Universidade Federal de Goiás

Instituto de Ciências Biológicas

Programa de Pós-graduação em Ecologia e Evolução

Tese de Doutorado

INCERTEZA NOS MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES

Geiziane Tessarolo

Orientador: Prof. Dr. Joaquin Hortal Muñoz

Co-orientador: Prof. Dr. Thiago Fernando Rangel

GOIÂNIA - GO

ABRIL DE 2014

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR AS TESES E DISSERTAÇÕES ELETRÔNICAS (TEDE) NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[ ] Dissertação** **[X] Tese**

2. Identificação da Tese ou Dissertação

Autor (a):	Geiziane Tessarolo				
E-mail:	geites@gmail.com				
Seu e-mail pode ser disponibilizado na página?					
Vínculo empregatício do autor					
Agência de fomento: CAPES		Coordenação de Aperfeiçoamento de Pessoal de Nível Superior		Sigla:	CAPES
País:	Brasil	UF: GO		CNPJ:	
Título:	Incerteza nos modelos de distribuição de espécies				
Palavras-chave:	Características das espécies, cobertura ambiental, desenho amostral, Incerteza, Modelos de distribuição de espécies.				
Título em outra língua:	Uncertainty in species distribution models				
Palavras-chave em outra língua:	Species traits, Environmental Completeness, Survey design, Uncertainty, Species distribution models.				
Área de concentração:	Ecologia				
Data defesa: (dd/mm/aaaa)	29/04/2014				
Programa de Pós-Graduação:	Ecologia e Evolução				
Orientador (a):	Dr. Joaquin Hortal Muñoz				
E-mail:	jhortal@mncn.csic.es				
Co-orientador	Dr. Thiago Fernando Valle Rangel				
E-mail:	rangel.tf@gmail.com				

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF ou DOC da tese ou dissertação.

O sistema da Biblioteca Digital de Teses e Dissertações garante aos autores, que os arquivos contendo eletronicamente as teses e ou dissertações, antes de sua disponibilização, receberão procedimentos de segurança, criptografia (para não permitir cópia e extração de conteúdo, permitindo apenas impressão fraca) usando o padrão do Acrobat.

_____

Data: ____ / ____ / _____

Assinatura do (a) autor (a)

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

Universidade Federal de Goiás

Instituto de Ciências Biológicas

Programa de Pós-graduação em Ecologia e Evolução

Tese de Doutorado

INCERTEZA NOS MODELOS DE DISTRIBUIÇÃO DE ESPÉCIES

Geiziane Tessarolo

Tese apresentada à Universidade Federal de Goiás, como parte das exigências do Programa de Pós-graduação em Ecologia e Evolução para obtenção do título de doutora.

Orientador: Prof. Dr. Joaquin Hortal Muñoz

Co-orientador: Prof. Dr. Thiago Fernando Rangel

GOIÂNIA - GO

ABRIL DE 2014

Dados Internacionais de Catalogação na Publicação (CIP)
GPT/BC/UFG

Tessarolo, Geiziane.
T338i Incerteza nos modelos de distribuição de espécies [manuscrito] : / Geiziane Tessarolo. - 2014.
151 f. : il.

Orientador: Prof. Dr. Joaquin Hortal Muñoz.
Tese (Doutorado) – Universidade Federal de Goiás, Instituto de Ciências Biológicas, 2014.
Bibliografia.
Inclui lista de figuras, abreviaturas, siglas e tabelas.
Apêndices.
1. Modelos de distribuição de espécies 2. Diversidade biológica 3. Cobertura ambiental I. Título.

CDU: 591.9

DEDICATÓRIA

Dedico esse trabalho aos meus pais, porque minhas atitudes e conquistas refletem as experiências vividas e não vividas, as quais eu devo em grande parte a eles.

"In a predestinate world, decision would be illusory; in a world of a perfect foreknowledge, empty, in a world without natural order, powerless. Our intuitive attitude to life implies non-illusory, non-empty, non-powerless decision. . . . Since decision in this sense excludes both perfect foresight and anarchy in nature, it must be defined as choice in face of bounded uncertainty". (George Shackle – Decision Order and Time in Human Affairs, 1961)

Agradecimentos

A conclusão de uma tese não é um mero resumo do ocorrido nos últimos quatro anos, ela reflete muito mais sobre nossas próprias expectativas, inseguranças e perseverança. Entretanto, o caminho percorrido para se aprender a lidar com elas não é ou pelos menos não deveria ser um caminho solitário. No meu caso esse caminho foi sempre acompanhado de pessoas que deram suporte e incentivo a persistir.

Agradeço ao meu orientador Joaquin Hortal Muñoz, primeiro pela coragem de aceitar uma doutoranda com só 3 anos de doutorado pela frente e sem projeto. Segundo, pela orientação, disponibilidade e esforço em discutir ideias, mesmo que por Skype às 4 da manhã. Você foi muito mais que meu orientador, você foi meu formador, e se a ciência hoje é muito mais agradável aos meus olhos, o mérito é todo seu, Joaquin!

Agradeço ao meu co-orientador Thiago Rangel, o qual muitas vezes eu fiz de orientador, pela disponibilidade, pelas discussões e por sempre ser uma fonte de ânimo, sempre saía das nossas conversas achando que eu ia conseguir fazer tudo e mais um pouco. Sem saber você acumulou funções de co (orientador) e psicólogo acadêmico.

Aos demais professores da pós por participarem direta ou indiretamente da minha formação.

Aos colegas do laboratório (115 rocks) pelo conhecimento transmitido, pelas conversas e discussões. Ao grupo de pesquisas do Professor Jorge Lobo, pelas contribuições nos trabalhos e pelas discussões que me permitiram compreender muito mais sobre ecologia e biodiversidade.

À todos meu amigos, parte fundamental nas nossas expectativas e frustações. Vocês deixaram o caminho mais ameno, cada um à sua forma. Camila, Dani e Cláudio o incentivo e apoio de vocês foram cruciais, sem ele nenhuma mudança significativa seria possível. Considerando que uma viagem é sempre mais divertida quando se está acompanhado, eu agradeço imensamente pelo prazer de pode ter dividido esse momento com Nathália Machado e Marina Zanin, porque amigos são aqueles que nos transformam em pessoas melhores. Aos meus amigos, companheiros da Espanha, Madrid não me matou, mas a saudade...

À minha família, pelo incentivo, apoio, refúgio, por me surpreenderem quando eu menos espero. Por entenderem a ausência e ainda assim tentarem me resgatar dela. Por ser a base para a formação contínua do meu caráter.

Agradeço à CAPES, pela bolsa concedida durante a execução do doutorado, à Universidade Federal de Goiás, pela bolsa concedida para a realização do Doutorado Sanduíche.

Summary

Chapter 3

General introduction

Uncertainty is in every aspect of nature, whether we look at the outcome of social processes, the strength of ecological interactions or the future changes in climate or economy. Consequently uncertainty is present in all your attempts to quantify or predict the real world (Regan et al., 2002). Although today it constitutes a fruitful field of research in many areas, in the 19th Century uncertainty was viewed as incompatible with science, been an obstacle to its progress, and as such, a feature that should be completely eliminated. This negative view neglected and discouraged the development of studies about the role of uncertainty within science for many years (Klir, 2006). However, in the beginning of the 20th Century this view started to change when the scientists facing the study of systems of great complexity accepted uncertainty as an useful and even essential component of many scientific queries (Peat, 2002). So although modern science tries to increase the precision in the descriptions of the systems it studies, we are already aware that some of these systems are so complex that it is impossible to remove all the uncertainty in their measurement (Halpern et al., 2006).

Ecological systems are highly variable and complex due to the great diversity of their components, the interactions between them and the variation of these interactions throughout different scales and the geographical space (Wu & David, 2002). These characteristics lead to the increase of uncertainty over time, which difficult the detailed knowledge of the dynamics of these systems (Mangel et al., 1996; Levin, 1999). This situation is further worsened by the fact that the information available about the functioning of these systems is often ambiguous, which makes their study, management and conservation a hard task (Harwood & Stokes, 2003; Halpern et al., 2006).

Models have been shown to be valuable tools to use the scarce information about ecological systems to understand their dynamics and make predictions about changes in their composition and functioning (Drechsler, 2004; Burgman et al., 2005). These models are

simplified representations of the key mechanisms and factors affecting ecological systems, and can include processes in different scales and hierarchical levels (Forbes et al., 2009; Schmolke et al., 2010). Although ecological models have received great attention in biodiversity conservation issues, decision makers are reluctant to use them (Brugnach et al., 2007). Possible reasons for such lack of confidence are the lack of certainty and the difficulty to validate the models (Brugnach et al., 2003; Young et al., 2014).

Data about species' presence and abundance are often necessary to address important ecological and evolutionary questions (Albert et al., 2010). Therefore, knowing the spatial distribution of species is essential for both understanding biodiversity patterns and informing conservation actions (Margules & Pressey, 2000; Jetz et al., 2012). However, the attempts to study these patterns stumble on the Wallacean shortfall, i.e. the widespread lack of data about the species distribution (Lomolino, 2004; Whittaker et al., 2005). This gap in knowledge occurs because obtaining good quality distributional data with usually requires large amounts of field work and financial resources. Given the common limitations of manpower and funds, distributional data are deficient or unavailable for most species (Hortal et al., 2008; Rocchini et al., 2011).

Species distribution models (SDM) arrived as a solution to this lack of distributional data (Guisan & Zimmermann, 2000). These models use statistical algorithms to relate variables that describe environmental factors with the occurrence/absence of the species in a series of sites, in order to define a multidimensional space of tolerance or requirements of species. By this correlation, SDM are used to make predictions about the distribution of the species (Araújo & Guisan, 2006; Rondinini et al., 2006). SDM results have therefore been used in many applications, like identifying areas of priority for conservation (Rodríguez-Soto et al., 2011; Selig et al., 2014), mapping rare and endangered species (Aitken et al., 2007; Marino et al., 2011), predicting the spread of invasive species (Steiner et al., 2008; Sheppard

et al., 2014), forecasting the impacts of climate change on species distribution (Thuiller et al., 2005; Huntley et al., 2008; Lemes et al., 2013), studying biodiversity patterns (Graham et al., 2006; Denslow et al., 2010), improving sampling (Guisan et al., 2006; Le Lay et al., 2010), testing biogegraphical hypotheses (Norris et al., 2006; Richards et al., 2007), or evaluating the extinction risk of species (Machado & Loyola, 2013), to name a few of their uses.

The increasing use of SDM in the recent years has been made possible mainly by the greater availability of large databases of species' occurrences (Graham et al., 2004), such as GBIF (Global Biodiversity Information Facility; http://www.gbif.org/) and SpeciesLink (http://splink.cria.org.br/), more reliable databases on environmental data and freely-available methods use supporting different types of data (Franklin & Miller, 2009). Due to the apparent wide availability of data, many researchers have focused in developing new modeling techniques with the main objective of generating more accurate predictions (Drake et al., 2006; Phillips et al., 2006; Prasad et al., 2006; Elith et al., 2008; Olden et al., 2008). Currently there is a large number of techniques available to model species' distributions, differring in complexity, assumptions and data requirements (presence only, presence\absence, abundance, etc; Syphard & Franklin, 2009). Many works have evaluated the differences in performance across SDM techniques (Segurado & Araújo, 2004; Tsoar et al., 2007; Syphard & Franklin, 2009). However, the results of these studies are conflicting (Elith et al., 2006; Peterson et al., 2007; Lobo, 2008; Hijmans, 2012). This occurs because, in practice, SDM resuls are influenced not only by the algorithms used, but also by many other factors such as scale (Peterson et al., 2007), prevalence in the training data (Meynard & Quinn, 2007), species' ecology (Syphard & Franklin, 2010), uncertainty of locality (Fernandez et al., 2009), sample size (Hernandez et al., 2006; Wisz et al., 2008; Feeley & Silman, 2011), ecological and geographical marginality (Chefaoui et al., 2011), quality of input data (Hortal et al., 2007, 2008) or the influence of historical and biogeographical processes (Hortal et al., 2012). The

uncertainty generated by all these factors diminishes the reliability of the SDM, making their use questionable for many proposals (Pearson & Dawson, 2003; Thuiller et al., 2003; Meynard & Quinn, 2007; Hortal et al., 2012).

The uncertainty associated to SDM can generate two undesired scenarios. At best, they may lose importance as useful tools for conservation assessment, leading to a decreasing use of these techniques. At worst, SDM may continue to be used without any concern about the reliability of their results due to the lack of alternative methods. Both options are worrying, since these techniques could actually provide quality distributional hypotheses. While the lack of distributional data is a major pitfall for many ecological and conservation applications, the indiscriminate use of SDMs yields results that may often be inappropriate for the purposes to which they are applied (Guisan et al., 2006). Uncertainty in SDM predictions comes from many different sources, that can be grouped in three main areas: data, model and prediction (Beale & Lennon, 2012). Uncertainty in the data refers to the variability arising from the deficiency in the knowledge about the patterns and processes studied (i.e. small or biased samples, quality and choice of covariates; (Rocchini et al., 2011; Beale & Lennon, 2012; Hortal et al., 2012). Model uncertainty arises from the differences in the assumptions, algorithms and parameterization across techniques, which results in variations of their ability to relate species distributions to climatic predictors (Segurado & Araújo, 2004; Heikkinen et al., 2006). Finally, prediction uncertainty comes from errors in the estimation of model parameters, that appear due to demographic stochasticity, fluctuations in ecological processes, randomness of dispersal and so on. These variations in model parameters impede determining whether any prediction provides a perfect fit to the species' geographic response; and even if it were, if it may not be so at a different time and/or place (Hastie et al., 2001; Dormann et al., 2008).

A variety of possible sources of uncertainty have been studied within each one of the abovementioned categories, including: quantity of data (Stockwell & Peterson, 2002; Lobo & Tognelli, 2011), positional error of data (Fernandez et al., 2009; Osborne & Leitão, 2009), bias in input data (Bean et al., 2012; Varela et al., 2014), choice of covariates (Synes & Osborne, 2011), spatial autocorrelation in covariates and results (Naimi et al., 2011; Crase et al., 2012; Record et al., 2013), modelling technique (Elith et al., 2006; Meynard & Quinn, 2007), species traits (Pöyry et al., 2008; Syphard & Franklin, 2010; Chefaoui et al., 2011), scale (Seo et al., 2009), or model success assessment metric (Hanberry et al., 2012), among others.

The limitations of distributional models caused by these sources of uncertainty are known since the beginning of the use of SDM (Booth et al., 2014). However, only recently researchers have been worried about the necessity of taking into account these uncertainties when analyzing SDM results (Barry & Elith, 2006). This has led to a seek for methodologies that allow assessing and managing the uncertainty associated to SDM, under the expectative that they can allow diminishing such uncertainty, at least for several sources, through, e.g., deciding whether it is necessary to collect more data and where; or by improving the estimation of model parameters. Nonetheless, even the intractable sources of uncertainty can be accounted for when known, helping to develop improved assessments of SDM precision (Beale & Lennon, 2012). More specifically, some studies point to the fact that the uncertainties related to SDM results can be clustered in space, presenting spatial autocorrelation (Hortal & Lobo, 2006; Hanspach et al., 2011). Therefore, there is a need for methodologies that map SDM uncertainty in a spatially explicit way (Guisan & Zimmermann, 2000; Elith & Leathwick, 2009). These methods will in turn allow detecting regions of higher uncertainty, as well as identifying the covariates important for species distributions that can relate to it (Lobo et al., 2007; Rocchini et al., 2011).

Thereby, identifying and quantifying the uncertainties associated with model predictions is now a key aspect for the improvement of SDM (Beale & Lennon, 2012). This will have great impact on conservation assessments, for it allows measuring the risk involved in each decision, in turn increasing the confidence of decision makers when using SDM results. Additionally, with an increased emphasis on uncertainty issues, researches may be stimulated to develop tools that allow the evaluation of prediction or even correct the results by uncertainty, improving the SDM predictions (Tessarolo, 2012).

Main objective

This thesis seeks to increase the knowledge about the uncertainties associated to Species Distribution Models, with the aim of generating useful information for the better understanding of the limitations of this research field, and ultimately helping to its progress. Following this main objective, we assess several of the main sources of SDM uncertainty, determining their relative importance and their impacts on SDM results. Further, we develop a new methodology to quantify uncertainty from the biogeographical data used on SDM.

General organization of the thesis

This thesis is composed of three chapters. In the first one we examine how different survey designs used to gather distributional data could affect SDM results, assessing also their relative importance when compared with other factors. In this chapter we seek to evaluate which survey designs lead to smaller impacts of data quality on SDM predictions. Also, we evaluate the magnitude of influence of the main sources of uncertainty in SDM: species, sample size, modelling technique and survey design. Further, we assessed the climatic bias associated with each survey design and the impact of prevalence and Relative Occurrence

Area (ROA) of species on SDM. To do this, we modeled the distribution of 34 Iberian endemic species of terrestrial vertebrates based on data of extremely good quality, using eight modelling techniques to calibrate SDMs from data simulating seven survey designs combined with five sample sizes. Survey design was the factor with less influence in SDM prediction, while the species being modeled was the most important. The bias in the environmental coverage of samples has greater impact on SDM results than its spatial structure.

The second chapter presents an evaluation of the influence of several species' attributes on SDM performance, including traits and geographical characteristics, as well as the coverage of the environmental response provided by the data. We used data from 57 Scarabaeidae dung beetle species from Madrid (Central Spain) to generate distribution maps. The influence of the nine species' attributes and environmental niche completeness was assessed with six different evaluation metrics; their importance for SDM results was evaluated with a Partial Least Square (PLS) analysis. Marginality and niche completeness were the most influential variables, as the more marginal the species and the higher the environmental completeness, the better are the models generated. When modelling various species is necessary to take into account species for which models can be poor due their characteristics. SDM from these species can be removed or downweighted, and/or improved with the collection of data to increase the coverage of their environmental niche.

Finally, in the third chapter all the information obtained in the previous chapters is used to develop a methodology to quantify the uncertainty in SDM arising from the distributional input data. Specifically we aim to quantify the uncertainty from various factors affecting the biogeographical knowledge and develop methods to summarize all these uncertainties in a way that can be displayed in a spatially-explicit fashion. For this purpose we create maps of biogeographical ignorance based on three main sources of ignorance: quality of surveys, spatial and temporal distance from survey points. Maps of ignorance were

generated for 57 Scarabaeidae species from the whole of the Iberian Peninsula and related with SDM generated using six different modelling techniques. Maps of Biogeographical ignorance allowed to evaluate the reliability of SDM predictions within the study area. We therefore suggest the SDM community to use these maps as additional information to diminish the subjectivity when interpreting the SDM results. The spatially-explicit visualization of uncertainties can improve the communication among scientists, as well as between scientists and decision markers.

References

Aitken M., Roberts D.W., & Shultz L.M. (2007) Modeling distributions of rare plants in the great basin, Western North America. *Western North American Naturalist*, **67**, 26–38.

Albert C.H., Yoccoz N.G., Edwards T.C., Graham C.H., Zimmermann N.E., & Thuiller W. (2010) Sampling in ecology and evolution - bridging the gap between theory and practice. *Ecography*, **33**, 1028–1037.

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

Barry S. & Elith J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Beale C.M. & Lennon J.J. (2012) Incorporating uncertainty in predictive species distribution modelling. *Philosophical transactions of the Royal Society of London*, **367**, 247–58.

Bean W.T., Stafford R., & Brashares J.S. (2012) The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, **35**, 250–258.

Booth T.H., Nix H. a., Busby J.R., & Hutchinson M.F. (2014) bioclim : the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, **20**, 1–9.

Brugnach M., Bolte J., & Bradshaw G.A. (2003) Determining the significance of threshold values uncertainty in rule-based classification models. *Ecological Modelling*, **160**, 63–76.

Brugnach M., Tagg A., Keil F., & J L.W. (2007) Uncertainty Matters : Computer Models at the Science – Policy Interface. *Water resour manage*, **21**, 1075–1090.

Burgman M.A., Lindenmayer D.B., & Elith J. (2005) Managing landscapes for conservation under uncertainty. *Ecology*, **86**, 2007–2017.

Chefaoui R.M., Lobo J.M., & Hortal J. (2011) Effects of species' traits and data characteristics on distribution models of threatened invertebrates. *Animal Biodiversity and Conservation*, **34**, 229–247.

Crase B., Liedloff A.C., & Wintle B. a. (2012) A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, **35**, 879–888.

Denslow M.W., Palmer M.W., & Murrell Z.E. (2010) Patterns of native and exotic vascular plant richness along an elevational gradient from sea level to the summit of the Appalachian Mountains, U.S.A. *The Journal of the Torrey Botanical Society*, **137**, 67–80.

Dormann C.F., Purschke O., García Márquez J.R., Lautenbach S., & Schröder B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–86.

Drake J.M., Randin C., & Guisan A. (2006) Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, **43**, 424–432.

Drechsler M. (2004) Model-based conservation decision aiding in the presence of goal conflicts and uncertainty. *Biodiversity and Conservation*, **13**, 141–164.

Elith J., H. Graham C., P. Anderson R., Dudík M., Ferrier S., Guisan A., J. Hijmans R., Huettmann F., R. Leathwick J., Lehmann A., Li J., G. Lohmann L., A. Loiselle B., Manion G., Moritz C., Nakamura M., Nakazawa Y., McC. M. Overton J., Townsend Peterson a., J. Phillips S., Richardson K., Scachetti-Pereira R., E. Schapire R., Soberón J., Williams S., S. Wisz M., & E. Zimmermann N. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith J. & Leathwick J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith J., Leathwick J.R., & Hastie T. (2008) A working guide to boosted regression trees. *The Journal of animal ecology*, **77**, 802–13.

Feeley K.J. & Silman M.R. (2011) Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, **17**, 1132–1140.

Fernandez M.A., Blum S.D., Reichle S., Holzman B., & Hamilton H. (2009) Locality uncertainty and the differential performance of four common niche-based modeling techniques. *Biodiversity Informatics*, **6**, 36–52.

Forbes V.E., Hommen U., Thorbek P., Heimbach F., Brink P.J. Van Den, Wogram J., Thulke H., Grimm V., Hommen À.U., & Thulke À.À.H. (2009) Ecological Models in Support of Regulatory Risk Assessments of Pesticides : Developing a Strategy for the Future Ecological Models in Support of Regulatory Risk Assessments of Pesticides :

Developing a Strategy for the Future. *Integrated Environmental Assessment and Management*, **5**, 167–172.

Franklin J. & Miller J.A. (2009) *Mapping species distributions: spatial inference and prediction.* Cambridge University Press Cambridge, UK, New York, New York, USA.

Graham C.H., Ferrier S., Huettman F., Moritz C., & Peterson A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in ecology & evolution*, **19**, 497–503.

Graham C.H., Moritz C., & Williams S.E. (2006) Habitat history improves prediction of biodiversity in rainforest fauna. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 632–6.

Guisan A., Lehmann A., Ferrier S., Austin M., Overton J.M.C., Aspinall R., & Hastie T. (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, **43**, 386–392.

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

Halpern B.S., Regan H.M., Possingham H.P., & McCarthy M.A. (2006) Accounting for uncertainty in marine reserve design. *Ecology letters*, **9**, 2–11; discussion 11–4.

Hanberry B.B., He H.S., & Dey D.C. (2012) Sample sizes and model comparison metrics for species distribution models. *Ecological Modelling*, **227**, 29–33.

Hanspach J., Kühn I., Schweiger O., Pompe S., & Klotz S. (2011) Geographical patterns in prediction errors of species distribution models. *Global Ecology and Biogeography*, **20**, 779–788.

Harwood J. & Stokes K. (2003) Coping with uncertainty in ecological advice: lessons from fisheries. *Trends in Ecology & Evolution*, **18**, 617–622.

Hastie T., Tibshirani R., & Friedman J.H. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer-Verlag, New York.

Heikkinen R.K., Luoto M., Araujo M.B., Virkkala R., Thuiller W., & Sykes M.T. (2006) Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography*, **30**, 751–777.

Hernandez P.A., Graham C.H., Master L.L., & Albert D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

Hijmans R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, **93**, 679–88.

Hortal J., Jimenez-Valverde A., Gómez J.F., Lobo J.M., & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Hortal J. & Lobo J.M. (2006) Towards a synecological framework for systematic conservation planning. *Biodiversity Informatics*, **3**, 16–45.

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 : the journal of the Society for Conservation Biology*, **21**, 853–63.

Hortal J., Lobo J.M., & Jiménez-Valverde A. (2012) Basic questions in biogeography and the (lack of) simplicity of species distributions: Putting species distribution models in the right place. *Natureza e Conservação*, **10**, 108–118.

Huntley B., Collingham Y.C., Willis S.G., & Green R.E. (2008) Potential impacts of climatic change on European breeding birds. *PloS one*, **3**, e1439.

Jetz W., McPherson J.M., & Guralnick R.P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in ecology & evolution*, **27**, 151–9.

Klir G.J. (2006) *Uncertainty and information : foundations of generalized information theory.* John Wiley & Sons, New Jersey.

Le Lay G., Engler R., Franc E., & Guisan A. (2010) Prospective sampling based on model ensembles improves the detection of rare species. *Ecography*, **33**, 1015–1027.

Lemes P., Melo A.S., & Loyola R.D. (2013) Climate change threatens protected areas of the Atlantic Forest. *Biodiversity and Conservation*, **23**, 357–368.

Levin S.A. (1999) *Fragile dominion: complexity and the commons.* Perseus Books,

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. (2007) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Lobo J.M. & Tognelli M.F. (2011) Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, **19**, 1–7.

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), Sinauer Associates, Inc., Sunderland, Massachussets.* pp. 293–296.

Machado N. & Loyola R.D. (2013) A comprehensive quantitative assessment of bird extinction risk in Brazil. *PloS one*, **8**, e72283.

Mangel M., Talbot L.M., Meffe G.K., Agardy M.T., Alverson D.L., Barlow J., Botkin D.B., Budowski G., Clark T., Cooke J., & others (1996) Principles for the conservation of wild living resources. *Ecological Applications*, **6**, 338–362.

Margules C.R. & Pressey R.L. (2000) Systematic conservation planning. *Nature*, **405**, 243–53.

Marino J., Bennett M., Cossios D., Iriarte A., Lucherini M., Pliscoff P., Sillero-Zubiri C., Villalba L., & Walker S. (2011) Bioclimatic constraints to Andean cat distribution: a modelling application for rare species. *Diversity and distributions*, **17**, 311–322.

Meynard C.N. & Quinn J.F. (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, **34**, 1455–1469.

Naimi B., Skidmore A.K., Groen T. a., & Hamm N. a. S. (2011) Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, **38**, 1497–1509.

Norris J.R., Jackson S.T., & Betancourt J.L. (2006) Classification tree and minimum-volume ellipsoid analyses of the distribution of ponderosa pine in the western USA. *Journal of Biogeography*, **33**, 342–360.

Olden J.D., Lawler J.J., & Poff N.L. (2008) Machine learning methods without tears: a primer for ecologists. *The Quarterly review of biology*, **83**, 171–93.

Osborne P.E. & Leitão P.J. (2009) Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions*, **15**, 671–681.

Pearson R.G. & Dawson T.P. (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361–371.

Peat F.D. (2002) *From certainty to uncertainty: the story of science and ideas in the twentieth century.* Joseph Henry Press, Washington, D.C.

Peterson A.T., Papeş M., & Eaton M. (2007) Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography*, **30**, 550–560.

Phillips S., Anderson R., & Schapire R. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

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.

Prasad A.M., Iverson L.R., & Liaw A. (2006) Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, **9**, 181–199.

Record S., Fitzpatrick M.C., Finley A.O., Veloz S., & Ellison A.M. (2013) Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change. *Global Ecology and Biogeography*, **22**, 760–771.

Regan H.M., Colyvan M., & Burgman M.A. (2002) A Taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, **12**, 618–628.

Richards C.L., Carstens B.C., & Lacey Knowles L. (2007) Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography*, **34**, 1833–1845.

Rocchini D., Hortal J., Lengyel S., Lobo J.M., Jimenez-Valverde A., Ricotta C., Bacaro G., & Chiarucci A. (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Rodríguez-Soto C., Monroy-Vilchis O., Maiorano L., Boitani L., Faller J.C., Briones M.Á., Núñez R., Rosas-Rosas O., Ceballos G., & Falcucci A. (2011) Predicting potential distribution of the jaguar (Panthera onca) in Mexico: identification of priority areas for conservation. *Diversity and Distributions*, **17**, 350–361.

Rondinini C., Wilson K. a, Boitani L., Grantham H., & Possingham H.P. (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology letters*, **9**, 1136–45.

Schmolke A., Thorbek P., DeAngelis D.L., & Grimm V. (2010) Ecological models supporting environmental decision making: a strategy for the future. *Trends in ecology & evolution*, **25**, 479–86.

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

Selig E.R., Turner W.R., Troëng S., Wallace B.P., Halpern B.S., Kaschner K., Lascelles B.G., Carpenter K.E., & Mittermeier R. a (2014) Global priorities for marine biodiversity conservation. *PloS one*, **9**, e82898.

Seo C., Thorne J.H., Hannah L., & Thuiller W. (2009) Scale effects in species distribution models: Implications for conservation planning under climate change. *Biology letters*, 39–43.

Sheppard C.S., Burns B.R., & Stanley M.C. (2014) Predicting plant invasions under climate change: are species distribution models validated by field trials? *Global Change Biology*, .

Steiner F.M., Schlick-Steiner B.C., VanDerWal J., Reuther K.D., Christian E., Stauffer C., Suarez A. V., Williams S.E., & Crozier R.H. (2008) Combined modelling of distribution and niche in invasion biology: a case study of two invasive Tetramorium ant species. *Diversity and Distributions*, **14**, 538–545.

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

Synes N.W. & Osborne P.E. (2011) Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. *Global Ecology and Biogeography*, **20**, 904–914.

Syphard A.D. & Franklin J. (2009) Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, **32**, 907–918.

Syphard A.D. & Franklin J. (2010) Species traits affect the performance of species distribution models for plants in southern California. *Journal of Vegetation Science*, **21**, 177–189.

Tessarolo G. (2012) Choosing the right path for species distribution modeling. *Frontiers of Biogeography*, **4**, 91–92.

Thuiller W., Araújo M.B., & Lavorel S. (2003) Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14**, 669–680.

Thuiller W., Lavorel S., Araújo M.B., Sykes M.T., & Prentice I.C. (2005) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 8245–8250.

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

Varela S., Anderson R.P., García-Valdés R., & Fernández-González F. (2014) Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, **37**, 1–8.

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.

Wisz M.S., Hijmans R.J., Li J., Peterson A.T., Graham C.H., & Guisan A. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Wu J. & David J.L. (2002) A spatially explicit hierarchical approach to modeling complex ecological systems: theory and applications. *Ecological Modelling*, **153**, 7–26.

Young J.C., Waylen K. a., Sarkki S., Albon S., Bainbridge I., Balian E., Davidson J., Edwards D., Fairley R., Margerison C., McCracken D., Owen R., Quine C.P., Stewart-Roper C., Thompson D., Tinch R., Hove S., & Watt A. (2014) Improving the science-policy dialogue to meet the challenges of biodiversity conservation: having conversations rather than talking at one-another. *Biodiversity and Conservation*, **23**, 387–404.

Chapter 1

Uncertainty associated with survey design in Species Distribution Models

Uncertainty associated with survey design in Species Distribution Models

Geiziane Tessarolo[1,2] , Thiago Fernando Rangel[1], Miguel B. Araújo[2,3] & Joaquín Hortal[1,2]

[1]Departamento de Ecologia, Instituto de Ciências Biológicas – ICB, Universidade Federal de Goiás – UFG Campus II, 74001-970, Goiânia, GO, Brazil

[2]Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales (CSIC), C/José Gutiérrez Abascal 2, 28006 Madrid, Spain

[3]Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY, Berks, United Kingdom

*Author for correspondence. e-mail: geites@gmail.com

Running title: Uncertainty in species distribution models

Word count: 5.658

ABSTRACT

Aim Species Distribution Models (SDM) can be used to predict the location of unknown populations from known species occurrences. It follows that how the data used to calibrate the models are collected can have a great impact on prediction success. We evaluated the influence of different survey designs and their interaction with the modelling technique on SDM performance.

Location Iberian Peninsula

Methods We examine how data recorded using seven alternative survey designs (random, systematic, environmentally stratified by class and environmentally stratified using *p*-median, biased due to accessibility, biased by human density aggregation and biased towards protected areas) could affect SDM predictions generated with nine modelling techniques (BIOCLIM, Gower distance, Mahalanobis distance, Euclidean distance, GLM, MaxEnt, ENFA and Random Forest). We also study how sample size, species' characteristics and modelling technique affected SDM predictive ability, using six evaluation metrics.

Results Survey design has a small effect on prediction success. Characteristics of species' ranges rank highest among the factors affecting SDM results: the species with lower relative occurrence area (ROA) are predicted better. Model predictions are also improved when sample size is large.

Main conclusions The species modelled – particularly the extent of its distribution – are the largest source of influence over SDM results. The environmental coverage of the surveys is more important than the spatial structure of the calibration data. Therefore, climatic biases in the data should be identified to avoid erroneous conclusions about the geographic patterns of species distributions.

Keywords: climatic bias, environmental niche models, Iberian Peninsula, model performance, relative occurrence area, sample size.

INTRODUCTION

Species Distribution Models (SDM) relate incomplete information about the occurrence of species (and sometimes about their absence) with environmental predictors, to generate spatially explicit predictions about their geographic distributions (Araújo & Guisan, 2006; Franklin & Miller, 2009; Peterson et al., 2011; Hortal et al., 2012). Despite their increasing use during the last 15 years (Thuiller et al., 2009; Lobo et al., 2010; Varela et al., 2014), SDMs pose many conceptual problems (Araújo & Guisan, 2006; Soberón, 2007, 2010; Jiménez-Valverde et al., 2008; Colwell & Rangel, 2009; Soberón & Nakamura, 2009; Hortal et al., 2012) and encompass a number of methodological uncertainties (Barry & Elith, 2006; Heikkinen et al., 2006; Rocchini et al., 2011; Beale & Lennon, 2012).

One of the assumptions of SDMs is that the data used for model calibration (i.e., the samples of presence, or presence and absence) are free of bias. When this assumption is violated by biases in the collection of data, the accuracy of model predictions can be affected (Hortal et al., 2008; Lobo, 2008; Loiselle et al., 2008; Rocchini et al., 2011). In fact, the design of the surveys can have a large impact on SDM performance (Hirzel & Guisan, 2002; Edwards Jr. et al., 2006; Albert et al., 2010; Braunisch & Suchant, 2010). It follows that an adequate spatial design of the surveys increases the value of biodiversity data collections to answer different questions in ecology, evolution and biogeography (Hortal & Lobo, 2005; Albert et al., 2010). Many methods to design surveys have been proposed with the objective of maximizing the amount of biodiversity captured, while incorporating time and cost limitations (Austin & Heyligers, 1989; Pereira & Itami, 1991; Hirzel & Guisan, 2002; Funk et

al., 2005; Hortal & Lobo, 2005; Medina et al., 2013). However, even planned sample designs can vary in the efficiency with which they detect biodiversity patterns (Sastre & Lobo, 2009).

Perhaps more importantly, good quality data coming from standardized surveys are rare or even lacking for most regions and species (Rocchini et al., 2011). Rather, biodiversity databases include heterogeneous information coming from inventories developed with a variety of objectives (Hortal et al., 2007). The absence of standardized sampling schemes often generates bias in the resulting distributional data. It is well known that the records of occurrence and/or absence of species are more frequent in more accessible locations (i.e., near major road routes, urban areas or the work centres of the taxonomists) and/or classical localities (e.g., national parks or other protected areas) that are repeatedly sampled over time (Dennis & Thomas, 2000; Kadmon et al., 2004; Romo et al., 2006; Hortal et al., 2007). These geographical biases in the survey effort may often result in historical and climatic biases (Hortal et al., 2007, 2008; Lobo et al., 2007a). Such biases can yield incomplete and potentially truncated characterizations of species realized niches (Hortal et al., 2008; Rocchini et al., 2011), with the consequent critical effect on the ability of SDM to describe environmental limits of species' distributions (Austin & Heyligers, 1989; Thuiller et al., 2004; Albert et al., 2010; Hortal et al., 2012).

Despite the potential effects of spatial biases in the design of the surveys on SDM prediction ability (Araújo & Guisan, 2006; Phillips et al., 2009) , this issue has not been addressed in the literature . Here we try to overcome this gap by systematically evaluating the influence of different survey design strategies on the performance of SDM predictions, as well as the potential interactions between survey designs and SDM techniques. Further, besides survey design we evaluate the magnitude of influence on predictive accuracy of other factors that may affect SDM performance, namely sample size, modelling technique and species modelled. To do this, we use distribution data for the 34 Iberian endemic terrestrial vertebrate

species, simulating calibration datasets with different survey designs and levels of sample size within the Iberian Peninsula, and evaluating the ability of different SDM techniques to interpolate their geographical distributions based on these datasets.

METHODS

Species data

We used data on the whole extent of the distribution of 34 Iberian endemic terrestrial vertebrate species (15 amphibian, 12 reptile and 7 mammal species) in the 5919 UTM cells of 10 x 10 km that conform the Iberian Peninsula. These data provide accurate representations of current distributions and were compiled from national databases in the context of a recent study examining climate change impacts on Iberian terrestrial fauna (Araújo et al., 2011). Cells where a species is not recorded were considered as a true absence for the species. We selected all endemic species with more than 15 records in the database; for these species the number of records ranged from 19 to 4738. We chose endemic species (or nearly endemic, species with occurrences in the north facing slopes of the Pyrenees) to minimize the risk of including spurious effects of truncations of the climatic niche (Thuiller et al., 2004). The complete list of species can be found in supporting information Table S1.

Environmental data

GIS data on climate and topography were obtained from Worldclim (Hijmans et al., 2005) and Global Resource Information Database - United Nations Environment Programme (UNEP/GRID, http://www.grid.unep.ch/data/data.php). To reduce collinearity among variables we used a principal components analysis (PCA) to guide the selection of a subset of variables among the 29 available (see also Baselga & Araújo, 2009). The first two axes

accounted for 73% of total variability; the rest of the axes were not significant according to a broken-stick criterion, so we discarded them. We therefore selected the two variables that accounted for most of the variability in each one of the first two axes: Mean temperature of the warmest quarter, and precipitation of the driest, coldest and wettest quarters.

These four environmental layers were used as the basis for the simulation of the stratified survey designs and the calibration of all SDMs. The location and density of roads was extracted from a commercial database maintained by Tele Atlas (http://www.tomtom.com/), and data on human population density (in year 2000) from the Gridded Population of the World Project (CIESIN & CIAT, 2005). Finally, data on the location and limits of protected areas were extracted from the online database of the Natura2000 European network (http://ec.europa.eu/environment/nature/nature2000/) and the World Database on Protected Areas (IUCN & UNEP, 2009). All these GIS data were reprocessed into the UTM 10x10 km Iberian grid developed by EDIT Geoplatform (http://edit.csic.es/; Sastre et al., 2009).

Survey designs

We evaluated the effectiveness of seven alternative survey design strategies: random, systematic, environmentally stratified using classes, environmentally stratified using the *p*-median approach, biased as a function of accessibility, biased as a function of human population density, and biased towards protected areas. The details of each survey design are as follows:

a) *Random.* This procedure simulates surveys taken without any constraint or bias. Each cell in the study area had equal probability of being sampled; once a cell is surveyed, it cannot be sampled again.

b) *Systematic.* It simulates a planned survey free of any bias that is conducted throughout the whole area. First, one cell was chosen randomly, and then the remaining cells were chosen at regular intervals starting from the initial cell throughout the study area.

c) *Stratified.* Surveys were stratified along environmental gradients without bias, following two strategies:

   i) *Stratified by class:* In a first step several non-overlapping environmental domains (i.e. environmental classes) were defined by *k*-means cluster analysis – 10 groups – using the environmental data of each cell as descriptors. Then, equal numbers of cells were randomly selected within each environmental class. The choice of the number of groups used to perform the *k*-means cluster analysis was based in a visual comparison of the real climatic map of Iberian Peninsula (AEMET & IMP, 2011) with maps generated from 2 to 12 clusters.

   ii) *p-median stratification*: In this survey design the cells to be sampled were selected in order to maximize the coverage of the environmental and spatial variation within the study area, as described by the matrices of environmental and spatial distances between all Iberian 10x10 km cells (Hortal & Lobo, 2005; Funk et al., 2005). According to this protocol, an initial number of cells (usually the well sampled sites) are chosen; then, the distances from each unselected site to the nearest selected (initial) cell are calculated based on the abovementioned matrices, and the distances are recalculated simulating that each site is added to the set of selected areas, summing the distances to such hypothetical new selected set from all unselected sites; the site that minimizes this sum is then incorporated to the group of already selected areas. At each iteration one grid cell is chosen and the distance and sum from the remaining cells to

the set of selected ones are recalculated (see Hortal & Lobo, 2005; Hortal et al., 2009). In our case, two cells were chosen randomly as initial set of cells.

d) *Biased.* In biased surveys some sites are less likely to be selected than others. We simulated the occurrence of bias in the selection of cells as caused by three different factors:

   i) *Accessibility*: Cells close to roads had a higher chance of being selected. To do this, the Euclidean geographic distance from each cell to the nearest national road was calculated. Then, these distance values were standardized to vary from 0.05 to 1, so the farthest cell had the smallest value (0.05); this provides a probability of being sampled to all cells, even the more distant ones. These standardized values were then used as weights while selecting cells at random.

   ii) *Aggregation by human density*: Here we simulate the greater amount of samples near to urban centres. An initial number of cells (anchor cells) were chosen based on population density. To choose the anchor cell the population density values were standardized to vary from 0.0001 to 1, so that cells with the higher population densities had higher values. These standardized values were used as weights on the probability of a random selection. To simulate aggregation of samples the remaining cells were selected based in their proximity to the anchor cells. To assign weights to all cells according to their proximity with the anchor cells we used a geographic Kernel estimator based on the geographic distance. This function simulates the aggregation by assigning weights to each cell that decreases with distance from the anchor cells. These weights were then used as probabilities in a new random selection procedure. In each step one of the anchor cells was chosen randomly and one neighbour cell to this anchor cell was selected. This process was repeated until the total number of cells to be surveyed was reached.

iii) *Protected areas*: Here, surveys were conducted only in protected areas, a common practice on biodiversity research. Cells that had at least 30% of their area protected were chosen (to ensure that protected areas cover the great majority of species present in the cell), and samples were randomly selected among them.

Selection and characteristics of the sample datasets

We extracted samples of five different sizes (1%, 5%, 10%, 20%, and 25%) from the full extent of the data domain (5919 10x10 km cells in the Iberian Peninsula), using each one of the seven survey designs described above. We chose 1% as the smaller sample size because, although small in relation to the regional extant, it can be the reality for the data actually used in many SDM applications, especially for rare species. Note that the distributions of many rare species are often modelled based on 5 or even less occurrences (see, eg. Kamino et al., 2012). We ran 50 simulations of each one of these (5 x 7) 35 combinations of survey design and sample size (Fig. 1), selecting a total of 1,750 subsets of cells within the domain. In each of these subsets we extracted presence-absence data for each species, as well as the environmental data for calibrating of the SDM. Due to the limitations of the survey designs, the final numbers of cells surveyed differed slightly from one strategy to another.

To describe the quality of the data provided by each class of survey design we measured the climatic bias of each sample. We also calculated two characteristics related to the species' distribution: prevalence and Relative Occurrence Area (ROA, Jiménez-Valverde et al., 2008; Lobo, 2008). These two characteristics are related, but not equivalent. Prevalence is a characteristic of the data, and reflects the proportion of occurrences from the total number of cases in the dataset – in this case, the number of presences related to the total number of cells within the domain of study. In contrast, ROA is an actual geographic trait of the species,

calculated as the ratio between the area occupied by the species and the entire study area – in this case, the Iberian Peninsula. Here, we estimated the area of occurrence of each species as the minimum convex polygon (i.e., the smallest polygon in which there is no internal angle is greater than 180 degrees) that includes all the occurrences from each species. Values of prevalence and ROA vary from 0 to 1; species with more restricted distributions (or less occurrences) present values near 0, while species with progressively larger ranges tend to 1.

We measured the climatic bias of each sample, in relation to the environmental conditions of the study area following Kadmon et al. (2003; see also Hortal et al., 2008). We created two density functions with each variable used to calibrate the SDM (mean temperature of warmest quarter, precipitation of driest quarter, precipitation of coldest quarter, and precipitation of wettest quarter) by dividing them in 10 equal-interval categories, considering all the variation in the Iberian Peninsula. The first density function created for each variable was based on the number of cells of the study area that fell into each category, and the second density function was created from the number of cells in each category captured in the samples. The value of climatic bias for each sample was thus calculated as the sum of differences of density functions for each climatic variable, standardized to vary from 0 to 1 (the less and most environmentally biased, respectively).

We also evaluated the number of times that a survey design failed to identify enough occurrences to generate models in each SDM technique. That is, every time that a model was not generated for a given species due to insufficient number of presences. We measured these failures for each combination of survey design, sample size and SDM technique. We measured these failures for each combination of survey design and sample size (50 runs) and SDM technique (8) (i.e. n = 400). Across these 400 simulations we calculated the mean number of species for which the survey design failed (ranging from 0 to a maximum of 34 species). This way we can evaluate the amount of biodiversity that may be overlooked during

the modelling process (considering all SDM techniques) for each specific survey design and sample size.

Species distribution modelling

To make predictions about the distributions of species we used an ensemble of eight SDM techniques, as well as -the combined consensus prediction (Araújo & New, 2007), using the BIOENSEMBLES platform for computer-intensive ensemble forecasting (Diniz-Filho et al., 2009; Rangel et al., 2009). SDM techniques were chosen to maximize the variety of strategies of statistical adjustment and type of input data currently available, and included: BIOCLIM (Busby, 1986), Gower distance (Carpenter et al., 1993), Mahalanobis distance (Farber & Kadmon, 2003), Euclidean distance, Generalized Linear Models (GLM; McCullagh & Nelder, 1989), Maximum Entropy Modelling (MaxEnt; Phillips et al., 2006), Ecological-Niche Factor Analysis (ENFA; Hirzel et al., 2002) and Random Forest (Breiman, 2001). These modelling techniques were used with their default settings, assuming that all data in each sample were true presences and true absences. The thresholds to convert the probabilities generated by the models into binary (i.e. presence/absence) predictions were chosen based on the prevalence of the species in each data calibration, so that the prevalence in the predictions reflects the prevalence of the training data (see Jiménez-valverde & Lobo, 2007). The consensuses were generated as the mean value of the outputs of all single-models (Araújo & New, 2007). To generate the species distributions maps with these techniques each combination of survey design and sample size was simulated 50 times (Fig. 1). Each of these simulations consisted in a new selection of calibration dataset following the rules of each type of survey design.

Statistical analyses

We used six metrics to analyze the performance of the different SDM techniques: sensitivity; specificity; kappa (Cohen, 1960), Area Under the ROC Curve (AUC; Fielding & Bell, 1997); true skill statistic (TSS; Allouche et al., 2006) and percentage of correctly classified instances (CCI; Fielding & Bell, 1997). These evaluation metrics include the most-used measures of SDM performance. We also evaluated the Klocation metric (Pontius, 2000; Geri et al., 2011), which is similar to kappa but takes into account the spatial location of errors. Given that its results were similar to kappa, we present only the results of this latter technique for the ease of comparison with former studies. All metrics were calculated based on the actual distributions of the species. Here, the evaluation was performed by comparing the binary predictions of presence and absence with the data from the distribution Atlases, after excluding the data used to calibrate the model.

The relationships between all evaluation metrics and survey design, species, sample size and modelling technique were assessed using a four-way ANOVA (Fig. 1), which analyzes the effect of categorical factors on a response by decomposing the variability in the response variable (dependent variable, here the evaluation metrics) amongst the different factors (independent variables, here survey design, species, sample size and SDM techniques). This analysis identifies which predictors have significant effects on the response variable, and how much of the variability in the response variable can be assigned to each factor (see also Diniz-Filho et al., 2009; Nenzén & Araújo, 2011). To ensure that the results of the analysis were not spuriously influenced by the different numbers of categories of each factor (which range from 5 to 34), the ANOVAs were performed by randomly choosing 5 categories in each factor. This procedure was repeated 10000 times for each assessment metric. Note that in this analysis the factor species summarizes a number of factors that could affect SDM performance, including data characteristics, and geographic and functional species' traits (see

Chefaoui et al., 2011). Including them would have resulted in an intractable ANOVA design. Therefore, we evaluated the influence of prevalence and ROA separately, by comparing them with the evaluation metrics through additional correlation analyses.

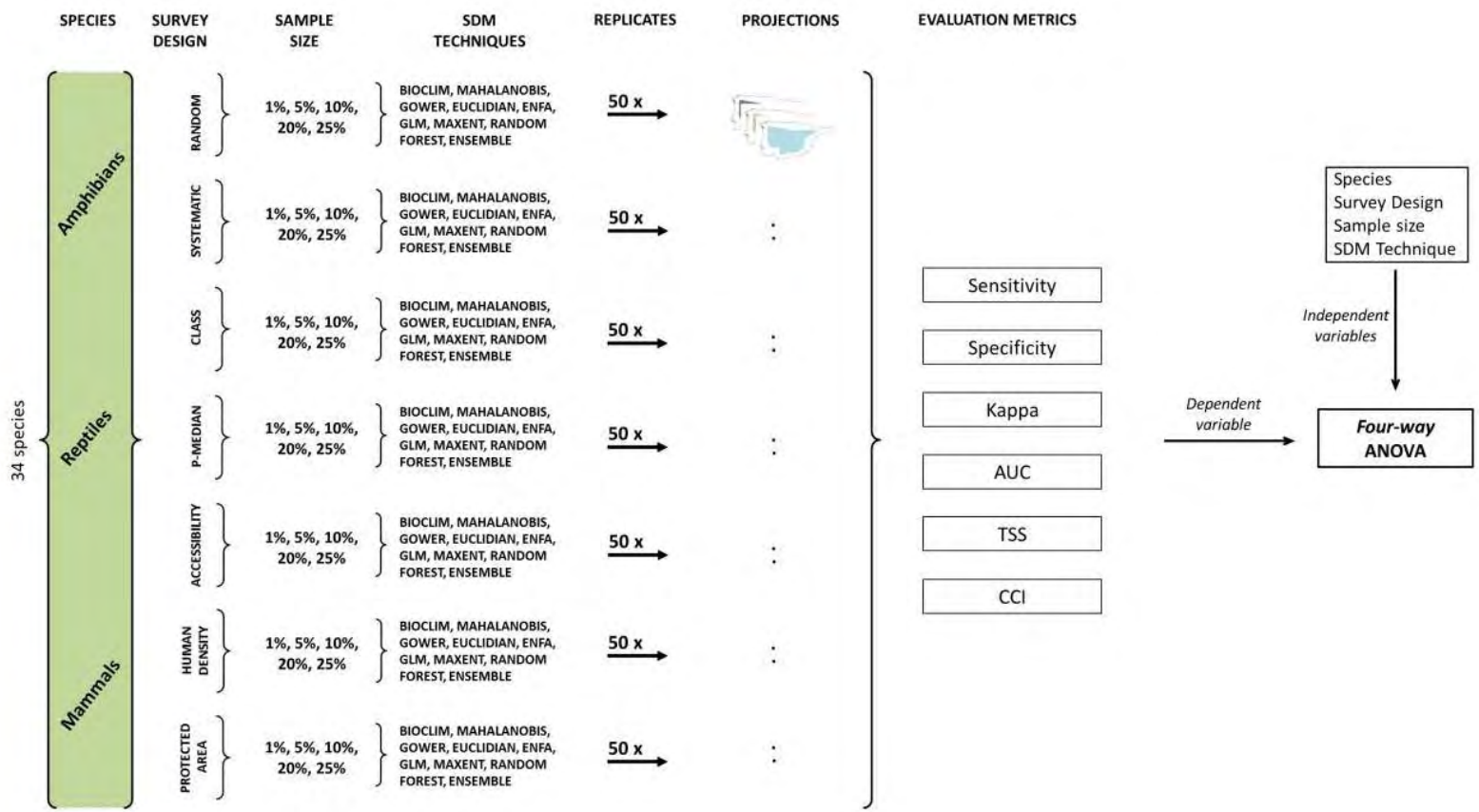


Figure 1. Schematic representation of the protocol used to generate predictions that were used in the evaluation of the effect of survey design on SDM. For each species we simulated 50 samples based on each survey design and sample size, and used these data to generate projections based on each modelling technique, as well as their ensemble. The six assessment metrics calculated for each map were used as dependent variable in six different four-way ANOVAs, using survey design, species, sample size and SDM technique as independent variables.

RESULTS

Factors affecting SDM performance varied across the different evaluation metrics. For five out of six evaluation metrics, the species modelled was the most important factor affecting model performance, followed by sample size and SDM technique (Table 1). The contribution of survey design, though significant, is much less important than any of these three other factors (Table 1). Contrary to our expectations, predictions of species potential distributions generated from different survey designs showed only comparatively small differences in their variance across evaluation metrics, compared to other factors. Interestingly, the degree of climatic bias was generally low for all survey designs. The most biased design was the one that follows human density (bias around 22%), while systematic surveys were the least biased (Fig. 2). Perhaps as a consequence, samples based on designs biased by human density performed worse for sensitivity, kappa, TSS and AUC, while those stratified by groups produced the better survey designs according to the same metrics (Fig. 3; Supporting information Fig. S1). Increasing sample size led to increasing SDM performance according to all metrics, although this pattern was clearer for Sensitivity and TSS (Fig. 4; Supporting information Fig. S2). Given that for most factors Sensitivity and Specificity were the most informative we show their results in the main text. The rest are available at the Supporting information (Figs. S1 to S5).

Table 1. Fvalues for each factor and assessment metric. The values show the mean of 10,000 ANOVA results per metric.

Factor	Evaluation metric					
	Sensitivity	Specificity	Kappa	TSS	CCI	AUC
Species	1893.15	9743.14	21114.05	19954.83	12582.46	3156
SDM technique	1280.90	799.66	1426.80	1389.54	804.46	103578.07
Sample Size	1653.14	607.40	2589.35	4718.27	1010.93	3673.84
Survey Design	1683.43	371.23	798.87	1499.87	150.15	196.26

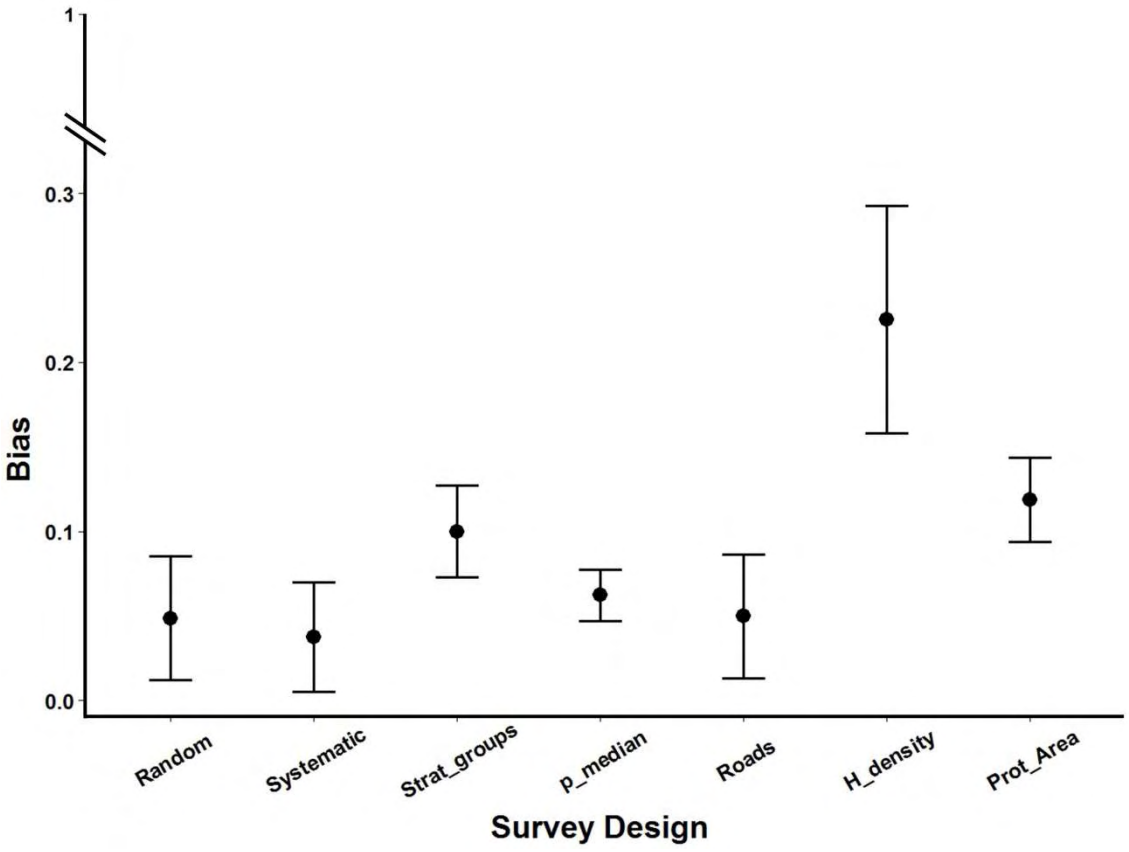


Figure 2. Median and standard deviation of the climatic bias (measured as the sum of differences in the frequencies) in the samples obtained with each survey design. Strat_groups = stratified by groups; p-median = Stratified by p-median; H_density = bias by human density, Prot. Area = bias by protected areas.

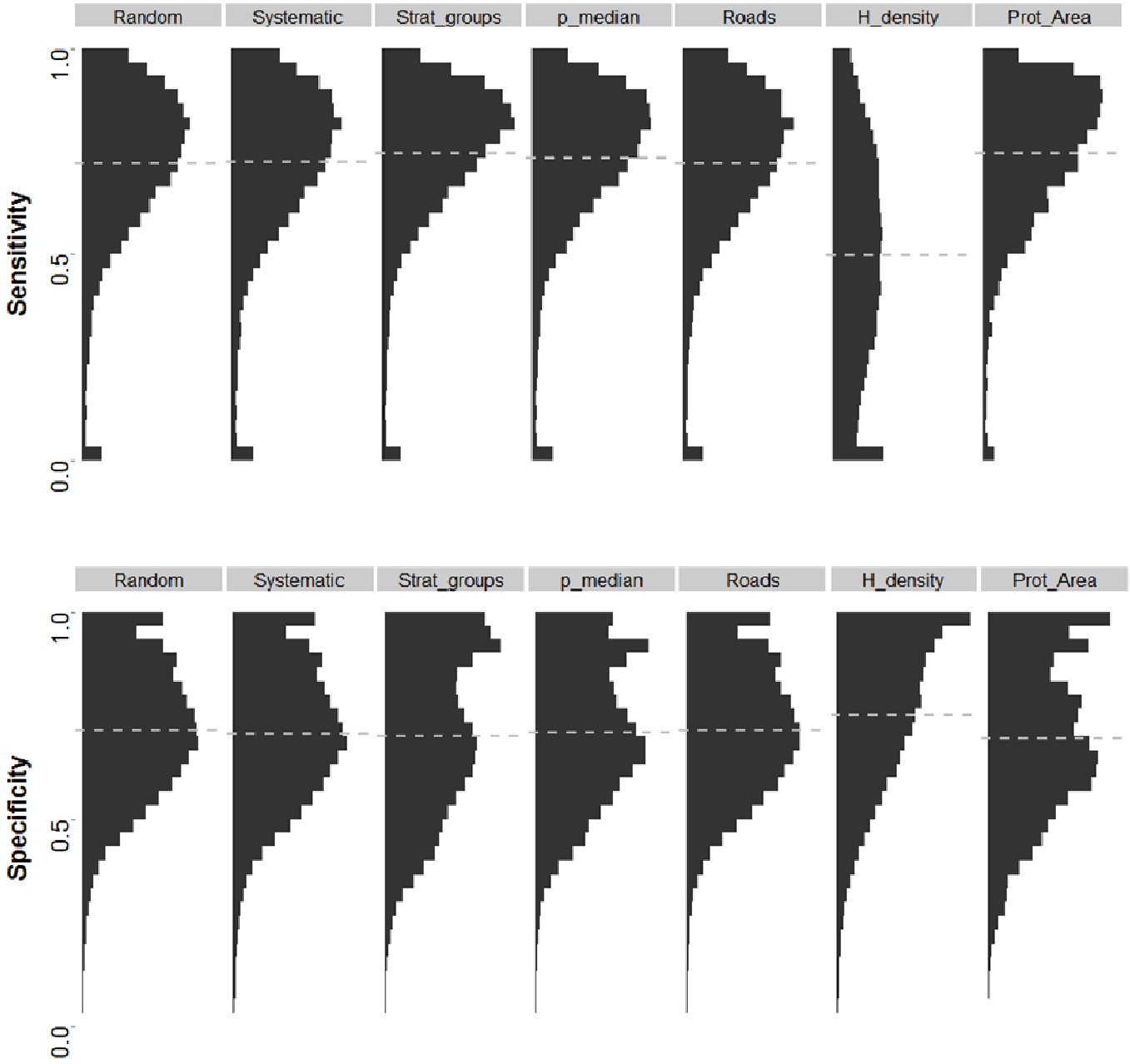


Figure 3. Performance of SDM predictions generated from data based on different survey designs, according to sensitivity and specificity. Survey design codes as in Fig. 2. The histograms represent the frequency in each class, and the grey lines indicate the mean. For other metrics of SDM performance see supporting information Fig. S1.

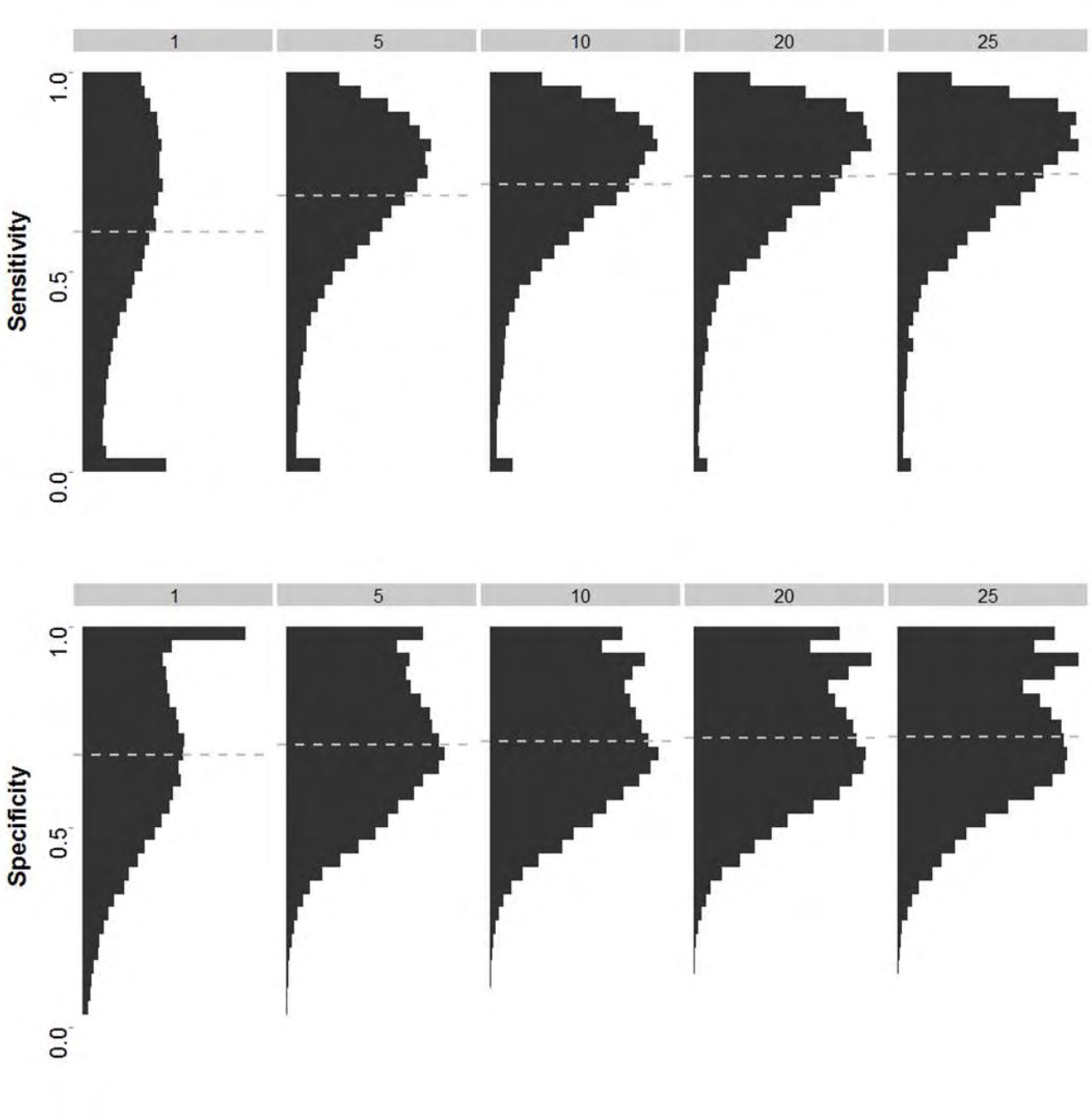


Figure 4. Performance of SDM predictions generated from datasets of five sample sizes (representing the percentage of cells selected from the whole study area, from 1 to 25%), according to sensitivity and specificity. The histograms represent the frequency in each class and the grey lines indicate the mean. For other metrics of SDM performance see supporting information Fig. S2.

With regard to SDM techniques, consensus predictions, GLM and MaxEnt ranked nearly always among the highest performing techniques. However, consensus predictions included some of the lowest performing predictions according to AUC, and GLM was below average for specificity (Fig. 5; Supporting information Fig. S3). Random Forest performed the best according to sensitivity and the worst for specificity, confirming a tendency for overfitting of the data. The opposite pattern was recorded for BIOCLIM, also confirming a tendency for underfitting (Supporting information Fig. S3). Random Forest was the technique the most affected by sample size, followed by MaxEnt and GLM, while consensus predictions were the least affected according to Specificity and CCI (Supporting information Fig. S4). Interestingly, the performance of some SDM techniques was unaffected by sample size across all metrics; according to Specificity and CCI, the interaction between sample size and technique causes a decrease in the performance of BIOCLIM, Gower distance and ENFA (Supporting information Fig. S4).

Despite the large influence of the species being modelled on model performance, we were not able to identify a particular group of species for which SDM performance was consistently better or worse than for others (Supporting information Fig. S5). We performed additional correlation analyses between Relative Occurrence Area (ROA), prevalence of species and all validation metrics (Table 2). Both ROA and prevalence showed negative correlations with all metrics except AUC, so that species with higher prevalences (i.e. more occurrences within the dataset) generated models with lower performance. Specificity, TSS and CCI detected much stronger negative effects of prevalence and ROA on predictions; this is perhaps not surprising, because these metrics also identify species as the main factor affecting SDM results, thus evidencing that the particular characteristics of each species' geographical range has large impacts on SDM performance. The number of times that a survey design did not generate sufficient occurrences to allow the modelling decreased with

the increase in sample size. In all cases, the values of these failures were highest for the biased human density design, particularly at the smaller sample size; almost 40% of the species did not present enough occurrences to be modelled at such sample size (Supporting information Fig. S6).

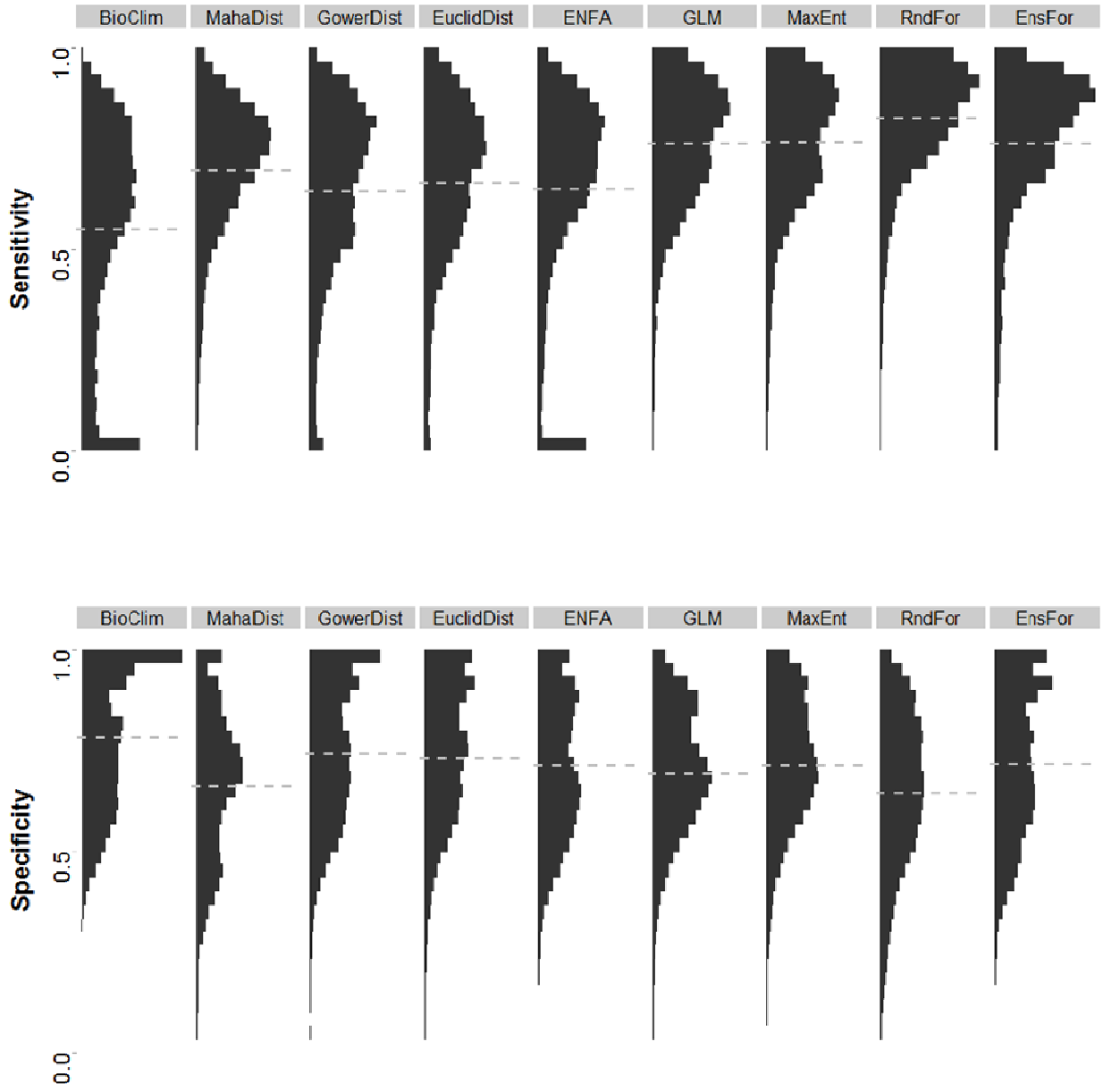


Figure 5. Predictive performance of different SDM techniques according to sensitivity and specificity. The histograms represent the frequency in each class and the grey lines indicate the mean. For the other metrics of SDM performance see supporting information Fig. S3.

Table 2. Pearson correlations (*r*) between the evaluation metrics and the prevalence on training dataset and ROA. All correlations were significant at $p > 0.001$

Metric	Prevalence		ROA	
	t	*R*	t	*r*
Sensitivity	-693.539	-0.09	-134.36	-0.19
Specificity	-317.905	-0.41	-525.7	-0.6
Kappa	-685.686	-0.097	-165.2	-0.23
TSS	-289.062	-0.38	-516.06	-0.59
CCI	-271.893	-0.36	-544.27	-0.61
AUC	13.58	0.02	21.27	0.03

DISCUSSION

How data are collected is supposed to be of critical importance for the development of species distribution models (Araújo & Guisan, 2006). The data used to calibrate models are known to lead to differences in SDM predictions thus being expected to affect their performance (e.g. Kadmon et al., 2003; Barry & Elith, 2006; Sánchez-Fernández et al., 2011). However, as shown here for the first time, the importance of data collection is dwarfed when compared with other factors affecting SDM predictions. Indeed, contrary to our own expectations, variation in survey design had relatively small, though significant, effects on the performance of SDM. Instead, species identity ranked highest among factors affecting the models, for all performance metrics (except AUC), followed by sample size and SDM technique.

The main assumption of SDMs is that species distributions are limited by environmental –typically climatic– factors (Araújo & Peterson, 2012). It follows that

restricting the range of environmental variation in which a species occurs could potentially affect the calculation of species-climate response curves which, in turn, could cause projections to be potentially erroneous (Thuiller et al., 2004). Surveys providing a comprehensive coverage of the environmental conditions matching species distributions, like the *p*-median or the stratified sampling by groups are suited for SDM applications because they reduce climatic biases in the data. However, neither the samples taken with these survey strategies showed the lowest values of climatic bias, nor did models based on them present significantly better predictive ability. Rather, most survey designs yielded models with very similar predictive accuracy and showed little differences in climatic bias (except for human density, see below).

One possible explanation for this unexpected result is that the spatial biases simulated in our analyses did not produce large differences in climatic biases. Surveys based in non-random and non-stratified designs are often expected to be spatially and environmentally biased (Hortal et al., 2007; Loiselle et al., 2008), but this is not always the case. In our simulated surveys, sampling along roads had values of climatic bias similar to random samples, and only slightly higher than systematic surveys. This implies that, at the scale and extent analyzed, the spatial location of roads in the Iberian Peninsula covers a large climatic variability, even with small sample sizes. Thus, the spatial bias in the surveys due to accessibility may not always result in climatic biases (Kadmon et al., 2004; McCarthy et al., 2012).

Importantly, the surveys that followed human density showed substantially higher climatic biases (as well as higher variability in such bias) compared to the other strategies. This is probably due to the fact that, following real situations, this simulation allows the cells that are surveyed at each stage to be located either further away or closer in the climatic space. This increases the variability in the values of climatic bias, and at the same time limits further

selections of cells to the neighbors of the already surveyed cells, thus increasing the climatic bias. Not surprisingly, this survey design was the one yielding the worst-performing SDM according to all metrics but specificity. Here, the contrasting behaviour of the values of specificity for the human density simulation (see Figure 3) can be due to the characteristics of this metric, which evaluates the capacity of the model to discriminate true absences, so the best models are those presenting less commission errors. Models built with biased climatic data can overfit to the climatic conditions that were effectively sampled (Lobo, 2008; Phillips et al., 2009), hence generating spatially-restricted predictions. This increases omission errors (false absences) as well as the number of true absences predicted by the model. Thus, metrics that use the number of true absences to assess SDM performance, such as specificity, will present higher values in climatically-biased survey designs, whereas metrics based on the number of omission errors can present low values for the same predictive maps. It is also worth noting that other biased survey designs did not show highly variable climatic bias values because both roads and protected areas are well distributed within the Iberian Peninsula, allowing them to cover the most important environmental gradients within this region. However, in areas where the roads and national parks (or any other spatial attribute influencing the surveys) are not widely distributed, samples following these biased designs can generate larger climatic bias, limiting the predictive ability of SDMs and consequently their utility. Should the climatic biases generated by our surveys be higher, we would have found larger effects of survey design. Having said this, our results point out to that the effects of climatic bias on SDM performance could be limited until a certain level, where model accuracy diminishes dramatically, a potential effect that deserves further investigation.

The importance of the particular species being modelled for SDM performance shown by our results may be due to differences in their geographic distributions. Many studies have highlighted that geographical and ecological species' characteristics affect SDM accuracy

(Berg et al., 2004; Segurado & Araújo, 2004; Guisan et al., 2007). Many spatial and ecological characteristics can affect SDM results (Brotons et al., 2004; Segurado & Araújo, 2004; Hernandez et al., 2006; Guisan et al., 2007; McPherson & Jetz, 2007; Newbold et al., 2009; Chefaoui et al., 2011), from which the proportion of the occupied area over the considered territory (i.e. ROA) and the prevalence are frequently reported (Brotons et al., 2004; Luoto et al., 2005; Lobo et al., 2007b; Chefaoui et al., 2011). These variations in performance due to the species characteristics can occur even when using the same SDM technique (Seoane et al., 2005; Hanspach et al., 2010), thus increasing their influence on SDM results. In our case, species with smaller geographical distributions are predicted better, as evidenced by the negative correlation between ROA and the assessment metrics. This is because it is easier for SDMs to capture the climate-distribution relationship for species of restricted range, for they occupy a smaller –and thus easier to classify– environmental domain within the study region (Segurado & Araújo, 2004; Jiménez-Valverde et al., 2008; Lobo, 2008). This implies that relatively small sample sizes but with a fair geographic coverage of surveys can provide data of sufficient quality to avoid spurious effects on SDM accuracy of geographically restricted species. As a consequence, the importance of survey design for SDM performance may be limited in our analysis because it is restricted to endemic species – which often present limited spatial distributions. It is however likely that the importance of survey design will increase for species with larger ranges, given that the more widely distributed is a species, the more likely is that biased survey designs are not able to include all the climatic conditions it inhabits, hence resulting in less accurate models.

Sample size is also known to have strong effects on SDM predictive accuracy (Hirzel & Guisan, 2002; Reese et al., 2005; Jiménez-Valverde et al., 2009; Chefaoui et al., 2011). For example, Araújo et al. (2009) showed that apparent failure of SDM to characterize European bird species-climate relationships in a high profile study (Beale et al., 2008) was due to

incomplete coverage of available data and, in another study, it was shown that the ability to predict range shifts of British birds under climate change was highly affected by the completeness of the data used to calibrate the models (Araújo et al., 2005). In our work increasing sample size led to increases in SDM performance and decreases in variability of model performance; predictions reached a large degree of stability with sample sizes of 10% of the studied area (592 cells) or larger. The measured effect of sample size may be in part related to its interaction with prevalence (Jiménez-Valverde et al., 2009); below a certain threshold (such as the 10% we found), the representation of the distribution of the species may be too poor to produce reliable models. Larger sample sizes would allow more accurate descriptions of the species' response to the environment, thus yielding better predictions (Wisz et al., 2008). Importantly, in our analyses larger sample sizes provide better coverage of the study area, rather than just increasing numbers of occurrences. Thus, the stability in the predictions can be due to the fair representation of the environmental variation within the region, a key factor for the accuracy of the description of species–environment relationships (Kadmon et al., 2003; Hortal et al., 2008, 2012; Hortal & Lobo, 2011) and therefore of higher importance for SDM performance than sample size *per se* (Newbold et al., 2009). Sastre & Lobo (2009) found that biased survey designs were not efficient in covering the true geographic pattern of species richness within a region (see also Hortal & Lobo, 2011). As showed by our analyses, the failures in covering the species distributions can be accentuated when biased survey designs are associated with smaller sample sizes. These failures are especially problematic when aiming to model the whole biodiversity of a region, and consequently for conservation actions based on many data-intensive strategies.

SDM techniques differ in their capacity to capture the relationship between species distributions and environmental variables. In general, complex techniques generate models with better performance within the training data (Elith et al., 2006; Tsoar et al., 2007; but see

Lobo, 2008; Hijmans, 2012). However, the results of model comparisons are conflicting across studies, due to strong variations in the modelling techniques and their possible interaction with data characteristics (Araújo et al., 2005, 2009). Although no SDM technique performs better in all cases, our results show that the ensemble forecasting approach could be a good strategy to enhance SDM performance, for it consistently ranks within the best-performing technique regardless of the validation metric used. However, in some cases the consensus predictions yielded models with lower performance than individual SDM, perhaps due to the inclusion of predictions from poorly-performing techniques. The accuracy of consensus predictions is highly constrained by the quality of individual predictions (Araújo et al., 2005). Therefore, to improve the performance of ensemble forecasting approaches it is necessary to modify the building-up of the consensus, either removing poor models or weighting individual models based on a previously selected performance metric (Araújo & New, 2007; Marmion et al., 2009; Garcia et al., 2012).

To summarize, our results support the view that the environmental coverage of the surveys is more important than the spatial structure of the calibration data. Spatially biased surveys do not always yield environmentally biased data, and therefore can generate models as satisfactory as those using carefully designed surveys. Nevertheless, we contend that the distributional data used for species distribution modelling should be evaluated to characterize their climatic bias and coverage (see Kadmon et al., 2004; Hortal et al., 2008; Hortal & Lobo, 2011). Whenever possible, contrasting estimated realized niches with estimated analogous dimensions of the fundamental niche could provide useful insights as to the possible biases in the estimates with SDM (Araújo et al., 2013). The largest source of variation in SDM predictions is the species being modelled, and more precisely the extent of its distribution within the studied area. Geographically restricted species (i.e. with lower ROA) yield more accurate results, but better predictions are also obtained with larger sample sizes and from

survey designs with low climatic biases. Further studies are necessary to reveal the specific geographic or functional traits that can affect the predictive power of SDM. Additionally, is important to investigate whether there is a minimum threshold of climatic bias behind which the quality of SDM predictions changes dramatically.

ACKNOWLEDGEMENTS

GT was funded by CAPES REUNI and PDSE grants nº11842121. TFR has been supported by several CNPq research grants (564718/2010-6, 474774/2011-2, 310117/2011-9). MBA was funded through the FCT PTDC/AAC-AMB/98163/2008 project and the Integrated Program of IC&DT Call Nº 1/SAESCTN/ALENT-07-0224-FEDER-00175. JH was supported by a Spanish DGCyT RyC Fellowship, by a Brazilian CNPq Visiting Researcher grant (400130/2010-6), and by the CSIC-CNPq cooperation project entitled *Developing tools to unify range dynamics and community-level processes into a single analytical framework* (2011BR0071).

REFERENCES

AEMET & IMP (2011) Iberian climate atlas. Air temperature and precipitation (1971-2000). Agencia Estatal de Meteorología and Instituto de Meteorologia de Portugal, Madrid.

Albert C.H., Yoccoz N.G., Edwards T.C., Graham C.H., Zimmermann N.E., & Thuiller W. (2010) Sampling in ecology and evolution - bridging the gap between theory and practice. *Ecography*, **33**, 1028–1037.

Allouche O., Tsoar A., & Kadmon R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Araújo M.B., Ferri-Yáñez F., Bozinovic F., Marquet P.A., Valladares F., & Chown S.L. (2013) Heat freezes niche evolution. *Ecology letters*, **16**, 1206–19.

Araújo M.B., Guilhaumon F., Neto D.R., Ortego I.P., & Calmaestra R.G. (2011) *Impactos, vulnerabilidad y adaptación al cambio climático de la biodiversidad española 2. fauna de vertebrados.* Dirección General de Medio Natural y Política Forestal. Ministerio de Medio Ambiente, y Medio Rural y Marino, Madrid.

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

Araújo M.B. & New M. (2007) Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22**, 42–47.

Araújo M.B. & Peterson T.A. (2012) Uses and misuses of bioclimatic envelope modeling. **93**, 1527–1539.

Araújo M.B., Thuiller W., & Yoccoz N.G. (2009) Reopening the climate envelope reveals macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences USA*, **106**, 106:E45–E46.

Araújo M.B., Whittaker R.J., Ladle R.J., & Erhard M. (2005) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, **14**, 529–538.

Austin M.P. & Heyligers P. (1989) Vegetation survey design for conservation: Gradsect sampling of forests in North-eastern New South Wales. *Biological Conservation*, **50**, 13–32.

Barry S. & Elith J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Baselga A. & Araújo M.B. (2009) Individualistic vs community modelling of species distributions under climate change. *Ecography*, **32**, 55–65.

Beale C.M. & Lennon J.J. (2012) Incorporating uncertainty in predictive species distribution modelling. *Philosophical transactions of the Royal Society of London*, **367**, 247–58.

Beale C.M., Lennon J.J., & Gimona A. (2008) Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences USA*, **105**, 14908–14912.

Berg Å., Gärdenfors U., & Von Proschwitz T. (2004) Logistic regression models for predicting occurrence of terrestrial molluscs in southern Sweden - importance of environmental data quality and model complexity. *Ecography*, **27**, 83–93.

Braunisch V. & Suchant R. (2010) Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. *Ecography*, **33**, 826–840.

Breiman L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.

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

Busby J.R. (1986) A biogeographical analysis of Notophagus cunninghamii (Hook.) in south-eastern Australia. *Australian Journal of Ecology*, **11**, 1–7.

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**, 667–680.

Chefaoui R.M., Lobo J.M., & Hortal J. (2011) Effects of species' traits and data characteristics on distribution models of threatened invertebrates. *Animal Biodiversity and Conservation*, **34**, 229–247.

CIESIN C.U. & CIAT (2005) *Gridded Population of the World, Version 3 (GPWv3). Socioeconomic Data and Applications Center (SEDAC), Columbia University – Center for International Earth Science Information Network (CIESIN), Columbia University and Centro Internacional de Agricultura.*

Cohen J. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement. **20**, 37–46.

Colwell R.K. & Rangel T.F. (2009) Hutchinson's duality: the once and future niche. *Proceedings of the National Academy of Sciences USA*, **106**, 19651–196588.

Dennis R.L.H. & Thomas C.D. (2000) Bias in butterfly distribution maps : the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, **4**, 73–77.

Diniz-Filho J.A.F., Bini L.M., Fernando Rangel T., Loyola R.D., Hof C., Nogués-Bravo D., & Araújo M.B. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897–906.

Edwards Jr. T., Cutler D., Zimmermann N., Geiser L., & Moisen G. (2006) Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, **199**, 132–141.

Elith J., H. Graham C., P. Anderson R., Dudík M., Ferrier S., Guisan A., J. Hijmans R., Huettmann F., R. Leathwick J., Lehmann A., Li J., G. Lohmann L., A. Loiselle B., Manion G., Moritz C., Nakamura M., Nakazawa Y., McC. M. Overton J., Townsend Peterson a., J. Phillips S., Richardson K., Scachetti-Pereira R., E. Schapire R., Soberón J., Williams S., S. Wisz M., & E. Zimmermann N. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Farber O. & Kadmon R. (2003) Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, **160**, 115–130.

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

Franklin J. & Miller J.A. (2009) *Mapping species distributions: spatial inference and prediction.* Cambridge University Press Cambridge, UK, New York, New York, USA.

Funk V.A., Richardson K.S., & Ferrier S. (2005) Survey-gap analysis in expeditionary research: where do we go from here? *Biological Journal of the Linnean Society*, **85**, 549–567.

Garcia R.A., Burgess N.D., Cabeza M., Rahbek C., & Araújo M.B. (2012) Exploring consensus in 21st century projections of climatically suitable areas for African vertebrates. *Global Change Biology*, **18**, 1253–1269.

Geri F., Amici V., & Rocchini D. (2011) Spatially-based accuracy assessment of forestation prediction in a complex Mediterranean landscape. *Applied Geography*, **31**, 881–890.

Guisan A., Zimmermann N.E., Elith J., Graham C.H., Phillips S., & Peterson a. T. (2007) What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, **77**, 615–630.

Hanspach J., Kühn I., Pompe S., & Klotz S. (2010) Predictive performance of plant species distribution models depends on species traits. *Perspectives in Plant Ecology, Evolution and Systematics*, **12**, 219–225.

Heikkinen R.K., Luoto M., Araujo M.B., Virkkala R., Thuiller W., & Sykes M.T. (2006) Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography*, **30**, 751–777.

Hernandez P.A., Graham C.H., Master L.L., & Albert D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

Hijmans R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, **93**, 679–88.

Hijmans R.J., Cameron S.E., Parra J.L., Jones P.G., & Jarvis A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

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

Hirzel A. & Guisan A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157**, 331–341.

Hortal J., Araújo M.B., & Lobo J.M. (2009) Testing the effectiveness of discrete and continuous environmental diversity as a surrogate for species diversity. *Ecological Indicators*, **9**, 138–149.

Hortal J., Jimenez-Valverde A., Gómez J.F., Lobo J.M., & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Hortal J. & Lobo J.M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913–2947.

Hortal J. & Lobo J.M. (2011) Can species richness patterns be interpolated from a limited number of well-known areas? Mapping diversity using GLM and kriging. *Natureza & Conservação*, **9**, 200–207.

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 : the journal of the Society for Conservation Biology*, **21**, 853–63.

Hortal J., Lobo J.M., & Jiménez-Valverde A. (2012) Basic questions in biogeography and the (lack of) simplicity of species distributions: Putting species distribution models in the right place. *Natureza e Conservação*, **10**, 108–118.

IUCN & UNEP (2009) *The World Database on Protected Areas (WDPA). UNEP-WCMC, Cambridge. Available at http://www.wdpa.org/.*

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.

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

Kadmon R., Farber O., & Danin A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, **13**, 853–867.

Kadmon R., Farber O., & Danin A. (2004) Effects of roadside bias on the accuracy of predictive maps produced by bioclimtic models. *Ecological Applications*, **14**, 401–413.

Kamino L.H.Y., Stehmann J.R., Amaral S., De Marco P., Rangel T.F., de Siqueira M.F., De Giovanni R., & Hortal J. (2012) Challenges and perspectives for species distribution modelling in the neotropics. *Biology letters*, **8**, 324–326.

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

Lobo J.M., Baselga A., Hortal J., Jiménez-Valverde A., & Gómez J.F. (2007a) 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*, **33**, 103–114.

Lobo J.M., Jiménez-Valverde A., & Real R. (2007b) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Loiselle B.A., Jørgensen P.M., Consiglio T., Jiménez I., Blake J.G., Lohmann L.G., & Montiel O.M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, **35**, 105–116.

Luoto M., Pöyry J., Heikkinen R., & Saarinen K. (2005) Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and biogeography*, **14**, 575–584.

Marmion M., Parviainen M., Luoto M., Heikkinen R.K., & Thuiller W. (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, **15**, 59–69.

McCarthy K., Fletcher R.J., Rota C.T., & Hutto R.L. (2012) Predicting species distributions from samples collected along roadsides. *Conservation Biology*, **26**, 68–77.

McCullagh P. & Nelder J.A. (1989) *Generalized Linear Models.* London.

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

Medina N.G., Lara F., Mazimpaka V., & Hortal J. (2013) Designing bryophyte surveys for an optimal coverage of diversity gradients. *Biodiversity and Conservation*, **22**, 3121–3139.

Nenzén H.K. & Araújo M.B. (2011) Choice of threshold alters projections of species range shifts under climate change. *Ecological Modelling*, **222**, 3346–3354.

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.

Pereira J.M.C. & Itami R.M. (1991) GIS-based habitat modeling using logistic multiple regression: a study of the Mt. Graham red squirrel. *Photogrammetric Engineering & Remote Sensing*, **57**, 1475–1486.

Peterson A.T., Soberón J., Pearson R.G., Anderson R.P., Martínez-Meyer E., Nakamura M., & Araújo M.B. (2011) *Ecological niches and geographic distributions.* Princeton University Press.

Phillips S., Anderson R., & Schapire R. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips S.J., Dudík M., Elith J., Graham C.H., Lehmann A., Leathwick J., & Ferrier S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, **19**, 181–97.

Pontius R.G. (2000) Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing*, **66**, 1011–1016.

Rangel T.F.L.V.B., Diniz-Filho J.A.F., & Araújo M.B. (2009) *BIOENSEMBLES 1.0. Software for computer intensive ensemble forecasting of species distributions under climate change.* Privately distributed. Madrid/Goiânia/Évora.

Reese G.C., Kenneth R.W., H J.A., & Flather C.H. (2005) Factos affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, **15**, 554–564.

Rocchini D., Hortal J., Lengyel S., Lobo J.M., Jimenez-Valverde A., Ricotta C., Bacaro G., & Chiarucci A. (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Romo H., Garcıa-Barros E., & Lobo J.M. (2006) Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography*, **29**, 873–885.

Sánchez-Fernández D., Lobo J.M., & Hernández-Manrique O.L. (2011) Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. *Diversity and Distributions*, **17**, 163–171.

Sastre P. & Lobo J. (2009) Taxonomist survey biases and the unveiling of biodiversity patterns. *Biological Conservation*, **142**, 462–467.

Sastre P., Roca P., & Lobo J.M. (2009) A Geoplatform for improving accessibility to environmental cartography. *Journal of Biogeography*, **36**, 568–568.

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

Seoane J., Bustamante J., & Díaz-Delgado R. (2005) Effect of expert opinion on the predictive ability of environmental models of bird distribution. *Conservation Biology*, **19**, 512–522.

Soberón J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology letters*, **10**, 1115–23.

Soberón J. & Nakamura M. (2009) Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences USA*, **106**, 19644–50.

Soberón J.M. (2010) Niche and area of distribution modeling: a population ecology perspective. *Ecography*, **33**, 159–167.

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.

Thuiller W., Lafourcade B., Engler R., & Araújo M.B. (2009) BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

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

Varela S., Anderson R.P., García-Valdés R., & Fernández-González F. (2014) Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, **37**, 1–8.

Wisz M.S., Hijmans R.J., Li J., Peterson A.T., Graham C.H., & Guisan A. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table S1 List of species used for the analyses, with numbers of presences and absences, degree of endemism and species code in the analysis.

Figure S1 Performance of SDM predictions from datasets based on different survey designs.

Figure S2 Performance of predictions from datasets of five sample sizes.

Figure S3 Predictive performance of SDM techniques according to Kappa, AUC, TSS and CCI validation metrics.

Figure S4 Interaction between sample size and model for specificity and CCI metrics.

Figure S5 SDM performance for all studied species according to Kappa, AUC, TSS and CCI validation metrics.

Figure S6 Number of times that a survey design does not get enough occurrences to generate models.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCH

Geiziane Tessarolo is a PhD student based at the Universidade Federal de Goiás, Brazil. Her research focuses on the assessment and evaluation of sources of uncertainty in species distribution modelling, particularly on data quality.

Thiago F. Rangel, Miguel B. Araújo and Joaquín Hortal are biogeographers studying the geographic distributions of species, among many other things.

Author contributions: G.T., J.H. and T.F.R. conceived the study; A.M.B. provided the data; G.T., J.H. and T.F.R. designed the analyses, with A.M.B.; G.T., J.H. and T.F.R. analyzed the data; G.T., J.H., T.F.R. and A.M.B. wrote the paper.

Supporting information:

Table S1. List of species available for samples, with number of presence and absence, degree of endemicity and species code in the analysis.

Species name	Species code	Presence	Absence	Endemicity
		Reptiles		
Algyroides marchi	sp1	25	5894	Endemic
Blanus cinereus	sp2	1400	4519	Endemic
Chalcides bedriagai	sp3	604	5315	Endemic
Chalcides striatus	sp4	1204	4715	Almost endemic
Iberolacerta bonnali	sp5	19	5900	Almost endemic
Iberolacerta monticola	sp6	127	5792	Endemic
Lacerta schreiberi	sp7	825	5094	Endemic
Podarcis bocagei	sp8	360	5559	Endemic
Podarcis carbonelli	sp9	57	5862	Endemic
Psammodromus hispanicus	sp10	1344	4575	Almost endemic
Rhinechis scalaris	sp11	2554	3365	Almost endemic
Vipera seoanei	sp12	360	5559	Endemic
			5894	
		Amphibians		
Alytes cisternasii	sp13	927	4992	Endemic
Alytes dickhilleni	sp14	131	5788	Endemic
Chioglossa lusitanica	sp15	342	5577	Endemic
Discoglossus galganoi	sp16	950	4969	Endemic
Discoglossus jeanneae	sp17	467	5452	Endemic
Euproctus asper	sp18	166	5753	Almost endemic
Hyla meridionalis	sp19	988	4931	Almost endemic
Lissotriton boscai	sp20	1222	4697	Endemic
Pelobates cultripes	sp21	1675	4244	Almost endemic
Pelodytes ibericus	sp22	507	5412	Endemic
Pleurodeles waltl	sp23	1400	4519	Almost endemic
Rana iberica	sp24	719	5200	Endemic
Rana perezi	sp25	4738	1181	Almost endemic
Rana pyrenaica	sp26	23	5896	Endemic
Triturus pygmaeus	sp27	616	5303	Endemic
			4992	
		Mammals		
Lynx pardinus	sp28	354	5565	Endemic
Talpa occidentalis	sp29	1056	4863	Endemic
Galemys pyrenaicus	sp30	706	5213	Endemic
Sorez granarius	sp31	171	5748	Endemic
Capra pyrenaica	sp32	648	5271	Endemic
Microtus cabrerae	sp33	361	5558	Endemic
Lepus castroviejoi	sp34	69	5850	Endemic

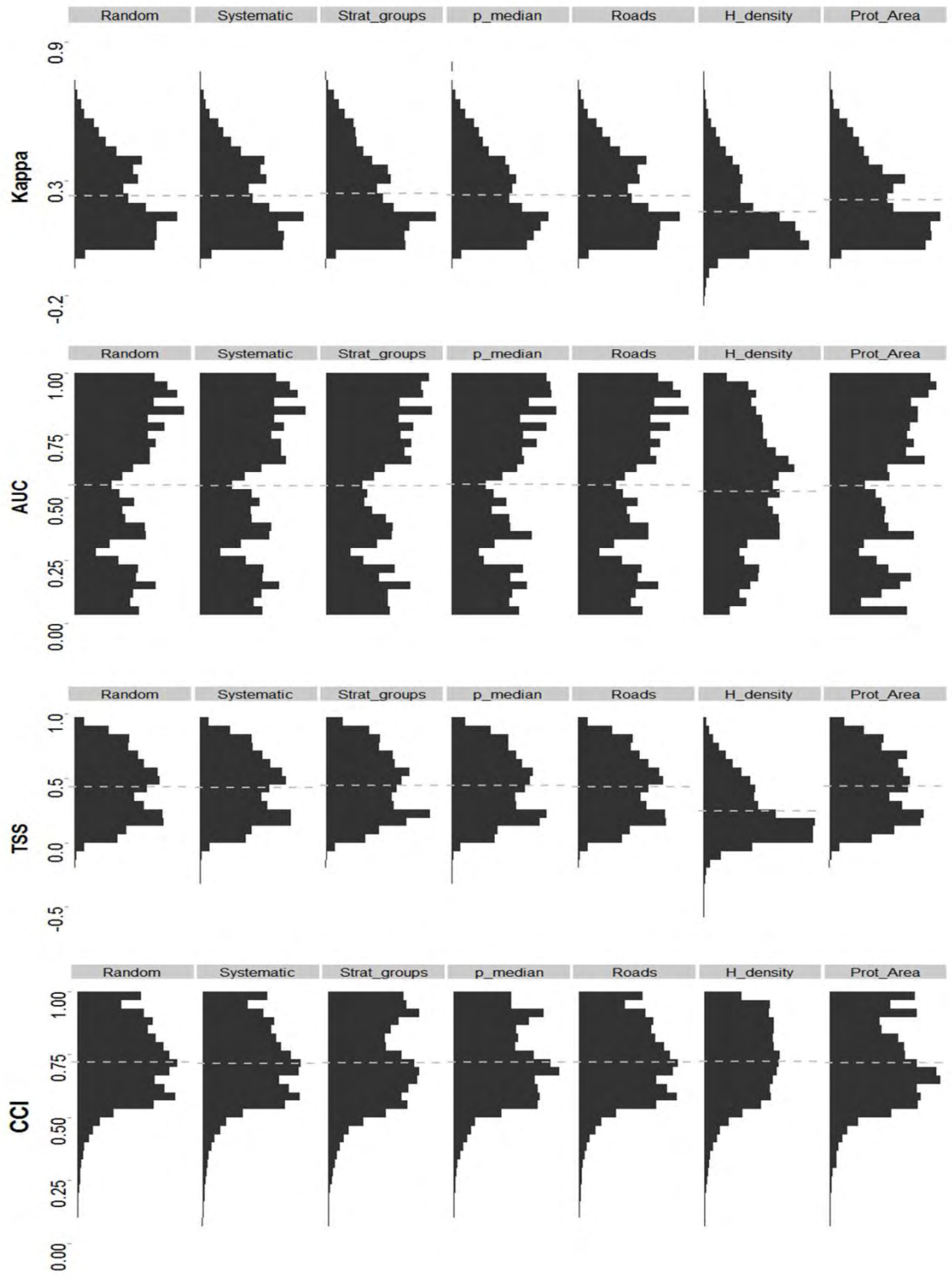


Fig. S1. Performance of SDM predictions from datasets based on different survey designs, according to Kappa, AUC,TSS and CCI validation metrics. H_density= bias by human density, p-median = Stratified by p-median, Prot. Area = bias by protected areas, Strat_groups = stratified by groups. The histograms represent the frequency in each class and the red lines indicate the mean.

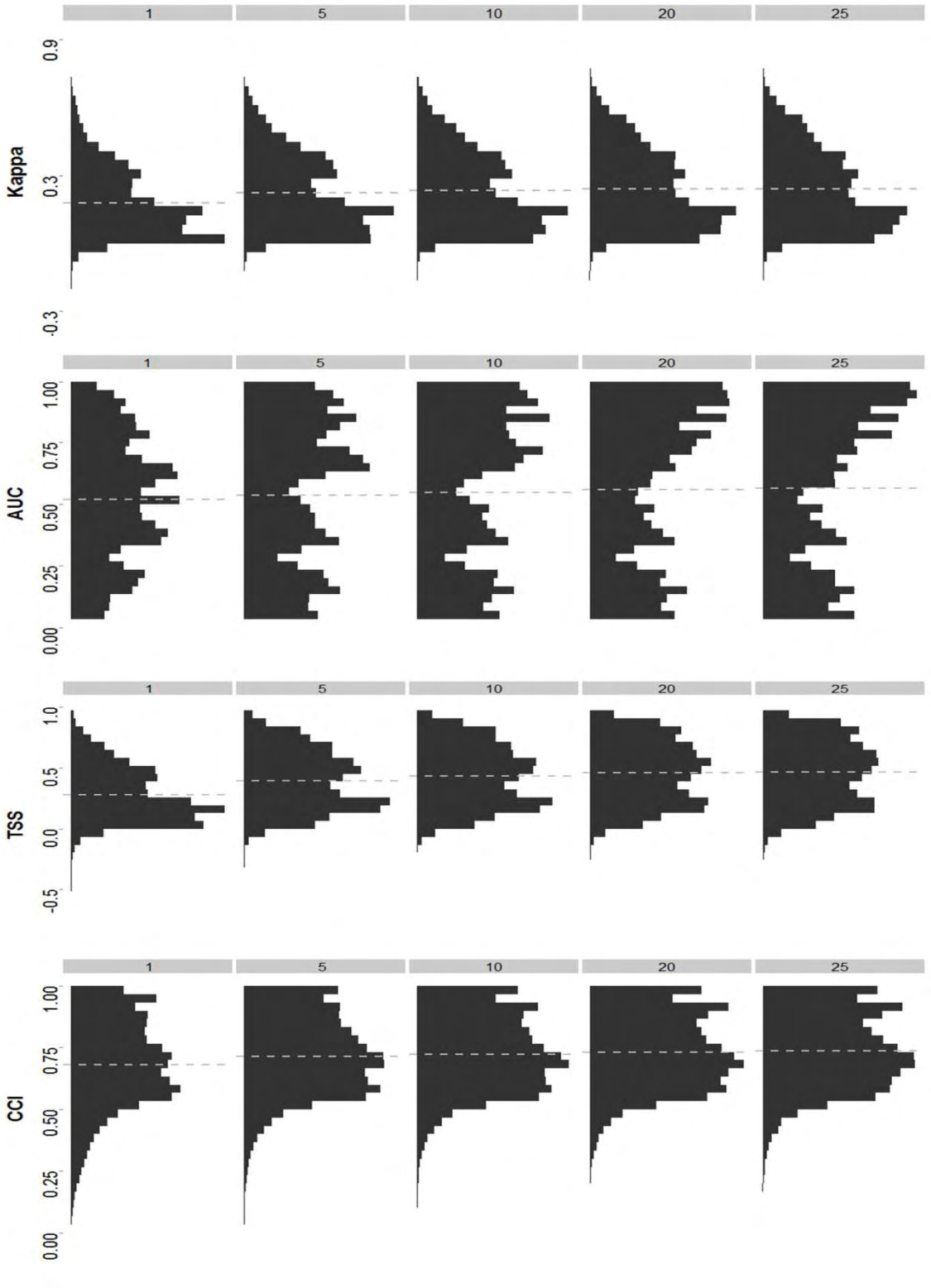


Fig. S2: Performance of prediction from datasets of five sample sizes (numbers represent the percentage of cells selected from the whole studied area), measured according to Kappa, AUC,TSS and CCI validation metrics. The histograms represent the frequency in each class and the red lines indicate thee mean.

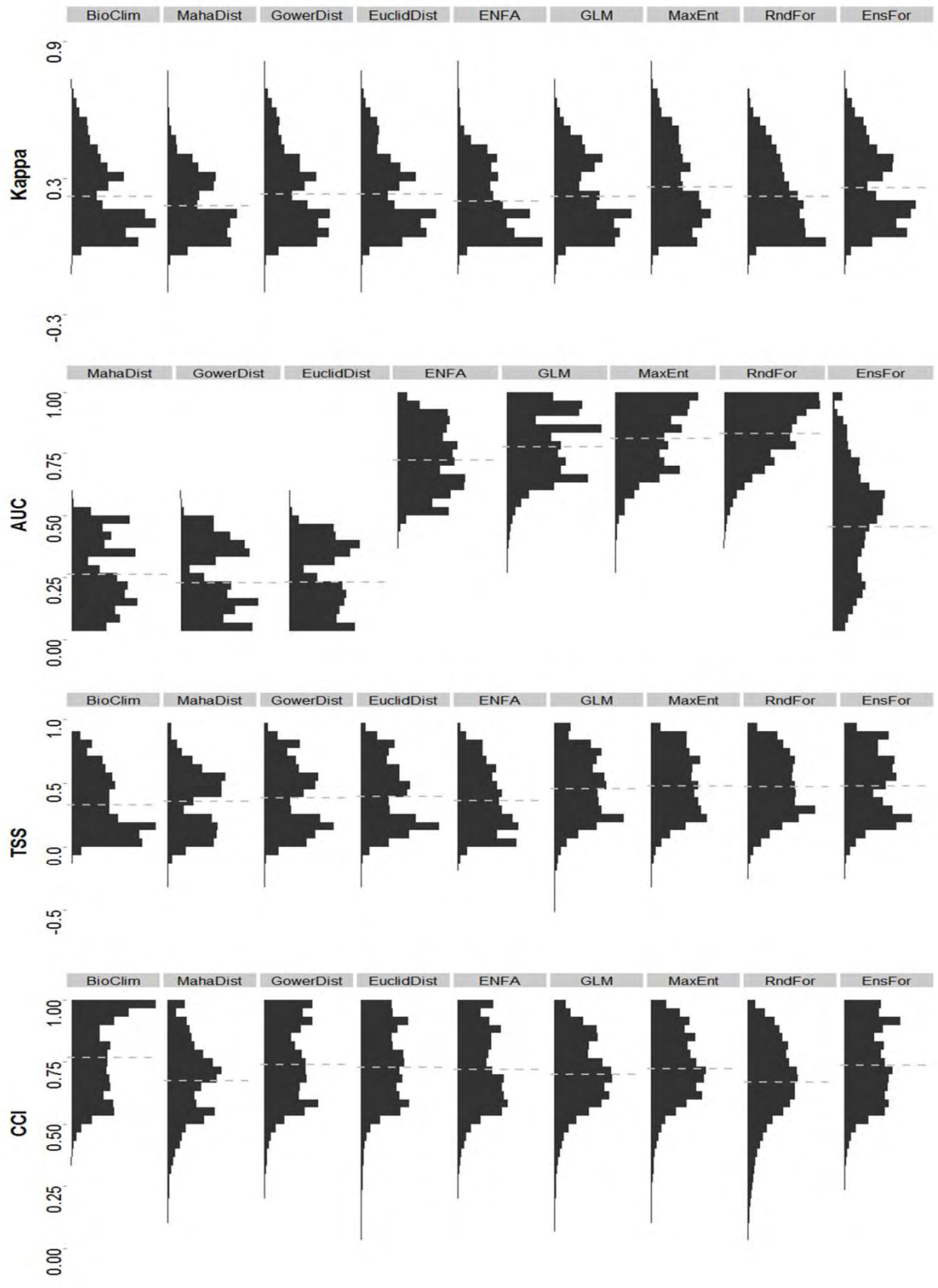


Fig. S3. Predictive performance of SDM techniques according to Kappa, AUC, TSS and CCI validation metrics. The histograms represent the frequency in each class and the red lines indicate the mean.

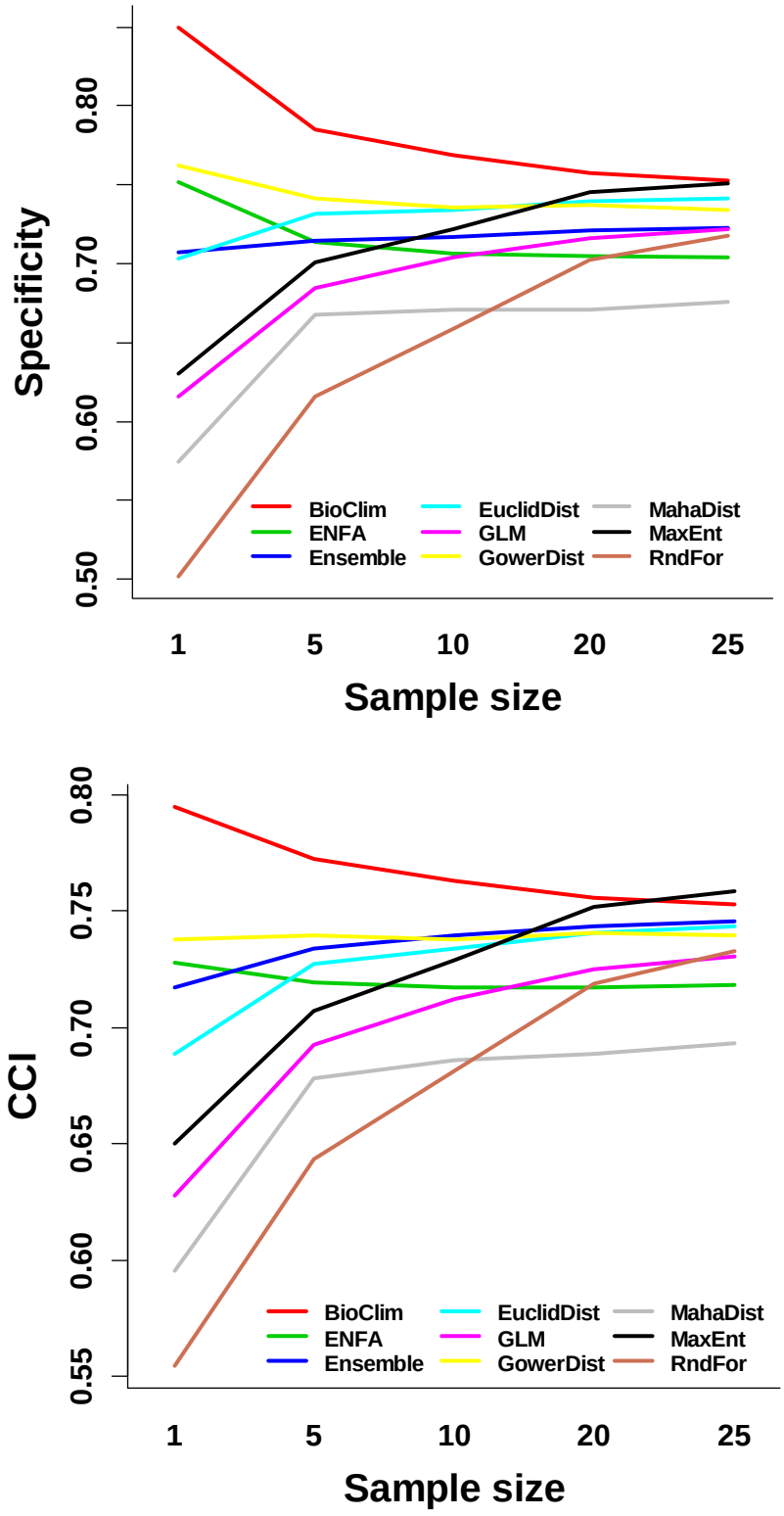


Fig. S4. Interaction between sample size and model for specificity and CCI metrics.

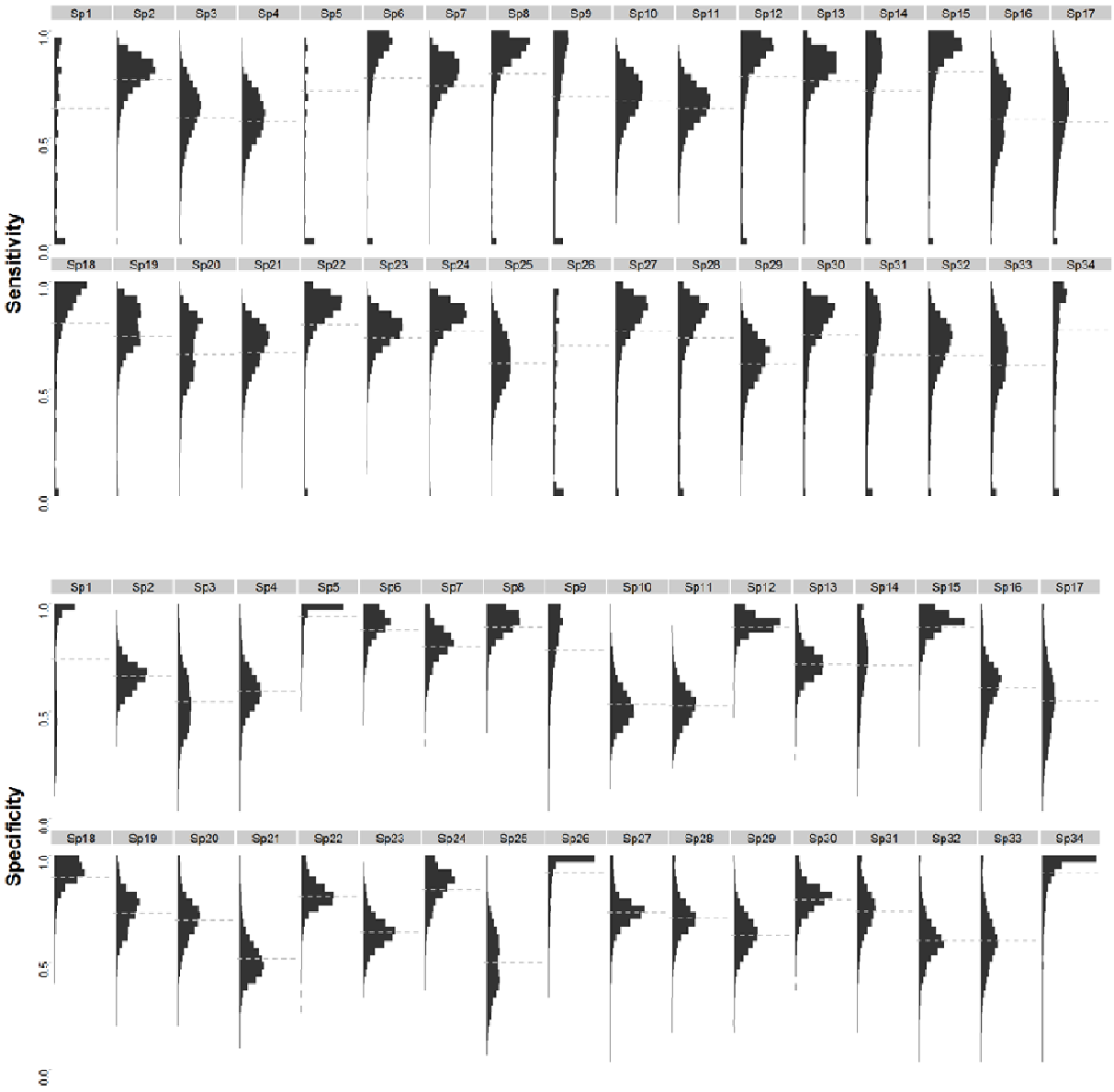
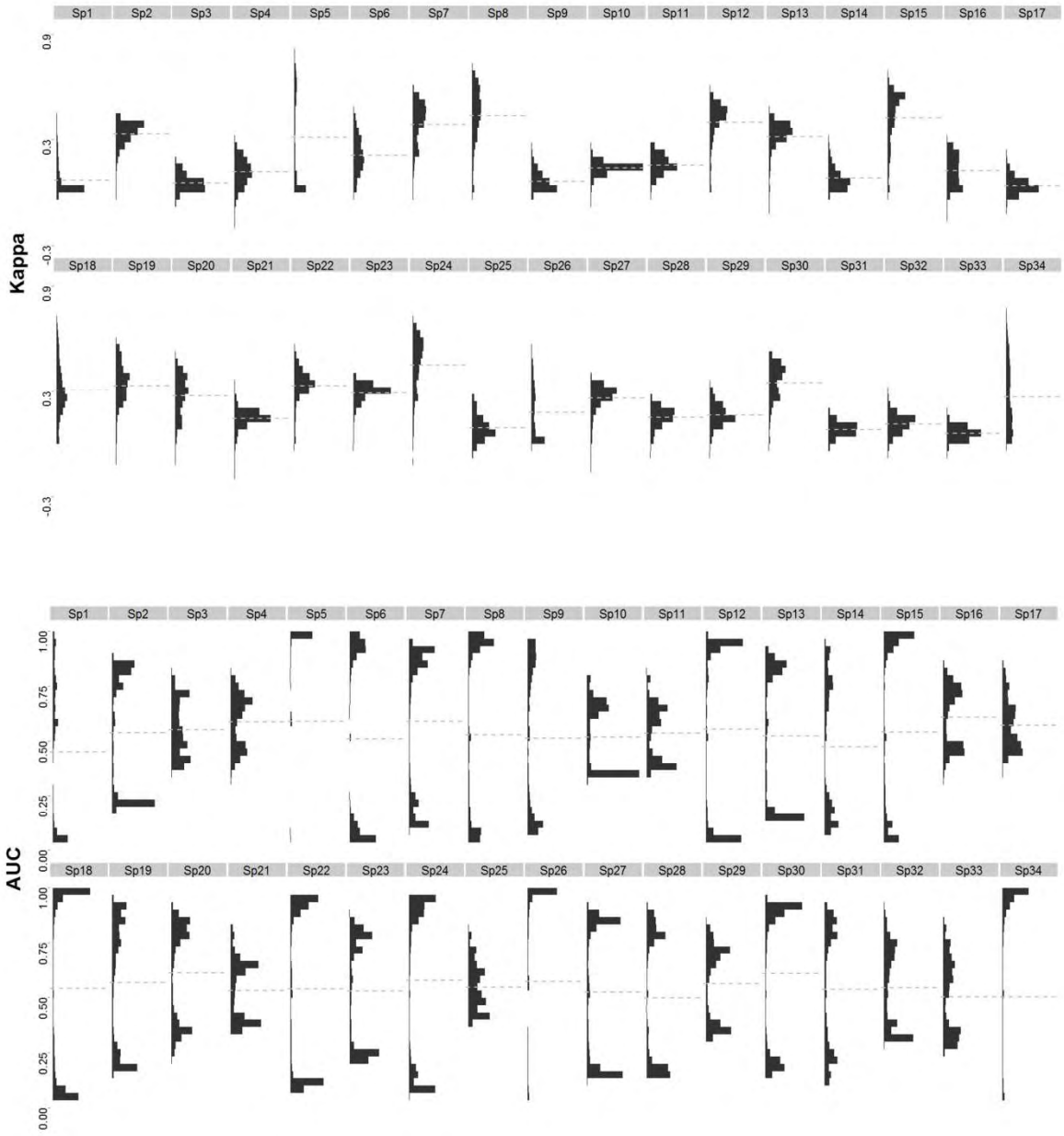
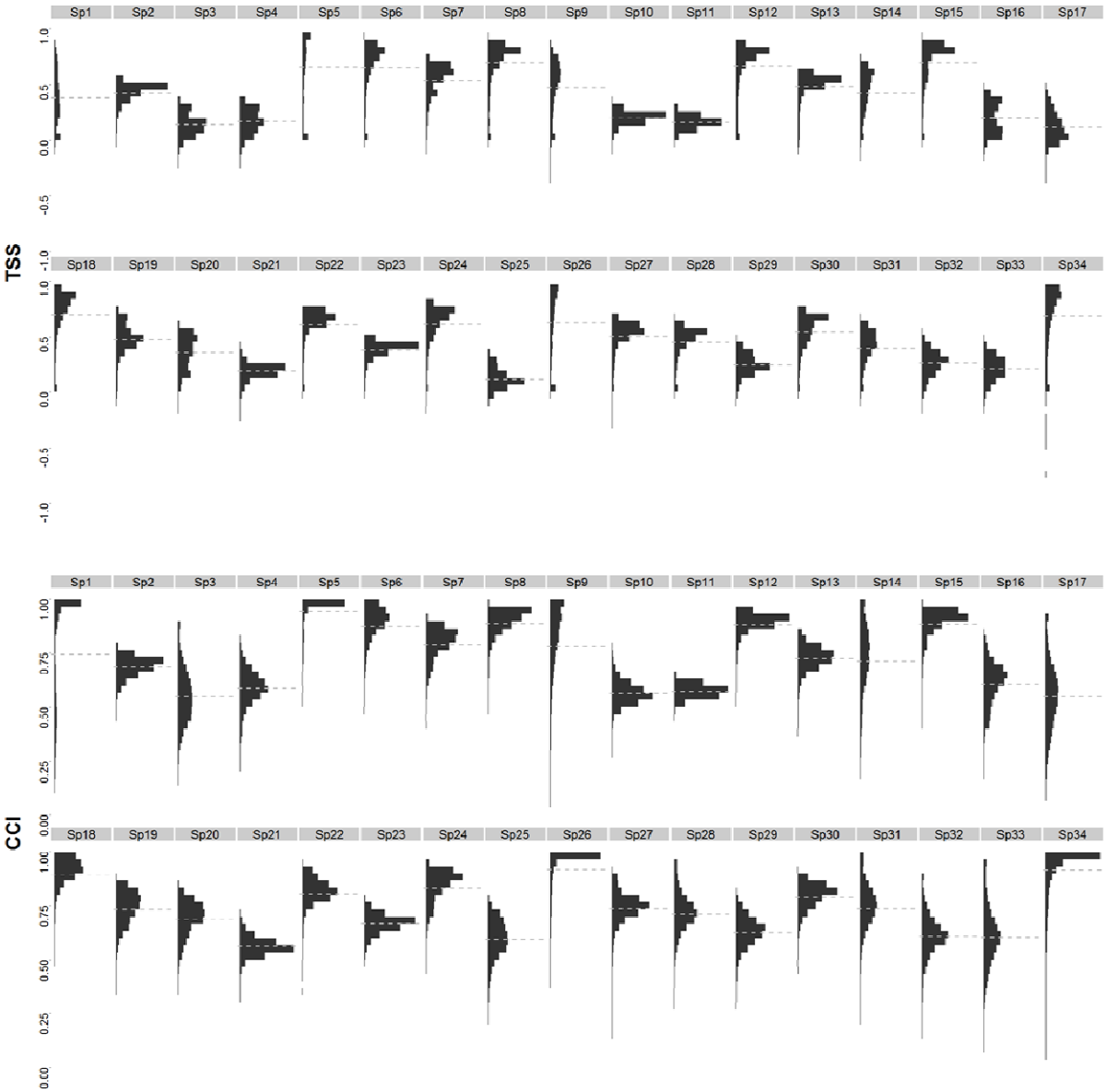


Fig. S5. SDM performance for all studied species according to Kappa, AUC, TSS and CCI validation metrics. The histograms represent the frequency in each class and the red lines indicate the mean. The species code can be found on supplementary appendix S1.

Continuation Figure S5.



Continuation Figure S5.


TSS
Sp1 Sp2 Sp3 Sp4 Sp5 Sp6 Sp7 Sp8 Sp9 Sp10 Sp11 Sp12 Sp13 Sp14 Sp15 Sp16 Sp17
1.0
0.5
0.0
-0.5
-1.0
Sp18 Sp19 Sp20 Sp21 Sp22 Sp23 Sp24 Sp25 Sp26 Sp27 Sp28 Sp29 Sp30 Sp31 Sp32 Sp33 Sp34
CCI
Sp1 Sp2 Sp3 Sp4 Sp5 Sp6 Sp7 Sp8 Sp9 Sp10 Sp11 Sp12 Sp13 Sp14 Sp15 Sp16 Sp17
1.00
0.75
0.50
0.25
0.00
Sp18 Sp19 Sp20 Sp21 Sp22 Sp23 Sp24 Sp25 Sp26 Sp27 Sp28 Sp29 Sp30 Sp31 Sp32 Sp33 Sp34

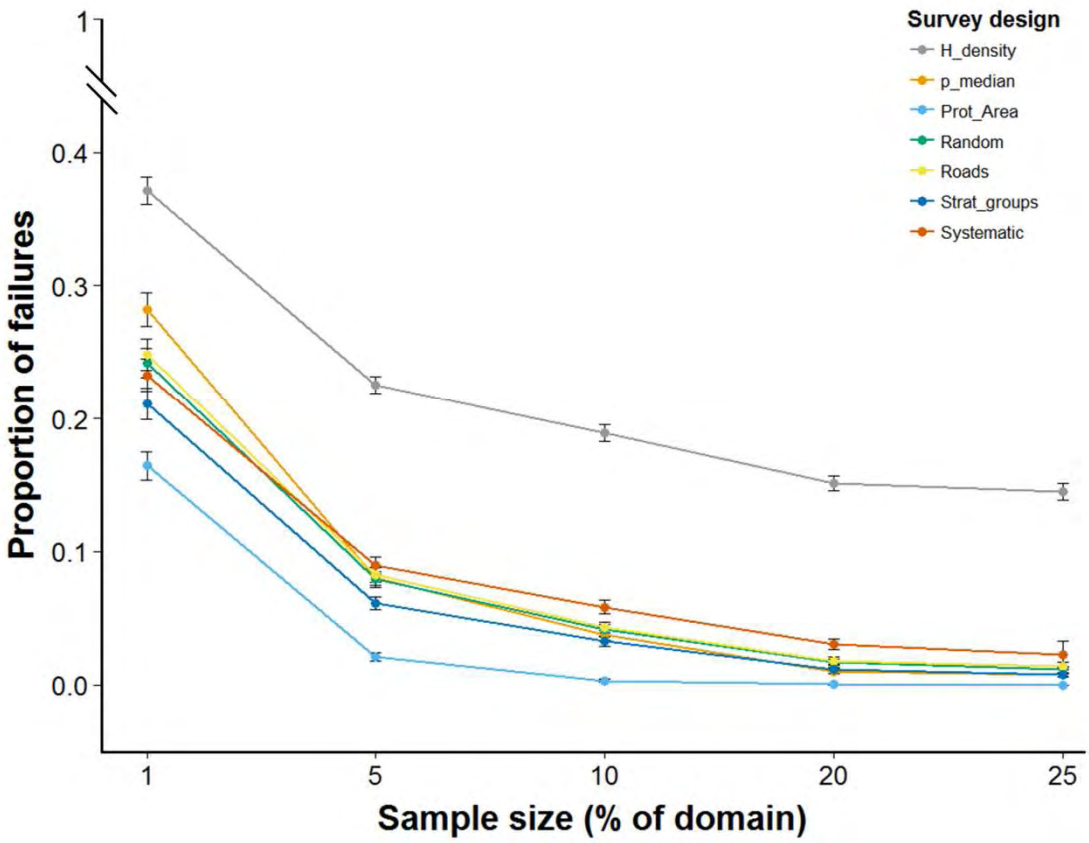


Fig. S6. Number of times that a survey design fails in get sufficient occurrences to generate models in each modelling technique. . Each estimated number is a mean based on 400 values (50 runs x 8 modelling technique) of missing species model.

Chapter 2

Ecological characteristics as source of uncertainty in species distribution models

Ecological characteristics as source of uncertainty in species distribution models

Geiziane Tessarolo, Thiago Fernando Rangel, Jorge M. Lobo & Joaquín Hortal

Abstract

Species distribution models (SDM) are widely used to describe species distributions for many purposes, including conservation actions. However, besides their potential for theoretical and applied purposes, SDMs are susceptible to various sources of uncertainty. The ecological characteristics of the species and their distributions can affect SDM results, often generating critical errors in the so-generated species distribution maps. In this work we assess the influence of ecological traits, the characteristics of species distributions and data quality on SDM performance for the distributions of dung beetle species in a well-known region of high environmental variability. In particular, we evaluate the impact of environmental coverage, a characteristic of the data that can covary with several ecological and distributional traits. To do this, we relate species traits and environmental completeness to variations on the performance of eight SDM techniques, based on the validation performed on an independent dataset. Marginality and environmental completeness were the variables that most influenced SDM results; here, species presenting higher marginality, environmental completeness and smaller values of relative occurrence area (ROA) were associated with better models. Our results evidence that to improve SDM performance, additional survey effort is needed for species that are likely to have their environmental niche less covered by the data. This highlights the importance of taking species and data characteristics into account when modelling and comparing large groups of species using SDM.

Keywords: Ecological traits, geographical traits, species distribution modelling, environmental completeness, marginality, ROA, Scarabaeoidea dung beetles, Iberian Peninsula

Introduction

Species distribution models (herein SDM) are a well-established set of techniques that model the potential distributions of species based on their known occurrences and environmental data. Despite being widely used (see Lobo et al., 2010), SDM results vary according to various sources and limitations, and can impose severe restrictions to their utility (Rocchini et al., 2011). The particular characteristics of each species are among the most important sources of uncertainty in SDM (Venier et al., 1999). This is due to the fact that the species differ in the principal characteristics that influence their distribution (Kadmon et al., 2003), which in turn causes variation in the detection of relation occurrences/absences and environmental variables across species (Brotons et al., 2004; Thuiller et al., 2010). Additionally, many species traits can determine how easy or not is to collect data on their occurrences (Boone & Krohn, 1999), which can therefore compromise the quality of the data (Ladle & Hortal, 2013).

Identify which characteristics affect SDM performance and determine the intensity of this influence will allow handpicking the species that are more suitable for SDM applications in specific situations (Guisan et al., 2007), allowing also to discard those with characteristics that may compromise such performance. Several species' characteristics are known to affect SDM results, such as body size (McPherson & Jetz, 2007a), life span (Hanspach et al., 2010), growth rate (Guisan et al., 2007), habitat specialization (Chefaoui et al., 2011), distribution range (Segurado & Araújo, 2004; Hernandez et al., 2006; Chefaoui et al., 2011), dispersal (Pöyry et al., 2008) and rarity (Franklin et al., 2009). Further, the particular characteristics of the species' distributions and their marginality with respect to the environmental conditions of

the study region can determine the accuracy of SDM predictions (Jiménez-Valverde et al., 2008; Lobo, 2008; Anderson & Raza, 2010). The use of the term trait can be equivocal, for it can refer to either characteristics of the individuals related with their ecological role and performance, or to general characteristics of the species such as its geographical distribution. For clarity, we will follow the spirit of (Violle et al., 2007; Díaz et al., 2013), and will call ecological traits or simply traits to the characteristics of the individuals that have ecological significance, geographic characteristics to the attributes of the distribution of the species, and data characteristics to those of the data itself.

In general, the results of former studies suggest that higher mobility and niche breadth, large sizes of the area occupied by the species (relative to the total area of the studied region) and the prevalence of occurrences in the data are negatively correlated with SDM performance (Berg et al., 2004; Hernandez et al., 2006; McPherson & Jetz, 2007a; Newbold et al., 2009). On the contrary, the distributions of species with higher rarity and wider life span and habitat specialization are often predicted better (Carrascal et al., 2006; Franklin et al., 2009; Mateo et al., 2010; Syphard & Franklin, 2010; Chefaoui et al., 2011). However, these results are conflicting and different studies reach different conclusions about the same traits and characteristics.

Variations in SDM performance have been attributed to the differences in the spatial resolutions of the analyses, the particular SDM technique used and the evaluation metric used to assess model performance (Segurado & Araújo, 2004; McPherson & Jetz, 2007b; Marmion et al., 2009; Santika, 2011; Hijmans, 2012). However, many of these studies use a small number of SDM techniques (Segurado & Araújo, 2004; McPherson & Jetz, 2007a; Pöyry et al., 2008; Marmion et al., 2009; Hanspach et al., 2010; Santika, 2011), or techniques that use only presence records (Segurado & Araújo, 2004; Guisan et al., 2007). Furthermore, they are often evaluated using AUC (the metric most commonly used to evaluate SDM performance),

which is known to present fatal flaws that would prevent its use for this task, and particularly to compare the results across a number of species with different data and geographic characteristics (Lobo et al., 2007b; Peterson et al., 2008; Hijmans, 2012; Jiménez-Valverde, 2012). Here note that data characteristics typically vary widely between species pertaining to the same dataset, which are often analyzed altogether with authomatized SDM protocols.

In particular, the completeness of the environmental response of the species that is covered by the data (i.e., environmental coverage) may vary within an homogeneous region and group of species (see Hortal et al., 2008), thus affecting the results of SDMs based on the same technique and overall dataset. Such environmental coverage is highly dependent on both sampling effort and survey success, and can in turn be influenced by some traits like showiness (i.e. conspicuousness) or habitat preferences (Pearce et al., 2001; Reese et al., 2005; Guisan et al., 2006). Many studies have stressed the importance of environmental coverage for SDM results (Kadmon et al., 2004; Thuiller et al., 2004; Hortal, 2008; Lobo et al., 2010). However, its effects on SDM performance across different species have not been formally evaluated yet.

In this work we evaluate the influence of a number of traits and species' characteristics on SDM performance. More specifically, our aim is to identify with traits have more impact on SDM results, and evaluate the influence of environmental coverage on the precision of these results. To do this, we use seven SDM techniques and the consensus among them to generate distribution maps for 57 dung beetle species (Coleoptera, Scarabaeoidea) that vary in their ecological traits and geographic and data characteristics in the Madrid region (central Spain), assessing their performance with six evaluation metrics.

Data and Methods

Species data

We used data on the distribution of Scarabaeoidea dung beetles in the Madrid region, central Spain. Data come from SCAMAD, a biological database that compiles all distributional information for dung beetles in Madrid and its surroundings that is available at natural history collections from museums, universities and private owners, as well as in the specialized literature (Hortal & Lobo, 2005; Hortal et al., 2006, 2008; Lobo & Hortal, 2006). We used two different versions of this database, to capture two different levels of knowledge. SCAMAD 1.0 contains data on 92.746 specimens of grouped in 5.364 records and 130 species, recorded from the year 1808 until 1998. This database was updated after an extensive survey conducted during 1999–2003, designed to cover all the environmental and spatial variability of the region (Hortal & Lobo, 2005). The current version of the database, SCAMAD 2.1, contains 7.116 records from 145.839 specimens and 133 species. These data are divided in 108 UTM cells of 10 x 10 km that conform the grid squares that had more than 5% of their territory in the Madrid administrative region.

Environmental data

We used two kinds of variables that have been previously related to the diversity of Iberian dung beetles (Hortal et al., 2001; Lobo & Martín-piera, 2002; Chefaoui et al., 2005): climate (mean, maximum and minimum temperature, total annual precipitation and total precipitation in summer) and substrate (Acid rocks, Acid deposits, basic rocks and deposits, Poorly developed soils, Soils in an early development stage, Soils with accumulation by illuviation, Soils with organic matter, Predominantly acid soils and Soils with accumulation of bases). The climatic data come from (Hortal & Lobo, 2005), the soil structure and composition are from (FAO, 1998) and bedrock data come from (ITGE, 1988). Substrate variables represent

the proportion of area of each grid cell covered by each category. These variables were re-scaled to UTM 10 x 10 km. See Chefaoui et al. (2005) and Hortal & Lobo (2005) for additional details on the origin of these variables. To reduce the collinearity among variables a principal components analysis (PCA) was used for each kind of variables. In total, four environmental factors were extracted, two related to climate (hereafter CF1 and CF2) and two to substrate (hereafter SF1 and SF2), that were chosen according to a broken-stick criterion. See Hortal et al. (2008) for more details about the PCA results. These four environmental factors were used to generate all species distribution models.

Species distribution modelling

Models of the distributions of all species were generated using data from SCAMAD 1.0. All species for which there were not enough occurrences to be used in SDM, or lacked information about the characteristics evaluated here (see below) were removed from the analyses. A total of 57 species were modelled (see appendix for a list of species). To reach as much as possible independence between the data used for calibration and evaluation, we divided the 108 cells in 5 groups selected randomly. Each one of these groups of cells was used to validate the results of SDMs calibrated with the remaining cells (~80% of the territory). Here, the data used for calibration came from SCAMAD 1.0, and that used for validation from SCAMAD 2.1. Thus, no data were used simultaneously for calibration and validation. And importantly, several cells present new information coming from additional surveys when used for validation. This procedure was repeated 10 times for each species.

The predictions about the distributions of species were generated using seven SDM techniques and an ensemble of them to make combined consensus prediction (Araújo & New, 2007). SDM techniques were implemented using the BIOENSEMBLES platform for computer-intensive ensemble forecasting (Diniz-Filho et al., 2009; Rangel et al., 2009), and include: Bioclim (Busby, 1986); Mahalanobis distance (Farber & Kadmon, 2003);

Generalized Linear Models (GLM; McCullagh & Nelder, 1989); Maximum Entropy Modelling (MaxEnt; Phillips et al., 2006); Ecological-Niche Factor Analysis (ENFA; Hirzel et al., 2002); Random Forest (Breiman, 2001); and Neural Networks (Nnet). The consensuses were generated by a mean of all single-models. The predictive performance of models was assessed using six different metrics: sensitivity; specificity; kappa (Cohen, 1960); true skill statistic (TSS; Allouche et al., 2006); percentage of correctly classified instances (CCI; Fielding & Bell, 1997); and Area Under the ROC Curve (AUC; Fielding & Bell, 1997).

Species' attributes

We gathered information about ten species' attributes related to their distribution and ecology, including three functional traits – body size (measured as body length), nesting behavior and food preferences (i.e. main food source); five geographic characteristics – rarity, range size, relative occurrence area (ROA), marginality and niche breadth, and; and two data characteristics – prevalence and environmental niche completeness. Information on body size, food preferences and nesting behaviour were collected from the specialized literature (Baraud, 1992; Martín-Piera, 2000) and complemented by consultation to specialists. The nesting behaviour is a key aspect of dung beetle life history (Halffter & Edmonds, 1982), and was classified in three categories: Paracoprid, Telecoprid and Endocoprid, which correspond respectively to nests buried in the ground beneath the excrement, buried after transportation (i.e. rolling behavior) and placed within the dung pat. Food preferences were classified in two major categories: coprophagous and detritivore.

Rarity was measured as the number of individuals of each species in SCAMAD 2.1, divided by the total number of individuals in this database. Range size is the number of 10 km grid cells occupied by the species in the study area. ROA is the ratio between the area occupied by the species and the total area of the study region (Lobo, 2008). Here, the

distribution of the species was estimated as the area of the minimum convex polygon (MPC, the smallest polygon in which there is no internal angle is greater than 180 degrees), generated with all the occurrences of each species (see Chefaoui et al., 2011). Marginality measures the difference between the mean of the environmental conditions in the study area and the optimum environmental conditions of the species and was calculated using Ecological Niche Factor Analysis (ENFA; Hirzel et al., 2002) based on the four environmental factors extracted from PCA (CF1,CF2, SF1,SF2). Niche breadth was measured as the volume of the smallest convex hull polygon (i.e. the minimal hyper-volume that encloses a set of input points in N-dimensional space) generated from all the occurrences of the species in the space defined by the four environmental factors, calculated using the geometry R package.

Prevalence is the ratio between the numbers of presence and absence cases in the training dataset. Environmental Niche Completeness is a measure of the proportion of the whole response of the species to the environment that is covered by the data (see Kadmon et al., 2004; Hortal et al., 2008). Here we assume that the surveys that led to SCAMAD 2.1, which came from a standardized protocol that allowed recovering data from the most important environmental gradients in the region (Hortal & Lobo, 2005), recovered the whole responses of all species to the environment (Hortal et al., 2008). Therefore, niche completeness was calculated by comparing the extents of the niche described by both SCAMAD 1.0 and 2.1 versions. To do this, we estimated the percentage of coverage that the data on the species in SCAMAD 1.0 offers from the the total extent of the occurrences from SCAMAD 2.1 in each environmental factor, and then calculated the total niche coverage from the relative ratios in these four factors (see Hortal et al., 2008 for more details).

Statistical analyses

The influence of species' attributes on SDM predictive performance was assessed by a partial least squares regression (PLS, Wold 1975). PLS is an extension of multiple linear regressions

where a linear model specifies the relationship between one or several response variables and a set of predictors (see Carrascal et al., 2009 for an discussion of its potential for ecological research). To establish these associations, a set of factors that maximizes the explanation of variance in the dependent variables is extracted from the predictors. In cases with multiple response variables (like in this study) PLS creates a series of latent factors that act as synthetic response variables. Due to the high collinearity between several predictors (ROA, prevalence, niche breadth and range size) in the PLS analysis we used only the two that were less correlated: Prevalence and ROA. Thus, the PLS was built in using body size, nesting behaviour, food preferences, marginality, rarity, prevalence, ROA, completeness and SDM technique as predictor variables, and the six assessment metrics (sensitivity, specificity, kappa, TSS, CCI, and AUC) as response variables. We also conducte separate regression tree analyses to evaluate the effects of all species' attributes on each one of the performance assessment metrics.

Results

The first two significant components ($p<0.05$) in the PLS analysis accounted for 33.16% of the variation in SDM predictive performance (Tables 1 and 2). The first component shows that species that with higher values of marginality and environmental niche completeness and lower values of ROA are predicted better. The second component also shows that species with higher niche completeness are predicted better, and in addition that Bioclim and ENFA SDM techniques are responsible for the worst predictions (Table 1). In general, PLS results show that, of all variables tested, the most influent on SDM results are niche completeness and marginality (Fig. 1a). Among SDM techniques, model ensembles, GLM and MaxEnt had the less influence on predictive performance whereas Random Forest and Bioclim were the most important (Fig. 1b), evidencing that in this particular dataset Random Forests improve model performance, while Bioclim decreases it. The regression tree analyses shows that niche

completeness has a strong influence on sensitivity (i.e. correct prediction of presences) TSS and AUC, the choice of the SDM technique exerted the highest influence on specificity (i.e. correct prediction of absences) and CCI (Table 2).

Table 1. Results of the partial least squares regression analysis, relating the contribution of each response variable (performance metrics) and predictors (species' attributes and SDM techniques) on the components.

Predictors	**1st Component**	**2nd Component**
Performance metrics		
Sensitivity	0.209798	0.559743
Specificity	0.489709	-0.404549
Kappa	0.291573	0.313540
TSS	0.433290	0.377732
CCI	0.505642	-0.373306
AUC	0.433290	0.377732
Species' attributes		
Body size	-0.032582	0.115369
Nesting: Endocoprid	0.069878	-0.035632
Nesting: Telecoprid	-0.037783	0.075875
Nesting: Paracoprid	-0.052602	-0.000809
Feeding: Coprophagous	0.114945	-0.016504
Feeding: Detritivore	-0.114945	0.016504
Rarity	0.096853	0.109634
Marginality	0.694205	0.061169
ROA	-0.339616	0.105748
Prevalence	-0.020628	0.223760
Niche completeness	0.394048	0.686245
SDM techniques		
Bioclim	0.270367	-0.369949
ENFA	0.155274	-0.339844
MahaDist	-0.035074	0.218525
Ensemble	0.000421	0.095606
NNet	-0.116876	0.067583
MaxEnt	0.025842	0.000998
GLM	-0.004177	0.002405
RndFor	-0.297174	0.342562
Explained variability	20.04%	13.12%

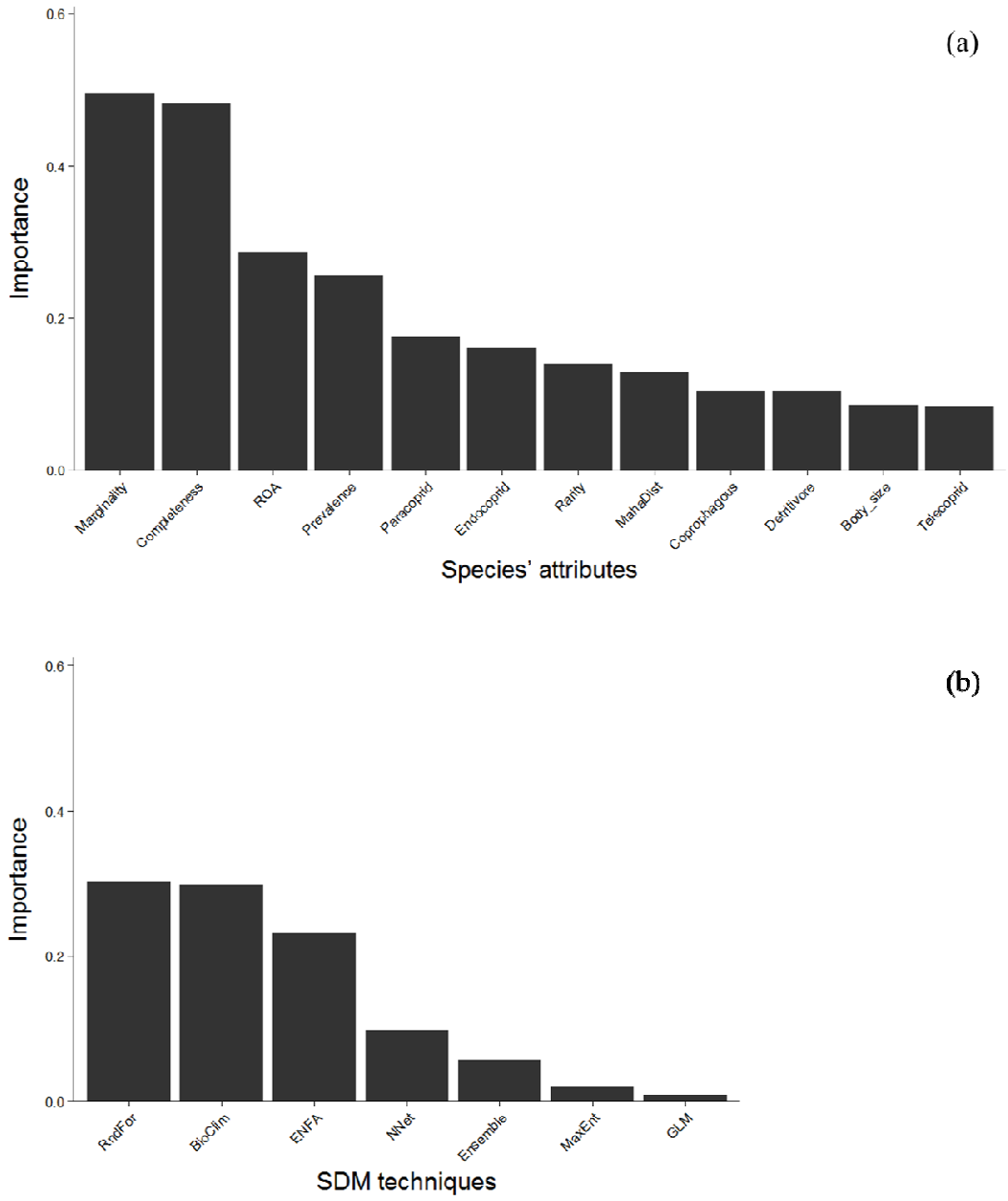


Fig 1. Variable importance from PLS analysis for the variability in the performance assessment metrics, according to both (a) species' attributes and (b) SDM techniques. Note that both types of factors have been analyzed altogether in a single PLS analysis, and therefore their importance values can be directly compared.

Table 2. Results of the regression tree analysis, depicting the importance of species' attributes and SDM technique for each performance assessment metric. The asterisk highlights the most important variable for each metric.

Metrics	Food preferences	Nesting behaviour	Body size	Rarity	Marginality	ROA	Prevalence	Niche completeness	Range size	Niche breadth	SDM technique
Sensitivity	-	-	47.15	32.75	86.20	74.79	19.05	194.75*	24.23	75.54	39.39
Specificity	-	-	3.55	7.20	25.49	22.08	15.05	6.23	19.60	28.98	49.28*
Kappa	0.11	0.05	0.82	3.24*	2.70	2.42	1.94	2.87	1.93	2.40	-
TSS	-	24.36	65.53	52.98	161.68	118.44	42.01	210.15*	47.88	121.11	-
CCI	-	-	3.54	6.80	26.94	23.78	14.12	2.59	19.56	32.87	43.38*
AUC	-	6.09	16.38	13.25	40.42	29.61	10.50	52.54*	11.97	30.28	-

Discussion

Our results show that the characteristics of both species and data affect SDM perfomance. Data and geographic characteristics exert the highest influence on model predictive performance, as indicated by the six evaluation metrics analyzes, although the SDM technique has also important influences for metrics more sensible to the correct classification of absences. These differences were expected, given that the different evaluation metrics assess different aspects of the model showing variations on importance of each factor (Mouton et al., 2010). But in general, the distributions of species for which the coverage of their environmental niche was incompletely covered by data are predicted worst, and those whose distributions are more environmentally marginal in the region are predicted better.

Not surprisingly, environmental niche completeness has a major influence on SDM results. The effectiveness of species distribution models relies on their capacity of providing an adequate representation of the responses of the species to the environment. The less is the environmental niche of a species represented in the occurrences, the less accurate is the shape of the response curve that will be captured by any SDM technique (Vaughan & Ormerod, 2003). Distortions on these response curves reduce the possible combinations of

environmental conditions on the model calibration process, diminishing their predictability and utility for present and future projections (Thuiller et al., 2004). The lower the environmental completeness, the most likely it is that a real presence is included within the selection of absences or pseudo-absences, thus reducing SDM performance.

Factors affecting SDM performance

A limited coverage of the species' realized response to environmental gradients could arise by collection bias, lack of sampling effort, or uneven survey designs. Although niche completeness is a characteristic of the data, it can be related to several traits such as conspicuousness or habitat preferences. Conspicuous species and those inhabiting easily accessible and/or interesting (to researchers) habitats are more likely to be recorded than those that are difficult to detect, or live in habitats that receive less attraction and/or in remote areas (Seoane et al., 2005; Romo et al., 2006; Lobo et al., 2007a). Or, simply, than species pertaining to taxonomic groups that do not receive the attention of specialists, providing a further interaction between the Linnean and Wallacean shortfalls (Whittaker et al., 2005; Rocchini et al., 2011). But as a general rule, niche completeness can be extremely sensible to uneven survey designs (Hortal et al., 2008) and the use of inadequate sampling techniques. Most data used for Species Distribution Models comes from biodiversity databases and distributional Atlases that gather data resulting from surveys designed in an informal way and conducted using a range of different methodologies (see Hortal et al., 2007; Hortal, 2008; Pineda & Lobo, 2009) it follows that it is highly likely that most data used in massive SDM exercises suffers for limited niche completeness for a significant number of species; almost half of the species had less than 100% of their response to the environment covered by data in Hortal et al., (2008). Further, given that the more different are sampling methodologies and/or survey designs, the higher the differences in niche completeness will be, the unevenness in

SDM performance can be more important when studying species from different taxa and/or regions.

The accuracy of species distribution models increases for species that are more marginal and with smaller ROA. Such negative effects of ROA on SDM predictions supports the assumption that the larger the geographical range of a species, the less accurate its predicted distribution will be. Arguably, this is due to the difficulty of identifying the actual relationship between occurrences and environmental gradients when species have wide climatic preferences. Being SDM essentially classification techniques, this makes difficult to distinguish suitable from unsuitable habitats (Syphard & Franklin, 2009; Grenouillet et al., 2011). In these cases, it seems necessary to include a selection of absence data that enhances the classification capacity of SDM techniques (Lobo et al., 2010). Besides, the populations of species with large geographical ranges (and less environmentally marginal) are more likely to show local or regional adaptations leading to differences in habitat preferences. Despite being part of the fundamental niche at the species level, modelling the distributions of such species with all their populations altogether can neglect the effects of local adaptation and consequently overestimate the actual distribution of the species, decreasing the predictive ability of the so-constructed models (Stockwell & Peterson, 2002; Hernandez et al., 2006).

We must note that several studies have found no relationship between marginality and SDM performance (Pöyry et al., 2008; Newbold et al., 2009). Here, Chefaoui et al., (2011) argue that the influence in model performance attributed to some ecological traits or geographical characteristics can in fact be related to ROA and data characteristics such as prevalence or sample size. As many evaluation metrics are known to be sensitive to data characteristics (Manel et al., 2001; McPherson et al., 2004; Allouche et al., 2006; Santika, 2011), the influence of these species traits could in fact be a statistical artifact. To account for this problem, here we used six evaluation metrics, some of them known to be sensible to data

characteristics, but others known to be insensitive to such issue, Given that we evaluated all of them together with the same weight in a single PLS model, we expect that the impact of eventual statistical artifacts due to data characteristic on the results would be minimal. In any case, given that marginality is the variable with the highest importance in our PLS analyses, we believe that, at least in this case, it truly has an impact on SDM performance.

None of the three species' ecological traits that we evaluated had a significant influence on SDM performance. This is likely to be related with the lack of relationship of these traits with the response of species to environmental gradients in a relatively small region such as Madrid. Although body size, nesting behavior and food preferences are well known to be associated to the ecological performance of dung beetles (Scholtz et al., 2009), and to their success in colonizing different environments, it is likely that their effects occur over larger geographical scales. At this finer geographical scale the traits exerting the most influence should be those reflecting the rate of climatic niche evolution (Kellermann et al., 2009; Lavergne et al., 2013) or to the stages of environmental equilibrium (Smith et al., 2013). That is, traits related to ecophysiological adaptations and specific habitat choices, so trait differences may reflect differences in the degree of fit of each species to the actual assumptions of SDM. Dung beetles are known to have a complex thermoregulatory behavior (Verdú & Lobo, 2008). Given the potential importance that the plasticity provided by such behaviour could have on SDM predictive ability (see above), it is likely that traits related to it (or to niche variability) can be more adequate to determine which species will show distributions that are more difficult to model accurately.

The modelling technique appears to play a secondary role on SDM performance. Only Bioclim and ENFA having a major influence on the second PLS component, in which they have a negative impact on SDM predictions. Despite being the technique ranking first on variable importance, Random Forests bear no significant impact on the two first PLS

components, and other SDM techniques seem to be even less important. This result is in agreement to former studies (Syphard & Franklin, 2010; Chefaoui et al., 2011; Dobrowski et al., 2011; Tessarolo et al., 2014) and supports the idea that when modelling various species it is more important to pay attention on the selection of adequate data and species with suitable characteristics than the choice of SDM technique (Hortal et al., 2012).

Practical implications

We have shown that the variation in data and species characteristics can affect SDM results. It is therefore likely that large analyses that involve using SDMs on large numbers of species are susceptible to suffer from large variations in SDM accuracy (McPherson & Jetz, 2007b; Pöyry et al., 2008). As a consequence, predictions of species diversity and composition based on the overlap of various single-species models can present high levels of error (Aranda & Lobo, 2010). Hence, we support the recommendation, made by prior studies, that species that are likely to show low SDM performance should be removed of the analyses (Pöyry et al., 2008; Hanspach et al., 2010). Importantly, special attention should be given to the possibility that several species have low levels of niche completeness, in particular for species that are inconspicuous and/or inhabit scarce and/or remote habitats. If that is the case, data quality shall be improved before these species are included in SDM-based studies.

The variations in SDM performance as a consequence of the species attributes are particularly problematic when these techniques are applied to large numbers of species for systematic conservation planning (Pöyry et al., 2008). Information about species distributions is of key importance for conservation planning. The lack of such data is generally ameliorated by using SDM results for species with a large variety of geographic and ecological attributes (Lemes et al., 2011; Machado & Loyola, 2013; Mateo et al., 2013). Variations in SDM performance between species can therefore result in significant amounts of errors due to the

inclusion of low quality models, hence leading to suboptimal or ineffective conservation actions (Hanspach et al. 2010, Zipkin et al, 2010). Fortunately, our results show a positive relationship between marginality and SDM performance. Marginality is generally linked with rarity, two characteristics that are associated with higher risk of extinction, and hence with species deeming higher priority for conservation. Further, species inhabiting restrict habitats suffer of low genetic variation and are more impacted by climate change, due the difficulties of adapting to new environmental conditions (Kellermann et al., 2009). The good performance of SDMs for these species may indicate that they can be a reliable source of information to be used as a guidance for conservation actions, provided that data quality and environmental niche coverage have been previously assessed (Kadmon et al., 2004; Hortal et al., 2008; Hortal & Lobo, 2011).

The increasing use of species distribution models as a source of information for practical and theoretical studies in ecology, evolution and conservation studies makes necessary to evaluate the effects of all potential sources of uncertainty on their performance. We therefore encourage that future studies evaluate other species attributes that can influence SDM performance. But most importantly, further work is needed on developing new methodologies capable to deal with the uncertainty caused by data characteristics. But in any case, both the coverage of the environmental responses of the species that is provided by the data and the marginality of each species stand out as two key determinants of the reliability of SDM results.

References

Allouche O., Tsoar A., & Kadmon R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Anderson R.P. & Raza A. (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography*, **37**, 1378–1393.

Aranda S.C. & Lobo J.M. (2010) How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. *Ecography*, **34**, 31–38.

Araújo M.B. & New M. (2007) Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22**, 42–47.

Baraud J. (1992) *Coléoptères S carabaeoidea d'Europe, Faune de France 78.* Fédération Française des Sociétés de Sciences Naturelles, Paris and Société Linnéenne de Lyon, Lyon.

Berg Å., Gärdenfors U., & Von Proschwitz T. (2004) Logistic regression models for predicting occurrence of terrestrial molluscs in southern Sweden - importance of environmental data quality and model complexity. *Ecography*, **27**, 83–93.

Boone R.B. & Krohn W.B. (1999) Modeling the occurrence of bird species: Are the errors predictable? *Ecological Applications*, **9**, 835–848.

Breiman L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.

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

Busby J.R. (1986) A biogeographical analysis of Notophagus cunninghamii (Hook.) in south-eastern Australia. *Australian Journal of Ecology*, **11**, 1–7.

Carrascal L.M., Galván I., & Gordo O. (2009) Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos*, **118**, 681–690.

Carrascal L.M., Seoane J., Palomino D., Alonso C.L., & Lobo J.M. (2006) Species-specific features affect the ability of census-derived models to map winter avian distribution. *Ecological Research*, **21**, 681–691.

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.

Chefaoui R.M., Lobo J.M., & Hortal J. (2011) Effects of species' traits and data characteristics on distribution models of threatened invertebrates. *Animal Biodiversity and Conservation*, **34**, 229–247.

Cohen J. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement. **20**, 37–46.

Díaz S., Purvis A., Cornelissen J.H.C., Mace G.M., Donoghue M.J., Ewers R.M., Jordano P., & Pearse W.D. (2013) Functional traits, the phylogeny of function, and ecosystem service vulnerability. *Ecology and evolution*, **3**, 2958–2975.

Diniz-Filho J.A.F., Bini L.M., Fernando Rangel T., Loyola R.D., Hof C., Nogués-Bravo D., & Araújo M.B. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897–906.

Dobrowski S.Z., Thorne J.H., Greenberg J.A., Safford H.D., Mynsberge A.R., Crimmins S.M., & Swanson A.K. (2011) Modeling plant ranges over 75 years of climate change in California , USA : temporal transferability and species traits. *Ecological Monographs*, **81**, 241–257.

FAO (1998) *Soil map of the World. - FAO/UNESCO.*

Farber O. & Kadmon R. (2003) Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, **160**, 115–130.

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

Franklin J., Wejnert K.E., Hathaway S. a., Rochester C.J., & Fisher R.N. (2009) Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California. *Diversity and Distributions*, **15**, 167–177.

Grenouillet G., Buisson L., Casajus N., & Lek S. (2011) Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, **34**, 9–17.

Guisan A., Broennimann O., Engler R., Vust M., Yoccoz N.G., Lehmann A., & Zimmermann N.E. (2006) Using niche-based models to improve the sampling of rare species. *Conservation Biology*, **20**, 501–511.

Guisan A., Zimmermann N.E., Elith J., Graham C.H., Phillips S., & Peterson a. T. (2007) What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, **77**, 615–630.

Halffter G. & Edmonds W.D. (1982) *The nesting behaviour of dung beetles. An ecological and evolutive approach.* Instituto de Ecología, MAB-UNESCO, México, D. F.

Hanspach J., Kühn I., Pompe S., & Klotz S. (2010) Predictive performance of plant species distribution models depends on species traits. *Perspectives in Plant Ecology, Evolution and Systematics*, **12**, 219–225.

Hernandez P.A., Graham C.H., Master L.L., & Albert D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

Hijmans R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, **93**, 679–88.

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

Hortal J. (2008) Uncertainty and the measurement of terrestrial biodiversity gradients. *Journal of Biogeography*, **35**, 1335–1336.

Hortal J., Jimenez-Valverde A., Gómez J.F., Lobo J.M., & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Hortal J. & Lobo J.M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913–2947.

Hortal J. & Lobo J.M. (2011) Can species richness patterns be interpolated from a limited number of well-known areas? Mapping diversity using GLM and kriging. *Natureza & Conservação*, **9**, 200–207.

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 : the journal of the Society for Conservation Biology*, **21**, 853–63.

Hortal J., Lobo J.M., & Jiménez-Valverde A. (2012) Basic questions in biogeography and the (lack of) simplicity of species distributions: Putting species distribution models in the right place. *Natureza e Conservação*, **10**, 108–118.

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**, 1343–1367.

Hortal J., Lobo J.M., & del Rey L. (2006) Distribuición y patrones de diversidad de los afódidos en la comunidad de Madrid (Coleoptera, Scarabaeoidea, Aphodiidae, Aphodinae y Psammodiinae). *Graellsia*, **62**, 439–460.

ITGE (1988) *Atlas Geocientífico y del Medio Natural de la Comunidad de Madrid. - Inst. Tecnológico GeoMinero de España.*

Jiménez-Valverde A. (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, **21**, 498–507.

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.

Kadmon R., Farber O., & Danin A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, **13**, 853–867.

Kadmon R., Farber O., & Danin A. (2004) Effects of roadside bias on the accuracy of predictive maps produced by bioclimtic models. *Ecological Applications*, **14**, 401–413.

Kellermann V., van Heerwaarden B., Sgrò C.M., & Hoffmann A. a (2009) Fundamental evolutionary limits in ecological traits drive Drosophila species distributions. *Science*, **325**, 1244–1246.

Ladle R.J. & Hortal J. (2013) Mapping species distributions : living with uncertainty. *Frontiers of Biogeography*, **5**, 8–9.

Lavergne S., Evans M.E.K., Burfield I.J., Jiguet F., & Thuiller W. (2013) Are species' responses to global change predicted by past niche evolution? *Philosophical transactions of the Royal Society of London.*, **368**, 1–11.

Lemes P., Faleiro F.A.M. V, Tessarolo G., & Loyola R.D. (2011) Refinando Dados Espaciais para a Conservação da Biodiversidade. *Natureza e Conservação*, **9**, 1–4.

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

Lobo J.M., Baselga A., Hortal J., Jiménez-Valverde A., & Gómez J.F. (2007a) 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. & Hortal J. (2006) Los Escarabeidos y Geotrúpidos de la Comunidad de Madrid: lista de especies, distribución geográfica y patrones de diversidad (Coleoptera, Scarabaeoidea, Scarabaeidae y Geotrupidae). *Graellsia*, **62**, 419–438.

Lobo J.M., Jiménez-Valverde A., & Hortal J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114.

Lobo J.M., Jiménez-Valverde A., & Real R. (2007b) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

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**, 158–173.

Machado N. & Loyola R.D. (2013) A comprehensive quantitative assessment of bird extinction risk in Brazil. *PloS one*, **8**, e72283.

Manel S., Williams H.C., & Ormerod S.J. (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.

Marmion M., Luoto M., Heikkinen R.K., & Thuiller W. (2009) The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling*, **220**, 3512–3520.

Martín-Piera F. (2000) Familia Scarabaeidae. *Coleoptera, Scarabaeoidea I.* (ed. by F. Martín-Piera and J.I. López-Colón), pp. 207–432. In: Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Madrid.

Mateo R.G., Felicísimo Á.M., & Muñoz J. (2010) Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. *Journal of Vegetation Science*, **21**, 908–922.

Mateo R.G., de la Estrella M., Felicísimo Á.M., Muñoz J., & Guisan A. (2013) A new spin on a compositionalist predictive modelling framework for conservation planning: A tropical case study in Ecuador. *Biological Conservation*, **160**, 150–161.

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

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

McPherson J.M. & Jetz W. (2007b) Type and spatial structure of distribution data and the perceived determinants of geographical gradients in ecology: the species richness of African birds. *Global Ecology and Biogeography*, **16**, 657–667.

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.

Mouton A.M., De Baets B., & Goethals P.L.M. (2010) Ecological relevance of performance criteria for species distribution models. *Ecological Modelling*, **221**, 1995–2002.

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.

Pearce J., Ferrier S., & Scotts D. (2001) An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. *Journal of environmental management*, **62**, 171–184.

Peterson A.T., Papeş M., & Soberón J. (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, **213**, 63–72.

Phillips S., Anderson R., & Schapire R. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Pineda E. & Lobo J.M. (2009) Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology*, **78**, 182– 190.

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.

Rangel T.F.L.V.B., Diniz-Filho J.A.F., & Araújo M.B. (2009) *BIOENSEMBLES 1.0. Software for computer intensive ensemble forecasting of species distributions under climate change.* Privately distributed. Madrid/Goiânia/Évora.

Reese G.C., Kenneth R.W., H J.A., & Flather C.H. (2005) Factos affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, **15**, 554–564.

Rocchini D., Hortal J., Lengyel S., Lobo J.M., Jimenez-Valverde A., Ricotta C., Bacaro G., & Chiarucci A. (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Romo H., Garcıa-Barros E., & Lobo J.M. (2006) Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography*, **29**, 873–885.

Santika T. (2011) Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, **20**, 181–192.

Scholtz C.H., Davis A.L. V, & Kryger U. (2009) *Evolutionary biology an conservation of dung beetles.* Pensoft Publishers, Sofia and Moscow.

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

Seoane J., Bustamante J., & Díaz-Delgado R. (2005) Effect of expert opinion on the predictive ability of environmental models of bird distribution. *Conservation Biology*, **19**, 512–522.

Smith A.B., Santos M.J., Koo M.S., Rowe K.M.C., Rowe K.C., Patton J.L., Perrine J.D., Beissinger S.R., & Moritz C. (2013) Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. *Ecography*, **36**, 1017–1031.

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

Syphard A.D. & Franklin J. (2009) Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, **32**, 907–918.

Syphard A.D. & Franklin J. (2010) Species traits affect the performance of species distribution models for plants in southern California. *Journal of Vegetation Science*, **21**, 177–189.

Tessarolo G., Rangel T.F., Araújo M.B., & Hortal J. (2014) Uncertainty associated with survey design in Species Distribution Models. Under review.

Thuiller W., Albert C.H., Dubuis A., Randin C., & Guisan A. (2010) Variation in habitat suitability does not always relate to variation in species' plant functional traits. *Biology letters*, **6**, 120–123.

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.

Vaughan I.P. & Ormerod S.J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17**, 1601–1611.

Venier L.A., Mckenney D.W., & Mckee J. (1999) Models of large-scale breeding-bird distribution as a function of macro-climate in Ontario, Canada. *Journal of Biogeography*, **26**, 315–328.

Verdú J.R. & Lobo J.M. (2008) Ecophysiology of thermoregulation in endothermic dung beetles: ecological and geographical implications. *Insect Ecology and Conservation* (ed. by S. Fattorini), pp. 1–28. Research Signpost, Kerala, India.

Violle C., Navas M.-L., Vile D., Kazakou E., Fortunel C., Hummel I., & Garnier E. (2007) Let the concept of trait be functional! *Oikos*, **116**, 882–892.

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.

Wold H. (1975) Soft modelling by latent variables; the nonlinear iterative partial least squares approach. *Perspectives in probability and statistics.* (ed. by J. Gani), pp. 117–142. Perspectives in probability and statistics. Papers in honour of M. S. Barlett. Academic Press.

Supporting information

Table S1. List of species modelled.

Species name	Family
Aphodius (Acanthobodilus) immundus	Aphodiidae
Aphodius (Acrossus) carpetanus	Aphodiidae
Aphodius (Acrossus) luridus	Aphodiidae
Aphodius (Agolius) bonvouloiri	Aphodiidae
Aphodius (Agrilinus) constans	Aphodiidae
Aphodius (Agrilinus) scybalarius	Aphodiidae
Aphodius (Ammoecius) frigidus	Aphodiidae
Aphodius (Anomius) castaneus	Aphodiidae
Aphodius (Aphodius) conjugatus	Aphodiidae
Aphodius (Aphodius) fimetarius	Aphodiidae
Aphodius (Aphodius) foetidus	Aphodiidae
Aphodius (Biralus) satellitius	Aphodiidae
Aphodius (Bodilus) ictericus	Aphodiidae
Aphodius (Bodilus) lugens	Aphodiidae
Aphodius (Chilothorax) distinctus	Aphodiidae
Aphodius (Chilothorax) lineolatus	Aphodiidae
Aphodius (Chilothorax) melanostictus	Aphodiidae
Aphodius (Chilothorax) sticticus	Aphodiidae
Aphodius (Colobopterus) erraticus	Aphodiidae
Aphodius (Coprimorphus) scrutator	Aphodiidae
Aphodius (Esymus) pusillus	Aphodiidae
Aphodius (Melinopterus) prodromus	Aphodiidae
Aphodius (Melinopterus) sphacelatus	Aphodiidae
Aphodius (Melinopterus) tingens	Aphodiidae
Aphodius (Nimbus) affinis	Aphodiidae
Aphodius (Nobius) bonnairei	Aphodiidae
Aphodius (Otophorus) haemorrhoidalis	Aphodiidae
Aphodius (Teuchestes) fossor	Aphodiidae
Aphodius (Trichonotulus) scrofa	Aphodiidae

Bubas bison	Scarabaeidae
Bubas bubalus	Scarabaeidae
Caccobius schreberi	Scarabaeidae
Cheironitis hungaricus	Scarabaeidae
Copris hispanus	Scarabaeidae
Copris lunares	Scarabaeidae
Euoniticellus fulvus	Scarabaeidae
Euonthophagus amyntas	Scarabaeidae
Geotrupes mutator	Geotrupidae
Gymnopleurus flagellatus	Scarabaeidae
Onitis belial	Scarabaeidae
Onthophagus (Furconthophagus) furcatus	Scarabaeidae
Onthophagus (Onthophagus) illyricus	Scarabaeidae
Onthophagus (Onthophagus) taurus	Scarabaeidae
Onthophagus (Palaeonthophagus) coenobita	Scarabaeidae
Onthophagus (Palaeonthophagus) fracticornis	Scarabaeidae
Onthophagus (Palaeonthophagus) grossepunctatus	Scarabaeidae
Onthophagus (Palaeonthophagus) latigena	Scarabaeidae
Onthophagus (Palaeonthophagus) lemur	Scarabaeidae
Onthophagus (Palaeonthophagus) nutans	Scarabaeidae
Onthophagus (Palaeonthophagus) opacicollis	Scarabaeidae
Onthophagus (Palaeonthophagus) ruficapillus	Scarabaeidae
Onthophagus (Palaeonthophagus) similis	Scarabaeidae
Onthophagus (Palaeonthophagus) stylocerus	Scarabaeidae
Onthophagus (Parentius) punctatus	Scarabaeidae
Silphotrupes escorialensis	Geotrupidae
Sisyphus schaefferi	Scarabaeidae
Typhaeus (Typhaeus) typhoeus	Geotrupidae

Chapter 3

Developing maps of biogeographical ignorance about species distributions

Developing maps of biogeographical ignorance about species distributions

Geiziane Tessarolo[1,2] , Richard Ladle[3], Jorge M. Lobo[2], Thiago Fernando Rangel[1] & Joaquín Hortal[2]

[1]Departamento de Ecologia, Instituto de Ciências Biológicas – ICB, Universidade Federal de Goiás – UFG Campus II, 74001-970, Goiânia, GO, Brazil

[2]Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales (CSIC), C/José Gutiérrez Abascal 2, 28006 Madrid, Spain

[3]Institute of Biological and Health Sciences, Federal University of Alagoas, Praça Afrânio Jorge, Prado, Maceió, AL, Brazil

Abstract

Species distribution models (SDMs) are widely used to generate spatially-explicit predictions about species' geographic ranges. However, SDM predictions are subject to many sources of variability during the modelling process, generating significant uncertainty in their results. These uncertainties diminish their reliability, being one of the most important factors limiting the use of SDM in ecological, evolutionary or conservation applications. The distributional data used to calibrate SDMs can have great impact on their predictions, being one of the main sources of SDM uncertainty. Here we propose a methodology to create Maps of Biogeographical Ignorance (MoBI) that can account for the uncertainty associated of the calibration data used in SDM. These maps reflect the taxonomic quality, spatial and environmental similarity and temporal decay in the degree of information provided by these data. We developed a set of tools for R that allow generating MoBI, and used data on Iberian Scarabaeidae dung beetles as an example. The maps allowed identifying geographic areas of large uncertainty (i.e. areas in which SDM results are less reliable); more than 60% of the study region presented high (>0.9) values of biogeographical uncertainty. The MoBI showed large similarity among different species. Although they can be improved adding more species-specific information in the methodology, it is likely that this pattern is common for most groups and regions. MoBI can also be adapted to depict the (lack of) knowledge about entire groups of species. This type of spatially-explicit approach can also be developed to account for other sources of uncertainty in SDM. And it may allow incorporating spatially-explicit accounts of error during the model building stage, hence improving SDM predictions. In any case, maps of biogeographical ignorance allow a fair acknowledgement of the uncertainty in SDM caused by data limitations, and are particularly useful for conservation assessments. We therefore suggest that MoBIs, or any similar account of degree of knowledge provided by

distributional data, should be routinely presented as support information about the reliability of SDM results.

Keywords: Ecological niche models, taxonomic quality, temporal decay, spatial decay, uncertainty, predictive accuracy, SDM performance, biodiversity data.

Introduction

Data on species distributions are the basis for many ecological and evolutionary studies. Thus, adequate information about where the species occur both in space and time is of fundamental importance to guarantee the utility and reliability of the results of these studies (Duputié et al., 2013). However, obtaining observational data about the complete distribution of a species is generally impracticable due the large temporal and spatial scales that are relevant to many researches (Blackburn, 2004). Despite current awareness of the need for collecting data about species distributions, our knowledge is biased to some taxa and regions (Bini et al., 2006; Hortal et al., 2007; Rocchini et al., 2011; Yang et al., 2013). The so-called Wallacean Shortfall (Lomolino, 2004) operates even for many well-known taxa. In fact, for most, if not all, species the information on their geographic distribution is not of enough quality to support robust analyses (Whittaker et al., 2005).

The difficulty to obtain reliable data about species distributions has resulted in the development of several methodologies to provide modelled distributional data, such as range maps and species distributions models (SDM). SDM are widely used in many research areas to study a variety of issues from biodiversity discovery to conservation planning (Peterson, 2006; Franklin & Miller, 2009). Although SDM research has experienced great technical

developments in what refers to the algorithms used for modeling, there are some conceptual, methodological and statistical issues still pending of discussion and advance (Araújo & Guisan, 2006; Hortal et al., 2012; Rangel & Loyola, 2012). Among them, one of the most outstanding concerns is the uncertainty associated with SDM predictions (Barry & Elith, 2006; Dormann et al., 2008; Aranda & Lobo, 2010; Tessarolo, 2012). Uncertainty is introduced in all stages of the modeling process (Heikkinen et al., 2006; Hortal et al., 2008; Diniz-Filho et al., 2009; Buisson et al., 2010; Grenouillet et al., 2011). There are three main sources of uncertainty in SDM: data quality and availability, the process of map construction (i.e. decisions related with the modelling technique) and the uncertainty in parameter estimations (due to the natural dynamics of species distributions) (Dormann et al., 2008; Rocchini et al., 2011; Beale & Lennon, 2012). Although the limitations imposed to SDM predictions by such uncertainty are known to be large, these models are generally presented and used as precise and accurate representations of the real distributions of species (Wilson, 2010). The reliability of SDM results is usually assessed based only on metrics of predictive performance that are not always consistent with their real accuracy (Lobo et al., 2007b; Hanberry et al., 2012; Jiménez-Valverde, 2012), and that are unable to provide information about the spatial distribution of the uncertainties (Fielding, 2002; Jiménez-Valverde et al., 2008; Ladle & Hortal, 2013). As a consequence, the geographical variations in the uncertainty associated with model predictions are rarely presented and/or taken into account in studies using SDM.

A few studies have evaluated the impacts of uncertainty through the different steps of the modelling process on SDM results (Syphard & Franklin, 2009; Lobo & Tognelli, 2011; Kramer-Schadt et al., 2013; Naimi et al., 2014). However, the quantification of these uncertainties is a topic underdeveloped in the SDM literature. Most studies proposing the quantification of uncertainty are focused on the statistical process of parameterization, while

others try to account for the uncertainty arising from the selection and fit of the models (Hartley et al., 2006; Swanson et al., 2013). There is however a lack of studies that propose methodologies to quantify the uncertainty from the input data used to calibrate the models, despite being recognized as an important source of error (Hirzel & Guisan, 2002; Edwards Jr. et al., 2006; Albert et al., 2010; Braunisch & Suchant, 2010). Once that SDM make use of prior knowledge on the geographical distribution of species, all the properties related to the characteristics and quality of such data can potentially influence SDM predictions (Rocchini et al., 2011; Kamino et al., 2012).

Recently, Rocchini et al. (2011) suggested the use of "maps of ignorance", based on Boggs' (1949) original idea of creating an atlas of the ignorance about spatially-explicit phenomena. These Maps of Biogeographical Ignorance (MoBI) will allow representing the geographic knowledge about the distribution of biodiversity. Ladle & Hortal (2013) propose conceptual ways to develop "maps of biogeographical ignorance" with the objective of assessing and communicating the uncertainty related to the quality, longevity and coverage of biogeographical data used to calibrate the models. Hence, the development of MoBI should take into account three main sources of biogeographical ignorance: the completeness of original survey data, the time passed since the last survey was completed, and the geographical (and environmental) distance between the point of prediction and the survey point. These components summarized in a map would provide information about the limitations and consequences of biogeographical data (Hortal et al., 2014, *in prep.*). When associated to SDM, MoBI may allow identifying areas where the reliability of the models is either known or unknown, or even to determine different levels of certainty about model predictions. Considering the importance of prior biogeographical knowledge for SDM, these maps will improve the interpretation of its model results. In additionally, MoBI can help evaluating the reliability of SDM predictions and identifying areas where more sampling

effort is required, thus offering a fundamental support to the development of informed conservation actions (Rocchini et al., 2011).

Although the conceptual basis to the development of maps of biogeographical ignorance is now well defined, no methodology has been yet suggested to create such maps. Here we propose a first methodological approach for the creation of MoBI about species distributions, which accounts for each source of ignorance highlighted on the theoretical model of Hortal et al. (2014, *in prep.*), namely data quality, and spatial and temporal decay of the knowledge provided by distributional data. This methodology is intended to be particularly suited to depict the spatial distribution of the uncertainty arising from the data used for SDM calibration. Therefore, we use data on fourteen Scarabaeidae dung beetle species from Iberian Peninsula to generate both species distribution models and the maps of biogeographical ignorance that are related to them.

Methods

Biological Data

We used data on the distribution of Scarabaeidae species coming from the latest version of BANDASCA (Lobo & Martín-piera, 1991). This database joints all distributional information about the species from this family available on the literature, museums and private collections and unpublished data. In total, it currently hosts information about c. 97,000 individuals from the 53 Iberian Scarabaeidae dung beetle species. From them, we chose the 14 species with higher numbers of records, to model their distribution and create maps of biogeographical ignorance.

Environmental data

Data on climate and topography were obtained from Worldclim (Hijmans et al., 2005) and the Global Resource Information Database - United Nations Environment Programme (UNEP/GRID, http://www.grid.unep.ch/data/data.php). A principal components analysis (PCA) was carried out with all the 29 variables in these datasets (see also Baselga & Araújo, 2009). A total of five PCA axes with eigenvalue >1 were selected, accounting for 91.04% of the climatic variation. For each one of these factors we selected the variable with highest loading value for all subsequent analyses. Hence, isothermality, mean temperature of wettest and warmest quarter, precipitation of coldest quarter and solar radiation monthly average were used to generated SDMs for all species, and to create the environmental distance matrix needed for the development of MoBIs (see below).

Species Distribution Models

We generated SDMs using six different algorithms and the combined consensus among them (Araújo & New, 2007). These algorithms were BIOCLIM (Busby, 1986), Gower distance (Carpenter et al., 1993), Generalized Linear Models (GLM; McCullagh & Nelder, 1989), Maximum Entropy (MaxEnt; Phillips et al., 2006), Generalized Additive Models (GAM; Hastie & Tibshirani, 1990) and Neural Networks (NNet; Ripley, 1996). The models were constructed using 75% of the data for calibration and 25% for validation. These datasets were selected randomly. This procedure was repeated 50 times, and these results were used to generate predictive maps of the occurrence of each species according to each algorithm, based on the mean value of the 50 predictions. The threshold used to convert the predicted probabilities into binary outcomes was set based on the prevalence ef each species in the data, so the predicted prevalence reflects the observed prevalence. The consensus among the six

SDM techniques was generated through the weighted means of the predictions of each algorithm; these weights were based on the True Skill Statistic (TSS) evaluation metric (Araújo & New, 2007).

Creating Maps of Biogeographical Ignorance

We developed a combination of tools that allow representing all the main sources of biogeographical ignorance suggested by Ladle & Hortal (2013) and Hortal et al., (2014, *in prep.*). Here we decompose these sources into four components: data completeness, taxonomic quality of the data, temporal decay in the information provided by these data, and spatial and environmental decay in such information. In our approach, the value of uncertainty for each cell of the MoBIs is calculated in two different ways, depending on whether the focal cell hosts data on the studied group or not. For cells with recorded data, such value is obtained as a composite of data completeness, taxonomic quality and temporal decay. Whereas for cells without data such value is obtained as an interpolation from the values of influent cells in which we add the space-environmental distance component (Fig. 1, see details below). All analyses were implemented in R (code available at Appendix X).

Data completeness

Failure to detect a species when it is actually present in a locality may come from insufficient sampling effort. We therefore characterize survey completeness (herein completeness for short) as the value of the final slope of the species accumulation curves for all cells that reached a minimum of individuals recorded. Species accumulation curves relate the number of species that are added to the inventory as the number of records increases. Here, the final slope value represents the current rate of accumulation of new species to the inventory. We

selected cells with a minimum of 50 individuals, generating 100 species accumulation curves by randomizing the order of entry of the individuals in the curve. For each curve the final slope value was calculated as the difference in the number of species observed between the two last records of the randomized curve. The mean of these 100 values, ranging from 0 to 1, is then used as value of completeness of the cell. For cells with less than 50 individuals the value of completeness is considered zero. See (Hortal & Lobo, 2005) for further details about the use of species accumulation curves in this context.

Taxonomic quality

One of the less acknowledged sources of uncertainty when characterizing species distributions is the taxonomic quality of the records (but see, e.g., Lozier et al., 2009). Further, the accuracy of the taxonomic identifications may decay with time as new taxonomic revisions are conducted, that may synonymize some species names while updating the knowledge about a given group (see (Baselga et al., 2010). We therefore estimate taxonomic quality according to the person responsible for the identification. We assigned a value of taxomonic quality to all the records in the database, according to the type of identifier (i.e., person responsible for the identification). We classified the identifiers in three classes: professional taxonomist, expert and amateur. These classes were used to define the initial value of quality. Thus, if the identifier is a taxonomist the quality value is bigger than for an expert, and the same follows from expert to amateur. Here we set up the quality values as 1, 0.9 and 0.75 for taxonomist, expert and amateur identifiers respectively.

Temporal decay in the information

The distributions of species, and therefore the composition of local assemblages, typically change in time. This implies that, as the time since each record was obtained passes, the likelihood that the recorded species is currently absent from the locality increases. To simulate such temporal decay in the information effectively provided by each occurrence record, we used a kernel Gaussian function. Here we assume that this function accounts for the non-linearity in such decay in effective knowledge. In our understanding the influence of time on the quality of the information provided by distributional data is not linear; rather, its influence may be much smaller in short periods of time. In the absence of data and/or hypotheses about the velocity of decrease in information quality, for simplicity we adjusted this decay function to the total period encompassed by the records available in the database. Therefore, the most recent and the oldest records in the database will be the most and less informative, respectively. The slope of the temporal decay curve will therefore vary according to the records available.

Spatial and environmental decay in the information

According to Tobler (1970) the first law of geography states that everything is related to everything else, but near things are more related than distant things. It follows that the information about any given phenomena in a certain location that is provided by other points in space will decrease as the distance to that point increases. In the case of the geographic distributions of species, it is well known that both environmental conditions and biogeographical and bionomic processes that leave a spatially-explicit pattern in the geographical space determine species distributions (Soberón, 2007, 2010; Hortal et al., 2010). Therefore, we assumed that the amount of information about the presence (or absence) of a species in the areas with insufficient survey effort provided by the sites with records diminishes with the spatial and environmental distance to these records. We therefore

modelled such spatial and environmental decay based on two distance matrices. The spatial distance matrix depicts the Euclidean geographic distances between all pairs of cells, whereas the environmental matrix was constructed using Gower multivariate distances based on the environmental variables that were used for SDM predictions (see above). These matrices were rescaled to range from 0 to 1, and afterwards combined as their product to take into account the interaction between both factors. This way, we generated a spatio-enviromental distance matrix that was later used to interpolate the degree of biogeographical knowledge (see below).

Assembling Maps of Biogeographical Ignorance

A Biogeographical Ignorance (BI) value was attributed to each cell of the map of the studied region according to a combination of the four components described above. For each record in the database the final amount of information value was calculated as the sum of the completeness, taxonomic quality and temporal decay. These amount of information values were rescaled to obtain values of Biogeographical Ignorance varying from 0 to 1 (complete ignorance and complete certainty, respectively). For cells with more than one record for the focal species the final BI value was set up as the maximum value of all these records, for such record provides the most reliable and up-to-date account of the occurrence of the species. For cells without records of the focal species, but with records of other species, the final BI value was calculated as the mean of the maximum BI values of each species recorded on the cell.

For cells without any information about the presence or absence of species, the BI value was based on an interpolation of the values of the neighboring influent cells. The interpolation values were calculated according to the spatio-environmental distance matrix by means of the Inverse Distance Weighting Average (IDWA). This is a local interpolation methodology in which the values of unknown points are estimated using a linear combination

of values at known data points (selected by a radius of influence) weighted by the inverse of the distance, under the assumption that the closer the estimated point is to a known data point, the stronger the influence of such data point will be. To define the radius of influence, we performed a correlogram with the environmental variables used on the SDM calibration. Here, the farther distance class with correlation above zero was used as the maximum distance at which a cell can be influent on others. Only cells placed closer to the estimated cell than this distance (388.4 Km in our analyses) were assumed to influence the interpolated value. Additionally, if there were recorded presences of the focal species within this radius, the corresponding cells were assumed to have a higher contribution to the interpolated value. Therefore, these cells were weighted to account for 50% of the interpolated value, whereas the remaining 50% weight was divided across the remaining cells.

In our approach, the distance used for IDWA calculations was the value of spatio-enviromental distance calculated by the combined spatial and environmental matrices. The IDWA is therefore calculated as:

$$Z_j = \frac{\sum_{i=1}^{n}\left(\frac{z_i}{(h_{ij})^b}\right)}{\sum_{i=1}^{n}\left(\frac{1}{(h_{ij})^b}\right)} \quad \text{(eq. 1)}$$

…where Z_j is the ignorace value interpolated to each cell j, h_{ij} is the spatio-environmental distance between the cell j and its neighbor i, b is the weighting exponent and n is the number of influent neighbours (i.e., those cells lying within the radius described above). By using this approach, we guarantee that the final biogeographical ignorance value for cells lacking previous information is more influenced by cells placed closer in the space and environment.

Results

For each species we generated seven SDM predictions. The model assessment values varied greatly among metrics; only the ensemble predictions presented values greater than 0.8 for all metrics (Table 1). Although SDM predictions vary widely among species, the geographic patterns of biogeographical ignorance were very similar for all species, showing similar trends in the Maps of Biogeographic Ignorance (Table 2; Supporting information Fig. S9). Therefore, for simplicity we only show the results of three species in the main text, these species representing the more discrepant distributional patterns modeled by consensus method – *Bubas bison*, *Copris lunaris* and *Onthophagus vacca*. The results for the rest of the species and SDM techniques are shown at the Supporting information Figs. S5 – S8. In general, the MoBIs show that Western (mostly Portugal) and Central-Eastern areas are the regions with larger degrees of uncertainty about the distribution of Scarabaeidae species in the Iberian Peninsula (Fig. 1). In fact, about 63% of the study area presented values of BI larger than 0.9 (Supporting information Table S5).

Table 1: Values of the predictive performance assessment metrics values for each species and modelling technique.

	Sensitivity							Specificity							Kappa							TSS						
Species	Bioclim	Gower	GLM	GAM	NNET	MaxEnt	Ensemble	Bioclim	Gower	GLM	GAM	NNET	MaxEnt	Ensemble	Bioclim	Gower	GLM	GAM	NNET	MaxEnt	Ensemble	Bioclim	Gower	GLM	GAM	NNET	MaxEnt	Ensemble
E. amyntas	0.28	0.26	0.32	0.31	0.32	0.31	0.90	0.29	0.34	0.33	0.40	0.35	0.39	0.99	0.01	0.02	0.03	0.05	0.04	0.05	0.90	0.29	0.30	0.33	0.35	0.33	0.35	0.94
B. bison	0.35	0.37	0.35	0.37	0.38	0.38	0.93	0.34	0.34	0.38	0.36	0.35	0.38	0.99	0.03	0.03	0.05	0.04	0.04	0.05	0.94	0.34	0.36	0.36	0.37	0.37	0.38	0.96
B. bubalus	0.33	0.33	0.35	0.34	0.35	0.36	0.96	0.28	0.31	0.37	0.36	0.35	0.36	0.98	0.02	0.02	0.05	0.05	0.05	0.05	0.94	0.31	0.32	0.36	0.35	0.35	0.36	0.97
O. fracticornis	0.31	0.27	0.41	0.33	0.30	0.38	0.82	0.35	0.33	0.34	0.43	0.42	0.42	0.99	0.03	0.02	0.04	0.09	0.08	0.09	0.85	0.33	0.30	0.37	0.38	0.36	0.40	0.91
E. fulvus	0.27	0.25	0.31	0.28	0.42	0.30	0.94	0.25	0.29	0.31	0.34	0.15	0.34	0.96	0.00	0.01	0.02	0.03	0.01	0.03	0.90	0.26	0.27	0.31	0.31	0.29	0.32	0.95
O. furcatus	0.30	0.29	0.30	0.29	0.25	0.30	0.87	0.27	0.27	0.31	0.35	0.39	0.36	0.98	0.01	0.01	0.03	0.04	0.05	0.05	0.87	0.28	0.28	0.31	0.32	0.32	0.33	0.92
C. hispanus	0.31	0.35	0.30	0.37	0.29	0.34	0.92	0.33	0.30	0.37	0.32	0.38	0.37	0.97	0.03	0.03	0.04	0.03	0.04	0.05	0.90	0.32	0.33	0.33	0.35	0.34	0.35	0.95
C. lunaris	0.29	0.33	0.32	0.34	0.24	0.34	0.84	0.32	0.28	0.33	0.35	0.40	0.36	0.99	0.02	0.01	0.03	0.03	0.04	0.04	0.86	0.30	0.30	0.33	0.34	0.32	0.35	0.91
O. opacicollis	0.25	0.31	0.27	0.29	0.24	0.27	0.90	0.29	0.28	0.36	0.35	0.40	0.38	0.98	0.00	0.01	0.02	0.02	0.04	0.03	0.89	0.27	0.29	0.31	0.32	0.32	0.33	0.94
O. punctatus	0.23	0.21	0.33	0.27	0.18	0.26	0.82	0.34	0.33	0.35	0.38	0.42	0.39	0.97	0.01	0.00	0.02	0.02	0.03	0.02	0.81	0.28	0.27	0.34	0.33	0.30	0.32	0.89
C. schreberi	0.27	0.29	0.31	0.32	0.25	0.30	0.80	0.29	0.30	0.32	0.37	0.39	0.40	0.98	0.01	0.02	0.02	0.05	0.04	0.06	0.82	0.28	0.30	0.31	0.34	0.32	0.35	0.89
O. similis	0.28	0.25	0.34	0.35	0.28	0.34	0.89	0.30	0.32	0.35	0.36	0.38	0.38	0.99	0.02	0.02	0.05	0.06	0.06	0.07	0.90	0.29	0.29	0.34	0.35	0.33	0.36	0.94
O. taurus	0.28	0.24	0.29	0.31	0.25	0.30	0.89	0.22	0.26	0.29	0.30	0.34	0.34	0.98	0.00	0.00	0.02	0.03	0.03	0.05	0.88	0.25	0.25	0.29	0.30	0.29	0.32	0.93
O. vacca	0.31	0.28	0.33	0.32	0.31	0.31	0.96	0.23	0.29	0.31	0.32	0.31	0.32	0.97	0.01	0.02	0.04	0.04	0.04	0.04	0.93	0.27	0.28	0.32	0.32	0.31	0.32	0.97

Table 2: Values of Pearson's correlation among the Maps of Biogeographic Ignorance of all species considered.

	E. amyntas	*B. bison*	*B. bubalus*	*O.fracticornis*	*E. fulvus*	*O. furcatus*	*C. hispanus*	*C. lunaris*	*O. opacicolis*	*O. punctatus*	*C. schreberi*	*O. similis*	*O. taurus*
B. bison	0.961												
B. bubalus	0.969	0.974											
O.fracticornis	0.976	0.962	0.964										
E. fulvus	0.966	0.965	0.972	0.967									
O. furcatus	0.968	0.958	0.968	0.961	0.972								
C. hispanus	0.962	0.977	0.973	0.960	0.966	0.961							
C. lunaris	0.974	0.959	0.968	0.976	0.969	0.962	0.962						
O. opacicolis	0.958	0.969	0.960	0.962	0.966	0.961	0.965	0.957					
O. punctatus	0.967	0.961	0.960	0.965	0.964	0.968	0.966	0.961	0.963				
C. schreberi	0.979	0.962	0.969	0.973	0.975	0.973	0.964	0.973	0.961	0.969			
O. similis	0.971	0.956	0.963	0.975	0.968	0.969	0.962	0.972	0.962	0.968	0.977		
O. taurus	0.969	0.965	0.973	0.963	0.977	0.975	0.963	0.964	0.968	0.964	0.973	0.97	
O. vacca	0.967	0.966	0.970	0.965	0.974	0.967	0.965	0.967	0.969	0.960	0.970	0.971	0.979

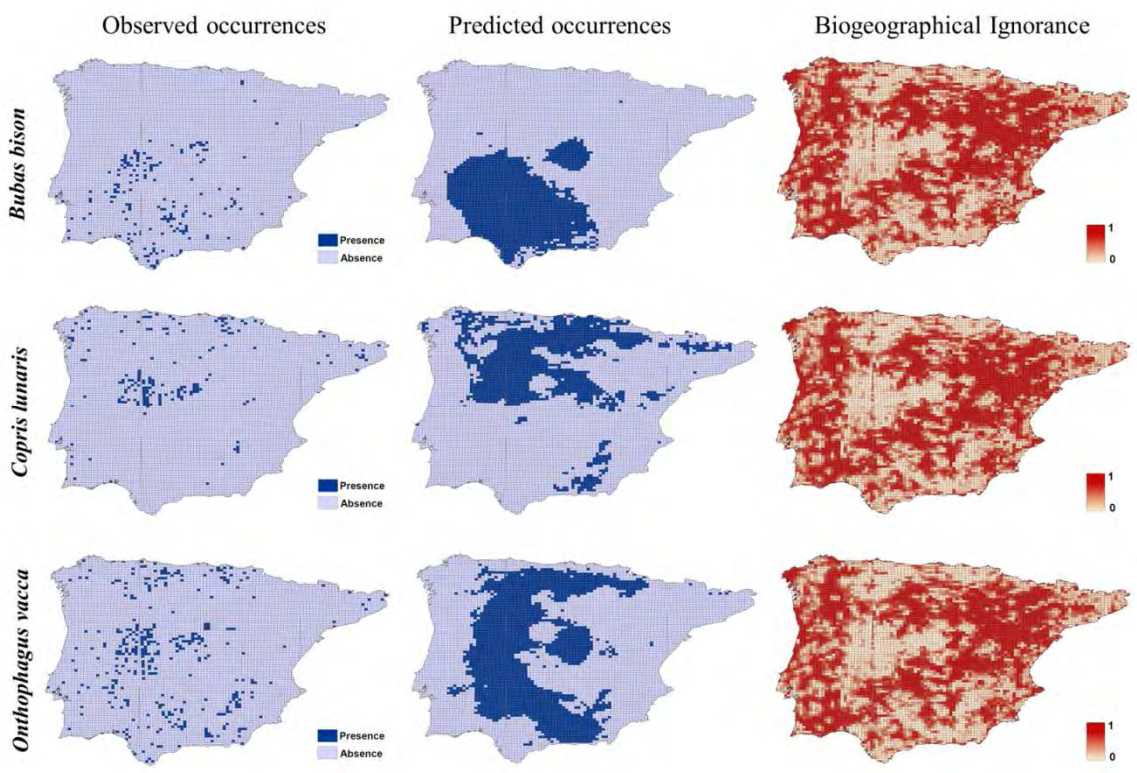


Figure 1. Observed occurrences from BANDASCA database, occurrences predicted by the ensemble SDM technique and Maps of Biogeographical Ignorance about the distribution of *Bubas bison*, *Copris hispanus* and *Onthophagus vacca*.

Discussion

Current loss of species is reaching values of magnitude comparable only to past events of extinction mass (Dirzo & Raven, 2003). The urgency fostered by the biodiversity crisis makes unadvisable to wait until complete data and/or accurate descriptions of the species responses to the environmental changes are available. However, it is likely that such kind of data is impossible to obtain for complex phenomena such as ecosystems or species distributions. Although species distribution models are commonly used to fill in this gap in knowledge, their predictions cannot be assumed to truly represent species

distributions, once we know that they host uncertainty coming from their input data and therefore in their estimates. In fact, their indiscriminate use can harm more than help the development of conservation strategies, as fundamentally flawed predictions of species distributions can result in failing to develop the most adequate strategies for biodiversity conservation. A fair and honest way of using SDM applications would therefore be to take into account their uncertainty when interpreting their results, to be aware about its potential impact on model results (Elith et al., 2002; Rocchini et al., 2011).

In this study we have created the first Maps of Biogeographical Ignorance associated to the calibration data used in SDM, taking into account the main factors influencing the knowledge (of lack of) about the geographic distributions of species. Our results suggest that the uncertainty due to lack of biogeographical knowledge may be widespread in, arguably, well-known regions such as the Iberian Peninsula, an area with fairly good –though uneven– knowledge on the distribution of Scarabaeidae species (Martín-Piera, 2000; Lobo et al., 2007a). As expected, the MoBIs reflect to some degree the bias in data collection. The regions with greater degrees of BI in the Iberian Peninsula are those placed further apart from urban centers, research institutes and universities, a bias which has been shown to occur to this group of species in the study area (Lobo & Martín-piera, 2002; Lobo et al., 2007a; Hortal & Lobo, 2011). Although MoBIs are conditioned by the spatial distribution of former surveys, the inclusion of the decay of information with increasing environmental distance partly disrupts such dependence. Further, such inclusion makes MoBIs more appropriate to describe the uncertainty associated to SDMs, which are calibrated using the same environmental variables. This dependence of data collection and the use of information about all the other species from the same group to calculate BI values are responsible to the similar patterns shown by MoBIs of different species.

Although several studies have attempted to quantify the uncertainty affecting SDMs arising from some sources (Hartley et al., 2006; Wilson, 2010; Synes & Osborne, 2011; Swanson et al., 2013), our study is the first to quantify the uncertainty caused from the quality and evenness of the distributional input data. Species distribution models are only as reliable as the data used to construct them. The quality of input data is one of the major sources of uncertainty for all kind of models. However, in SDM in particular such uncertainty can have an even larger impact than in other types of models, for they can be fitted with little information about the species' presence and little or even no information about their absence (Lobo et al., 2010; Beale & Lennon, 2012; Kamino et al., 2012). The widespread uncertainty associated to the distributional data can therefore have a critical impact on the reliability of SDM projections, once that more than half of the studied area presented values of biogeographical ignorance (i.e. uncertainty) above 0.9. Nonetheless, we must note that the high-uncertainty area can dramatically increase if the purpose of the study needs a more conservative value of uncertainty, so these figures may be much higher than the reasonably acceptable for the design of conservation schemes.

We must note that this is a first approximation to the creation of MoBIs. Rather than being a definitive proposal, the analyses implemented here constitute a necessarily simple and preliminary approach to a complex goal that may need further technical and practical improvements. The MoBIs for single species can be improved by the inclusion of more specific data about the different components of the uncertainty in distributional data. For example, in the taxonomic quality component, the rate of temporal decay could be based on the actual rates of synonymyzation and the numbers of revisions published after the individuals of each particular record were identified and/or revised for the last time. Further, the values of similarity used to characterize the environmental

and spatial decay in the information can be rescaled to more complex curves using parameters reflecting the size of the species' distributional range and/or its dispersal ability. The addition of species-specific data can increase the discrepancy between the MoBIs of different species, making them more appropriate for their use in assessments of the reliability of SDM results for unique species. However, it is also possible to increase the generality of MoBIs, in a way that makes them adequate for quantifying the uncertainty of any given group of species as a whole. The potential value of such approach is supported by the high similarity in the MoBIs of different species shown by our results. The generalization of the maps of biogeographical ignorance to account for the uncertainty associated to the data on a given group of species as a whole can be useful to include rare species with little data available in the assessment, as well as on applications based in either species richness models or in conservation assessments based on calibrating SDMs for many species in an automated fashion (Wilson et al., 2005).

Information visualization is a valuable way to explore and communicate data, because it allows that the relevant information is easily highlighted and understood. This can allow identifying processes that were yet unknown (Elith et al., 2002; Griethe & Schumann, 2005). Further, visualization is a crucial step not only for the development of science, but also to engage the general public and the policy makers into exploring and understanding scientific results and model projections (McInerny et al., 2014). This approach can be particularly useful in the analyses of errors, once that the spatial mapping of the degree of knowledge allows identifying both areas with higher uncertainties and the covariates that could be possibly linked to these uncertainties (Lobo & Martín-piera, 2002; Barry & Elith, 2006; Kühn et al., 2006). Here we used maps to summarize and display the information about the uncertainty coming

from the distributional data used to calibrate SDMs, showing that MoBIs can be a relevant source of information for modelers. These maps provide information about the areas of failure of the SDMs and allow identifying areas where improvements in data quality is critical. Further, their presentation to end users will make them aware about the degree of uncertainty associated to SDM predictions, allowing a fair interpretation of their results. In the context of systematic conservation planning, MoBIs may help solving the problem of lack of communication of uncertainty that is commonly acknowledged by decision makers (Regan et al., 2005; Burgman & Yemshanov, 2013; Young et al., 2014), allowing an explicit consideration of this uncertainty when making management decisions.

It is arguably impossible to address all sources of uncertainty in a single anaysis (Elith et al., 2002). Therefore, the primary focus of data quality assessments should rely in one or a few important sources. Although our proposed methodology considers only one of the several important sources of uncertainty affecting SDM, the approach used here can be can adapted and widely applied to other sources of ignorance. As mentioned above, the information about the uncertainty associated to the distributional data can help in assessing SDM reliability. But it can also help improving the models. The development of new methods that quantify the uncertainty in SDM, will make necessary the concomitant development of methods to handle this uncertainty. A promising area of research would be the implementation of methods that allow accounting for the variations in uncertainty as an spatially-explicit error term during the modeling process, so that model predictions are corrected for variations in uncertainty over space (see (McInerny & Purves, 2011; Beale & Lennon, 2012). Thus, we encourage the development of methodologies to account for, map and ultimately handle other sources of uncertainty in SDM. More importantly, we suggest that the here presented Maps of

Biogeographical Ignorance, or more developed approaches, should be routinely used as an indispensable part for any assessment and interpretation of SDM. The joint visualization of SDM predictions and their associated uncertainty (such as in Figure 1) may allow a fair interpretation of the reliability of these predictive maps. Here we agree with Elith et al. (2002) in their interpretation of Draper (1995): "… it is difficult to deal with uncertainty, but in the long run it is better to address and treat the uncertainty in an attempt to include the true values within a model's predictions than to ignore uncertainty and miss the true values completely." The more honest we are with the limitations of our descriptions of nature, the more sound will the conclusions of our research be.

References

Albert C.H., Yoccoz N.G., Edwards T.C., Graham C.H., Zimmermann N.E., & Thuiller W. (2010) Sampling in ecology and evolution - bridging the gap between theory and practice. *Ecography*, **33**, 1028–1037.

Aranda S.C. & Lobo J.M. (2010) How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. *Ecography*, **34**, 31–38.

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

Araújo M.B. & New M. (2007) Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22**, 42–47.

Barry S. & Elith J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Baselga a., Lobo J.M., Hortal J., Jiménez-Valverde a., & Gómez J.F. (2010) Assessing alpha and beta taxonomy in eupelmid wasps: determinants of the probability of describing good species and synonyms. *Journal of Zoological Systematics and Evolutionary Research*, **48**, 40–49.

Baselga A. & Araújo M.B. (2009) Individualistic vs community modelling of species distributions under climate change. *Ecography*, **32**, 55–65.

Beale C.M. & Lennon J.J. (2012) Incorporating uncertainty in predictive species distribution modelling. *Philosophical transactions of the Royal Society of London*, **367**, 247–58.

Bini L.M., Diniz-Filho J.A.F., Rangel T.F.L.V.B., Bastos R.P., & Pinto M.P. (2006) Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. *Diversity and Distributions*, **12**, 475–482.

Blackburn T.M. (2004) Method in macroecology. *Basic and Applied Ecology*, **5**, 401–412.

Boggs W.S. (1949) An atlas of ignorance: a needed stimulus to honest thinking and hard work. *Proceedings of the american philosophical society*, **93**, 253–258.

Braunisch V. & Suchant R. (2010) Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. *Ecography*, **33**, 826–840.

Buisson L., Thuiller W., Casajus N., Lek S., & Grenouillet G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145–1157.

Burgman M. a. & Yemshanov D. (2013) Risks, decisions and biological conservation. *Diversity and Distributions*, **19**, 485–489.

Busby J.R. (1986) A biogeographical analysis of Notophagus cunninghamii (Hook.) in south-eastern Australia. *Australian Journal of Ecology*, **11**, 1–7.

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**, 667–680.

Diniz-Filho J.A.F., Bini L.M., Fernando Rangel T., Loyola R.D., Hof C., Nogués-Bravo D., & Araújo M.B. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897–906.

Dirzo R. & Raven P.H. (2003) Global state of biodiversity and loss. *Annual Review of Environment and Resources*, **28**, 137–167.

Dormann C.F., Purschke O., García Márquez J.R., Lautenbach S., & Schröder B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–86.

Draper D. (1995) Assesment and propagation of model uncertainty. *Journal of the Royal Statistical Society*, **57**, 45–97.

Duputié A., Zimmermann N.E., & Chuine I. (2013) Where are the wild things? Why we need better data on species distribution. *Global Ecology and Biogeography*, n/a–n/a.

Edwards Jr. T., Cutler D., Zimmermann N., Geiser L., & Moisen G. (2006) Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, **199**, 132–141.

Elith J., Burgman M.A., & Regan H.M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, **157**, 313–329.

Fielding A.H. (2002) What are the appropriate characteristics of an accuracy measure? *Predicting Species Occurrences: Issues of Accuracy and Scale (eds Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. & Samson, F.B.). Island Press, Covelo, CA.* pp. 271–280.

Franklin J. & Miller J.A. (2009) *Mapping species distributions: spatial inference and prediction.* Cambridge University Press Cambridge, UK, New York, New York, USA.

Grenouillet G., Buisson L., Casajus N., & Lek S. (2011) Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, **34**, 9–17.

Griethe H. & Schumann H. (2005) Visualizing uncertainty for improved decision making. *Proceedings of the 4th International Conference on Business Informatics Research, Skövde, Sweden*, .

Hanberry B.B., He H.S., & Dey D.C. (2012) Sample sizes and model comparison metrics for species distribution models. *Ecological Modelling*, **227**, 29–33.

Hartley S., Harris R., & Lester P.J. (2006) Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. *Ecology letters*, **9**, 1068–79.

Hastie T. & Tibshirani R.J. (1990) Generalized Additive Models. , vol. 43. *Monographs on Statistics and Applied Probability*, **43**, 335 pp.

Heikkinen R.K., Luoto M., Araujo M.B., Virkkala R., Thuiller W., & Sykes M.T. (2006) Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography*, **30**, 751–777.

Hijmans R.J., Cameron S.E., Parra J.L., Jones P.G., & Jarvis A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hirzel A. & Guisan A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157**, 331–341.

Hortal J., Jimenez-Valverde A., Gómez J.F., Lobo J.M., & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Hortal J. & Lobo J.M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913–2947.

Hortal J. & Lobo J.M. (2011) Can species richness patterns be interpolated from a limited number of well-known areas? Mapping diversity using GLM and kriging. *Natureza & Conservação*, **9**, 200–207.

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 : the journal of the Society for Conservation Biology*, **21**, 853–63.

Hortal J., Lobo J.M., & Jiménez-Valverde A. (2012) Basic questions in biogeography and the (lack of) simplicity of species distributions: Putting species distribution models in the right place. *Natureza e Conservação*, **10**, 108–118.

Hortal J., Roura-Pascual N., Sanders N.J., & Rahbek C. (2010) Understanding (insect) species distributions across spatial scales. *Ecography*, **33**, 51–53.

Hortal J., Tessarolo G., & Ladle R.J. (2014) Mapping biogeographical ignorance. *in prep.*, in prep.

Jiménez-Valverde A. (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, **21**, 498–507.

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.

Kamino L.H.Y., Stehmann J.R., Amaral S., De Marco P., Rangel T.F., de Siqueira M.F., De Giovanni R., & Hortal J. (2012) Challenges and perspectives for species distribution modelling in the neotropics. *Biology letters*, **8**, 324–326.

Kramer-Schadt S., Niedballa J., Pilgrim J.D., Schröder B., Lindenborn J., Reinfelder V., Stillfried M., Heckmann I., Scharf A.K., Augeri D.M., Cheyne S.M., Hearn A.J., Ross J., Macdonald D.W., Mathai J., Eaton J., Marshall A.J., Semiadi G., Rustam R., Bernard H., Alfred R., Samejima H., Duckworth J.W., Breitenmoser-Wuersten C., Belant J.L., Hofer H., & Wilting A. (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 1–14.

Kühn I., Bierman S.M., Durka W., & Klotz S. (2006) Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. *The New phytologist*, **172**, 127–39.

Ladle R.J. & Hortal J. (2013) Mapping species distributions : living with uncertainty. *Frontiers of Biogeography*, **5**, 8–9.

Lobo J.M., Baselga A., Hortal J., Jiménez-Valverde A., & Gómez J.F. (2007a) 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*, **33**, 103–114.

Lobo J.M., Jiménez-Valverde A., & Real R. (2007b) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Lobo J.M. & Martín-piera F. (1991) La creación de un banco de datos zoológico sobre los Scarabaeidae Ibero-Baleares. *Elytron*, **5**, 31–37.

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**, 158–173.

Lobo J.M. & Tognelli M.F. (2011) Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, **19**, 1–7.

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), Sinauer Associates, Inc., Sunderland, Massachussets.* pp. 293–296.

Lozier J.D., Aniello P., & Hickerson M.J. (2009) Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. *Journal of Biogeography*, **36**, 1623–1627.

Martín-Piera F. (2000) Familia Scarabaeidae. *Coleoptera, Scarabaeoidea I.* (ed. by F. Martín-Piera and J.I. López-Colón), pp. 207–432. In: Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Madrid.

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

McInerny G.J., Chen M., Freeman R., Gavaghan D., Meyer M., Rowland F., Spiegelhalter D.J., Stefaner M., Tessarolo G., & Hortal J. (2014) Information visualisation for science and policy: engaging users and avoiding bias. *Trends in Ecology & Evolution*, **20**, 148–157.

McInerny G.J. & Purves D.W. (2011) Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

Naimi B., Hamm N. a. S., Groen T. a., Skidmore A.K., & Toxopeus A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, **37**, 191–203.

Peterson A.T. (2006) Uses and requirements of ecological niche models and related dsitributional models. *Biodiversity Informatics*, **3**, 59–72.

Phillips S., Anderson R., & Schapire R. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Rangel T.F. & Loyola R.D. (2012) Labeling Ecological Niche Models. *Natureza & Conservação*, **10**, 119–126.

Regan H.M., Ben-haim Y., Langford B., Wilson W.G., Lundberg P., Andelman S.J., & Burgman M.A. (2005) Robust decision-making under severe incertainty for conservation management. *Ecological Applications*, **15**, 1471–1477.

Ripley B.D. (1996) Pattern recognition and neural networks. .

Rocchini D., Hortal J., Lengyel S., Lobo J.M., Jimenez-Valverde A., Ricotta C., Bacaro G., & Chiarucci A. (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Soberón J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology letters*, **10**, 1115–23.

Soberón J.M. (2010) Niche and area of distribution modeling: a population ecology perspective. *Ecography*, **33**, 159–167.

Swanson A.K., Dobrowski S.Z., Finley A.O., Thorne J.H., & Schwartz M.K. (2013) Spatial regression methods capture prediction uncertainty in species distribution model projections through time. *Global Ecology and Biogeography*, **22**, 242–251.

Synes N.W. & Osborne P.E. (2011) Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. *Global Ecology and Biogeography*, **20**, 904–914.

Syphard A.D. & Franklin J. (2009) Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, **32**, 907–918.

Tessarolo G. (2012) Choosing the right path for species distribution modeling. *Frontiers of Biogeography*, **4**, 91–92.

Tobler W.R. (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, **46**, 234–240.

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.

Wilson K.A. (2010) Dealing with Data Uncertainty in Conservation Planning. *Natureza e Conservação*, **8**, 145–150.

Wilson K.A., Westphal M.I., Possingham H.P., & Elith J. (2005) Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, **122**, 99–112.

Yang W., Ma K., & Kreft H. (2013) Geographical sampling bias in a large distributional database and its effects on species richness-environment models. *Journal of Biogeography*, **40**, 1415–1426.

Young J.C., Waylen K. a., Sarkki S., Albon S., Bainbridge I., Balian E., Davidson J., Edwards D., Fairley R., Margerison C., McCracken D., Owen R., Quine C.P., Stewart-Roper C., Thompson D., Tinch R., Hove S., & Watt A. (2014) Improving the science-policy dialogue to meet the challenges of biodiversity conservation: having conversations rather than talking at one-another. *Biodiversity and Conservation*, **23**, 387–404.

Supporting information:

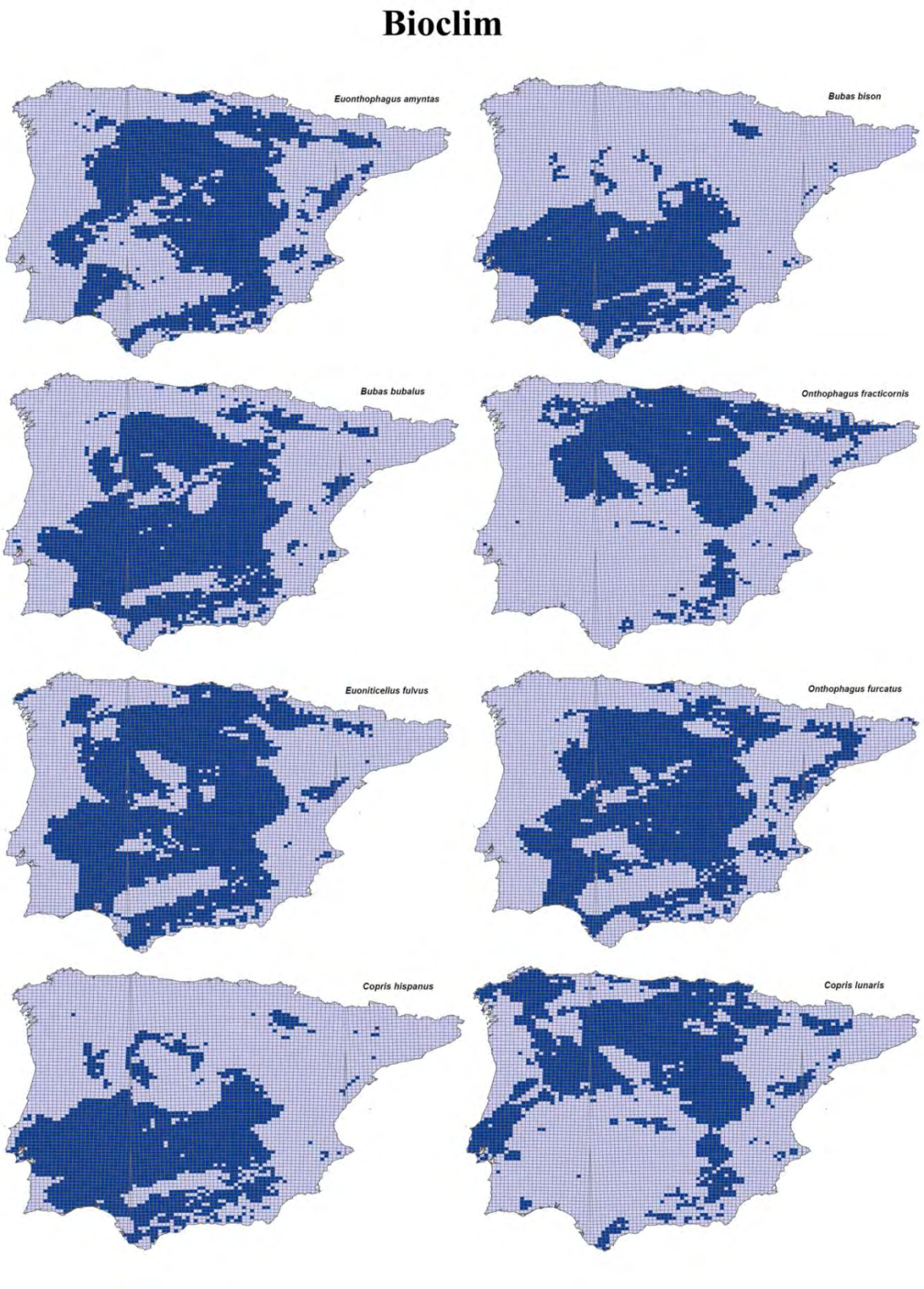
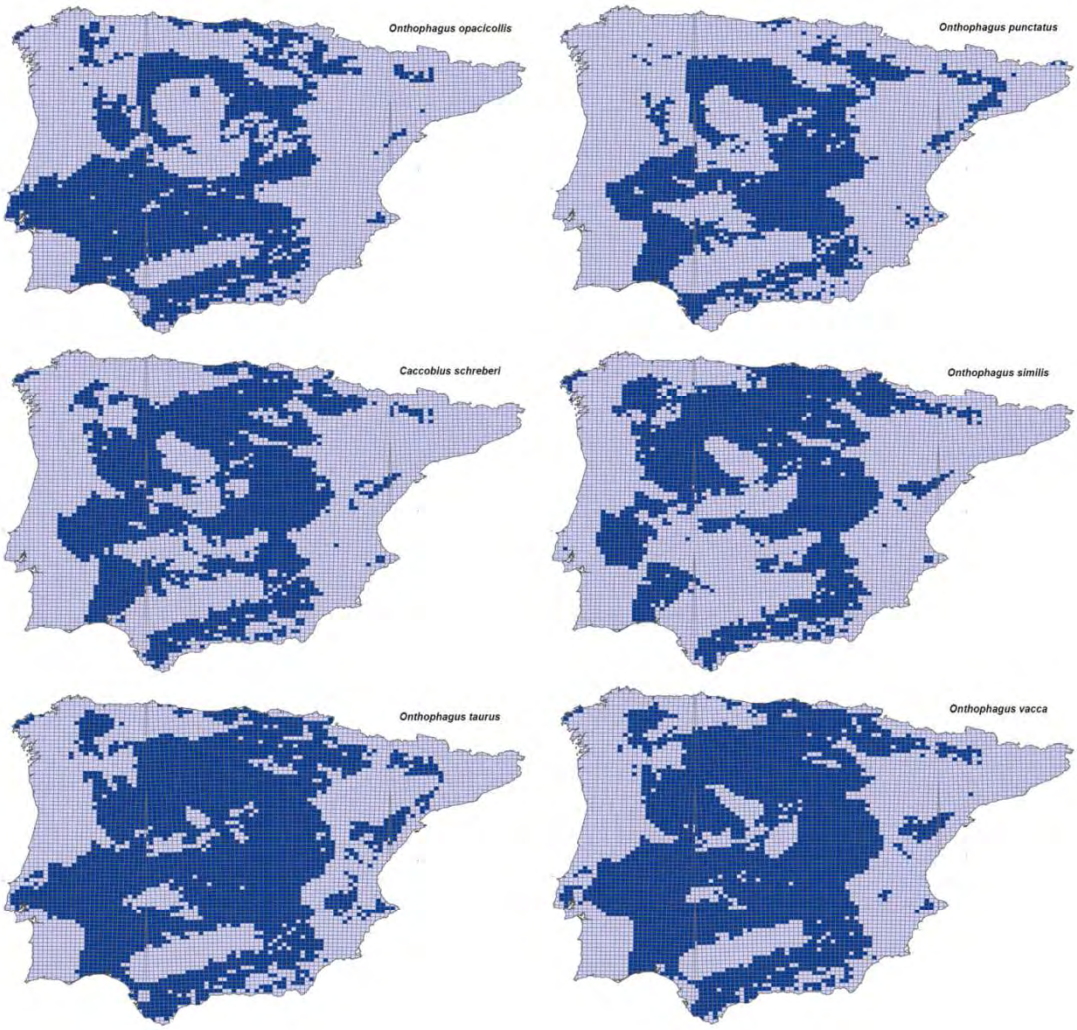


Fig S1: Occurrences predicted by Bioclim technique.

Bioclim



Cont. Fig S1

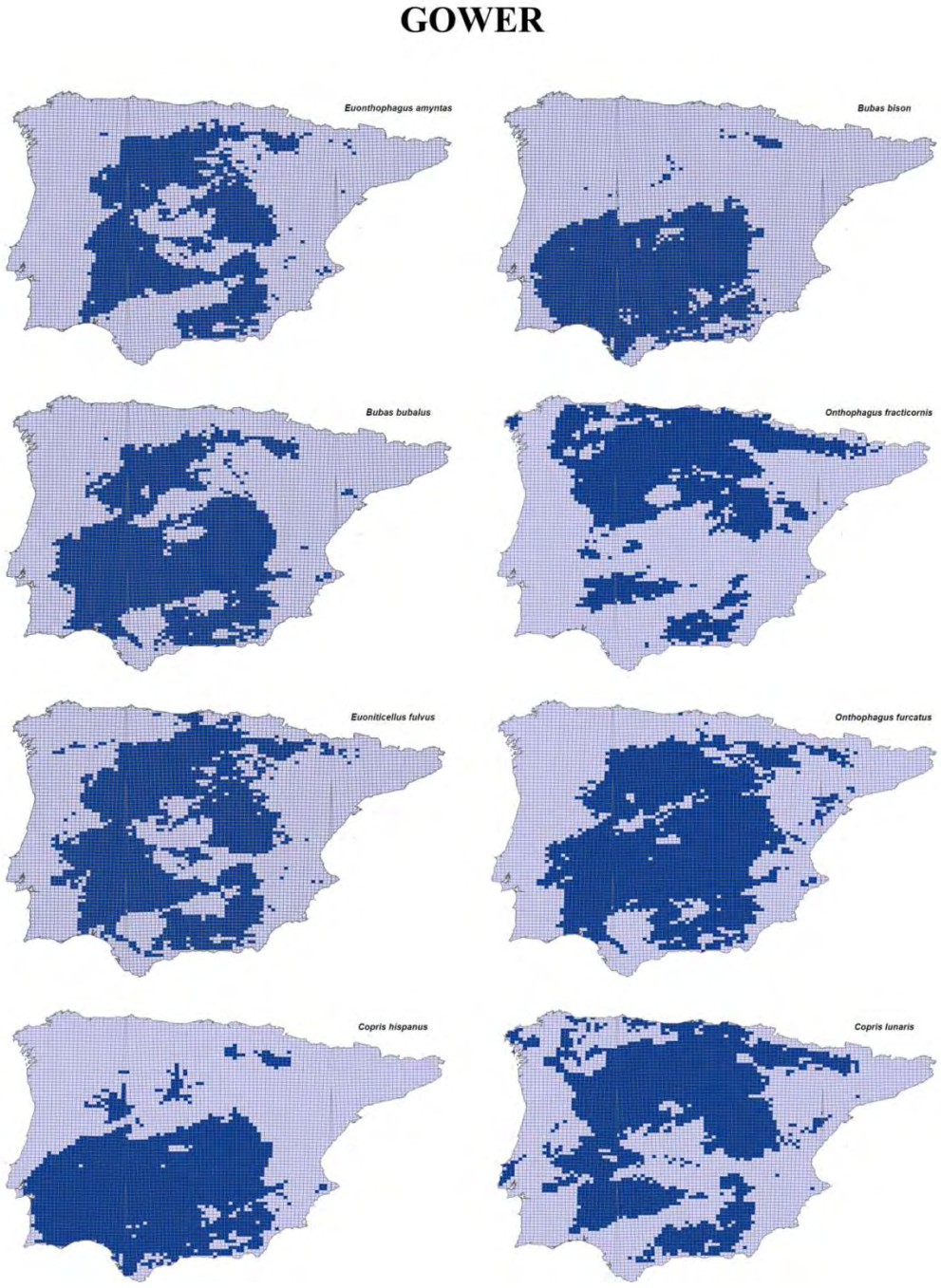
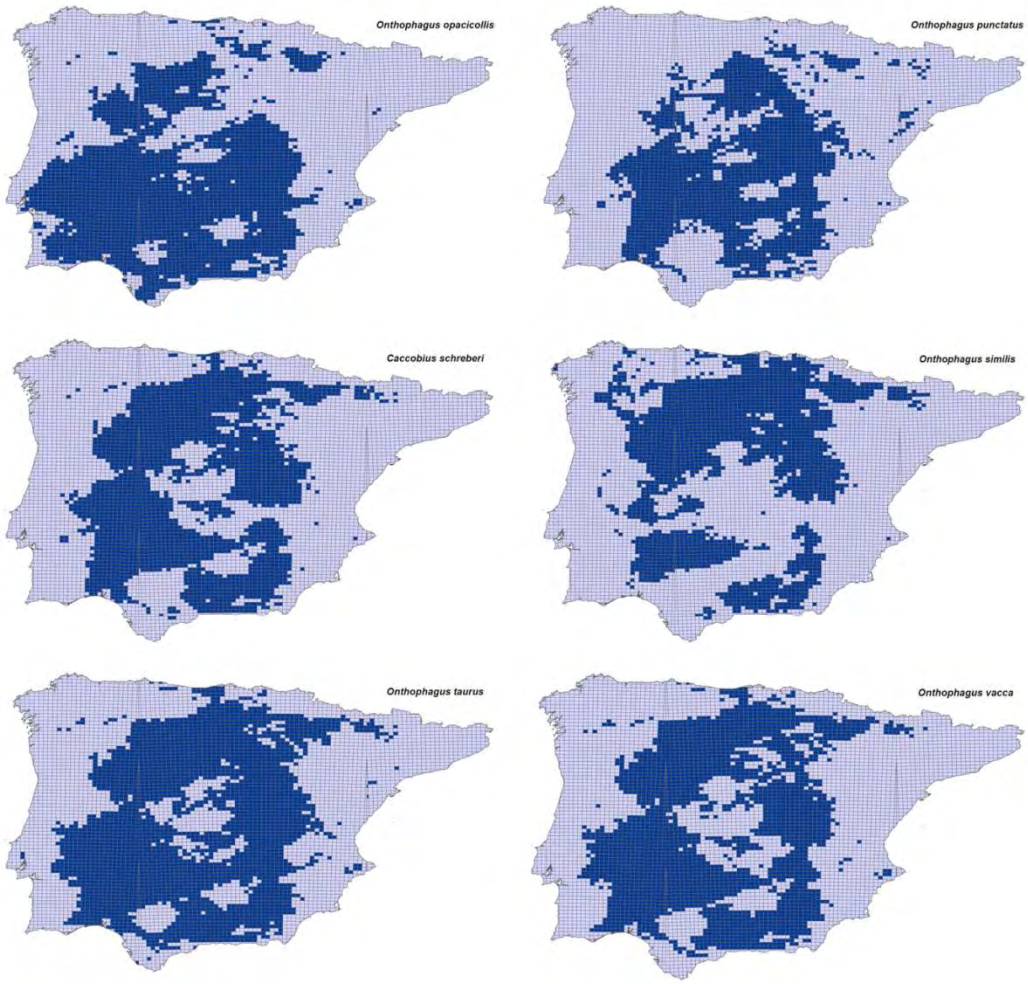


Fig. S2. Predicted occurrence by Gower distance technique.

GOWER



Cont. Fig.S2

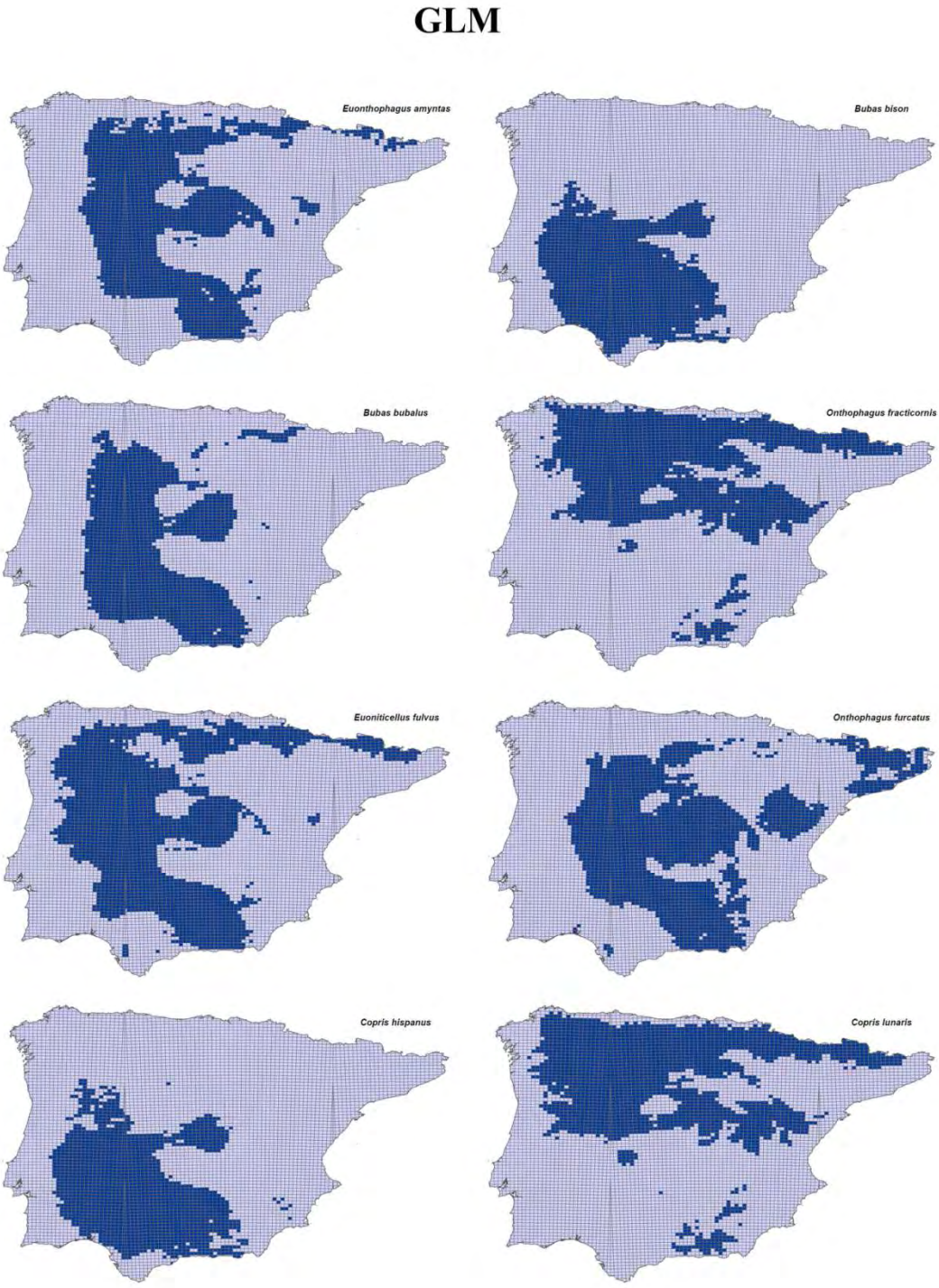
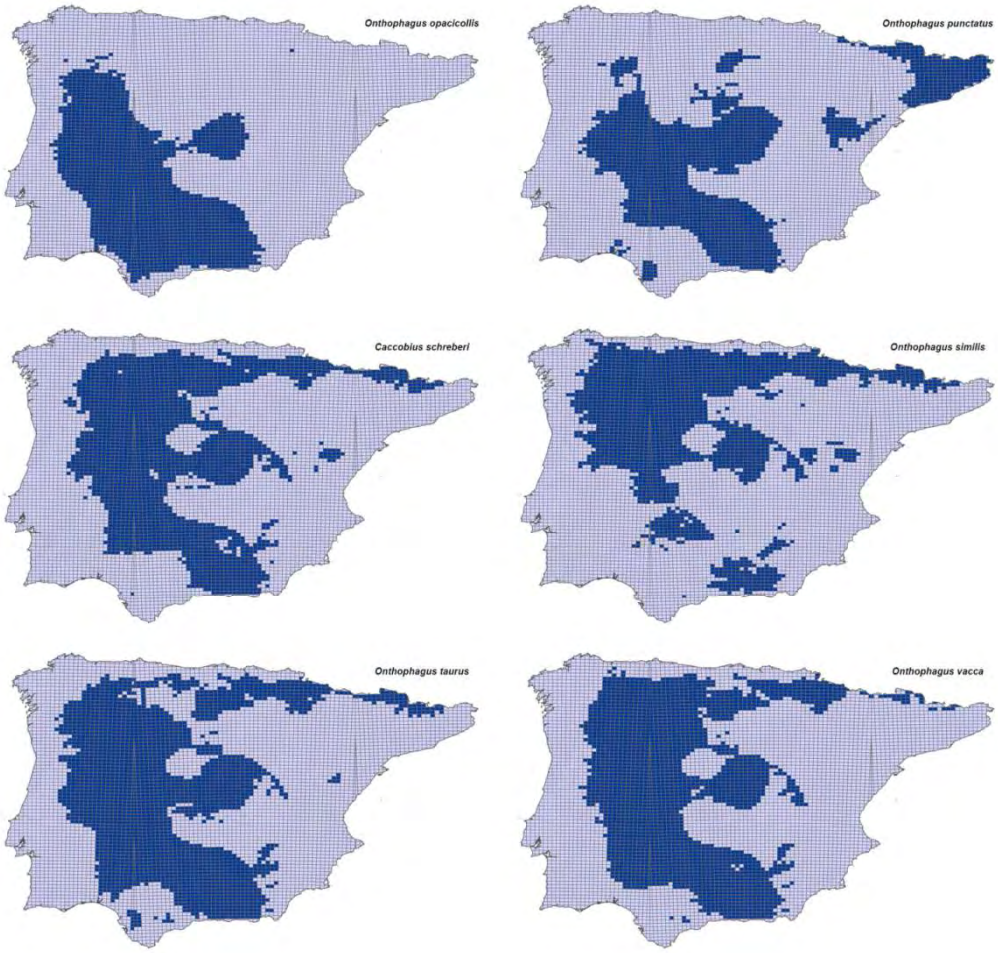


Fig. S4. Predicted occurrence by GLM distance technique.

GLM



Cont. Fig. S4

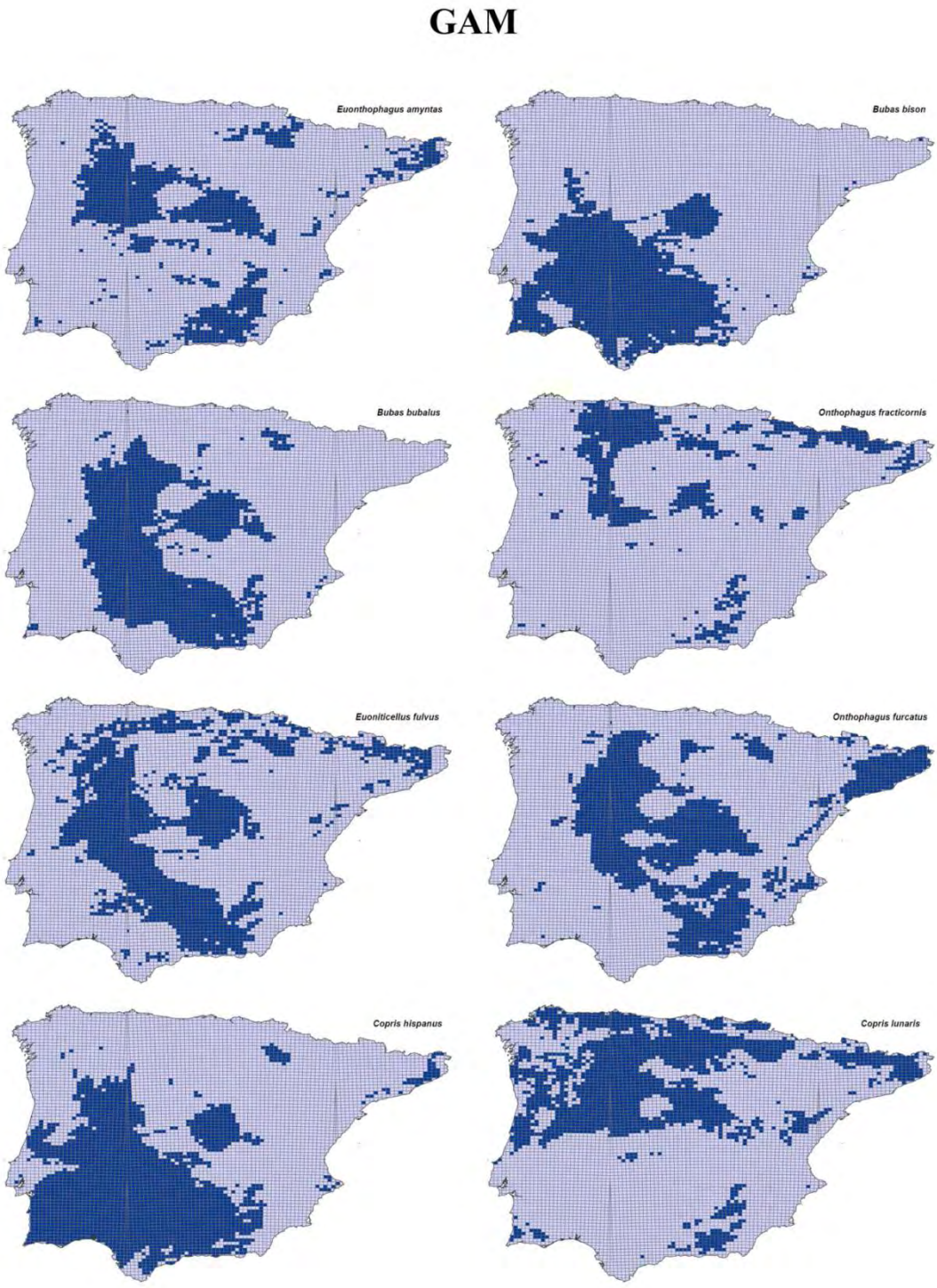
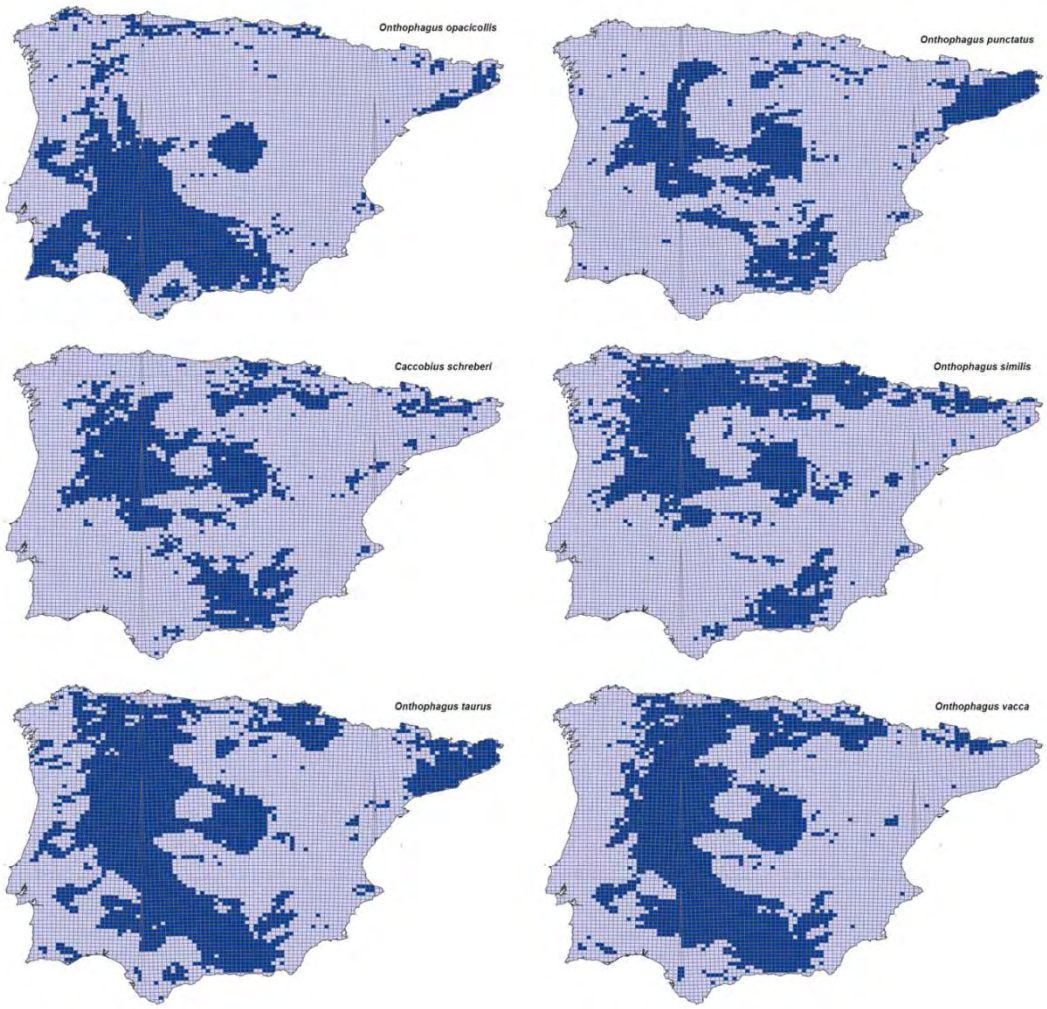


Fig.S5. Predicted occurrence by GAM distance technique.

GAM



Cont. Fig. S5

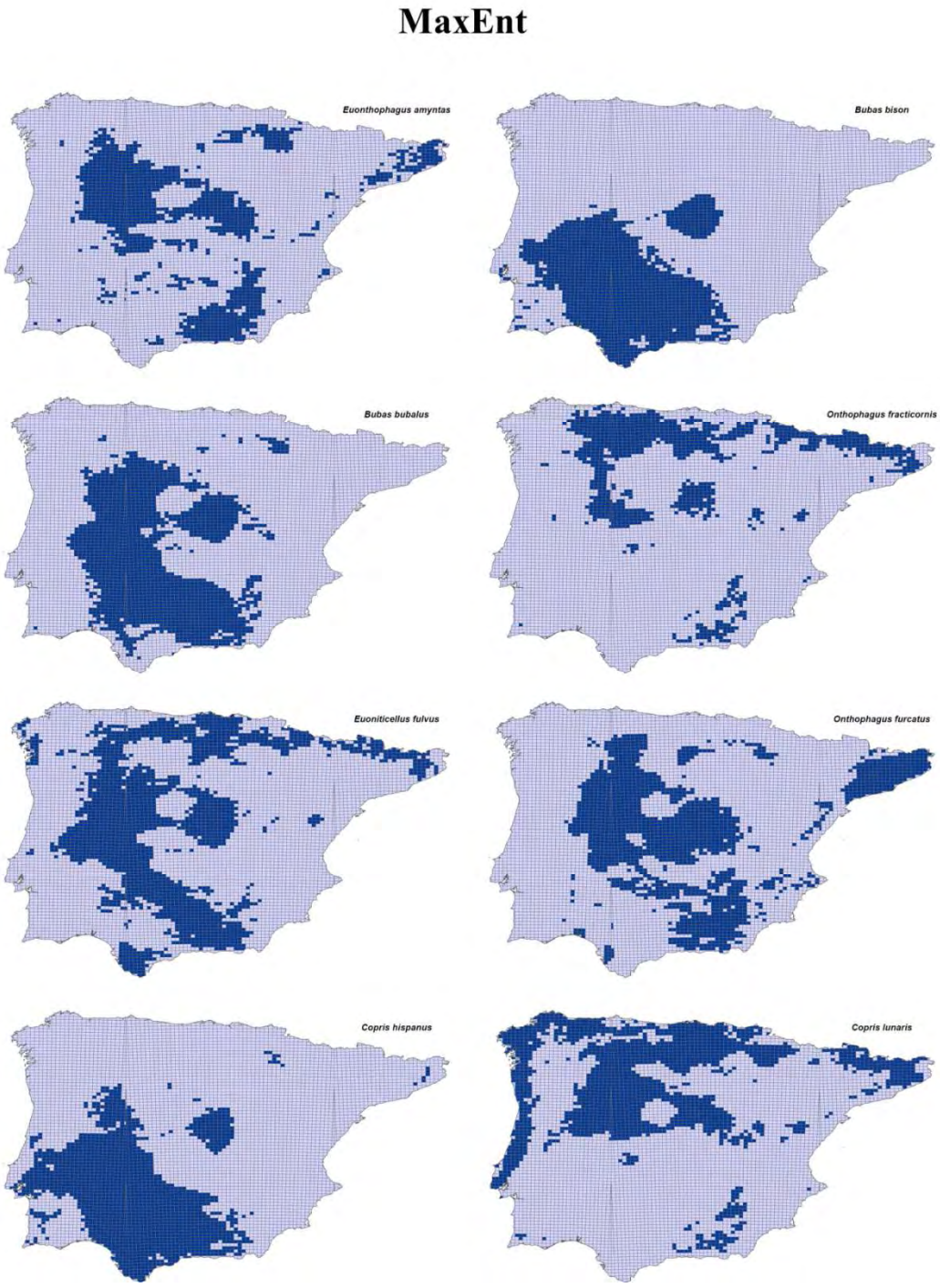
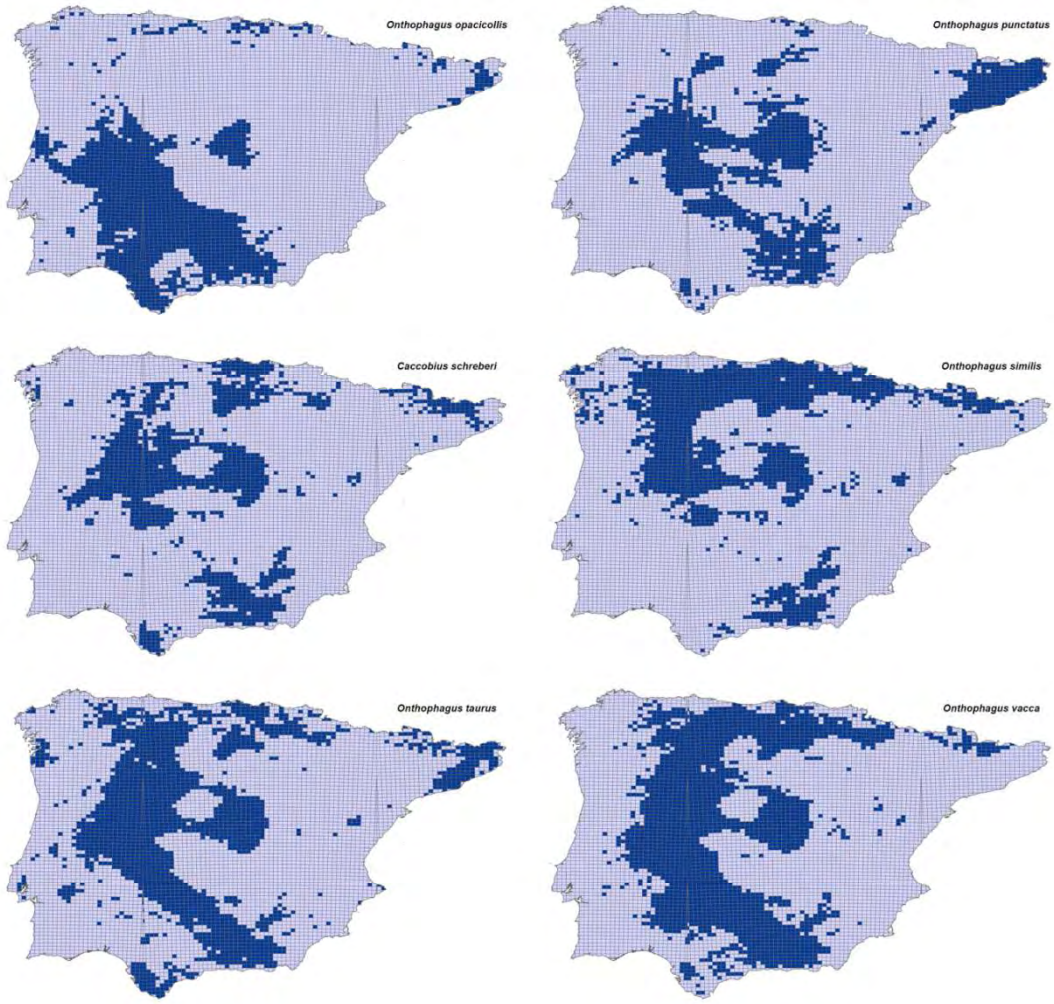


Fig. S6. Predicted occurrence by MaxEnt technique.

MaxEnt



Cont. Fig. S6

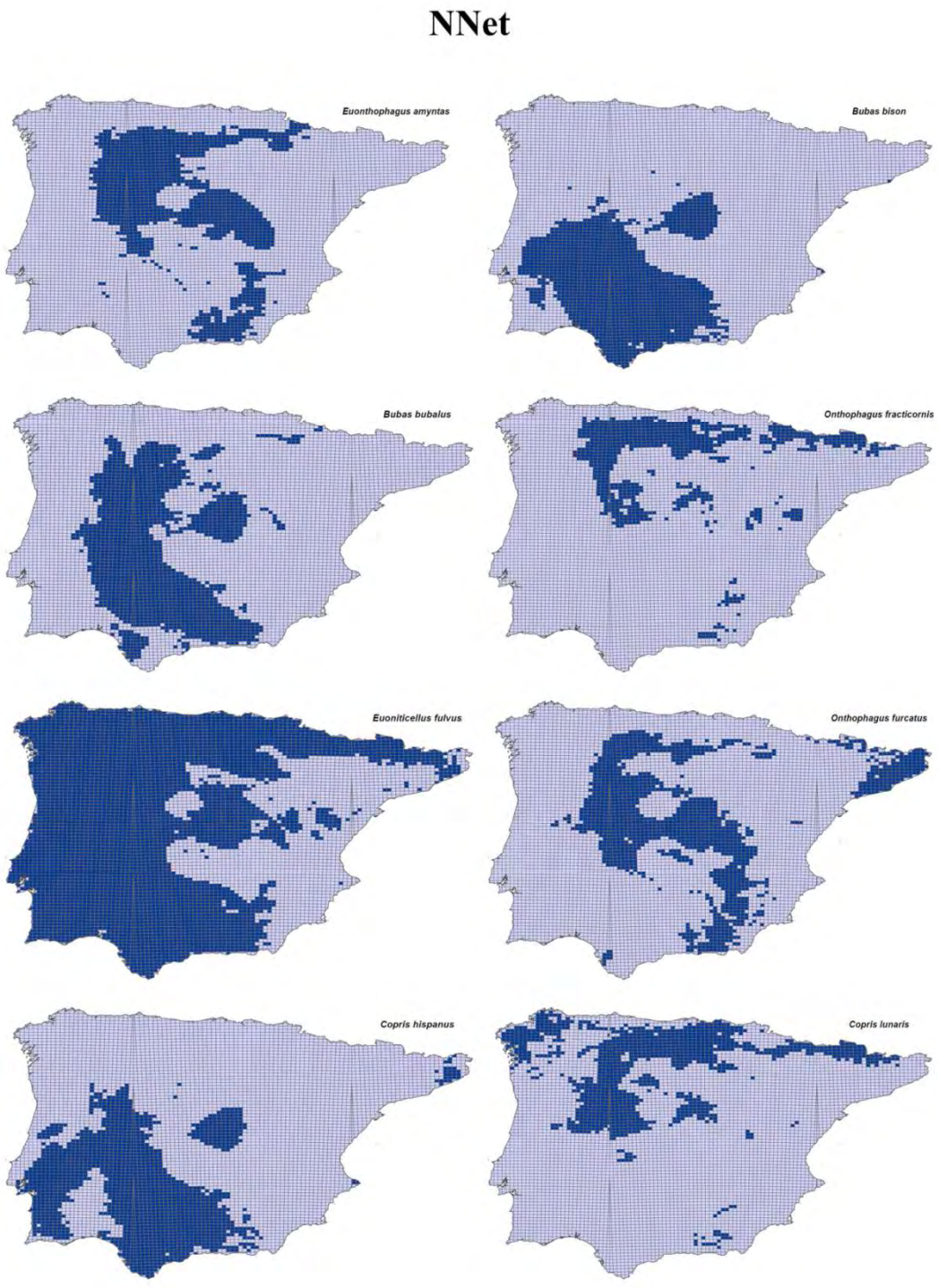
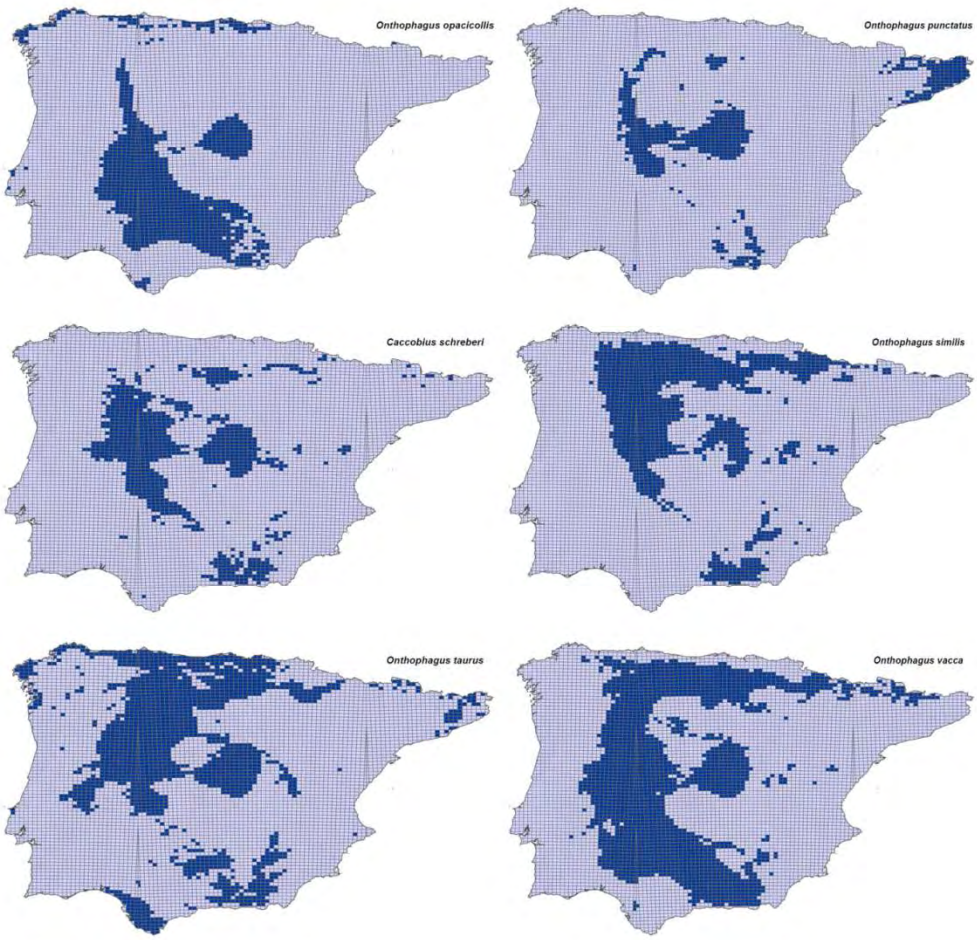


Fig. S7. Predicted occurrence by NNET distance technique.

Cont. fig xx

NNet



Cont. Fig. S7

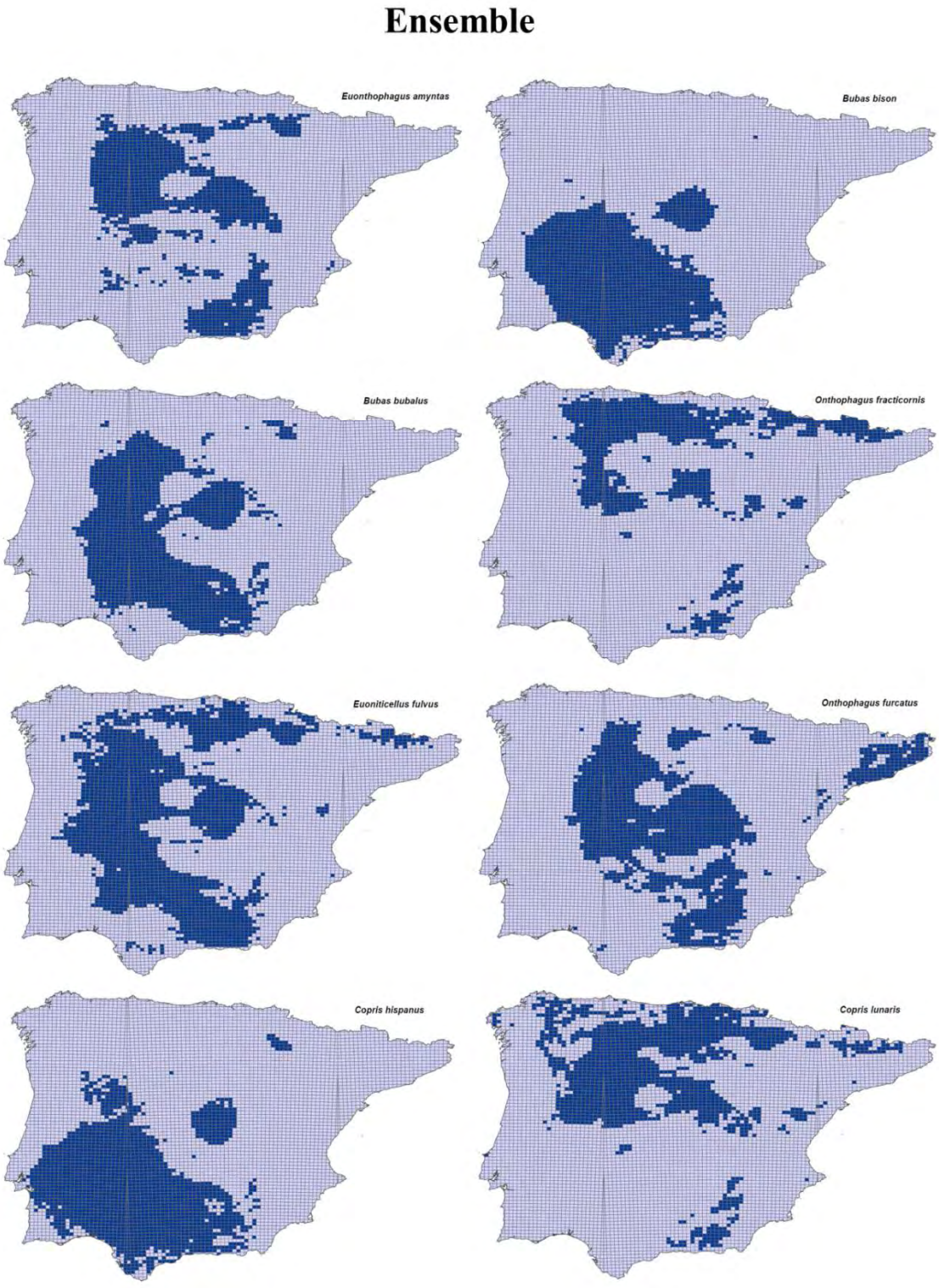
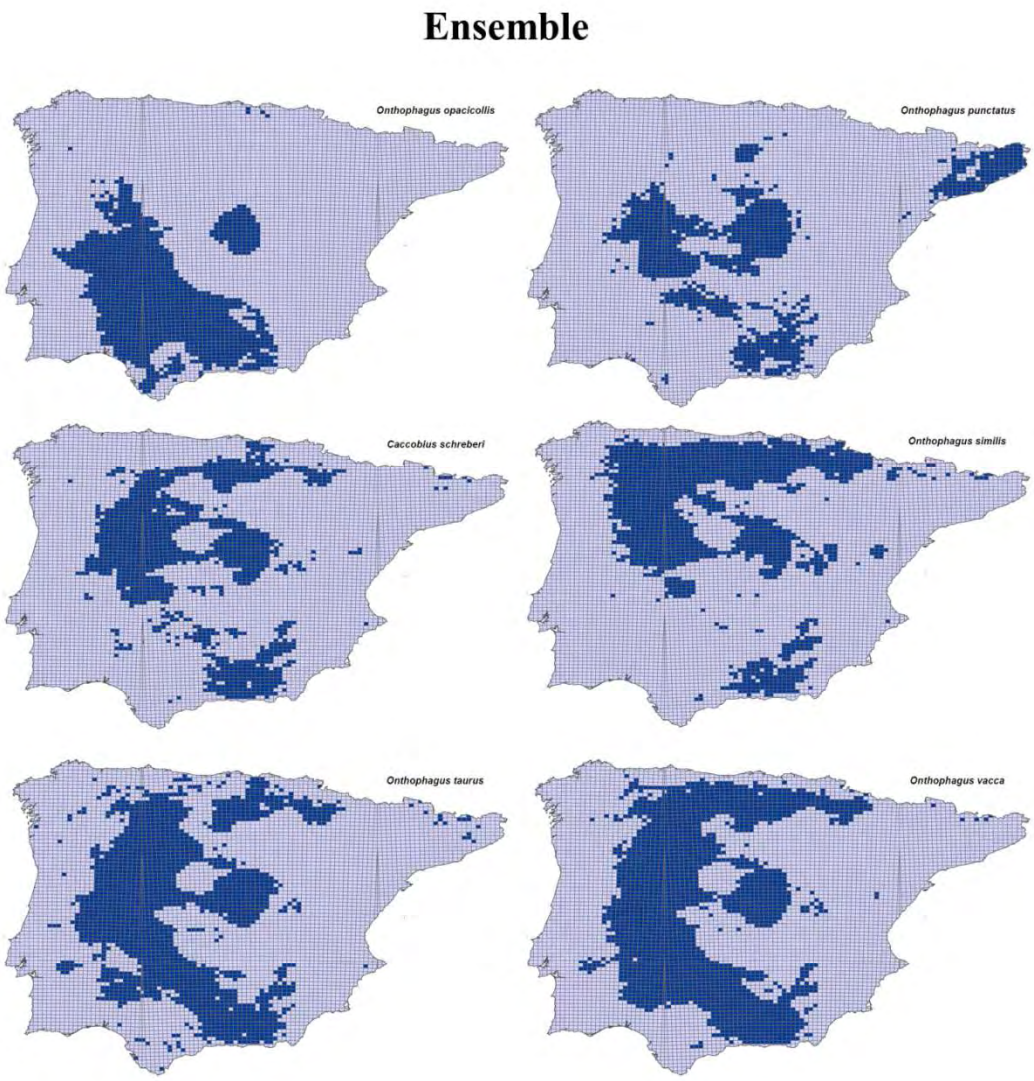


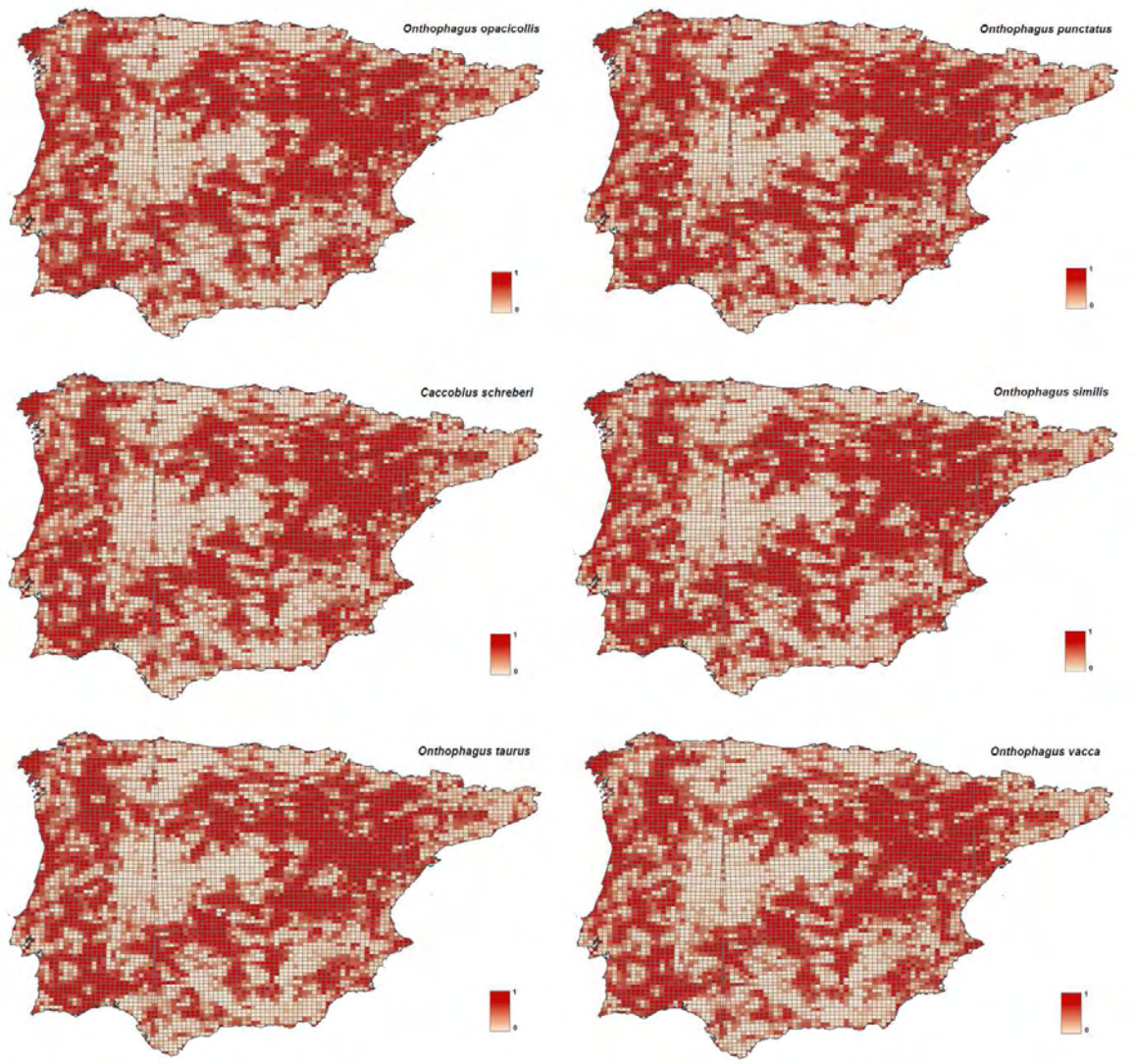
Fig. S8. Predicted occurrence by NNET technique.

Cont. fig xx



Cont. Fig. S8

Fig. S9. Maps of biogeographical ignorance for each species.


Onthophagus opacicollis
Onthophagus punctatus
Caccobius schreberi
Onthophagus similis
Onthophagus taurus
Onthophagus vacca

Cont. Fig. S9

Table S5: Number and correspondent percentage of cell in the study area with BI values greater than 0.9.

Species	Nº_cells	% of area
amyntas	3744	63.25
bison	3720	62.85
bubalus	3715	62.76
fracticornis	3786	63.96
fulvus	3706	62.61
furcatus	3636	61.43
hispanus	3708	62.65
lunaris	3754	63.42
opacicollis	3720	62.85
punctatus	3748	63.32
schreberi	3732	63.05
similis	3702	62.54
taurus	3629	61.31
vacca	3652	61.70

Appendix 1: R code for generation of Maps of Biogeographical ignorance (MoBI).

```
#########################################################
#MoBI - Maps of Biogeographical Ignorance
#########################################################
# data - Base de dados contendo nessa ordem: id da célula,specie,número de indivíduos, ano de coleta, qualificação do identificador (1 a 3).
# nr = valor mínimo de número de indivíduos na célula para que seja selecionada para o cálculo da pendente.
# trd = valor de decaymento linear para o calculo da qualidade taxonômica
      # pode ser um valor específico definido pelo usuário (valor entre 0 e 1).
      # ou calculado pela função se usada a opção trd=="procurar" (função linear onde o maior número de revisões terá o valor de zero na qualidade taxonômica.
      # se não há dados sobre o número de revisões o valor será baseado só no identificador, para isso usar a opção trd="na" e colocar a coluna referente ao numero de revisões na tabela preenchido com zero).

# DU - Data de utilização dos dados, se não for especificada a função usa o ano da data em que a análise é feita no R.
# Var - tabela com código das células, longitude, latitude e variáveis que serão utilizadas para calcular o correlograma espaço-ambiental.
# sp - vetor com o nome das espécies foco para as quais os mapas serão gerados.


MoBI<-function (data,TRD="procurar",DU="auto",sp,var){

 ##########################################################################
 # Calcular distância espaço-ambiental
 ##########################################################################

 packages<-function(x, repos="http://cran.r-project.org", ...){
      x <- deparse(substitute(x))
      if (!require(x,character.only=TRUE)){
           install.packages(pkgs=x, repos=repos, ...)
           require(x,character.only=TRUE)
      }
 }
 packages(fields)
 packages(cluster)
 # Gerar as matrizes de distância ambiental e espacial
 id<-var[,1]
 coord<-var[,2:3]
 vari<-var[,4:ncol(var)]

 # Distância ambiental
 library (cluster)
 var_dist<-daisy(vari,metric="gower")
 #write.table(as.matrix(var_dist),"var_dist.txt",sep="\t",row.names=F)
```

```
 # Distância espacial
 library(fields)
 ind<-t(combn(nrow(coord),2))
 pwc<-apply(ind,1,function(x)res<-
c(coord[x[1],1],coord[x[1],2],coord[x[2],1],coord[x[2],2]))
 pwc<-t(pwc)

 # Convert degrees to radians
 deg2rad <- function(deg) return(deg*pi/180)
 rad_cor<-deg2rad(pwc)

 # Calculates the geodesic distance between two points specified by radian
latitude/longitude using the
 # Haversine formula (hf)
 gcd.hf <- function(long1, lat1, long2, lat2) {
      R <- 6371 # Earth mean radius [km]
      delta.long <- (long2 - long1)
      delta.lat <- (lat2 - lat1)
      a <- sin(delta.lat/2)^2 + cos(lat1) * cos(lat2) * sin(delta.long/2)^2
      c <- 2 * asin(min(1,sqrt(a)))
      d = R * c
      return(d) # Distance in km
 }
 dist_km<-
apply(rad_cor,1,function(x)gcd.hf(long1=x[1],lat1=x[2],long2=x[3],lat2=x[4]))
 int<-cbind(ind,dist_km)
 spt_dist<-matrix(NA, nrow(coord),nrow(coord))
 rownames(spt_dist)<-id
 colnames(spt_dist)<-id
 for(i in 1:nrow(ind)){
      spt_dist[ind[i,2],ind[i,1]]<-int[i,3]
 }
 spt_dist<-as.dist(spt_dist)
 spt_dist<-as.matrix(spt_dist)
 var_dist<-as.matrix(var_dist)

 #padronizar as matrizes de distância p variar de 0 a 1

 #padroniza matriz de variávies ambientais
 var_standar<-var_dist# matriz nova p receber os dados da padronização
 mn_var<-min(var_dist)             # encontra a menor densidade
 mx_var<-max(var_dist)
 for (i in 1:length(var_dist)){ # para cada célula do mapa
      var_standar[i]<- ((var_dist[i])- mn_var)/(mx_var-mn_var) # calcula o novo
valor da célula variando de 0 a 1.
 }

 #padroniza a matriz espacial
```

```
 spt_standard<-spt_dist
 mn_spt<-min(spt_dist)
 mx_spt<-max(spt_dist)
 for (i in 1:length(spt_dist)){
      spt_standard[i]<- ((spt_dist[i])- mn_spt)/(mx_spt-mn_spt)
 }
 # multiplicar as duas distâncias
 spt.envir.dist<-spt_standard*var_standar
 write.table(var_dist,"var_dist.txt",sep="\t",row.names=F,quote=F)
 write.table(spt_dist,"spt_dist.txt",sep="\t",row.names=F,quote=F)
 write.table(spt.envir.dist,"spt.envir.dist.txt",sep="\t",row.names=F,quote=F)

 ########################################
 # Calcular o correlograma das variáveis
 ########################################
 packages(vegan)
 library(vegan)
 ge<-mantel.correlog(var_dist,spt_dist,nperm=200)
 for (g in 1: length(ge$mantel.res[,3])){
      if(ge$mantel.res[g,3]<0){
           max.dis<-ge$mantel.res[(g-1),1]
           break
      }
 }
 ######################
 # COMPLETENESS
 ######################
 Acum_curve<-function(data,nr){
      data<-data[complete.cases(data[,3]),]

      #cria tabelas com número de espécies e número de registros por célula
      cel1<-sort(unique(data[,1]))
      sp_name<-sort(unique(data[,2]))

      abund<-matrix(0,length(cel1),(length(sp_name)))
      abund<-data.frame(cel1,abund,stringsAsFactors=FALSE)
      colnames(abund)<-c("Id",sp_name)

      regist<-matrix(0,length(cel1),(length(sp_name)))
      regist<-data.frame(cel1,regist,stringsAsFactors=FALSE)
      colnames(regist)<-c("Id",sp_name)

      if((sp_name[1] %in% names(abund))==FALSE){
           print("You should import your data without factors. You can use
[as.is=T] when importing")
           stop
      }
      for(i in 1:length(data[,1])){
           for (j in 1:length(abund[,1])){
```

```
            if(data[i,1]==abund[j,1]){
                  for(s in 1:length(sp_name)){
                        if(sp_name[s]==data[i,2]){
                              abund[j,(s+1)]<-abund[j,(s+1)]+data[i,3]
                              regist[j,(s+1)]<-regist[j,(s+1)]+1
                        }
                  }
            }
      }
}
#calcula número de espécies por célula
pre_abs<-regist[,-1]
pre_abs<-ifelse(pre_abs>0,1,0)
pre_abs2<-data.frame(regist[,1],pre_abs)
colnames(pre_abs2)<-c("id",colnames(pre_abs))
write.table(pre_abs2,"pres_abes.txt",sep="\t",row.names=F,quote=F)
numb_sp<-rowSums(pre_abs)
#calcula número de registro por célula
num_regis<-rowSums(regist[,-1])

#Cria matriz para receber valores das pendentes
Compl<-matrix(NA,nrow(abund),4)
colnames(Compl)<-c("UTM10","Complet_ind","Complet_reg","Number_regist")

#Adicionar número de registros por célula e número de registro por
espécie por célula
Compl[,4]<-num_regis

indx<-which(Compl[,4]>=nr)
cal_pend<-Compl[indx,]
calc_abund<-abund[indx,]
calc_regist<-regist[indx,]
#########################################################################
#curva de acumulação por individuo

for(i in 1: nrow(calc_abund)){
      ve<-numeric(0)
      for(j in 2:ncol(calc_abund)){
            re<-rep(j,calc_abund[i,j])
            ve<-c(ve,re)
      }
      pend_I<-numeric(0)
      for (v in 1:100){
            tab_I<-matrix(NA,length(ve),2)
            for(a in 1:length(ve)){
                  ale_I<-unique(sample(ve,a))
                  riq_I<-length(ale_I)
                  lin_I<-c(a,riq_I)
                  tab_I[a,]<-lin_I
```

```
                }
                #calcula a pendente
                pend_I[v]<-as.numeric((tab_I[length(tab_I[,1]),2]-
tab_I[(length(tab_I[,1])-1),2]))
          }
          p_fin<-1-(mean(pend_I))
          Compl[indx[i],1:2]<-c(calc_abund[i,1],p_fin)
     }
     ##########################################################################
     #curva de acumulação por registro
     for(i in 1: nrow(calc_regist)){
          ve<-numeric(0)
          for(j in 2:ncol(calc_regist)){
                re<-rep(j,calc_regist[i,j])
                ve<-c(ve,re)
          }
          pend_R<-numeric(0)
          for (v in 1:100){
                tab_R<-matrix(NA,length(ve),2)
                for(a in 1:length(ve)){
                     ale_R<-unique(sample(ve,a))
                     riq_R<-length(ale_R)
                     lin_R<-c(a,riq_R)
                     tab_R[a,]<-lin_R
                }
                #calcula a pendente
                pend_R[v]<-as.numeric((tab_R[length(tab_R[,1]),2]-
tab_R[(length(tab_R[,1])-1),2]))
          }
          p_reg<-1-(mean(pend_R))
          Compl[indx[i],3]<-p_reg
     }
     Compl[,1]<-cel1
     Compl<-Compl[indx,]
     #Exporta a tabela com as pendentes p as células elecionadas pelo critério
do usuário
     write.table(Compl,"Completeness_final.txt",sep="\t",row.names=F)
     return(Compl)
 }

 ############################
 # Decay taxonomic quality
 ############################


 Tax_qual<-function(data,trd="procurar"){
     options("scipen"=500)
     cel1<-unique(data[,1])
     quali_final<-matrix(NA,length(cel1),2)
```

```
      colnames(quali_final)<-c("Id",paste("Tax_qual_",sep=""))
      quali_final[,1]<-cel1
      if(trd=="procurar"){
            trd<-1/(max(data[,6]))
      }

      if(trd=="na"){
            trd<-1
      }

      qual<-numeric(0)

      for (i in 1:nrow(data)){
            if (data[i,5]==3){
                  qua<-1-(trd*data[i,6]) # qualidade iniciando em 1
                  ifelse(qua<0,qual[i]<-0,qual[i]<-qua)
            }
            if(data[i,5]==2){
                  qua<-0.9-(trd*data[i,6]) # qualidade inciiando em 0.9
                  ifelse(qua<0,qual[i]<-0,qual[i]<-qua)
            }
            if(data[i,5]==1){
                  qua<-0.75-(trd*data[i,6]) # qualidade iniciando em 0.75
                  ifelse(qua<0,qual[i]<-0,qual[i]<-qua)
            }
      }
      qual<-cbind (data[,1],qual)
      qual<-qual[order(qual[,1]),]
      write.table(qual,"Taxonomic_quality_s_revisao.txt",sep="\t",row.names=F)
      return(qual)
}

########################
# TEMPORAL DECAY
########################
# DU - ano do estudo, ou data de utilização dos dados, se não for especificada
# a função usa o ano corrente,o ano da data em que a análise é feita no R.

Temp.dec<-function(data, DU="auto"){
      options("scipen"=500)
      if(DU=="auto"){#Encontra o ano atual
            DU<- as.numeric(format(Sys.Date(), "%Y"))
      }
      dif<-sapply(data[,4],function(x,DU)dif<-(DU-x),DU=DU)

      #plot(dif,res)
      h<-max(dif,na.rm=T)
      res<-sapply(dif,function(x,h)(1-((x/h)^2))^2,h=h)#Calcula a função kernel
bisquared
```

```
      Temp_qual<-cbind(data[,1],res)
      colnames(Temp_qual)<-c("id","temp_qual")
      write.table(Temp_qual,"Temp_qual.txt",sep="\t",row.names=F)
      return(Temp_qual)
  }

  ##############################
  # FINAL TEMPORAL DECAY
  ##############################

  completenes<-Acum_curve(data,nr=50)
  taxononomic_quality<-Tax_qual(data,trd=trd)
  time_decay<-Temp.dec (data, DU=DU)
  comple<-completenes
  part<-
as.data.frame(cbind(data[,1:2],as.numeric(taxononomic_quality[,2]),as.numeric(ti
me_decay[,2])))
  part[, c("compl_ind","compl_regis")]<-NA
  for (i in 1:nrow(completenes)){
      indx<-which(part[,1]==completenes[i,1])
      part[indx,5:6]<-rep(as.numeric(completenes[i,2:3]),each=length(indx))
  }
  colnames(part)<-c("ID","sp","Tax_qual","Time_qual","Com_ind","Com_reg")
  part[is.na(part)]<-0
  part$CB<-NA
  part$CB<-rowSums(part[,c(3,4,5)])

  write.table(part,"part.txt",sep="\t",row.names=F)

  options("scipen"=999)

  for (e in 1: length(sp)){
      selec<-part[which(part[,2]==sp[e]),]
      selec<-selec[,-c(3,4)]
      cepre<-unique(selec[,1])
      adce<- data.frame(matrix(NA,length(var[,1]),(ncol(part))))
      adce[,1]<-levels(var[,1])
      adce[,2]<-"nsp"
      colnames(adce)<-colnames(part)
      rem<-which(adce[,1] %in% cepre)
      no_prese<-adce[-rem,]
      no_prese<-no_prese[,-c(3,4)]

      for(i in 1:nrow(no_prese)){
          cur<-part[which(part[,1]==no_prese[i,1]),]
          if(nrow(cur)!=0){
              tg<-unique(cur[,2])
              me<-numeric(0)
```

```
            for(y in 1:length(tg)){
                  baba<-which(cur[,2]==tg[y])
                  masp<-max(cur[baba,7])
                  me[y]<-masp
            }

            no_prese[i,5]<-mean(me)
            no_prese[i,3]<-cur[1,5]
            no_prese[i,4]<-cur[1,6]
      }
}
no_prese[is.na(no_prese)]<-0
colnames(no_prese)<-colnames(selec)
infi<-rbind(selec,no_prese)
infi<-infi[order(infi[,1]),]
dupli<-rle(infi[,1])
bi_part<-data.frame(matrix(NA, length(dupli$values),2))
colnames(bi_part)<-c("Id","BI")
for (t in 1:length(dupli$lengths)){
      bi_part[t,1]<-dupli$values[t]
      if(dupli$lengths[t]>1){
            calc<-infi[which(infi[,1]==dupli$values[t]),]
            fivalue<-max(calc[,5])
            bi_part[t,2]<-fivalue
      }else{
            calc<-infi[which(infi[,1]==dupli$values[t]),]
            bi_part[t,2]<-calc[1,5]
      }
}
name<-paste("bi_part_",sp[e],".txt",sep='')
write.table(bi_part,name,sep="\t",row.names=F)

###############
# Interpolação
###############

bio.tot<-as.data.frame(matrix(NA,length(id),3))#cria matriz p receber os
dados finais
colnames(bio.tot)<-c("id","IgB","complete")

for (i in 1:length(id)){
      indr<-which(bi_part[,1]==id[i])
      bio.tot[i,1]<-bi_part[indr,1]
      bio.tot[i,2]<-bi_part[indr,2]

}
b<-5
bi.interpol<-matrix(NA,nrow(bio.tot),4)
colnames(bi.interpol)<-c("id","IDW_todas","IDW_com_CB","IDW_com_compl")
```

```
	bi.interpol[,1]<-levels(id)
	for (s in 1:nrow(comple)){
		indcom<-which(bio.tot[,1]==comple[s,1])
		if(length(indcom)>0){
			bio.tot[indcom,3]<-comple[s,2]
		}
	}
	bio.tot[is.na(bio.tot)]<-0
	bio.tot[,3]<-as.numeric(bio.tot[,3])
	bio.tot$presence<-0
	for (p in 1:length(cepre)){
		indt<-which(bio.tot[,1]==cepre[p])
		if(length(indt)>0){
			bio.tot[indt,4]<-1
		}
	}

	# A interpolação está baseada só nas células do entorno dentro do raio da
max.dis

	for (c in 1:nrow(bio.tot)){
		indx2<-which(spt_dist[c,]<= max.dis & spt_dist[c,]!=0)
		cur<-data.frame(bio.tot[indx2,],spt.envir.dist[indx2,c])
		cur<-cur[complete.cases(cur),]

		if(length(cur[,1])>0){
			divt<-length(cur[,1])/2
			indprest<-which(cur[,4]==1)
			numet<-0
			denot<-0

			if(length(indprest)<divt & length(indprest)>0){
				pesoprest<-divt/length(indprest)
				pesoabest<-divt/(length(cur[,1])-length(indprest))

				for(t in 1:length(cur[,1])){
					if(cur[t,4]==1){
						nct<-(pesoprest/(cur[t,5]^b))*cur[t,2]
						dct<-(pesoprest/(cur[t,5]^b))
					}else{
						nct<-(pesoabest/(cur[t,5]^b))*cur[t,2]
						dct<-(pesoabest/(cur[t,5]^b))
					}
					numet<-numet+nct
					denot<-denot+dct
				}
			}
			if(length(indprest)>divt || length(indprest)==0){
				for(t in 1:length(cur[,1])){
```

```
                nct<-(as.numeric(cur[t,2])/(cur[t,5]^b))
                dct<-1/(cur[t,5]^b)
                numet<-numet+nct
                denot<-denot+dct
           }
     }
     z.inter_t<-numet/denot
     bi.interpol[c,2]<-z.inter_t
}else{
     bi.interpol[c,2]<-0
}

indbi<-which(cur[,2]>0)
curbi<-cur[indbi,]
if(length(curbi[,1])>1){
     divb<-length(curbi[,1])/2
     indpresb<-which(curbi[,4]==1)
     numeb<-0
     denob<-0

     if(length(indpresb)<divb & length(indpresb)>0){
          pesopresb<-divb/length(indpresb)
          pesoabesb<-divb/(length(curbi[,1])-length(indpresb))
          for(z in 1:length(curbi[,1])){
                if(curbi[z,4]==1){
                     ncb<-(pesopresb/(curbi[z,5]^b))*curbi[z,2]
                     dcb<-(pesopresb/(curbi[z,5]^b))
                }else{
                     ncb<-(pesoabesb/(curbi[z,5]^b))*curbi[z,2]
                     dcb<-(pesoabesb/(curbi[z,5]^b))
                }
                numeb<-numeb+ncb
                denob<-denob+dcb
           }
     }
     if(length(indpresb)>divb || length(indpresb)==0){
          for(z in 1:length(curbi[,1])){
                ncb<-(as.numeric(curbi[z,2])/(curbi[z,5]^b))

                dcb<-1/(curbi[z,5]^b)
                numeb<-numeb+ncb
                denob<-denob+dcb
           }
     }
     z.inter_b<-numeb/denob
     bi.interpol[c,3]<-z.inter_b
}else{
     bi.interpol[c,3]<-0
}
```

```
            indc<-which(cur[,3]>0)
            curco<-cur[indc,]
            if(length(curco[,1])>1){
                  divc<-length(curco[,1])/2
                  indpresc<-which(curco[,4]==1)
                  numec<-0
                  denoc<-0

                  if(length(indpresc)<divc & length(indpresc)>0){
                        pesopresc<-divc/length(indpresc)
                        pesoabesc<-divc/(length(curco[,1])-length(indpresc))
                        for(x in 1:length(curco[,1])){
                              if(curco[x,4]==1){
                                    ncc<-(pesopresc/(curco[x,5]^b))*curco[x,2]
                                    dcc<-(pesopresc/(curco[x,5]^b))
                              }else{
                                    ncc<-(pesoabesc/(curco[x,5]^b))*curco[x,2]
                                    dcc<-(pesoabesc/(curco[x,5]^b))
                              }
                              numec<-numec+ncc
                              denoc<-denoc+dcc
                        }
                  }
                  if(length(indpresc)>divc || length(indpresc)==0){
                        for(x in 1:length(curco[,1])){
                              ncc<-(as.numeric(curco[x,2])/(curco[x,5]^b))

                              dcc<-1/(curco[x,5]^b)
                              numec<-numec+ncc
                              denoc<-denoc+dcc
                        }
                  }
                  z.inter_c<-numec/denoc
                  bi.interpol[c,4]<-z.inter_c
            }else{
                  bi.interpol[c,4]<-0
            }
      }
      write.table(bi.interpol,"bi.interpol_2.txt",sep="\t",row.names=F)
      bi.interpol<-bi.interpol[,-1]
      FINAL_VALUE<-cbind(bio.tot,bi.interpol)
      nome<-paste("BI_FINAL_VALUE_2pesos_",sp[e],".txt",sep="")
      write.table(FINAL_VALUE,nome,sep="\t",row.names=F,quote=F)
 }
} #FIM
```