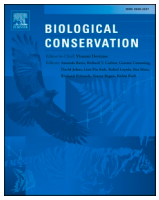
Biological Conservation 284 (2023) 110207

Contents lists available at ScienceDirect

Biological Conservation

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



Effects of occurrence data density on conservation prioritization strategies

Marlon E. Cobos [a,*], Claudia Nuñez-Penichet [a], Peter D. Campbell [a], Jacob C. Cooper [a,b], Fernando Machado-Stredel [a], Narayani Barve [c], Uzma Ashraf [d,e], Abdelghafar A. Alkishe [a], Eric Ng'eno [a], Rahul Raveendran Nair [a], P. Joser Atauchi [f,g], Adeola Adeboje [a], A. Townsend Peterson [a]

[a] Department of Ecology and Evolutionary Biology & Biodiversity Institute, University of Kansas, Lawrence, KS 66045, USA
[b] Department of Biology, University of Nebraska at Kearney, Kearney, NE 68849, USA
[c] Florida Museum of Natural History, University of Florida, Gainesville, FL, USA
[d] Department of Land, Air and Water Resources, University of California, Davis, USA
[e] Wild Energy Initiative, Institute of the Environment, University of California, Davis, USA
[f] Museo de Historia Natural Cusco (MHNC), Universidad Nacional de San Antonio Abad del Cusco. Paraninfo s/n, Cusco, Peru
[g] Instituto para la Conservación de Especies Amenazadas, Cusco, Peru

ARTICLE INFO

Keywords:
Occurrence data
Data density
Ecological niche model
Birds
Species distribution models
Zonation

ABSTRACT

Place-prioritization analyses are a means by which researchers can translate information on the geographic distributions of species into quantitative prioritizations of areas for biodiversity conservation action. Although several robust algorithms are now available to support this sort of analysis, their vulnerability to biases deriving from incomplete and imbalanced distributional information is not well understood. In this contribution, we took a well-sampled group (i.e., Icteridae or New World blackbirds) in an intensively sampled region (the contiguous continental United States), and developed a set of pseudo-experimental manipulations of occurrence data density—in effect, we created situations in which data density was reduced 10- or 100-fold, and situations in which data density varied 100-fold from region to region. The effects were marked: priority areas for conservation shifted, appeared, and disappeared as a function of our manipulations. That is, differences in density of data can affect the position and complexity of areas of high conservation priority that are identified using distributional areas of species derived from ecological niche modeling. The effects of data density on prioritizations become more diffuse when considerations of existing protected areas and costs related to human intervention are taken into account, but changes are still manifested. Appropriate considerations of sampling density when constructing ecological niche models to identify distributional areas of species are key to preventing artifactual biases from entering into and affecting results of analyses of conservation priority.

1. Introduction

An important paradigm in biodiversity conservation is that of quantitative prioritizations of geographic regions for biodiversity conservation efforts (Brooks et al., 2006). In these initiatives, distributional information for a group of species of conservation concern is fed in—often in the form of raster-format maps that represent outputs of ecological niche modeling or species distribution modeling efforts—to algorithms that optimize area selection to produce solutions that are maximally efficient in protecting distributional areas of species (Nori et al., 2016; Zhang et al., 2012). These analyses have now been enabled by development of diverse analytical platforms (Moilanen et al., 2005; Watts et al., 2009), and also allow incorporation of additional details, such as existing protected areas networks or human-perturbed areas that are not fertile options for conservation action (Justus and Sarkar, 2002; Moilanen et al., 2009). Place-prioritization analyses have been implemented for many taxa in many parts of the world, such that they have provided valuable insights into geographic priorities for conservation effort (Gardner et al., 2018; Nori et al., 2016; Velazco et al., 2022).

These place-prioritization efforts, however, are developed against a backdrop of massive imbalances and contrasts in primary biodiversity data availability (Peterson and Soberón, 2018). Regions such as North America, Australia, South Africa, and Europe enjoy massive densities of biodiversity data, whereas other regions have data resources that are

* Corresponding author at: Biodiversity Institute, University of Kansas, 1345 Jayhawk Blvd., Lawrence, Kansas 66045, USA.
E-mail address: manubio13@gmail.com (M.E. Cobos).

https://doi.org/10.1016/j.biocon.2023.110207
Received 2 January 2023; Received in revised form 13 July 2023; Accepted 18 July 2023
Available online 23 July 2023

considerably more sparse. These data disparities exist in some cases because such information is lacking (Daru et al., 2018; Funk et al., 2005), in other cases because data exist but in formats that are not readily readable or usable (Peterson et al., 2018), and in still other cases because sociological or political will prevents data sharing (Huang et al., 2012; Scoble, 2000). A relevant area of inquiry is therefore whether these regional or national imbalances in biodiversity data density affect the outcomes of place-prioritization analyses. More generally, we are, in effect, assessing the effects of the Wallacean knowledge shortfall (Beck et al., 2013; Bini et al., 2006) on the outcomes of place-prioritization analyses.

This contribution aims to test the proposition that geographic patterns of occurrence data density, and biases in that density that interact with Wallacean knowledge shortfalls, have consistent and predictable effects that propagate through the conservation prioritization process. That is, regions or species with low data densities will manifest simpler and less detailed potential distributional areas reconstructed using ecological niche modeling or species distribution modeling. These simpler model summary outputs in turn propagate through the conservation prioritization process, yielding prioritization schemes that differ markedly as a function of data density. We discuss the implications of these phenomena for development of effective biodiversity conservation strategies on regional, continental, and global scales (e.g., Nori et al., 2020; Zeller et al., 2013).

2. Methods

This paper consists of a lengthy and complicated sequence of analyses. To maximize their replicability, we have executed all steps on the R platform (R Core Team, 2022). We have provided the data and R scripts used in a Figshare repository (accessible at doi:https://doi.org/10.6084/m9.figshare.21787226), in the hopes that others will be able to take maximum advantage of the methods that we have followed in developing these analyses.

2.1. Data and data manipulations

2.1.1. Occurrence data

We used New World blackbirds (Icteridae; hereafter referred to as "blackbirds") across the contiguous continental ("Lower 48") United States as a hypothetical example of a lineage with reasonable species diversity that might be the subject of a conservation prioritization effort. Our goal in this paper is not to assemble strategies for blackbird conservation, but rather to test and replicate the effects of different data densities on such prioritization efforts. As such, we present analyses of blackbirds simply to illustrate the effects of occurrence data density on such prioritization exercises.

Blackbird species across the Lower 48 United States offer an example of a suite of densely sampled species on which to perform a series of manipulations in which data density is reduced experimentally. Blackbirds are widespread, easily detectable, and (mostly) easily identifiable across the entire study region. Blackbird species also cover a wide range of conservation categories, with taxa in this group ranging from among the most numerous of all North American birds (i.e., *Agelaius phoeniceus*, with an estimated 1.5×10^8 individuals across the United States and Canada) to some of the most imperiled in the United States (e.g., *A. tricolor* is Endangered, with estimates of $<2 \times 10^5$ individuals left in California; Beedy et al., 2020; Meese, 2017; Neff, 1937; Rosenberg et al., 2016). Furthermore, blackbirds are for the most part easily detectable, and the more localized taxa (e.g., *Icterus parisorum*) are highly sought by birdwatchers, such that data densities are high for all species. With this rich data resource for US blackbirds, we could then proceed to manipulate data density across the Lower 48 United States to mimic the data imbalances that are manifested across many borders (e.g., between countries) around the world.

We obtained occurrence records from eBird (Sullivan et al., 2009) for the period 1 January 2002 through 30 June 2022 (eBird Basic Dataset, 2022). We filtered the data to exclude records with no coordinates or that represented exact duplicates of other records; we retained records from the months May through July, to keep a set of records related to breeding periods. To exclude erroneous records (e.g., vagrants), we removed records outside areas considered to be the actual breeding range of each species based on consultation of key literature references (Del Hoyo et al., 2011). Considering that our aim was to test effects of reduced data density in conservation prioritizations, we applied a spatial filter of points per species to keep only one record per pixel (pixel size 4 km, see below), but did not apply any other spatial rarefaction to our data. We performed these analyses using base R functions and the terra package (Hijmans, 2022).

2.1.2. Data manipulations

To assess effects of density of occurrence data on results from exercises to determine areas of priority for conservation, as are very frequent across international borders around the world, we manipulated data based on a suite of random sampling protocols. Three of the treatments were as follows: (1) the complete set of records (i.e., 100 % data density), (2) a 10 % random subsample of available records; and (3) a 1 % random subsample of available records. Note that with the latter, most extreme reduction, it was necessary to remove two species (*Icterus graduacauda* and *I. gularis*), as only single records were available for them at these reduced densities.

Two other treatments explored the effect of differential data density across a region on the pattern of conservation priority areas that is reconstructed. To this end, we mimicked an imaginary national border, with 100-fold more or less data on either side. We used 101.77° W longitude as the position of this border, as it represents a relative trough in data density across North America, running more or less continuously from the Mexican border to the Canadian border (Fig. 1). As such, in one treatment, (4) we used full data density west of this border and 1 % data density to the east, and in the other, (5) we used full data density east of this border and 1 % data density to the west.

As such, we have created a series of illustrative and informative pseudo-experimental manipulations of data densities across the United States. Comparing conservation prioritizations based on treatments (1) and (2) shows the effect of a 90 % reduction in data availability, and comparing treatments (1) and (3) shows the effect of a 99 % reduction in data availability, on prioritizations. Comparing treatments (1) and (3) with treatments (4) and (5) will illustrate the importance of regional imbalances in data density.

2.1.3. Environmental data

As environmental data inputs, we used the following climate variables: precipitation, temperature (mean, minimum, and maximum), and vapor pressure deficit (maximum and minimum), from PRISM (PRISM Climate Group, 2022). We used 30-year normals (1991–2020) to represent climatic conditions across the Lower 48 United States during May–July (spatial resolution 4 km). To reduce dimensionality and multicollinearity, we performed a principal component analysis (PCA) on these variables, and retained the first four principal components (PCs), which together reflected 96.6 % of the total variance in the overall dataset for further analysis (Table S1). We downloaded environmental data using the R package prism (Hart and Bell, 2015), and performed PCA with the package kuenm (Cobos et al., 2019a).

2.2. Geographic distributions of species

For each blackbird species, we created a customized hypothesis of the accessible area (**M** in the BAM diagram; Soberón and Peterson, 2005), which is most appropriate as a delimitation of an area for model calibration (Barve et al., 2011). The BAM framework gives a theoretical basis by which to understand scenarios of how biotic (**B**), abiotic (**A**), and dispersal (**M**, mobility) factors determine species distributions

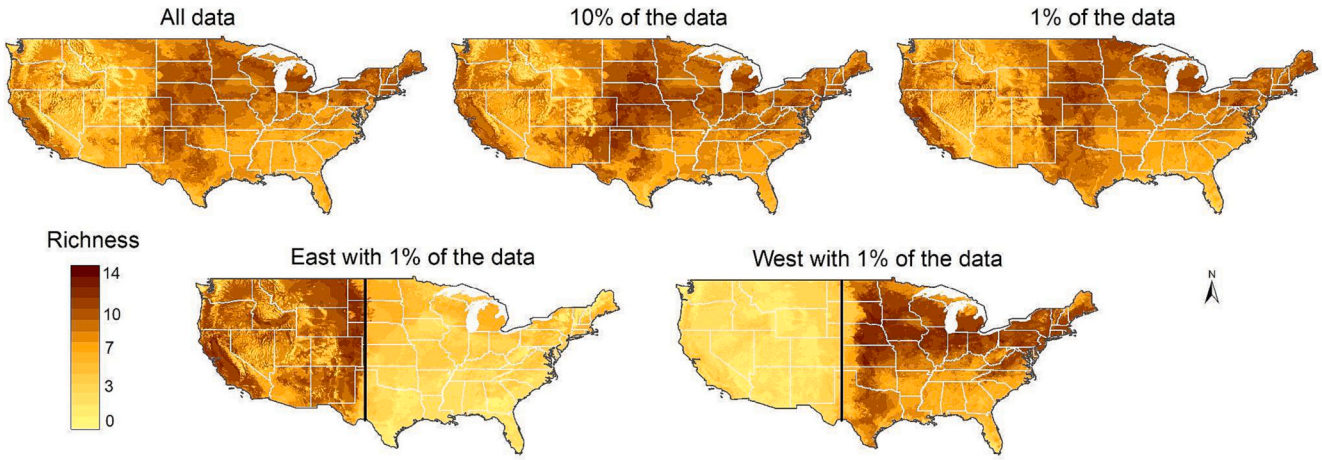


Fig. 1. Summary of geographic patterns of blackbird species richness across the Lower 48 United States under five distinct treatments of the available occurrence data, as propagated through detailed ecological niche modeling steps.

(Soberón and Peterson, 2005). Our **M** hypotheses were based on dispersal simulations using the grinnell R package (Machado-Stredel et al., 2021). We used occurrence records spatially thinned to a minimum point-to-point distance of ~30 km. We also used raw environmental variables (i.e., not the PCs), as required for calculations of suitability using simple ellipsoid models in these simulations. We tested different values for two parameters: number of dispersal events (125 and 250) and dispersal kernel standard deviation (1, 2, 3, 4, and 5 pixels); simulations were run under stable environmental conditions, using a normal dispersal kernel, with four as a maximum number of dispersers, and 5 % as the threshold for accessibility. We selected calibration areas to be used in model selection based on agreement among multiple parameter settings (i.e., relative invariance with respect to parameter values), as well as general correspondence to known biogeographic breaks. Model calibration was performed using only the environmental information inside the **M** area for each species.

For each of the five sets of occurrence data corresponding to the five treatments described above, we used ecological niche modeling with a maximum entropy algorithm (Phillips et al., 2017; Phillips et al., 2006) to derive suitability layers and estimate distributional areas, as follows. We started with model selection processes, in which we tested distinct candidate models created with variations of algorithm parameter combinations and sets of predictor variables. We produced 198 candidate models, representing all possible combinations of 3 feature classes (i.e., lq, lp, lqp; l = linear, q = quadratic, p = product), 6 regularization multiplier values (0.1, 0.3, 0.6, 1, 2, 3), and 11 sets of predictors representing all combinations of two or more of the PCs. Feature classes determine the expected type of response of suitability to each of the predictors (the combinations used aim for simple, unimodal responses), and regularization multiplier controls how adjusted the response is to the observed values of presence (higher values make for more relaxed adjustments). We evaluated models based on an ordered set of three criteria (Cobos et al., 2019a): (1) statistical significance based on partial ROC analyses (Peterson et al., 2008), (2) model predictive performance as reflected in omission rates (using a maximum acceptable omission rate criterion of $E = 5$ %; Anderson et al., 2003), and (3) low model complexity using the Akaike Information Criterion corrected for small sample sizes (AICc; Warren and Seifert, 2011). We therefore selected models that were statistically significant, had omission rates <0.05, and that had AICc values within 2 AICc units of the minimum value among the significant and well-performing models (Cobos et al., 2019b).

For each species, we created final models using all parameter settings and sets of variables selected as described above, with 5 replicates via bootstrap, and clog-log outputs. We calculated a consensus model using the median across all results from the final models for each species. We binarized consensus results via minimum training presence thresholding approaches with a maximum allowable omission rate of E = 5 % (Anderson et al., 2003). To use ecological niche models to approximate the occupied distributional area ($\mathbf{G}_O$) of each species (Loiselle et al., 2003), we retained areas meeting the threshold with a value of 1 within **M**, but set to 0 at all sites outside of **M**, as such sites would not be accessible to the species for dispersal and colonization. Steps of model calibration and projections, as well as development of consensus models, were performed in R using the kuenm package; raster processing to restrict distributional areas was performed using the package terra (Hijmans, 2022).

2.3. Conservation area prioritizations

We used Zonation v4 (Moilanen et al., 2014; Moilanen et al., 2005) to run analyses to prioritize areas as more or less important for conservation efforts, aimed at protecting a set percentage of the range area of each of our species. Analyses were performed using the conservation model "basic core-area," in two ways. (1) We used only the thresholded (binarized) distributional hypothesis for each species, as described above. In the second approach, (2) we considered existing protected areas as a starting point for conservation action ("mask" layer), and existing human-modified areas as areas in which such action is not feasible ("cost" layer), in tandem with the binarized distributional area hypotheses, with details as follows. For existing protected areas, we used all categories of terrestrial protected areas from the layer of world protected areas (UNEP-WCMC and IUCN, 2022) updated as of November 2022 (available at www.protectedplanet.net) as areas already under protection. To summarize areas already beyond hope of protection, we used information from the "human footprint" data layer (available at https://sedac.ciesin.columbia.edu/; see details of data management below; Venter et al., 2016).

We rasterized the layer of protected areas to match the extent and resolution of the raster layers representing distributions of species. Values in this raster layer were one and two representing non-protected and protected regions, respectively. The layer of human footprint was aggregated and masked to match the resolution and extent of all layers to be used in prioritizations. The original values in this later layer ranged 0–50 in terrestrial areas; we replaced values of zero by 1.1×10^{-8}, as Zonation requires the cost layer not to contain values of zero. As our data came from multiple sources, we created a consensus layer to mask all raster inputs to be used in further analyses. This consensus layer grouped cells with no values from all raster layers used, which allowed us to mask out those cells that had information in some layers but not in others. All prioritization analyses were performed for the five data treatments

described above using the package zonator (Lehtomaki, 2018) in R. The zonator package allows users to run Zonation with most of its features from the R interface.

To visualize priority areas identified as an outcome of the different data manipulations, we needed to binarize the prioritization results to a single threshold with a similar "meaning" across different treatments. To that end, we set three priority thresholds to achieve at least 1 %, 5 %, and 10 % of range representation for all of the species in the analysis in each prioritization exercise. We derived these threshold values from the curves representing the relationship of percent range representation versus prioritization value for all species, produced within zonator. Specifically, we determined the prioritization value that yielded ≥1 %, 5 %, and 10 % representation for all species, and used that value as a threshold for creating a binary output for the prioritization. Raster processing to prepare data and process results in this section were done using the terra package in R.

3. Results

Initial occurrence datasets covered 21 blackbird species in the Lower 48 United States, and yielded a grand total of 10,699,387 occurrence records, with per-species totals ranging from 3257 (*I. graduacauda*) to 3,562,404 (*A. phoeniceus*). With data cleaning and reduction steps, these numbers were reduced to a total of 532,906, with per-species totals ranging from 123 (*I. graduacauda*) to 112,130 (*A. phoeniceus*). The pseudo-experimental treatments reduced the per-species numbers in different ways. For instance, the 90 % data reduction ranged from 12 (*I. graduacauda*) to 11,205 (*A. phoeniceus*) records, and the 99 % data reduction ranged from 1 (*I. graduacauda*) to 1120 (*A. phoeniceus*) records. Further, the 99 % data reduction in the eastern half of the country resulted in 1 (*I. graduacauda*) to 27,472 (*A. phoeniceus*) records, and the 99 % data reduction in the western half of the country gave 11 (*A. tricolor*) to 85,696 (*A. phoeniceus*) records. Indeed, for three of the five manipulations, two species (*I. graduacauda* and *I. gularis*) were reduced to single records; these species were therefore excluded from subsequent analyses, to assure comparability among prioritization efforts. As such, our different manipulations of the occurrence data had significant implications for data densities of occurrence data for blackbird species that would be fed into ecological niche models.

Ecological niche modeling steps were conducted independently for each species in the study—i.e., model calibration areas were established via simulations, and optimal parameter values were chosen by detailed, species-specific model selection exercises. Given that the details of these steps are not central to the point of this paper, which focuses on the effects of data density on conservation prioritizations, we have provided the species-specific results of the **M** simulations (Table S2; Figs. S1–S17), model-selection exercises (Tables S3–S7), and model geographic projections (Figs. S18-S22) in the Supplementary Materials.

For the three initial treatments of data density, ecological niche model outputs translated into maps of species richness that showed complex patterns across the country (Fig. 1, top row). At full data density, estimated species richness was highest in central and southern California, across much of the Great Plains, and eastward through the Great Lakes region to New England. Those patterns seem largely unaffected by data density, except that areas of higher species richness shifted somewhat northward at lowest data density. The two treatments in which data density was only reduced in one-half of the country showed a reduction in species richness in the half where data density was reduced, reaching a maximum of only 7 species.

As the occurrence data were passed through the workflow of **M** simulation in grinnell, ecological niche model evaluation and selection in kuenm, and area prioritization in Zonation, patterns changed and differences among treatments became more pronounced. That is, using all data available (Fig. 2, top-left panel), priority regions were in central and southern California, New Mexico, Minnesota, northern Iowa, and along the Gulf Coast into Florida. For the 10 % data-density analyses (90 % data reduction), however, the Gulf Coast and Florida priority regions disappeared, and priority was elevated in southern New England. With 1 % data density, priority regions consisted of odd, north-to-south zigzag patterns across the central and southern United States, clearly reflecting very simple distributional summaries underlying the prioritizations.

Priority patterns were considerably more complex when protected areas and human-modified areas were included in the place-prioritization analyses (Fig. 2, bottom row), as expected given the interaction between those "mask" and "cost" areas and patterns of presence and absence of species. Differences do exist—for instance, priority areas in New Mexico and westernmost Texas in the two higher-data-density prioritizations shift eastward in Texas in the 1 % data density prioritization. Overall, though, the differences are less dramatic—or at least are less clear and obvious—in the analyses that take into account existing protected areas and existing patterns of human dominance of landscapes.

In the treatments in which we created data disparities longitudinally across the Lower 48 United States, we observed many of the same contrasts (Fig. 3). Major observations were that the priority zones

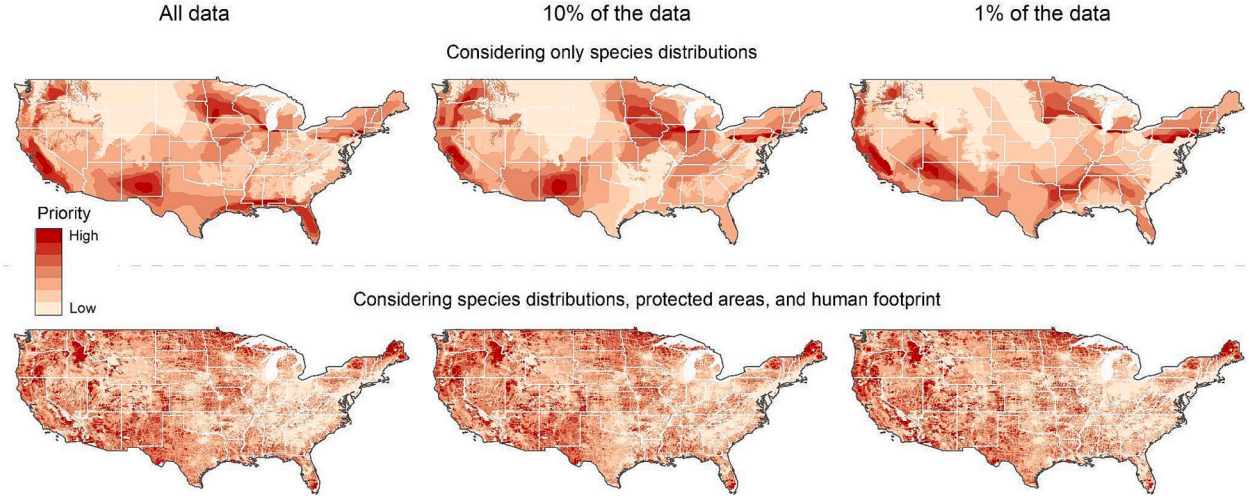


Fig. 2. Conservation prioritization regions as a function of (1) occurrence data density (columns left-to-right represent prioritizations based on all occurrence data, 10 % data density, and 1 % data density), and (2) inclusion of information on existing protected and human-modified areas in the analyses (i.e., top row does not include such information and depends only on species' distributional patterns, bottom row includes that information).

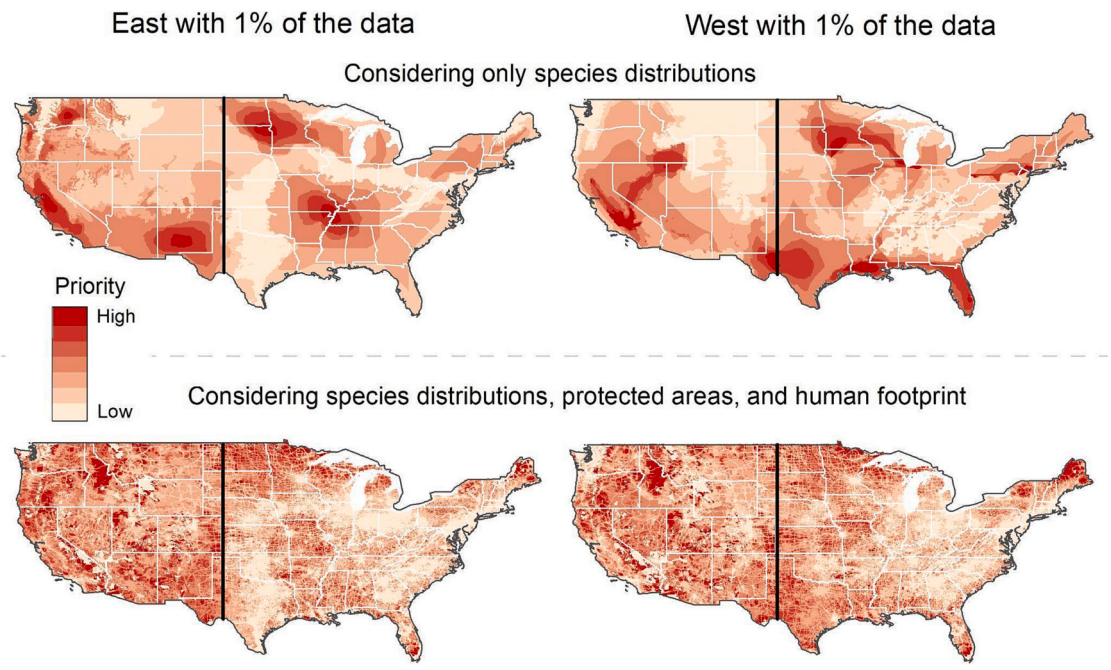


Fig. 3. Conservation prioritization zones based on 1:100 data density imbalances between eastern and western halves of the country. The top row of maps shows prioritizations developed based only on species' distributional information. The second row of maps shows prioritizations developed based on the species' distributional information, plus information on current protected areas and current areas of high human impact.

shifted longitudinally towards the region with greater data density (see, e.g., Texas). Also, the Gulf Coast and Florida priority area disappears in the treatment in which the eastern half of the country is at 1 % data density. Certain instabilities are noteworthy, such as a very "cold" spot in the central Midwest when the east is at full density, turning into a very "hot" spot when the east is at 1 % density. These contrasts are manifested also—albeit possibly less visibly—when protected areas and human-impacted areas are included in the place prioritization exercise (Fig. 3, bottom row).

Areas of greatest instability in the prioritization exercises were distributed nonrandomly (Fig. 4), and more or less in accord with the descriptions of the individual prioritization outcomes above. That is, areas of greatest instability in the manipulations of overall data density were in the central Midwest and Gulf Coast regions of the eastern United States (Fig. 4, top row). Areas of instability in the east-west data density manipulations were similarly in the eastern United States, as well as in

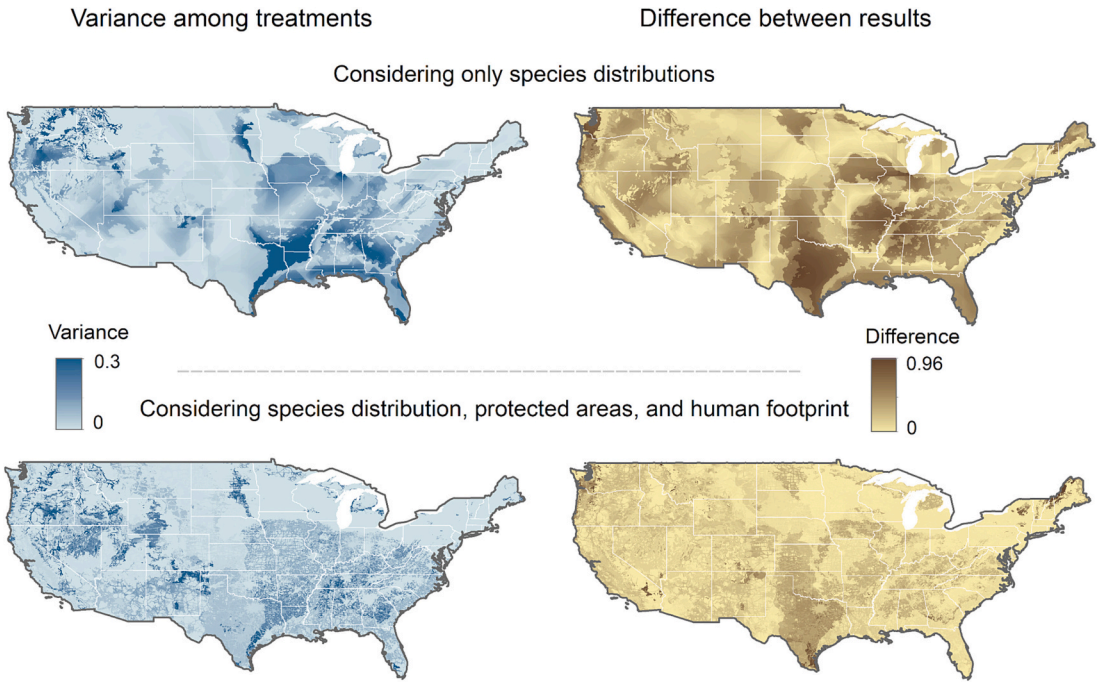


Fig. 4. Variability in prioritizations deriving from different treatments to the data and information used as input to place-prioritization analyses. The left column shows variance calculations across the three data-density treatments (1 %, 10 %, and 100 %), whereas the right column shows the difference in prioritization value between the two manipulations (i.e., east reduced, west reduced) in the regional analyses.

the Pacific Northwest and in the southern Great Plains and Texas, as described above (Fig. 4, bottom row).

We established priority-value cutoffs (thresholds) for each of the 10 prioritizations that were developed based on ≥1 %, 5 %, and 10 % range representation for all species (examples for a 5 % threshold in Figs. S23–S32). Threshold values across the 10 prioritizations ranged from 0.80 to 0.98 (Table S8), and the species that determined the threshold value found varied across our multiple treatments. The resulting maps of high-priority areas (Fig. 5, Fig. S33) illustrate the points above more clearly, showing how the high-priority areas shift among different treatments.

Fig. 5. Areas of priority for conservation according to distinct data used in prioritizations and threshold values to protect different percentages of species ranges. The threshold values used are 1 %, 5 %, and 10 %, and they represent values of priority resulting from zonation exercises that allow protection of at least those given percentages of all species ranges. Results for treatments using all species data, and data reduced to 10 % and 1 % are shown here. Results for treatments reducing data in half of the country are presented in Fig. S33.

4. Discussion

This paper set out very specifically to test effects of data density on outcomes of place-prioritization analyses. That is, instead of implementing analyses on blackbirds or conservation strategies for birds of the United States, we rather focus on the subtle (or not so subtle) biases that manifest in situations of low data density, or in regions of low density even when high-density regions exist nearby, and as such this study assesses phenomena of interest in many regions around the world. This study is intended to guide users of such tools towards effective use of place-prioritization analyses in situations in which data-density effects have the potential to affect the results of their work.

The outcomes of our analyses were as predicted at the outset of the project, but with some added complexity. That is, when ecological niche models are properly calibrated to create the distributional summaries for species in place-prioritization exercises, models based on relatively few occurrence points will often produce more generalized and simpler summaries of ranges of species (see also Muscatello et al., 2021). The simpler nature of these range summaries propagates through the place-prioritization analyses to produce conservation prioritizations that have less detail; indeed, at times, they even take on a geometric and very non-biological aspect (Figs. 2, 5). Of perhaps greater concern is that the differences are not solely in terms of detail, but rather are large-scale differences in which an area of high priority at full data density is ignored at lower data densities—these more qualitative differences are of greatest concern, as they reflect deep instability in place-prioritization analyses with respect to occurrence data density (Fig. 4).

Our analyses, in which we created a 1:100 difference in data density across an imaginary "border" in the middle of the US, also demonstrated deep instabilities and dramatic effects caused by the differences in data density (Fig. 3–4). Such "borders" associated with dramatic differences in biodiversity data density are actually quite common around the world (Hughes et al., 2021). As examples, we have illustrated biodiversity data densities along the border between the United States and Mexico, the border between Western Europe and Eastern Europe and the Former Soviet Union, contrasts between Australia and New Guinea, and contrasts between South Africa and neighboring countries in Fig. 6. As such, the effect that we created with an artificial "border" within the United States is indeed a phenomenon that exists in many places around the world.

Exploring the ways in which these data-density effects are manifested, and reflecting a bit on the species-richness maps (Fig. 1, bottom row), it is clear that models for widespread species are being truncated spatially at the dividing line between low- and high-density regions. This large-scale effect might seem to be something that would get detected in some data- or model-quality assessment step (e.g., explorations of spatial correlation of data; Crase et al., 2012), were this study to be a real-world prioritization effort. In fact, one would expect that it would be noticed immediately, as this study is being developed in the United States with a taxon as well-known as birds, although niche-related analyses for United States vertebrate taxa have made such errors in the past, as in the case of niche-centroid calculations for a rodent species in the southwestern United States [(Dallas et al., 2017); see Soberón et al., 2018]. Nonetheless, for many taxa in many regions of the world, the Wallacean Shortfall (Lomolino, 2004) is significant, such that range-estimate truncations may not be noticed at all, particularly in large-scale prioritization analyses based on hundreds or thousands of species (e.g., Brum et al., 2017; Nori et al., 2020). Although previous results suggest that including more species in prioritizations will increase the stability of spatial prioritizations in this type of analyses (Kujala et al., 2018), biases from using poor characterizations of the distributional potential of species will remain in the data used for such analyses and propagate through the analyses into the final prioritizations.

At first glance, our results would appear to be positive in nature as regards the addition of protected areas and human-influenced areas to the analysis. That is, the effects of our manipulations of data densities

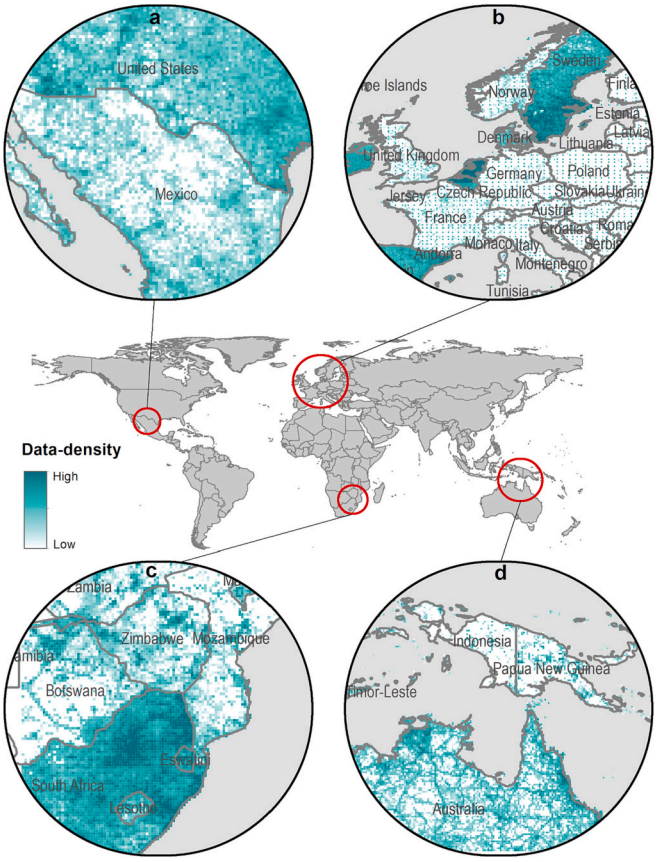


Fig. 6. Example world regions showing variations in data density across "borders," such as those mimicked in our regional data-density manipulations. For these examples, we used all records of birds (class Aves) from the Global Biodiversity Information Facility (GBIF.org, 2022a, 2022b, 2022c, 2022d). a) the United States and Mexico; b) the transition from Western Europe through Eastern Europe to the Former Soviet Union; c) northern South Africa and neighboring parts of Mozambique, Botswana, and Zimbabwe; and d) northern Australia and New Guinea, Indonesia, and off-lying islands. Data densities are visualized as linear, $\log_{10}$-based color ramps from no data (white) to high data density (dark teal). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

were far more pronounced in analyses in which protected areas and human-influenced areas were not included, such that one might conclude that the inclusion of those considerations moderated the effects of uneven data density. Although the inclusion of such considerations is indeed necessary to produce more realistic and effective prioritizations (Kukkala and Moilanen, 2013), we, come to the conclusion that, in our study, the more detailed nature of the results from analyses incorporating existing protected and disturbed areas simply obscures the same effects that are more easily perceived in the more basic analyses. This effect is significant for the outcomes of real-world analyses that are designed to result in conservation action: genuinely concerning effects of low data density may not be visible in such detailed analyses.

The question, then, is what to *do* about such data imbalances. On the simplest level, as biases detected in our prioritizations derive from data density imbalances, researchers using such approaches should therefore inspect their input data carefully via graphical explorations to identify if such problems exist. Data cleaning steps and appropriate applications of spatial thinning could help solve some of the imbalances. Balancing the data via differential sampling according to administrative borders or according to borders that show distinct sampling effort should be considered if problems persist after initial steps of data processing (e.g., Ingenloff et al., 2017; Nuñez-Penichet et al., 2021). Part of our methods

included delimitation of areas for model calibration according to dispersal factors, and model calibration to select algorithm parameterizations that produce good models. These two methodological steps help to create models that are appropriately fitted to the data (i.e., non overfitted) and prevent increasing biases if unnoticeable density imbalances remain. Nonetheless, to the degree that data imbalances are dramatic, it may prove impossible to thin data sufficiently to remove the negative effects of the imbalance.

When data reduction is not an option, and perhaps more positive in general is the step of improving data densities in the under-represented regions. That is, on a project-by-project basis, it is possible to scour the literature and other biodiversity data resources for additional data from the undersampled regions. This step can provide additional occurrence data for poorly represented regions, and as such may alleviate some of the problems of imbalances in data density. Most productively in terms of creating permanent science resources, however, is the much larger task of improving biodiversity data resources for undersampled regions. In this sense, the example of Mexico stands out: a country invested heavily in marshaling, improving, and sharing data about its biodiversity, and transformed itself from undersampled to the status of being a global leader (Soberón, 2022). Strategic assessments may identify crucial points at which a relatively small investment may free up or generate large amounts of data that would be immediately useful (Peterson et al., 2018). Although this process is not rapid, it will have the most permanent and pervasive impacts on the ability of researchers to conduct analyses of this sort, and thus improve the status of biodiversity science for the region in question.

In sum, in this contribution, we have explored how data density regarding primary biodiversity data characterizations of species' distributions propagates through analysis workflows to affect results of place-prioritization analyses for biodiversity conservation. Reduced data density—either overall or regionally—can lead researchers to use overly simple distributional estimates for species in prioritization efforts. The errors in distributional summaries propagate through the place-prioritization algorithms to produce prioritizations that are distinct from the outcomes when full data density is used, likely an effect of the relative simplicity and lack of detail in the input range summaries. Indeed, major priority areas can "blink on" and "blink off" in the face of changes in data density. As such, and given that the world is characterized by major disparities and imbalances in biodiversity data density, place-prioritization efforts must necessarily consider data densities available with which to characterize species distributions rigorously and in sufficient detail, as well as any disparities across a region of analyses. Failing to do so will result in unanticipated bias in results and interpretation, affecting decision making in conservation efforts.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Marlon E. Cobos: Investigation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Visualization. **Claudia Nuñez-Penichet:** Formal analysis, Writing – review & editing, Visualization. **Peter D. Campbell:** Formal analysis, Visualization. **Jacob C. Cooper:** Investigation, Methodology, Writing – review & editing. **Fernando Machado-Stredel:** Investigation, Methodology, Writing – review & editing. **Narayani Barve:** Formal analysis, Visualization. **Uzma Ashraf:** Formal analysis, Visualization. **Abdelghafar A. Alkishe:** Formal analysis, Visualization. **Eric Ng'eno:** Formal analysis, Writing – review & editing. **Rahul Raveendran Nair:** Formal analysis. **P. Joser Atauchi:** Formal analysis. **Adeola Adeboje:** Formal analysis. **A. Townsend Peterson:** Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that no competing interests exist.

Data availability

A link to data and code has been shared.

Acknowledgments

We thank the remaining members of the KUENM working group for their ideas and input in the development of this contribution.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biocon.2023.110207.

References

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232. https://doi.org/10.1016/S0304-3800(02)00349-6.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J., Villalobos, F., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecol. Model. 222, 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011.

Beck, J., Ballesteros-Mejia, L., Nagel, P., Kitching, I.J., 2013. Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? Divers. Distrib. 19, 1043–1050. https://doi.org/10.1111/ddi.12083.

Beedy, E.C., Hamilton, W.J., Meese, R.J., Airola, D.A., Pyle, P., 2020. In: Rodewald, P.G. (Ed.), Birds of the World. Cornell Lab of Ornithology, Ithaca, New York. https://doi.org/10.2173/bow.tribla.01. Tricolored blackbird (Agelaius tricolor), version 1.0.

Bini, L.M., Diniz-Filho, J.A.F., Rangel, T.F.L.V.B., Bastos, R.P., Pinto, M.P., 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. Divers. Distrib. 12, 475–482. https://doi.org/10.1111/j.1366-9516.2006.00286.x.

Brooks, T.M., Mittermeier, R.A., da Fonseca, G.A.B., Gerlach, J., Hoffmann, M., Lamoreux, J.F., Mittermeier, C.G., Pilgrim, J.D., Rodrigues, A.S.L., 2006. Global biodiversity conservation priorities. Science 313, 58–61. https://doi.org/10.1126/science.1127609.

Brum, F.T., Graham, C.H., Costa, G.C., Hedges, S.B., Penone, C., Radeloff, V.C., Rondinini, C., Loyola, R., Davidson, A.D., 2017. Global priorities for conservation across multiple dimensions of mammalian diversity. PNAS USA 114, 7641–7646. https://doi.org/10.1073/pnas.1706461114.

Cobos, M.E., Peterson, A.T., Barve, N., Osorio-Olvera, L., 2019a. Kuenm: an R package for detailed development of ecological niche models using Maxent. PeerJ 7, e6281. https://doi.org/10.7717/peerj.6281.

Cobos, M.E., Peterson, A.T., Osorio-Olvera, L., Jiménez-García, D., 2019b. An exhaustive analysis of heuristic methods for variable selection in ecological niche modeling and species distribution modeling. Ecol. Inform. 53, 100983 https://doi.org/10.1016/j.ecoinf.2019.100983.

Crase, B., Liedloff, A.C., Wintle, B.A., 2012. A new method for dealing with residual spatial autocorrelation in species distribution models. Ecography 35, 879–888. https://doi.org/10.1111/j.1600-0587.2011.07138.x.

Dallas, T., Decker, R.R., Hastings, A., 2017. Species are not most abundant in the Centre of their geographic range or climatic niche. Ecol. Lett. 20, 1526–1533. https://doi.org/10.1111/ele.12860.

Daru, B.H., Park, D.S., Primack, R.B., Willis, C.G., Barrington, D.S., Whitfeld, T.J.S., Seidler, T.G., Sweeney, P.W., Foster, D.R., Ellison, A.M., Davis, C.C., 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytol. 217, 939–955. https://doi.org/10.1111/nph.14855.

Del Hoyo, J., Elliott, A., Christie, D. (Eds.), 2011. Handbook of the Birds of the World. Vol.16: Tanagers to New World Blackbirds, 1st, edition. ed. Lynx Edicions, Barcelona.

eBird Basic Dataset, Jun 2022. Version: EBD_relJun-2022. Cornell Lab of Ornithology, Ithaca, New York.

Funk, V.A., Richardson, K.S., Ferrier, S., 2005. Survey-gap analysis in expeditionary research: where do we go from here? Biol. J. Linn. Soc. 85, 549–567. https://doi.org/10.1111/j.1095-8312.2005.00520.x.

Gardner, C.J., Nicoll, M.E., Birkinshaw, C., Harris, A., Lewis, R.E., Rakotomalala, D., Ratsifandrihamanana, A.N., 2018. The rapid expansion of Madagascar's protected area system. Biol. Conserv. 220, 29–36. https://doi.org/10.1016/j.biocon.2018.02.011.

GBIF.org (09 December 2022a) GBIF Occurrence Download doi:10.15468/dl.qneyyj.

GBIF.org (09 December 2022b) GBIF Occurrence Download doi:10.15468/dl.qaaa5a.

GBIF.org (13 December 2022c) GBIF Occurrence Download doi:10.15468/dl.9u4tpp.

GBIF.org (20 December 2022d) GBIF Occurrence Download doi:10.15468/dl.fr35nw.

Hart, E., Bell, K., 2015. Prism: access data from the Oregon state PRISM climate project. R package version 0, 2. https://CRAN.R-project.org/package=prism.

Hijmans, R., 2022. terra: Spatial data analysis. R package version 1, 6–17. https://CRAN.R-project.org/package=terra.
Huang, X., Hawkins, B.A., Lei, F., Miller, G.L., Favret, C., Zhang, R., Qiao, G., 2012. Willing or unwilling to share primary biodiversity data: results and implications of an international survey. Conserv. Lett. 5, 399–406. https://doi.org/10.1111/j.1755-263X.2012.00259.x.
Hughes, A.C., Orr, M.C., Ma, K., Costello, M.J., Waller, J., Provoost, P., Yang, Q., Zhu, C., Qiao, H., 2021. Sampling biases shape our view of the natural world. Ecography 44, 1259–1269. https://doi.org/10.1111/ecog.05926.
Ingenloff, K., Hensz, C.M., Anamza, T., Barve, V., Campbell, L.P., Cooper, J.C., Komp, E., Jimenez, L., Olson, K.V., Osorio-Olvera, L., Owens, H.L., Peterson, A.T., Samy, A.M., Simões, M., Soberón, J., 2017. Predictable invasion dynamics in north American populations of the Eurasian collared dove Streptopelia decaocto. Proc. R. Soc. B 284, 20171157. https://doi.org/10.1098/rspb.2017.1157.
Justus, J., Sarkar, S., 2002. The principle of complementarity in the design of reserve networks to conserve biodiversity: a preliminary history. J. Biosci. 27, 421–435. https://doi.org/10.1007/BF02704970.
Kujala, H., Moilanen, A., Gordon, A., 2018. Spatial characteristics of species distributions as drivers in conservation prioritization. Methods Ecol. Evol. 9, 1121–1132. https://doi.org/10.1111/2041-210X.12939.
Kukkala, A.S., Moilanen, A., 2013. Core concepts of spatial prioritization in systematic conservation planning. Biol. Rev. 88, 443–464. https://doi.org/10.1111/brv.12008.
Lehtomaki, J., 2018. Zonator: Utilities for Zonation Spatial Conservation Prioritization software. R package version 0, 6. https://CRAN.R-project.org/package=zonator.
Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G., Williams, P.H., 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conserv. Biol. 17, 1591–1600. https://doi.org/10.1111/j.1523-1739.2003.00233.x.
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, Sunderland, Massachusetts, pp. 293–296.
Machado-Stredel, F., Cobos, M.E., Peterson, A.T., 2021. A simulation-based method for identifying accessible areas as calibration areas for ecological niche models and species distribution models. Front. Biogeogr. 13, e48814 https://doi.org/10.21425/F5FBG48814.
Meese, R.J., 2017. Results of the 2017 tricolored blackbird statewide survey. In: Nongame Wildlife Program Report. California Department of Fish and Wildlife, Wildlife Branch, Sacramento, California.
Moilanen, A., Franco, A.M.A., Early, R.I., Fox, R., Wintle, B., Thomas, C.D., 2005. Prioritizing multiple-use landscapes for conservation: methods for large multi-species planning problems. Proc. R. Soc. B 272, 1885–1891. https://doi.org/10.1098/rspb.2005.3164.
Moilanen, A., Arponen, A., Stokland, J.N., Cabeza, M., 2009. Assessing replacement cost of conservation areas: how does habitat loss influence priorities? Biodivers. Conserv. 142, 575–585. https://doi.org/10.1016/j.biocon.2008.11.011.
Moilanen, A., Montesino Pouzols, F., Meller, L., Veach, V., Arponen, A., Leppanen, J., Kujala, H., 2014. Zonation - spatial conservation planning methods and software. In: User Manual. Version 4. C-BIG Conservation Biology Informatics Group Department of Biosciences University of Helsinki, Finland.
Muscatello, A., Elith, J., Kujala, H., 2021. How decisions about fitting species distribution models affect conservation outcomes. Conserv. Biol. 35, 1309–1320. https://doi.org/10.1111/cobi.13669.
Neff, J.A., 1937. Nesting distribution of the tri-colored red-wing. Condor 39, 61–81. https://doi.org/10.2307/1363776.
Nori, J., Torres, R., Lescano, J.N., Cordier, J.M., Periago, M.E., Baldo, D., 2016. Protected areas and spatial conservation priorities for endemic vertebrates of the Gran Chaco, one of the most threatened ecoregions of the world. Divers. Distrib. 22, 1212–1219. https://doi.org/10.1111/ddi.12497.
Nori, J., Loyola, R., Villalobos, F., 2020. Priority areas for conservation of and research focused on terrestrial vertebrates. Conserv. Biol. 34, 1281–1291. https://doi.org/10.1111/cobi.13476.
Nuñez-Penichet, C., Osorio-Olvera, L., Gonzalez, V.H., Cobos, M.E., Jiménez, L., DeRaad, D.A., Alkishe, A., Contreras-Díaz, R.G., Nava-Bolaños, A., Utsumi, K., Ashraf, U., Adeboje, A., Peterson, A.T., Soberon, J., 2021. Geographic potential of the world's largest hornet, Vespa mandarinia Smith (Hymenoptera: Vespidae), worldwide and particularly in North America. PeerJ 9, e10690. https://doi.org/10.7717/peerj.10690.
Peterson, A.T., Soberón, J., 2018. Essential biodiversity variables are not global. Biodivers. Conserv. 27, 1277–1288. https://doi.org/10.1007/s10531-017-1479-5.
Peterson, A.T., Papeş, M., Soberón, J., 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol. Model. 213, 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008.
Peterson, A.T., Asase, A., Canhos, D., Souza, S. de, Wieczorek, J., 2018. Data leakage and loss in biodiversity informatics. Biodivers. Data J 6, e26826. doi:https://doi.org/10.3897/BDJ.6.e26826.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.
Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E., 2017. Opening the black box: an open-source release of Maxent. Ecography 40, 887–893. https://doi.org/10.1111/ecog.03049.
PRISM Climate Group, (2022), Oregon State University, https://prism.oregonstate.edu, accessed 27 September 2022.
R Core Team, 2022. R: a language and environment for statistical computing. Version 4 (2), 2. https://www.r-project.org/.
Rosenberg, K.V., Kennedy, J.A., Dettmers, R., Ford, R.P., Reynolds, D., Alexander, J.D., Beardmore, C.J., Blancher, P.J., Bogart, R.E., Butcher, G.S., Camfield, A.F., Demarest, D.W., Easton, W.E., Giocomo, J.J., Keller, R.H., Mini, A.E., Panjabi, A.O., Pashley, D.N., Rich, T.D., Ruth, J.M., Stabins, H., Stanton, J., Will, T., 2016. Partners in Flight Landbird Conservation Plan: 2016 Revision for Canada and Continental United States (Partners in Flight Science Committee).
Scoble, M.J., 2000. Costs and benefits of web access to museum data. Trends Ecol. Evolut. 15, 374. https://doi.org/10.1016/S0169-5347(00)01895-4.
Soberón, J., 2022. Biodiversity informatics for public policy. The case of CONABIO in Mexico. Biodiv. Inform. 17, 96–107.
Soberón, J., Peterson, A.T., 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiv. Inform. 2, 1–10. https://doi.org/10.17161/bi.v2i0.4.
Soberón, J., Peterson, A.T., Osorio-Olvera, L., 2018. A comment on "Species are not most abundant in the centre of their geographic range or climatic niche". Rethink. Ecol. 3, 13–18. https://doi.org/10.3897/rethinkingecology.3.24827.
Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: a citizen-based bird observation network in the biological sciences. Biol. Conserv. 142, 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006.
UNEP-WCMC and IUCN, 2022. Protected Planet: The World Database on Protected Areas (WDPA). UNEP-WCMC and IUCN, Cambridge, UK. Available at: www.protectedplanet.net.
Velazco, S.J.E., Bedrij, N.A., Rojas, J.L., Keller, H.A., Ribeiro, B.R., De Marco, P., 2022. Quantifying the role of protected areas for safeguarding the uses of biodiversity. Biol. Conserv. 268, 109525 https://doi.org/10.1016/j.biocon.2022.109525.
Venter, O., Sanderson, E.W., Magrach, A., Allan, J.R., Beher, J., Jones, K.R., Possingham, H.P., Laurance, W.F., Wood, P., Fekete, B.M., Levy, M.A., Watson, J.E. M., 2016. Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. Nat. Commun. 7, 12558. https://doi.org/10.1038/ncomms12558.
Warren, D.L., Seifert, S.N., 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecol. Appl. 21, 335–342. https://doi.org/10.1016/j.ecolmodel.2011.02.011.
Watts, M.E., Ball, I.R., Stewart, R.S., Klein, C.J., Wilson, K., Steinback, C., Lourival, R., Kircher, L., Possingham, H.P., 2009. Marxan with zones: software for optimal conservation based land- and sea-use zoning. Environ. Model. Softw. 24, 1513–1521. https://doi.org/10.1016/j.envsoft.2009.06.005. Special issue on simulation and modelling in the Asia-Pacific region.
Zeller, K.A., Rabinowitz, A., Salom-Perez, R., Quigley, H., 2013. The jaguar corridor initiative: A range-wide conservation strategy. In: Ruiz-Garcia, M., Shostell, J.M. (Eds.), Molecular Population Genetics, Evolutionary Biology, and Biological Conservation of Neotropical Carnivores. Nova Publishers, New York, pp. 629–657.
Zhang, M.-G., Zhou, Z.-K., Chen, W.-Y., Slik, J.W.F., Cannon, C.H., Raes, N., 2012. Using species distribution modeling to improve conservation and land use planning of Yunnan, China. Biol. Conserv. 153, 257–264. https://doi.org/10.1016/j.biocon.2012.04.023.