ECOGRAPHY

Research

Using maps of biogeographical ignorance to reveal the uncertainty in distributional data hidden in species distribution models

Geiziane Tessarolo, Richard J. Ladle, Jorge M. Lobo, Thiago Fernando Rangel and Joaquín Hortal

G. Tessarolo (https://orcid.org/0000-0003-1361-0062) ✉ (geites@gmail.com), T. F. Rangel (https://orcid.org/0000-0002-2001-7382) and J. Hortal (https://orcid.org/0000-0002-8370-8877), Depto de Ecologia, Inst. de Ciências Biológicas – ICB, Univ. Federal de Goiás – UFG Campus II, Goiânia, GO, Brazil. GT also at: Laboratório de Biogeografia e Ecologia Aquática, Univ. Estadual de Goiás, CCET, Anápolis, GO, Brazil. – R. J. Ladle, CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Univ. do Porto, Vairão, Portugal and BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Vairão, Portugal. – J. M. Lobo (https://orcid.org/0000-0002-3152-4769) and JH, Dept of Biogeography and Global Change, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain.

Ecography
44: 1743–1755, 2021
doi: 10.1111/ecog.05793

Subject Editor:
Christine N. Meynard
Editor-in-Chief: Miguel Araújo
Accepted 14 September 2021

Species distribution models (SDMs) are subject to many sources of uncertainty, limiting their application in research and practice. One of their main limitations is the quality of the distributional data used to calibrate them, which directly influences the accuracy of model predictions. We propose a standardized methodology to create maps, describing the limitations of occurrence data for covering the distribution of a species. We develop a set of tools based on the general framework of Maps of Biogeographical Ignorance to describe the main sources of data-driven uncertainty: taxonomic stability, environmental similarity, geographical proximity and temporal decay of the underlying biodiversity data. The so-derived indicators of data-driven uncertainty account for inventory completeness, taxonomic quality, time since the surveys and geographical (and environmental) distance to localities with information. These indicators form the basis of ignorance maps, which can be used to visualize the reliability of SDM projections in geographical space, to estimate the uncertainty of these predictions and to identify target survey areas. To demonstrate the application of our approach, we use data on fourteen Iberian species of Scarabaeidae dung beetles. Data-driven uncertainty is widespread even for this well-surveyed group; more than 60% of the region has distributional uncertainty values higher than 0.6, and 30% higher than 0.7. Ignorance maps can be jointly evaluated with SDM predictions to generate spatially explicit maps of uncertainty, identifying where predictions are reliable/unreliable. Neglecting such uncertainty can severely affect SDM effectiveness, as it can introduce biases and inaccuracies into the measured species–environment relationships. These errors could result in incorrect theoretical or practical applications, including ill-advised conservation actions. We therefore advocate for the routine use of ignorance maps or similar techniques as supporting information in SDM applications.

Keywords: bias, biodiversity data, distributional uncertainty, ecological niche models, predictive accuracy, SDM performance, spatial decay, taxonomic quality, temporal decay, Wallacean shortfall

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Conservation science has been framed as a crisis discipline, where rapid decisions often need to be made on the basis of old, limited and incomplete data (Soulé 1985). This challenge becomes increasingly difficult over large temporal and spatial scales due to the limited resources available for biodiversity research (Yoccoz et al. 2001) and the practical difficulties of collecting observational data (Blackburn 2004). As a consequence, our knowledge of species distributions, known as the Wallacean shortfall (Lomolino 2004, Hortal et al. 2015), is often strongly biased with significant deficits for hyperdiverse taxa and regions (Bini et al. 2006, Hortal et al. 2007, Rocchini et al. 2011, Yang et al. 2013, Stropp et al. 2016).

In the absence of reliable and widespread observational data, scientists have developed several methods to estimate distributions (Hortal 2008), the most widely used of which are the so-called species distributions models (SDMs). SDMs are functions relating a response variable representing the observed distribution of a species against a set of environmental predictors, and are used to study a wide variety of issues, from species discovery to conservation planning (Peterson 2006, Richards et al. 2007, Williams et al. 2009, Austin and Van Niel 2011, Arias-González et al. 2012, Bhatt et al. 2013, Guisan et al. 2013, Fagundes et al. 2016). SDMs are closely related to Environmental Niche Models and Habitat Suitability Models; although these three terms are often treated as synonyms, we favour SDMs as it unequivocally indicates that the modelled phenomena are species distributions. Despite their widespread use and increasing sophistication, SDMs still have conceptual, methodological and statistical issues that are yet to be fully resolved (Araújo and Guisan 2006, Austin 2007, Hortal et al. 2012, Peterson and Soberón 2012, Rangel and Loyola 2012, Lobo 2016). One outstanding concern is the acceptable degree of uncertainty associated with SDM predictions (Barry and Elith 2006, Dormann et al. 2008, Aranda and Lobo 2010, Beale and Lennon 2012, Tessarolo 2012). Uncertainty is unavoidable in any statistical process in which a prediction is made from data. In the case of SDMs, uncertainties are introduced in all stages of the modelling process, including the spatial precision and selection of distributional data and the ancillary variables used as predictors, the algorithms used for modelling, model evaluation, model projection into geographical space, or the application of thresholds to distinguish areas of predicted presence (Heikkinen et al. 2006, Hortal et al. 2008, Diniz-Filho et al. 2009, Buisson et al. 2010, Grenouillet et al. 2011, Watling et al. 2015, Tessarolo et al. 2021). Such issues are not solely of academic interest; SDM outputs are frequently taken at face value (Wilson 2010), potentially leading to the ineffective use of limited conservation resources.

The quantification, visualization and effective communication of the uncertainties of the data underpinning SDMs is thus a critical step to evaluate the outputs of these ubiquitous conservation tools (Rocchini et al. 2011, Beale and Lennon 2012, Gould et al. 2014, McInerny et al. 2014, Araújo et al. 2019). Indeed, evaluating uncertainty in model predictions has attracted much attention (Barry and Elith 2006, Graham et al. 2007, Wisz et al. 2008, Syphard and Franklin 2009, Lobo and Tognelli 2011, Naimi et al. 2011, 2014, Kramer-Schadt et al. 2013, Swanson et al. 2013, Tessarolo et al. 2014, 2021). These studies cover different stages of the SDM process, including model selection, parametrization and selection of predictors. However, few studies propose methodologies to quantify the effect of the quality of the input data used to calibrate the models, despite being widely recognized as an important source of error (Hirzel and Guisan 2002, Edwards Jr. et al. 2006, Albert et al. 2010, Braunisch and Suchant 2010, Naimi et al. 2014, Niamir et al. 2019). In this study, we use uncertainty as a concept related with building a statistical model. However, although the characteristics and quality in the input data can influence SDM predictions (Rocchini et al. 2011, Kamino et al. 2012, Graham and Kimble 2019), they should be treated as ignorance about the predicted response variable. One way to counter misplaced perceptions of certainty in spatial reconstructions is through the parallel construction of 'maps of ignorance' for the modelled response variable. This idea can be traced back to Boggs' (1949) remarkable proposal for creating an 'atlas of the ignorance' about a wide range of spatially explicit phenomena, to represent the geographical variations in the level and accuracy of knowledge about the thematic maps that were typically depicted in most contemporary atlases. More recently, Rocchini et al. (2011) suggested that maps identifying where species distribution data are reliable and where they are uncertain would be ideal to highlight uncertainties in the predictions of SDMs. Extending this idea, Ladle and Hortal (2013) outlined the conceptual basis for 'Maps of Biogeographical Ignorance' incorporating measures of the quality, longevity and coverage of distributional data (Meyer et al. 2015, 2016, Stropp et al. 2016).

The development of maps that depict the limitations of distributional information should address at least four main sources of biogeographical ignorance (Ladle and Hortal 2013): 1) inventory completeness of original survey data; 2) taxonomic quality of data; 3) the time elapsed since the last survey was completed; and 4) the geographical (and environmental) distance between the localities with and without information on the occurrence of the studied taxa. Maps that integrate these four sources of ignorance could provide clear information on the consequences of using distributional data indiscriminately (Ladle and Hortal 2013). On the one hand, they can provide an intermediate step for improving the coverage and representativeness of biodiversity data, as they can be used to identify areas where more sampling effort would provide maximum benefit for conservation modelling efforts (Ronquillo et al. 2020, Sobral-Souza et al. 2021). On the other, they can be used in association with SDMs to assess and visualize the uncertainty related to the data used to generate these models. Specifically, they can be used to identify areas where model reliability is high/low according to how well the original dataset represents the conditions in these areas and to determine different levels of certainty about model predictions; such analyses can then be used to

support the development of informed conservation actions (Rocchini et al. 2011).

Here, we propose a methodology for creating Maps of Biogeographical Ignorance (MoBIs) to measure spatial reliability of knowledge about species distributions. Further, we show how to use these maps in combination with SDMs to provide an account of the uncertainty driven by distributional data that is inherent to their model projections. It is important to note that our approach does not cover other sources of uncertainty, such as positional errors and scaling (Moudrý and Šímová 2012), environmental data not considered (Beale and Lennon 2012), demographic variations (Chen et al. 2019) or species characteristics (Tessarolo et al. 2021). Building on the conceptual framework of Ladle and Hortal (2013), our analytical approach integrates the uncertainty due to variations in the quality and the spatial and temporal decay of the information provided by distributional data. As proof of concept, we use occurrence data to generate MoBIs, SDMs and the associated maps, taking into account this lack of knowledge or quality in the primary data, for 14 Scarabaeidae dung beetle species from the Iberian Peninsula.

Material and methods

Biological data

We used data on the distribution of Scarabaeidae species in the Iberian Peninsula derived from the latest version of BANDASCA (Lobo and Martín-Piera 1991), which can be freely downloaded from GBIF (Data availability statement). This database combines all distributional information about Iberian dung beetle species of this family available in the literature, museums and private collections and unpublished data. In total, BANDASCA currently hosts information on about ca 97 000 records from the 53 currently recognized Iberian Scarabaeidae dung beetle species. We then selected the 14 species with the highest numbers of records (Supporting information), for which we model their distribution, create MoBIs and represent jointly model results and uncertainty due to the quality in the primary data at a resolution of 10 km width UTM cells.

Environmental data

Data on climate and topography were obtained from WorldClim (Fick and Hijmans 2017) and the Global Resource Information Database – United Nations Environment Programme (UNEP/GRID, <www.grid.unep.ch/data/data.php>). All variables were rescaled to a resolution of 10 km (EPSG:4230 – ED50, projection: longlat, Ellipsoid: International 1924). Then, a principal components analysis (PCA) was carried out with all 29 variables included in these two datasets (Baselga and Araújo 2009). Five PCA axes with eigenvalues >1 were selected, accounting for 91.04% of the climatic and topographic variation in the Iberian Peninsula (Supporting information). For each of these factors, we selected the variable with highest factor loading for subsequent analyses (Supporting information). Hence, isothermality, mean temperature of wettest and warmest quarter, precipitation of coldest quarter and solar radiation monthly average were used to generate SDMs for all species and to create the environmental distance matrix needed for the development of MoBIs (below).

Species distribution models

Species distribution models were generated using five widely used and standard modelling techniques, namely: BIOCLIM (Busby 1986; following the method of percentile range); generalized linear models (GLM; McCullagh and Nelder 1989; using a logistic regression only with linear terms allowed); maximum entropy (MaxEnt; Phillips et al. 2006; set as 'Auto feature', in which feature types – linear, quadratic, product, threshold and hinge – and regularization parameters are automatically selected); generalized additive models (GAM; Hastie and Tibshirani 1987; allowing any quadratic penalized GLM and using penalized regression splines with GCV smoothness method); and random forest (Breiman 2001; with number of trees = 1000). We also generated the combined consensus among them (Araújo and New 2007).

The occurrence data used to calibrate and validate the SDMs were scaled to the same resolution of the environmental variables (10 × 10 km). Thus, for each cell, only one record of occurrence was considered for each species. As all the SDM techniques except BIOCLIM require the use of absence data, we followed the standard procedure of relating presences versus randomly selected absences (i.e. pseudo-absences) for the training and validation process in order to provide results comparable to those in other published studies. Thus, pseudo-absences were selected at random for each species within the Iberian Peninsula avoiding presence points, choosing the same number of presences as absences to avoid effects of prevalence (Jiménez-Valverde et al. 2009, Barbet-Massin et al. 2012).

The data (records and pseudo-absences) were randomly split into calibration and validation datasets (totalling 75% and 25% of the data, respectively), and the models were constructed using the first dataset and validated with the latter. This procedure was repeated 10 times, and the results were used to generate predictive maps of the probability of occurrence of each species according to each SDM technique, based on the mean value of the 10 predictions. Split and replication procedures were performed to increase the chance that all presences and different sets of pseudo-absences would be used to generate the models. The threshold used to convert the predicted probabilities into binary outcomes to calculate accuracy metrics was the one that maximizes the sum of sensitivity and specificity (Jiménez-Valverde and Lobo 2007). The accuracy of the binary models was assessed by sensitivity, specificity, kappa and true skill statistic (TSS) metrics. Here we have to highlight that, in this case, specificity, kappa and TSS do not indicate the capacity of predictions to reflect the actual distribution of the species, as real absence data is

lacking. Thus, they are calculated only to provide a measure of internal validation (as obtained frequently in SDMs studies). The consensus among the five SDM techniques was generated through the sensitivity-weighted mean (Araújo and New 2007). We chose the sensitivity statistic because it takes into account only the true positive rate. The sdm R package (Naimi and Araújo 2016) was used to perform all SDM analyses.

Creating Maps of Biogeographical Ignorance

We used a combination of analytical and visualization tools to represent the main sources of biogeographical ignorance identified by Ladle and Hortal (2013). Specifically, we decomposed these sources into four quantifiable components: data completeness, taxonomic quality of data, temporal decay in the information provided by these data and the spatial and environmental distance to a surveyed site.

Data completeness

Failure to detect a species when it is actually present in a locality may be due to insufficient sampling effort. Therefore, we characterize survey completeness (herein completeness for short) through species accumulation curves. Species accumulation curves relate the number of species that are added to the inventory as the number of database records increases. Here, the slope value represents the current rate of accumulation of new species in the inventory. We used data about all species in the database to generate the species accumulation curves. We generated 100 species accumulation curves per cell (with a minimum of 50 individuals, below) by randomizing the order of entry of the individuals in the curve. The mean of these 100 slope values was subtracted from 1 (so they ranged from 0 to 1, with 1 indicating a complete inventory) and then used as value of completeness of the cell. As species accumulation curves are sensitive to low sample sizes, we considered that all cells with less than 50 recorded individuals had a completeness of zero. Hortal and Lobo (2005) for further details about the use of species accumulation curves in this context. The use of data of all the species of a high-rank taxonomic group to generate accumulation curves produces low values of ignorance for well-surveyed cells in which the studied species were not found; such cells probably reflect true absence data for the non-recorded species. It is important to note that the number of individuals in a cell may be influenced by the availability of adequate habitat for the studied group. Thus, cells with a lower amount of appropriate habitat will probably have a lower number of individuals even if they have been well sampled. We performed a sensitivity analysis (below) by changing the number of individuals needed to consider a cell as sufficiently well-sampled to generate species accumulation curves and calculate completeness values.

Taxonomic quality

One of the less acknowledged sources of uncertainty in species distributions is the taxonomic quality of the records. The accuracy of the taxonomic identifications may vary with the experience of the identifier (Shea et al. 2011, Mackay-Smith and Roberts 2019, Egli et al. 2020), but also decay with time as new taxonomic revisions (e.g. splits and clumps) come into force; older taxonomic identifications may be synonymized by updates in the taxonomic knowledge of a given group (Baselga et al. 2010). Therefore, we estimate taxonomic quality for all the records in the database, according to the type of identifier (i.e. the person responsible for the identification). We classified identifiers into three different classes: 1) taxonomists – who work on systematics, describing species and revising their taxonomic status; 2) experts – who have extensive hands-on experience of the studied group, and are therefore used to conducting taxonomic identifications, and; 3) amateurs – who do not work with the studied group, or do it only occasionally and have limited experience with taxonomic identifications. Thus, a taxonomist is considered to produce higher quality identifications than an expert, and the same follows from expert to amateur. Drawing on personal experience working with amateurs, students and specialists who study this group, we set the taxonomic quality values as 1, 0.9 and 0.75 for taxonomist, expert and amateur identifiers, respectively.

Temporal decay in the information

The distributions of species (and therefore the composition of local assemblages) change through time for a variety of reasons (Buckley 2013, Diekmann et al. 2014, Del Vecchio et al. 2015, Van der Sande et al. 2016, Pöysä et al. 2019). Consequently, as the time since each record was obtained increases, the likelihood that the recorded species is currently absent from a locality increases as well (Tessarolo et al. 2017). Furthermore, a species' realized niche may also change over time (Boitani et al. 2011, Silva et al. 2016, Tessarolo et al. 2017). Thus, the older a record used to generate SDM, the higher the uncertainty associated with the descriptions of current species niche and distribution that are based on it. To simulate such temporal decay in the information provided by each occurrence record, we used a kernel Gaussian function to increase uncertainty with the increment in years since the recording. Here we assume that this function accounts for the non-linearity in such decay in effective knowledge. In other words, the influence of time on the quality of the information provided by distributional data is not linear; rather, its influence may be lower over shorter periods of time. In the absence of data and/or hypotheses about the rate of decrease in information quality, we adjusted this decay function to the total period encompassed by the records available in the database, i.e. the most recent and the oldest records in the database will be the most and least informative, respectively. These values ranged from 0 to 1, with 1 indicating the highest quality.

Spatial and environmental decay in the information

The first law of geography states that 'everything is related to everything else, but near things are more related than distant things' (Tobler 1970). It follows that the quality of the information about any given phenomenon in an unsampled

location is dependent on the quality of the information in the sampled locations that are spatially close and/or environmentally similar. In the case of species' geographic ranges, it is well known that species distributions are determined by both environmental conditions and biogeographical and ecological processes that leave a spatially explicit pattern in the geographical space (Nekola and White 1999). Therefore, we assumed that the amount of information about the presence (or absence) of a species in an insufficiently surveyed location diminishes when the spatial and environmental distance to surveyed locations increases. Such spatial and environmental decay was estimated based on two distance matrices: a spatial matrix and an environmental matrix. The spatial distance matrix was calculated using the Haversine geographic distances between all pairs of cells (thus accounting for the curvature of Earth's surface), whereas the environmental matrix was constructed using Euclidean multivariate distances based on the same environmental variables that were used for SDM predictions (above). These matrices were rescaled to range from 0 to 1, and were combined afterwards by multiplication with exponential weighting (0.5) – thus providing the same weight for both matrices (Eq. 1).

$$(Space.Dist_{ij})^{0.5} \times (Envi.Dist_{ij})^{0.5} = Spt.Env.Dist_{ij} \quad (1)$$

where: $Space.Dist_{ij}$ is the spatial distance between the cells i and j, $Envi.Dist_{ij}$ is the spatial distance between the cells i and j and $Spt.env.Dist_{ij}$ is the spatio-environmental distance between the cells i and j. The spatio-environmental distance matrix generated this way was later used to interpolate the degree of biogeographical knowledge (below).

Each one of the components described above was calculated for each record in the database. Thus, each record has independent values of taxonomic quality and temporal decay, but records in the same cell have the same completeness value. Having a value of each component for each record allowed us to calculate a biogeographical ignorance value for each record (below).

Assembling Maps of Biogeographical Ignorance

A biogeographical ignorance value for each record was attributed according to a combination of the four components described above (Fig. 1; Supporting information). For each record in the database, the final amount of information value was calculated as the sum of the completeness, taxonomic quality and temporal decay values. Here, we consider that all factors are equally important in providing knowledge about species distributions. It should be noted that these factors and their parameters can vary in their influence for specific species groups and areas (Discussion). The summed values were standardized in such a way that the lowest value became zero and the highest value one. As all the components are initially calculated so that a score of 1 indicates high quality, the standardized value was scaled from 1, to obtain biogeographical ignorance scores that vary from 0 to 1 (complete knowledge and complete ignorance, respectively; Fig. 1). For cells with more than one record for the same species, the biogeographical ignorance value was set as the minimum value of all these records, assuming that such record provides the most reliable and up-to-date account of the occurrence of the species (above).

After calculating the biogeographical ignorance value for each record, we generated MoBIs for each species (focal species). A biogeographical ignorance value for a focal species was attributed to each cell of the studied in one of three ways depending on whether the cell hosts original data on the focal species and the studied taxonomic group or not. First, if the cell hosts records for the focal species, the biogeographical ignorance value for the cell is that from the most reliable record for the focal species. That is, the lowest biogeographical ignorance value among all the records for the focal species is used to represent the highest degree of information on the distribution of the species. Second, if the cell holds records for other species of the taxonomic group, but not for the focal species, the biogeographical ignorance value for the cell is calculated as the mean of minimum biogeographical ignorance values of each species in the cell (Fig. 1). The reasoning behind the use of biogeographical ignorance of other species as a measure of biogeographical ignorance for focal species is that, if enough data have been collected for other species in the group (organisms that share similar sampling protocol), but the focal species have not been recorded, this indicates a likely 'real' absence for the focal species. Thus, data collection of other species also provides information about the distribution of the focal species, justifying the use of biogeographical ignorance of other species to understand the biogeographical uncertainty of the focal species. Finally, for cells without original data for the studied group, environmental and spatial proximity to cells with data is the only basis on which to calculate ignorance. Specifically, for cells without data on focal taxa, ignorance values are interpolated from the values of influent cells using spatio-environmental distance (details above). Thus, ignorance values for cells without data depend on their environmental and spatial proximity to influent cells (see below for details about the definition of influent cells). The influent cells may contain both cells with occurrence of focal species and cells with no record of this species but with presences of other species belonging to the studied group. Also, for influent cells without data, the biogeographical ignorance value for interpolation is assumed to be zero. Interpolated values were calculated according to the spatial–environmental distance matrix by means of the inverse distance weighting average (IDWA). This is a local interpolation method in which the values of unknown points are estimated using a linear combination of values at known data points (selected by a radius of influence) weighted by the inverse of the distance. This method assumes that the closer the estimated point is to a known data point, the stronger the influence of such known data point will be. To define the radius of influence, we performed a correlogram with the environmental variables used in the SDM calibration. Here, the farthest distance class with correlation above zero was used as the maximum distance at which a cell can

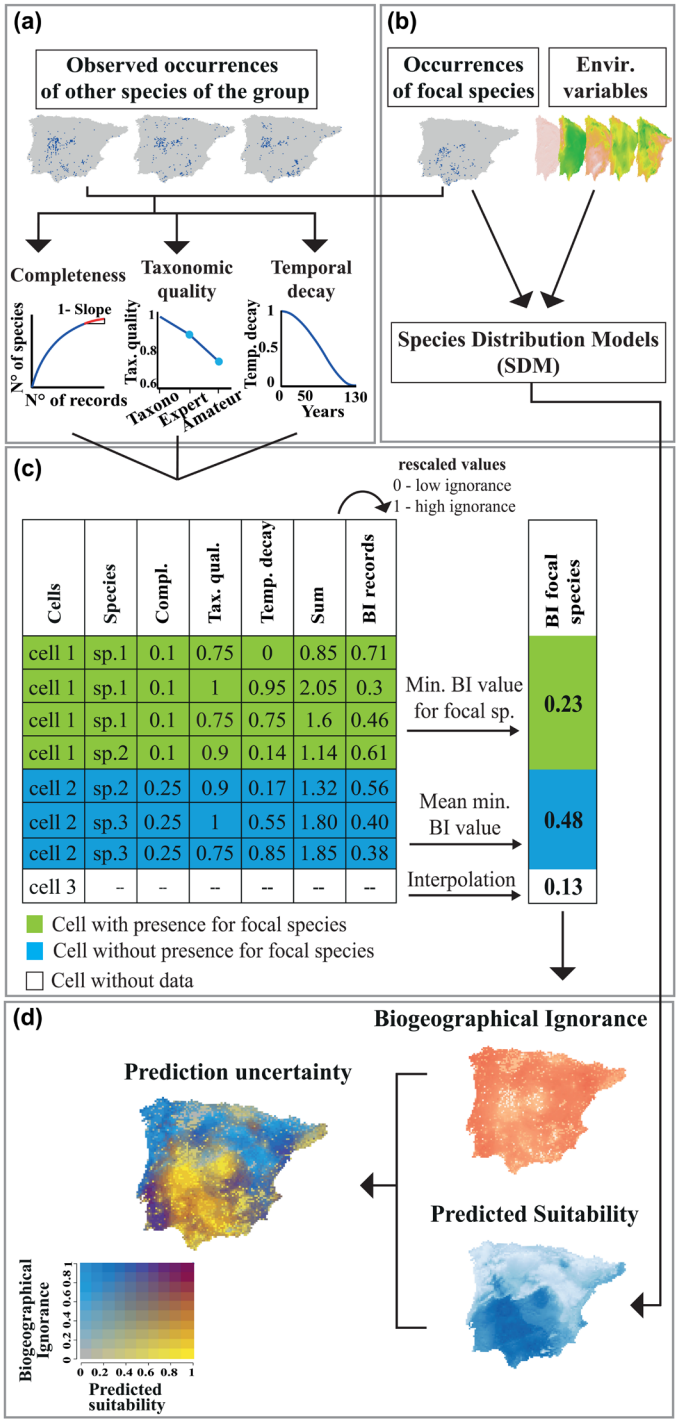

Figure 1. Methods for assembling Maps of Biogeographical Ignorance and jointly visualizing predictions and uncertainty. (a) Data on selected environmental variables and observed occurrences for the focal species are used to generate SDMs. (b) Completeness is calculated for each grid cell in the studied region, and taxonomic quality and temporal decay are calculated for each record in the database, including records for the focal species and for other species of the taxonomic group. (c) Values of each component are summed and then rescaled to vary from 0 to 1 (low–high biogeographical ignorance; BI). Records are separated by cells. For cells with more than one record for the focal species (sp. 1; green records for sp. 1 in cell 1), the final BI value is set up as the minimum value of all these records (note that the record for sp. 2 is ignored, as it is not related to the focal species). For cells without records for the focal species, but with records for other species of the taxonomic group (blue records in cell 2), the final biogeographical ignorance value is the mean of the minimum value of each species in the cell. For cells without any data on species distribution, final biogeographical ignorance value is based on interpolation considering values of neighbouring influent cells and the spatio-environmental distance (cell 3, white). (d) Values of predicted suitability and biogeographical ignorance are overlaid to generate bivariate maps of prediction uncertainty.

be influent on others. Only cells placed closer to the estimated cell than this distance (388.4 km in our analyses) were assumed to influence the interpolated value. Additionally, if there were recorded presences of the focal species within this radius, the corresponding cells were assumed to have a higher contribution to the interpolated value. Therefore, these cells were weighted (see q_i values below) to account for 50% of the interpolated value, whereas the remaining 50% weight was divided across the remaining cells. All these calculations are limited to the geographical domain analysed (in our example the Iberian Peninsula), discarding unsuitable territories such as sea or large water extensions.

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

$$Z_j = \frac{\sum_{i=1}^{n}\left(\frac{Z_i.q_i}{\left(h_{ij}\right)^b}\right)}{\sum_{i=1}^{n}\left(\frac{q_i}{(h_{ij})^b}\right)} \quad (2)$$

…where Z_j is the ignorance value interpolated to cell j, Z_i is the known value of biogeographical ignorance for the influent i cell, h_{ij} is the spatio-environmental distance between the cell j and i, the b is the weighting exponent (here we used p = 2, known as inverse distance squared weighted interpolation), q_i is the additional weight corresponding to cell i (here used to set different importance for cells depending on they have presence of the focal species or not) and n is the number of influent neighbours (i.e. those cells lying within the radius described above). By using this approach, we ensure that the final biogeographical ignorance value for cells lacking previous information is more influenced by cells located closer in space and/or in the environment and with presence records for the focal species. The final MoBI for each species were generated with the following parameters: completeness calculation for cells with a minimum of 50 individuals (Comple50 in the Supporting information); taxonomic quality of 1, 0.9 and 0.75 for taxonomist, expert and amateur identifiers, respectively (Tax_qual1 in the Supporting information); temporal decay based on a kernel Gaussian function (T_decay1 in the Supporting information) and spatio-temporal matrix generated by use of equally exponential weighting (0.5) for space and environmental distance matrix (red lines in the Supporting information). All analyses were implemented in R environment (<www.r-project.org>). The R-script used to calculate biogeographical ignorance values can be found in the Supporting information.

Sensitivity analyses

As the choice of the parameters for the generation of MoBIs is subjective, we carried out a series of sensitivity analyses to evaluate how parameter configurations may affect the resulting MoBIs. For the completeness component, we changed the number of individuals needed to consider a cell as well-sampled enough to construct species accumulation curves and calculate completeness values for it. Thus, we created MoBIs with completeness calculated for cells with a minimum of 10, 20, 30, 40 and 50 individuals (Supporting information). For the taxonomic quality component, we set values for identification quality with 0.5, 1 and 1.5 quality of difference among identifiers classes and the original values (1, 0.9 and 0.75 for taxonomist, expert and amateur, respectively; see above and Supporting information). We evaluated five different types of temporal decay curves (Supporting information). These curves were set in a way to allow a faster or slower velocity decay in the initial years. Finally, considering that the effects of either climatic or spatial distance can have different importance in determining both species occurrence and biogeographical knowledge, we set up seven combinations of exponential weights for their matrices when constructing the spatio-environmental matrix used in biogeographical ignorance interpolation (Supporting information). For these sensitivity analyses, only one parameter was changed at a time with the others maintained at their original values. Density curves and coefficients of variation were calculated among MoBIs of each species to assess the variations generated by the parameter configuration.

Joint mapping of data-driven distributional uncertainty and SDM results

To identify areas where models for each species are reliable (or not) due to the knowledge (and ignorance) derived from distributional data and the quality of these data, we visually combined the maps representing ensemble suitability predictions with the MoBIs. Suitability predictions and biogeographical ignorance values were divided into deciles and then different colours were assigned to different combinations of deciles from both values; i.e. each colour represents a decile shift of each variable. This generates bivariate maps of uncertainty, showing areas of high and low suitability together with the uncertainty associated to their estimates (Fig. 2). Additionally, we generated binary uncertainty maps choosing thresholds of 0.7 and 0.6 of biogeographical ignorance and two thresholds for suitability: according to prevalence in the training data (in this case 0.5, classifying all areas with predicted values greater than 0.5 as presence and all areas below 0.5 as absences), and using the lowest presence threshold (LPT) in which the minimum predicted value of suitability for training presence points is considered the threshold to convert continuous predictions in binary ones. We used the suitability values predicted by the ensemble forecast to find the LTP for each species ensemble. We used these thresholds simply to illustrate how uncertainty maps change with different threshold choices. The R-script to generate bivariate maps of prediction uncertainty is provided in the Supporting information.

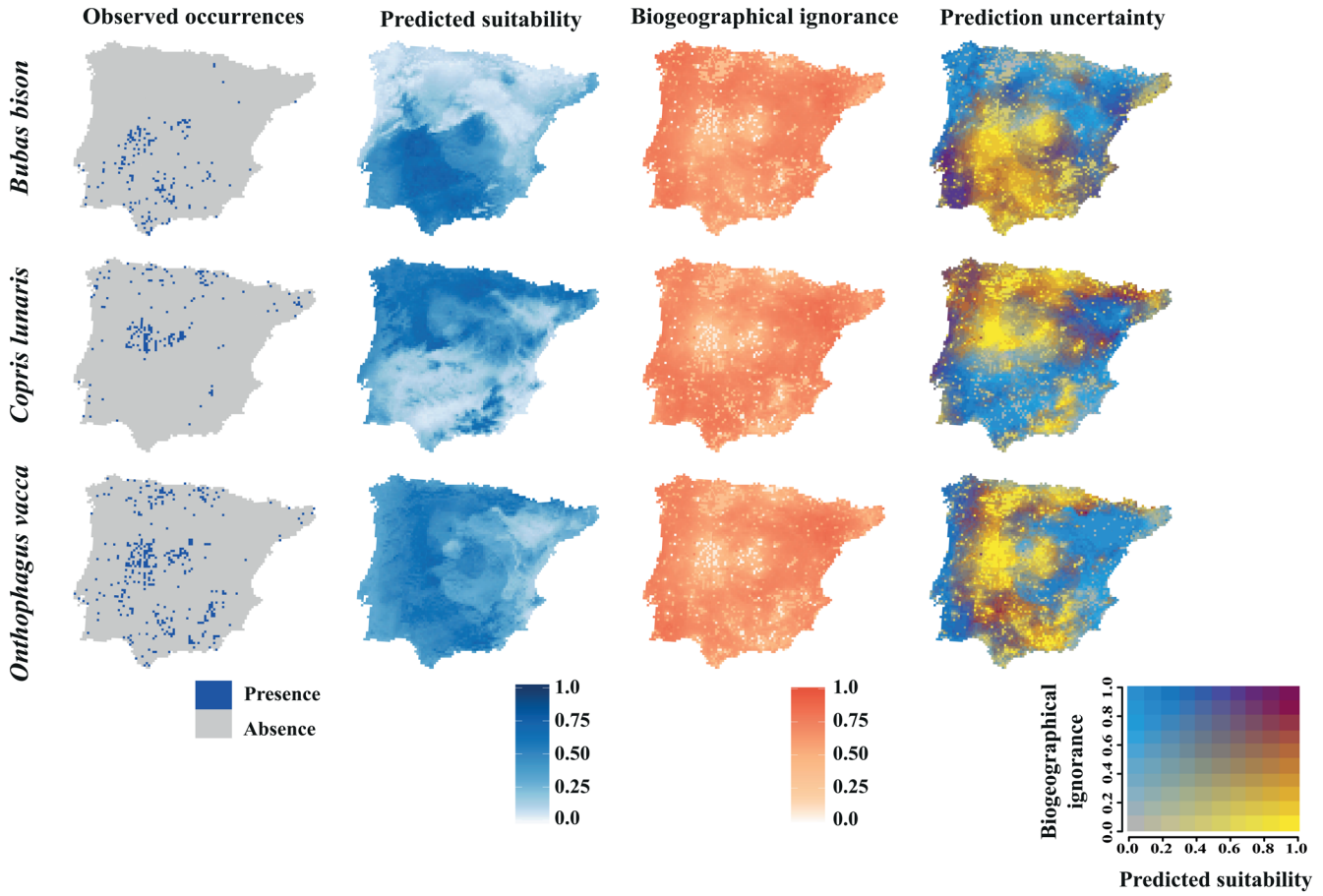

Figure 2. Maps on the known and predicted distribution of three dung beetle species (*Bubas bison*, *Copris hispanus* and *Onthophagus vacca*) in the Iberian Peninsula, and their associated biogeographical ignorance. Maps depict the observed occurrences from BANDASCA database, the overall suitability values predicted by the ensemble SDM technique and the corresponding Maps of Biogeographical Ignorance and the composite maps depicting both SDM projections and prediction uncertainty.

Results

For each species we generated 51 SDM predictions (50 single models based on different training and validation datasets and one ensemble including all these model projections). The geographic patterns of biogeographical ignorance were very similar for all species (Supporting information). Therefore, for simplicity, we only show in the main text the results of the three species that present the more discrepant SDM predictions according to the consensus method: *Bubas bison*, *Copris lunaris* and *Onthophagus vacca* (Fig. 2). The results for the other species and the values of the validation of SDM predictions are provided in the Supporting information. In general, MoBIs indicate that western (mostly Portugal) and central-eastern Iberia are the regions with higher ignorance (i.e. the most uncertain) about the distribution of Scarabaeidae species in the Iberian Peninsula (Fig. 2). Strikingly, about 30% of the study area presented biogeographical ignorance values larger than 0.7, and more than 60% presented values higher than 0.6 (Supporting information).

Despite the similarities in the MoBIs for different species (Fig. 2), their derived uncertainty maps show different spatial patterns for each species due to differences in their predicted distributions. In these maps, four areas related to the combination of extreme (high or low) values of suitability and ignorance are highlighted (Fig. 2):

- Areas with both high suitability and high ignorance values (purple to red cells), i.e. areas where SDM projections suggest that the species is present, but as data-driven uncertainty is high, no or little support exists for these predictions.
- Areas of high suitability and low ignorance (yellow cells), reflecting areas where the prediction that the species is likely present is accompanied by a high degree of certainty.
- Areas with both low suitability and low ignorance values (grey cells), which have the highest likelihood of having been accurately assigned as absences.
- Areas of low suitability and high ignorance (blue cells), where the low suitability (i.e. predicted absence) has little support based on the limited information provided by the data.

Binary uncertainty maps follow the same patterns as continuous uncertainty maps. However, the amount of each area

may vary according to the selected suitability and ignorance thresholds (Supporting information).

The sensitivity analyses conducted on the construction of the spatio-environmental matrix and the selection of the temporal decay parameters generated the largest variation in the general distribution of biogeographical ignorance values (as assessed by density curves), while the values of the coefficient of variation (CV) for these parameters ranged between 0–0.39 and 0–0.92, respectively (Supporting information). Here, the higher the weight given to the spatial distance matrix, the higher were the biogeographical ignorance values; and as slower is the temporal decay, the lower are the so-generated biogeographical ignorance values. The general distribution of biogeographical ignorance values was little affected by changes in taxonomic quality values, which had the lowest CV range (0–0.27; Supporting information). Completeness variation also had little influence on the distributions of biogeographical ignorance values (Supporting information), despite showing the higher range of CV (0–1.72; Supporting information); this is due to a few cells being strongly affected by changes in the calculation of completeness while most cells were little affected, thus generating high CV values but with little effect on the final MoBIs.

Discussion

We successfully created Maps of Biogeographical Ignorance (MoBIs) for species occurrences considering the three key dimensions affecting quality of biodiversity data (taxonomy, space and time; Meyer et al. 2016, Kissling et al. 2018). These maps depict data-driven uncertainty and are designed to be used in parallel with SDMs, allowing researchers and conservationists to simply and intuitively evaluate the quality of underlying distributional data.

The quality of input data is one of the major sources of uncertainty for all kinds of models. However, the impact of data-driven uncertainty may be particularly high for SDMs compared to other types of models, because they can be (and often are) fitted with relatively little information about a species' presence, and sparse (or no) information about their absence (Lobo et al. 2010, Beale and Lennon 2012). Our results suggest that the distributions of even relatively well-surveyed groups (Scarabaeidae) and regions (Iberian Peninsula) (Martín-Piera 2000, Lobo et al. 2007) are often weakly supported by data, as about 30% of the study area had values of biogeographical ignorance (and thus distributional uncertainty) above 0.7.

MoBIs partially reflect bias in data collection, as the regions with greater degrees of biogeographical ignorance in the Iberian Peninsula are those located further away from large urban areas, research institutes and universities, a bias that has previously been shown to occur to this taxon in the study area (Lobo et al. 2007). This dependence of data collection and the use of information about all the other species from the same group to calculate biogeographical ignorance values are responsible for the similar patterns in MoBIs of different species. However, although the distribution of data-driven uncertainty is largely dependent on the spatial allocation of former surveys, the inclusion of the decay of information with increasing environmental distance in MoBIs partly disrupts the survey dependence. Further, including this process makes them more appropriate to describe the uncertainty associated with SDMs, which are calibrated using the same environmental variables.

The parameter values used in the spatial and temporal decay curves had a weak influence on the final maps of ignorance. This could be a result of both the database and the selection of the records used for model calibration. BANDASCA is a well-curated and updated database, so the quality of its records is relatively homogeneous, which could minimize the effect of changing parameters for the generation of MoBIs. Data coming from more heterogeneous data sources, such as GBIF, may be more sensitive to differences in parameter values, as they will result in a higher variability in biogeographical ignorance values (and the uncertainty maps derived from them). Furthermore, the relevance of these parameters is partly associated with the potential for adapting them to the particularities of the area, species and/or questions studied. For example, the temporal decay curves can be set to have a higher decay for more mobile species, as high immigration can cause rapid loss of information over time. Similarly, estimates of temporal decay rates can be set higher for areas with higher rates of habitat conversion, as the probability that occurrence data is obsolete due to local extinction and immigration events is increased in these areas. However, it should be noted that the temporal changes in recording bias described by Lobo et al. (2007) for this dataset seemingly had little effect on data-driven uncertainty. The utility of this particular aspect of biogeographic ignorance may therefore be limited in the absence of rapid landscape changes, as the selection of data from a well-defined temporal window to match climate data will already minimize the uncertainty coming from the temporal decay in information value (Tessarolo et al. 2017).

To construct the first MoBIs we gave the same weight to both the decay of information along environmental and spatial distance. That is, we assumed that environment and space had equal importance in determining the degree (and quality) of the information provided by progressively more distant cells. However, this is just an assumption; the characteristics of the group and study region may change this weighting and, by extension, differentially influence the ability of SDMs to extrapolate species distributions outside of the universe provided by the data. In this case, relative weights for both matrices could be set accordingly. The spatial component may have greater importance for information quality of data for groups in which either dispersal plays an important role in determining the observed distribution, or whose distributions have not yet reached an equilibrium with the environment. In these cases, the spatial matrix should have higher weight. Conversely, for groups where most species ranges are in equilibrium with the environment, climatic distance may have a greater impact and therefore receive a

higher weight. In the same way, the decay in taxonomic quality could be adjusted for taxa or species that are easily identifiable and have more stable taxonomies.

Our analyses should not be viewed as a definitive solution for mapping the uncertainty associated with distributional data. Rather, we have outlined a preliminary approach that may need major technical and practical improvements before its wide-scale implementation. Specifically, as a first demonstration of the method, we did not provide strict validation and more research will clearly be required. For example, studies using datasets obtained from exhaustive surveys and containing presence and absence data could be used to generate MoBIS based on simulations of different subsets of presence points, which can be compared with biogeographical ignorance derived from the entire dataset. This procedure would allow evaluation of the robustness of the method under different scenarios of ignorance by comparing congruence between areas of high ignorance (in the simulated MoBIs) with areas of low sampling effort and/or older records. That said, while the use of presence and absence data would facilitate estimating ignorance maps, the availability of absences is not a prerequisite for delimiting biogeographical ignorance because absences are rarely estimated and ignorance maps should be considered whatever the quality of the data.

MoBIs for single species could be improved by the inclusion of more specific data about the different components of the uncertainty in distributional data. For example, with regards to taxonomic quality the rate of decay could be based on observed rates of synonymization and the numbers of revisions published after the individuals of each particular record were identified and/or revised for the last time (Tessarolo et al. 2017). Further, the values of similarity used to characterize the environmental and spatial decay in the information can be rescaled for more complex curves using parameters reflecting the size of the species' distributional range, dispersal ability or the observed distance decay of similarity (Nekola and White 1999). The addition of species-specific data and/or models accounting for conspicuousness and/or detectability and how distinguishable the species is from similar species (Guillera-Arroita 2017) can also improve estimates of ignorance. This will, in turn, increase the discrepancy between the MoBIs of different species, making them more appropriate for use in assessments about the reliability of SDMs or other geographic range maps such as those developed by IUCN assessments.

Information visualization is key for exploring and communicating data and, if done well, has the potential to identify unrecognized processes (Elith et al. 2002, Griethe and Schumann 2005, McInerny et al. 2014). Visualization is also crucial to fully engage both the general public and the policy makers, facilitating their exploration and understanding of scientific results and model projections (Barry and Elith 2006, Rocchini et al. 2019). MoBIs may be particularly useful in revealing errors by identifying areas with both high uncertainties and covariates that could be possibly linked with these uncertainties. We used maps to communicate information about the uncertainty coming from the distributional data used to calibrate SDMs, with the objective of demonstrating how MoBIs can be a relevant source of information for modellers and practitioners. Our maps provide information about areas where SDMs may fail, and can identify regions where improvements in data quality and coverage could significantly improve results. Significantly, the clear visual representation of uncertainty in the form of maps is a powerful way to make end users aware of the degree of uncertainty associated with the predictions of their SDMs, allowing a more nuanced interpretation of their results.

It is probably impossible to address all sources of uncertainty in SDMs with a single analysis (Elith et al. 2002). Thus, data quality assessments should focus on one or a few sources. Although our proposed method considers only some of the several important sources of uncertainty affecting SDM results, our approach is flexible and can be easily adapted and applied to other sources of ignorance. Information about uncertainty can help in assessing SDM reliability but can also help improve the models themselves. The development of new methods that quantify the uncertainty in SDMs necessitates the concomitant development of methods to handle this uncertainty. It follows that a promising area of research would be the improvement and development of methods that account for variations in uncertainty as a spatially explicit error term during the modelling process, so that, model predictions are corrected for variations in uncertainty over space (McInerny and Purves 2011, Wenger et al. 2013, Stoklosa et al. 2015, Stolar and Nielsen 2015, Niamir et al. 2019). We would strongly encourage the development and improvement of such methods for the different sources of uncertainty affecting SDMs.

More generally, we strongly advocate the routine use of MoBIs as an indispensable part of any assessment and interpretation of SDM results. The joint visualization of SDM predictions and their associated uncertainty (such as in Fig. 2) generates profound insights into the reliability of these commonly used predictive maps. Here, we agree with Elith et al. (2002) in their interpretation of Draper (1995): '… it is difficult to deal with uncertainty, but in the long run it is better to address and treat the uncertainty in an attempt to include the true values within a model's predictions than to ignore uncertainty and miss the true values completely'. The more honest we are about the limitations of our descriptions of nature, the more effective will be the conservation actions we base on them.

More generally, we strongly advocate the routine use of MoBIs as an indispensable part of any assessment and interpretation of SDM results. The joint visualization of SDM predictions and their associated uncertainty (such as in Fig. 2) generates profound insights into the reliability of these commonly used predictive maps. Here, we agree with Elith et al. (2002) in their interpretation of Draper (1995): '… it is difficult to deal with uncertainty, but in the long run it is better to address and treat the uncertainty in an attempt to include the true values within a model's predictions than to ignore uncertainty and miss the true values completely'. The more honest we are about the limitations of our descriptions of nature, the more effective will be the conservation actions we base on them.

Acknowledgements – We thank Christine N. Meynard and anonymous reviewers for their useful comments on earlier versions of this manuscript. We also would like to thank Sara Varela who led us to the "Atlas of Ignorance" paper of Sam Boggs, which inspired the development of this paper.
Funding – This work has been funded by the Brazilian CNPq PVE grant 314523/2014-6 and the Brazilian National Inst. for Science and Technology in Ecology, Evolution and Biodiversity Conservation (INCT-EECBio), supported by MCTIC/CNPq (465610/2014-5) and the Fundação de Amparo à Pesquisa do Estado de Goiás (201810267000023). GT was supported by CAPES REUNI doctorate fellowship and PDSE grant no. 11842121. RJL is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 854248.

Author contributions

Geiziane Tessarolo: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Methodology (equal); Project administration (lead); Writing – original draft (lead); Writing – review and editing (equal). **Richard J. Ladle**: Conceptualization (equal); Writing – original draft (equal); Writing – review and editing (equal). **Jorge M. Lobo**: Conceptualization (equal); Data curation (lead); Methodology (equal); Writing – review and editing (equal). **Thiago Rangel**: Conceptualization (equal); Supervision (equal); Writing – review and editing (equal). **Joaquín Hortal**: Conceptualization (lead); Formal analysis (equal); Investigation (equal); Methodology (lead); Supervision (lead); Writing – original draft (equal); Writing – review and editing (lead).

Transparent Peer Review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.05793>.

Data availability statement

All data used for this work comes from open access sources, that can be accessed online as follows: species distribution data comes from BANDASCA, a GBIF dataset (available at <www.gbif.org/dataset/7b8f1304-f762-11e1-a439-00145eb45e9a>), and environmental data comes from WorldClim and UNEP/GRID databases (available at <www.worldclim.org/> and <https://geodata.grid.unep.ch/>, respectively). The R script for calculation of values of biogeographical ignorance is available from Zenodo (<https://doi.org/10.5281/zenodo.5555453>).

References

Albert, C. H. et al. 2010. Sampling in ecology and evolution – bridging the gap between theory and practice. – Ecography 33: 1028–1037.
Aranda, S. C. and Lobo, J. M. 2010. How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. – Ecography 34: 31–38.
Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.
Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.
Araújo, M. B. et al. 2019. Standards for distribution models in biodiversity assessments. – Sci. Adv. 5: eaat4858.
Arias-González, J. E. et al. 2012. Predicting spatially explicit coral reef fish abundance, richness and Shannon–Weaver index from habitat characteristics. – Biodivers. Conserv. 21: 115–130.
Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. – Ecol. Model. 200: 1–19.
Austin, M. P. and Van Niel, K. P. 2011. Improving species distribution models for climate change studies: variable selection and scale. – J. Biogeogr. 38: 1–8.
Barbet-Massin, M. et al. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.
Barry, S. and Elith, J. 2006. Error and uncertainty in habitat models. – J. Appl. Ecol. 43: 413–423.
Baselga, A. and Araújo, M. B. 2009. Individualistic vs community modelling of species distributions under climate change. – Ecography 32: 55–65.
Baselga, A. et al. 2010. Assessing alpha and beta taxonomy in eupelmid wasps: determinants of the probability of describing good species and synonyms. – J. Zool. Syst. Evol. Res. 48: 40–49.
Beale, C. M. and Lennon, J. J. 2012. Incorporating uncertainty in predictive species distribution modelling. – Phil. Trans. R. Soc. B 367: 247–258.
Bhatt, S. et al. 2013. The global distribution and burden of dengue. – Nature 496: 504–507.
Bini, L. M. et al. 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. – Divers. Distrib. 12: 475–482.
Blackburn, T. M. 2004. Method in macroecology. – Basic Appl. Ecol. 5: 401–412.
Boggs, W. S. 1949. An atlas of ignorance: a needed stimulus to honest thinking and hard work. – Proc. Am. Phil. Soc. 93: 253–258.
Boitani, L. et al. 2011. What spatial data do we need to develop global mammal conservation strategies? – Phil. Trans. R. Soc. B 366: 2623–2632.
Braunisch, V. and Suchant, R. 2010. Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. – Ecography 33: 826–840.
Breiman, L. 2001. Random forests. – Machine learning 45: 5–32.
Buckley, B. A. 2013. Rapid change in shallow water fish species composition in an historically stable antarctic environment. – Antartic Sci. 25: 676–680.
Buisson, L. et al. 2010. Uncertainty in ensemble forecasting of species distribution. – Global Change Biol. 16: 1145–1157.
Busby, J. R. 1986. A biogeographical analysis of *Notophagus cunninghamii* (Hook.) in south-eastern Australia. – Aust. J. Ecol. 11: 1–7.
Chen, X. et al. 2019. Uncertainty analysis of species distribution models. – PloS One 14: e0214190.
Del Vecchio, S. et al. 2015. Changes in plant species composition of coastal dune habitats over a 20-year period. – AoB Plants 7: 1–10.
Diekmann, M. et al. 2014. Long-term changes in calcareous grassland vegetation in north-western Germany – no decline in spe-

cies richness, but a shift in species composition. – Biol. Conserv. 172: 170–179.
Diniz-Filho, J. A. F. et al. 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. – Ecography 32: 897–906.
Dormann, C. F. et al. 2008. Components of uncertainty in species distribution analysis: a case study of the great grey shrike. – Ecology 89: 3371–3386.
Draper, D. 1995. Assesment and propagation of model uncertainty. – J. R. Stat. Soc. 57: 45–97.
Edwards Jr, T. et al. 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. – Ecol. Model. 199: 132–141.
Egli, L. et al. 2020. Taxonomic error rates affect interpretations of a national-scale ground beetle monitoring program at National Ecological Observatory Network. – Ecosphere 11: e03035.
Elith, J. et al. 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. – Ecol. Model. 157: 313–329.
Fagundes, C. K. et al. 2016. Testing the efficiency of protected areas in the Amazon for conserving freshwater turtles. – Divers. Distrib. 22: 123–135.
Fick, S. E. and Hijmans, R. J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. – Int. J. Climatol. 37: 4302–4315.
Gould, S. F. et al. 2014. A tool for simulating and communicating uncertainty when modelling species distributions under future climates. – Ecol. Evol. 4: 4798–4811.
Graham, C. H. et al. 2007. The influence of spatial errors in species occurrence data used in distribution models. – J. Appl. Ecol. 45: 239–247.
Graham, J. and Kimble, M. 2019. Visualizing uncertainty in habitat suitability models with the hyper-envelope modeling interface, ver. 2. – Ecol. Evol. 9: 251–264.
Grenouillet, G. et al. 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. – Ecography 34: 9–17.
Griethe, H. and Schumann, H. 2005. Visualizing uncertainty for improved decision making. – In: Proceedings of the 4th international conference on business informatics research, Skövde, Sweden. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.7326&rep=rep1&type=pdf>.
Guillera-Arroita, G. 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. – Ecography 40: 281–295.
Guisan, A. et al. 2013. Predicting species distributions for conservation decisions. – Ecol. Lett. 16: 1424–1435.
Hastie, T. and Tibshirani, R. 1987. Generalized additive models: some applications. – J. Am. Stat. Assoc. 82: 371–386
Heikkinen, R. K. et al. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. – Prog. Phys. Geogr. 30: 751–777.
Hirzel, A. and Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. – Ecol. Model. 157: 331–341.
Hortal, J. 2008. Uncertainty and the measurement of terrestrial biodiversity gradients. – J. Biogeogr. 35: 1335–1336.
Hortal, J. and Lobo, J. M. 2005. An ED-based protocol for optimal sampling of biodiversity. – Biodivers. Conserv. 14: 2913–2947.
Hortal, J. et al. 2007. Limitations of biodiversity databases: case study on seed–plant diversity in Tenerife, Canary Islands. – Conserv. Biol. 21: 853–863.
Hortal, J. et al. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. – Oikos 117: 847–858.
Hortal, J. et al. 2012. Basic questions in biogeography and the (lack of) simplicity of species distributions: putting species distribution models in the right place. – Nat. Conserv. 10: 108–118.
Hortal, J. et al. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. – Annu. Rev. Ecol. Evol. Syst. 46: 523–549.
Jiménez-Valverde, A. and Lobo, J. 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. – Acta Oecol. 31: 361–369.
Jiménez-Valverde, A. et al. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Commun. Ecol. 10: 196–205.
Kamino, L. H. Y. et al. 2012. Challenges and perspectives for species distribution modelling in the neotropics. – Biol. Lett. 8: 324–326.
Kissling, W. D. et al. 2018. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. – Biol. Rev. 93: 600–625.
Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. – Divers. Distrib. 19: 1366–1379.
Ladle, R. J. and Hortal, J. 2013. Mapping species distributions: living with uncertainty. – Front. Biogeogr. 5: 8–9.
Lobo, J. M. 2016. The use of occurrence data to predict the effects of climate change on insects. – Curr. Opin. Insect Sci. 17: 62–68.
Lobo, J. M. and Martín-Piera, F. 1991. La creación de un banco de datos zoológico sobre los Scarabaeidae Ibero-Baleares. – Elytron 5: 31–37.
Lobo, J. M. and Tognelli, M. F. 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. – J. Nat. Conserv. 19: 1–7.
Lobo, J. M. et al. 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? – Divers. Distrib. 13: 772–780.
Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.
Lomolino, M. V. 2004. Conservation biogeography. – In: Lomolino, M. V. and Heaney, L. R. (eds), Frontiers of biogeography: new directions in the geography of nature. Sinauer Associates, Inc., Sunderland, Massachussets, pp. 293–296.
Mackay-Smith, T. H. and Roberts, D. L. 2019. Accuracy in the identification of orchids of the genus *Angraecum* by taxonomists and non-taxonomists. – Kew Bull. 74: 27.
Martín-Piera, F. 2000. Familia Scarabaeidae. – In: Martín-Piera, F. and López-Colón, J. I. (eds), Coleoptera, Scarabaeoidea I. Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, pp. 207–432.
McCullagh, P. and Nelder, J. A. 1989. Generalized linear models, 2nd edn. – Chapman and Hall.
McInerny, G. J. and Purves, D. W. 2011. Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. – Methods Ecol. Evol. 2: 248–257.

McInerny, G. J. et al. 2014. Information visualisation for science and policy: engaging users and avoiding bias. – Trends Ecol. Evol. 29: 148–157.
Meyer, C. et al. 2015. Global priorities for an effective information basis of biodiversity distributions. – Nat. Commun. 6: 8221.
Meyer, C. et al. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. – Ecol. Lett. 19: 992–1006.
Moudrý, V. and Šímová, P. 2012. Influence of positional accuracy, sample size and scale on modelling species distributions: a review. – Int. J. Geogr. Inform. Sci. 26: 2083–2095.
Naimi, B. and Araújo, M. B. 2016. sdm: a reproducible and extensible R platform for species distribution modelling. – Ecography 39: 368–375.
Naimi, B. et al. 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. – J. Biogeogr. 38: 1497–1509.
Naimi, B. et al. 2014. Where is positional uncertainty a problem for species distribution modelling? – Ecography 37: 191–203.
Nekola, J. C. and White, P. S. 1999. The distance decay of similarity in biogeography and ecology. – J. Biogeogr. 26: 867–878.
Niamir, A. et al. 2019. Incorporating knowledge uncertainty into species distribution modelling. – Biodivers. Conserv. 28: 571–588.
Peterson, A. T. 2006. Uses and requirements of ecological niche models and related dsitributional models. – Biodivers. Inform. 3: 59–72.
Peterson, A. T. and Soberón, J. 2012. Species distribution modeling and ecological niche modeling: getting the concepts right. – Nat. Conserv. 10: 102–107.
Phillips, S. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Pöysä, H. et al. 2019. Changes in species richness and composition of boreal waterbird communities: a comparison between two time periods 25 years apart. – Sci. Rep. 9: 1725.
Rangel, T. F. and Loyola, R. D. 2012. Labeling ecological niche models. – Nat. Conserv. 10: 119–126.
Richards, C. L. et al. 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. – J. Biogeogr. 34: 1833–1845.
Rocchini, D. et al. 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. – Prog. Phys. Geogr. 35: 211–226.
Rocchini, D. et al. 2019. Cartogramming uncertainty in species distribution models: a Bayesian approach. – Ecol. Complexity 38: 146–155.
Ronquillo, D. et al. 2020. Assessing spatial and temporal biases and gaps in the publicly available distributional information of Iberian mosses. – Biodivers. Data J. 8: e53474.
Shea, C. P. et al. 2011. Misidentification of freshwater mussel species (Bivalvia:Unionidae): contributing factors, management implications and potential solutions. – J. N. Am. Benthol. Soc. 30: 446–458.
Silva, D. P. et al. 2016. Contextualized niche shifts upon independent invasions by the dung beetle Onthophagus taurus. – Biol. Inv. 18: 3137–3148.
Sobral-Souza, T. et al. 2021. Knowledge gaps hamper understanding the relationship between fragmentation and biodiversity loss: the case of Atlantic Forest fruit-feeding butterflies. – PeerJ 9: e11673.
Soulé, M. E. 1985. What is conservation biology? – BioScience 35: 727–734.
Stoklosa, J. et al. 2015. A climate of uncertainty: accounting for error in climate variables for species distribution models. – Methods Ecol. Evol. 6: 412–423.
Stolar, J. and Nielsen, S. E. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. – Divers. Distrib. 21: 595–608.
Stropp, J. et al. 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa. – Global Ecol. Biogeogr. 25: 1085–1096.
Swanson, A. K. et al. 2013. Spatial regression methods capture prediction uncertainty in species distribution model projections through time. – Global Ecol. Biogeogr. 22: 242–251.
Syphard, A. D. and Franklin, J. 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. – Ecography 32: 907–918.
Tessarolo, G. 2012. Choosing the right path for species distribution modeling. – Front. Biogeogr. 4: 91–92.
Tessarolo, G. et al. 2014. Uncertainty associated with survey design in species distribution models. – Divers. Distrib. 20: 1258–1269.
Tessarolo, G. et al. 2017. Temporal degradation of data limits biodiversity research. – Ecol. Evol. 7: 6863–6870.
Tessarolo, G. et al. 2021. High uncertainty in the effects of data characteristics on the performance of species distribution models. – Ecol. Indic. 121: 107147.
Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit Region. – Econ. Geogr. 46: 234–240.
Van der Sande, M. T. et al. 2016. Old-growth Neotropical forests are shifting in species and trait composition. – Ecol. Monogr. 86: 228–243.
Watling, J. I. et al. 2015. Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. – Ecol. Model. 309–310: 48–59.
Wenger, S. J. et al. 2013. Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. – Global Change Biol. 19: 3343–3354.
Williams, J. N. et al. 2009. Using species distribution models to predict new occurrences for rare plants. – Divers. Distrib. 15: 565–576.
Wilson, K. A. 2010. Dealing with data uncertainty in conservation planning. – Nat. Conserv. 8: 145–150.
Wisz, M. S. et al. 2008. Effects of sample size on the performance of species distribution models. – Divers. Distrib. 14: 763–773.
Yang, W. et al. 2013. Geographical sampling bias in a large distributional database and its effects on species richness-environment models. – J. Biogeogr. 40: 1415–1426.
Yoccoz, N. G. et al. 2001. Monitoring of biological diversity in space and time. – Trends Ecol. Evol. 16: 446–453.