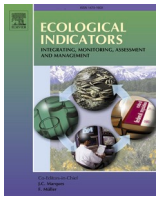
Ecological Indicators 121 (2021) 107147

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



High uncertainty in the effects of data characteristics on the performance of species distribution models

Geiziane Tessarolo [a,b,*], Jorge M. Lobo [c], Thiago Fernando Rangel [a], Joaquín Hortal [a,c]

[a] Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, Brazil
[b] Programa de Pós-graduação em Recursos Naturais do Cerrado, Universidade Estadual de Goiás, Anápolis, Brazil
[c] Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales (CSIC), C/José Gutiérrez Abascal 2, 28006 Madrid, Spain

ARTICLE INFO

Keywords:
Ecological traits
Marginality
ROA
Scarabaeoidea dung beetles
Species distribution modelling
Uncertainty

ABSTRACT

Species distribution models (SDM) are widely used as indicators of different aspects of geographical ranges for many purposes, from conservation to biogeographical and evolutionary analyses. However, these techniques are susceptible to various sources of uncertainty. Data coverage, species' ecology, and the characteristics of their geographic distributions can affect SDM results, often generating critical errors in predicted distribution maps. We assess the influence of data quality, the characteristics of species distributions, and ecological traits on SDM performance. We predict the distributions of dung beetle species in Madrid region (central Spain) using six SDM techniques and validate them on an independent dataset. We relate variations in model performance with environmental completeness, data characteristics, and species traits through a partial least squares analysis. In this analysis, body size, nesting behaviour, marginality, rarity, data prevalence, Relative Occurrence Area (ROA), range size, niche breadth, and completeness are used as predictors of six assessment metrics (sensitivity, specificity, kappa, TSS, CCR, and AUC). Marginality and data prevalence were the variables that most influenced SDM performance, followed by range size, ROA, and niche breadth: species presenting higher marginality and data prevalence, and smaller ROA and niche breadth were associated with better models. Nesting behaviour, rarity, niche completeness, and body size had minor importance for SDM performance. Our results highlight the importance of taking species' and data characteristics into account when modelling and comparing large groups of species using SDM. This implies that estimates of species richness and composition based on stacked SDMs can show high levels of error if they are constructed for groups of species with diverse ecological traits and types of geographic distributions. We suggest that the species holding characteristics that lead to poor SDM performance should not be included when constructing composite biodiversity variables. Further effort is needed to develop SDM methodologies and protocols that account for such source of uncertainty.

1. Introduction

Species Distribution Models (herein SDMs, also known as Environmental Niche Models) include a set of algorithms and general data management procedures that seek to model the potential or realized geographic distributions of species, typically based on their known occurrences and several environmental predictors (Guisan et al., 2017). SDMs are widely used as indicators of different aspects of species' geographical ranges, from current distributions to areas that could be potentially occupied by invader species or host relocated or dispersing populations under climate change scenarios. Despite the wide popularity of SDMs, the reliability of their results depends on the characteristics and quality of the occurrence data used to train them. It follows that the limitations in the primary biodiversity data used for modelling can impose severe restrictions on their utility (Araújo et al., 2019; Duputié et al., 2014; Guillera-Arroita et al., 2015; Rocchini et al., 2011).

Knowledge of species distributions is limited (i.e. the so-called Wallacean shortfall; Lomolino 2004) and, in general, geographical and environmentally biased (Beck et al., 2014; Hortal et al., 2015, 2008; Oliveira et al., 2016). In addition, the particular characteristics of each species, like those related to their ecological role and geographical distribution, may also seriously influence the uncertainty of SDM predictions (Chefaoui et al., 2011; Gábor et al., 2019). Species often differ in

* Corresponding author at: Programa de Pós-graduação em Recursos Naturais do Cerrado, Universidade Estadual de Goiás, Anápolis, Brazil.
E-mail address: geites@gmail.com (G. Tessarolo).

https://doi.org/10.1016/j.ecolind.2020.107147
Received 8 April 2020; Received in revised form 16 October 2020; Accepted 31 October 2020
Available online 20 November 2020

the biological traits determining their distributions, hindering the automatic detection of the "true" relationship between species occurrences (and, eventually, absences) and environmental predictors (Brotons et al., 2004; Thuiller et al., 2010). Further, many species traits –such as conspicuousness or habitat preferences–determine how often their occurrence is effectively recorded during surveys (Boone and Krohn, 1999), diminish the quality of the data used for training SDMs, and with it compromising their predictions (Ladle and Hortal, 2013; Sheth et al., 2008).

Identifying which species' characteristics or traits affect the predictive capacity and performance of SDMs, and determining the intensity of their eventual influences, could allow selecting the species that are more suitable for SDM applications under specific situations (Guisan et al., 2007), allowing also to discard the use of these techniques (to model the distributions of species with characteristics that may compromise such performance. By species characteristics we refer to either attributes of the individuals related to their ecological role (i.e. functional traits), or general features of the species such as its geographical distribution. Here we borrow the terminology used in functional ecology, considering as *ecological traits* (or simply *traits* for short) the characteristics of the individuals that have ecological significance (Díaz et al., 2013; Moretti et al., 2016; Violle et al., 2007). We distinguish traits from both *eco-geographic characteristics* –which refer to the attributes of the distribution of the species, and *data characteristics* –the general attributes of the data itself.

Several ecological traits are known to affect SDM results, including body size (França and Cabral, 2016; Zamorano et al., 2019), life span (Hanspach et al., 2010; McCune et al., 2020), growth rate (Guisan et al., 2007), habitat specialization (Marshall et al., 2015; Regos et al., 2019) or dispersal ability (McCune et al., 2020). But the accuracy of SDM predictions is also determined by species' eco-geographic characteristics such as geographic range size (Chefaoui et al., 2011; Guo et al., 2015; Wogan, 2016), rarity (Franklin et al., 2009), the marginality of their occurrences with respect to the environmental conditions of the study region (Gábor et al., 2020; Jiménez-Valverde et al., 2008; Lobo, 2008), or the proportion of their potential distributions that is effectively occupied by species' populations (Grenié et al., 2020). In general, SDM performance seems to be poorer in species with higher mobility, broader niche breadth and wide distribution ranges (Guo et al., 2015; Hortal et al., 2008; Newbold et al., 2009; Tessarolo et al., 2014). On the contrary, the distributions of rare species, and those with larger lifespan and less habitat specialization are often predicted better (Chefaoui et al., 2011; Henckel et al., 2020; Marshall et al., 2015; Mateo et al., 2010; McCune et al., 2020). Finally, data characteristics are also a decisive factor influencing SDM performance. For example, the size of the area occupied by the species relative to the total area of the studied region (Relative Occurrence Area, ROA) determines predictive success, as the rate of well-predicted absences increases when the extent of analyses is much larger than the species distribution, hence inflating performance metrics (Lobo, 2008; Lobo et al., 2008; Moudrý and Šímová, 2013).

Our ignorance about the environmental response of most species (i.e. the Hutchinsonian shortfall *sensu* Hortal et al., 2015) also hinders assessing the effects of species traits on SDM performance. The coverage of such environmental responses by the occurrence data (i.e., environmental coverage) may vary among the species present in the same region (see Hortal et al., 2008), thus affecting SDM results. 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 (Guisan et al., 2006; Pearce et al., 2001; Reese et al., 2005). Although many studies have stressed the importance of environmental coverage to obtain accurate models of species distributions (Hortal et al., 2008; Lobo et al., 2010; Thuiller et al., 2004), its effects on SDM performance have been seldom evaluated (see, e.g. Grenié et al., 2020).

Studies on the influence of species' characteristics on SDM performance have often provided conflicting results, leading to different conclusions for the same species traits and characteristics. For example, while Stockwell and Peterson (2002) reported a negative influence of range size on model performance, Garrison and Lupo (2002) found opposite results. Also, species with lower prevalence were better modeled in Syfert et al. (2013) and van Proosdij et al. (2016), while better models were achieved for species with higher prevalence in Meynard and Quinn (2007) and Sor et al. (2017). Similarly, some studies found positive impacts on model performance of larger niche breadths (Seoane et al., 2005), others that more specialist species render better models (Connor et al., 2018; Hernandez et al., 2006), while others did not find any relationship between niche breadth and model performance (Newbold et al., 2009; Pöyry et al., 2008). These disagreements may come from the different attributes evaluated in each study, but also from the covariation between functional traits, eco-geographic, and data characteristics, which often vary together giving rise to syndromes (i.e., values for certain traits and characteristics that appear repeatedly). Nonetheless, how SDM performance is calculated is crucial to accurately compare the different results of comparative studies. 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 (Hijmans, 2012; Marmion et al., 2009; Santika, 2011; Wogan, 2016). Many of the studies assessing the influence of species traits in SDM use a small number of modelling techniques (Hanspach et al., 2010; Marmion et al., 2009; McCune et al., 2020; Santika, 2011). Furthermore, previous studies often based their comparisons on AUC, which despite being the discrimination metric most commonly used to evaluate SDM performance, it is also well-known to be unable to provide fair comparisons of model results across species with different data and eco-geographic characteristics (Hijmans, 2012; Jiménez-Valverde, 2012; Lobo et al., 2008; Phillips, 2008).

In this study we evaluate the influence of several species traits and eco-geographic and data characteristics on SDM performance, as measured by several discrimination metrics. More specifically, our aims are to identify which traits impact SDM results the most and evaluate the influence of environmental coverage on the precision of these results. To do this, we use historical and standardized data as calibration and validation datasets, respectively, to assess the performance of the most commonly used SDM techniques to predict the distribution of 93 dung beetle species in Madrid (central Spain), and measure the success of these predictions using a set of standard performance metrics. We compare these performance metrics with the functional traits, and the eco-geographic and data characteristics of each species altogether using Partial Least Squares analyses to both assess their relevance for SDM results and identify their eventual covariation.

2. Methods

2.1. Species data

We used data on the distribution of dung beetles (Coleoptera, Scarabaeoidea) in central Spain from SCAMAD (Hortal et al., 2020), a biological database that compiles all distributional information for this group in Madrid province (8,022 km^2) and its surroundings (Hortal and Lobo, 2005). Records in this database come from natural history collections from museums, universities, and private owners, as well as from specialized literature (Hortal et al., 2008, 2006; Hortal and Lobo, 2005; Lobo and Hortal, 2006). We used two different versions of the database to consider two different levels of knowledge: (i) historical surveys from multiple naturalists gathered throughout the years with no planned sampling design, and therefore subject to the biases typically shown by the historical data present in biodiversity databases and (ii) standardized planned surveys designed to provide good geographical coverage of the spatial and environmentally-driven variations of biodiversity within the studied territory. Historical data comes from SCAMAD 1.0, a database containing data on 92,368 specimens grouped in 5,296 records

belonging to 133 species, recorded from 1808 to 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 and Lobo, 2005). This updated version of the database, SCAMAD 2.1, was used to account for standardized surveys. It contains data on 145,839 specimens, 7,117 records of both presence and absence gathered between 1808 and 2003, and 135 species. After removing records with missing spatial reference and those from empty traps, we retained 5,184 and 6,969 records from SCAMAD 1.0 and SCAMAD 2.1, respectively. All these data are divided in 108 UTM cells of 10 × 10 km that conform the grid squares that had more than 5% of their territory in the Madrid administrative region (bounding box defined by the UTM 30 N coordinates 361069.1, 4410324.7 and 501174.7, 4560437.8; see Hortal et al., 2008 for further details). Species occurrences in these cells for SCAMAD 1.0 and 2.1 are shown in Tables S1 and S2, respectively.

2.2. Environmental data

We used two kinds of variables that have been previously related to geographic variations in the distribution and diversity of Iberian dung beetles (Chefaoui et al., 2005; Hortal et al., 2001; Lobo and Martín-Piera, 2002): climate (mean, maximum and minimum temperature, total annual precipitation, and total summer precipitation) and substrate (edaphic and geologic variables; 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). Climatic data were specifically interpolated for this region from monthly temperature and precipitation scores for 41 stations in central Iberia (means from 30-year data), more details in Hortal and Lobo (2005). Substrate variables represent the proportion of area of each grid cell covered by each category, based on data on soil structure and composition from Food and Agriculture Organization (FAO - UNESCO, 1988) and bedrock data from Instituto Tecnológico y Geominero de España (ITGE, 1988). See Chefaoui et al. (2005) and Hortal and Lobo (2005) for additional details on the origin of these variables. All variables were resampled to a resolution of 10 × 10 km UTM grid cells.

To reduce the collinearity among predictors a principal component analysis (PCA) was calculated for each kind of variable, selecting PCA factors for further analyses according to a broken-stick criterion. In total, four environmental factors were extracted, two related to climatic variables (hereafter CF1 and CF2) and two to substrate variables (hereafter SF1 and SF2). These factors explain 94.4% of total variability in climatic variables (CF1 = 84.0%, CF2 = 10.4%), and 82.5% of total variability in substrate variables (SF1 = 52.7%, SF2 = 29.8%). See Hortal et al. (2008) for more details about the PCA results. The scores of these four PCA factors were retained to be subsequently used as predictors in SDMs.

2.3. Species distribution models

All species that lacked information about the characteristics evaluated here (see below) were removed from the analyses. A total of 93 species were finally modelled (see Supplementary materials Table S3 for a complete list). Species distribution models were generated using data from SCAMAD 1.0 and validated with the data from SCAMAD 2.1. To reach as much independence between the data used for calibration and evaluation as possible, we divided the 108 cells into 5 groups of 21 or 22 UTM cells selected randomly, ensuring that each group contains 20% of the presences for the species. Each one of these groups of cells (containing presences and pseudo-absences) was used to validate the results of SDMs calibrated with the remaining cells (~80% of the territory). Besides not using the same data for calibration and validation, the predictive capacity of SDMs was tested as consequence of the new distributional information and additional surveys included in SCAMAD 2.1. This validation procedure was repeated 10 times for each species.

The predictions about the distributions of species were generated using five modelling techniques that were summarized in a combined consensus prediction through ensemble forecasting (Araújo and New, 2007). SDM techniques and consensus were implemented using the SDM package (Naimi and Araújo, 2016), and include: Bioclim (Busby, 1986); Generalized Linear Models (GLM; McCullagh and Nelder, 1991); Maximum Entropy Modelling (MaxEnt; Phillips et al., 2006); Random Forest (Breiman, 2001); and support vector machines (SVM; Cortes and Vapnik, 1995). The consensus ensembles were generated by a weighted mean using TSS metrics of all single-models.

2.4. Response performance variables

The predictive performance of models was assessed with six different metrics typically used to measure the discrimination capacity of SDMs: sensitivity (the proportion of presences correctly predicted); specificity (the proportion of absences correctly identified); kappa (Cohen, 1960); true skill statistic (TSS; Allouche et al., 2006); percentage of correct classification rate (CCR; Fielding and Bell, 1997); and Area Under the ROC Curve (AUC; Fielding and Bell, 1997). These metrics were selected to provide comparability with most, if not all, SDM applications. Here we must point out that, although we acknowledge that model training and evaluation would benefit from selecting reliable absences coming from well-sampled cells (Lobo et al., 2010; Warren et al., 2020), we deliberately use similar modelling procedures as those generally found in the literature, to provide a fair comparison with the common SDM use standards. Thus, we generated random pseudoabsences by considering all cells without information of presence for a species as pseudo-absences. Pseudo-absences were split into 5 groups jointly with the presences (see above) and all performance metrics requiring absence information were calculated using these pseudoabsences.

2.5. Species characteristics used as predictors of SDM performance

We examined the association with SDM performance metrics of nine characteristics: three functionally-relevant species traits – body size (measured as body length), nesting behaviour, and demographic rarity; three eco-geographic characteristics – range size, environmental marginality, and niche breadth; and three data characteristics – relative occurrence area (ROA), data prevalence, and environmental niche completeness. Here note that many of these nine attributes are related among them, their values are often correlated (see Table S4), and some of them could be assigned to several categories. For example, depending on the reliability of the inventories, ROA can be classified as either a geographical attribute depicting how restricted is a species distribution, or a property of the data that relates the observed occurrences with the extent of the analysis. Therefore, their assignment to ecological, eco-geographic, or data categories should be taken as a general orientation rather than as a hard classification.

Information on body size and nesting behaviour were collected from the dung beetle 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 and Edmonds, 1982), and was classified in three categories: Paracoprid, Telecoprid, and Endocoprid, which correspond respectively to nests of dung buried in the ground beneath the excrement, buried after transportation (i.e. rolling behaviour) and placed within the dung pat. These three categories were included as dummy variables (1/0). Demographic rarity was measured as the inverse of the number of collected individuals of each species (in Log + 1).

Range size is the number of 10 × 10 km grid cells occupied by the species in the study area. Environmental marginality measures the difference between the mean of the environmental conditions in the study area and the optimum environmental conditions of each species and was calculated using Ecological Niche Factor Analysis (ENFA; Hirzel et al., 2002) based on the four formerly mentioned PCA factors extracted from the considered environmental variables (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 the set of input points in N-dimensional space) generated from all the occurrences of the species in the environmental space defined by the four PCA factors, calculated using *geometry* R package (Habel et al., 2015). All these eco-geographic characteristics were calculated using the data from SCAMAD 2.1. database.

ROA was calculated as the ratio between the area occupied by the species and the total area of the study region (Lobo, 2008). As in the case of the environmental space, the area occupied for each species was estimated as the smallest convex hull polygon delimited by the available occurrence points. Data prevalence is the ratio between the numbers of presences and the total number of cells used for model training. Finally, environmental niche completeness is a measure of the proportion of the whole environmental niche of the species that is covered by their data (Hortal et al., 2008; Kadmon et al., 2004). 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 and 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 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).

2.6. Statistical analyses

For each species we calculated the mean of the ten values obtained for each performance metric (one for each ensemble prediction), relating them with the nine species traits and characteristics through partial least-squares regression analysis (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 that are often related among them (see Carrascal et al., 2009). Thus, PLS allows detecting orthogonal combinations of many collinear predictors able to account for the explained variability of more than one response variable that acts as synthetic response variables. Therefore, a PLS analysis was conducted using body size, nesting behaviour, demographic rarity, environmental marginality, range size, niche breadth, data prevalence, ROA, and niche completeness as predictor variables, and six performance metrics (sensitivity, specificity, kappa, TSS, CCR, and AUC) as response variables. To construct the PLS we used a leave-one-out cross-validation. We only retained those PLS components explaining more than 5% of the original variance in the response variables and presenting goodness of prediction on test data (Q^2) > 0, *i.e.*, only components that when added do not decrease the model prediction ability. Additionally, we calculated the variable importance (VIP) of each predictor variable to the model. The VIP values inform about the overall contribution of each predictor variable to the whole PLS model, as they can be ranked by their importance. VIP was calculated as:

$$VIP_{j=}\sqrt{N\left(\frac{\sum_{j=1}^{k} w_{cj}^2 SS_c}{SS_{total}}\right)}$$

where:

N = number of predictors

k = number of PLS dimensions extracted

w_{cj} = weight of variable j for PLS dimension c

SS_c = explained sum of squares of the PLS dimension c

$SS_{Total} = \sum_{c=1}^{k} SS_c$

3. Results

The two first significant components extracted from the PLS analysis accounted for 69.48% of the variability in the response variables. All performance metrics have positive weights in the two PLS components (Table 1), indicating a positive covariation among them. Sensitivity and TSS had the greatest contribution to the first component (29.4% and 27.9%, respectively), while the contribution of all the other performance metrics varied from 5.4% to 13.9% as reflected by their squared weights. The amount of variability of this PLS component that can be explained by the linear combination of the predictors is 38.4%, underlining the high relevance of marginality and, to a less extent, niche breadth and ROA (explaining 43.2%, 21.6%, and 21.5% of this amount, respectively). All the other predictors accounted for a relatively low explanatory capacity (from 0 to 6.7%; Table 1). Marginality is positively correlated with this component while the correlations of niche breadth and ROA are negative, indicating that the higher the marginality and the lower the niche breadth and ROA, the better the performance as measured by sensitivity and TSS (Fig. 1A).

Kappa, CCR, AUC, and specificity present the largest contributions to the second PLS component (28.8%, 21.1%, 21.7% and 18.5%, respectively; see Table 1). In this case, 31.0% of the variability is explained by the linear combination of the predictors, highlighting the explanatory capacity of data prevalence (27.0% of this amount), marginality (24.2%), range size (20.9%), and demographic rarity (16.0%). All these predictors are positively related to the values of this component, except in the case of rarity. All other predictors account for less than 4% of the total variability of this component. Thus, discrimination performance measured by Kappa, CCR, AUC, or specificity is higher when demographic rarity is lower, but data prevalence and the marginality of the species are higher, and their range sizes are larger (see Fig. 1B).

Environmental marginality is the variable that contributes the most for the explanation for the assessment metrics considering complete results, followed by data prevalence, range size, niche breadth, and ROA – which represent similar contributions considering their VIP values (Fig. 2). In contrast, demographic rarity, the three nesting behaviour strategies, niche completeness, and body size seem to have minor effects on performance when classic discrimination metrics are used to examine the predictive capacity of SDMs following standard procedures.

Table 1
Weights of predictors obtained by applying a Partial Least Squares Regression analysis (PLS), relating the two main components of the response variables (performance metrics) and predictors (species' and data characteristics) on the two significant PLS components.

	1st Component	2nd Component
Response variables		
Sensitivity	0.543	0.078
Specificity	0.331	0.431
Kappa	0.232	0.537
TSS	0.528	0.304
CCR	0.354	0.459
AUC	0.372	0.466
Predictor variables		
Body size	−0.019	0.024
Marginality	0.657	0.492
Abundance rarity	0.047	−0.400
Range size	−0.259	0.457
Data prevalence	−0.200	0.519
Niche breadth	−0.465	0.114
ROA	−0.464	0.129
Niche completeness	0.108	0.089
Nesting: Endocoprid	0.049	−0.071
Nesting: Paracoprid	0.014	0.187
Nesting: Telecoprid	−0.112	−0.202
Explained variability	38.37%	30.99%

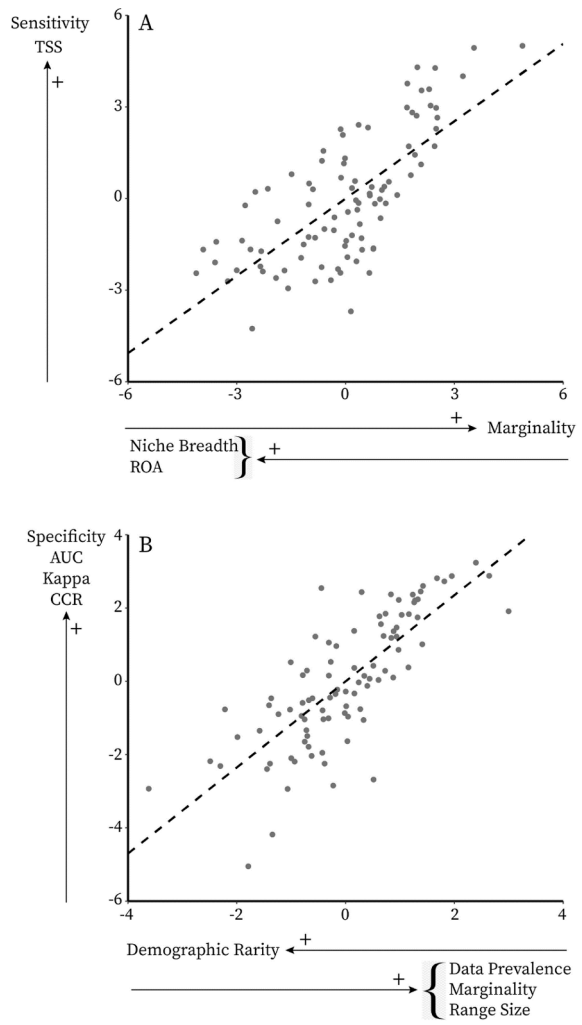


Fig. 1. Relationship between the X and Y scores (predictors and response variables, respectively) for the first (A) and second (B) PLS components, showing the performance metrics and species' and data characteristics that significantly influence these components and their signs.

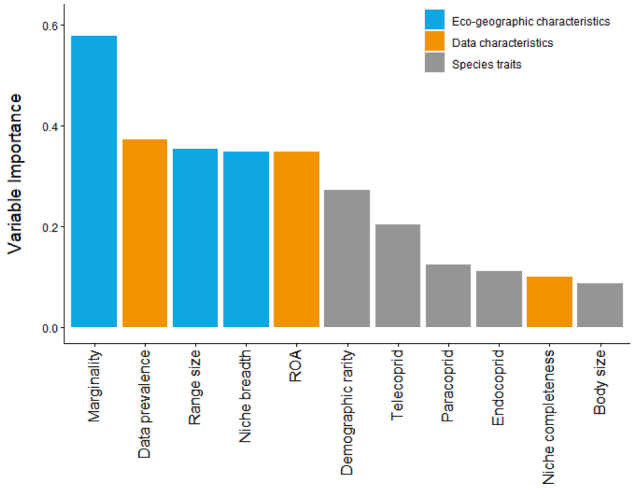


Fig. 2. Importance of species' and data characteristics for determining the variability in the metrics of model performance, according to the PLS analysis.

4. Discussion

Our results show that eco-geographic and data characteristics influence the accuracy of SDM predictions, as indicated by the six performance metrics analysed here. Evidence for a significant effect of ecological traits is however less clear in our study. Understanding the nature of these influences is difficult because all these characteristics are intricately related among them, so their relative causal effects cannot be dissociated. Indeed, demographic rarity –the only trait showing significant effects on some performance metrics– is negatively and curvilinearly related to eco-geographic variables such as range size and niche breadth, as well as with data characteristics such as ROA and data prevalence (Table S4; Fig. S1).

Environmental marginality shows a higher association with the predictive capacity of the SDMs. Those species whose occurrences are placed farther from the average conditions of the studied territory are predicted better, according to the standard SDMs and the performance metrics used here (see below). This variable, together with range size and data prevalence, is associated with specificity –the capacity of correctly predicting absence data– and therefore influences most performance metrics. To understand these results it is necessary to consider that the mean number of cells occupied by the studied species cover, on average, only 14.9% ± 1.7% of the total extent of analysis (mean ± 95% CI). In consequence, predicting absences correctly is much more decisive than predicting correctly the presences for obtaining better performance metrics. Range size and data prevalence are curvilinearly and negatively correlated with environmental marginality (see Table S4 and Fig. S1), so that the capacity of correctly predict absences is greater when the range size and the data prevalence of the species are comparatively higher (the maximum percentage of occupied cells in the study region is 37.9%). Niche breadth and ROA influence the predictive capacity of presences; the less the coverage of the total area of the study region provided by the observations, and the narrower the niche of the species, the greater is the apparent capacity of SDM techniques to predict accurately species occurrences. However, these two variables are also highly (and curvilinearly) correlated with environmental marginality (see Table S4 and Fig. S1).

Several studies found no relationship between marginality and SDM performance (Newbold et al., 2009; Pöyry et al., 2008), and others report opposite results about the influence of data prevalence and range size (Marmion et al., 2009; Sor et al., 2017; Syfert et al., 2013; van Proosdij et al., 2016). Those discrepancies have been attributed to the metrics used in each analysis, as many evaluation metrics are known to be sensitive to data characteristics (Allouche et al., 2006; Chefaoui et al., 2011; Leroy et al., 2018; Santika, 2011). We however expect that these effects do not affect the conclusions of our study, since although some of the metrics we evaluate are sensible to data characteristics, others are known to be insensitive to such an issue, and our results hold back for the combination of all of them. Rather, discrepancies with other studies may arise from treating species characteristics as independent despite the high degree of correlation among them (see Table S4). Indeed, the fact that certain traits covary altogether forming data syndromes makes difficult to identify which ones determine SDM accuracy the most, as they may interact among them, and probably most, if not all, influence model performance.

The effect of data and species' eco-geographic characteristics on SDM performance implies that predictions of species diversity and composition based on the overlap of various single-species models can present high levels of error due to the inclusion of low-quality models, hence leading to suboptimal or ineffective conservation actions (Aranda and Lobo, 2010; Hanspach et al., 2010; Zipkin et al., 2010). Therefore, we support the recommendation that species that are likely to show low SDM performance should be removed from the analyses (Hanspach et al., 2010; Pöyry et al., 2008). Fortunately, our results show a positive relationship between marginality and SDM performance. Marginality is generally linked with rarity, two characteristics that are associated with

a higher risk of extinction, and hence with species deeming higher priority for conservation. Further, species inhabiting uncommon habitats suffer from low genetic variation and are more impacted by climate change, due to 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 guide for conservation actions, provided that data quality has been previously assessed (Duputié et al., 2014; Hortal and Lobo, 2011). However, it is necessary to be cautious when a high discriminatory capacity is attributed to the SDMs of these rare species because high sensitivity, specificity, and AUC scores can be obtained although the distribution area is over-predicted (Lobo, 2008).

It is important to highlight that advancing in SDM accuracy requires investing in improving data quality. Our study has been carried out following the most common approach in predicting species distributions; i.e. using pseudoabsences, cross-validations, and ensemble models (Hao et al., 2019; Zurell et al., 2020). But the lack of reliable absence information in the data used for training SDMs unavoidably implies that an unknown number of absences apparently well predicted are in fact erroneous, thereby spuriously inflating predictive success according to most performance metrics. Besides the obvious need for further surveys, analyses of survey completeness may help to bridge this gap, as they may allow providing a higher likelihood of the absences coming from well-sampled areas, giving them more weight in model performance assessments.

The increasing use of species distribution models for applied and theoretical studies in ecology, evolution, and conservation studies requires a careful evaluation of the effects of sources of uncertainty on model performance. The importance of data characteristics found in our study calls for increasing the quality and completeness of data used in SDMs. Well-designed surveys and new methods to deal with the uncertainty will improve the performance of SDM predictions.

CRediT authorship contribution statement

Geiziane Tessarolo: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Jorge M. Lobo:** Methodology, Formal analysis, Writing - review & editing. **Thiago Fernando Rangel:** Resources, Writing - review & editing. **Joaquín Hortal:** Conceptualization, Methodology, Formal analysis, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the Brazilian CNPq project PVE 314523/2014-6. GT was funded by CAPES REUNI and PDSE grant n° 11842121.

Data accessibility

The datasets used for the analysis are freely available and can be accessed online as described below.

1 – Species distribution data come from SCAMAD, a database on the distribution of Madrid dung beetles. The two versions of this database used in this article, SCAMAD 1.0 (collected before 1999) and SCAMAD 2.1 (including collections until 2003), are are freely available at digital.csic open access repository (https://doi.org/10.20350/digitalCSIC/12534).

2 – Soil structure and composition are available at www.fao.org.

3 - Bedrock data are available at www.igme.es.

Funding

This work was funded by the Brazilian CNPq project PVE 314523/2014-6. GT was funded by CAPES REUNI and PDSE grant n° 11842121.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2020.107147.

References

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

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 (Cop.) 34, 31–38. https://doi.org/10.1111/j.1600-0587.2010.06134.x.

Araújo, M.B., Anderson, R.P., Barbosa, A.M., Beale, C.M., 2019. Standards for distribution models in biodiversity assessments. Sci. Adv. 5, eaat4858.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22, 42–47. https://doi.org/10.1016/j.tree.2006.09.010.

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

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Inform. 19, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Boone, R.B., Krohn, W.B., 1999. Modeling the occurrence of bird species: are the errors predictable? Ecol. Appl. 9, 835–848. https://doi.org/10.1890/1051-0761(1999)009[0835:MTOOBS]2.0.CO;2.

Breiman, L., 2001. Random Forests. Mach. Learn.

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

Busby, J.R., 1986. A biogeographical analysis of Notophagus cunninghamii (Hook.) in south-eastern Australia. Aust. J. Ecol. 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. https://doi.org/10.1111/j.1600-0706.2008.16881.x.

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. Biol. Conserv. 122, 327–338. https://doi.org/10.1016/j.biocon.2004.08.005.

Chefaoui, R.M., Lobo, J.M., Hortal, J., 2011. Effects of species' traits and data characteristics on distribution models of threatened invertebrates. Anim. Biodivers. Conserv. 34, 229–247.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Measure. 20, 37–46.

Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J., Wang, F., Liu, J., 2018. Effects of grain size and niche breadth on species distribution modeling. Ecography (Cop.) 41, 1270–1282. https://doi.org/10.1111/ecog.03416.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1023/A:1022627411411.

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. Ecol. Evol. 3, 2958–2975. https://doi.org/10.1002/ece3.601.

Duputié, A., Zimmermann, N.E., Chuine, I., 2014. Where are the wild things? Why we need better data on species distribution. Glob. Ecol. Biogeogr. 23, 457–467. https://doi.org/10.1111/geb.12118.

FAO - UNESCO, 1988. The FAO-UNESCO Soil Map of the World.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49. https://doi.org/10.1017/S0376892997000088.

França, S., Cabral, H.N., 2016. Predicting fish species distribution in estuaries: influence of species' ecology in model accuracy. Estuar. Coast. Shelf Sci. 180, 11–20. https://doi.org/10.1016/j.ecss.2016.06.010.

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. Divers. Distrib. 15, 167–177. https://doi.org/10.1111/j.1472-4642.2008.00536.x.

Gábor, L., Moudrý, V., Barták, V., Lecours, V., 2019. How do species and data characteristics affect species distribution models and when to use environmental filtering? Int. J. Geogr. Inf. Sci. 00, 1–18. https://doi.org/10.1080/13658816.2019.1615070.

Gábor, L., Moudrý, V., Lecours, V., Malavasi, M., Barták, V., Fogl, M., Šímová, P., Rocchini, D., Václavík, T., 2020. The effect of positional error on fine scale species distribution models increases for specialist species. Ecography (Cop.) 43, 256–269. https://doi.org/10.1111/ecog.04687.

Garrison, B.A., Lupo, T., 2002. Accuracy of bird range maps based on habitat maps and habitat relationship models. In: Scott, J.M., Heglund, B., Morrison, M.L., Wall, W.A.,

Haufler, J. (Eds.), Predicting Species Occurrences – Issues of Accuracy and Scale. Island Press, pp. 367–375.
Grenié, M., Violle, C., Munoz, F., 2020. Is prediction of species richness from stacked species distribution models biased by habitat saturation? Ecol. Indic. 111, 105970 https://doi.org/10.1016/j.ecolind.2019.105970.
Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R., Wintle, B.A., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr. 24, 276–292. https://doi.org/10.1111/geb.12268.
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. Conserv. Biol. 20, 501–511. https://doi.org/10.1111/j.1523-1739.2006.00354.x.
Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781139028271.
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? Ecol. Monogr. 77, 615–630. https://doi.org/10.1890/06-1060.1.
Guo, C., Lek, S., Ye, S., Li, W., Liu, J., Li, Z., 2015. Uncertainty in ensemble modelling of large-scale species distribution: effects from species characteristics and model techniques. Ecol. Modell. 306, 67–75. https://doi.org/10.1016/j.ecolmodel.2014.08.002.
Habel, K., Grasman, R., Gramacy, R., Stahel, A., Sterratt, D.C., 2015. geometry: Mesh generation and surface tesselation. R Packag. version 0.3-6.
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. Perspect. Plant Ecol. Evol. Syst. 12, 219–225. https://doi.org/10.1016/j.ppees.2010.04.002.
Hao, T., Elith, J., Guillera-Arroita, G., Lahoz-Monfort, J.J., 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Divers. Distrib. 25, 839–852. https://doi.org/10.1111/ddi.12892.
Henckel, L., Bradter, U., Jönsson, M., Isaac, N.J.B., Snäll, T., 2020. Assessing the usefulness of citizen science data for habitat suitability modelling: opportunistic reporting versus sampling based on a systematic protocol. Divers. Distrib. 26, 1276–1290. https://doi.org/10.1111/ddi.13128.
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 (Cop.) 29, 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x.
Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93, 679–688.
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. https://doi.org/10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2.
Hortal, J., de Bello, F., Diniz-filho, J.A.F., Lewinsohn, T.M., Lobo, J.M., Ladle, R.J., 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annu. Rev. Ecol. Evol. Syst. 46, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400.
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. https://doi.org/10.1111/j.2008.0030-1299.16434.x.
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. Nat. Conserv. 9, 200–207. https://doi.org/10.4322/natcon.2011.026.
Hortal, J., Lobo, J.M., 2005. An ED-based protocol for optimal sampling of biodiversity. Biodivers. Conserv. 14, 2913–2947. https://doi.org/10.1007/s10531-004-0224-z.
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.
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). Biodivers. Conserv. 10, 1343–1367.
Hortal, J., Lobo, J.M., Martín-Piera, F., 2020. SCAMAD - Base de datos corológicos y fenológicos acerca de la distribución de los escarabeidos coprófagos (Col. Scarabaeoidea) de Madrid. https://doi.org/10.20350/digitalCSIC/12534.
ITGE, 1988. Atlas Geocientífico y del Medio Natural de la Comunidad de Madrid. - Inst. Tecnológico GeoMinero de España., ITGE.
Jiménez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol. Biogeogr. 21, 498–507. https://doi.org/10.1111/j.1466-8238.2011.00683.x.
Jiménez-Valverde, A., Lobo, J.M., Hortal, J., 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Divers. Distrib. 14, 885–890. https://doi.org/10.1111/j.1472-4642.2008.00496.x.
Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol. Appl. 14, 401–413. https://doi.org/10.1890/02-5364.
Kellermann, V., Van Heerwaarden, B., Sgrò, C.M., Hoffmann, A.A., 2009. Fundamental evolutionary limits in ecological traits drive Drosophila species distributions. Science (80-.) 325, 1244–1246. https://doi.org/10.1126/science.1175443.
Ladle, R.J., Hortal, J., 2013. Mapping species distributions : living with uncertainty. Front. Biogeogr. 5, 8–9.
Leroy, B., Delsol, R., Hugueny, B., Meynard, C.N., Barhoumi, C., Barbet-Massin, M., Bellard, C., 2018. Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. J. Biogeogr. 45, 1994–2002. https://doi.org/10.1111/jbi.13402.
Lobo, J.M., 2008. More complex distribution models or more representative data? Biodivers. Informatics 5, 14–19.
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. https://doi.org/10.3989/graellsia.2006.v62.iExtra.126.
Lobo, J.M., Jiménez-Valverde, A., Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography (Cop.) 33, 103–114. https://doi.org/10.1111/j.1600-0587.2009.06039.x.
Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x.
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. Conserv. Biol. 16, 158–173.
Lomolino, M.V., 2004. Conservation Biogeography. In: Lomolino, M.V., Heaney, L.R. (Eds.), Frontiers of Biogeography: New Directions in the Geography of Nature. Sinauer Associates Inc, Sunderland, Massachussets, pp. 293–296.
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. Ecol. Modell. 220, 3512–3520. https://doi.org/10.1016/j.ecolmodel.2008.10.019.
Marshall, L., Carvalheiro, L.G., Aguirre-Gutiérrez, J., Bos, M., de Groot, G.A., Kleijn, D., Potts, S.G., Reemer, M., Roberts, S., Scheper, J., Biesmeijer, J.C., 2015. Testing projected wild bee distributions in agricultural habitats: predictive power depends on species traits and habitat type. Ecol. Evol. 5, 4426–4436. https://doi.org/10.1002/ece3.1579.
Martín-Piera, Fermín, 2000. Familia Scarabaeidae, in: Martín-Piera, F, López-Colón, J.I. (Eds.), Coleoptera, Scarabaeoidea I. In: Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Madrid, pp. 207–432.
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. J. Veg. Sci. 21, 908–922. https://doi.org/10.1111/j.1654-1103.2010.01198.x.
McCullagh, P., Nelder, J.A., 1991. Generalized linear models, Second ed. Chapman and Hall, London, UK.
McCune, J.L., Rosner-Katz, H., Bennett, J.R., Schuster, R., Kharouba, H.M., 2020. Do traits of plant species predict the efficacy of species distribution models for finding new occurrences? Ecol. Evol. 10, 5001–5014. https://doi.org/10.1002/ece3.6254.
Meynard, C.N., Quinn, J.F., 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. J. Biogeogr. 34, 1455–1469. https://doi.org/10.1111/j.1365-2699.2007.01720.x.
Moretti, M., Dias, A., de Bello, F., Altermatt, F., Chown, S.L., Azcárate, F.M., Bell, J.R., Fournier, B., Hedde, M., Hortal, J., Ibanez, S., Öckinger, E., Souza, J.P., Ellers, J., Berg, M.P., 2016. Handbook of protocols for standardized measurement of terrestrial invertebrate functional traits. Funct. Ecol. in press.
Moudrý, V., Šímová, P., 2013. Relative importance of climate, topography, and habitats for breeding wetland birds with different latitudinal distributions in the Czech Republic. Appl. Geogr. 44, 165–171. https://doi.org/10.1016/j.apgeog.2013.08.001.
Naimi, B., Araújo, M.B., 2016. sdm: a reproducible and extensible R platform for species distribution modelling. Ecography (Cop.) 39, 368–375. https://doi.org/10.1111/ecog.01881.
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. Biodivers. Conserv. 18, 3629–3641. https://doi.org/10.1007/s10531-009-9668-5.
Oliveira, U., Paglia, A.P., Brescovit, A.D., de Carvalho, C.J.B., Silva, D.P., Rezende, D.T., Leite, F.S.F., Batista, J.A.N., Barbosa, J.P.P.P., Stehmann, J.R., Ascher, J.S., de Vasconcelos, M.F., De Marco, P., Löwenberg-Neto, P., Dias, P.G., Ferro, V.G., Santos, A.J., 2016. The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. Divers. Distrib. 22, 1232–1244. https://doi.org/10.1111/ddi.12489.
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. J. Environ. Manage. 62, 171–184. https://doi.org/10.1006/jema.2001.0425.
Phillips, S., Anderson, R., Schapire, R., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Modell. 190, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.
Phillips, S.J., 2008. Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). Ecography (Cop.). 31, 272–278. https://doi.org/10.1111/j.2007.0906-7590.05378.x.
Pöyry, J., Luoto, M., Heikkinen, R.K., Saarinen, K., 2008. Species traits are associated with the quality of bioclimatic models. Glob. Ecol. Biogeogr. 17, 403–414. https://doi.org/10.1111/j.1466-8238.2007.00373.x.
Reese, G.C., Wilson, K.R., Hoeting, J.A., Flather, C.H., 2005. Factos affecting species distribution predictions: a simulation modeling experiment. Ecol. Appl. 15, 554–564.
Regos, A., Gagne, L., Alcaraz-Segura, D., Honrado, J.P., Domínguez, J., 2019. Effects of species traits and environmental predictors on performance and transferability of ecological niche models. Sci. Rep. 9, 1–14. https://doi.org/10.1038/s41598-019-40766-5.
Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A., 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Prog. Phys. Geogr. 35, 211–226. https://doi.org/10.1177/0309133311399491.

Santika, T., 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. Glob. Ecol. Biogeogr. 20, 181–192. https://doi.org/10.1111/j.1466-8238.2010.00581.x.

Seoane, J., Carrascal, L.M., Alonso, C.L., Palomino, D., 2005. Species-specific traits associated to prediction errors in bird habitat suitability modelling. Ecol. Modell. 185, 299–308. https://doi.org/10.1016/j.ecolmodel.2004.12.012.

Sheth, S.N., Lohmann, L.G., Consiglio, T., Jiménez, I., 2008. Effects of detectability on estimates of geographic range size in Bignonieae. Conserv. Biol. 22, 200–211. https://doi.org/10.1111/j.1523-1739.2007.00858.x.

Sor, R., Park, Y.-S., Boets, P., Goethals, P.L.M., Lek, S., 2017. Effects of species prevalence on the performance of predictive models. Ecol. Modell. 354, 11–19. https://doi.org/10.1016/j.ecolmodel.2017.03.006.

Stockwell, D., Peterson, T.A., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Modell. 148, 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X.

Syfert, M.M., Smith, M.J., Coomes, D. a., 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. PLoS One 8, 1–10. https://doi.org/10.1371/journal.pone.0055158.

Tessarolo, G., Rangel, T.F., Araújo, M.B., Hortal, J., 2014. Uncertainty associated with survey design in Species Distribution Models. Divers. Distrib. 20, 1258–1269. https://doi.org/10.1111/ddi.12236.

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. Biol. Lett. 6, 120–123. https://doi.org/10.1098/rsbl.2009.0669.

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 (Cop.) 27, 165–172. https://doi.org/10.1111/j.0906-7590.2004.03673.x.

van Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J., Raes, N., 2016. Minimum required number of specimen records to develop accurate species distribution models. Ecography (Cop.) 39, 542–552. https://doi.org/10.1111/ecog.01509.

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. https://doi.org/10.1111/j.2007.0030-1299.15559.x.

Warren, D.L., Matzke, N.J., Iglesias, T.L., 2020. Evaluating presence-only species distribution models with discrimination accuracy is uninformative for many applications. J. Biogeogr. 47, 167–180. https://doi.org/10.1111/jbi.13705.

Wogan, G.O., 2016. Life history traits and niche instability impact accuracy and temporal transferability for historically calibrated distribution models of North American birds. PLoS One 11, 1–22. https://doi.org/10.1371/journal.pone.0151024.

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

Zamorano, D., Labra, F.A., Villarroel, M., Lacy, S., Mao, L., Olivares, M.A., Peredo-Parada, M., 2019. Assessing the effect of fish size on species distribution model performance in southern Chilean rivers. PeerJ 7, e7771. https://doi.org/10.7717/peerj.7771.

Zipkin, E.F., Andrew Royle, J., Dawson, D.K., Bates, S., 2010. Multi-species occurrence models to evaluate the effects of conservation and management actions. Biol. Conserv. 143, 479–484. https://doi.org/10.1016/j.biocon.2009.11.016.

Zurell, D., Franklin, J., König, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J.J., Leitão, P.J., Park, D.S., Peterson, A.T., Rapacciuolo, G., Schmatz, D.R., Schröder, B., Serra-Diaz, J.M., Thuiller, W., Yates, K.L., Zimmermann, N.E., Merow, C., 2020. A standard protocol for reporting species distribution models. Ecography (Cop.) ecog.04960. https://doi.org/10.1111/ecog.04960.