Spatially continuous identification of beta diversity hotspots using species distribution models

Advisory Committee Document

Gabriel Dansereau

November 25, 2019

Advisory Committee Document

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 therefore show that it is 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, thus 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 sites with high conservation value in poorly sampled regions.

Species distribution models (SDMs) already allow to make predictions on continuous spatial scales which could be used to calculate LCBD indices. Theses methods, also known as bioclimatic envelope models (Araújo and Peterson 2012), aim to predict species presence or absence based on observation of occurrences at known locations (Poisot et al. 2019). This way, they generate novel ecological insights, and represent an approach yet to be applied to LCBD. We believe that such an approach of generating novel ecological insights for unsampled or lesser-known locations could be an interesting new perspective in the study. Through them, we would be able to expand community information already available, and thus work on a much larger community matrix than in typical LCBD studies.

Appropriate data to expand measures of exceptional biodiversity through space is 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 growing citizen-contributed database of worldwide bird observations. Both of these are commonly used in SDMs, and offer relevant information on extended spatial scales. Hence, we believe that we could use them 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 doesn't replace a full sampling within the community, it does provide relevant ecological insights. For instance, the method could help identify unsampled sites with potential conservation value which should be targeted as soon as possible in future studies. We also believe that our 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. Again, this method would be more relevant as an informative approach to suggest 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 research project. The preparation part of the project, including data collection and manipulation, has already been done, and 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, and presents two main advantages over other large scale datasets (Johnston et al. 2019): 1) data is structured as checklist and users can explicitly specify their observations as "complete checklists" when all detected species were reported, which allows to infer information on species absences, 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. We chose to focus specifically on the Warblers family, as it is a diverse group, popular among birders, with over 30 million observations.

We decided to restrict our analyses to North America and collected climate data available in the WorldClim 2 database (Fick and Hijmans 2017). We believe North America represents a suitable scale, large enough to cover a lot of variation in environmental variables and community structure, as well as phenomenons such as species migration. We also expect such extent of the spatial scale to cover for imprecision in estimated species ranges. The WorldClim data consists of spatially interpolated monthly climate data for global areas, available for resolutions from 10 arc-minutes to 30 arc-seconds. The variables used are provided in Table 1, and consists of different measures of temperature and precipitation. We chose to use the coarser 10 arc-minutes resolution in our analyses, again to cover for imprecision, and because we believe it is sufficient for proof of concept.

Table 1: 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

Variable	Description
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

2. Data Manipulation

WorldClim variables and eBird occurrence data are provided in different formats, so they require some manipulation 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 corresponds 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 location. These entries can easily be matched to the 2-D 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 later created a presence-absence community matrix Y, with the sites being the grid cells. We also applied the Hellinger transformation on the raw presence-absence data, although the most appropriate method remains to be determined, especially since the data has to be compared with the SDM predictions. 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.

3. SDM – The BIOCLIM method

We used the BIOCLIM method to predict species distributions. BIOCLIM, first introduced by (Nix 1986), is considered as the classic "climate-envelope-model", and is now available to users through the `dismo` package in R (Hijmans et al. 2017). 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 theory (Booth et al. 2014; Hijmans et al. 2017). It is also a method designed for presence-only data, which does not require information on absences, nor take them into account if provided (as in our case). 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.

Briefly, the BIOCLIM method defines species potential range as a multidimensional environmental hypervolume bounded by the minimum and maximum values of all presences (Franklin 2010). For each species, the algorithm establishes the percentile distribution of the values of each environmental variables at the known locations of occurrences (Hijmans et al. 2017). The environmental variables of all sites are then compared to those percentile distributions and given scores between 0 (1st percentile) and 1 (100th percentile). The median or 50th percentile is considered as the most suitable location and both tails (e.g. 10th and 90th percentile) are not distinguished, the values larger than 0.5 being subtracted from 1. The minimum percentile score across all environmental variables is selected as the prediction value for each site and multiplied by 2 so values are between 0 and 1 (Hijmans et al. 2017). It should be noted that the limiting variable is thus not necessarily the same for all sites. Values of 1 are rare, as it would mean a perfectly median site on all variables, and values of 0 are frequent, since they are assigned whenever an environmental value is outside the range of the observed values (Hijmans et al. 2017). Finally, before calculating richness or beta diversity metrics, we transformed the predictions back to a presence-absence format, where all predictions greater than one are considered as presence. This might tend to overestimate species ranges and create some sort of border effect, but we believe the effects will be mitigated given the spatial extent and coarse scale of our study.

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), although we did not investigate it for

now. First, the presence-absence matrix Y had to be transformed in an appropriate way, as mentioned earlier. We chose to apply the Hellinger transformation to the raw data and no transformation on the SDM predictions for now, as the most appropriate one still needs to be determined. 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 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.

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, so will likely require different methods. Many metrics are well documented in the literature to test SDM predictions, such as the Kappa index (Franklin 2010), and could be used for the BIOCLIM predictions. Another possible way would to separate the data into a training and testing dataset, with 70% and 30% of the data for instance, which is a common approach in machine learning techniques. However, this approach reduces the amount of data that can be used in the model, and raises the issue of making sure that the datasets are both random and representative of the data, as well as the community dynamics. Also, in this framework, the testing data cannot be considered as independent, which prevents using it in certain tests of significance. One interesting approach, suggested by (Elith et al. 2006) for SDMs, would be to find independent, well-structured presence-absence datasets for validation, on which beta diversity metrics has or could be calculated. This validation might not cover the entire extent of the predictions, but it might bring interesting perspectives if combined with other validation methods, mostly because it would bring a closer comparison to the way LCBD metrics are used at the moment.

6. Alternative methods

Other methods could possibly outperform BIOCLIM for the predictions, as have already proven by Elith et al. (2006). 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 for instance. The most obvious alternative to BIOCLIM is MAXENT (Phillips, Anderson, and Schapire 2006), another presence-only method that has come to be one of the most widely used methods. Machine learning methods would be also be interesting alternatives that have been proven to outperform BIOCLIM (Franklin 2010). Random Forests, especially are simple methods to put in place, allow for quantification of the variables importance in explaining variation, and offer intrinsic testing metrics. Neural networks could also be an interesting alternative. 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 Warblers species, as no appropriate information on their interaction is available to our knowledge. Joint species distribution models (JSDMs) might be an interesting way to encompass those, as they attempt to model species cooccurence, rather than the distribution of single species distributions (Pollock et al. 2014). A different taxonomic group and data datasets could also be used with more details on interactions could also be used, though having a method that can be applied to any taxonomic group would be more interesting. Yet, such an approach might prove to be beyond the scope of the present research.

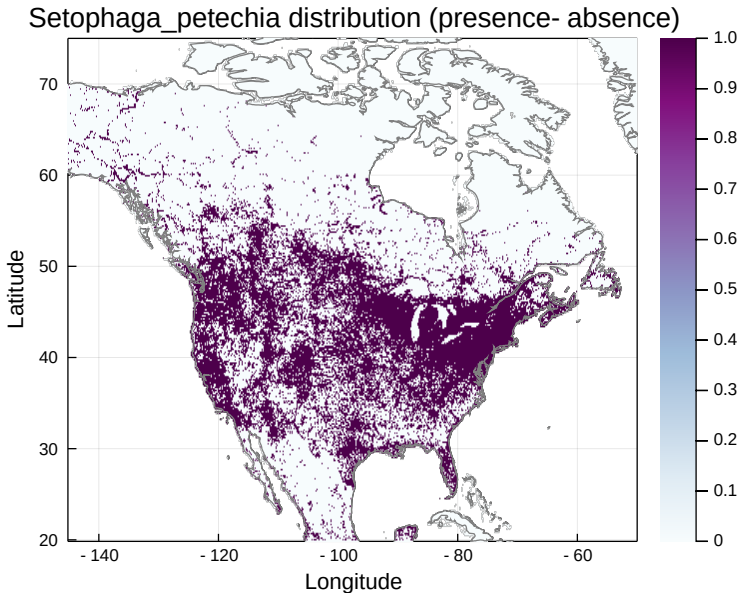
7. Climate change scenarios & temporal beta diversity

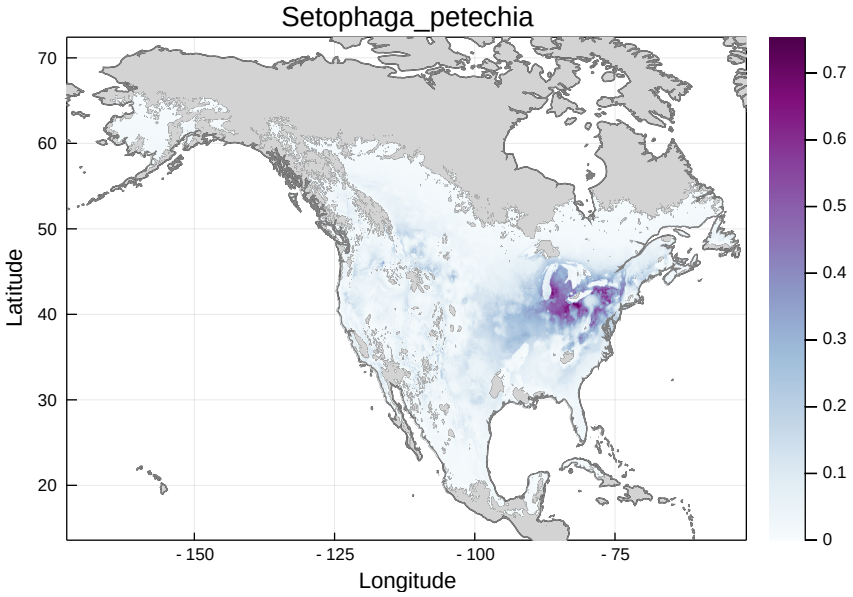
We aim to apply our method to environmental conditions from climate change scenarios, first to model community compositions after climate change on continuous scales through SDMs, and then to identify the sites where the community has changed in the most exceptional ways. This can be done through LCBD values, but also through temporal beta diversity indices (TBI) (Legendre 2019), which 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 each site compared to all other ones, TBI measure changes in community composition for a single site between two surveys, and can also be decomposed into species losses and gains. Moreover, TBI can be tested for significance using a permutation test. An approach similar to that of (Legendre and Condit 2019) would be most interesting to

follow: they first computed LCBD indices and compared the sites that were significant for two surveys 30 years apart, highlighting a swamp region where important changes seemed to have occurred, and then used TBI indices to confirm the sites with significant changes, decompose those into losses and gains and identify the species that had changed the most. Such an approach could be highly informative with our data, although the permutation tests and corrections to apply might cause problems given the number of sites that would be implied in our study. The possibility of using climate change scenarios in the SDMs also needs to be investigated in more details. We did not try to download nor find the appropriate data for now, but we found that the interpolated variables are sometimes different than those used in Worldclim 2.0. The SDM models and predictions might therefore be slightly different than those used for the LCBD calculations, 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, which could return highly informative results for conservation purposes.

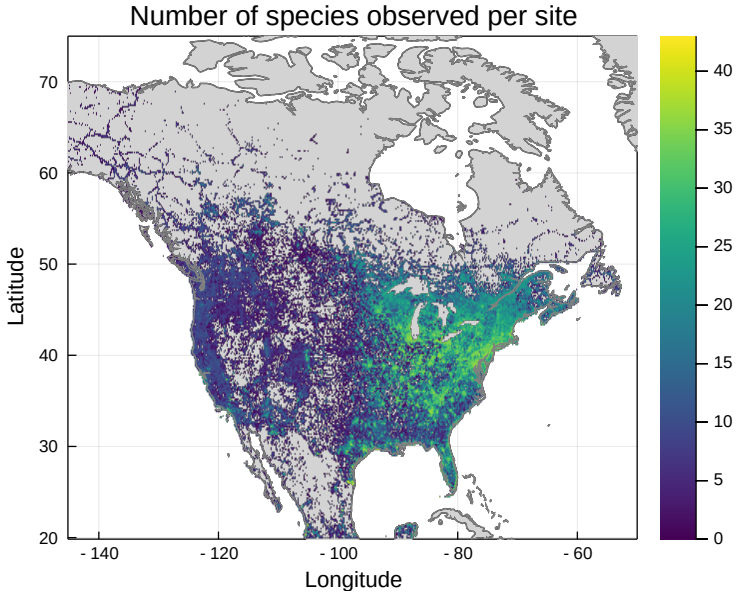
Preliminary Results

Our preliminary results mainly compare raw data statistics to prediction statistics. (Raw & SDM figures will be presented side-by-side)

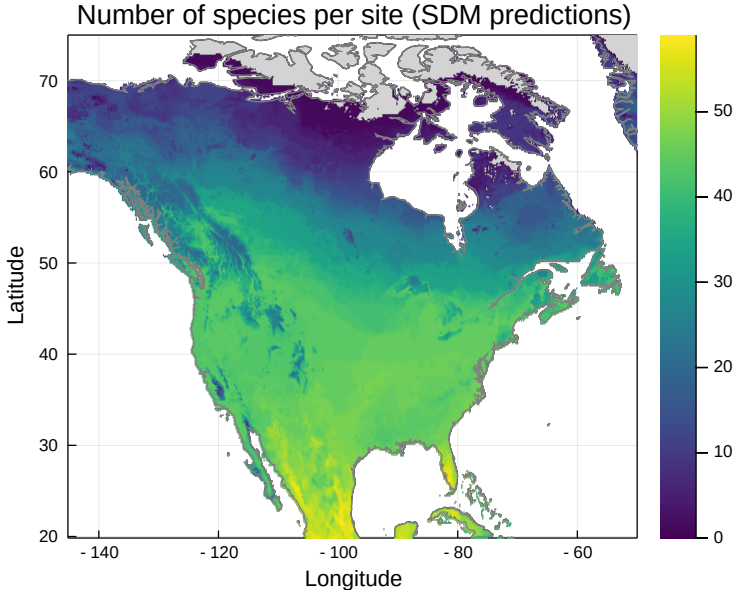



Setophaga_petechia
Latitude
Longitude
70
60
50
40
30
20
- 150
- 125
- 100
- 75
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

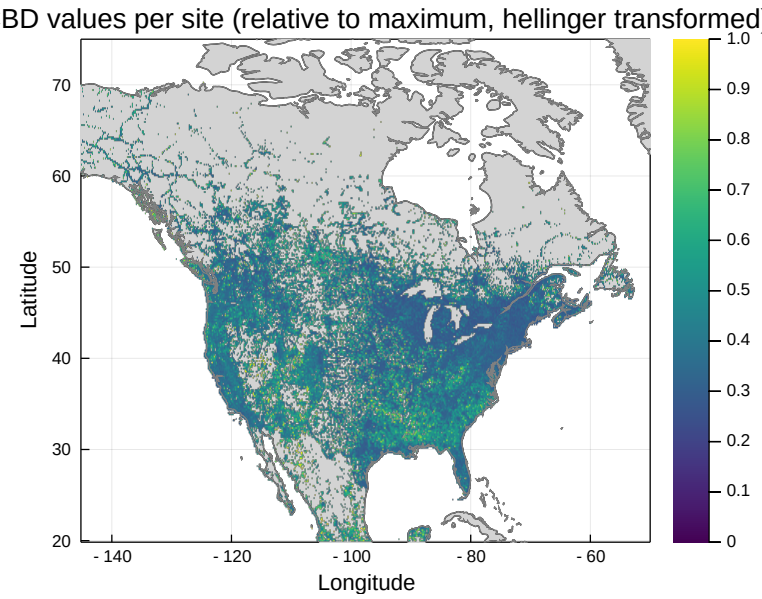
223


Number of species observed per site
Latitude
Longitude
70
60
50
40
30
20
- 140
- 120
- 100
- 80
- 60
40
35
30
25
20
15
10
5
0

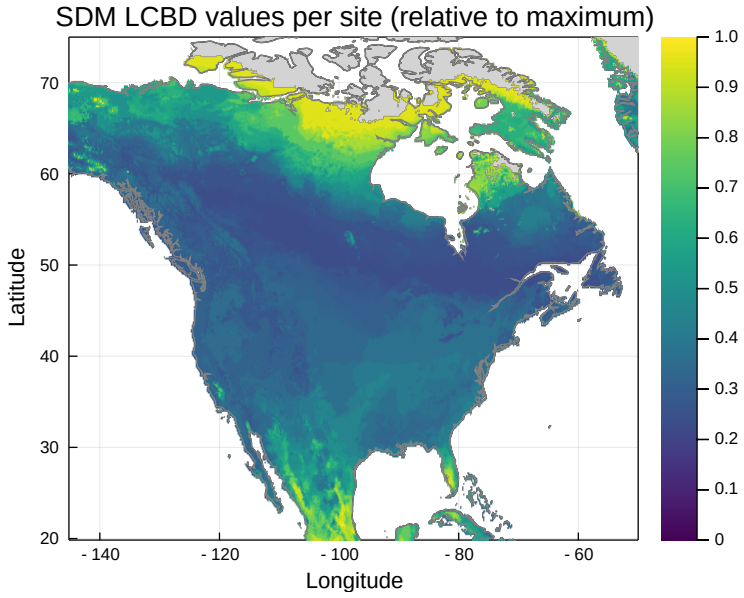
224


Number of species per site (SDM predictions)
Latitude
Longitude
70
60
50
40
30
20
- 140
- 120
- 100
- 80
- 60
50
40
30
20
10
0

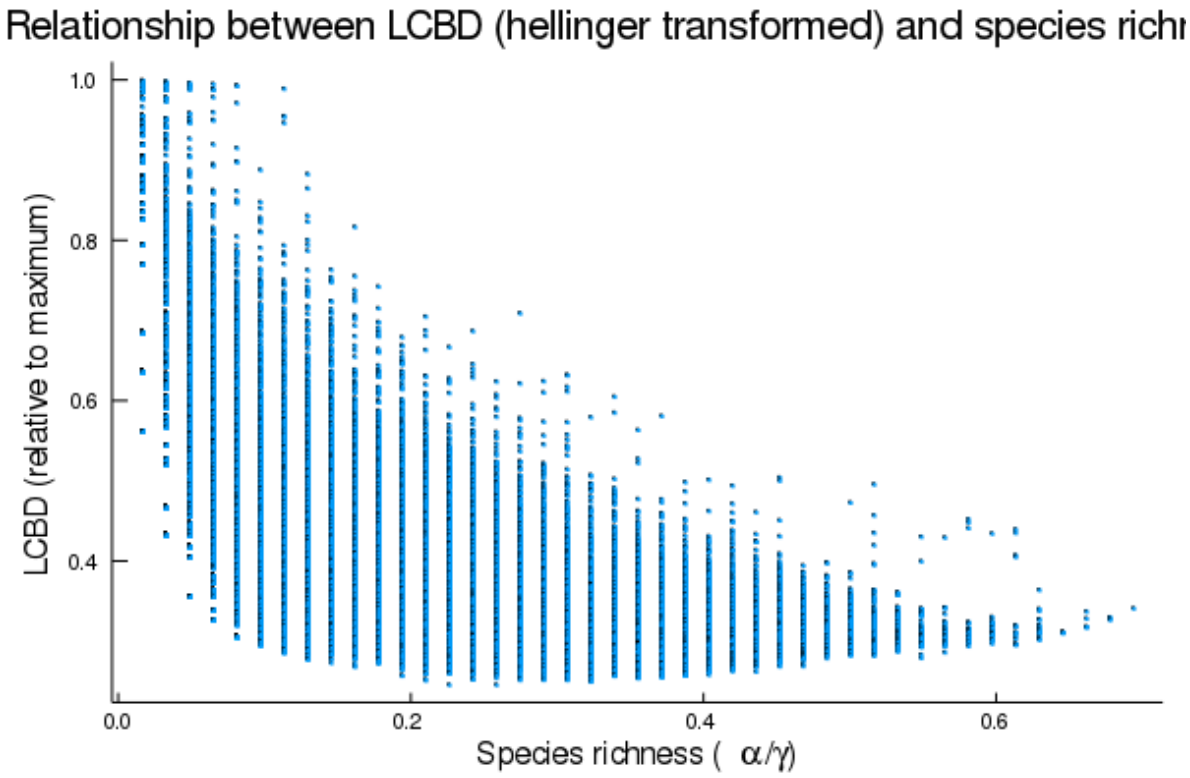
225


LCBD values per site (relative to maximum, hellinger transformed)
Latitude
70
60
50
40
30
20
- 140
- 120
- 100
- 80
- 60
Longitude
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

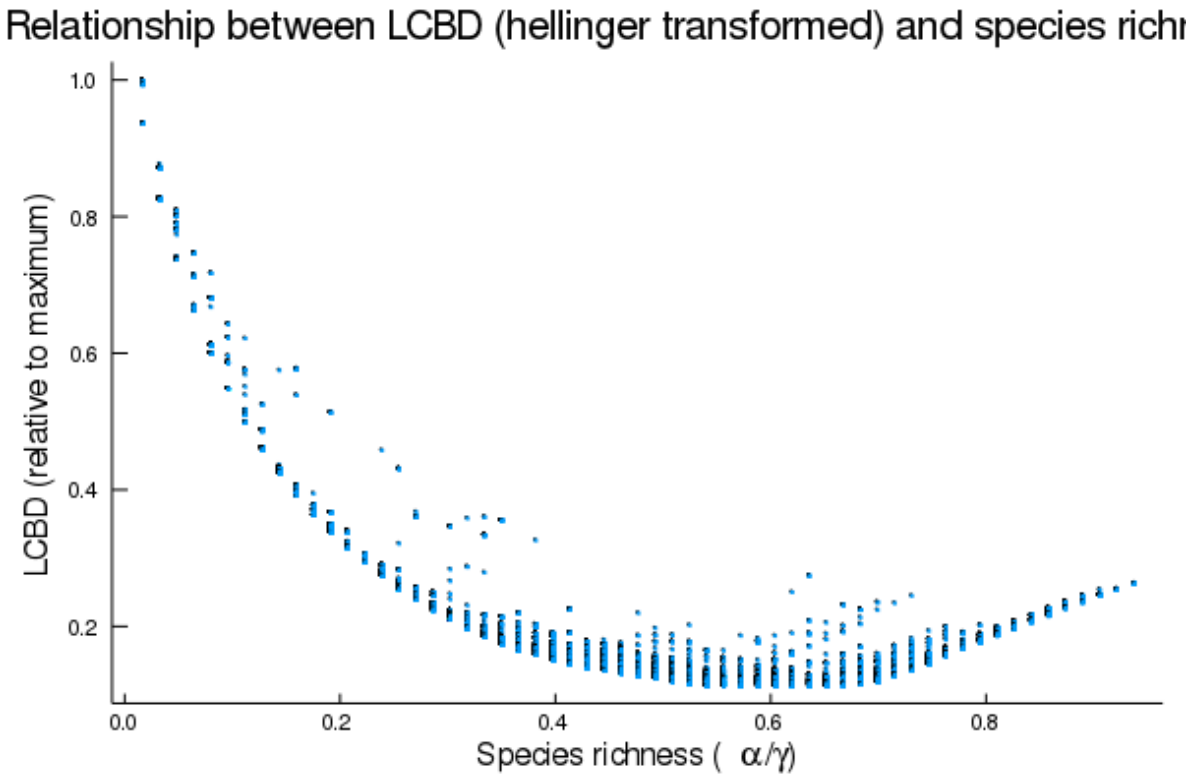
226


SDM LCBD values per site (relative to maximum)
Latitude
70
60
50
40
30
20
- 140
- 120
- 100
- 80
- 60
Longitude
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

227


Relationship between LCBD (hellinger transformed) and species richı
LCBD (relative to maximum)
1.0
0.8
0.6
0.4
0.0
0.2
0.4
0.6
Species richness (α/γ)

228


Relationship between LCBD (hellinger transformed) and species richı
LCBD (relative to maximum)
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
Species richness (α/γ)

229

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.

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.12180.

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.