ORIGINAL ARTICLE

Delimiting the geographical background in species distribution modelling

Pelayo Acevedo[1,2,3*], Alberto Jiménez-Valverde[1], Jorge M. Lobo[4] and Raimundo Real[1]

1 - Biogeography, Diversity, and Conservation Research Team, Faculty of Sciences, University of Málaga, E-29071 Málaga, Spain.

2 - CIBIO/UP – Centro de Investigação em Biodiversidade e Recursos Genéticos da Universidade do Porto, Campus Agrario de Vairão, 4485-661 Vairão, Portugal.

3 - Instituto de Investigación en Recursos Cinegéticos (CSIC-UCLM-JCCM), E-13071 Ciudad Real, Spain.

4 – Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales, E-28006 Madrid, Spain.

* Correspondence: Pelayo Acevedo, Biogeography, Diversity, and Conservation Research Team, Faculty of Sciences, University of Málaga, E-29071 Málaga, Spain.

E-mail: pacevedo@uma.es

Running title: The geographical background in species distribution modelling

ABSTRACT

Aim The extent of the study area (geographical background, GB) can strongly affect the results of species distribution models (SDMs), but as yet we lack objective and practicable criteria for delimiting the appropriate GB. We propose an approach to this problem using trend surface analysis (TSA) and provide an assessment of the effects of varying the GB extent on the performance of SDMs for four species.

Location Mainland Spain.

Methods Using data for four well-known wild ungulate species and different GBs delimited with a TSA, we assessed the effects of the GB extent on the predictive performance of the SDMs; specifically, on model calibration (Miller's statistic) and discrimination (AUC, sensitivity and specificity), and on the tendency of the models to predict environmental potential when they are projected beyond their training area.

Results In the training area, discrimination significantly increased and calibration decreased as the GB was enlarged. In contrast, as GB was enlarged, both discriminatory power and calibration decreased when assessed in the core area of the species distributions. When models trained using small GBs were projected beyond their training area, they showed a tendency to predict higher environmental potential for the species than those models trained using large GBs.

Main conclusions By restricting the GB extent using a geographical criterion, model performance in the core area of the species distribution can be significantly improved. Large GBs make models demonstrate high discriminatory power but are barely informative. By delimiting GB using a geographical criterion, the effect of historical events on model parameterization may be reduced. Thus, purely environmental models are obtained which, when projected onto a new scenario, depict the potential

distribution of the species. We therefore recommend the use of TSA in geographically delimiting the GB for use in SDMs.

Keywords **Calibration, discrimination, environmental potential, extent, geographical background, historical factors, Spain, species distribution models, trend surface analysis, ungulate distributions.**

INTRODUCTION

Recent studies have shown that the extent of the study area – or geographical background (GB) – in species distribution modelling (SDM) has a strong effect on the parameterization and evaluation of the models (Barve *et al.*, 2011). If the GB is too small to fully represent the ranges of the species, then the importance of coarse-scale factors such as climate may be underestimated when one delimits the species distribution (Jiménez-Valverde *et al.*, 2011a; Sánchez-Fernández *et al.*, 2011). On the other hand, if the GB is very large then the ability of the models to tease out the fine-scale conditions that actually determine species distribution will be limited (Lobo *et al.*, 2010). VanDerWal *et al.* (2009) showed that as the GB extent decreases, so does the number of variables included in the models, which in turn affects the predicted spatial patterns.

The effects of the extent of the GB on the discriminatory power of a model, i.e. the effectiveness of the scoring rule (suitability value in a broad sense) in separating instances of presence from those of absence, are noteworthy (Lobo *et al.*, 2008; Barve *et al.*, 2011). Higher and more significant discriminatory values can be obtained simply by increasing the GB extent such that the number of uninhabited and unsuitable localities under consideration increases. In this way, it is easy to obtain models with high discriminatory power but with little informational content (Jiménez-Valverde *et al.*, 2008).

Anderson & Raza (2010), working with sister species, demonstrated the effects of the GB extent on model transferability, and discussed their results within a potential versus realized distribution framework (see Jiménez-Valverde *et al.*, 2008). They argued that if unoccupied but environmentally suitable areas for the species are considered for model training, then the capacity to predict the species' potential distribution will be reduced. On the other hand, if the models are trained using a small area in which the species may have a high probability of being at equilibrium with the environment, then the models will be able to identify other potential occurrence areas when transferred. Barve *et al.* (2011) went a step further and argued that the appropriate GB for model training, validation, and comparison should comprise the set of localities that a species has 'sampled' over its history, i.e. 'the parts of the world that have been accessible to the species via dispersal over relevant periods of time' (Barve *et al.*, 2011, p. 1811). This accessible area is called 'M' in the Biotic, Abiotic and Movement (BAM) diagram terminology (*sensu* Soberón & Peterson, 2005; see also Barve *et al.*, 2011) and it is important to realize that it is specific to each species. Both Anderson & Raza (2010) and Barve *et al.* (2011) recognized that delimiting the appropriate GB is generally not feasible because the biological information required for this purpose is rarely available for most species.

Barve *et al.* (2011) discussed several methods for delimiting the proper GB for an SDM analysis. They suggested that the most workable approach could be to use biotic regions, i.e. climatic and geographic units with organisms that share broad environmental adaptations and history. Another method would be to use SDMs in a two-step procedure: using the results of a first round of modelling to help define the appropriate GB to be considered in a second round. Finally, they noted that the most interesting but also the most challenging approach would be to use information related to the dispersal capacity of the species, phylogeographic data and palaeoclimatic data. However, Barve *et al.* (2011) recognized the excessive simplicity, the circularity, and the lack of operability, respectively, of their proposals. In this study, we propose and assess a species-specific, practicable procedure to delimit the GB based on the global surface-fitting procedure known as trend surface analysis (TSA; Legendre & Legendre, 1998).

We argue that to develop purely environmental models in SDM (i.e. the so-called ecological niche models *sensu* Soberón, 2010), a GB which maximizes the likelihood that the targeted species is interacting with the environment should first be delimited. This can be done by controlling for the broad-scale geographical structure of the data that may be caused by numerous factors such as dispersal limitation, geographical characteristics or historical events, among others (McGlone, 1996; Soberón & Peterson, 2005; Svenning & Skov, 2005). As the present distributional range of a species is determined by its past distribution and population dynamics, the geographical universe delimited with the TSA may be considered a reflection of the history of the ecological interactions of the species (e.g., Real *et al*., 2003). With the TSA, we can delimit the area that has the highest probability of being accessible to a species given its present distributional pattern and at the same time avoid the inclusion of geographical regions that, due to their spatial remoteness, may be uninformative for an ecological model (Lobo *et al*., 2010). By accounting for the broad-scale spatially structured variation of species occurrence data, the GB on which SDMs should be trained is defined. Because SDMs are eco-geographical, once the broad-scale geographical structure has been accounted for, the models parameterized within the GB (delimited with the TSA) can be considered to be largely environmental, and these are the models that can be projected onto new territories to identify favourable locations for the species.

In this context, our main objective is to propose an approach to delimiting the GB by using TSA to objectively identify the area within which SDMs should be built. Based on this approach, we also assess the effects of the GB extent on the predictive performance of the SDMs, and specifically on model calibration and discrimination. To the best of our knowledge, the effects of the GB extent on model calibration have not been evaluated. However, this is not surprising because calibration, i.e. how closely the predicted probabilities match the observed proportions of occurrence (Pearce & Ferrier, 2000), is rarely assessed in SDM. We also evaluate the effects of the GB extent on the tendency of the models to predict environmental potential when they were projected onto a new scenario. To this end, we modelled the distribution of four mammal species with well known and contrasting distribution patterns in mainland Spain.

MATERIALS AND METHODS

The data

The study area was mainland Spain, an environmentally heterogeneous territory with a complex geological history (Font, 2000; Hevia, 2004). For modelling purposes we used the Universal Transverse Mercator (UTM) 10 km × 10 km squares as territorial units (n = 4684 squares in the study area). We modelled the distribution of four well-known native species (see Fig. 1): red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*), and Pyrenean chamois (*Rupicapra pyrenaica*). The red deer is a common species (n = 1530 presences), and is widely distributed throughout the study area. The roe deer is also a common species (n = 1782 presences) in the northern half of Spain. The Iberian wild goat is mainly distributed in the eastern mountain ranges (n = 621 presences). Finally, the Pyrenean chamois has a very limited distribution (n = 173 presences) restricted to the northern mountain ranges. Presence data for these species were extracted from Palomo *et al*. (2007), and information on the Iberian wild goat was updated using data from Acevedo & Cassinello (2009). The rate of false absence data can be considered negligible.

Delimiting GB

A third-degree TSA was fitted, as this is recommended for exploring processes that occur at the same or a higher spatial scale than the study area (Legendre & Legendre, 1998; p. 742). The saturated functions of TSA, i.e., without the selection of predictors, were used to obtain comparable models for the different species. For each species, seven GBs of different extent were delimited using the TSA predicted values. The first extent was delimited by the lowest TSA value assigned to a presence (GB_{LOW}); the reasoning behind this GB is that it seems logical to select a GB that includes all the presence records currently known for the species. Next, the GB was restricted by selecting as thresholds the TSA values which correspond to excluding 1% (GB_{-1}), 5% (GB_{-5}) and 10% (GB_{-10}) of the presences with the lowest TSA values. Similarly, the extent was enlarged including 1% (GB_{+1}), 5% (GB_{+5}) and 10% (GB_{+10}) of the absences that had the highest TSA values lower than the values for any presence. Finally, the total study area (mainland Spain) was also included as an additional extent (GB_{MS}). In summary, eight GBs of different extents were considered for each species and these GBs were each used to assess the effects of the GB extent on the performance of the models.

Species distribution models

Logistic regressions were performed for each species and criterion (n = 32 models) with 28 environmental predictors related to topography (2 variables), climate (22 variables), human activity (3 variables), and lithology (1 variable; see Table 1). Variables were chosen on the basis of availability and potential predictive power for wild ungulates in Spain (see Acevedo & Real, 2011). As investigating the environmental factors that modulate species distribution was not the aim of this study, we have not described the variables further; details can be found in Barbosa *et al*. (2003).

Logistic models were forward–backward stepwise fitted using a 0.05 significance threshold for the inclusion of the variables and 0.10 for their exclusion. Models were trained on each extent (eight different GBs) and projected onto mainland Spain for each species. To compare the results of each model obtained from species with different prevalences, the favourability function was applied to convert logistic probabilities (*P*) into favourability values (*F*) that are independent of sample prevalence (for further details about this function see Real *et al*., 2006).

Training and evaluation data sets

For each species and GB, a distribution model was trained using a 70% random sample of the data. The predictive performance of the models was assessed on three evaluation data sets: (1) on independent data and within the GB considered in the training process, i.e., on the remaining 30% of the data (evaluation in the training area); (2) only on the independent data that are included within GB_{-10} (evaluation in the core area – independent data); and (3) in order to avoid problems because of a small sample size in the previous evaluation data set, models were also evaluated using all the localities included in GB_{-10} for each species (evaluation in the core area – full data). Different evaluation data sets were selected to analyse the effects of GB extent in different contexts of the distribution of a species (core area in relation to the complete training area). The ‘core area – full data’ data set was constant across GBs for each species and therefore provided a way of comparing performance in a quasi-standardized manner between all models for each species.

Predictive performance

Sensitivity (*Se*, the ratio of correctly predicted presences to the total number of presences), specificity (*Sp*, the ratio of correctly predicted absences to the total number of absences), and the AUC (area under the curve of the receiver operating characteristic plot) were computed to assess the discriminatory power of the models on each evaluation data set. *Se* and *Sp* were calculated using a cut-off of $F = 0.5$ according to the favourability concept (Real *et al.*, 2006). Calibration of the *P*-values was assessed using Miller’s statistic, which is based on the hypothesis that the calibration line – perfect calibration – has an intercept of zero and a slope of one (for details see Miller *et al.*, 1991; Pearce & Ferrier, 2000). The R script provided by Wintle *et al.* (2005) was used for calculating Miller’s statistic. Finally, for each species and GB, the number of territorial units predicted as presences ($F > 0.5$) in the whole study area was calculated as a proxy of the environmental potential predicted by models.

Assessing the effect of the GB extent

The effects of the GB extent on each of the performance measures and evaluation data sets were assessed using general linear mixed models because performance measures are not independent (Zuur *et al.*, 2009). The species was included as a random factor and the GB extent – measured as number of territorial units – as a covariable. The normality of the residuals of each model was determined using the Kolmogorov–

Smirnov test (Zar, 1999). All models were assessed using the statistical package SPSS 18.

RESULTS

The results of TSA provided evidence of broad-scale spatial trends in the distribution of the four species (Fig. 2). The favourability maps obtained from the 32 models are shown in Appendix S1 in the Supporting Information; the case of the roe deer is presented as an example (Fig. 3). In general, visual assessment of the geographical patterns shows that the predictions of the models for each species are quite similar in the core area (GB_{-10}), and that the highest variability between the different models is obtained when making predictions outside the training data sets (see Appendix S1).

The results of the statistical models used to assess the effect of the GB extent on the measures of model performance are summarized in Table 2 (statistical parameters are given in Appendix S2). The residuals of each model were normally distributed ($P >$ 0.05 in all cases). In most cases, the GB extent was significantly associated with the four performance measures. There was a negative association between the discriminatory power – AUC and *Sp* – and the GB extent when models were assessed in the core areas and a positive association when the evaluation was performed on the training area. The relationship with *Se* was positive in all cases, although it was not always significant. Miller's statistic, in which high values indicate poorly calibrated models, was positively associated with the GB extent when the models were assessed on both the training and the core area data sets. Finally, there was a negative association between the GB extent and the area predicted as suitable for mainland Spain ($F_{1,27}$ = 6.023, P = 0.021; species was included as a random factor: $F_{1,27}$ = 62.022, P < 0.001).

In summary, as the GB extent increases, the discriminatory power within the considered GB improves but on closer inspection, when only the performance in the core area is assessed, the discriminatory capacity worsens due to overprediction. Calibration is always negatively affected by increasing the GB extent. Furthermore, when the models are projected beyond their training area, the smaller the GB extent, the higher the capacity to predict environmental potential becomes (Fig. 3; see also Appendix S1).

DISCUSSION

Our results demonstrate that the GB selected has visible effects on the parameters used to measure the predictive performance of SDMs, namely model discrimination (Lobo *et al*., 2008; VanDerWal *et al*., 2009; Barve *et al*., 2011), calibration, and the model's capacity to predict environmental potential (Anderson & Raza, 2010). Unfortunately, the GB has usually been defined using geopolitical criteria with no real biological justification (Meyer & Thuiller, 2006). Sufficient evidence has now been accumulated showing the effects of GB on SDMs, and steps to delimit it are beginning to be contemplated; the approach that is proposed in this study is a practical and objective way to do so.

The inclusion of absences from beyond the geographical domain of the species, i.e., increasing GB, is an easy way to obtain models with a high capacity to discriminate between instances of presence and instances of absence (Lobo *et al*., 2008; Barve *et al*., 2011). This is corroborated by the positive association between the GB extent and the discrimination measures obtained when the models were evaluated on the training area data sets. However, discriminatory power decreased as GB increased when assessed on the core area of the species distribution (i.e. GB_{-10}, the minimum extent). On the other hand, calibration improved when models were built using smaller GBs. In other words, if absences from beyond the geographical domain of the species are included, then the models will not effectively reflect the probability of presence. In summary, increasing GB produces apparently better (in terms of discrimination) but barely informative models (see also Lobo *et al*., 2010).

Our results indicate that larger favourable areas in mainland Spain were predicted when using smaller GBs (see Fig. 3 and Appendix S1). Although there is no objective way to assess the accuracy of the estimations of environmental potential, these were consistent with expert opinions for the studied species. Anderson & Raza (2010) found a similar pattern and explained that using a large GB in SDM could make the models prone to overfit the environmental conditions present in the region occupied by the species. This may happen because the algorithm recognizes spurious environmental differences between the inhabited localities and localities that could be inhabited but are not, which may be due, for example, to barriers preventing species dispersion or other historical events restricting the species current distribution. By delimiting the GB using geographical criteria, we may be excluding – or at least minimizing – the effect of historical events on model parameterization. Thus, we may be able to obtain environmental models which, when projected onto a new scenario, may help to depict the potential distribution of the species more reliably. The extrapolation of models is risky and requires caution and careful consideration (Jiménez-Valverde *et al.*, 2011b). For instance, it is necessary to highlight the areas that have environmental values that are beyond those shown in the training region, because the predictions there are more uncertain (Elith *et al.*, 2010; Jiménez-Valverde *et al.*, 2011c). It is also advisable to check for maintenance of the correlation structure among the independent variables in the new geographical area with respect to the training region (see Jiménez-Valverde *et al.*, 2011c). It is also interesting to highlight that the patterns obtained in this study using presence–absence data and logistic regression follow the same trend as those obtained by other authors using presence–background data and MAXENT (Anderson & Raza, 2010; Barve *et al.*, 2011), which suggests that they do not depend on the modelling technique.

To our knowledge, only Barve *et al.* (2011) have previously presented a framework for thinking about and estimating the GB in the context of SDM; they suggested several potential approaches to objectively delimit 'M' (see Introduction). The most promising approach would be to use information related to the dispersal capacity and history of the species, but the data required are rarely available. More feasible procedures such as the use of biogeographical regions overlook the species-specificity of the GB and may not be entirely satisfactory. In this study, we propose a simple but practical and species-specific way to delimit GB using purely geographical criteria. TSA is a simple method that accounts for broad-scale spatial structures and shows the main geographical trends in the data (see Legendre & Legendre, 1998). Thus, we argue that TSA is a useful method for use in delimiting the area in order to maximize the likelihood that the target species is currently interacting with the environment. At the same time, it minimizes the probability of including regions that are suitable for the species but that are uninformative for an ecological model due to their spatial remoteness from the current geographical range (see Lobo *et al.*, 2010). The TSA should be considered a working procedure intended to minimize the role played by the factors that operate beyond the area inhabited by the species. Strictly speaking, the spatial pattern generated with the TSA cannot be considered a geographical representation of 'M' because the concepts that underlie each approximation do not necessarily converge on the same geographical space. Most likely, the longer the species has been present in the accessible area, the lower the similarity between the spatial patterns yielded by the two approximations will be. Nevertheless, under such extreme circumstances (long time periods), estimating ‘M’ is very difficult if not impossible. Whether ‘sampled’ unoccupied localities that are far away from the present distribution range should be considered in the modelling process, or should be excluded because they are not informative about the interaction of the species with the environment (see Lobo *et al*., 2010), is debatable. We show that TSA is a practical approach that can be used as a reference for future studies aimed at developing new ideas in delimiting GB. We also anticipate that other spatial pattern analytical procedures may merit future investigation, and that the delimitation of the GB is a promising line of research and debate.

ACKNOWLEDGEMENTS

This study was presented at the International Biogeography Society Crete meeting. We are grateful to A. Townsend Peterson and one anonymous referee for their useful comments and suggestions on previous versions of the manuscript. Lucía Maltez kindly reviewed the English. P.A. and A.J.-V. were supported by the Juan de la Cierva research program awarded by the Ministerio de Ciencia e Innovación–Fondo Social Europeo, and partly by the project CGL2009-11316/BOS—Fondos FEDER. P.A. is in Portugal thanks to a José Castillejo fellowship (2010–2011) granted by the Ministerio de Ciencia e Innovación.

REFERENCES

Acevedo, P. & Cassinello, J. (2009) Biology, ecology and status of Iberian ibex *Capra pyrenaica*: a critical review and research prospectus. *Mammal Review*, **39**,17–32.

Acevedo, P. & Real, R. (2011) Biogeographical differences between the two Capra *pyrenaica* subspecies, *C. p. victoriae* and *C. p. hispanica*, inhabiting the Iberian Peninsula: implications for conservation. *Ecological Modelling*, **222**, 814–823.

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.

Barbosa, A.M., Real, R., Olivero, J. & Vargas, J.M. (2003) Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biological Conservation*, **114**, 377–387.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J. & Villalobos, F. (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, **222**, 1810–1819.

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Font, I. (1983) *Atlas climático de España*. Instituto Nacional de Meteorología, Madrid.

Font, I. (2000) *Climatología de España y Portugal*. Ediciones Universidad de Salamanca, Salamanca.

Hevia, I.M. (2004) *Geología de España. Una historia de seiscientos millones de años*. Rueda S. L., Madrid.

IGME (1979) *Mapa hidrogeológico Nacional. Explicación de los mapas de lluvia útil, de reconocimiento hidrogeológico y de síntesis de los sistemas acuíferos*, 2nd ed. Instituto Geológico y Minero de España (IGME), Madrid.

IGN (1999) *Mapa de carreteras. Península Ibérica, Baleares y Canarias*. Instituto Geográfico Nacional (IGN)/Ministerio de Fomento, Madrid.

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

Jiménez-Valverde, A., Lira-Noriega, A., Barve, N., Maher, S.P., Nakazawa, Y., Papes, M., Soberón, J., Sukumaran, J. & Peterson, A.T. (2011a) Dominant climate influences on North American birds. *Global Ecology and Biogeography*, **20**, 114–118.

Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J., Aragón, P. & Lobo, J. M. (2011b) Use of niche models in invasive species risk assessments. *Biological Invasions*, **13**, 2785–2797.

Jiménez-Valverde, A., Decae, A.E. & Arnedo, M.A. (2011c) Environmental suitability of new reported localities of the funnelweb spider *Macrothele calpeiana*: an assessment using potential distribution modelling with presence-only techniques. *Journal of Biogeography*, **38**, 1213–1223.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd English edn. Elsevier Science, Amsterdam.

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

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

McGlone, M. (1996) When history matters: scale, time, climate and tree diversity. *Global Ecology and Biogeography*, **5**, 309–314.

Meyer, C.B. & Thuiller, W. (2006) Accuracy of resource selection functions across spatial scales. *Diversity and Distributions*, **12**, 288–297.

Miller, M.E., Hui, S.L. & Tierney, W. (1991) Validation techniques for logistic regression models. *Statistics in Medicine*, **10**, 1213–1226.

Montero de Burgos, J.L. & González-Rebollar, J.L. (1974) *Diagramas bioclimáticos*. ICONA, Madrid.

Palomo, L.J., Gisbert, J. & Blanco, J.C. (2007) *Atlas y Libro Rojo de los mamíferos terrestres de España*. Ministerio de Medio Ambiente, Madrid.

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

Real, R., Barbosa, A.M., Porras, D., Kin, M.S., Márquez, A.L., Guerrero, J.C., Palomo, L.J., Justo, E.R. & Vargas, J.M. (2003) Relative importance of environment, human activity and spatial situation in determining the distribution of terrestrial mammal diversity in Argentina. *Journal of Biogeography*, **30**, 939–947.

Real, R., Barbosa, A.M. & Vargas, J.M. (2006) Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, **13**, 237–245.

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.

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

Soberón, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distribution areas. *Biodiversity Informatics*, **2**, 1–10.

Svenning, J.C. & Skov, F. (2005) The relative roles of environment and history as controls of tree species composition and richness in Europe. *Journal of Biogeography*, **32**, 1019–1033.

VanDerWal, J., Shoo, L.P., Graham, C. & Williams, S.E. (2009) Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecological Modelling*, **220**, 589–594.

Wintle, B.A., Elith, J. & Potts, J.M. (2005) Fauna habitat modelling and mapping in an urbanising environment: a case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, **30**, 719–738.

Zar, J.H. (1999) *Biostatistical analysis*, 5th edn. Pearson, New York.

Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A. & Smith, G.M. (2009) *Mixed effects models and extensions in ecology with R*. Springer, New York.

SUPPORTING INFORMATION

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

Appendix S1 Favourability maps for each species and geographical background (GB) extent.

Appendix S2 Statistical parameters of the mixed models used to assess the effect of the GB extent on the calibration and discrimination parameters.

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

Pelayo Acevedo is a researcher at the University of Málaga. His interests include the study of factors affecting the distribution and abundance of pathogens and their hosts and vectors, through fragmented habitats.

Author contributions: P.A. and R.R. conceived the ideas; P.A. and A.J.-V. analysed the data; P.A., A.J.-V, J.M.L. and R.R participated in the discussion of the results and wrote the manuscript.

Editor: Miles Silman

Table 1 Variables used to model the distribution of wild ungulates.

Factor	**Variable description**
Orography	Mean elevation (m) *
	Mean slope (degrees; calculated from mean altitude)
Climatology	Mean annual precipitation (mm) [*P*] [†]
	Maximum precipitation in 24 h (mm) [*MP24*] [†]
	Relative maximum precipitation (= *MP24/P*)
	Mean annual number of days with precipitation ≥ 0.1 mm [†]
	Mean annual number of hail days [†]
	Mean annual number of foggy days [†]
	Inter-annual pluviometric irregularity [‡]
	Mean annual potential evapotranspiration (mm) [*PET*] [†]
	Mean annual actual evapotranspiration (mm) (= min [*P*,*PET*])
	Mean relative air humidity in July at 07:00 h (%) [*HJL*] [†]
	Mean relative air humidity in January at 07:00 h (%) [*HJN*] [†]
	Annual humidity range (%) (=*HJL*–*HJN*)
	Mean annual solar radiation (kW h m^{-2} day^{-1}) [†]
	Mean temperature in July (ºC) [*TJL*] [†]
	Mean temperature in January (ºC) [*TJN*] [†]
	Annual temperature range (ºC) (=*TJL*–*TJN*)
	Mean annual temperature (ºC) [†]
	Mean annual number of frost days (minimum temperature ≤ 0 ºC) [†]
	Continentality index [¶]
	Humidity index [¶]
	Mean annual insolation (hours $year^{-1}$) [†]
	Mean annual runoff (mm) [‡]
Human activity	Distance to the nearest town with more than 100,000 inhabitants (km) [§]
	Distance to the nearest town with more than 500,000 inhabitants (km) [§]
	Distance to the nearest highway (km) [§]
Lithology	Soil permeability [**]
Third-degree polynomial of the trend surface analysis	Mean latitude (ºN) [*LA*] [§]
	Mean longitude (ºE) [*LO*] [§]
	$LALO = LA \times LO$
	$LOLA^2 = LO \times LA^2$
	$LO^2LA = LO^2 \times LA$
	$LA^2 = LA \times LA$
	$LO^2 = LO \times LO$

	$LA^3 = LA^2 \times LA$
	$LO^3 = LO^2 \times LO$

Sources: * http://www.etsimo.uniovi.es/~feli/data/datos.html. [†] Font (1983). [‡] Montero de Burgos & González-Rebollar (1974). [¶] Font (2000). [§] IGN (1999); data on the number of inhabitants of urban centres taken from the Instituto Nacional de Estadística (http://www.ine.es). ** IGME (1979).

Table 2 Summary of the results of the general linear mixed models used to assess the effect of the extent of the geographical background (GB) on the performance (calibration and discrimination) of species distribution models for four species in mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*). Species was included as a random factor. The predictive performance of the models was evaluated on different data sets (see text for details). Statistical parameters are reported in the Supporting Information (Appendix S2).

Evaluation data set	**Parameter (dependent variable)**	**GB extent (covariable)**
Training area	Miller's statistic	(+)**
	Sensitivity	(+) ns
	Specificity	(+)**
	AUC	(+)*
Core area – independent data	Miller's statistic	(+)*
	Sensitivity	(+) ns
	Specificity	(–)**
	AUC	(–) #
Core area – full data	Miller's statistic	(+)**
	Sensitivity	(+)**
	Specificity	(–)**
	AUC	(–)*

ns, non-significant; #, $P < 0.08$; *, $P < 0.05$; **, $P < 0.01$.

FIGURE LEGENDS

Figure 1 Current distribution of the focus species in mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*). Presence data were referred to UTM 10 × 10 km grid cells. These were taken from Palomo *et al*. (2007) and Acevedo & Cassinello (2009).

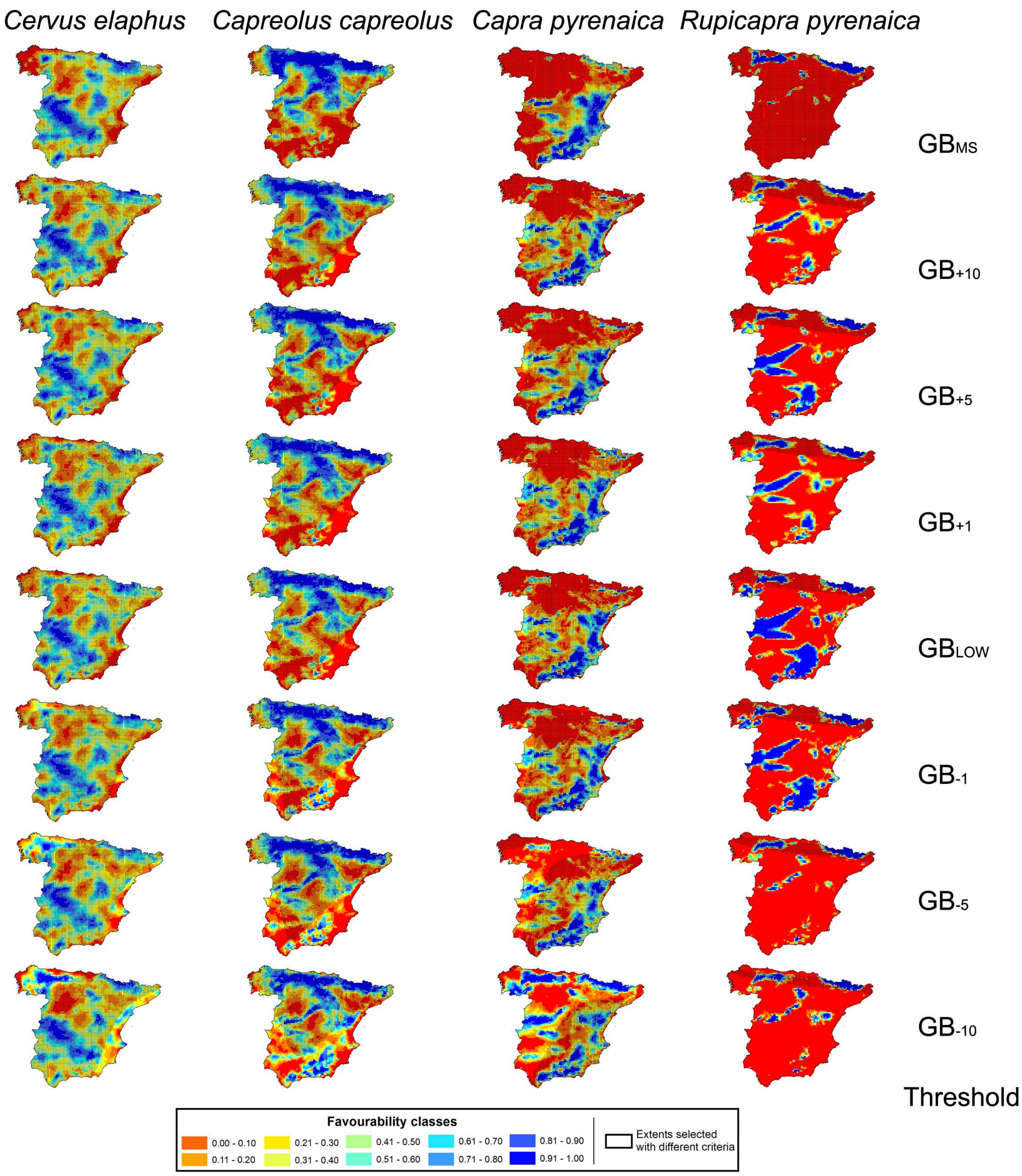
Figure 2 Results of the trend surface analysis (TSA) using a third-degree polynomial of the spatial coordinates applied to the occurrence localities of the four species in mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*).

Figure 3 Species distribution models (favourability values) obtained using different criteria to delimit the geographical background (GB) extent using roe deer (*Capreolus capreolus*) as an example (see also Appendix S1). 'GB_{MS}' indicates the model that included the complete study area (mainland Spain) as a training data set. 'GB_{LOW}' indicates the model in which the training area was delimited by the lowest trend surface analysis (TSA) value assigned to a presence (see text for details). 'GB_{+10}' indicated the model that included 10% of the absences that, having TSA values lower than any presence, had the highest TSA values. Finally, 'GB_{-10}' is similar to 'GB_{LOW}' but excludes 10% of the presences. The dashed line marks the area delimited with 'GB_{-10}'.

SUPPORTING INFORMATION -- *Journal of Biogeography*

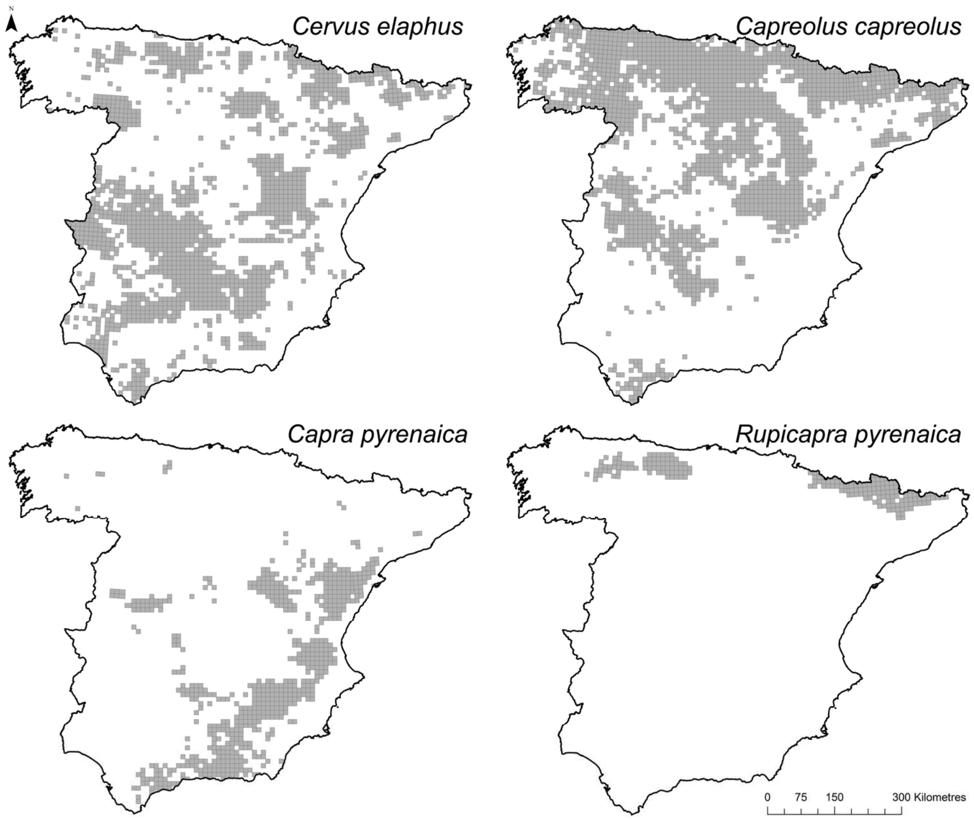
Delimiting the geographical background in species distribution modelling, by P. Acevedo, A. Jiménez-Valverde, J.M. Lobo and R. Real

Appendix S1 Species distribution models (favourability values) for each of the focus species: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*). For each species, seven geographical backgrounds (GBs) of different extent were delimited. The first extent (GB_{LOW}) was delimited by the lowest trend surface analysis (TSA) value assigned to a presence. Next, the extent was constrained by selecting as thresholds the TSA values which correspond to excluding 1% (GB_{-1}), 5% (GB_{-5}) and 10% (GB_{--10}) of the presences with the lowest TSA values. In a similar way, the extent was enlarged including 1% (GB_{+1}), 5% (GB_{+5}) and 10% (GB_{+10}) of the absences that had the highest TSA values lower than that of any presence. Finally, we also included the complete study area (mainland Spain) as an additional GB (GB_{MS}).

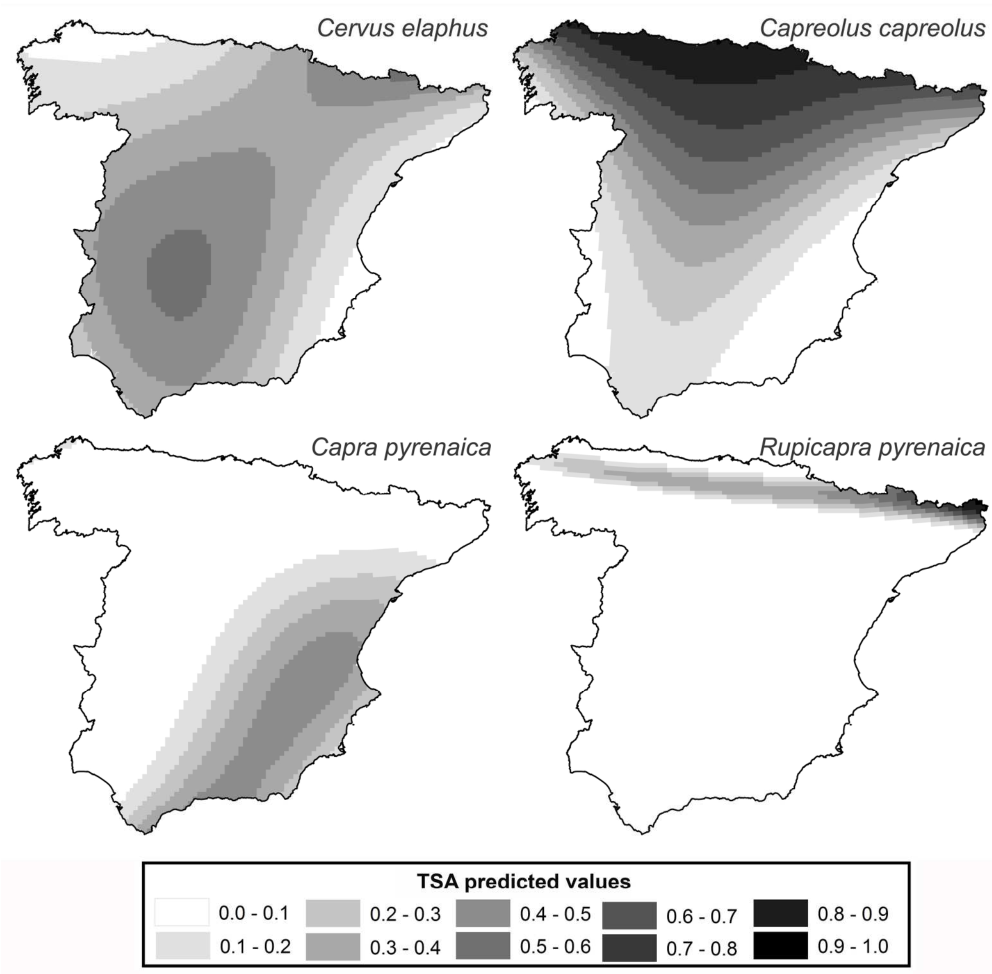


Appendix S2 Statistical parameters of the mixed models carried out to assess the effect of the extent of the geographical background (GB) on the performance (calibration and discrimination) of the species distribution models for four species in mainland Spain: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), Iberian wild goat (*Capra pyrenaica*) and Pyrenean chamois (*Rupicapra pyrenaica*). Species was included as a random factor. The predictive performance of the models was evaluated on different data sets (see text for details).

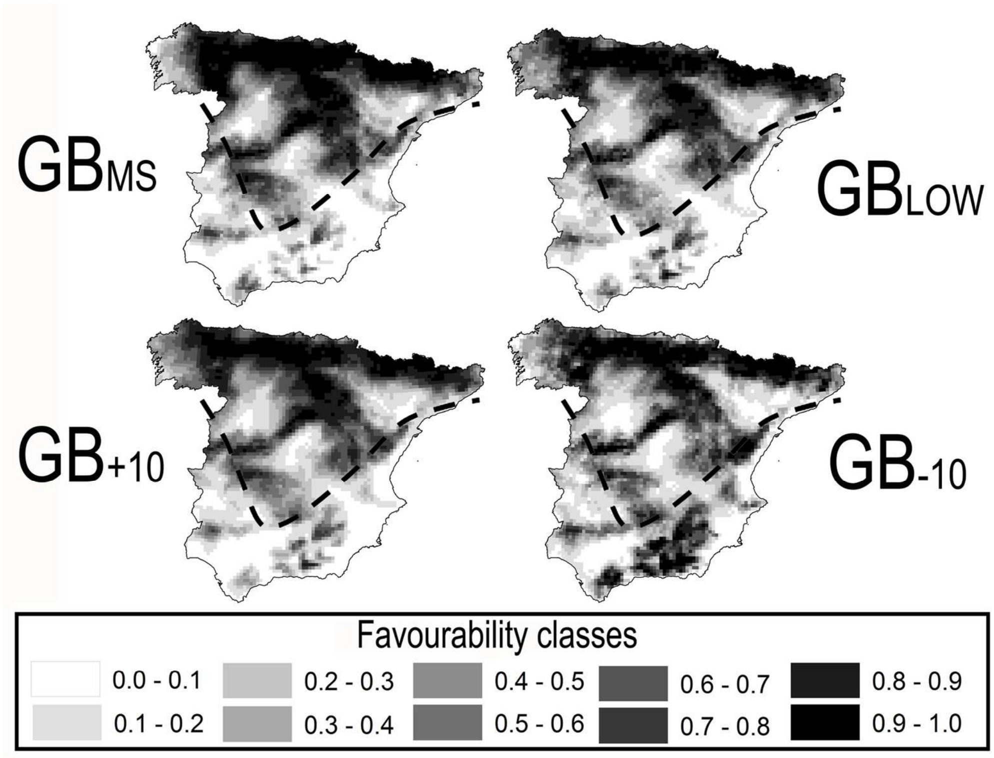
Data set	**Parameter (dependent variable)**	**Intercept**	**Species (random factor)**	**GB extent (covariable)**
Training area	Miller's statistic	$F_{1,15.16}$ 4.42, P 0.043	$F_{3,27} = 8.22$, $P < 0.001$	(+) $F_{1,27} = 15.11$, $P = 0.001$
	Sensitivity	$F_{1,5.19} = 342.70$, $P < 0.001$	$F_{3,27} = 38.77$, $P < 0.001$	(+) $F_{1,27} = 1.52$, $P = 0.230$
	Specificity	$F_{1,3.56} = 392.16$, $P < 0.001$	$F_{3,27} = 139.25$, $P < 0.001$	(+) $F_{1,27} = 11.81$, $P = 0.002$
	AUC	$F_{1,3.66} = 483.38$, $P < 0.001$	$F_{3,27} = 119.07$, $P < 0.001$	(+) $F_{1,27} = 7.12$, $P = 0.013$
Core area – independent data	Miller's statistic	$F_{1,20.72} = 9.24$, $P = 0.089$	$F_{3,27} = 5.38$, $P = 0.005$	(+) $F_{1,27} = 2.58$, $P = 0.012$
	Sensitivity	$F_{1,6.82} = 222.17$, $P < 0.001$	$F_{3,27} = 23.52$, $P < 0.001$	(+) $F_{1,27} = 1.63$, $P = 0.213$
	Specificity	$F_{1,14.04} = 536.90$, $P < 0.001$	$F_{3,27} = 9.04$, $P < 0.001$	(–) $F_{1,27} = 9.02$, $P = 0.006$
	AUC	$F_{1,3.91} = 706.30$, $P < 0.001$	$F_{3,27} = 87.20$, $P < 0.001$	(–) $F_{1,27} = 2.35$, $P = 0.077$
Core area – full data	Miller's statistic	$F_{1,7.90} = 3.77$, $P = 0.092$	$F_{3,27} = 18.89$, $P < 0.001$	(+) $F_{1,27} = 26.92$, $P < 0.001$
	Sensitivity	$F_{1,3.87} = 270.75$, $P < 0.001$	$F_{3,27} = 90.77$, $P < 0.001$	(+) $F_{1,27} = 14.40$, $P = 0.001$
	Specificity	$F_{1,17.51} = 1491.59$, $P < 0.001$	$F_{3,27} = 6.83$, $P = 0.001$	(–) $F_{1,27} = 45.75$, $P < 0.001$
	AUC	$F_{1,3.19} = 898.98$, $P < 0.001$	$F_{3,27} = 403.79$, $P < 0.001$	(–) $F_{1,27} = 5.56$, $P = 0.026$



Current distribution of the focus species in mainland Spain: red deer (Cervus elaphus), roe deer (Capreolus capreolus), Iberian wild goat (Capra pyrenaica) and Pyrenean chamois (Rupicapra pyrenaica). Presence data were referred to UTM 10 × 10 km grid cells. These were taken from Palomo et al. (2007) and Acevedo & Cassinello (2009).
165x136mm (300 x 300 DPI)



Results of the trend surface analysis (TSA) using a third-degree polynomial of the spatial coordinates applied to the occurrence localities of the four species in mainland Spain: red deer (Cervus elaphus), roe deer (Capreolus capreolus), Iberian wild goat (Capra pyrenaica) and Pyrenean chamois (Rupicapra pyrenaica).
196x192mm (300 x 300 DPI)



Species distribution models (favourability values) obtained using different criteria to delimit the geographical background (GB) extent using roe deer (Capreolus capreolus) as an example (see also Appendix S1). 'GBMS' indicates the model that included the complete study area (mainland Spain) as a training data set. 'GBLOW' indicates the model in which the training area was delimited by the lowest trend surface analysis (TSA) value assigned to a presence (see text for details). 'GB+10' indicated the model that included 10% of the absences that, having TSA values lower than any presence, had the highest TSA values. Finally, 'GB−10' is similar to 'GBLOW' but excludes 10% of the presences. The dashed line marks the area delimited with ‘GB−10’.
152x116mm (300 x 300 DPI)