Pre-methods

- Introduction
  - LCBD as useful measure for ecological uniqueness
  - Restriction to small scales & few sites, unlike BD itself
  - Restriction to known composition
  - Potential to fill-in gaps through SDMs
  - Poses question of applicability on large, continuous scales
  - Potential varying relationship between richness-LCBD, given dependence of BD on scale

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. To this day, LCBD indices have mostly been used on regional, smaller scales with relatively few sites. Furthermore, as beta diversity implies a comparison among the sites of a given region, their use is typically restricted to strictly sampled sites with known species composition, hence to discontinuous spatial scales. Here, we investigate the variation of LCBD indices on extended spatial scales including both species-poor and species-rich regions, and investigate their applicability for continuous scales and unsampled sites through the use of species distribution models (SDMs). To this aim, we used Bayesian additive regression trees (BARTs) to model species composition on continuous scales based on observation data from the eBird database.

Methods

- Methods
  - eBird & transformation to presence-absence
  - WorldClim & Copernicus, variable selection
  - BARTs with `embarcadero`
  - Richness, LCBD with hellinger transformation, relationship
  - Subareas (rich & poor)
  - 3 scales

We decided to focus our analyses on the Warblers family (*Parulidae*)in North America (Canada,

US, Mexico). We collected all occurrence data available in the eBird database, which represented roughly 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), which could be seen in our data (tbl. **??**). 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. **??**), 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. We also collected land cover data from the Copernicus Global land service (Buchhorn et al. 2019). These data consisted of 10 variables for the main land cover classes, each represented by their percentage of cover fraction. The Copernicus data is available at a 100m spatial resolution, finer than for the WorldClim data, hence we coarsened it it to the same resolution by averaging the cover fraction values.

Figures

Figure: Distribution of species richness in North America, defined as the number of Warblers species per site (10 arc-minutes pixels). The raw occurrence observations from eBird (left) and the SDM predictions from the single-species BART models (right) were both transformed into presence-absence data per species before calculating richness.

Figure: Distribution of the LCBD values in North America, calculated from the variance of the site-by-species community matrix Y and scaled to the maximum value observed. Occurrence observations from eBird (left) and single-species SDM predictions (right) were converted into presence-absence data per species, then the Hellinger transformation was applied before computing the LCBD indices.

Figure: Relationship between the species richness and the LCBD value of each site based on the occurrence observations from eBird (left) and the SDM predictions (right). LCBD values were scaled to the maximum value observed after applying Hellinger transformation.

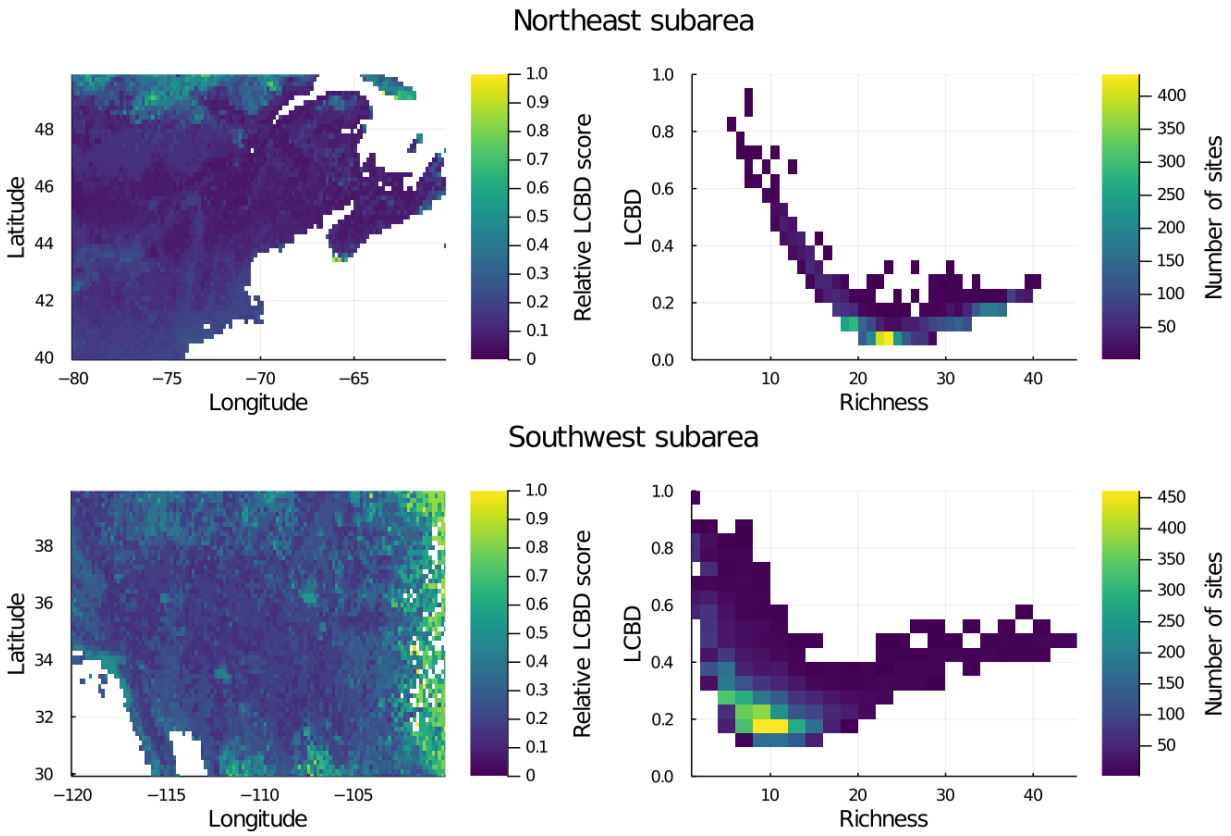

Figure 1: NE subareas

Figure: Comparison between a species-rich region (Northeast) and a species-poor one (Southwest) at a given scale, based on the SDM predictions.

This figure highlights the difference between species-rich and species-poor subareas. In a species-rich region, such as the Northeastern region of our study extent (North America), LCBD scores display a decreasing relationship with species richness. Hence, the sites with the highest LCBD values, the most unique ones in terms of species composition, are the species-poor sites, while the species-rich sites display low scores. Thus, our results show that the only way to stand out in such a region is by having few species. Since most sites comprise between 20 to 30 warblers species, the most species-rich ones with 40 species do not stand out and are not as exceptional as the ones with 10 species or fewer. The Southwest subarea, on the other hand, shows a different relationship. The sites with the highest LCBD values are once again the poorest ones in terms of species richness. However, since most sites only comprise around 10 species, the decreasing relationship with richness is initially much sharper, and displays a more important increase as richness reaches 20 species. This way, species-rich sites comprising around 40 species stand out more and are more exceptional in species-poor regions than in species-rich ones.

This result higlights an important aspect regarding the LCBD measure: contrary to previous findings, it does not simply decrease with species richness. In fact, the relationship with richness is not constant – it actually depends on the general profile of the region on which it is applied, and whether or not this region is species-poor or species-rich. A parabolic relationship was actually expected when the measure was introduced, as both extremes should normally stand out. An explanation for the previously observed results could be that extremely rich sites are just much less ecologically possible. It is unlikely that all species could be found in a single site given their different niche preferences, while poor sites are much more feasible. These sites will almost always contribute more to the variance, as measured by LCBD values.

Figure: Effect of scaling and full region extent size on the relationship between site richness and LCBD value. LCBD values are re-calculated at each scale based on the sites in this region only.

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

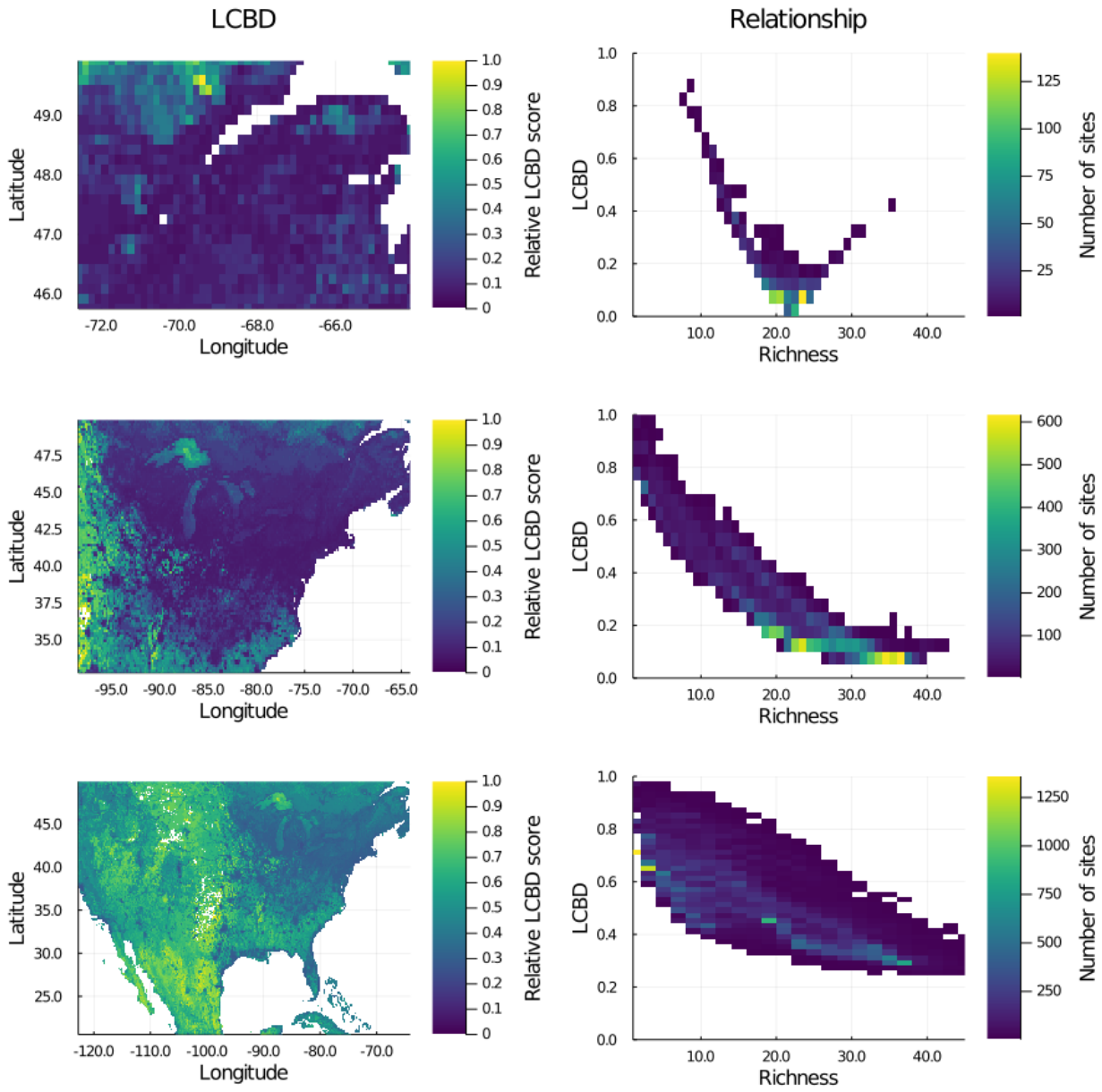

Figure 2: 3 scales

Current MaxEnt Studies." *Diversity and Distributions* 20 (1): 1–9. https://doi.org/10.1111/ddi.12144.

Buchhorn, Marcel, Bruno Smets, Luc Bertels, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, and Steffen Fritz. 2019. "Copernicus Global Land Service: Land Cover 100m: Epoch 2015: Globe." Zenodo. https://doi.org/10.5281/zenodo.3243509.

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.

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.

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.