both types of factors coincided. Because host plant distribution, generally wider than that of most lepidopteran species (Gutiérrez, 1997), also depends on environmental factors, it seems reasonable to begin with models based on environmental variables, then filter them with vegetation variables, rather than incorporate these latter variables at the beginning of the modeling process. Obtaining reliable models including solely vegetation variables would be possible if sufficient information about the studied species' host plant is available. The major difficulties are finding a species whose nutritive requirements are well known a priori along with access to precise vegetation maps.

The possible expansion of *G. isabelae* through reforested areas (Soria *et al.*, 1986; Robredo, 1988; López-Sebastián *et al.*, 2002) is not supported by our results.

Four old records found in reforested pine woods (two in 1943 and two in 1974) have not increased much in number over time (three in 1987 and just one in 2001). Hence there was no evidence supporting the expansion hypothesis, because these more recent records may be due to a greater sampling effort rather than expansion. Historic records and (or) nonreforested pine woods are found in each habitat patch, with the exception of patch 5 (Table 3.5), which is located in a reforested wood, perhaps the unique recent expansion.

We found lithology and spatial variables to be less relevant to *G. isabelae* distribution prediction. *G. isabelae* was linearly and negatively related with calcareous stony soil area but positively related with acidic sediments. The biological implication of these relationships is obscure, but the inclusion of two lithology variables in the final model based on all the variables considered confirmed their slight relevance (Table 3.1). We suggest that the slight negative influence of calcareous soils may be due to their poorer water retention, as well as their high mineral content. Lastly, the minor relevance of spatial variables showed that, after the inclusion of all aforementioned variables, no other spatially structured factors aided in accounting for potential species distribution.

The Nonequilibrium Distribution

As may be common, *G. isabelae* distribution is not in equilibrium with environmental conditions because models define only potential species distributions in which currently occupied and suitable areas still not colonized, or with extinct populations, are mixed. The suitable area was around 2.7 times larger than the occupied area and, interestingly, favorable areas lacking presence data fall mainly in northwestern Iberia and, to a lesser

extent, in some few southernmost Iberian localities. The current nonequilibrium state of this species suggests that other factors (mainly historical) may help to explain this

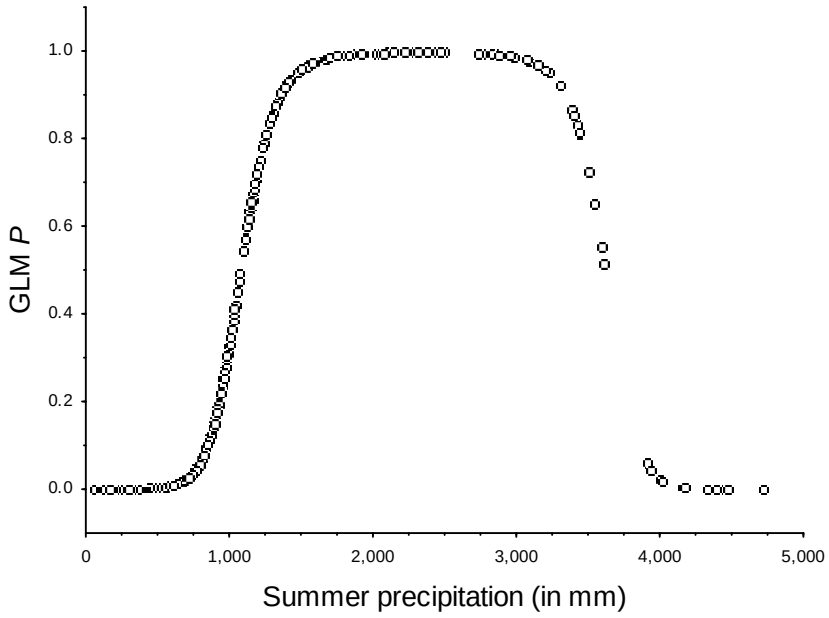


Fig. 3.2 - Relationship between summer precipitation and generalized linear model derived probability scores performed for *Graellsia isabelae* with Spanish locations data since 1849.

spatially biased distribution, whereas other contemporary ecological factors (predators, parasites, competitive interactions, or random extinctions) are more unlikely. We suggest that the current *G. isabelae* distribution could be associated with the dynamism of its host plants during glacial periods.

Pleistocene climate oscillations are known to have severely influenced the distribution patterns of most European animal and plant species (Hewitt, 2000; Schmitt & Krauss, 2004). Among the three main European refuges of Scots pine (the Iberian Peninsula, the Alps, and the Balkans; Bennett *et al.*, 1991) existing during the last glacial period, the Iberian one seems to have had populations that may have remained

isolated until now, without migration and expansion outside of the glacial refuges (Peñalba, 1994; Sinclair *et al.*, 1999; Soranzo *et al.*, 2000; Cheddadi *et al.*, 2006). The current endemicity of *G. isabelae* may be partially due to its association with these Iberian pine populations. Mitochondrial DNA and allozyme variation suggest that Scots pine survived in the Iberian Peninsula during the Pleistocene glaciations in central and eastern Iberian refuges (Sinclair *et al.*, 1999; Soranzo *et al.*, 2000). Pollen fossil evidence and recent potential range simulations indicate that both eastern and northwestern Iberian refuges existed for Scots pine during the last glacial maximum (Blanco *et al.*, 1998; Benito Garzón *et al.*, 2006; Cheddadi *et al.*, 2006). Interestingly, the map of Iberian *P. sylvestris* refuges recently established by Cheddadi and collaborators as well as the predicted distribution obtained by Benito Garzón *et al.* (2006) basically coincide with the potential range established by us for *G. isabelae*. If both suitable climatic and favorable host plant conditions exist for the presence of *G. isabelae* in the western area of Iberian Central System and the northwestern Cantabric Mountains, why is *G. Isabelae* currently absent of these territories? Amid such an arid, cold climate, highland pine woods were one of the most important arboreal elements during the glacial maximum, so the moth distribution should have been wider. However, recent charcoal data (Figueiral & Carcaillet, 2005) demonstrate that northwestern Iberian Scots pine populations decreased dramatically from Holocene times as a consequence of climate warming, competition with either angiosperms or other pine species (Blanco *et al.*, 1998), and anthropic factors as fires and grazing, probably causing the extinction of these *Graellsia* populations.

Conservation

Distribution maps drawn up from ten (Soria *et al.*, 1986) or six (Masó & Ylla, 1989) different Spanish populations of *G. isabelae* do not show the eight habitat patches, belonging to four main regions, well separated and with historical observational records, identified by our study. The major parts of these patches are surrounded by suitable, possibly connecting, areas for current populations. Probably the most disturbing situation exists in the Iberian Central System, where *G. isabelae* in neighboring suitable areas are scarcest. Although a high proportion, around 66%, of the area of this region is currently protected by Natura 2000 reserve design, <33% of the suitable habitat in all regions is protected. Although all regions would be partially protected by SCIs, just 38.8% of the total area is currently so; even worse, <20% of regions such as 2 and 5 are now protected. In our opinion, the conservation of woods located in protected sites does not seem sufficient to preserve current populations; preservation of surrounding suitable habitats is also necessary.

Some known causes for the decline in *Graellsia* populations are predation by predators such as European robin (*Erithacus rubecula*; Masó & Ylla, 1989), parasites, and entrapment in resins or phytosanitaries used to combat pine processionary moth (*Thaumetopoea pityocampa*; Soria *et al.*, 1986). However, habitat loss is the most obvious cause for depletion of lepidopteran populations. Short-term causes of habitat loss in Iberian Peninsula are urbanization (which has increased by 25% from 1990 to 2000; Ministerio de Medio Ambiente, 2005), forest fires (2.1 million ha burned from 1991 to 2004; WWF/Adena 2006), and logging. Reintroduction programs in suitable habitats can contribute to the recovery of populations. Successful establishment of the

species once reintroduced has already been reported in France (Hautes Alpes) in 1922 and Madrid (Viejo, 1992).

MANAGEMENT IMPLICATIONS

Our study proved a direct relation among *G. isabelae* and Scots pine and Austrian pine woods; therefore we emphasize the importance of preserving such forests. Conservation measures must be focused mainly on the preservation of Scots pine and Austrian pine woods. We recommend managing woodlands properly and periodic monitoring of the status of each population. Sites of Community Importance should be wider in habitat patches found in Catalonia and Teruel regions. Similarly, suitable habitats around patches should be preserved as they could connect different *G. isabelae* populations. Because *G. isabelae* is a sedentary and no expansive species, we recommend reintroduction programs in suitable habitats.

ACKNOWLEDGEMENTS

Special thanks to E. García-Barros and J. Hortal for their valuable suggestions. The comments from Y. Wiersma and an anonymous referee improved a former version of the manuscript. This paper was supported by a Fundación Banco Bilbao Vizcaya Argentaria Project (Diseño de una red de reservas para la protección de la Biodiversidad en América del sur austral utilizando modelos predictivos de distribución con taxones hiperdiversos) and a Ministerio de Educación y Ciencia Project (CGL20040-04309).

REFERENCES

Agenjo, R. 1943. Ensayo sobre *Graellsia isabelae* (Graells). El lepidóptero más bello de Europa. *EOS* **XIX**, 311-411.

Anderson, R. P., Lew, D. & Peterson, A. T. 2003. Evaluating predictive models of species´distributions: criteria for selecting optimal models. *Ecological Modelling* **162**, 211-232.

Araújo, M. B. & Pearson, R. G. 2005. Equilibrium of species' distributions with climate. *Ecography* **28**(5), 693-695.

Austin, M. P. 1980. Searching for a model for use in vegetation analysis. *Vegetatio* **42**(1-3), 11-21.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecological Modelling* **157**(2-3), 101-118.

Baixeras, J. 2001. Seguimiento de poblaciones de *Graellsia isabelae* en zonas de actuación del proyecto Life-habitats. Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, Valencia.

Benito Garzón, M., Blazek, R., Neteler, M., Sánchez de Dios, R., Sainz Ollero, H. & Furlanello, C. 2006. Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula. *Ecological Modelling* **197**, 383-393.

Bennett, K. D., Tzedakis, P. C. & Willis, K. J. 1991. Quaternary refugia of north European trees. *Journal of Biogeography* **18**, 103-115.

Birks, H. J. B. 1996. Statistical approaches to interpret diversity patterns in the Norwegian mountain flora. *Ecography* **19**(3), 332-340.

Blanco, E., Casado, M. A., Costa, M., Escribano, R., García, M., Génova, M., Gómez, A., Gómez, F., Moreno, J. C., Morla, C., Regato, P. & Sainz, H. 1998. *Los Bosques Ibéricos. Una interpretación geobotánica*. Planeta, Barcelona.

Burnham, K. P. & Anderson, D. R. 2002. *Model selection and inference: a practical information-theoretical approach*. Springer-Verlag, New York, USA.

Cheddadi, R., Vendramin, G. G., Litt, T., François, L., Kageyama, M., Lorentz, S., Laurent, J.-M., Beaulieu, J.-L., Sadori, L., Jost, A. & Lunt, D. 2006. Imprints of glacial refugia in the modern genetic diversity of *Pinus sylvestris*. *Global Ecology and Biogeography* **15**(3), 271-282.

Chefaoui, R. M., Hortal, J. & Lobo, J. M. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation* **122**(2), 327-338.

Chefaoui, R. M. & Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling* **210**(4), 478-486.

Clark, L. 2000. Global Change Data Archive Vol. 3. 1 km Global Elevation Model. Clark University.

Clark, L. 2003. Idrisi Kilimanjaro, Worcester, MA.

Collingham, Y. C., Wadsworth, R. A., Huntley, B. & Hulme, P. E. 2000. Predicting the spatial distribution of alien riparian species: issues of spatial scale and extent. *Journal of Applied Ecology* **37**, 13-27.

Dennis, R. L. H. & Hardy, P. B. 1999. Targeting squares for survey: predicting species richness and incidence of species for a butterfly atlas. *Global Ecology and Biogeography* **8**(6), 443-454.

Engler, R., Guisan, A. & Rechsteiner, L. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* **41**(2), 263-274.

Fernández-Vidal, E. H. 1992. Comentarios acerca de la distribución geográfica francesa y notas taxonómicas sobre *Graellsia isabelae* (Graells, 1849). *SHILAP Revista de Lepidopterología* **20**(77), 29-49.

Fielding, A. H. 2002. What are the appropiate characteristics of an accuracy measure? In: *Predicting Species Occurrences: Issues of Accuracy and Scale* (edited by J. M. Scott, P. J. Heglund, J. B. Haufler, M. L. Morrison, M. G. Raphael, W. A. Wall & F. B. Samson). Island Press, Washington, DC, 271-280.

Figueiral, I. & Carcaillet, C. 2005. A review of Late Pleistocene and Holocene biogeography of highland Mediterranean pines (Pinus type sylvestris) in Portugal, based on wood charcoal. *Quaternary Science Reviews* **24**, 2466-2476.

Galante, E. & Verdú, J. R. 2000. *Los Artrópodos de la "Directiva Hábitat" en España*. Ministerio de Medio Ambiente, Madrid.

García-Barros, E. & Herranz, J. 2001. Nuevas localidades de *Proserpinus proserpina* (Pallas, 1772) y *Graellsia isabelae* (Graells, 1849) del centro peninsular. *SHILAP Revista de Lepidopterología* **29**(114), 183-184.

Gómez-Bustillo, M. R. & Fernández-Rubio, F. 1974. Consideraciones sobre la planta nutricia de *Graellsia isabelae* (Graells, 1849) y descripción de una nueva subsp. española (Lep. Syssphingidae). *SHILAP Revista de Lepidopterología* **2**(7), 183-189.

Graells, M. P. 1849. Description d´un Lépidoptère nouveau de la tribu des Saturnides, appartenant à la Faune entomologique espagnole. *Revue et Magasin de Zoologie Pure et Appliquée* (1), 601-602.

Guisan, A. & Theurillat, J. P. 2000. Equilibrium modeling of alpine plant distribution: how far can we go? *Phytocoenologia* **30**, 353-384.

Guisan, A. & Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters* **8**, 993-1009.

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

Gutiérrez, D. 1997. Importance of historical factors on species richness and composition of butterfly assemblages (Lepidoptera: Rhopalocera) in a northern Iberian mountain range. *Journal of Biogeography* **24**, 77-88.

Hastie, T., Tibshirani, T. & Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hewitt, G. M. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405**(6789), 907-913.

Hirzel, A. H. & Arlettaz, R. 2003. Modelling habitat suitability for complex species distributions by the environmental-distance geometric mean. *Environmental Management* **32**(5), 614-623.

Instituto Geográfico Nacional. *Atlas Nacional de España. Volume 1 and 2*, Madrid, Spain.

Jiménez-Valverde, A. & Lobo, J. M. 2006. The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions* **12**(5), 521-524.

King, G. & Zeng, L. 2000. Logistic regression in rare events data. The Global Burden of Disease 2000 in Aging Populations, Research Paper No. 2 (edited by Havard Burden of Disease Unit, C. f. P. a. D. S.). http://www.hsph.harvard.edu/burdenofdisease/publications/papers/Logistic%20Regression.pdf

Krauss, J., Steffan-Dewenter, I. & Tscharntke, T. 2003. How does landscape context contribute to effects of habitat fragmentation on diversity and population density of butterflies? *Journal of Biogeography* **30**(6), 889-900.

Legendre, P. & Legendre, L. 1998. *Numerical Ecology*. Elsevier, Amsterdam.

Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. 2005. Selecting thresholds of ocurrence in the prediction of species distributions. *Ecography* **28**(3), 385-393.

Lobo, J. M., Verdú, J. R. & Numa, C. 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions* **12**(2), 179-188.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G. & Williams, P. H. 2003. Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology* **17**(6), 1591-1600.

López-Sebastián, E., López, J. C., Juan, M. J. & Selfa, J. 2002. Primeras citas de mariposa isabelina en la Comunidad Valenciana. *Quercus*(193), 10-13.

MacNally, R. 2000. Regression and model-building in conservation biology, biogeography and ecology: The distinction between -and reconciliation of- "predictive" and "explanatory" models. *Biodiversity and Conservation* **9**(5), 655-671.

MacNally, R. 2002. Multiple regression and inference in ecology and conservation biology: further comments on retention of independent variables. *Biodiversity and Conservation* **11**, 1397-1401.

Masó i Planas, A. & Ylla i Ullastre, J. 1989. Consideraciones sobre la ecología, comportamiento, alimentación y biogeografía de *Graellsia isabelae* (Graells). *SHILAP Revista de Lepidopterología* **17**(65), 49-60.

McCullagh, P. & Nelder, J. A. 1989. *Generalized Linear Models*. Chapman & Hall, London.

Ministerio de Medio Ambiente. 2005. Perfil ambiental de España 2005. http://www.mma.es/portal/secciones/calidad_contaminacion/indicadores_ambientales/perfil_ambiental_2005.

Montoya, R. & Hernández, R. 1975. Graellsia isabelae (Graells). *Vida Silvestre*(12), 207-211.

Oberthür, C. 1923. Découverte en France, dans les Hautes Alpes, par le Docteur Cleu, de la *Saturnia (Graellsia) Isabellae* Graells. *Etudes de Lépidoptérologie comparée*(20), 161-173.

Olden, J. D., Jackson, D. A. & Peres-Neto, P. 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society* **131**, 329-336.

Ott, L., Larson, R. F. & Mendenhall, W. 1983. *Statistics: A Tool for the Social Sciences*. Duxbury Press, Boston.

Pearson, R. G. & Dawson, T. E. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography* **12**, 361-372.

Pearson, R. G., Dawson, T. P., Berry, P. M. & Harrison, P. A. 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. *Ecological Modeling* **154**, 289-300.

Peñalba, M. C. 1994. The history of Holocene vegetation in northern Spain from pollen analysis. *Journal of Ecology* **82**, 815-832.

Reutter, B., Helfer, V., Hirzel, A. H. & Vogel, P. 2003. Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography* **30**(4), 581-590.

Robredo, F. 1988. *Estudios sobre los tratamientos forestales con diflubenzurón y su incidencia sobre la fauna*. Ministerio de Agricultura, Madrid.

Ruíz de la Torre, J. 2002. Mapa forestal de España. Ministerio de Medio Ambiente. Organismo Autónomo Parque Nacionales, Madrid.

Schmitt, T. & Krauss, J. 2004. Reconstruction of the colonization route from glacial refugium to the northern distribution range of the European butterfly Polyommatus coridon (Lepidoptera: Lycaenidae). *Diversity and Distributions* **10**(4), 271-274.

Schröder, B. 2004. ROC Plotting and AUC Calculation Transferability Test. Institute for Geoecology, Potsdam University. Potsdam. http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html.

Sinclair, W. T., Morman, J. D. & Ennos, R. A. 1999. The postglacial history of Scots pine (Pynus sylvestris L) in western Europe: evidence from mitochondrial DNA variation. *Molecular ecology* **8**(1), 83-88.

Skov, F. & Svenning, J. C. 2004. Potential impact of climate change on the distribution of forest herbs in Europe. *Ecography* **27**(3), 366-380.

Sokal, R. R. & Rohlf, F. J. 1981. Biometry. Freeman, New York.

Soranzo, N., Alia, R., Provan, J. & Powell, W. 2000. Patterns of variation at a mitochondrial sequencetagged-site locus provides new insights into the postglacial history of European Pinus sylvestris populations. *Molecular ecology* **9**(9), 1205-1211.

Soria, S., Abos, F. & Martín, E. 1986. Influencia de los tratamientos con diflubenzurón ODC 45% sobre pinares en las poblaciones de *Graellsia isabelae* (Graells) (Lep. Syssphingidae) y reseña de su biología. *Boletín de sanidad vegetal*(12), 29-50.

StatSoft, I. 2001. STATISTICA (data analysis software system), version 6.

Steffan-Dewenter, I. & Tscharntke, T. 2000. Butterfly community structure in fragmented habitats. *Ecology Letters* **3**(5), 449-456.

Svenning, J. C. & Skov, F. 2004. Limited filling of the potential range in European tree species. *Ecology Letters* **7**(7), 565-573.

Thomas, C. D. 2000. Dispersal and extinction in fragmented landscapes. *Proceedings of the Royal Society of London* **267**(1439), 139-145.

Verdú, J. R. & Galante, E. 2002. Climatic stress, food availability and human activity as determinants of endemism patterns in the Mediterranean region: the case of dung beetles (Coleoptera, Scarabeoidea) in the Iberian Peninsula. *Diversity and Distributions* **8**, 259-274.

Viejo, J. L. 1992. Biografía de un naturalista y biología del lepidóptero por él descrito. Graells y la *Graellsia*. *Quercus*(74), 22-30.

Vives, A. 1994. Catálogo sistemático y sinonímico de los Lepidópteros de la Península Ibérica y Baleares (Insecta; Lepidoptera). Ministerio de Agricultura, Pesca y Alimentación, Madrid.

Vuattoux, R. 1984. *Pinus pinea*: une nouvelle plante nourricière pour *Graellsia isabelae*. *Bulletin Société Sciences Naturelles* **43**, 11.

Wahlberg, N., Klemetti, T. & Hanski, I. 2002. Dynamic populations in a dynamic landscape: the metapopulation structure of the marsh fritillary butterfly. *Ecography* **25**(2), 224-232.

Warren, M. S., Hill, J. K., Thomas, J. A., Asher, J., Fox, R., Huntley, B., Roy, D. B., Telfer, M. G., Jeffcoate, S., Harding, P., Jeffcoate, G., Willis, S. G., Greatorex-Davies, J. N., Moss, D. & Thomas, C. D. 2001. Rapid responses of British butterflies to opposing forces of climate and habitat change. *Nature* **414**(1), 65-69.

Wenzel, M., Schmitt, T., Weitzel, M. & Seitz, A. 2006. The severe decline of butterflies on western German calcareous grasslands during the last 30 years: A conservation problem. *Biological Conservation* **128**(4), 542-552.

White, P. S., Wilds, S. P. & Stratton, D. A. 2001. The distribution of heath balds in the Great Smoky Mountains, North Carolina and Tennessee. *Journal of Vegetation Science* **12**(4), 453-466.

WWF/Adena. 2006. Grandes Incendios Forestales. http://assets.wwfes.panda.org/downloads/informe_incendios_06.pdf.

Ylla i Ullastre, J. 1997. *Historia natural del lepidòpter Graellsia isabelae (Graells, 1849)*. Institut d´Estudis Catalans, Barcelona.

Zaniewski, A. E., Lehmann, A. & Overton, J. M. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* **157**(2-3), 261-280.

Zweig, M. H. & Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**(4), 561-577.

Efectos de las características ecológicas y de los datos en el comportamiento de los modelos de distribución de especies de invertebrados protegidos

RESUMEN

La poca calidad de los datos disponibles acerca de la distribución de especies amenazadas dificulta el diseño de estrategias para su conservación. En este capítulo, evaluamos los efectos de las características de los datos y de las especies en la precisión de los modelos de distribución de invertebrados protegidos, elaborados a partir de datos de atlas y colecciones de historia natural. Se obtienen los modelos de 20 especies de invertebrados amenazados de la Península Ibérica y Baleares que tienen diferentes características con GAM, GLM y NNET, usando datos de museos y atlas complementados con pseudo-ausencias. A continuación, se explora la relación entre la precisión de los modelos y las características particulares de cada especie.

Las dos características asociadas a los datos examinadas, el tamaño de muestra (N) y el área de ocurrencia relativa (ROA), afectaron de forma significativa a la precisión de los modelos. La marginalidad, el tipo de hábitat, el nivel trófico y la detección del hábitat también presentan correlación con la precisión de los modelos de distribución; mientras que la extensión de la ocurrencia de la especie y la capacidad de vuelo no parecen influir en ninguna medida de precisión de los modelos.

Concluimos que realizar modelos a partir de datos de museos y atlas usando pseudo-ausencias parece un buen método para predecir la distribución de especies

amenazadas, siempre que tengamos suficiente cantidad de datos de presencia. En general, las especies con las que se consiguen modelos de distribución más precisos son aquellas con mayor tamaño de muestra o menor ROA. Las especies asociadas a hábitats difícilmente reconocibles en capas digitales, tales como son los hábitats riparios y húmedos, parecen obtener peores predicciones. Además, las características asociadas a los datos parecen interaccionar con otras características ecológicas, como la predictibilidad del hábitat y el grupo trófico afectando al comportamiento de los modelos.

Palabras clave: modelos de distribución, Península Ibérica, precisión del modelo, características de la especie, invertebrados protegidos.

Este capítulo ha sido enviado a publicar como:

Chefaoui, R. M., Lobo J. M. & Hortal, J. "Assessing the effects of data characteristics and ecological traits on species distribution models of threatened invertebrates." *Diversity and Distributions*.

Assessing the effects of data characteristics and ecological traits on species distribution models of threatened invertebrates.

ABSTRACT

The poor quality of distributional data on threatened species hampers designing conservation strategies. Here we evaluate the effects of several data characteristics and species' traits on the accuracy of species distribution models (SDM) for red-listed invertebrates, generated from museum and atlas data. We apply three SDM techniques (GAM, GLM and NNET) using pseudoabsences to model the distribution of 20 threatened Iberian invertebrates. We correlate the accuracy of the obtained models with several data characteristics and species' ecological traits.

Both data characteristics examined, the amount of data (N) and the relative occurrence area (ROA), significantly affected the accuracy of the models. Marginality, habitat type, trophic level and habitat detectability were also correlated with model accuracy, whereas the total extent of the distribution range and flight capacity were not.

SDM based on museum or herbarium data complemented with pseudoabsences may be used to characterize the distribution of threatened species, provided that enough distributional data is available. In general, the species whose distributions are modelled more accurately are those with greater sample size or smaller ROA. Species related to habitats that are problematic to detect using GIS data, such as riparian or humid areas, seem to obtain worse predictions. Also, data characteristics interact with several

ecological traits, such as habitat detectability and trophic level, to affect model performance.

Keywords: Data quality, ecological characteristics, Iberian Peninsula, predictive accuracy, sample size, species distribution modelling.

INTRODUCTION

Including rare and threatened species in conservation planning assessments is a major challenge in conservation science. Being endangered species usually rare as well, including them in area prioritization exercises is a complex task, due to the low spatial congruence among the rarest species (Grenyer *et al*., 2006). Such complexity is further increased by the difficulty of mapping the distribution of these rare and threatened species. Here, data scarcity is often ameliorated with the help of GIS-based models and analytical techniques. Species distribution modelling (SDM) is nowadays a well-established research topic (see Lobo *et al.*, 2010), and many studies use data from museum collections and the literature to model the distributions of species from many living groups (e.g., Reutter *et al*., 2003; Brotons *et al*., 2004; Elith & Leathwick, 2007). This approach is especially important for the hyperdiverse invertebrates, where the difficulty of developing extensive surveys makes biodiversity databases based on data from museums and atlases the necessary alternative to obtain presence records for mapping distributions (e.g., Chefaoui *et al.*, 2005; Lobo *et al.*, 2006, 2010; Chefaoui & Lobo, 2007). Unfortunately, many of the databases that are often used to model species distributions present important limitations due to sampling bias or deficient survey

effort (Hortal *et al*., 2007). This is likely the case for many invertebrate groups (Lobo *et al.*, 2007; Hortal *et al.*, 2008). Under these circumstances, systematic conservation planning for invertebrate taxa entails modelling species with diverse characteristics and ecological requirements generally using poor quality data, often with no time for detailed "species-by-species" expert assessments (see Cabeza *et al.*, 2010). Hence, using authomatized SDM protocols to predict the distribution of invertebrates from presence-only data is hampered by: (i) the use of heterogeneous biological data sources generally without any survey effort measure; (ii) the environmental and spatially biased character of this information; (iii) the lack of accurate absence data; (iv) the difficulty of identifying the best predictor variables for each species; and (v) the high diversity in the accuracy scores of these predictions according to the considered species.

Many studies have addressed how varying sample size, geographic ranges and other ecological characteristics of the species affect model accuracy, as an attempt to understand the limitations and possibilities of SDM techniques. An increase in model accuracy has been related to greater sample sizes (Stockwell & Peterson, 2002; McPherson *et al*., 2004), as well as to species with more specialized requirements (Brotons *et al*., 2004; Seoane *et al*., 2005), less mobility (Pöyry *et al*., 2008) and smaller geographic ranges (Stockwell & Peterson, 2002; Segurado & Araújo, 2004; Hernández *et al*., 2006). Nevertheless, the relationships found among model performance and species characteristics are strongly dependent on the modelling technique, but also on the characteristics of the used data itself, namely sample size and the proportion of the occupied area over the total area of the considered territory (the relative occurrence area or ROA; Lobo, 2008; Lobo *et al*., 2008). Thus, a better understanding on how species traits and data characteristics influence the results of different modelling methods could

help refining the use of SDM, particularly when modelling species with heterogeneous characteristics (Segurado & Araújo, 2004; Marmion *et al*., 2008).

The general aim of this study is to determine how species and data traits influence the ability of SDM to model invertebrate species. More precisely, we examine the relationship between several measures of model accuracy (AUC, sensitivity and specificity) and (i) two characteristics of the data used, namely sample size and ROA, and (ii) several species traits, including niche specialization (marginality), the total extent of the distribution range (herein TER), dispersal ability, trophic group, habitat type and habitat detectability. To do this, we apply three SDM procedures (Generalized Linear Models, GLMs; Generalized Additive Models, GAMs; and Neural Network Models, NNETs) to model the distribution in the Iberian Peninsula of 20 threatened invertebrate species (mostly arthropods) with dissimilar ecological characteristics, using presence data from museum collections and atlases, and pseudo-absences (Zaniewski *et al*., 2002; Chefaoui & Lobo, 2008).

METHODS

Study area

The study area was the Ibero-balearic region (western Mediterranean) which comprises 587,663 km^2, divided into 6,150 cells of 10 km x 10 km, that constitute the units of analysis. All biological and environmental data was referred to these cells.

Biological Data

We selected 20 threatened invertebrates, mostly arthropods, with dissimilar biological and ecological traits from the Red Book of the Spanish Invertebrates (Verdú & Galante,

2006) and the Spanish Inventory of species included in the "Habitats Directive" (Galante & Verdú, 2000). Occurrence data was obtained from these two references, and completed with other diverse bibliographic sources (Soria *et al*., 1986; Castillejo, 1990; Rosas *et al*., 1992; Viejo, 1992; Grosso-Silva, 1999; Grupo de Trabajo sobre Lucanidae Ibéricos, 2000; García-Barros & Herranz, 2001; Pérez-Bote *et al*., 2001; Raimundo *et al*., 2001; López-Sebastián *et al*., 2002; Martínez-Orti, 2004).

Because accurate absence data were not available, we previously identified those pseudo-absences located outside the climatic domain defined by the available presences (see Lobo *et al*., 2010). For that, a multidimensional envelope was carried out for each species (Busby, 1991; Lobo *et al*., 2006) using the climatic variables mentioned below. Subsequently, ten times more pseudo-absences than presences (prevalence = 0.1) were randomly selected from the area outside each envelope to avoid biases caused by the inclusion of an extremely high number of absences (King & Zeng, 2000; Dixon *et al*., 2005; Jiménez-Valverde & Lobo, 2006; Jiménez-Valverde *et al.*, 2009). This way of extracting pseudo-absences does not allow accounting for environmentally suitable localities from where the species is absent, either because it has not been able to colonize them or because it became recently extinct. Hence, the so-obtained geographical predictions will tend to approximate the potential distributions of the studied species (Chefaoui & Lobo, 2008; Jiménez-Valverde *et al*., 2008; Lobo *et al.*, 2010).

Predictor variables

Due to the heterogeneity in the ecological roles, life histories and adaptations of the invertebrates studied, we selected the best predictor variables for each species from a

range of topographic, climatic, lithological and spatial variables (Table 4.1). We extracted topographic variables from a global digital elevation model with 1-km spatial resolution (Clark Labs, 2000); elevation range was calculated as the difference between maximum and minimum elevation in each cell. Temperature and precipitation data at 1-km resolution were provided by the Spanish Instituto Nacional de Meteorología (http://www.aemet.es/). We calculated aridity as $Ia = 1/ (P/T + 10) \times 10^2$, where P is the mean annual precipitation and T the mean annual temperature (see Verdú & Galante, 2002). We digitized four lithology variables from a lithology map (Instituto Geográfico Nacional, 1995), calculating the area of calcareous deposits, siliceous sediments, stony acidic soils and calcareous soils on each cell. Finally, we extracted two spatial variables per cell: latitude (Lat) and longitude (Lon) of the centroid of each cell; in addition, we derived the nine terms of their third order polynomial expression (i.e., Trend Surface Analysis). The inclusion of these spatial variables after the environmental predictors can help to consider the effect of unaccounted-for predictors and/or other factors that have been able to generate spatial patterns in species distributions (see Legendre & Legendre, 1998).

All predictor variables referred to the same 10 x10 km UTM grid cells were extracted and handled using IDRISI Kilimanjaro GIS software (Clark Labs, 2003). All these variables (including latitude and longitude) were standardized to zero mean and one standard deviation to eliminate the effect of varying measurement scales.

To select the best explanatory variables for each species, we used an individual logistic regression of presence-absence of species against each one of the explanatory variables by means of the Statistica software (Statsoft, 2001). We evaluated the linear, quadratic and cubic functions of each variable, in order to account for possible

curvilinear relationships (Austin, 1980). In addition, we chose the most appropriate spatial variables for each species after a backward-stepwise elimination of non-significant terms from the third-degree polynomial of latitude and longitude.

Predictor variables	**Minimum - Maximum values**
Topographic variables	
Maximum elevation (meters)	1 - 3399
Mean elevation (meters)	1 - 2721
Minimum elevation (meters)	0 - 2521
Elevation range (meters)	0 - 2291
Climatic variables	
Winter precipitation (Jan., Feb., March) (mm)	491 - 9579
Spring precipitation (April, May, June) (mm)	463 - 6236
Summer precipitation (July, August, Sept.) (mm)	66 - 4724
Autumn precipitation (Oct., Nov., Dec.) (mm)	607 - 6140
Temperature range (ºC)	11 - 32
Maximum Winter Temperature (ºC)	1 - 18
Mean Winter Temperature (ºC)	-4 - 13
Minimum Winter Temperature (ºC)	-8 - 10
Maximum Spring Temperature (ºC)	6 - 23
Mean Spring Temperature (ºC)	0 - 17
Minimum Spring Temperature (ºC)	-5 - 12
Maximum Summer Temperature (ºC)	19 - 35
Mean Summer Temperature (ºC)	10 - 26
Minimum Summer Temperature (ºC)	2 - 20
Maximum Autumn Temperature (ºC)	9 - 25
Mean Autumn Temperature (ºC)	2 - 21
Minimum Autumn Temperature (ºC)	-3 - 15
Aridity	0 - 1.64
Lithological variables	
% Acid soil	0 - 100
% Calcareous soil	0 - 100
% Acid sediments	0 - 100
% Calcareous sediments	0 - 100
Spatial variables	
Latitude (Y)	3990000 - 4860000
Longitude (X)	-20000 - 1060058

Table 4.1 - Predictor variables used to generate the species distribution models. The appropriate variables for each species were previously selected by individual logistic regression analyses.

Species Distribution Models

Data on presences, pseudo-absences and the selected predictor variables for each species were used to generate predictive functions by means of three different and widely used

species distribution modelling techniques: Generalized Linear Models (GLMs), Generalized Additive Models (GAMs) and Neural Network Models (NNETs). GLMs (McCullagh & Nelder, 1989) were elaborated assuming a logistic relationship between the dependent and the explanatory variables (i.e., link function), and a binomial error distribution of the dependent variable; GAMs using penalized regression splines (Wood & Augustin, 2002); and NNET fitting a single-hidden-layer neural network, with skip-layer connections (Ripley, 1996). All SDM were fitted in R (R Development Core Team, 2008).

Measures of model performance

Given that for some species the sample size was not large enough to split it into representative training and evaluation datasets, we implemented a "leave-one-out" jack-knife procedure (Olden *et al*., 2002). Here, each observation is excluded and the model parameterized using the remaining $n - 1$ observations to obtain a predicted probability score for the excluded observation; this procedure yields relatively unbiased estimates of model performance (Olden *et al*., 2002). After repeating this procedure n times (one per observation), we used these new jack-knife probabilities to calculate three measures of model performance: (i) the area under the ROC function (AUC; Zweig & Campbell, 1993; Schröder, 2004), (ii) sensitivity (proportion of correctly predicted presences) and (iii) specificity (proportion of correctly predicted absences). Sensitivity and specificity were calculated fixing the threshold probability according to the prevalence of the data (0.1; see Jiménez-Valverde & Lobo, 2006). Thus, these performance measures indicate the discrimination power of the models when they are validated with data not used in

the training process (Fielding & Bell, 1997). All measures range from 0 (poor quality model) to 1 (excellent prediction).

Data characteristics

We evaluated the influence on model performance of two characteristics of the input data: sample size (N) and the relative occurrence area. ROA is the ratio between the area of the distribution range of the species within the studied region, and the total area of such region (Lobo, 2008; Jiménez-Valverde *et al*., 2008). Here, the area of the study region is thus the whole area of the Ibero-balearic region (see above), and the distribution range of the species within such region was estimated as the minimum convex polygon (i.e. the smallest polygon in which no internal angle exceeds 180 degrees) that contains all presence sites (also called convex-hull; Burgman & Fox, 2003). ROA measures whether the allocation of presence points in the study area shows a relatively wide distribution (as ROA value tends to 1) or a more restricted pattern.

Species' traits

We examined six species characteristics for their correlation with model accuracy: niche marginality, the total extent of the distribution range (TER), habitat type, habitat detectability, trophic group, and dispersal ability. Raw data on these characteristics were collected from published information on their life histories and biogeography, and were then classified into categories. The degree of specialization of each species was estimated from their marginality scores obtained with ENFA (Hirzel *et al*., 2007). ENFA measures the average position of the species' niche according to the observed localities of presence, in relation to the average environmental conditions in the study area; high marginality values indicate a tendency to inhabit extreme conditions

regarding the overall conditions in the considered region. TER is a qualitative variable with three categories that represent the total extent and the general distribution of the species: Iberian and Ibero-Maghrebian species (Cat. I), European species (Cat. II), and Euroasiatic species (Cat. III). The type of habitat generally inhabited by the species was also classified in three categories: Type I (woodlands and mountainous habitats), Type II (open habitats such as grasslands, rocky slopes, etc) and Type III (humid and riparian conditions). Habitat detectability refers to the easiness of detecting the suitable habitat patches for each species using GIS-based data. Each species was classified according to its membership to habitats of either low- or high-detectability. Low-detectability habitats are those habitats that are usually smaller than the resolution employed in GIS data on land cover, including microhabitats such as very specific host plants, under stones or river banks; conversely, high-detectability habitats are those easily identifiable using GIS data, such as extensive woodlands, grasslands or mountainous areas. Besides, species were classified into two trophic groups according to their trophic range, phytophagous (P) or non-phytophagous (NP) species. Finally, the dispersal ability of the species was measured as a binary variable accounting for whether they are able to fly or not (Table 4.2).

Evaluation of the influence on model performance

We examined individually whether any of the data characteristics or species' traits were correlated with the measures of model performance. The influence of continuous traits (N, ROA and marginality) was assessed through Spearman rank correlations (*Rs*) with each one of the accuracy measures (AUC, sensitivity and specificity). Here, Partial Correlation Analysis was also used in order to estimate the single contribution of N and

	Data characteristics		Species traits					
Species	**N**	**ROA**	**Marginality**	**TER**	**Habitat Type**	**Habitat detectability**	**Trophic level**	**Flight capacity**
Cerambyx cerdo	152	0.796	0.768	Euroasiatic (C.III)	Woods (T.I)	High	Polyphagous (P)	Yes
Coenagrion mercuriale	87	0.629	0.455	Ibero-Maghrebian (C.I)	Riparian (T.III)	Low	Carnivorous (NP)	Yes
Cupido lorquinii	87	0.267	1.201	Ibero-Maghrebian (C.I)	Grassland (T.II)	High	Omnivorous (NP)	Yes
Elona quimperiana	41	0.141	2.869	Ibero-Maghrebian (C.I)	Mixed (T.II)	Low	Omnivorous (NP)	No
Eriogaster catax	12	0.067	2.538	Euroasiatic (C.III)	Woods (T.I)	High	Polyphagous (P)	Yes
Euphydryas aurinia	749	0.851	1.154	Euroasiatic (C.III)	Woods (T.I)	High	Oligophagous (P)	Yes
Geomalacus maculosus	37	0.114	2.397	European (C.II)	Humid (T.III)	Low	Polyphagous (P)	No
Graellsia isabelae	138	0.212	2.240	Iberian (C.I)	Woods (T.I)	High	Oligophagous (P)	Yes
Lucanus cervus	456	0.625	1.915	Euroasiatic (C.III)	Woods (T.I)	High	Polyphagous (P)	Yes
Macromia splendens	10	0.436	1.797	Ibero-Maghrebian (C.I)	Riparian (T.III)	Low	Carnivorous (NP)	Yes
Macrothele calpeiana	92	0.076	1.624	Iberian (C.I)	Mixed (T.II)	Low	Carnivorous (NP)	No
Maculinea alcon	49	0.212	2.528	European (C.II)	Grassland (T.II)	High	Omnivorous (NP)	Yes
Maculinea arion	166	0.310	3.397	Euroasiatic (C.III)	Grassland (T.II)	High	Omnivorous (NP)	Yes
Maculinea nausithous	17	0.041	4.584	European (C.II)	Grassland (T.II)	High	Omnivorous (NP)	Yes
Oxygastra curtisi	21	0.612	1.971	European (C.II)	Riparian (T.III)	Low	Omnivorous (NP)	Yes
Parnassius apollo	314	0.459	3.600	Euroasiatic (C.III)	Mountain (T.I)	Low	Polyphagous (P)	Yes
Parnassius mnemosyne	42	0.017	5.897	Euroasiatic (C.III)	Mountain (T.I)	High	Oligophagous (P)	Yes
Rosalia alpina	47	0.132	3.656	Euroasiatic (C.III)	Woods (T.I)	High	Polyphagous (P)	Yes
Vertigo moulinsiana	20	0.064	1.261	European (C.II)	Humid (T.III)	Low	Carnivorous (NP)	No
Zerynthia rumina	1107	0.927	0.376	Ibero-Maghrebian (C.I)	Rocky slopes (T.II)	Low	Oligophagous (P)	Yes

Table 4.2 - Data characteristics and species traits that may influence model accuracy. N: Sample size; ROA: relative occurrence area; TER: total extent of the distribution range, in three categories from more restricted to wider distribution: C.I = Iberian and Ibero-Maghrebian, C.II = European and C.III = Euroasiatic distribution; Habitat types: T.I = woods and mountainous habitats, T.II = grasslands and varied habitats, T. III = riparian and humid habitats; trophic level categories: P = phytophagous, NP = non phytophagous.

ROA on the variation of accuracy measures. The degree of association between model accuracy measures and the qualitative traits considered (TER, habitat type, habitat detectability, trophic group and dispersal ability) was established by using non-parametric statistical tests such as Kruskal-Wallis or Mann-Whitney U.

In addition, to eliminate the eventual influence of data characteristics, we regressed each accuracy measure against N and ROA and extracted the residuals of these relationships. The so-obtained residual values of the accuracy measures were later submitted to new correlation, Kruskal-Wallis or Mann-Whitney U tests to evaluate their relationships with the species' traits, applying both a standard significance level (p<0.05) and a Bonferroni correction for multiple comparisons (p = 0.05/9 = 0.006).

RESULTS

On average, the three modelling techniques achieved quite high accuracy in their predictions (mean AUC ± SD = 0.951 ± 0.013; mean specificity = 0.917 ± 0.017; mean sensitivity = 0.903 ± 0.017) (Table 4.3). Neither AUC nor specificity or sensitivity values differ significantly among the three SDM techniques (Kruskal-Wallis test; n = 60; AUC: H = 3.98, p = 0.14; specificity: H = 3.26, p = 0.20; sensitivity: H = 4.20, p = 0.10). The area estimated for the potential distribution of the studied species did not differ significantly among the three modeling techniques either (Kruskal-Wallis test; n = 60; H = 0.31, p = 0.86).

Among the data characteristics and species' traits examined, N, ROA and, to a lesser extent, marginality significantly (p<0.006) affected the accuracy of distribution models (Table 4.4). Several species traits (habitat type, trophic group and habitat detectability) were also associated with model accuracy measures (p<0.05), although

their influence was much lower than data characteristics and dropped below significance when a Bonferroni correction is applied. In contrast, TER and flight capacity did not seem to influence any measure of model accuracy.

Species with greater N obtained higher model accuracies; AUC values above 0.98 and sensitivity scores higher than 0.94 were achieved when models were elaborated with sample sizes larger than 200 records (see Fig. 4.1 and Appendix I). Partial correlation analyses of both data traits (N and ROA) on accuracy measures showed that while sample size was always positively and significantly correlated with model accuracy, ROA was negatively correlated in most part of the occasions (seven out of nine; see Table 4.4).

The species traits showed less influence on model performance. Marginality values are significantly correlated with accuracy, but only when the effect of N and ROA is not considered. The only trait that may remain relevant for accuracy measures after accounting for data characteristics is habitat type; species associated to humid and riparian conditions seem to be predicted worse (see Fig. 4.2, Appendix II and Fig. 4.3). However, this association is not statistically significant when a Bonferroni correction is applied. Other associations such as the trophic range of species and GLM accuracy or habitat detectability and GAM performance, also cease to be significant under the more restrictive Bonferroni significance levels.

	AUC			Specificity			Sensitivity			Area (in grid cells)		
Species	**GAM**	**GLM**	**NNET**	**GAM**	**GLM**	**NNET**	**GAM**	**GLM**	**NNET**	**GAM**	**GLM**	**NNET**
Cerambyx cerdo	0.9402	0.9557	0.8353	0.8620	0.8746	0.7565	0.8618	0.8750	0.7565	4328	4458	2507
Coenagrion mercuriale	0.9196	0.9414	0.8020	0.8275	0.8735	0.7298	0.8275	0.8735	0.7356	3857	4165	1943
Cupido lorquinii	0.9730	0.8936	0.9788	0.9563	0.9827	0.9310	0.9540	0.8045	0.9310	1021	859	822
Elona quimperiana	0.9866	0.9692	0.9885	0.9512	0.9439	0.9463	0.9512	0.9512	0.9512	594	551	398
Eriogaster catax	0.9430	0.9062	0.9840	0.9083	0.9583	0.9166	0.9166	0.8333	0.9166	4845	4866	5234
Euphydryas aurinia	0.9896	0.9909	0.9835	0.9524	0.9571	0.9508	0.9519	0.9572	0.9506	4127	4159	4019
Geomalacus maculosus	0.9554	0.9654	0.9385	0.8918	0.9189	0.8918	0.8918	0.9189	0.8918	784	747	676
Graellsia isabelae	0.9934	0.9927	0.9709	0.9594	0.9507	0.9275	0.9565	0.9492	0.9275	1021	962	998
Lucanus cervus	0.9926	0.9924	0.9818	0.9700	0.9649	0.9547	0.9692	0.9649	0.9539	2983	3243	3056
Macromia splendens	0.8830	0.8030	0.8570	0.7600	0.7000	0.7900	0.8000	0.7000	0.8000	1361	848	266
Macrothele calpeiana	0.9933	0.9321	0.9626	0.9565	0.9782	0.9130	0.9565	0.8804	0.9130	454	369	318
Maculinea alcon	0.9729	0.9668	0.9536	0.9346	0.9265	0.9183	0.9387	0.9183	0.9183	1182	938	1024
Maculinea arion	0.9927	0.9912	0.9785	0.9596	0.9698	0.9337	0.9578	0.9698	0.9337	1205	1141	1074
Maculinea nausithous	0.9861	0.9081	0.9892	0.9411	0.9882	0.9411	0.9411	0.8235	0.9411	214	172	285
Oxygastra curtisi	0.8749	0.8544	0.8920	0.8904	0.9190	0.8095	0.8095	0.7619	0.8095	677	493	557
Parnassius apollo	0.9932	0.9917	0.9869	0.9722	0.9746	0.9726	0.9713	0.9745	0.9713	1850	1622	1884
Parnassius mnemosyne	0.9953	0.9462	0.9977	0.9761	0.9880	0.9976	0.9761	0.9047	1.0000	230	147	125
Rosalia alpina	0.9881	0.9426	0.9838	0.9446	0.9148	0.9361	0.9361	0.9148	0.9361	545	485	341
Vertigo moulinsiana	0.9745	0.9288	0.8575	0.9350	0.8550	0.8050	0.9500	0.8500	0.8000	405	444	1326
Zerynthia rumina	0.9860	0.9888	0.9660	0.9428	0.9551	0.9265	0.9430	0.9548	0.9268	5422	5596	5455
Mean	0.9666	0.9430	0.9444	0.9245	0.9296	0.8974	0.9230	0.8890	0.8982	1855.2	1813.2	1615.4
± SD	± 0.03	± 0.04	± 0.05	± 0.05	± 0.06	± 0.07	± 0.05	± 0.07	± 0.07	±1716.3	±1823.7	±1638.4

Table 4.3 - Accuracy measures and resulting area size for each studied species and modelling technique used. All areas are measured as the number of grid cells (of 100 km^2 each). GAM: Generalized Additive Models; GLM: Generalized Linear Models; NNET: Neural Network Models. SD: standard deviation.

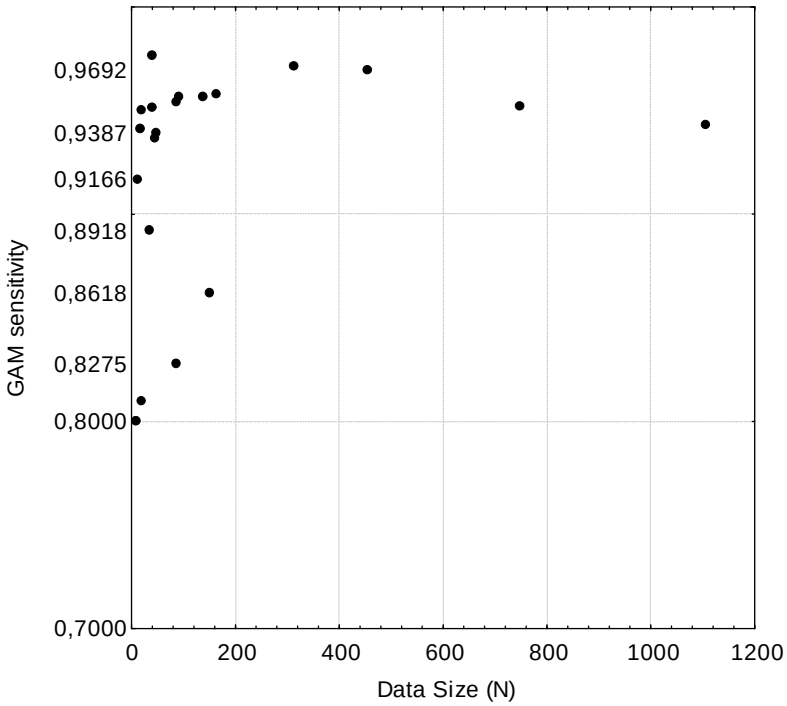
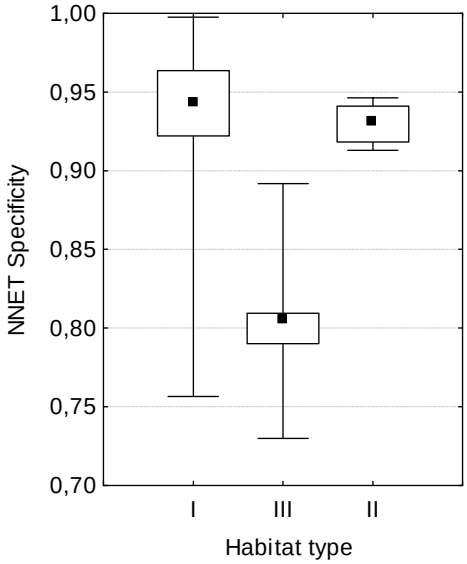


Fig. 4.1 - Correlation between sensitivity of GAM models and data size (N), an example of how the number of occurrences influences on accuracy scores. Similar results were obtained with specificity and AUC metrics (see Appendix I).

Fig. 4.2 - Specificity of NNET results by habitat type (I = Woods and Mountainous habitats, II = Grasslands and varied habitats, III = Riparian and humid habitats). Less accurate models are obtained for species associated to riparian and humid habitats. Similar results were obtained with sensitivity and AUC metrics (see Appendix II). The middle point shows the median response for each habitat type and specificity score combination. The bottom and top of the box show the 25 and 75 percentiles respectively. The whiskers show minimum and maximum values.



DISCUSSION

Several species traits and data characteristics have been formerly shown to influence SDM performance (e.g. Brotons *et al*., 2004; Segurado & Araújo, 2004; Seoane *et al*., 2005; Hernández *et al*., 2006; Marmion *et al*., 2008). In general, the prevalence in the dataset is thought to affect the accuracy of models (e.g. McPherson *et al*., 2004; Seoane *et al*., 2005; Marmion *et al*., 2008), although these effects might only appear on extreme prevalence values (see Jiménez-Valverde *et al.*, 2009). To avoid this possible statistical artefact we have removed the effect of prevalence, finding that the performance of SDM techniques varies according to the particular dataset used and, to a lesser extent, to the specific characteristics of the species.

Larger sample sizes have been previously shown to increase model accuracy (Stockwell & Peterson, 2002; McPherson *et al*., 2004; Hernández *et al*., 2006); worse results are obtained when low number of observations are used (Jiménez-Valverde *et al.*, 2009). In this work, sample size had a highly significant effect on model performance although further work would be needed to confirm the robustness of each particular SDM technique to low sample sizes. In spite of the significant relationship between sample size and ROA (Spearman Rank Order Correlation: R_S = 0.65, p = 0.0017; Fig. 4.4), the relative occurrence area (ROA) did not show any direct relationship with model performance when analyzed individually (but see below).

		Data characteristics		Species traits					
		N	ROA	Marginality	TER	Habitat Type	Trophic level	Flight capacity	Habitat detectability
GAM	AUC	R= 0.47	R= -0.25	R= 0.41	H= 3.39	H= 7.85	Z= 1.36	Z= 0.28	Z= 1.25
		p= 0.03(*)	*p*= 0.28	*p*= 0.08	*p*= 0.2	*p*= 0.02(*)	*p*= 0.2	*p*= 0.8	*p*= 0.21
		R_P= 0.75	**R_P= -0.74**	R= 0.16	H= 5.98	H= 8.52	Z= 1.21	Z= 0.66	Z= 2.16
		***p*<0.001**	***p*<0.001**	*p*= 0.48	*p*= 0.05	*p*= 0.01(*)	*p*= 0.23	*p*= 0.5	*p*= 0.03(*)
	Sensitivity	R= 0.52	R= -0.16	R= 0.30	H= 3.20	H= 7.13	Z= 0.94	Z= 0.14	Z= 1.21
		p= 0.02(*)	*p*= 0.48	*p*= 0.19	*p*= 0.20	*p*= 0.03(*)	*p*= 0.34	*p*= 0.88	*p*= 0.22
		R_P= 0.77	**R_P=-0.78**	R= 0.11	H= 4.05	H= 6.17	Z= 0.45	Z= 0.66	Z= 2.01
		***p*<0.001**	***p*<0.001**	*p*= 0.63	*p*= 0.13	*p*= 0.04(*)	*p*= 0.65	*p*= 0.5	*p*= 0.04(*)
	Specificity	R= 0.51	R= -0.15	R= 0.38	H= 4.11	H= 8.21	Z= 1.21	Z= 0.18	Z= 1.40
		p= 0.019(*)	*p*= 0.51	*p*= 0.09	*p*= 0.13	*p*= 0.02(*)	*p*= 0.22	*p*= 0.85	*p*= 0.16
		R_P= 0.66	**R_P= -0.66**	R= 0.25	H= 3.16	H= 3.76	Z= 0.38	Z= 1.23	Z= 1.56
		***p*= 0.002**	***p*= 0.002**	*p*= 0.29	*p*= 0.2	*p*= 0.15	*p*= 0.7	*p*= 0.22	*p*= 0.12
GLM	AUC	**R= 0.75**	R= 0.33	R= 0.08	H= 2.75	H= 5.04	Z= 2.19	Z= 0.28	Z= 0.87
		***p*<0.001**	*p*= 0.15	*p*= 0.72	*p*= 0.25	*p*= 0.08	*p*= 0.02(*)	*p*= 0.77	*p*= 0.38
		R_P= 0.54	R_P= -0.31	R= 0.19	H= 2.63	H= 2.71	Z= 0.98	Z= 0.00	Z= 1.18
		p= 0.016(*)	*p*= 0.2	*p*= 0.42	*p*= 0.27	*p*= 0.25	*p*= 0.32	*p*= 1.00	*p*= 0.24
	Sensitivity	**R= 0.74**	R= 0.29	R= 0.15	H= 3.72	H= 4.52	Z= 2.04	Z= 0.09	Z= 0.42
		***p*<0.001**	*p*= 0.21	*p*= 0.52	*p*= 0.15	*p*= 0.10	*p*= 0.04(*)	*p*= 0.92	*p*= 0.68
		R_P= 0.57	R_P= -0.37	R= 0.27	H= 2.93	H= 2.57	Z= 1.06	Z= -0.28	Z= 0.57
		p= 0.01(*)	*p*= 0.12	*p*= 0.24	*p*= 0.23	*p*= 0.27	*p*= 0.29	*p*= 0.77	*p*= 0.57
	Specificity	R= 0.26	R= -0.26	R= 0.39	H= 1.29	H= 9.15	Z= 0.38	Z= 0.75	Z= 1.71
		p= 0.26	*p*= 0.26	*p*= 0.08	*p*= 0.52	*p*= 0.01(*)	*p*= 0.71	*p*= 0.45	*p*= 0.09
		R_P= 0.51	R_P= -0.49	R= 0.26	H= 1.93	H= 2.34	Z= -0.22	Z= 1.42	Z= 1.4
		p= 0.023(*)	*p*= 0.03(*)	*p*= 0.27	*p*= 0.38	*p*= 0.31	*p*= 0.82	*p*= 0.15	*p*= 0.16
NNET	AUC	R= 0.035	R= -0.39	**R= 0.69**	H= 3.45	H= 8.75	Z= 1.29	Z= 0.56	Z= 1.78
		p= 0.88	*p*= 0.08	***p*<0.001**	*p*= 0.17	*p*= 0.01(*)	*p*= 0.19	*p*= 0.57	*p*= 0.07
		R_P= 0.70	**R_P= -0.72**	R= 0.42	H= 4.04	H= 4.53	Z= 0.45	Z= 1.32	Z= 1.63
		***p*<0.001**	***p*<0.001**	*p*= 0.07	*p*= 0.13	*p*= 0.1	*p*= 0.65	*p*= 0.18	*p*= 0.10
	Sensitivity	R= 0.32	R= -0.15	**R= 0.59**	H= 4.12	H= 8.84	Z= 1.43	Z= 0.80	Z= 1.71
		p= 0.17	*p*= 0.53	***p*<0.001**	*p*= 0.13	*p*= 0.01(*)	*p*= 0.15	*p*= 0.42	*p*= 0.08
		R_P= 0.73	**R_P= -0.74**	R= 0.49	H= 4.66	H= 5.33	Z= 0.68	Z= 1.51	Z= 1.86
		***p*<0.001**	***p*<0.001**	*p*= 0.02(*)	*p*= 0.09	*p*= 0.07	*p*= 0.49	*p*= 0.13	*p*= 0.06
	Specificity	R= 0.34	R=-0.14	**R= 0.57**	H= 4.38	H= 8.94	Z= 1.51	Z= 0.85	Z= 1.78
		p= 0.14	*p*= 0.56	***p*<0.001**	*p*= 0.11	*p*= 0.01(*)	*p*= 0.13	*p*= 0.39	*p*= 0.07
		R_P= 0.73	**R_P= -0.74**	R= 0.48	H= 5.03	H= 6.09	Z= 0.91	Z= 1.42	Z= 1.94
		***p*<0.001**	***p*<0.001**	*p*= 0.03(*)	*p*= 0.08	*p*= 0.04(*)	*p*= 0.36	*p*= 0.16	*p*= 0.05

Table 4.4 - Relationships between the three measures of model accuracy (AUC, Sensitivity and Specificity) and data characteristics or species traits. Spearman rank correlation coefficients (R) are used to assess the effect of continuous variables (upper rows); partial correlations (R_P) are used to asses the individual relevance of N and ROA (lower rows). The effects of qualitative species traits were assessed using Kruskal-Wallis (H) and Mann-Whitney U test (Z) on either the direct values of the accuracy measures (upper rows), or on the residuals of regressing them on N and ROA (lower rows). (*): Statistically significant relationships (p<0.05). Variables significant after applying a Bonferroni correction (p< 0.0060), are shown in bold. TER is the total extent of the distribution range (see text and Table 2

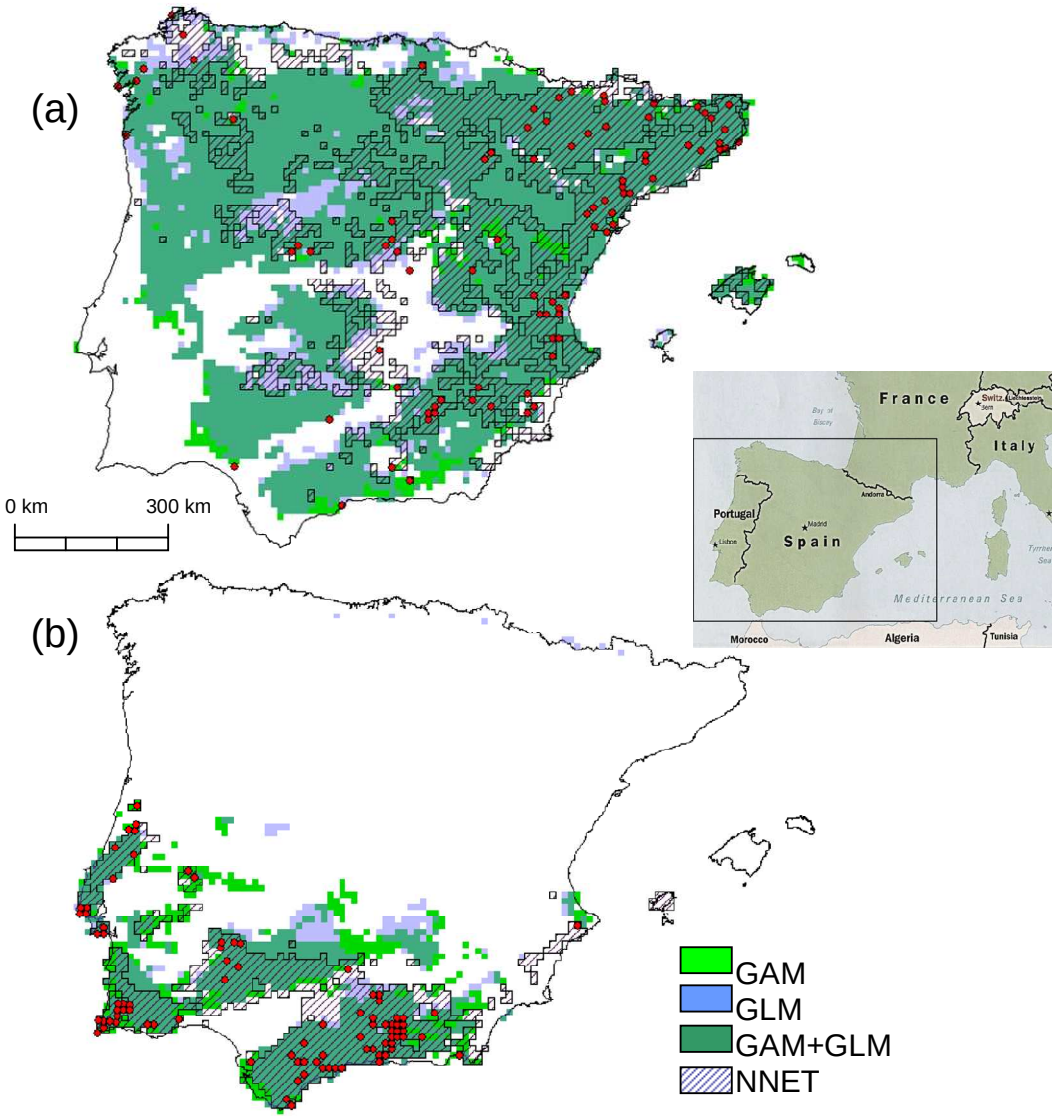


Fig. 4.3 - Differences between the predictive maps produced for a riparian species, *Coenagrion mercuriale* (a) and a species not linked to riparian habitats, *Cupido lorquinii* (b). Although data for both species have the same sample size (N = 87), GAM and NNET models performed better for *C. lorquinii* than for *C. mercuriale*.

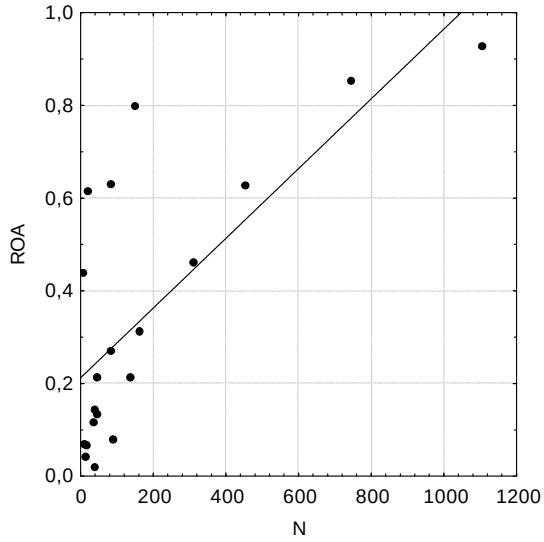


Fig. 4.4 - Relationship between the values of sample size (N) and the relative occurrence area (ROA).

Several species' traits seem to influence model performance to a lesser degree. The species with more restricted ecological requirements (i.e., the most marginal species) were modelled more accurately than those with less restricted requirements, but only in the case of NNET. In contrast to other studies (Brotons *et al*., 2004; Segurado & Araújo, 2004; Luoto *et al*., 2005), we did not find any strong relationship between the performance of GAM and GLM models and niche specialization (i.e., marginality). This is in agreement with Pöyry *et al*. (2008) and Newbold *et al*. (2009), who were not able to detect any effect of the niche width of butterflies on model accuracy. Besides, all SDM techniques, but specially GAM and NNET, seem to perform better with species not associated with riparian and humid conditions, a result also found by McPherson & Jetz (2007). Such poor performance may be associated with a poorer localization of wetlands in land cover maps, which hampers the inclusion of environmental variables related with the quality of aquatic habitats as predictors, thereby impeding using the actual determinants of the distribution of riparian species. Finally, we did not detect any influence on model accuracy of the variables measuring flight capacity and the total extent of the distribution range of the species. Hence, it can be assumed that the SDM

techniques used are not sensible to either how widespread is the species outside of the study area, nor to its dispersal capacity.

Both sample size and ROA altogether seem to interact with the influence of other species traits on model accuracy. Once their effect is removed some effects of species' traits appear, disappear or are reinforced. In particular, the residual analyses reveal a consistent, though weak, relationship between model performance and habitat detectability; species associated to easy-to-detect habitats are predicted more accurately by GAM models than those whose preferred habitats are smaller than the resolution of the available GIS layers. This also agrees with the results obtained by McPherson & Jetz (2007), where habitat detectability also had a secondary role on model accuracy. Besides, this result supports the idea commented above: the low detectability of riparian and humid habitats is associated with a poorer performance. On the other hand, the weak relationship between the better performance of GLMs for phytophagous species (in comparison with non-phytophagous) disappears after removing the effect of N and ROA, revealing that this minor relationship could be a spurious statistical artefact. Further analyses are needed to evaluate whether other species traits not considered in this work are important for the performance of SDM, beyond the mere limitations of data characteristics such as N or ROA.

Our results confirm that although species traits may affect SDM performance, prediction accuracy is mostly affected by the characteristics of the data. In fact, given the overall good results obtained by the three methods, we believe that more attention should be given to assessing the quality and/or adequacy of the data, rather than to electing a particular SDM technique. In general, species are modelled more accurately when sample sizes are larger, no matter the technique used. Moreover, ROA has an

additive effect to that of sample size, evidencing that selecting coarse extents of analysis to model the distribution of geographically restricted species may result in trivial models, able to discriminate such restricted distributions within a large geographical context and therefore obtaining high accuracy measures, but unable to capture the environmental response of the species with precision (Lobo, 2008; Jiménez-Valverde *et al*., 2008). The separate effects of N and ROA are difficult to determine due to the almost inevitable correlation between them (the species with more presence data have higher probabilities of being widely distributed in the considered region). Due to this, an unknown proportion of the effect on model performance generally attributed to low sample sizes may be due to a lesser relative occurrence area of presence data in the studied region; i.e. the inability of selecting reliable absences outside the environmental domain known to be used by the species when the number of observations is low (Austin & Meyers, 1996).

Based on these results, we advice researchers and conservation planners using SDM to ensure that the amount of data available is enough to obtain accurate models, and that the geographical focus (i.e., extent) of the analysis is the adequate to recover the environmental response of each particular species. In addition, special care should be taken while modelling species inhabiting inconspicuous habitats or strongly affected by interactions occurring at small spatial scales (see Hortal *et al.,* 2010). The problems associated to the prediction of the distributions of these species should be tackled by either using more precise predictors or resizing the scale (i.e., grain) of the analyses.

ACKNOWLEDGEMENTS

This research was supported by the Spanish "Ministerio de Ciencia e Innovación"

through the project CGL2008-03878.

REFERENCES

Austin, M. P. 1980. Searching for a model for use in vegetation analysis. *Vegetatio* **42**, 11-21.

Austin, M. P. & Meyers, J. A. 1996. Current approaches to modelling the environmental niche of eucalypts: implications for management of forest biodiversity. *Forest Ecology and Management* **85,** 95-106.

Brotons, L., Thuiller, W., Araujo, M. B. & Hirzel, A. H. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* **27**, 437-448.

Burgman, M. A. & Fox, J. C. 2003. Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation* **6**, 19-28.

Busby, J. R. 1991. BIOCLIM- a bioclimate analysis and prediction system. In: *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* (edited by Margules, C. R. & Austin, M. P.). CSIRO, Melbourne, 64-68.

Cabeza, M., Arponen, A., Jäättelä, L., Kujala, H., van Teeffelen, A., & Hanski, I. 2010. Conservation planning with insects at three different spatial scales. *Ecography*, doi:10.1111/j.1600-0587.2009.06040.x.

Castillejo, J. 1990. Babosas de la Península Ibérica. I. Los Ariónidos. Catálogo crítico y mapas de distribución. (Gastrópoda, Pulmonata, Arionidae). *Iberus* **9**, 331-345.

Chefaoui, R. M. & Lobo, J. M. 2007. Assessing the conservation status of an Iberian moth using pseudo-absences. *The Journal of Wildlife Management* **8**, 2507-2516.

Chefaoui, R. M. & Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling* **210**, 478-486.

Chefaoui, R. M., Hortal, J., & Lobo, J. M. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation* **122**, 327-338.

Clark Labs. 2000. *Global Change Data Archive Vol. 3. 1 km Global Elevation Model*. Clark University.

Clark Labs. 2003. *Idrisi Kilimanjaro*. Clark Labs, Worcester, MA.

Dixon, P. M., Ellison, A. M. & Gotelli, J. 2005. Improving the precision of estimates of the frequency or rare events. *Ecology* **86**, 1114-1123.

Elith, J. & Leathwick, J. R. 2007. Predicting species' distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* **13**, 165-175.

Engler, R., Guisan, A. & Rechsteiner, L. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* **41**, 263-274.

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

Galante, E. & Verdú, J. R. 2000. *Los Artrópodos de la "Directiva Hábitat" en España*. Ministerio de Medio Ambiente, Madrid.

García-Barros, E. & Herranz, J. 2001. Nuevas localidades de *Proserpinus proserpina* (Pallas, 1772) y *Graellsia isabelae* (Graells, 1849) del centro peninsular. *SHILAP Revista de Lepidopterología* **29**, 183-184.

Grenyer, R., Orme, C. D. L., Jackson, S. F., Thomas, G. H., Davies, R. G., Davies, T. J., Jones, K. E., Olson, V. A., Ridgely, R. S., Rasmussen, P. C., Ding, T-S., Bennett, P. M., Blackburn, T. M., Gaston, K. J., Gittleman, J. L. & Owens, I. P. F. 2006. Global distribution and conservation of rare and threatened vertebrates. *Nature* **444**, 93-96.

Grosso-Silva, J. M. 1999. Contribuição para o conhecimento dos lucanídeos (Coleoptera, Lucanidae) de Portugal. *Boletín de la S.E.A.* **25**, 11-15.

Grupo de Trabajo sobre Lucanidae Ibéricos 2000. Proyecto Ciervo Volante (PCV). *Boletín de la S.E.A.* **27**, 108-109.

Hernández, 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**, 773-785.

Hirzel, A. H., Hausser, J. & Perrin, N. 2007. *Biomapper 4.0.* Division of Conservation Biology, University of Bern, Laussane.

Hortal, J., Lobo, J. M. & Jimenez-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., & Baselga, A. 2008. Historical bias in biodiversity inventories affects the observed realized niche of the species. *Oikos* **117**, 847-858.

Hortal, J., Roura-Pascual, N., Sanders, N. J., & Rahbek, C. 2010. Understanding (insect) species distributions across spatial scales. *Ecography*, doi:10.1111/j.1600-0587.2009.06428.x.

Instituto Geográfico Nacional 1995. *Mapa de suelos del Atlas Nacional de España (Edafología)*. CSIC/IRNAS, Sevilla.

Jiménez-Valverde, A. & Lobo, J. M. 2006. The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions* **12**, 521-524.

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

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology* **10**, 196-205.

King, G. & Zeng, L. 2000. Logistic regression in rare events data. The Global Burden of Disease 2000 in Aging Populations, Research Paper No. 2 (edited by Havard Burden of Disease Unit, C. f. P. a. D. S.). http://www.hsph.harvard.edu/burdenofdisease/publications/papers/Logistic%20Regression.pdf

Legendre, P. & Legendre, L. 1998. *Numerical Ecology*. Elsevier, Amsterdam.

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

Lobo, J. M., Jiménez-Valverde, A. & Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* **17**, 145-151.

Lobo, J. M., Verdú, J. R. & Numa, C. 2006. Environmental and geographical factors affecting the Iberian distribution of flightless Jekelius species (Coleoptera: Geotrupidae). *Diversity and Distributions* **12**, 179-188.

Lobo, J. M., Baselga, A., Hortal, J., Jiménez-Valverde, A., & Gómez, J. F. 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Diversity and Distributions* **13**, 772-780.

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. 2010. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, doi:10.1111/j.1600-0587.2009.06039.x.

López-Sebastián, E., López, J.C., Juan, M. J. & Selfa, J. 2002. Primeras citas de mariposa isabelina en la Comunidad Valenciana. *Quercus* **193**, 10-13.

Luoto, M., Poyry, J., Heikkinen, R. K., & Saarinen, K. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography* **14**, 575-584.

Marmion, M., Luoto, M., Heikkinen, R. K. & Thuiller, W. 2008. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling* **220**, 3512-3520.

Martínez-Orti, A. 2004. Descripción de los moluscos terrestres del Valle del Najerilla. *Noticiario de la Sociedad Española de Malacología* **41**, 30-32.

McCullagh, P. & Nelder, J.A. 1989. *Generalized Linear Models*. Chapman & Hall, London.

McPherson, J. M. & Jetz, W. 2007. Effects of species' ecology on the accuracy of distribution models. *Ecography* **30**, 135-151.

McPherson, J. M., Jetz, W. & Rogers, D. J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology* **41**, 811-823.

Newbold, T., Reader, T., Zalat, S., El-Gabbas, A. & Gilbert, F. 2009. Effect of characteristics of butterfly species on the accuracy of distribution models in an arid environment. *Biodiversity and Conservation* **18**, 3629-3641.

Olden, J. D., Jackson, D. A. & Peres-Neto, P. 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society* **131**, 329-336.

Pérez-Bote, J. L., García, J. M., Ferri, F. & Moreno, J. A. 2001. Nueva cita de *Lucanus cervus* (Linnaeus, 1758) en Extremadura (Coleoptera: Lucanidae). *Boletín de la S.E.A.* **28**, 130.

Pöyry, J., Luoto, M., Heikkinen, R. K. & Saarinen, K. 2008. Species traits are associated with the quality of bioclimatic models. *Global Ecology and Biogeography* **17**, 403-414.

R Development Core Team 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.

Raimundo, R., Algarvio, R., Casas Novas, P. & Figuereiredo, D. 2001. Cartography of some species of xilophagous and xilomicetophagous insects in Quercus suber and Quercus rotundifolia in Alentejo using GIS. *Suplemento ao Boletim da Sociedade Portuguesa de Entomología* **6**, 459-468.

Reutter, B., Helfer, V., Hirzel, A. H. & Vogel, P. 2003. Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography* **30**, 581-590.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks.* Cambridge University Press.

Rosas, G., Ramos, M. A. & García Valdecasas, A. 1992. *Invertebrados españoles protegidos por convenios internacionales*. ICONA, Madrid.

Schröder, B. 2004. ROC Plotting and AUC Calculation Transferability Test. Institute for Geoecology, Potsdam University. Potsdam. http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html.

Segurado, P. & Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. *Journal of Biogeography* **31**, 1555-1568.

Seoane, J., Carrascal, L. M., Alonso, C. L. & Palomino, D. 2005. Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling* **185**, 299-308.

Soria, S., Abos, F. & Martín, E. 1986. Influencia de los tratamientos con diflubenzurón ODC 45% sobre pinares en las poblaciones de Graellsia isabelae (Graells) (Lep. Syssphingidae) y reseña de su biología. *Boletín de sanidad vegetal* **12**, 29-50.

StatSoft, I. 2001. *Statistica.* Data analysis software system, Tulsa.

Stockwell, D. R. B. & Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling* **148**, 1-13.

Verdú, J. R. & Galante, E. 2002. Climatic stress, food availability and human activity as determinants of endemism patterns in the Mediterranean region: the case of dung beetles (Coleoptera, Scarabeoidea) in the Iberian Peninsula. *Diversity and Distributions* **8**, 259-274.

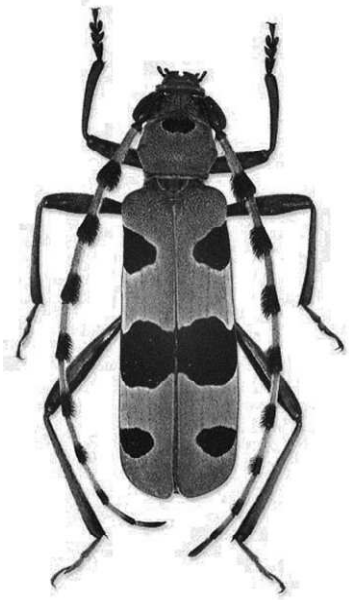
Verdú, J. R. & Galante, E. (eds.) 2006. *Libro Rojo de los Invertebrados de España*. Ministerio de Medio Ambiente, Madrid.

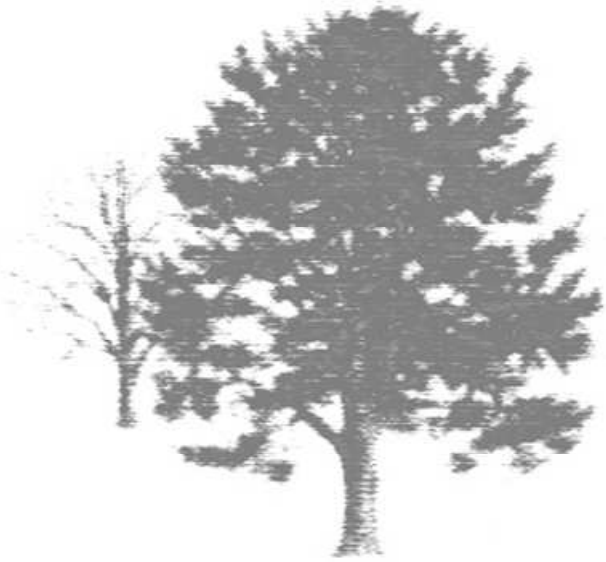
Viejo, J. L. 1992. Biografía de un naturalista y biología del lepidóptero por él descrito. Graells y la Graellsia. *Quercus* **74**, 22-30.

Wood, S. N. & Augustin, N. H. 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling* **157**, 157-177.

Zaniewski, A. E., Lehmann, A. & Overton, J. M. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* **157**, 261-280.

Zweig, M. H. & Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**, 561-577.





CONCLUSIONES Y FUTURAS LÍNEAS DE TRABAJO

CONCLUSIONES

I. Evaluación del comportamiento de distintas técnicas predictivas empleadas para la modelización de la distribución potencial de invertebrados usando datos de presencia disponibles en museos, atlas y bases de datos.

Métodos que usan exclusivamente presencias: Ecological Niche Factor Analysis (ENFA) y Modelo de Envoltura Ambiental (MDE):

- Estos métodos, al trabajar sin ausencias verdaderas, alcanzan peores resultados de validación y tienen tendencia a la sobrepredicción en comparación con las técnicas que usan ausencias fiables, por lo que deben ser usados con precaución y siempre que se quiera obtener representaciones geográficas que se aproximen a la distribución potencial. Sin embargo, pueden ser muy útiles como método de obtención de pseudo-ausencias.
- Además de calcular mapas de porcentaje de probabilidad de presencia, ENFA permite descubrir la respuesta ambiental de las especies, ofreciendo buenos resultados incluso con un número bajo de presencias.
- Para evitar introducir falsas presencias en el modelo, es necesario revisar la validez de las citas antiguas cuando exista la sospecha de que se haya producido un cambio en el hábitat o no se haya vuelto a registrar la presencia de la especie en muestreos recientes.

Método de obtención de pseudo-ausencias:

- Se han comprobado diversas formas de obtención de pseudo-ausencias: al azar, mediante MDE (y MDE expandido) o a partir de ENFA (con diferentes umbrales de adecuación de hábitat). Los resultados de validación para los modelos obtenidos mediante cualquiera de ellos han sido muy altos ya que todos ellos son, inevitablemente, capaces de discriminar con fiabilidad los lugares de presencia de las ausencias ambientales.
- Se han obtenido modelos más precisos al seleccionar las pseudo-ausencias con ENFA o MDE a partir de zonas ambientalmente alejadas del óptimo nicho ecológico, que seleccionándolas al azar. Además, a mayor distancia ambiental entre las presencias y las pseudo-ausencias, mayor capacidad explicativa y precisión de los modelos. Este resultado puede dar lugar a equivoco, ya que la capacidad explicativa de estos modelos no significa que las hipótesis de distribución generadas sean más fiables.
- Seleccionar un número diez veces superior de pseudo-ausencias que de presencias parece conveniente para la elaboración de los modelos.

Métodos basados en ausencias y presencias: Modelos Lineales Generalizados (GLM), Modelos Generalizados Aditivos (GAM) y Modelos de Redes Neuronales (NNET):

- No se han hallado diferencias significativas respecto a la precisión de los modelos ni el área predicha entre las distintas técnicas.
- Al haberse empleado pseudo-ausencias obtenidas de zonas ambientalmente alejadas del óptimo de la especie, las predicciones generadas por estas técnicas tienen tanto porcentajes de variabilidad explicada como valores de medidas de precisión altos.

Sin embargo, estos resultados no garantizan la misma fiabilidad que si se tratara de modelos obtenidos a partir de ausencias verdaderas.

- El método por el cual se hayan obtenido previamente las pseudo-ausencias influye en los porcentajes de variabilidad explicada resultantes, en las medidas de validación y, sobre todo, en el grado de restricción de la distribución estimada (ver Fig. I). Esta información permite al investigador decidir la táctica necesaria según el tipo de predicción que se desee obtener, si más cercana a la distribución potencial o a la real, simplemente eligiendo uno u otro método de obtención de pseudo-ausencias.
- Los modelos que usan unicamente presencias (ENFA, MDE) tienen tendencia a sobrepredecir y generar una distribución potencial, mientras que al incorporar las ausencias en el modelo (GLM) se restringe esta predicción hasta que se va acercando a la "distribución real" a medida que esas ausencias están más cercanas al espacio ambiental de las presencias. Una predicción de la distribución real en la que las fuerzas contingentes de restricción tales como los factores históricos, las interacciones bióticas o las limitaciones de dispersión juegan un papel importante, necesitaría obligadamente incorporar ausencias de aquellas localidades favorables pero inhabitadas.
- La principal dificultad se encuentra en predecir distribuciones próximas a la realidad cuando las especies no están en equilibrio, puesto que los datos de presencia y de ausencia pueden ser posibles bajo similares condiciones ambientales y por tanto los modelos son menos precisos.
- Cuando el tamaño de muestra disponible para la especie es pequeño, como suele suceder con invertebrados protegidos, dividirlo para obtener un grupo de entrenamiento y otro de validación no es conveniente, por lo que es aconsejable usar un procedimiento Jack-knife para obtener las medidas de validación.

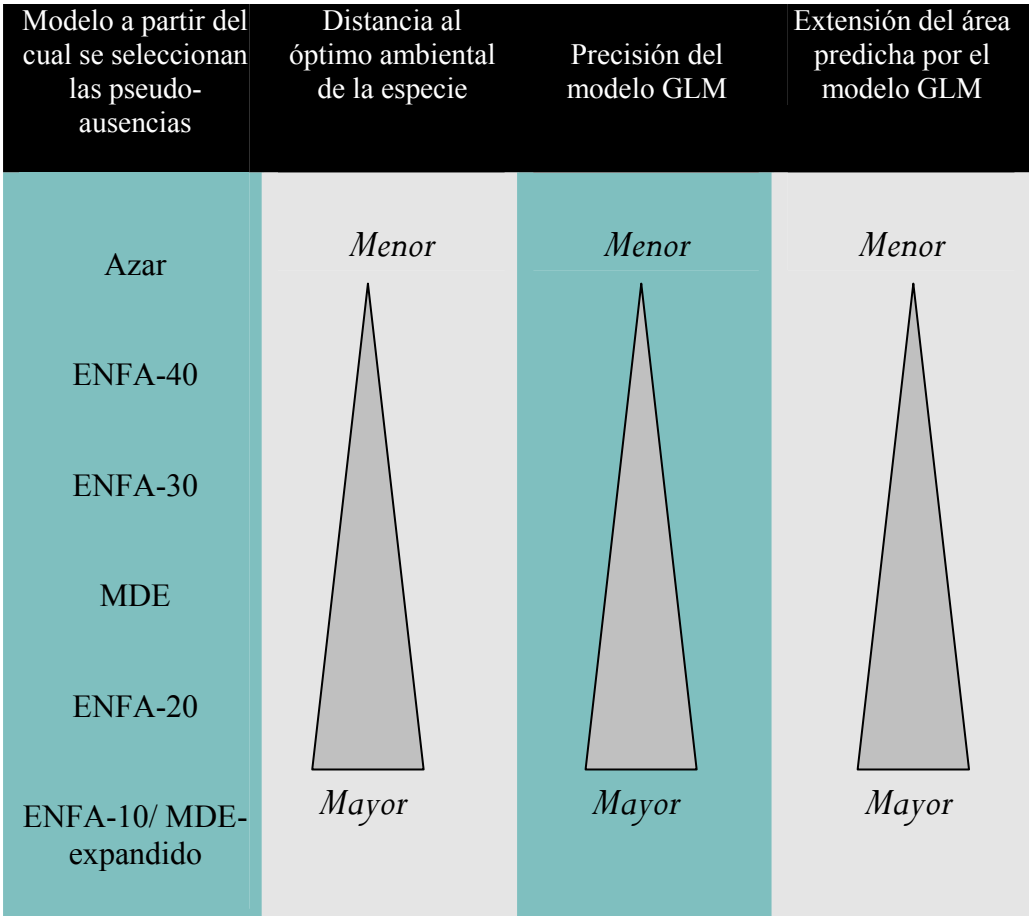


Fig. I - Características de las predicciones obtenidas por los modelos generalizados lineales (GLM) en función del método previo de selección de pseudo-ausencias.

II. Delimitación del nicho ambiental ocupado por una especie y las variables que afectan en mayor medida a su presencia mediante modelos predictivos.

- En esta Tesis se ha comprobado que es posible identificar las principales variables explicativas responsables de la distribución de especies de invertebrados, así como delimitar su nicho ambiental a partir de la información que proporcionan los datos de presencia disponibles en museos, atlas y bases de datos.
- Al modelizar invertebrados de diversas características ecológicas, las variables más significativas para la presencia de cada especie se seleccionan previamente mediante un análisis de regresión logística individual (GLM). Posteriormente, las variables pueden ser clasificadas en grupos para testar modelos que combinen diferentes tipos de variables. También es posible introducir variables espaciales (longitud, latitud) para determinar la importancia de factores no considerados y desconocidos que produzcan un patrón espacial.
- La radiación solar y la presencia de suelos calcáreos resultan determinantes para la presencia de *Copris hispanus*, mientras que *Copris lunaris* requiere suelos silíceos y elevadas precipitaciones. Dentro de la Comunidad de Madrid *C. hispanus* se encuentra en áreas de clima mediterráneo-seco, en tanto que *C. lunaris* prefiere áreas de clima alpino-húmedo; ambas especies comparten un espacio ambiental intermedio dentro de este gradiente.
- Por otra parte, las variables que condicionan en mayor medida la presencia de *Graellsia isabelae* son: la precipitación estival (en un rango de 1250 mm a 3250 mm), la aridez y la altitud media. Esta especie prefiere hábitats con condiciones montañosas de rango medio.

- Los modelos predictivos también han permitido determinar que *Pinus sylvestris* y *Pinus nigra* son las plantas más relacionadas con la presencia de *G. isabelae*.
- En esta Tesis se propone que, una vez obtenidos los modelos de distribución de especies fitófagas mediante variables ambientales, la información relativa a la distribución de las plantas nutricias sea utilizada para filtrar y ajustar aún más estas predicciones, en lugar de ser introducidas en las primeras fases de elaboración del modelo.

III. Evaluación de los efectos de los datos y las características ecológicas de las especies en la precisión de los modelos de distribución de invertebrados.

- Las características más influyentes en la precisión de las diferentes técnicas predictivas examinadas (GLM, GAM y NNET) han sido las relacionadas con las características de los datos: el tamaño de muestra y el área de ocurrencia relativa (ROA). Los modelos más precisos se han obtenido para las especies de las que se disponía de mayor tamaño de muestra (a partir de 200 presencias se obtienen valores de AUC superiores a 0.98 y de sensitividad mayores de 0.94) o menor ROA.
- Todas las técnicas predictivas han obtenido buenos valores de precisión sin diferencias significativas entre ellas. Estos resultados de precisión son previsibles debido al uso de pseudo-ausencias seleccionadas fuera del espacio ambiental de la especie. Por lo tanto, hay que subrayar la conveniencia de prestar mayor atención a la calidad de los datos que a la elección de la técnica predictiva.
- En menor medida, otras características de las especies tales como la marginalidad, el tipo de hábitat, el nivel trófico y la detectabilidad del hábitat parecen influir en las medidas de precisión de los modelos. La especialización del nicho (marginalidad) parece influir unicamente a los modelos NNET, que actúan con mayor precisión con las especies con requerimientos ecológicos restringidos. Todas las técnicas coinciden en predecir mejor las especies no asociadas a condiciones húmedas y riparias, posiblemente debido a la dificultad de introducir variables predictivas de tales hábitats en los modelos.
- Por otra parte, las características de los datos (el área de ocurrencia relativa junto al tamaño de muestra) parecen crear un artefacto estadístico, interaccionando con otras

características ecológicas como, por ejemplo, la detección del hábitat (revela una mejor predicción de las especies que pertenecen a hábitats detectables) y el grupo trófico, para afectar a la calidad de los modelos de distribución.

IV. Obtención de explicaciones biogeográficas que justifican la distribución actual de las especies.

- Al describir la distribución potencial de *G. isabelae* mediante modelos basados en presencias se ha encontrado que ésta es 2.7 veces mayor que la real. La especie no ocupa toda el área apropiada para su presencia en la zona oeste de la Península Ibérica, probablemente por causas distintas a los factores ambientales, por lo que se encuentra en una situación de no equilibrio con el clima.
- Se sugiere que la actual distribución de la especie está asociada con el dinamismo de sus plantas nutricias durante los periodos glaciares del Holoceno en los que los bosques de *Pinus sylvestris* disminuyeron considerablemente en la parte noroeste de la Península.

V. Evaluación del estado de conservación de determinadas especies de invertebrados protegidos examinando la distribución de sus poblaciones y la eficacia de las reservas para preservarlas.

- Se han encontrado deficiencias en la red de Espacios Naturales Protegidos para la protección de las poblaciones de *C. hispanus* y *C. lunaris*, pues no representaban una extensión suficiente de hábitat adecuado para ellas. Sin embargo, la Red Natura 2000 ha incrementado la extensión y la conectividad de hábitats favorables para estas dos especies de *Copris.*
- A partir de los modelos de distribución obtenidos para *Graellsia isabelae*, se han identificado ocho poblaciones separadas y con citas históricas. De entre ellas, las poblaciones de Cataluña y Montalbán (Teruel) no se encuentran suficientemente protegidas por los Lugares de Interés Comunitario (LIC). Por otra parte, nuestro estudio descarta que la especie se encuentre en expansión gracias a las reforestaciones.
- La conservación de *G. isabelae* depende de la de los bosques de *Pinus sylvestris* y *P. nigra* situados tanto en el interior de los LIC como en sus inmediaciones. La reintroducción de la especie en estos hábitats mejoraría su conservación.

LÍNEAS DE FUTURO

En lo relativo al comportamiento de los modelos predictivos:

En esta Tesis se han identificado algunos aspectos sobre el funcionamiento de los modelos predictivos que aún quedan por evaluar de manera más precisa. Sería conveniente seguir estudiando el comportamiento de las técnicas en función de la variación de la cantidad de datos de presencia y del area de ocurrencia relativa con un número mayor de especies para determinar la influencia de las características de los datos de forma más exacta.

Asimismo, convendría perfeccionar la predicción de especies asociadas a hábitats dificilmente caracterizables con capas digitales, estudiando posibilidades tales como cambios de escala (grano) de los análisis o explorando variables más precisas (p. ej. para especies riparias o de hábitats húmedos).

En relación a la conservación de las especies de invertebrados protegidos:

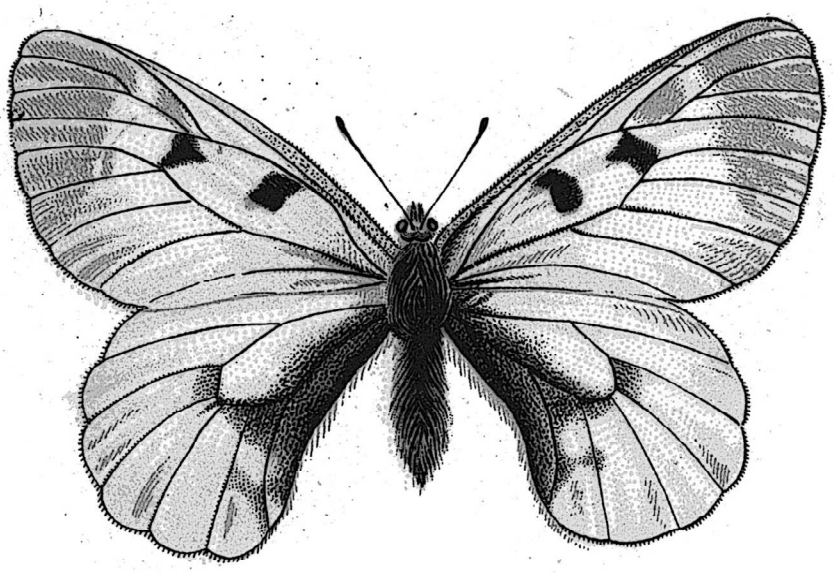
Siguiendo la Directiva Hábitats, cada comunidad autónoma deberá emprender acciones de mejora y restauración de hábitats y de protección de especies concretas. Aunque aún son escasas, se empiezan a emprender acciones para determinar el estado de las poblaciones y medidas de conservación para estas especies (p. ej. como el de la Junta de Extremadura, http://xtr.extremambiente.es/artropodos/; o en el LIC Aiako Harria; http://www.lifeaiakoharria.net/datos/documentos/AH_invertebrados%20directiva.pdf).

De forma similar, aquí se proponen las siguientes líneas futuras de trabajo:

- Con respecto a *G. isabelae*, sería apropiado realizar muestreos en zonas donde la presencia sea probable para poder constatar su ausencia (como ya se está haciendo

con las poblaciones francesas http://www.orleans.inra.fr/orleans_eng/les_unites/ur_zoologie_forestiere/). Además, debería realizarse un seguimiento de sus poblaciones para determinar la abundancia del insecto en ellas, efectuándose reintroducciones en caso necesario.

- El Parque Regional de la Cuenca Alta del Manzanares, señalado "hot-spot" de biodiversidad, destaca como un ENP conveniente para la conservación de las dos especies del género *Copris* presentes en la Comunidad de Madrid. Sería conveniente identificar lugares con esas mismas características para estudiar las especies presentes en ellos y proponer su protección.
- En el caso de no ser posible hacer un seguimiento especie por especie, sería recomendable realizar un análisis "gap" del conjunto de especies amenazadas con el propósito de evaluar la protección que ofrece la Red Natura 2000 y estimar posibles mejoras en el diseño de los LIC.

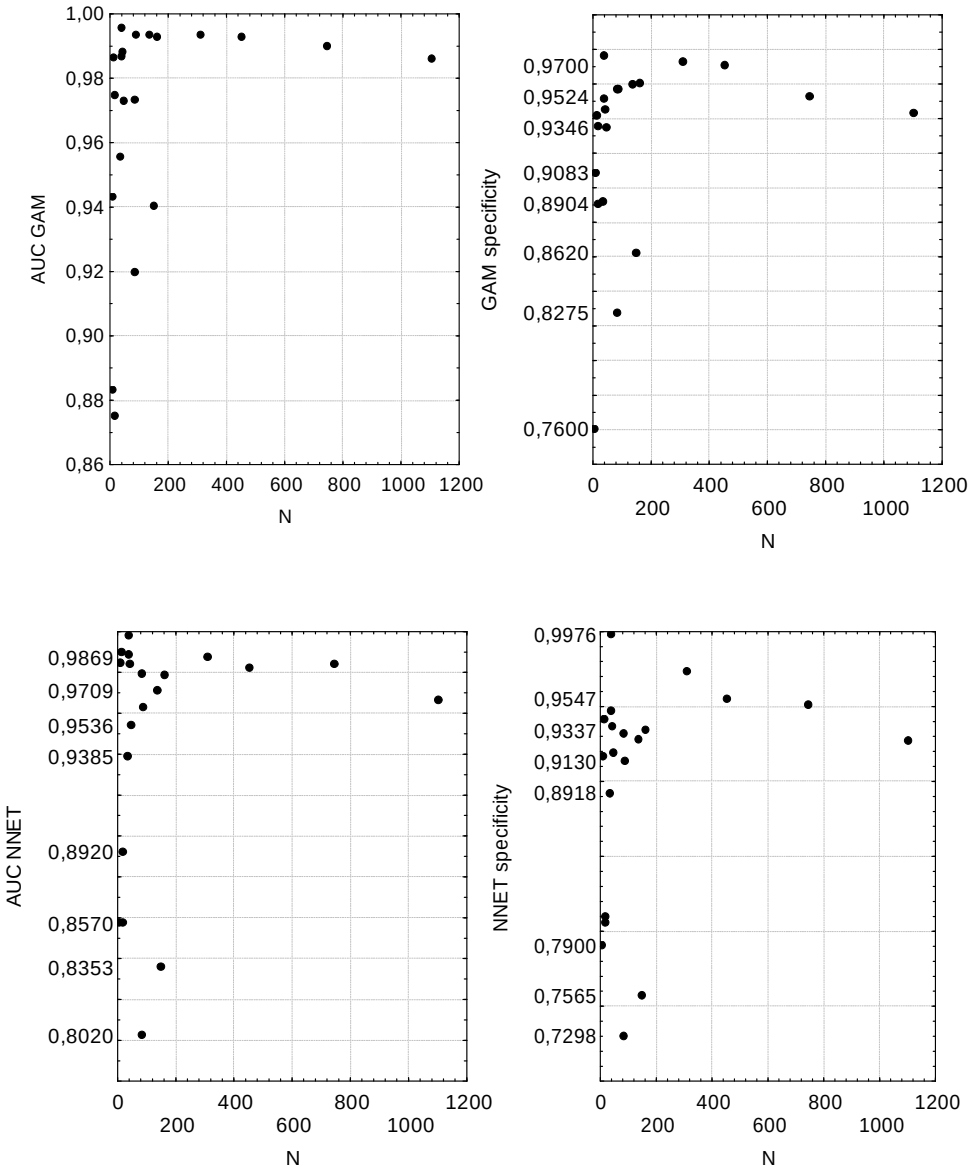


ANEXOS

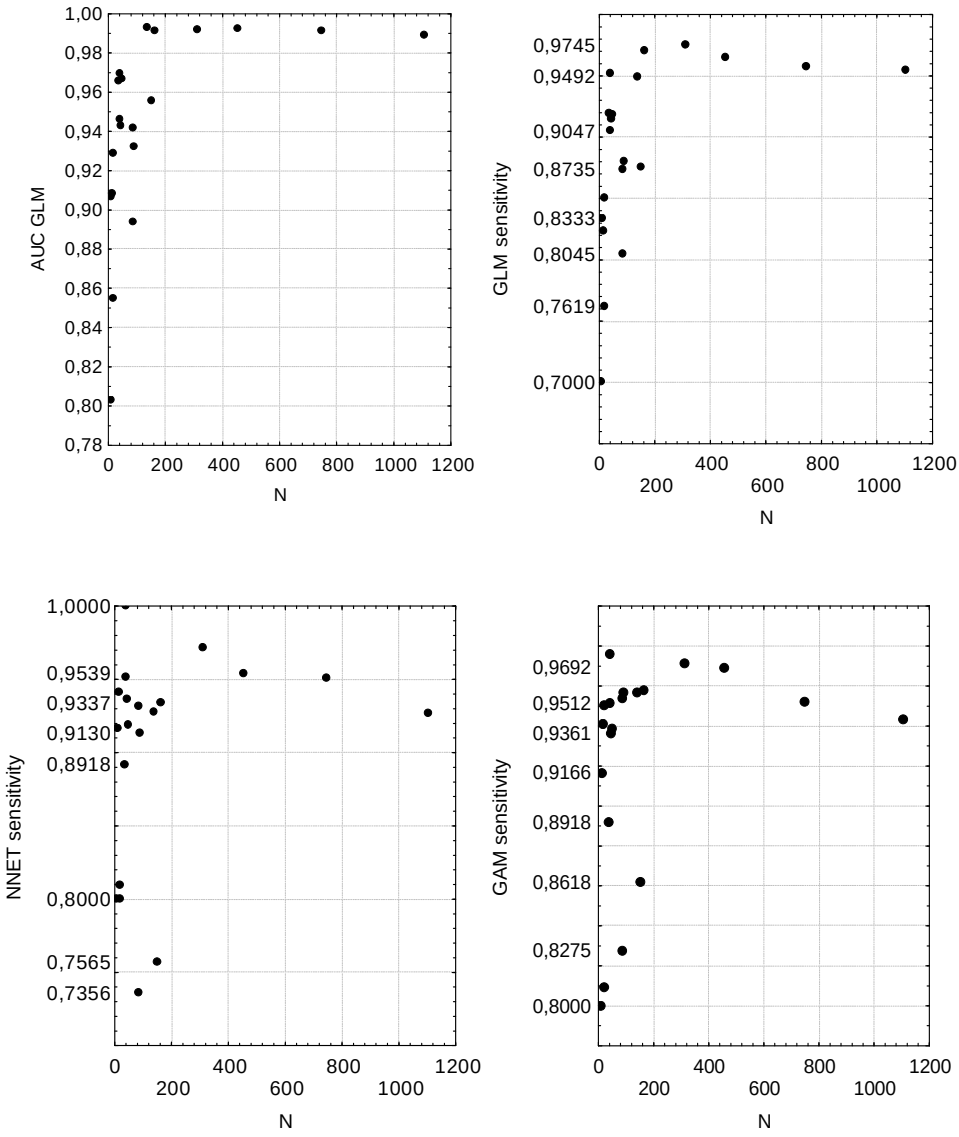
ANEXO I

Apéndice I (correspondiente al Capítulo IV) – Diagramas de puntos ("scatterplots") de los análisis de correlación significativos entre las medidas de precisión de los modelos (AUC, sensitividad y especificidad) y el tamaño de muestra (N).

Appendix I - Scatterplots of significant correlation analyses between accuracy measures (AUC, sensitivity and specificity) and data size (N).



Appendix I - Scatterplots of significant correlation analyses between accuracy measures (AUC, sensitivity and specificity) and data size (N).

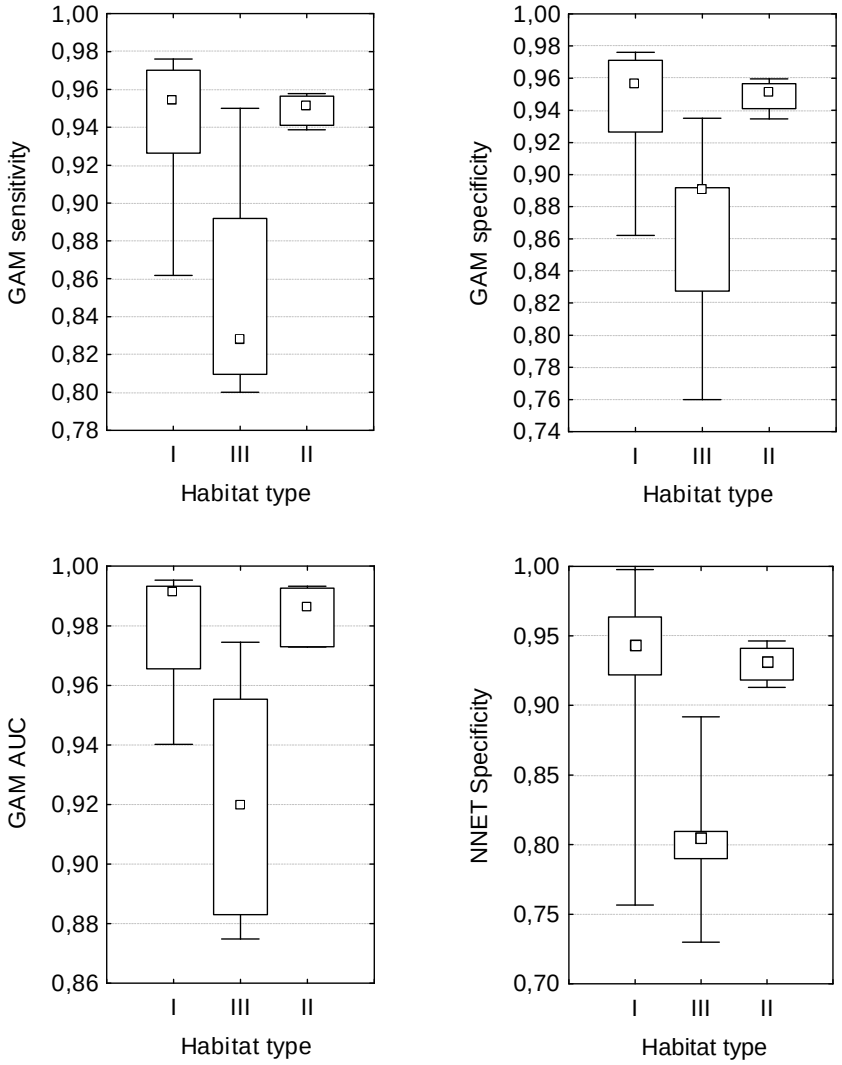


ANEXO II

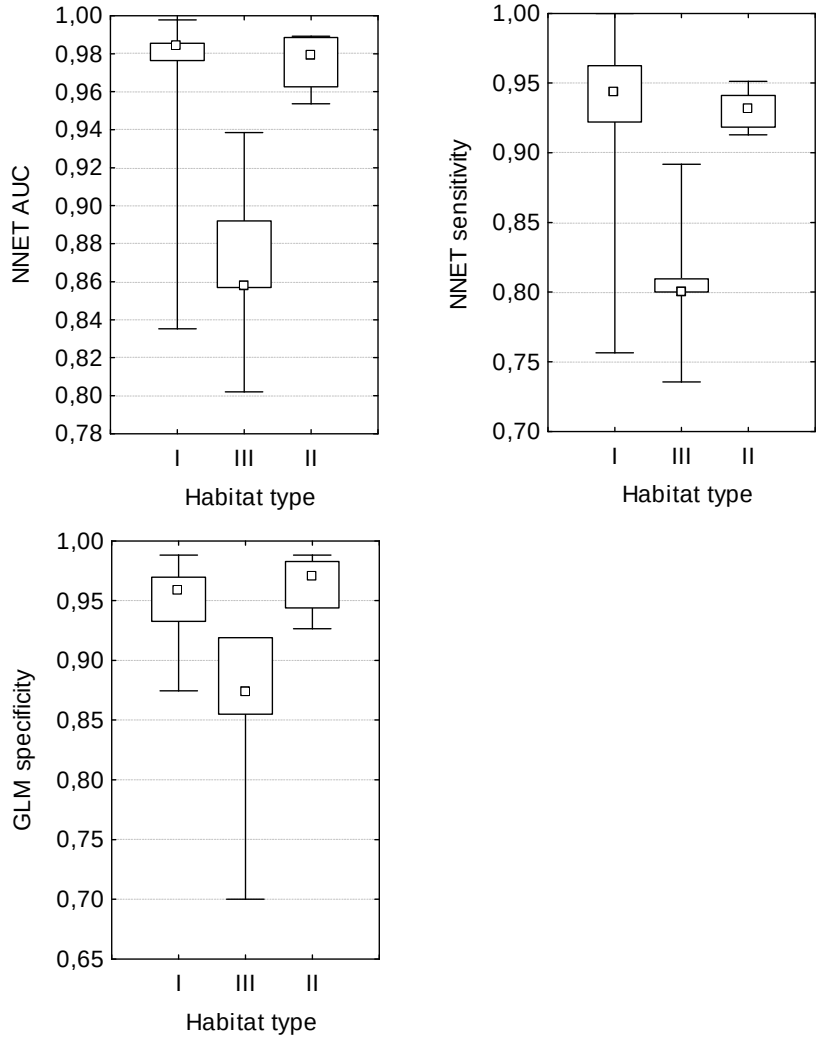
Apéndice II (correspondiente al Capítulo IV) – Diagramas de caja ("Box-plots") que representan la relación entre las medidas de precisión de los modelos y el tipo de hábitat de las especies del Capítulo IV. Las especies asociadas a hábitats riparios o húmedos obtienen modelos menos precisos.

El punto medio muestra la mediana de la respuesta de cada tipo de hábitat frente a cada medida de precisión. Los extremos inferiores y superiores de la caja ("box") señalan los percentiles del 25% y del 75% respectivamente. Los bigotes ("whiskers") señalan los valores máximos y mínimos. (I = Hábitats boscosos y montañosos; II = Estepas y hábitats mixtos; III = Hábitats riparios y húmedos).

Appendix II - Accuracy measures results by habitat type (I = Woods and Mountainous habitats, II = Grasslands and varied habitats, III = Riparian and humid habitats). Less accurate models are obtained for species associated to riparian and humid habitats. The middle point shows the median response for each habitat type and score combination. The bottom and top of the box show the 25 and 75 percentiles respectively. The whiskers show minimum and maximum values.



Appendix II - Accuracy measures results by habitat type (I = Woods and Mountainous habitats, II = Grasslands and varied habitats, III = Riparian and humid habitats). Less accurate models are obtained for species associated to riparian and humid habitats. The middle point shows the median response for each habitat type and score combination. The bottom and top of the box show the 25 and 75 percentiles respectively. The whiskers show minimum and maximum values.



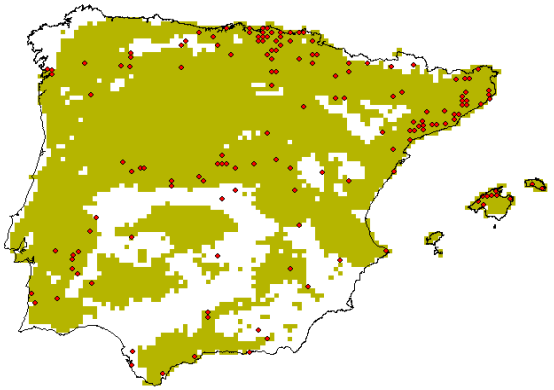
ANEXO III

Puntos de presencia y distribuciones potenciales de las 20 especies de invertebrados protegidos.

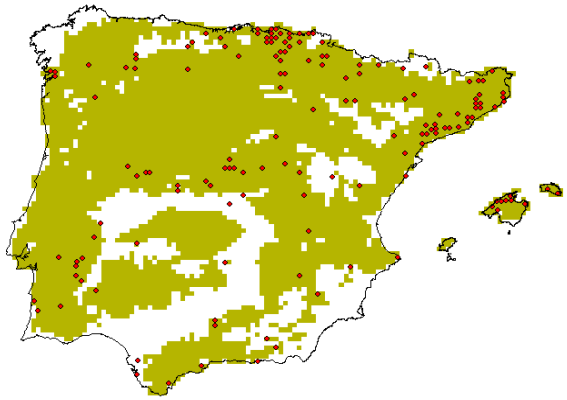
Cerambyx cerdo (Linnaeus, 1758; Coleoptera: Cerambycidae)

Presencias observadas (n=152)

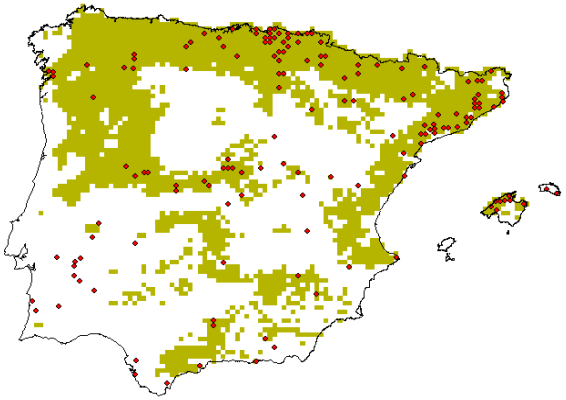
Distribución potencial (GAM)



Distribución potencial (GLM)



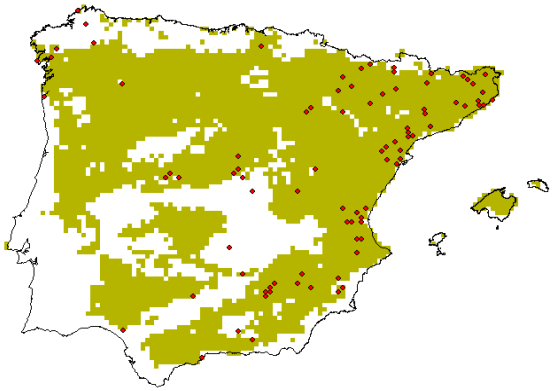
Distribución potencial (NNET)



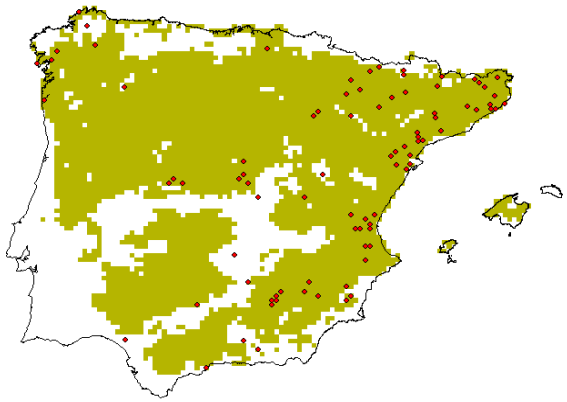
Coenagrion mercuriale (Charpentier,1840; Odonata: Coenagrionidae)

Presencias observadas (n=87)

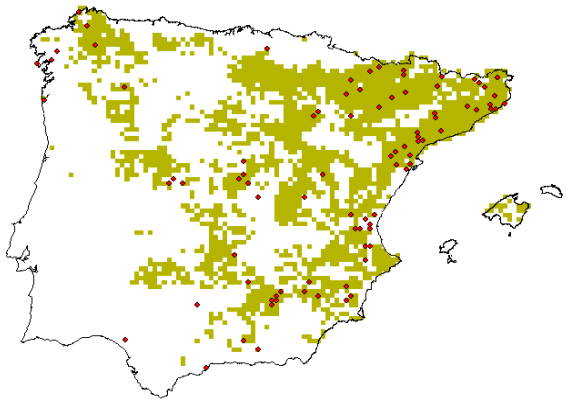
Distribución potencial (GAM)



Distribución potencial (GLM)

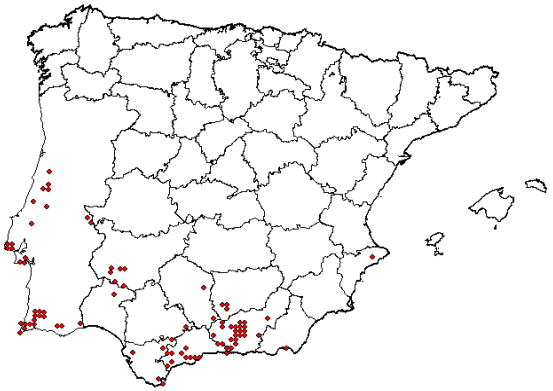


Distribución potencial (NNET)

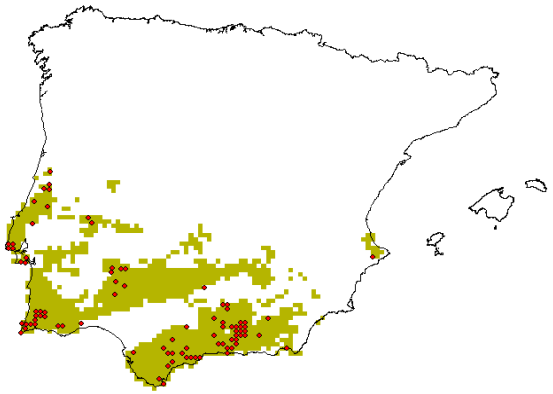


Cupido lorquinii (Herrich-Schäffer, 1847; Lepidoptera: Lycaenidae)

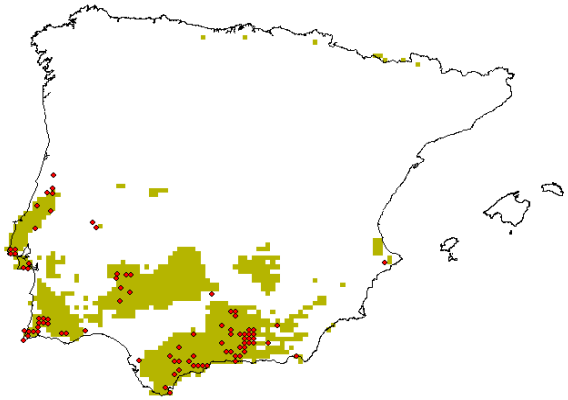
Presencias observadas (n=87)



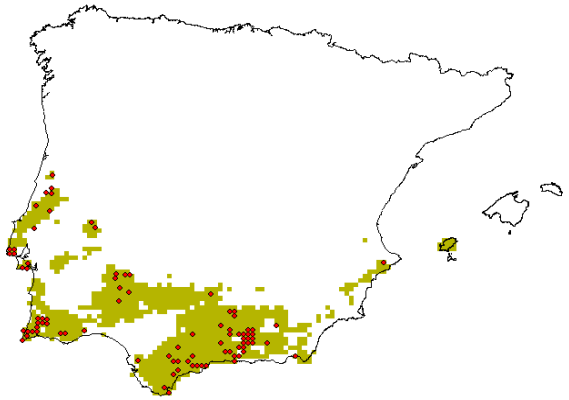
Distribución potencial (GAM)



Distribución potencial (GLM)

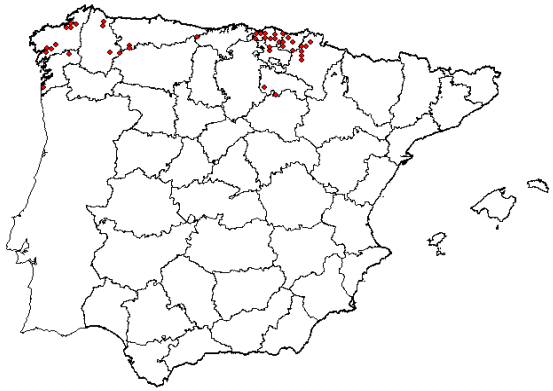


Distribución potencial (NNET)

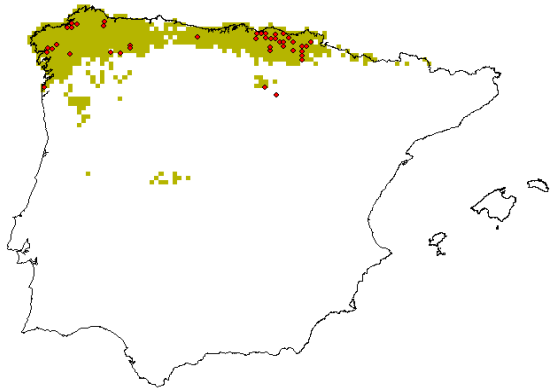


Elona quimperiana (Férusac, 1821; Pulmonata: Elonidae)

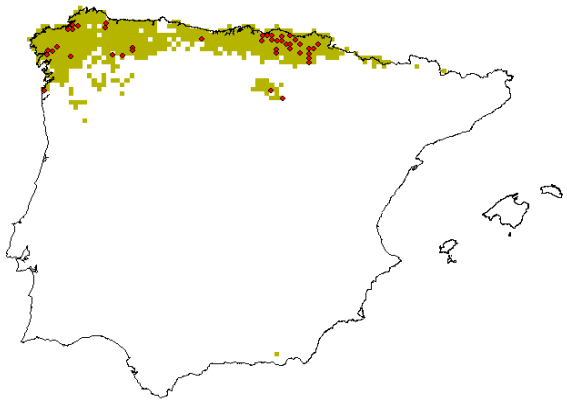
Presencias observadas (n=41)



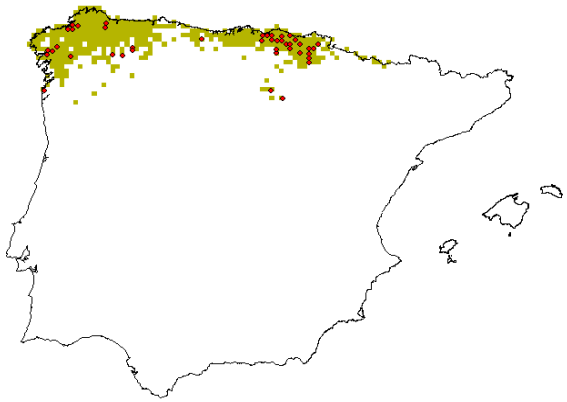
Distribución potencial (GAM)



Distribución potencial (GLM)



Distribución potencial (NNET)



Eriogaster catax (Linnaeus, 1758; Lepidoptera: Lasiocampidae)

Presencias observadas (n=12)

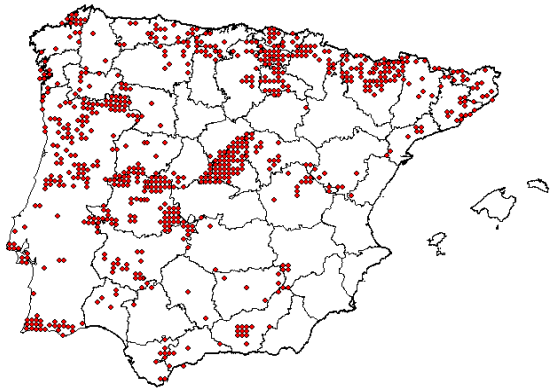
Distribución potencial (GAM)

Distribución potencial (GLM)

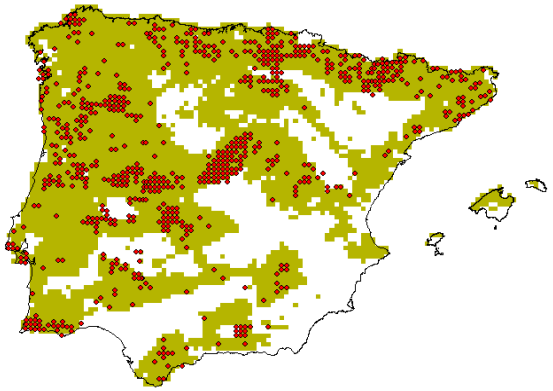
Distribución potencial (NNET)

Euphydryas aurinia (Rottemburg, 1775; Lepidoptera: Nymphalidae)

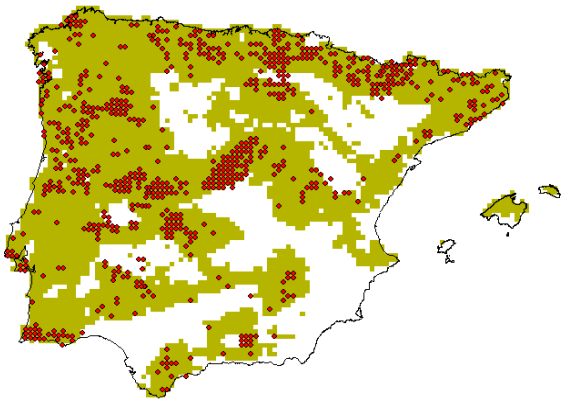
Presencias observadas (n=749)



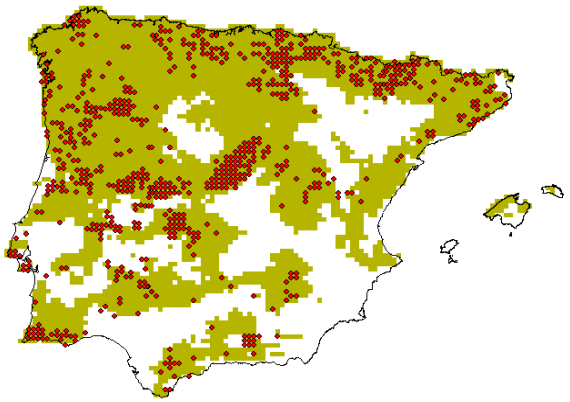
Distribución potencial (GAM)



Distribución potencial (GLM)

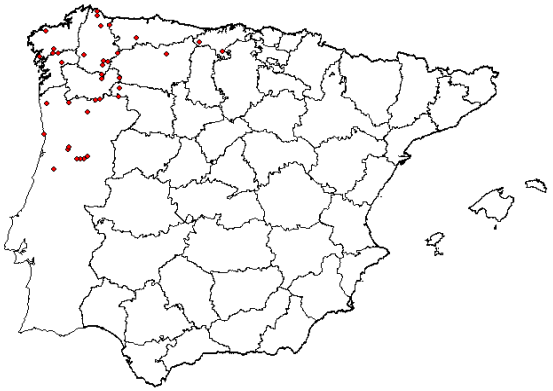


Distribución potencial (NNET)

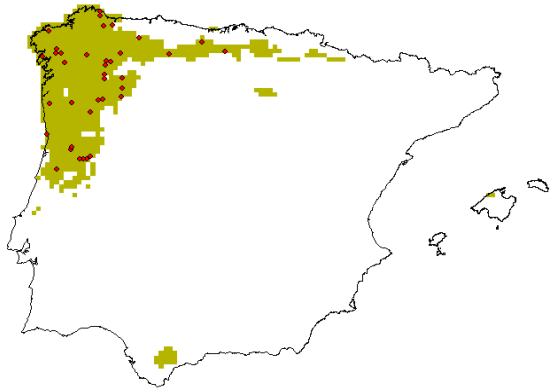


Geomalacus maculosus (Allman, 1843; Pulmonata: Arionidae)

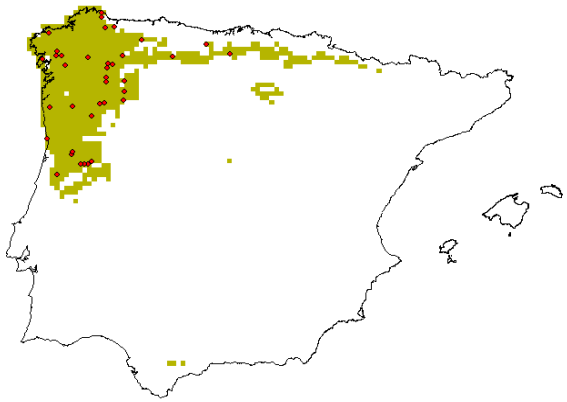
Presencias observadas (n=37)



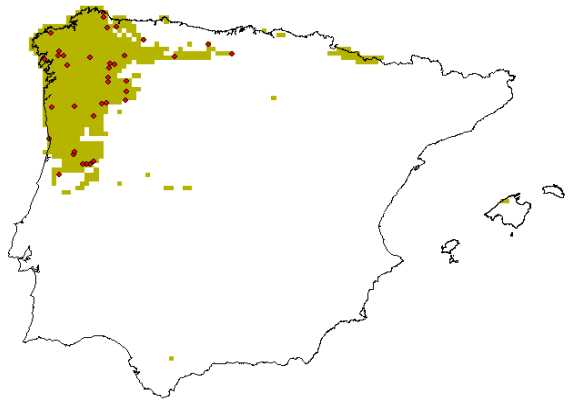
Distribución potencial (GAM)



Distribución potencial (GLM)

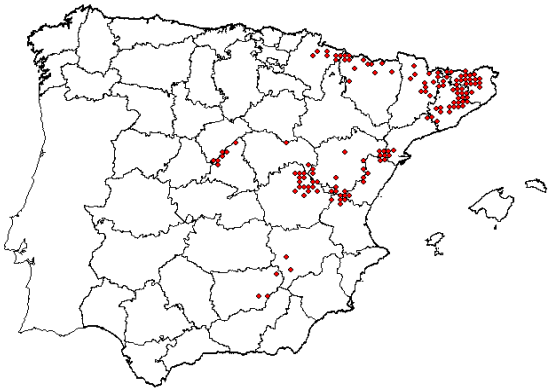


Distribución potencial (NNET)

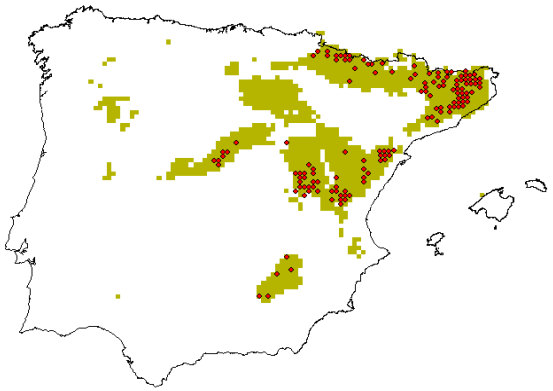


Graellsia isabelae (Graells, 1849; Lepidoptera: Saturniidae)

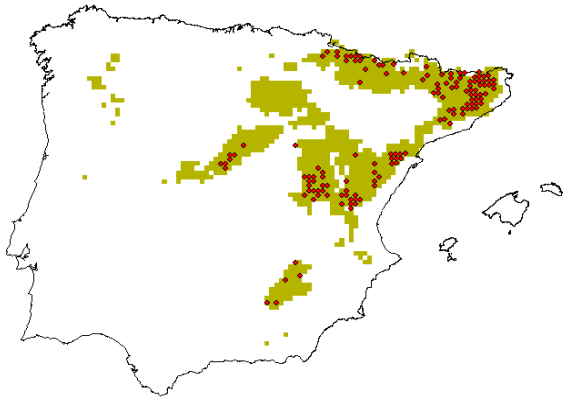
Presencias observadas (n=138)



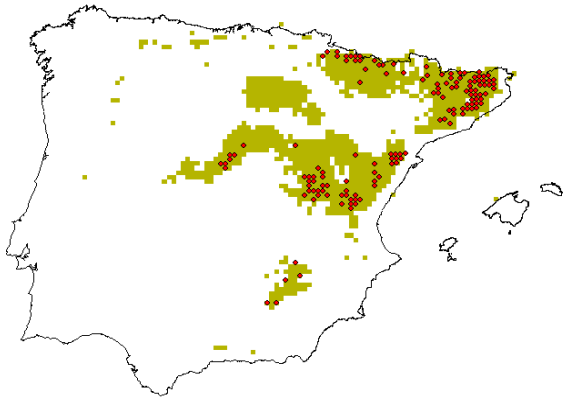
Distribución potencial (GAM)



Distribución potencial (GLM)

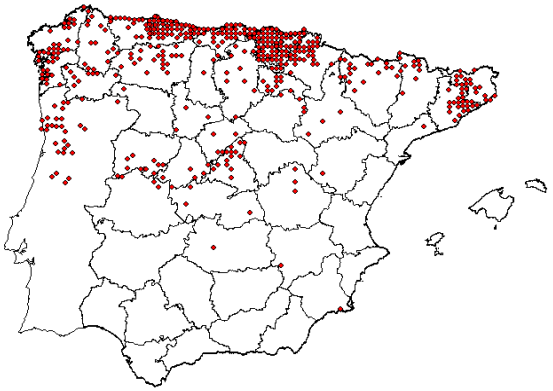


Distribución potencial (NNET)

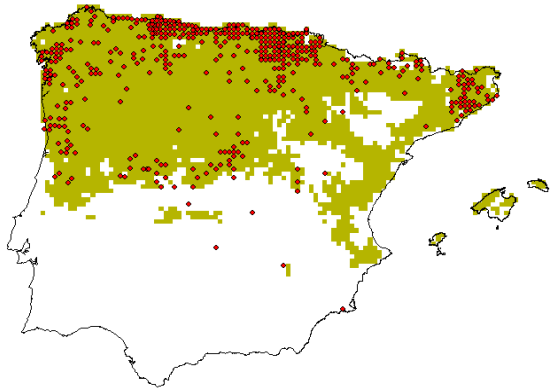


Lucanus cervus (Linnaeus, 1758; Coleoptera: Lucanidae)

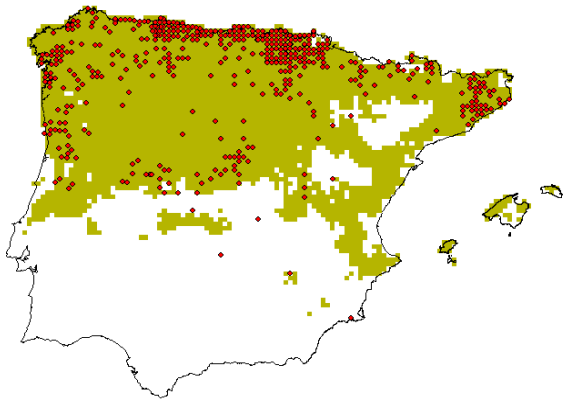
Presencias observadas (n=456)



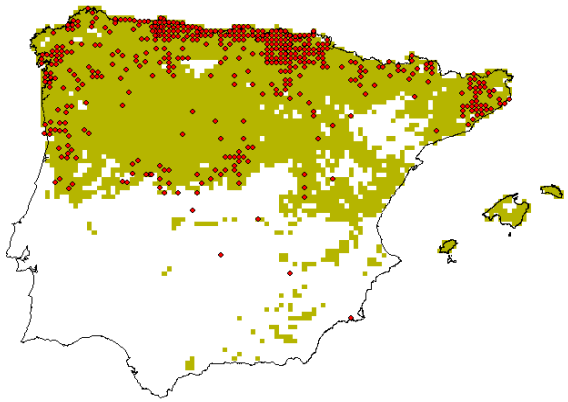
Distribución potencial (GAM)



Distribución potencial (GLM)



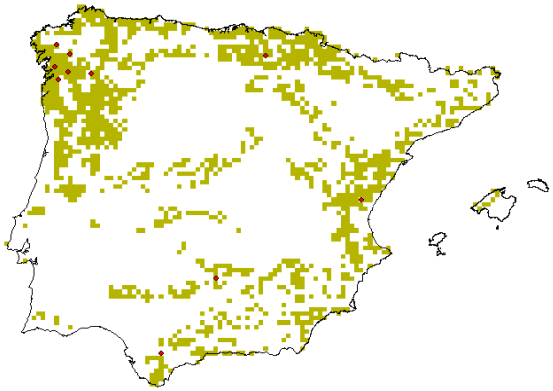
Distribución potencial (NNET)



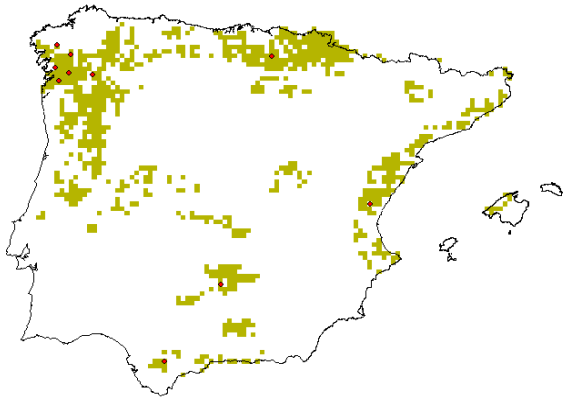
Macromia splendens (Pictet, 1843; Odonata: Corduliidae)

Presencias observadas (n=10)

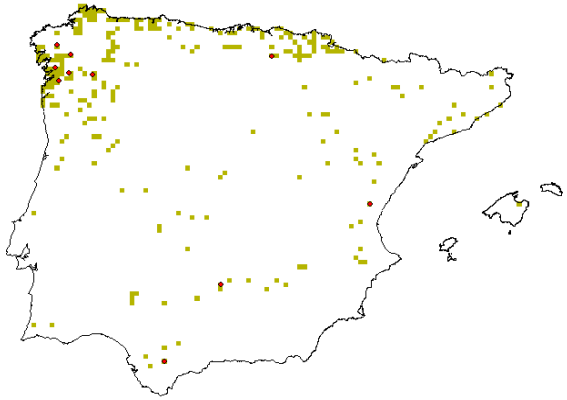
Distribución potencial (GAM)



Distribución potencial (GLM)

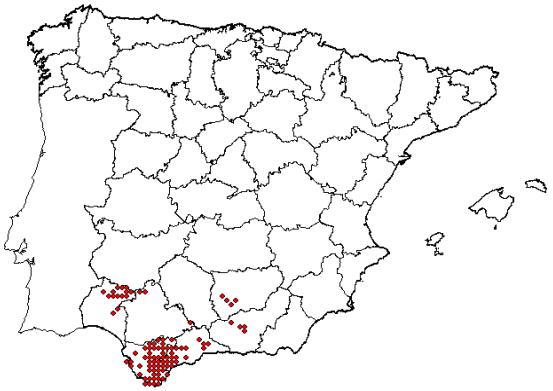


Distribución potencial (NNET)

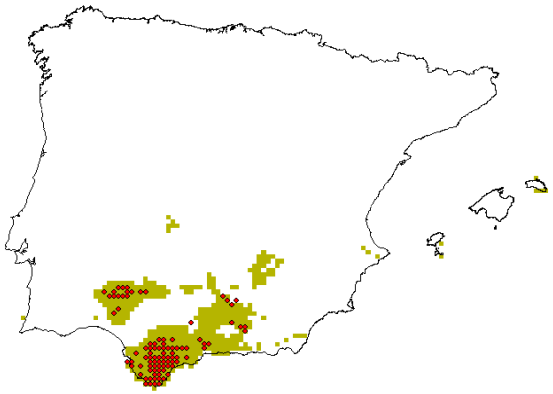


Macrothele calpeiana (Walckenaer, 1805; Araneae: Hexathelidae)

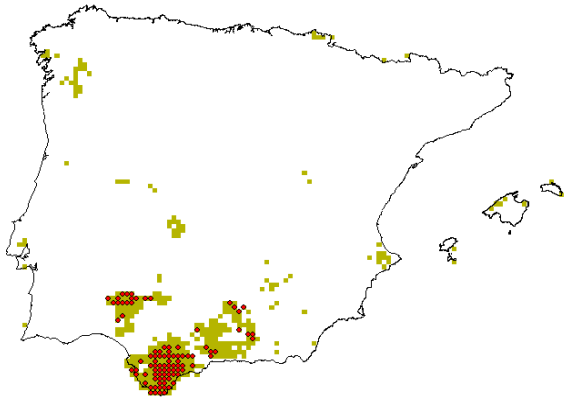
Presencias observadas (n=92)



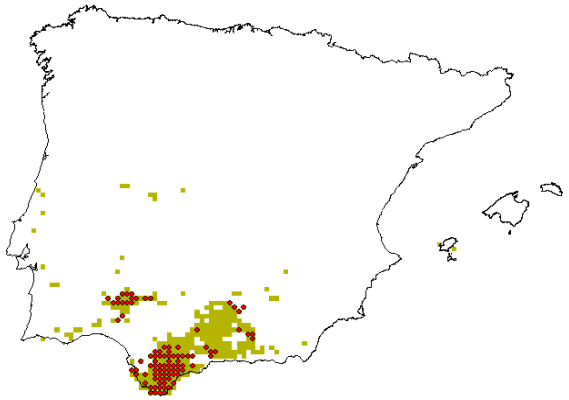
Distribución potencial (GAM)



Distribución potencial (GLM)

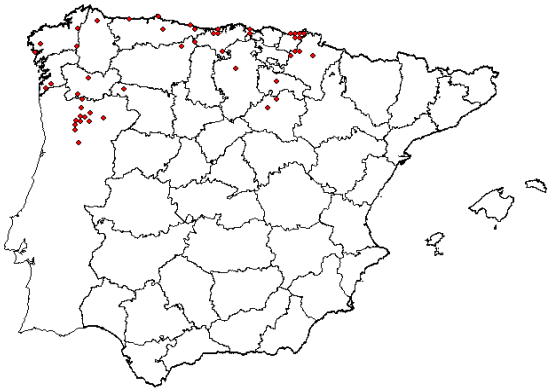


Distribución potencial (NNET)

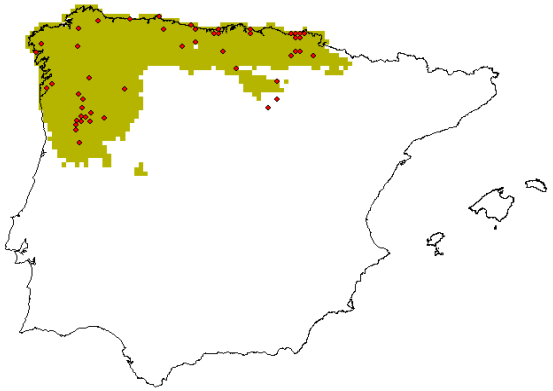


Maculinea alcon (Denis & Schiffermüller, 1775; Lepidoptera: Lycaenidae)

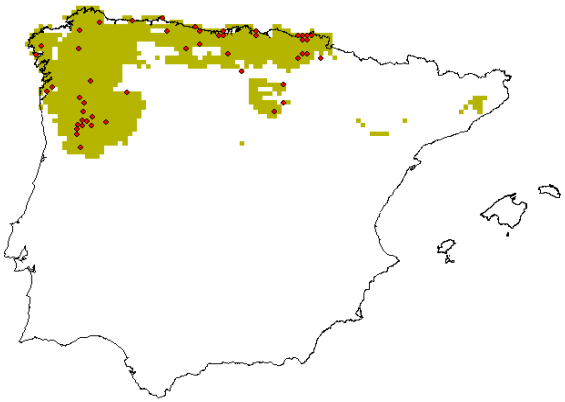
Presencias observadas (n=49)



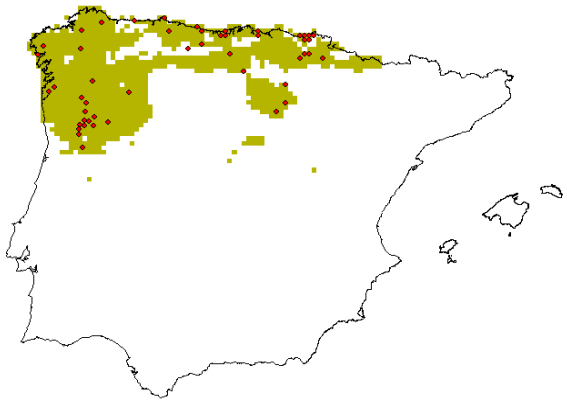
Distribución potencial (GAM)



Distribución potencial (GLM)

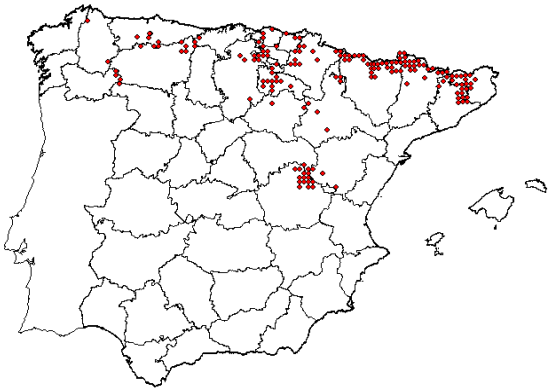


Distribución potencial (NNET)

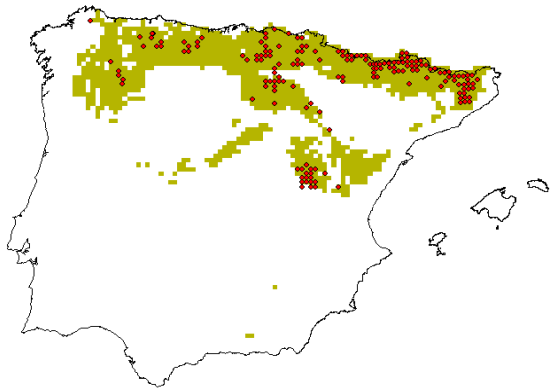


Maculinea arion (Linnaeus, 1758; Lepidoptera: Lycaenidae)

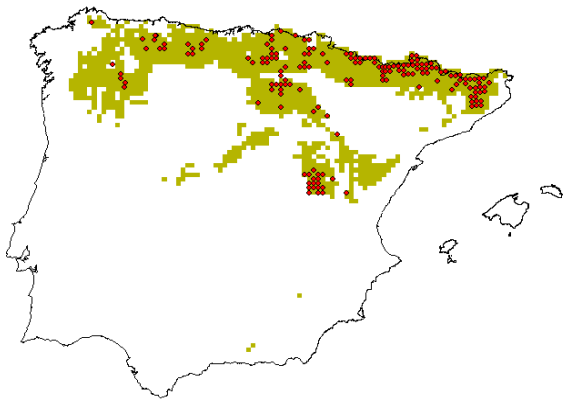
Presencias observadas (n=166)



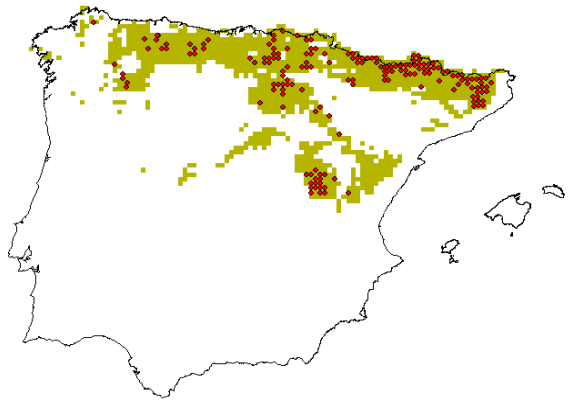
Distribución potencial (GAM)



Distribución potencial (GLM)

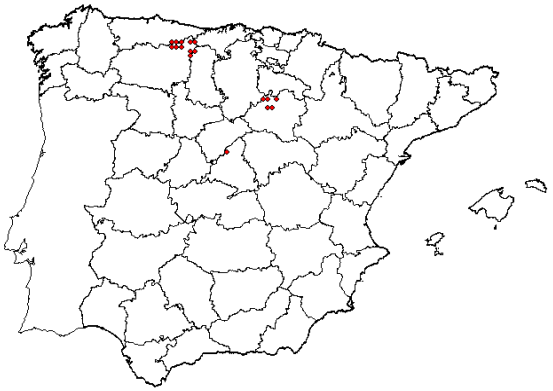


Distribución potencial (NNET)

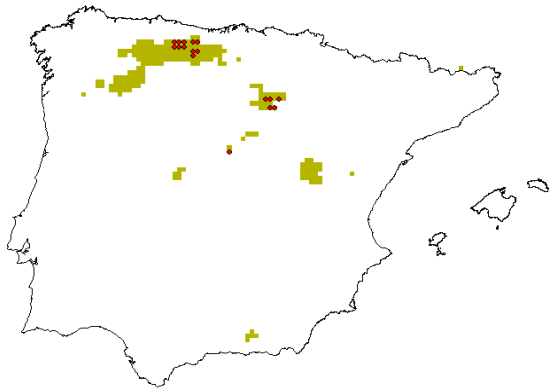


Maculinea nausithous (Bergsträsser, 1779; Lepidoptera: Lycaenidae)

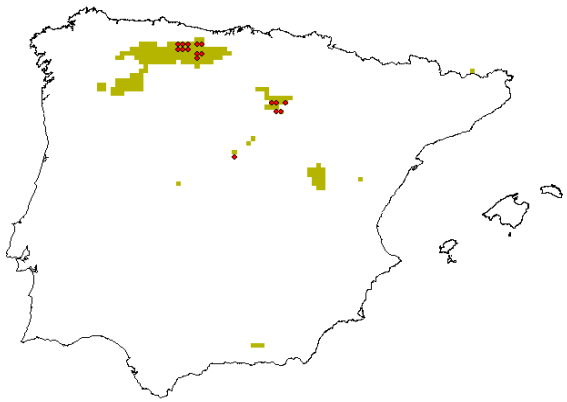
Presencias observadas (n=17)



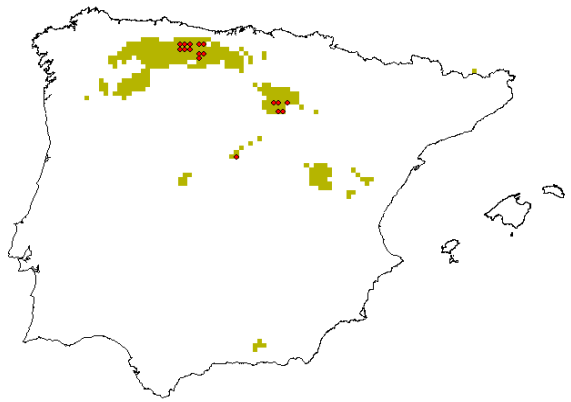
Distribución potencial (GAM)



Distribución potencial (GLM)



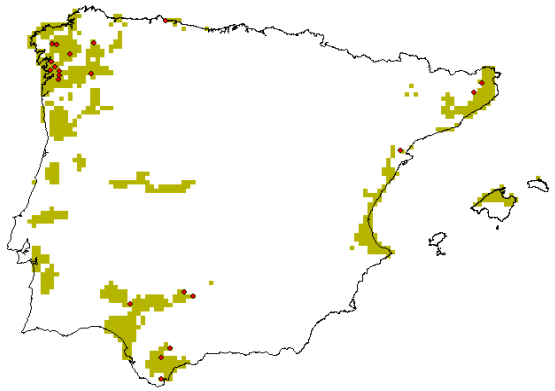
Distribución potencial (NNET)



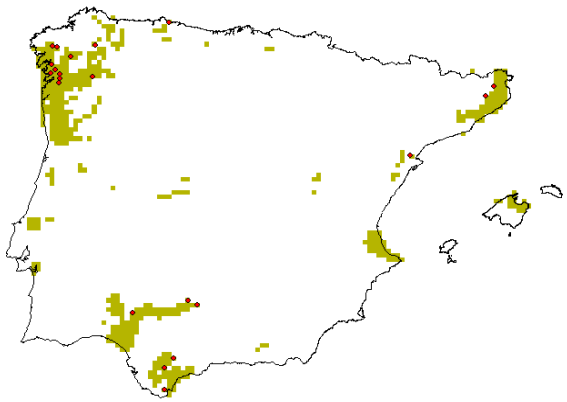
Oxygastra curtisi (Dale, 1834; Odonata: Corduliidae)

Presencias observadas (n=21)

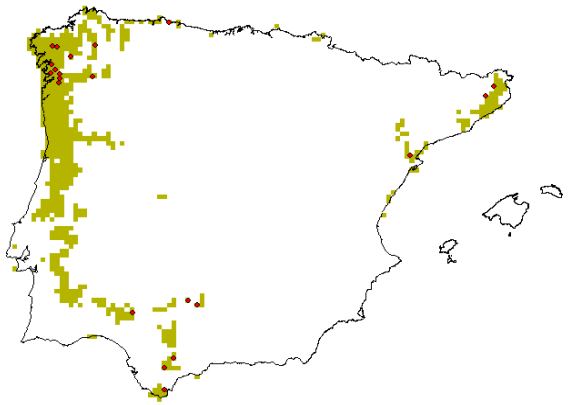
Distribución potencial (GAM)



Distribución potencial (GLM)

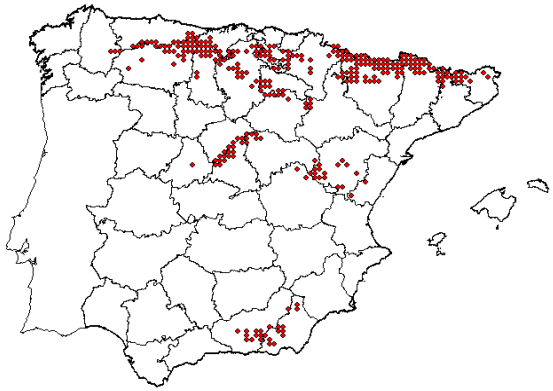


Distribución potencial (NNET)

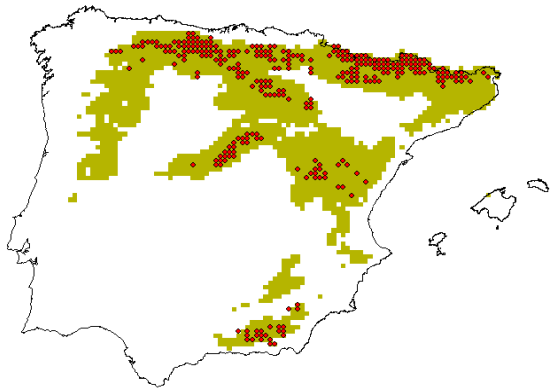


Parnassius apollo (Linnaeus, 1758; Lepidoptera: Papilionidae)

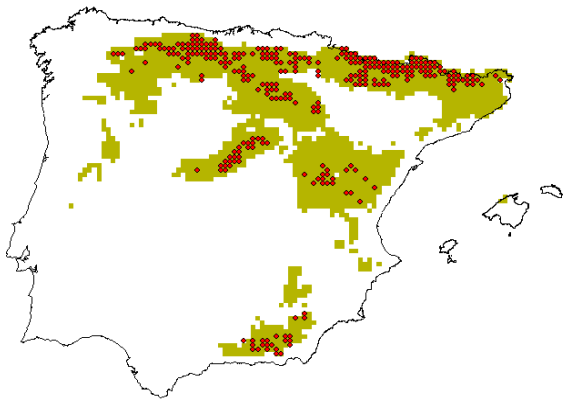
Presencias observadas (n=314)



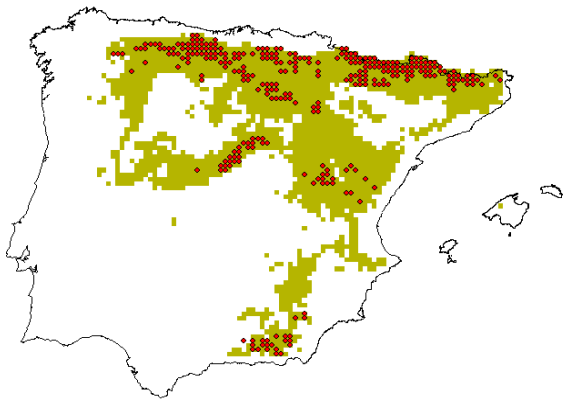
Distribución potencial (GAM)



Distribución potencial (GLM)



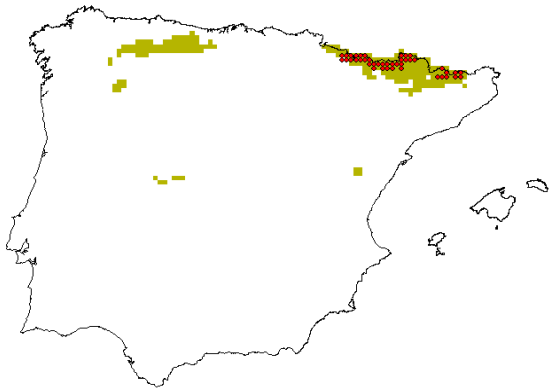
Distribución potencial (NNET)



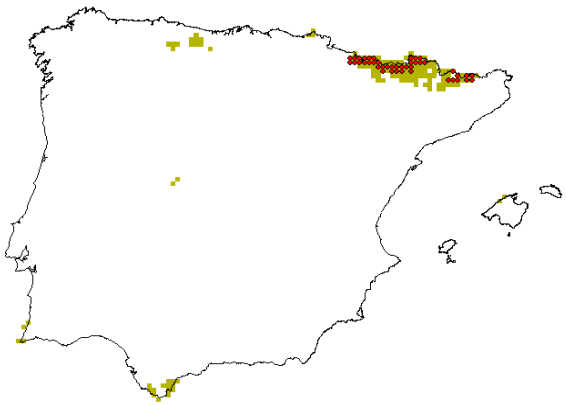
Parnassius mnemosyne (Linnaeus, 1758; Lepidoptera: Papilionidae)

Presencias observadas (n=42)

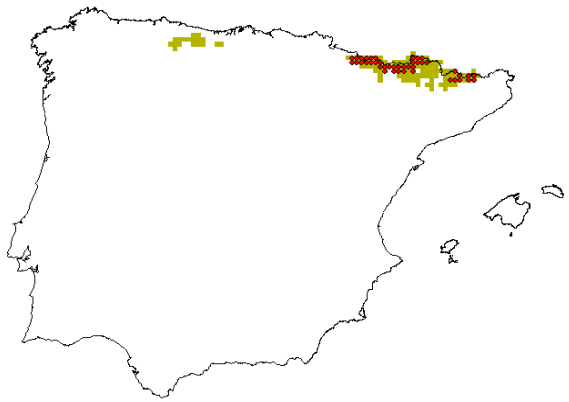
Distribución potencial (GAM)



Distribución potencial (GLM)

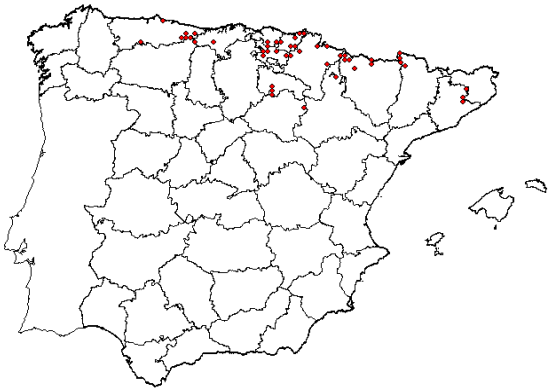


Distribución potencial (NNET)

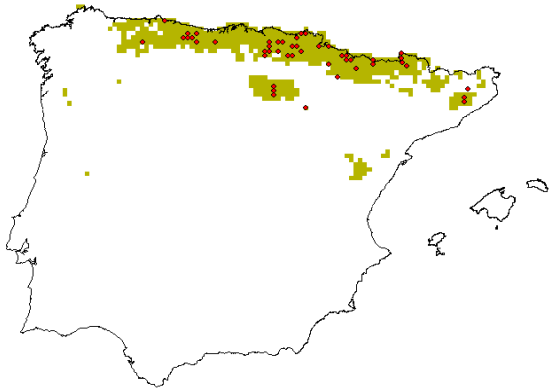


Rosalia alpina (Linnaeus, 1758; Coleoptera: Cerambycidae)

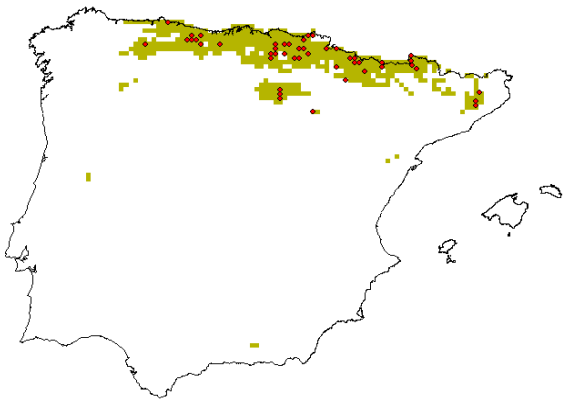
Presencias observadas (n=47)



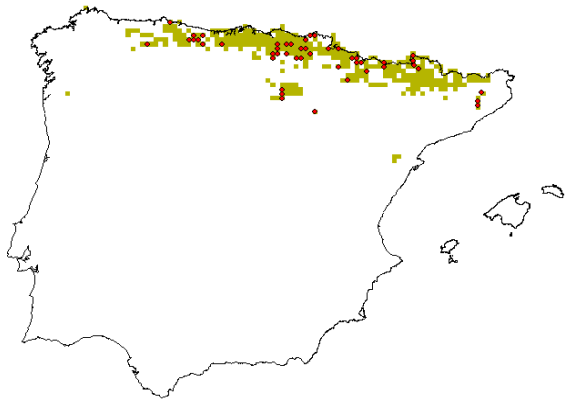
Distribución potencial (GAM)



Distribución potencial (GLM)



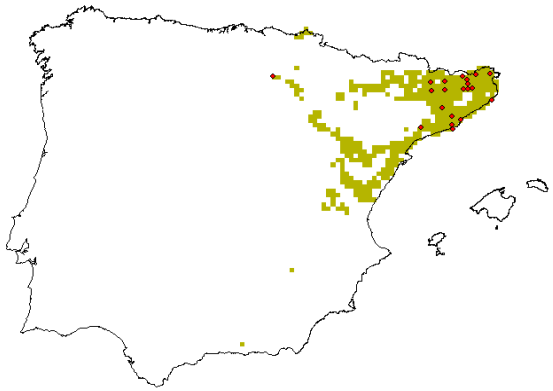
Distribución potencial (NNET)



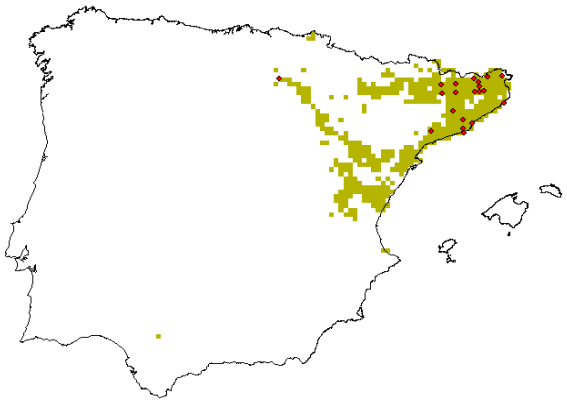
Vertigo moulinsiana (Dupuy, 1849; Pulmonata: Vertiginidae)

Presencias observadas (n=20)

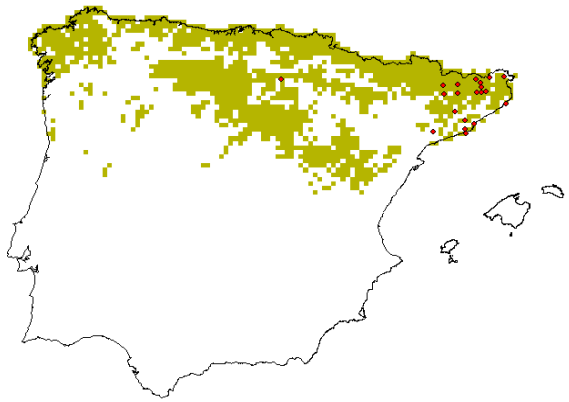
Distribución potencial (GAM)



Distribución potencial (GLM)

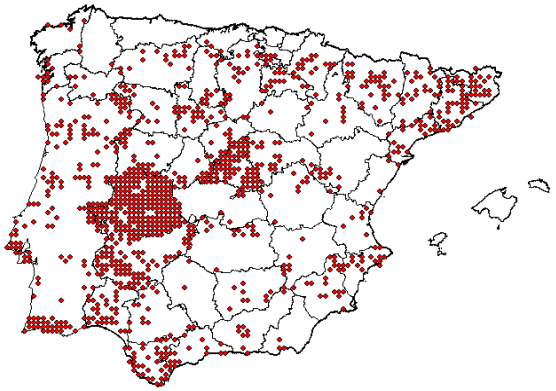


Distribución potencial (NNET)

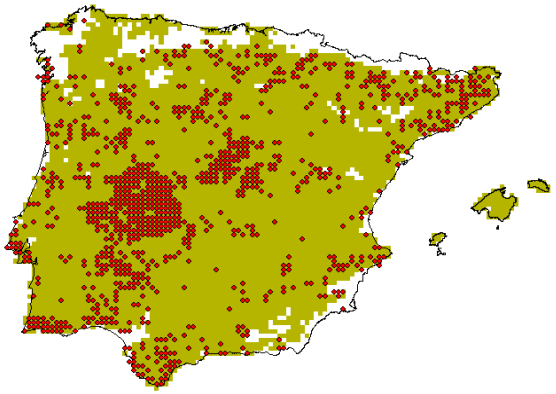


Zerynthia rumina (Linnaeus, 1758; Lepidoptera: Papilionidae)

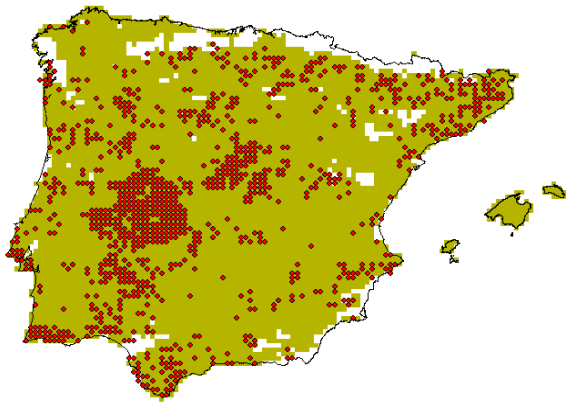
Presencias observadas (n=1107)



Distribución potencial (GAM)



Distribución potencial (GLM)



Distribución potencial (NNET)