

1

Université de Montréal

2

Spatially continuous identification of beta diversity hotspots using species distribution models

3

4

By

5

Gabriel Dansereau

6

20147609

7

Département de sciences biologiques

8

Faculté des arts et des sciences

9

Advisory Committee Meeting

10

November 26, 2019

¹¹ **Contents**

¹²	Abstract	2
¹³	Introduction	3
¹⁴	Methods	4
¹⁵	1. Data Collection	4
¹⁶	2. Data Manipulation	6
¹⁷	3. SDM – The BIOCLIM method	6
¹⁸	4. LCBD calculation	7
¹⁹	5. Prediction validity	8
²⁰	6. Alternative methods	8
²¹	7. Climate change scenarios & temporal beta diversity	9
²²	Preliminary Results	10
²³	References	15

²⁴ **List of Tables**

²⁵	1 WorldClim 2 climate variables used in the analyses	5
---------------	--	---

²⁶ **List of Figures**

²⁷	1 Single Species Distributions - Raw	11
²⁸	2 Single Species Distributions - SDM	11
²⁹	3 Species Richness - Raw	12
³⁰	4 Species Richness - SDM	12
³¹	5 LCBD values - Raw (transformed)	13
³²	6 LCBD values - SDM	13
³³	7 LCBD-richness relationship - Raw	14
³⁴	8 LCBD-richness relationship - SDM	15

³⁵ **Abstract**

³⁶ Beta diversity is an essential measure to describe the organization of biodiversity in space.
³⁷ The calculation of local contributions to beta diversity (LCBD), specifically, allows for the
³⁸ identification of sites with exceptional diversity within a region of interest, which is useful
³⁹ for both community ecology and conservation purposes. However, beta diversity implies a
⁴⁰ comparison among the sites of a given region, thus, its use is restricted to sites with known
⁴¹ species composition, and to discontinuous spatial scales. We therefore propose a method
⁴² to calculate LCBD indices on continuous scales for a whole region of interest, including
⁴³ unsampled sites. First, species distributions can be predicted on continuous scales using
⁴⁴ species distribution models (SDM). These models, such as the BIOCLIM method, use the
⁴⁵ environmental conditions at sampled sites to predict the presence or absence of each species at
⁴⁶ unsampled locations. Second, LCBD statistics can then be computed on the SDM predictions.
⁴⁷ We therefore show that it is possible to identify beta diversity hotspots on spatially continuous
⁴⁸ and extended scales. Our results confirm that LCBD values are related to species richness,
⁴⁹ and that species-poor sites contribute most to beta diversity.

50 **Introduction**

51 Beta diversity, defined as the variation in species composition among sites in a geographic
52 region of interest (Legendre, Borcard, and Peres-Neto 2005), is an essential measure to
53 describe the organization of biodiversity in space. Total beta diversity within a community
54 can be partitioned into local contributions to beta diversity (LCBD) (Legendre and De Cáceres
55 2013), which allows for the identification of sites with exceptional species composition,
56 hence exceptional biodiversity. Such a method is useful for both community ecology and
57 conservation biology, as it highlights sites that are most important for their research or
58 conservation values. However, LCBD calculation methods require complete information
59 on community composition, such as a community composition matrix Y , thus they are
60 inappropriate for partially sampled or unsampled sites. To our knowledge, these methods
61 have mostly been applied on community data from sampled sites, hence on discontinuous
62 spatial scales, e.g. at intervals along a river stream (Legendre and De Cáceres 2013). This
63 raises the following questions: 1) could LCBD indices be extended to continuous spatial
64 scales, and 2) could this provide novel ecological insights in poorly sampled regions? We
65 aim to answer these questions by combining the LCBD calculation methods with predictive
66 biogeography approaches, and suggest that this would allow for the identification of sites
67 with high conservation value in poorly sampled regions.

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

77 Appropriate data to expand measures of exceptional biodiversity through space is increasingly
78 available online. For instance, the Worldclim 2.0 database (Fick and Hijmans 2017) provides
79 interpolated climate data for global land areas at very high spatial resolution, and the eBird

80 platform (Sullivan et al. 2009) provides a growing citizen-contributed database of worldwide
81 bird observations. Both of these are commonly used in SDMs, and offer relevant information
82 on extended spatial scales. Hence, we believe that we could use them to predict community
83 composition and calculate LCBD indices on continuous spatial scales, and that the result
84 would be representative of the true community structure.

85 The predictive approach we suggest would be especially useful in poorly sampled regions,
86 or in regions with only sparse sampling. While it doesn't replace a full sampling within the
87 community, it does provide relevant ecological insights. For instance, the method could help
88 identify unsampled sites with potential conservation value which should be targeted as soon
89 as possible in future studies. We also believe that our method could also be combined with
90 IPCC climate change scenarios, which provide projections for climate variables, in a way that
91 would allow us to model beta diversity changes with climate change and to identify the sites
92 where the changes in the community will be most important. Again, this method would be
93 more relevant as an informative approach to suggest sites to prioritize for future conservation
94 and more structured research.

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

101 **Methods**

102 **1. Data Collection**

103 We decided to focus our analyses on bird species and collected the data available on eBird
104 for the Warblers family. The complete database contains nearly 600 million observations,
105 and presents two main advantages over other large scale datasets (Johnston et al. 2019): 1)
106 data is structured as checklist and users can explicitly specify their observations as "complete
107 checklists" when all detected species were reported, which allows to infer information on
108 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

¹²³ **2. Data Manipulation**

¹²⁴ WorldClim variables and eBird occurrence data are provided in different formats, so they
¹²⁵ require some manipulation to be combined together. WorldClim variables are provided in
¹²⁶ a 2-dimensional grid format, useful for large scale analyses and visualization, where each
¹²⁷ cell or pixel corresponds to the resolution of 10 arc-minutes. Each of the 19 variables forms a
¹²⁸ different grid. On the other hand, eBird records are occurrence-based, so each entry in the
¹²⁹ dataset corresponds to an observation of a single species at a given location. These entries can
¹³⁰ easily be matched to the 2-D grid format of the WorldClim variables through their spatial
¹³¹ coordinates, which we found more useful for large scale analyses and visualization. Hence,
¹³² for each species, we matched all occurrences in eBird to the grid format of the WorldClim
¹³³ variables, and later created a presence-absence community matrix Y , with the sites being the
¹³⁴ grid cells. We also applied the Hellinger transformation on the raw presence-absence data,
¹³⁵ although the most appropriate method remains to be determined, especially since the data
¹³⁶ has to be compared with the SDM predictions. All data manipulations and further analyses
¹³⁷ were realized in *Julia v1.2.0* (Bezanson et al. 2017) with the basic structure built around the
¹³⁸ soon-to-be-released `SimpleSDMLayers.jl` package.

¹³⁹ **3. SDM – The BIOCLIM method**

¹⁴⁰ We used the BIOCLIM method to predict species distributions. BIOCLIM, first introduced by
¹⁴¹ (Nix 1986), is considered as the classic “climate-envelope-model”, and is now available to
¹⁴² users through the `dismo` package in R (Hijmans et al. 2017). It has long been outperformed
¹⁴³ by other methods (Elith et al. 2006), but it is still commonly used for its simplistic approach

144 and ease of understanding, as well as its simple relation to niche theory (Booth et al. 2014;
145 Hijmans et al. 2017). It is also a method designed for presence-only data, which does not
146 require information on absences, nor take them into account if provided (as in our case).
147 Despite that, we chose this method for our preliminary analyses as it was easier to implement
148 and because we believe it to be sufficient for proof-of-concept. We discuss possible alternatives
149 in the “Alternative methods” section below.

150 Briefly, the BIOCLIM method defines species potential range as a multidimensional envi-
151 ronmental hypervolume bounded by the minimum and maximum values of all presences
152 (Franklin 2010). For each species, the algorithm establishes the percentile distribution of
153 the values of each environmental variables at the known locations of occurrences (Hijmans
154 et al. 2017). The environmental variables of all sites are then compared to those percentile
155 distributions and given scores between 0 (1st percentile) and 1 (100th percentile). The median
156 or 50th percentile is considered as the most suitable location and both tails (e.g. 10th and
157 90th percentile) are not distinguished, the values larger than 0.5 being subtracted from 1.
158 The minimum percentile score across all environmental variables is selected as the prediction
159 value for each site and multiplied by 2 so values are between 0 and 1 (Hijmans et al. 2017).
160 It should be noted that the limiting variable is thus not necessarily the same for all sites.
161 Values of 1 are rare, as it would mean a perfectly median site on all variables, and values
162 of 0 are frequent, since they are assigned whenever an environmental value is outside the
163 range of the observed values (Hijmans et al. 2017). Finally, before calculating richness or beta
164 diversity metrics, we transformed the predictions back to a presence-absence format, where
165 all predictions greater than one are considered as presence. This might tend to overestimate
166 species ranges and create some sort of border effect, but we believe the effects will be mitigated
167 given the spatial extent and coarse scale of our study.

168 4. LCBD calculation

169 We calculated the LCBD statistics through the total variance of the matrix Y for both the raw
170 data and SDM predictions. Legendre and De Cáceres (2013) showed that LCBD coefficients
171 can be calculated directly through the total variance of matrix Y , or through a matrix of
172 dissimilarities among sampling units. We chose the first approach, as it also allows to
173 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 be 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 cooccurrence, 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

233 follow: they first computed LCBD indices and compared the sites that were significant for
234 two surveys 30 years apart, highlighting a swamp region where important changes seemed
235 to have occurred, and then used TBI indices to confirm the sites with significant changes,
236 decompose those into losses and gains and identify the species that had changed the most.
237 Such an approach could be highly informative with our data, although the permutation tests
238 and corrections to apply might cause problems given the number of sites that would be
239 implied in our study. The possibility of using climate change scenarios in the SDMs also
240 needs to be investigated in more details. We did not try to download nor find the appropriate
241 data for now, but we found that the interpolated variables are sometimes different than those
242 used in Worldclim 2.0. The SDM models and predictions might therefore be slightly different
243 than those used for the LCBD calculations, and potentially less reliable. Nonetheless, we
244 believe it will be possible to do some kind of time analysis linking beta diversity, climate
245 change, and species distribution modelling, which could return highly informative results for
246 conservation purposes.

247 **Preliminary Results**

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

Setophaga_petechia distribution (presence- absence)

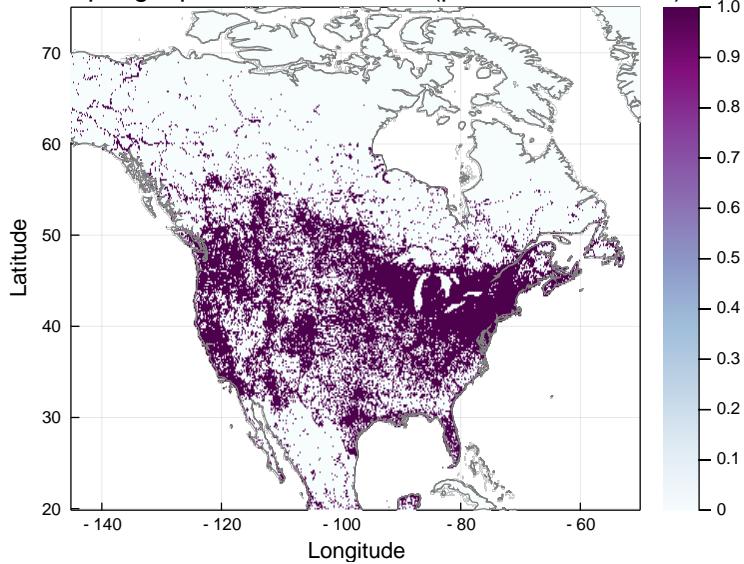


Figure 1: Single Species Distributions - Raw

Setophaga_petechia

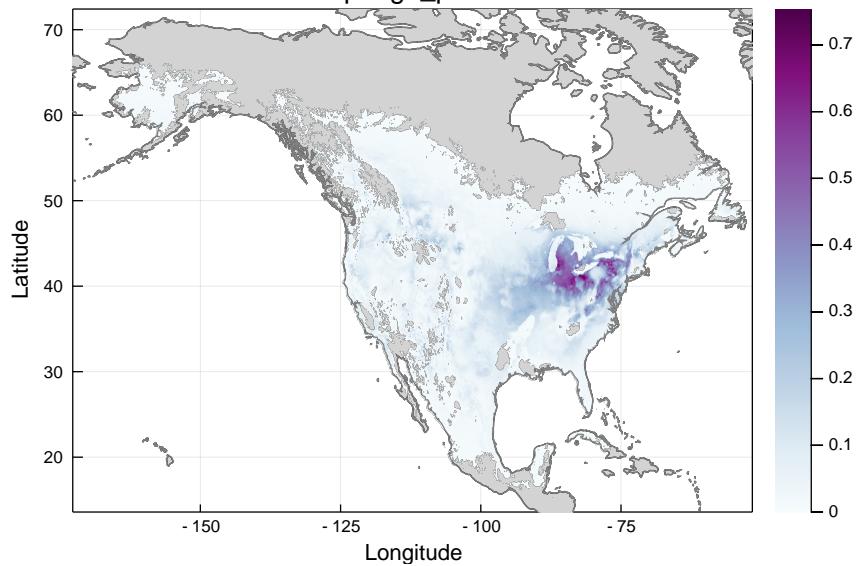


Figure 2: Single Species Distributions - SDM

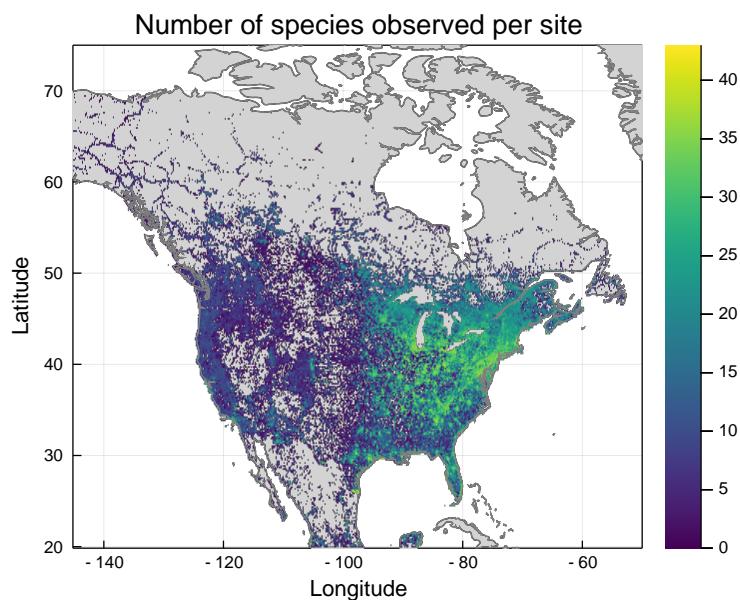


Figure 3: Species Richness - Raw

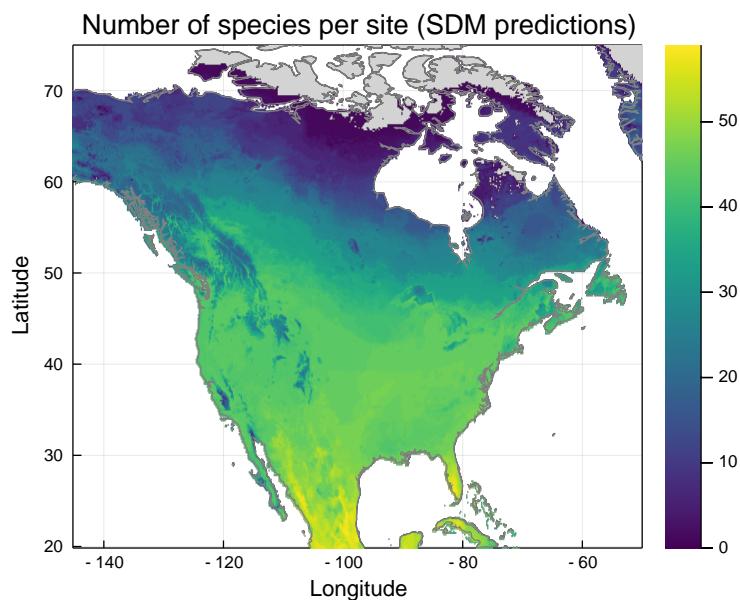


Figure 4: Species Richness - SDM

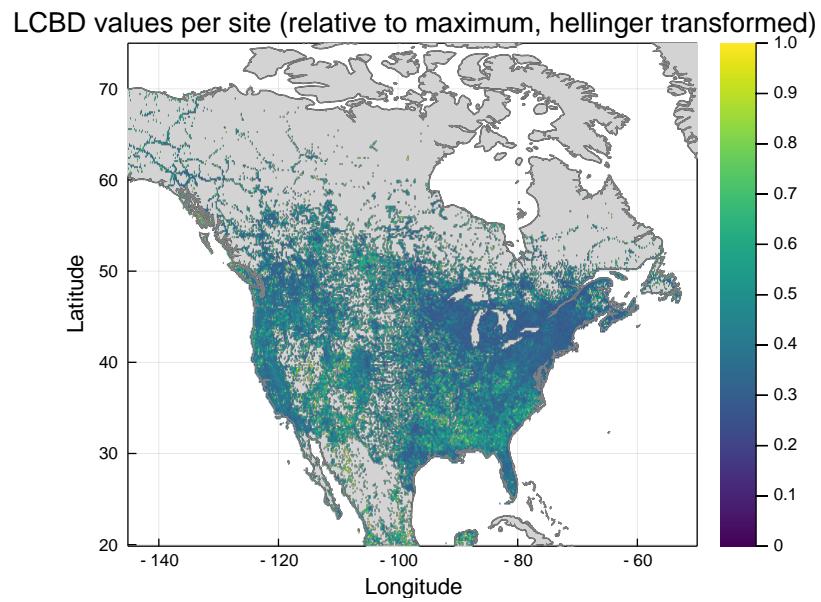


Figure 5: LCBD values - Raw (transformed)

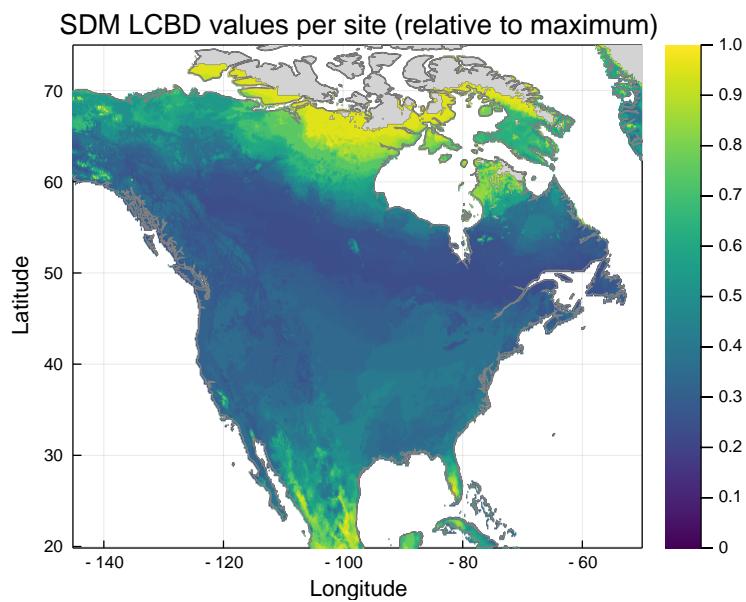


Figure 6: LCBD values - SDM

Relationship between LCBD (hellinger transformed) and species richness

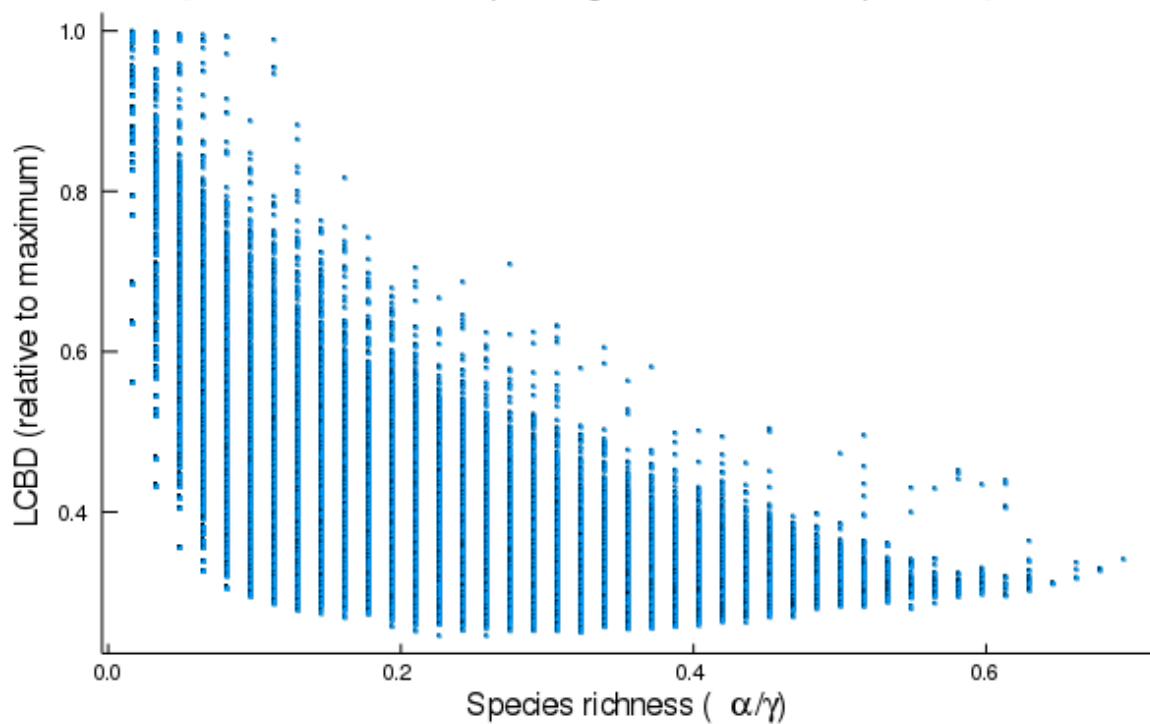


Figure 7: LCBD-richness relationship - Raw

Relationship between LCBD (hellinger transformed) and species richness

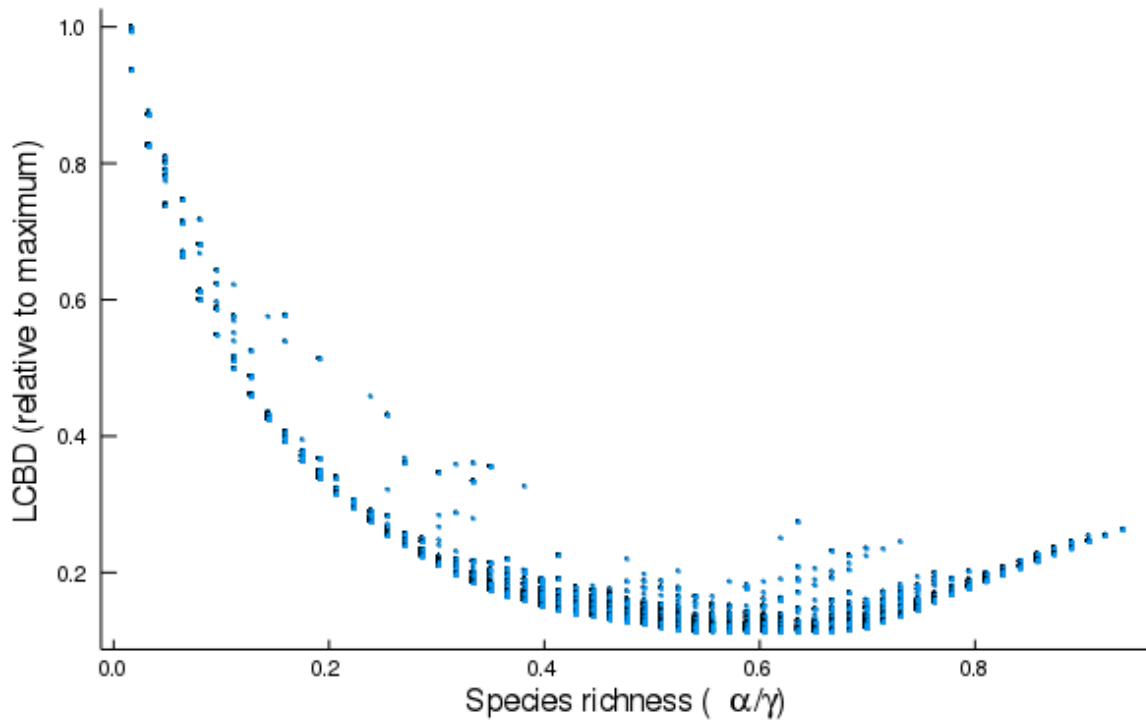


Figure 8: LCBD-richness relationship - SDM

250 References

- 251 Araújo, Miguel B., and A. Townsend Peterson. 2012. “Uses and Misuses of Bioclimatic
252 Envelope Modeling.” *Ecology* 93 (7): 1527–39. <https://doi.org/10.1890/11-1930.1>.
- 253 Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2017. “Julia: A Fresh
254 Approach to Numerical Computing.” *SIAM Review* 59 (1): 65–98. <https://doi.org/10.1137/141000671>.
- 255
- 256 Booth, Trevor H., Henry A. Nix, John R. Busby, and Michael F. Hutchinson. 2014. “BIOCLIM:
257 The First Species Distribution Modelling Package, Its Early Applications and Relevance to
258 Most Current MaxEnt Studies.” *Diversity and Distributions* 20 (1): 1–9. <https://doi.org/10.1111/ddi.12144>.
- 259
- 260 Elith, Jane, Catherine H. Graham, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine
261 Guisan, Robert J. Hijmans, et al. 2006. “Novel Methods Improve Prediction of Species’
262 Distributions from Occurrence Data.” *Ecography* 29 (2): 129–51. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>.
- 263

- 264 Fick, Stephen E., and Robert J. Hijmans. 2017. "WorldClim 2: New 1-Km Spatial Resolution
265 Climate Surfaces for Global Land Areas." *International Journal of Climatology* 37 (12):
266 4302–15. <https://doi.org/10.1002/joc.5086>.
- 267 Franklin, Janet. 2010. "Moving Beyond Static Species Distribution Models in Support of
268 Conservation Biogeography: Moving Beyond Static Species Distribution Models." *Diversity*
269 and *Distributions* 16 (3): 321–30. <https://doi.org/10.1111/j.1472-4642.2010.00641.x>.
- 270 Hijmans, Robert J., Steven Phillips, John Leathwick, and Jane Elith. 2017. *Dismo: Species*
271 *Distribution Modeling*. <https://CRAN.R-project.org/package=dismo>.
- 272 Johnston, A., W. M. Hochachka, M. E. Strimas-Mackey, V. Ruiz Gutierrez, O. J. Robinson,
273 E. T. Miller, T. Auer, S. T. Kelling, and D. Fink. 2019. "Best Practices for Making Reli-
274 able Inferences from Citizen Science Data: Case Study Using eBird to Estimate Species
275 Distributions." *bioRxiv*, March, 574392. <https://doi.org/10.1101/574392>.
- 276 Legendre, Pierre. 2019. "A Temporal Beta-Diversity Index to Identify Sites That Have
277 Changed in Exceptional Ways in Space-Time Surveys." *Ecology and Evolution* 9 (6): 3500–
278 3514. <https://doi.org/10.1002/ece3.4984>.
- 279 Legendre, Pierre, Daniel Borcard, and Pedro R. Peres-Neto. 2005. "Analyzing Beta Diver-
280 sity: Partitioning the Spatial Variation of Community Composition Data." *Ecological*
281 *Monographs* 75 (4): 435–50. <https://doi.org/10.1890/05-0549>.
- 282 Legendre, Pierre, and Richard Condit. 2019. "Spatial and Temporal Analysis of Beta Diversity
283 in the Barro Colorado Island Forest Dynamics Plot, Panama." *Forest Ecosystems* 6 (1): 7.
284 <https://doi.org/10.1186/s40663-019-0164-4>.
- 285 Legendre, Pierre, and Miquel De Cáceres. 2013. "Beta Diversity as the Variance of Community
286 Data: Dissimilarity Coefficients and Partitioning." *Ecology Letters* 16 (8): 951–63. [//doi.org/10.1111/ele.12141](https://doi.org/10.1111/ele.12141).
- 288 Nix, Henry A. 1986. "A Biogeographic Analysis of Australian Elapid Snakes." *Atlas of Elapid*
289 *Snakes of Australia* 7: 4–15.
- 290 Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. 2006. "Maximum Entropy
291 Modeling of Species Geographic Distributions." *Ecological Modelling* 190 (3): 231–59.

- 292 <https://doi.org/10.1016/j.ecolmodel.2005.03.026>.
- 293 Poisot, Timothée, Richard LaBrie, Erin Larson, Anastasia Rahlin, and Benno I. Simmons. 2019.
- 294 “Data-Based, Synthesis-Driven: Setting the Agenda for Computational Ecology.” *Ideas in*
- 295 *Ecology and Evolution* 12 (July). <https://doi.org/10.24908/iee.2019.12.2.e>.
- 296 Pollock, Laura J., Reid Tingley, William K. Morris, Nick Golding, Robert B. O’Hara, Kirsten
- 297 M. Parris, Peter A. Vesk, and Michael A. McCarthy. 2014. “Understanding Co-Occurrence
- 298 by Modelling Species Simultaneously with a Joint Species Distribution Model (JSDM).”
- 299 *Methods in Ecology and Evolution* 5 (5): 397–406. <https://doi.org/10.1111/2041-210X.12180>.
- 300
- 301 Sullivan, Brian L., Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and
- 302 Steve Kelling. 2009. “eBird: A Citizen-Based Bird Observation Network in the Biological
- 303 Sciences.” *Biological Conservation* 142 (10): 2282–92. <https://doi.org/10.1016/j.biocon.2009.05.006>.
- 304