Université de Montréal

Spatially continuous identification of beta diversity hotspots using species distribution models

By

Gabriel Dansereau

20147609

Département de sciences biologiques

Faculté des arts et des sciences

Advisory Committee Meeting

November 29, 2019

Contents

List of Tables

List of Figures

Abstract

Beta diversity is an essential measure to describe the organization of biodiversity in space. The calculation of local contributions to beta diversity (LCBD), specifically, allows for the identification of sites with exceptional diversity within a region of interest, which is useful for both community ecology and conservation purposes. However, beta diversity implies a comparison among the sites of a given region, thus, its use is restricted to sites with known species composition, and to discontinuous spatial scales. We therefore propose a method to calculate LCBD indices on continuous scales for a whole region of interest, including unsampled sites. First, species distributions can be predicted on continuous scales using species distribution models (SDM). These models, such as the BIOCLIM method, use the environmental conditions at sampled sites to predict the presence or absence of each species at unsampled locations. Second, LCBD statistics can then be computed on the SDM predictions. We show that it is therefore possible to identify beta diversity hotspots on spatially continuous and extended scales. Our results confirm that LCBD values are related to species richness, and that species-poor sites contribute most to beta diversity.

Introduction

Beta diversity, defined as the variation in species composition among sites in a geographic region of interest (Legendre, Borcard, and Peres-Neto 2005), is an essential measure to describe the organization of biodiversity in space. Total beta diversity within a community can be partitioned into local contributions to beta diversity (LCBD) (Legendre and De Cáceres 2013), which allows for the identification of sites with exceptional species composition, hence exceptional biodiversity. Such a method is useful for both community ecology and conservation biology, as it highlights sites that are most important for their research or conservation values. However, LCBD calculation methods require complete information on community composition, such as a community composition matrix Y, thus they are inappropriate for partially sampled or unsampled sites. To our knowledge, theses methods have mostly been applied on community data from sampled sites, hence on discontinuous spatial scales, e.g. at intervals along a river stream (Legendre and De Cáceres 2013). This raises the following questions: 1) could LCBD indices be extended to continuous spatial scales, and 2) could this provide novel ecological insights in poorly sampled regions? We aim to answer these questions by combining the LCBD calculation methods with predictive biogeography approaches, and suggest that this would allow for the identification of hotspots with high conservation value in poorly sampled regions.

Species distribution models (SDMs) already allow to make predictions on continuous spatial scales, and these predictions could therefore be used to calculate LCBD indices. SDMs, also known as bioclimatic envelope models (Araújo and Peterson 2012), aim to predict species presence or absence based on previous observations of occurrence, and the environmental conditions at which these were made (Poisot et al. 2019). Examples of uses include climate change impact and invasion risk assessment, reserve selection and design, and discovery of new populations (Araújo and Peterson 2012). This way, they generate novel ecological insights for unsampled or lesser-known locations (Poisot et al. 2019), an approach yet to be applied to the LCBD framework. We believe that a predictive approach such as this one would bring a new perspective to biodiversity study and community ecology. By using SDMs, we would be able to expand community information already available, and thus work on a much larger community matrices than in typical LCBD studies, which might highlight new

diversity hotspots.

Climate and biodiversity data on extended spatial scales are increasingly available online. For instance, the Worldclim 2.0 database (Fick and Hijmans 2017) provides interpolated climate data for global land areas at very high spatial resolution, and the eBird platform (Sullivan et al. 2009) provides a rapidly growing, citizen-contributed database of worldwide bird observations. Both of these are commonly used in SDMs, and offer relevant information on extended spatial scales. Therefore, we believe that these datasets could be used to predict community composition and calculate LCBD indices on continuous spatial scales, and that the result would be representative of the true community structure.

The predictive approach we suggest would be especially useful in poorly sampled regions, or in regions with only sparse sampling. While it does not replace a full sampling within the community, predictions and exploratory analyses do provide relevant ecological insights that could be used in different ways. For instance, our method could help identify unsampled sites with potential conservation value which should be targeted as soon as possible in future studies. We believe that the method could also be combined with IPCC climate change scenarios, which provide projections for climate variables, in a way that would allow us to model beta diversity changes with climate change and to identify the sites where the changes in the community will be most important. Once again,this would prove very relevant in an informative approach, suggesting sites to prioritize for future conservation and more structured research.

In this document, we cover in more details the methods that we suggest for this M.Sc. research project. The preparation part of the project, including data collection and manipulation, has already been done. A workflow for the analyses, including code implementation, has been defined as well. We also detail preliminary analyses and results intended as proof-of-concept for the approach, which, of course, needs to be refined. Finally, we discuss methods that we intend to use in future analyses, and whose feasibility is not as clearly stated.

Methods

1. Data Collection

We decided to focus our analyses on bird species and collected the data available on eBird for the Warblers family. The complete database contains nearly 600 million observations. We chose to focus specifically on the Warblers family, as it is a diverse group, popular among birders, with over 30 million observations. Global citizen-contributed databases often present additional challenges compared to conventional datasets due to their lack of structure, as well as spatial and taxonomic biases (Johnston et al. 2019). For instance, there was a clear bias in our data towards the United States, where there were many more observations and sampling events (tbl. 1). However, eBird offers two advantages over other large scale datasets (Johnston et al. 2019): 1) the data is structured as checklists and users can explicitly specify their observations as "complete checklists" when all detected species were reported, which allows to infer information on species absences, and 2) the dataset is semi-structured and checklists are associated with metadata describing sampling effort, such as duration of search, distance travelled and number of observers, which can be used as controls in the analyses. Hence, model performance can be improved by inferring absences and subsampling checklists, while spatial bias can be compensated by including effort covariates in the model (Johnston et al. 2019). Therefore, we believe the dataset can be appropriately used to achieve our objective of expanding measures of exceptional biodiversity through space.

We collected the data available in the WorldClim 2 database (Fick and Hijmans 2017) for North America, to which we decided to restrict our analyses. The WorldClim data consists of spatially interpolated monthly climate data for global areas, available for resolutions from 10 arc-minutes to 30 arc-seconds (around 18 km^2 and 1 km^2 at the equator). Since the release of the first version of the database in 2005 (Hijmans et al. 2005), it became the most common source of climate data for SDM studies (Booth et al. 2014). The variables we used were different measures of temperature and precipitation (tbl. 2), which very high global cross-validation coefficients (> 0.99 and 0.86 respectively) (Fick and Hijmans 2017). We chose to use the coarser 10 arc-minutes resolution in our preliminary analyses, as we believed it was sufficient for proof of concept of our method. However, Hijmans et al. (2005) showed high within-grid cell variation in the 10 arc-minutes data, and therefore recommended the

use of the finer resolution, which hid less of the variation known to the model. Given this, we might reconsider the resolution to use in our final analyses.

We chose to restrict our analyses to North America given the high amount of data available in eBird. We believed it represented a suitable scale for our models, large enough to cover a lot of variation in environmental variables and community structure, as well as phenomenons such as species migration. We also expected such extent of the spatial scale to cover for imprecision in estimated species ranges.

2. Data Manipulation

WorldClim variables and eBird occurrence data are provided in different formats, so they required some manipulations to be combined together. WorldClim variables are provided in a 2-dimensional grid format, useful for large scale analyses and visualization, where each cell or pixel has a size corresponding to the resolution of 10 arc-minutes. Each of the 19 variables forms a different grid. On the other hand, eBird records are occurrence-based, so each entry in the dataset corresponds to an observation of a single species at a given time and location. These entries can easily be matched to the 2D grid format of the WorldClim variables through their spatial coordinates, which we found more useful for large scale analyses and visualization. Hence, for each species, we matched all occurrences in eBird to the grid format of the WorldClim variables, and then created a presence-absence community matrix Y, taking all the grid cells as sites. At the 10 arc-minutes resolution, we obtained 39 024 sites with occurrences and 62 species in total. All data manipulations and further analyses were realized in *Julia v1.2.0* (Bezanson et al. 2017), with the basic structure built around the soon-to-be-released `SimpleSDMLayers.jl` package.[1]

3. SDM – The BIOCLIM Method

We predicted species distributions using the BIOCLIM method (Nix 1986), a climate-envelope model, considered a classic in the field. This method simply relates a species' distribution to the ranges of bioclimatic variables at known locations (Booth et al. 2014). It has long been outperformed by other methods (Elith et al. 2006), but it is still commonly used for its simplistic approach and ease of understanding, as well as its simple relation to niche

[1] https://github.com/EcoJulia/SimpleSDMLayers.jl

theory (Booth et al. 2014; Hijmans et al. 2017). It is also primarily designed for presence-only data. Despite that, we chose this method for our preliminary analyses as it was easier to implement and because we believe it to be sufficient for proof-of-concept. We discuss possible alternatives in the "Alternative Methods" section below.

The BIOCLIM method defines species potential ranges as a multidimensional environmental hypervolume bounded by the minimum and maximum values for all occurrences (Franklin 2010). For each species, we established the percentile distribution of each environmental variable at the known locations of occurrence (Hijmans et al. 2017). All sites were then compared to those percentile distributions and given a score per variable according to their ranking between 0.0 (1st percentile) and 1.0 (100th percentile). The median or 50th percentile was considered the most suitable value of the variable, and values larger than 0.5 were subtracted from 1. Therefore, both tails were considered the same. The minimum percentile score across all environmental variables was then selected as the predicted value for each site. Values were multiplied by 2 and could therefore be interpreted as probabilities of species occurrence (Hijmans et al. 2017). Predictions of 1 should be rare by definition, as they require a perfectly median site on all variables, and values of 0 should be frequent, since they occur whenever an environmental value is outside the range of the observed ones (Hijmans et al. 2017).

The final step was to convert the probabilities into presence-absence data, so they could be compared with the raw occurrence data. We transformed the probabilities into zeros and ones by converting all values greater than zero to one. Although it might tend to overestimate species ranges, such a transformation is common in SDMs and can be accounted for during result validation with specific methods (Franklin 2010). We also considered applying a threshold determined by sensitivity analysis, but we haven't done it yet. In any case, converting into presence-absence data allowed easier calculation of the richness and beta diversity metric.

4. LCBD Calculation

We calculated the LCBD statistics through the total variance of the matrix Y for both the raw data and SDM predictions. Legendre and De Cáceres (2013) showed that LCBD coefficients

can be calculated directly through the total variance of matrix Y, or through a matrix of dissimilarities among sampling units. We chose the first approach as it also allows to compute species contributions to beta diversity (SCBD), which could also prove useful for conservation purposes, but we did not investigate these for now. Before computing the LCBD statistics, the presence-absence matrix Y had to be transformed in an appropriate way (Legendre and De Cáceres 2013). We chose to apply the Hellinger transformation to the raw data and no transformation on the SDM predictions for now, although we did not investigate these in detail. The most appropriate transformation still needs to be determined, especially for the SDM predictions. We then computed a matrix S of squared deviations from column means and summed all the values of S to obtain the total sum of squares (SS) of the species composition data (Legendre and De Cáceres 2013). LCBD coefficients are then computed as $LCBD_i = SS_i/SS_{Total}$, where SS_i is the sum of squares of a sampling unit i. Finally, since our matrix Y is very large, the LCBD coefficients are very small, so we scaled them to the maximum value observed.

5. Prediction Validity

The exact way of testing the validity of the predictions remains to be determined, and will also depend on the exact methods used to make the SDM predictions. A key element to note is that both SDM predictions and LCBD values will have to be validated, hence they might require different methods. Metrics that measure the accuracy of categorical or probabilistic predictions in SDMs are well documented, and take various forms. Some require absence data to test against, and can be used on probabilistic predictions directly (area-under-curve, AUC) or after a conversion of the predictions to binary presence-absence using a given threshold (Kappa index, measuring the difference between observed and chance agreement in a confusion matrix) (Franklin 2010). Other methods are appropriate for presence-only data, such as the Boyce Index. In any case, measuring prediction error is only one part of the validation. Finding appropriate data for evaluation is also critical (Franklin 2010), especially since we aim to describe community structure. Separating the data into training and testing datasets, with 70% and 30% of the observations for instance, is an approach common in machine learning methods. However, all of the available observations might be needed in some cases (Franklin 2010). An interesting approach, suggested by Elith et al.

(2006) for SDMs, would be to find independent, well-structured presence-absence datasets for validation, on which both SDM predictions and beta diversity metrics could be tested. This approach has the advantage that the testing data is truly independent of the training one, hence it could be used with certain tests of significance. Although it might not cover the entire extent of the predictions in a single test, this method would bring a closer comparison to the way LCBD metrics are used in most studies. Therefore, it would provide interesting perspectives if combined with other, full-extent validation methods.

6. Alternative methods

Many methods generally outperform BIOCLIM for the predictions, as shown by Elith et al. (2006). In our case, better predictions will come by two different means: 1) approaches that are better than BIOCLIM to model the relationship between species presence-absence (or even abundance) and environmental variables, and 2) approaches that account for other drivers of species distributions, such as ecological interactions and species migration. Machine learning methods, especially, would be interesting alternatives to consider. MAXENT (Phillips, Anderson, and Schapire 2006), another presence-only method, has come to be one of the most widely used methods in SDM studies, often with WorldClim variables (Booth et al. 2014). Similarly, Random Forests are simple to put in place, take into account both presence and absence data, allow for quantification of the variables importance in explaining variation, and offer intrinsic testing metrics (Franklin 2010). However, while those methods might return more accurate predictions, they do not implicitly model other drivers of species distribution, among which species interactions and functional niche. Integrating those factors might prove more difficult given our dataset and our focus on Warblers species, as no appropriate information on their interaction is available. Joint species distribution models (JSDMs) might be an interesting way to encompass those, as they attempt to model species co-occurrence, rather than the distribution of single species (Pollock et al. 2014). Also, a different taxonomic group and dataset with more details on interactions could simply be used. On the other hand, a method that could be applied to any taxonomic group, especially those well represented in large citizen-contributed datasets, would be most useful for research and conservation purposes.

7. Climate Change Scenarios and Temporal Beta Diversity

We aim to apply our method to environmental conditions from IPCC climate change scenarios. First, community compositions after climate change could be modelled on continuous scales through SDMs. Second, we could identify the sites where the community has changed in the most exceptional ways. This identification can be done by looking at the variation in LCBD values, but also through the use of temporal beta diversity indices (TBI) (Legendre 2019). TBI indices allow to study changes in community composition through time from repeated surveys at given sites. Whereas LCBD values essentially measure the contribution to beta diversity of one site compared to all others, TBI measure changes in community composition site-wise between two surveys. Moreover, TBI indices can be decomposed into species losses and gains, and can be tested for significance using a permutation test (Legendre 2019). An approach similar to that of Legendre and Condit (2019) would be interesting to follow in our case. First, they computed LCBD indices and compared the location of the sites with exceptional compositions between two surveys 30 years apart. The comparison showed that important changes seemed to have occurred in a specific swamp region. Then, they used TBI indices to confirm the sites with significant changes, decompose these changes into losses and gains, and identify the species that had changed the most. An approach such as this one could be highly informative with our data, although the permutation tests and corrections required might cause some problems given the number of sites in our study.

The possibility of using climate change scenarios in the SDMs also needs to be assessed. We did not try to download nor find the appropriate data for now. However, interpolated climate change variables are sometimes different than the ones in WorldClim. Therefore, the SDM models to use and the resulting predictions might have to be different too, and potentially less reliable. Nonetheless, we believe it will be possible to do some kind of time analysis linking beta diversity, climate change and species distribution modelling, and that it could return highly informative results for conservation purposes.

Preliminary Results

Our preliminary results consisted of comparisons between the raw occurrence data and the SDM predictions for the four following elements: single-species distribution (fig. 1),

species richness (fig. 2), LCBD coefficients (fig. 3), as well as the relationship between the species richness and LCBD coefficients (fig. 4). Two main results emerged from them: 1) the models provided community composition results for poorly sampled regions, both expected species-poor and species-rich, and 2) the relationship between species richness and species distribution models was in line with previous studies for species poor sites, but the SDM models captured a new association for very rich sites.

First, the example of the Yellow Warbler (*Setophaga petechia*), one of the most observed species, showed that the single-species models predicted a broad distribution covering poorly sampled areas, with notable patches of absence across the continent (fig. 1). Likewise, species richness, defined as the number of species present per site, showed a clear latitude gradient, with the poorest sites to the North and the richest to the South (fig. 2). A form of altitude gradient could also be observed, with the Rockies and other mountains well delimited by their lower values. In both cases, the results make intuitive sense and highlight the models ability to predict species presence despite poor or no sampling. Mexico, for example, has much sparser sampling and fewer observations, but the models predict Yellow Warblers presence in most areas nonetheless, as well as higher species richness than on the highly sampled Atlantic Coast, which make sense for a more southern location. We believe these to be valid insights on poorly sampled locations, but there is still a need for an appropriate method of validation to confirm our intuition, as well as a thoughtful consideration of factors such as species migration.

Second, our preliminary LCBD results seemed to confirmed the association between species richness and LCBD coefficients from previous studies, but the SDM predictions captured a new association for extremely rich sites. Indeed, raw occurrence data showed a negative relationship between species richness and LCBD coefficients (fig. 4), as observed previously by Heino and Grönroos (2017), with no clear geographic pattern (fig. 3a). On the other hand, SDM predictions showed a clear geographic pattern, with the highest values to the northern and southern extremes (fig. 3b). Moreover, the richness-LCBD relationship showed a quadratic form, with the LCBD coefficients re-increasing beyond a richness of 0.6 (fig. 4), which approximately corresponds to the maximum richness observed in the raw occurrence data. Therefore, the SDM predictions captured a new association that could not be seen in the

occurrence data, possibly because there were no rich enough sites to display it. By definition, LCBD indices should highlight the most exceptional species compositions, both species poor or species rich. Thus, the quadratic relationship we observed makes sense, on the condition that extremely rich sites can realistically exist. These richest sites should likely have been in the southernmost locations such as Mexico, which is heavily undersampled. For instance, all but two species were seen there at least once, but the maximum number of species recorded in a single checklist was lower than in the US and in Canada tbl. 1, which was a little surprising. Hence, there might have been extremely rich communities that were not sampled sufficiently to reveal their true community structure. On the other hand, our models might have been too optimistic in predicting the existence of such rich sites. In any case, our method did provide relevant and novel ecological insights, as we expected. The concurrence of our SDM predictions for intermediate and species-poor sites with the raw occurrence data, as well as the results of Heino and Grönroos (2017), is promising. The possibility of finding new associations should therefore only encourage to push its use even further.

Finally, one disappointing aspect of our method is that the result failed to identify patterns on finer scales. The trends shown by the SDMs for both the species richness and LCBD coefficients were large-scale, latitude-related patterns. Except for mountains, few exceptional sites are actually shown in the middle of the landscape. While it might have been unrealistic to expect such results from a coarse analysis like ours, it would be useful for conservation purposes to be able the identify precise sites within smaller regions. This might be achieved by using a finer resolution, which we should probably reconsider in light of these results, or by using a different technique, such as training the models and predicting species distributions on large scales, but computing and scaling LCBD values on finer local ones, which might highlight regional differences in a new way.

References

Araújo, Miguel B., and A. Townsend Peterson. 2012. "Uses and Misuses of Bioclimatic Envelope Modeling." *Ecology* 93 (7): 1527–39. https://doi.org/10.1890/11-1930.1.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2017. "Julia: A Fresh Approach to Numerical Computing." *SIAM Review* 59 (1): 65–98. https://doi.org/10.1137/141000671.

Booth, Trevor H., Henry A. Nix, John R. Busby, and Michael F. Hutchinson. 2014. "BIOCLIM: The First Species Distribution Modelling Package, Its Early Applications and Relevance to Most Current MaxEnt Studies." *Diversity and Distributions* 20 (1): 1–9. https://doi.org/10.1111/ddi.12144.

Elith, Jane, Catherine H. Graham, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, et al. 2006. "Novel Methods Improve Prediction of Species' Distributions from Occurrence Data." *Ecography* 29 (2): 129–51. https://doi.org/10.1111/j.2006.0906-7590.04596.x.

Fick, Stephen E., and Robert J. Hijmans. 2017. "WorldClim 2: New 1-Km Spatial Resolution Climate Surfaces for Global Land Areas." *International Journal of Climatology* 37 (12): 4302–15. https://doi.org/10.1002/joc.5086.

Franklin, Janet. 2010. "Moving Beyond Static Species Distribution Models in Support of Conservation Biogeography: Moving Beyond Static Species Distribution Models." *Diversity and Distributions* 16 (3): 321–30. https://doi.org/10.1111/j.1472-4642.2010.00641.x.

Heino, Jani, and Mira Grönroos. 2017. "Exploring Species and Site Contributions to Beta Diversity in Stream Insect Assemblages." *Oecologia* 183 (1): 151–60. https://doi.org/10.1007/s00442-016-3754-7.

Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis. 2005. "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *International Journal of Climatology* 25 (15): 1965–78. https://doi.org/10.1002/joc.1276.

Hijmans, Robert J., Steven Phillips, John Leathwick, and Jane Elith. 2017. *Dismo: Species Distribution Modeling*. https://CRAN.R-project.org/package=dismo.

Johnston, A., W. M. Hochachka, M. E. Strimas-Mackey, V. Ruiz Gutierrez, O. J. Robinson, E. T. Miller, T. Auer, S. T. Kelling, and D. Fink. 2019. "Best Practices for Making Reliable Inferences from Citizen Science Data: Case Study Using eBird to Estimate Species Distributions." *bioRxiv*, March, 574392. https://doi.org/10.1101/574392.

Legendre, Pierre. 2019. "A Temporal Beta-Diversity Index to Identify Sites That Have Changed in Exceptional Ways in Space–Time Surveys." *Ecology and Evolution* 9 (6): 3500–3514. https://doi.org/10.1002/ece3.4984.

Legendre, Pierre, Daniel Borcard, and Pedro R. Peres-Neto. 2005. "Analyzing Beta Diversity: Partitioning the Spatial Variation of Community Composition Data." *Ecological Monographs* 75 (4): 435–50. https://doi.org/10.1890/05-0549.

Legendre, Pierre, and Richard Condit. 2019. "Spatial and Temporal Analysis of Beta Diversity in the Barro Colorado Island Forest Dynamics Plot, Panama." *Forest Ecosystems* 6 (1): 7. https://doi.org/10.1186/s40663-019-0164-4.

Legendre, Pierre, and Miquel De Cáceres. 2013. "Beta Diversity as the Variance of Community Data: Dissimilarity Coefficients and Partitioning." *Ecology Letters* 16 (8): 951–63. https://doi.org/10.1111/ele.12141.

Nix, Henry A. 1986. "A Biogeographic Analysis of Australian Elapid Snakes." *Atlas of Elapid Snakes of Australia* 7: 4–15.

Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. 2006. "Maximum Entropy Modeling of Species Geographic Distributions." *Ecological Modelling* 190 (3): 231–59. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Poisot, Timothée, Richard LaBrie, Erin Larson, Anastasia Rahlin, and Benno I. Simmons. 2019. "Data-Based, Synthesis-Driven: Setting the Agenda for Computational Ecology." *Ideas in Ecology and Evolution* 12 (July). https://doi.org/10.24908/iee.2019.12.2.e.

Pollock, Laura J., Reid Tingley, William K. Morris, Nick Golding, Robert B. O'Hara, Kirsten M. Parris, Peter A. Vesk, and Michael A. McCarthy. 2014. "Understanding Co-Occurrence by Modelling Species Simultaneously with a Joint Species Distribution Model (JSDM)." *Methods in Ecology and Evolution* 5 (5): 397–406. https://doi.org/10.1111/2041-210X.121

80.

Sullivan, Brian L., Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and Steve Kelling. 2009. "eBird: A Citizen-Based Bird Observation Network in the Biological Sciences." *Biological Conservation* 142 (10): 2282–92. https://doi.org/10.1016/j.biocon.2009.05.006.

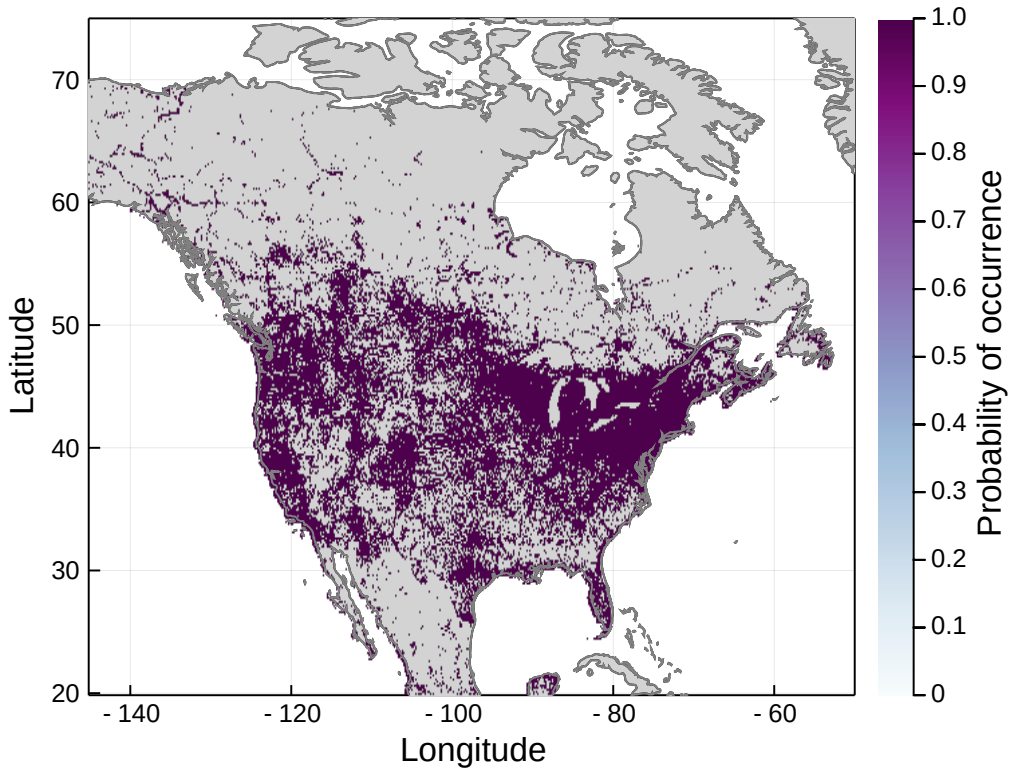
Appendix

Table 1: Structure of the Warblers data in the eBird checklists for the countries used in the analyses

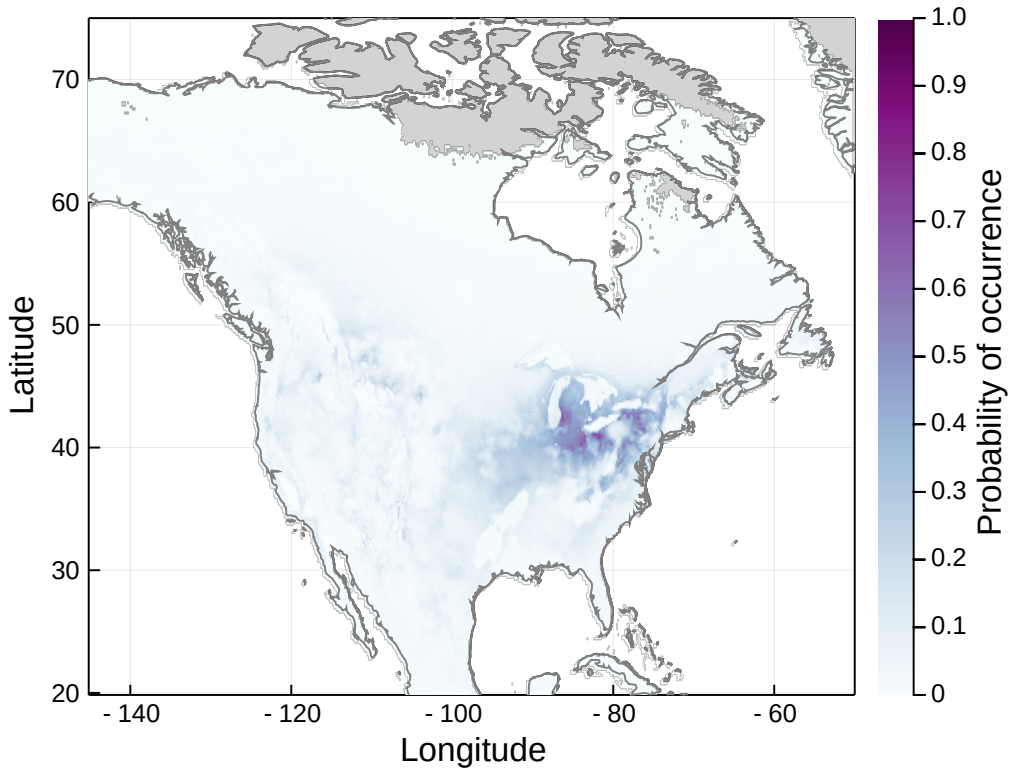
Country	Observations	Checklists	Species	Species per checklist (mean)	Species per checklist (median)	Species per checklist (maximum)
US	19 206 453	7 840 526	56	2.450	2.0	34
CA	3 360 650	1 115 625	45	3.012	2.0	31
MX	407 227	147 599	61	2.759	2.0	21
Total	22 974 330	9 103 750	63	2.523	2.0	34

Table 2: Descripion of the WorldClim 2 climate variables used in the analyses

Variable	Description
1	Annual Mean Temperature
2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
3	Isothermality (BIO2/BIO7) (* 100)
4	Temperature Seasonality (standard deviation *100)
5	Max Temperature of Warmest Month
6	Min Temperature of Coldest Month
7	Temperature Annual Range (BIO5-BIO6)
8	Mean Temperature of Wettest Quarter
9	Mean Temperature of Driest Quarter
10	Mean Temperature of Warmest Quarter
11	Mean Temperature of Coldest Quarter
12	Annual Precipitation
13	Precipitation of Wettest Month
14	Precipitation of Driest Month
15	Precipitation Seasonality (Coefficient of Variation)
16	Precipitation of Wettest Quarter
17	Precipitation of Driest Quarter
18	Precipitation of Warmest Quarter
19	Precipitation of Coldest Quarter

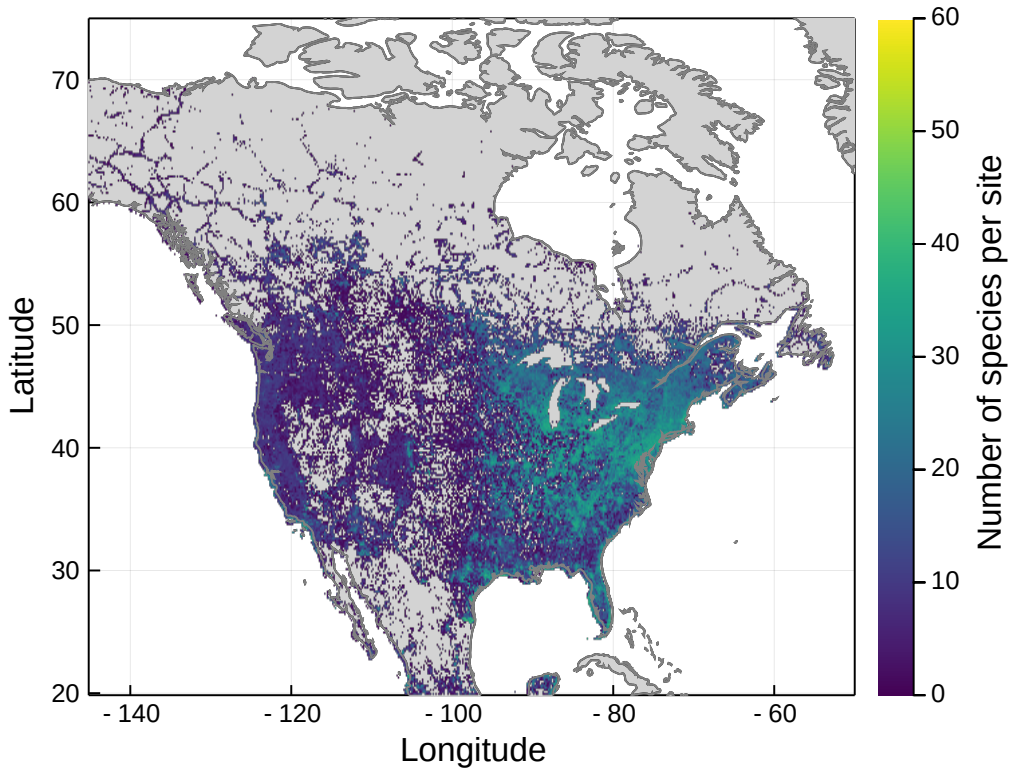


(a) Raw occurrence data

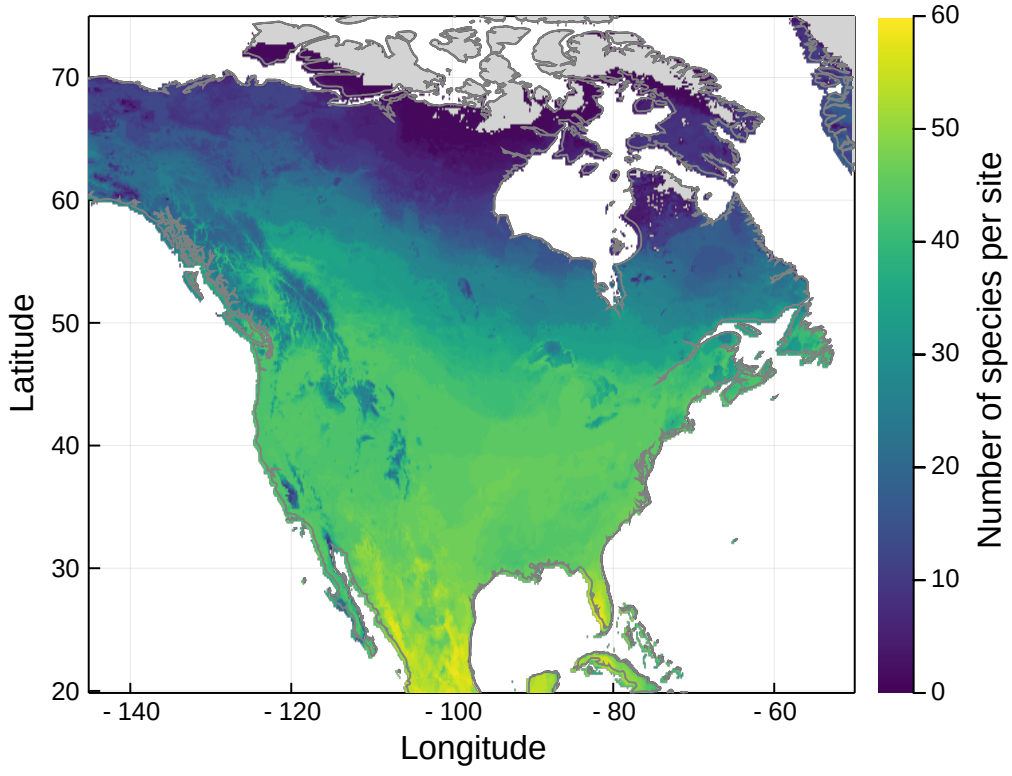


(b) SDM predictions

Figure 1: Distribution of a single species, the Yellow Warbler (*Setophaga petechia*), based on the raw occurrence data (fig. 1a) and on the probabilistic SDM predictions from the BIOCLIM model (fig. 1b). Purple spots in fig. 1a represent sites where the species was observed. fig. 1b present the probabilities of occurrence as a gradient ranging from 0.0 (species absent) to 1.0 (species present).

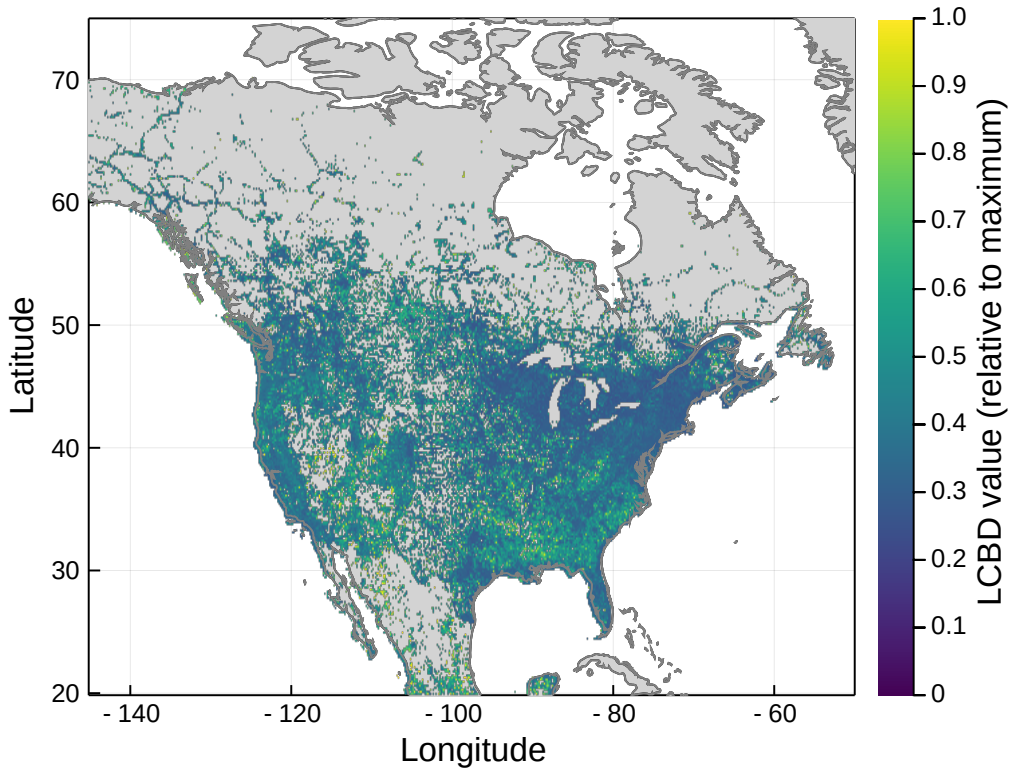


(a) Raw occurrence data

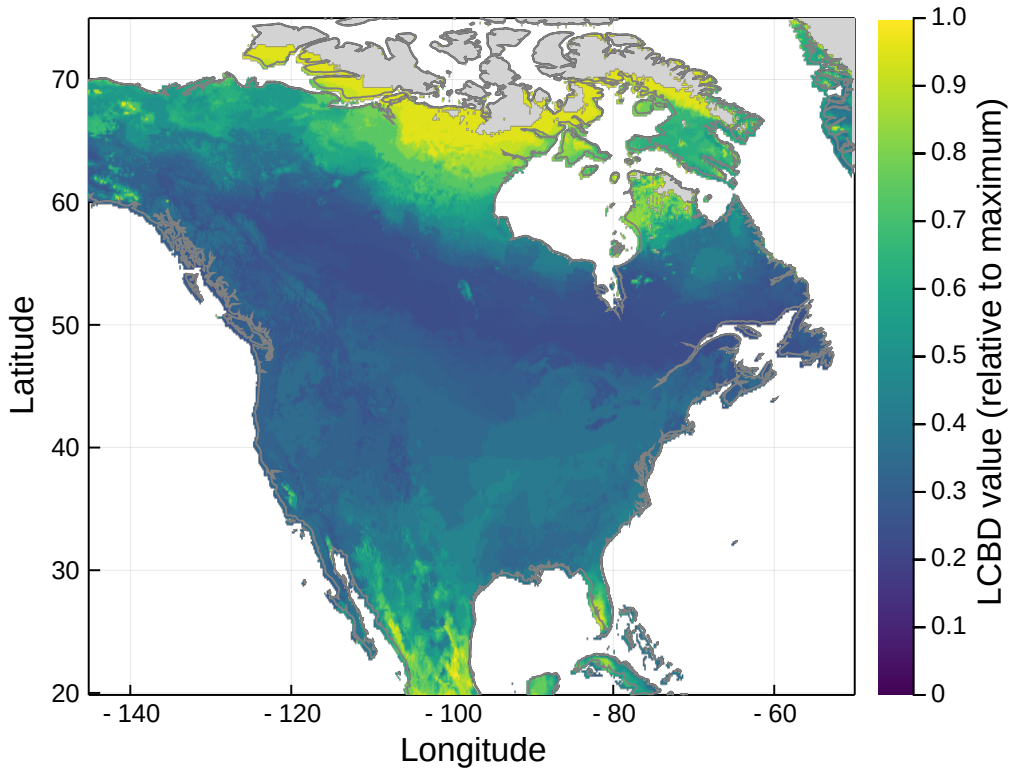


(b) SDM predictions

Figure 2: Distribution of species richness in North America, defined as the number of Warblers species per site. The raw occurrence observations from eBird (fig. 2a) and the SDM predictions from the BIOCLIM model (fig. 2b) were both transformed into presence-absence data per species before calculating richness.



(a) Raw occurrence data (Hellinger transformed)



(b) SDM predictions

Figure 3: Distribution of the LCBD values in North America, calculated from the variance of the community matrix Y and scaled to the maximum value observed. The Hellinger transformation was applied on the raw occurrence data (fig. 3a) before calculating the LCBD indices. SDM predictions (fig. 3b) were converted into presence-absence data, but no transformation was applied before calculating the LCBD indices.

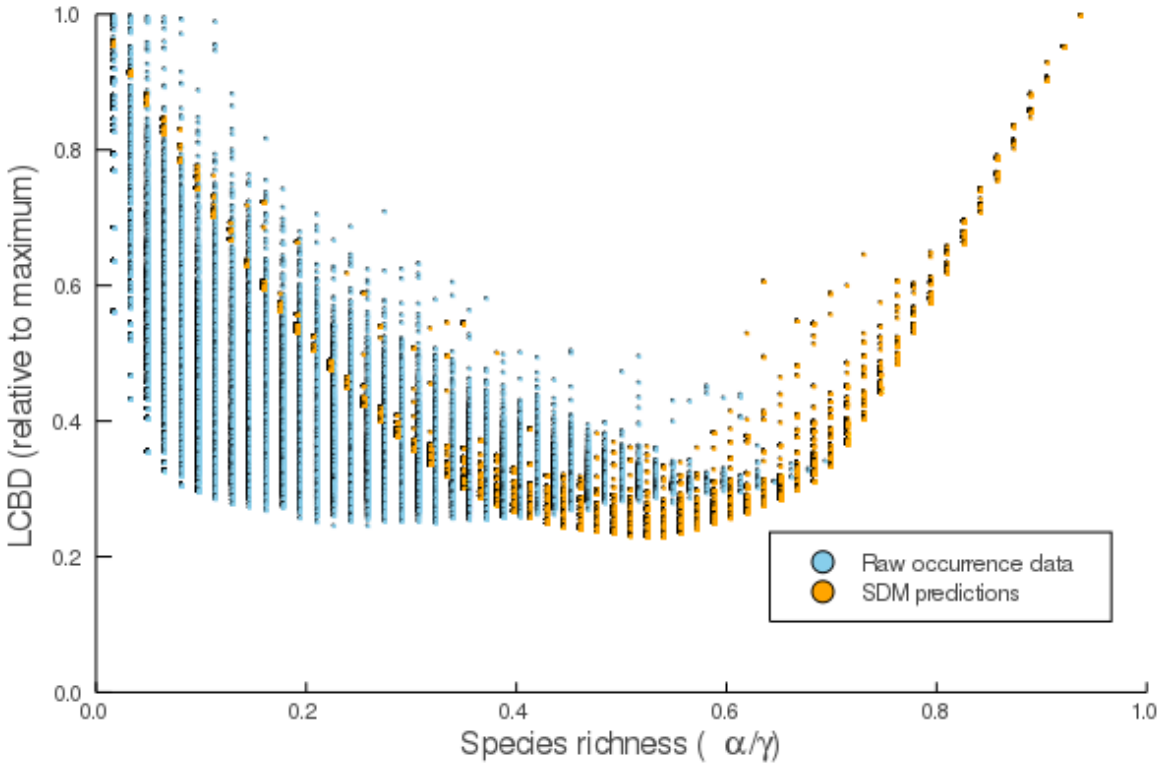


Figure 4: Relationship between the species richness and the LCBD value of the each site for raw occurrence data (blue) and SDM predictions (orange). Species richness was calculated as the number of species in a site (α), divided by the total number of species (γ). LCBD values were scaled to the maximum value observed. Hellinger transformation was applied on the raw occurrence data before calculating LCBD indices.