

**Université de Montréal**

**Évaluation de l'unicité écologique à grande étendue spatiale à  
l'aide de modèles de répartition d'espèces**

par

**Gabriel Dansereau**

Département de sciences biologiques

Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)

en sciences biologiques

8 juillet 2021



# Université de Montréal

Faculté des arts et des sciences

---

Ce mémoire intitulé

## **Évaluation de l'unicité écologique à grande étendue spatiale à l'aide de modèles de répartition d'espèces**

présenté par

**Gabriel Dansereau**

a été évalué par un jury composé des personnes suivantes :

*Anne-Lise Routier*

---

(présidente)

*Timothée Poisot*

---

(directeur de recherche)

*Pierre Legendre*

---

(codirecteur)

*Élise Filotas*

---

(membre du jury)



## Résumé

---

La diversité bêta est une mesure essentielle pour décrire l'organisation de la biodiversité dans l'espace. Le calcul des contributions locales à la diversité bêta (LCBD), en particulier, permet d'identifier des sites à forte unicité écologique et montrant une diversité exceptionnelle au sein d'une région d'intérêt. Jusqu'à présent, l'utilisation des LCBD s'est restreinte à des échelles locales ou régionales avec un petit nombre de sites dont la composition en espèces est entièrement connue. Dans ce mémoire, j'ai examiné la variation des LCBD sur de grandes étendues spatiales, incluant des régions riches et pauvres en espèces. Pour ce faire, j'ai utilisé les arbres de régression additifs bayésiens (BARTs) pour prédire la répartition des parulines en Amérique du Nord à partir d'observations provenant de la base de données eBird. Mes résultats ont montré que la relation entre la richesse et les LCBD varie selon la région et l'étendue spatiale où celle-ci est mesurée et qu'elle est influencée par la proportion d'espèces rares dans les communautés. Les modèles de répartition d'espèces ont fourni des estimations fortement corrélées avec les données observées, mais montrant de l'autocorrélation spatiale. Les sites identifiés comme uniques sur de grandes étendues spatiales peuvent donc varier en fonction de la richesse régionale, de l'étendue totale et de la proportion d'espèces rares. Les modèles de répartition d'espèces peuvent être utilisés pour prédire l'unicité écologique sur de grandes étendues spatiales, ce qui pourrait permettre d'identifier des points chauds de diversité bêta et d'importantes cibles de conservation au sein de régions non échantillonnées.

**Mots clés :** diversité bêta, unicité écologique, contributions locales à la diversité bêta, modèles de répartition d'espèces, échelle spatiale étendue, eBird.



## Abstract

---

Beta diversity is an essential measure to describe the organization of biodiversity through space. The calculation of local contributions to beta diversity (LCBD), specifically, allows the identification of sites with high ecological uniqueness and exceptional diversity within a region of interest. To this day, LCBD indices have primarily been used on regional and smaller scales, with relatively few sites. Furthermore, their use is typically restricted to strictly sampled sites with known species composition, leading to gaps in spatial coverage on broad extents. Here, I investigated the variation of LCBD indices over broad spatial extents, including species-poor and species-rich regions, and examined their applicability for spatially continuous data and unsampled sites through species distribution modelling (SDM). To this aim, I used Bayesian additive regression trees (BARTs) to model warbler species composition on spatially continuous data based on observations recorded in the eBird database. My results showed that the relationship between richness and LCBD values varies according to the region and the spatial extent at which it is applied. It is also affected by the proportion of rare species in the community. Species distribution models provided highly correlated estimates with observed data, although spatially autocorrelated. Sites identified as unique over broad spatial extents may therefore vary according to regional richness, total extent size, and the proportion of rare species. Species distribution modelling can be used to predict ecological uniqueness over broad spatial extents, which could help identify beta diversity hotspots and important targets for conservation purposes in unsampled locations.

**Keywords:** beta diversity, ecological uniqueness, local contributions to beta diversity, species distribution modelling, broad spatial scale, eBird.





# Table des matières

---

<b>Résumé .....</b>	<b>5</b>
<b>Abstract .....</b>	<b>7</b>
<b>Liste des tableaux .....</b>	<b>13</b>
<b>Table des figures.....</b>	<b>15</b>
<b>Liste des sigles et des abréviations .....</b>	<b>19</b>
<b>Remerciements.....</b>	<b>21</b>
<b>Introduction.....</b>	<b>23</b>
0.1. Mise en contexte .....	23
0.2. Biodiversité et diversité bêta.....	24
0.2.1. Définition de la diversité bêta .....	25
0.2.2. Méthodes de calcul .....	26
0.2.3. Utilisation comme mesure spatialement explicite.....	28
0.3. Modèles prédictifs.....	30
0.3.1. Modèles de répartition d'espèces.....	30
0.3.2. Extension des modèles aux communautés .....	32
0.4. Données.....	33
0.4.1. Développement de nouvelles bases de données et de méthodes associées .....	33
0.4.2. Science citoyenne .....	35

0.4.3. Le problème des données d'absence .....	36
0.5. Enjeux spatiaux .....	37
0.5.1. Relation avec la richesse spécifique et la proportion d'espèces rares .....	37
0.5.2. Utilité en conservation .....	38
0.6. Objectifs de l'étude .....	39

## **First Article. Evaluating ecological uniqueness over broad spatial extents using species**

<b>distribution modelling .....</b>	<b>41</b>
1. Introduction .....	43
2. Methods .....	48
Occurrence data .....	48
Environmental data .....	48
Species distribution models .....	49
Quantification of ecological uniqueness .....	50
Comparison of observed and predicted values .....	51
Investigation of regional and scaling variation .....	52
Proportion of rare species .....	52
Software .....	53
3. Results .....	53
Species distribution models generate relevant community predictions .....	53
Uniqueness displays regional variation as two distinct profiles .....	54
Uniqueness depends on the scale on which it is measured .....	57
Uniqueness depends on the proportion of rare species .....	58
4. Discussion .....	61

5. Acknowledgments.....	64
<b>Conclusion .....</b>	<b>65</b>
<b>Bibliographie.....</b>	<b>69</b>
<b>Second Article. SimpleSDMLayers.jl and GBIF.jl: A Framework for Species</b>	
<b>Distribution Modelling in Julia.....</b>	<b>83</b>
Summary.....	83
Statement of need .....	84
Basic structure .....	85
Feature overview .....	87
Examples.....	89
Spatial operations .....	89
GBIF integration.....	90
Acknowledgements .....	91



## Liste des tableaux

---



## Table des figures

---

1	Comparison of species richness and LCBD scores from observed and predicted warbler occurrences in North America. Values were calculated for sites representing ten arcminute pixels. We measured species richness after converting the occurrence data from eBird (a) and the SDM predictions from our single-species BART models (b) to a presence-absence format per species. We applied the Hellinger transformation to the presence-absence data, then calculated the LCBD values from the variance of the community matrices, separately for the occurrence data (c) and the SDM predictions (d). LCBD values ranged between 1.444e-5 and 5.860e-5 for observation data and between 5.788e-6 and 1.706e-5 for SDM data. The total beta diversity was 0.608 for the observation data and 0.775 for the SDM data. Areas in light grey (not on the colour scale) represent mainland sites with environmental data but without any warbler species present. ....	55
2	Comparison between observed and predicted estimates of species richness (a) and ecological uniqueness (b). The difference values represent the estimate from the predicted data set minus the estimate from the observed data set. The difference values for richness ranged between -38 and 45 (a). LCBD values were recalculated for the same set of sites with observations in both data sets. Recalculated LCBD ranged between 1.450e-5 and 5.910e-5 for observation data and between 1.117e-5 and 5.132e-5 for SDM data. The difference values for LCBD scores ranged between -4,060e-5 and 3.297e-5 (b).....	56

- 3 Comparison of the regression residuals between the observed and predicted estimates of species richness (a) and ecological uniqueness (b). The estimate from the predicted data set was used as the dependent variable and the estimate from the observed data set as the independent variable. A negative binomial regression with a log link function was used for species richness, and a beta regression with a logit link function was used for uniqueness. The deviance residuals for richness ranged between -3,591 and 4,654 (a). LCBD values were recalculated for the same set of sites with observations in both data sets. The deviance residuals for LCBD scores ranged between -4,866 and 2,799 (b). ..... 57
- 4 Comparison between a species-rich region (Northeast, a) and a species-poor one (Southwest, b) based on the SDM predictions for warbler species in North America. The richness-LCBD relationship displayed contrasting profiles for the subregions according to their general richness. Total beta diversity was higher in the Southwest subregion than in the Northeast one. The left-side figures represent the LCBD scores for the assembled presence-absence predictions, calculated separately in each region after applying the Hellinger transformation. The right-side figures represent the decreasing relationship between LCBD values and species richness, with the number of sites in the bins of the 2-dimensional histogram. The vertical and horizontal dashed lines respectively represent the median richness and LCBD value in each region. LCBD values ranged between  $7.032\text{e-}5$  and  $1.333\text{e-}3$  for the Northeast subregion and between  $6.035\text{e-}5$  and  $5.236\text{e-}4$  for the Southwest one. .... 58
- 5 Effect of extent size on the relationship between site richness and LCBD values based on the SDM predictions for warbler species in North America. The relationship progressively broadens and displays more variance when scaling up while total beta diversity increases. The LCBD values were recalculated at each scale based on the sites in this region. LCBD values ranged between  $2.195\text{e-}4$  and  $5.209\text{e-}3$  at the finest scale,



	between 1.478e-4 and 3.500e-3 at the intermediate one, and between 1.179e-05 and 5.218e-05 at the broadest one. The vertical and horizontal dashed lines respectively represent the median richness and LCBD value in each region.....	59
6	Proportion of rare species in the ascending and descending portions of the LCBD-richness relationship for the Northeast (a) and Southwest (b) subregions. The left side figures show the geographic distribution of the sites from each group. Sites were assigned to the ascending portion if their species richness was higher than the richness of the site with the lowest LCBD value, which corresponds to the inflection point of the right side figures of Fig. 4, and in the descending portion otherwise. The right side figures are density plots of the proportion of rare species in each group. Species were classified as rare when they occurred in fewer than 40% of the sites in the subregion. The proportion of rare species was then calculated for every site.....	60
A1	Map of the average annual temperature data from WorldClim 2.1, accessed as a layer through SimpleSDMLayers.jl .....	88
A2	Latest belted kingfisher occurrences from the GBIF database displayed over the temperature data through the integration between SimpleSDMLayers.jl and GBIF.jl...	91



## Liste des sigles et des abréviations

---

BART	Arbres de régression additifs bayésiens ( <i>Bayesian additive regression trees</i> )
BN	Réseaux bayésiens ( <i>Bayesian networks</i> )
BRT	Arbres de régression fortifiés ( <i>Boosted regression trees</i> )
HMSC	Modélisation hiérarchique des communautés d'espèces ( <i>Hierarchical modelling of species communities</i> )
JSMD	Modèles conjoints de répartition d'espèces ( <i>Joint species distribution models</i> )
LCBD	Contributions locales à la diversité bêta ( <i>Local contributions to beta diversity</i> )

MCMC	Méthode de Monte-Carlo par chaînes de Markov ( <i>Markov Chain Monte-Carlo method</i> )
MEM	Modèles macro-écologiques ( <i>Macroecological models</i> )
RF	Forêts d'arbres décisionnels ( <i>Random Forests</i> )
SDM	Modèles de répartition d'espèces ( <i>Species distribution models</i> )
SESAM	Modélisation spatialement explicite des assemblages d'espèces ( <i>spatially explicit species assemblage modelling</i> )
S-SDM	Modèles de répartition d'espèces superposés ( <i>Stacked species distribution models</i> )
TSS	<i>True skill statistic</i>

## Remerciements

---

Mon parcours à la maîtrise a été une expérience des plus stimulantes. La très grande partie du mérite pour ces expériences positives revient à ceux et celles que j'ai eu la chance de côtoyer.

Merci en premier à mes co-directeurs, Timothée et Pierre. Merci pour vos conseils, merci de m'avoir guidé dans mon parcours, merci d'avoir toujours eu des suggestions pour stimuler mes réflexions en recherche, et merci d'avoir su me dire quand il fallait m'arrêter. J'ai énormément appris grâce à vous et je vous en serai toujours reconnaissant.

Ensuite, merci à tous mes collègues avec qui j'ai eu la chance de vivre cette expérience. Sans vous, mon parcours aurait été beaucoup plus difficile. Je suis ressorti grandi du contact avec chacun et chacune d'entre vous. Merci à Daphnée, Eva et Mathilde pour vos conseils, votre expérience et votre présence très précieuse au début de ma maîtrise. Merci à Francis et Gracielle pour les nombreux projets stimulants, loufoques parfois, complexes à d'autres moments, mais toujours agréables en votre compagnie. Merci à tous les collègues du labo, Andréanne, Fares, Kiri, Marie-Andrée, Miléna, Norma, Philippe, Salomé, Sandrine, Tanya et Valentine.

Merci à Andrew d'avoir été tel un guide pour un voyageur intergalactique en recherche. Discuter avec toi donnait parfois l'impression d'aller jusqu'à la fin du monde, sur des sujets aussi variés que la vie, l'univers et tout le reste. Merci cependant pour tous les poissons, et sache que tes conseils, loin d'être généralement inoffensifs, ont toujours été utiles.

Merci au FRQNT et à BIOS<sup>2</sup>, dont le soutien financier a permis la réalisation de ces travaux.

Merci finalement à mes parents pour leur éternel soutien. Merci d'être présents dans tous les moments, difficiles comme heureux.



# Introduction

---

## 0.1. Mise en contexte

L'identification des zones clés de biodiversité est l'une des priorités pour la conservation et la gestion des aires protégées. En particulier, il y a actuellement un besoin de développer des méthodes permettant d'identifier les sites les plus importants pour la biodiversité de façon efficace sur de grandes étendues spatiales. Or, identifier de tels endroits implique plusieurs questions complexes. En premier, il est nécessaire de définir ce que constituent des zones clés de biodiversité. Plusieurs définitions et plusieurs mesures ont été suggérées à ce sujet, mais elles varient généralement quant à l'étendue spatiale ou aux régions ciblées. Ensuite, au-delà de la définition de la biodiversité, il est nécessaire de trouver des données qui permettent d'évaluer avec justesse le caractère unique ou exceptionnel de la biodiversité à des sites donnés. La récolte de données en écologie est parfois difficile à réaliser à certains endroits, notamment en région éloignée. Les connaissances actuelles des différents milieux ne sont pas équivalentes non plus, alors que certains endroits plus proches des villes ou d'intérêt écologique particulier sont beaucoup mieux connus. Lorsque nécessaire, les observations directes peuvent parfois être remplacées par des prédictions réalisées à partir de données plus générales. Par contre, une panoplie de méthodes prédictives existent et la plupart d'entre elles n'ont pas été évaluées spécifiquement avec certaines mesures de biodiversité. Finalement, il est également nécessaire d'adapter à la fois les mesures de biodiversité et les méthodes prédictives aux grandes étendues spatiales. La biodiversité varie parfois différemment en fonction des échelles et il en est de même quant à la performance des mesures. Intégrer le tout peut donc s'avérer complexe et implique d'avoir une compréhension développée des définitions de la

biodiversité, des données et des méthodes disponibles, ainsi que des facteurs pouvant influencer la biodiversité en fonction des échelles spatiales.

Dans mon mémoire, je me suis intéressé à cette question en cherchant à vérifier l'applicabilité d'une mesure donnée, celle des contributions locales à la diversité bêta, pour identifier les zones de biodiversité exceptionnelle à grande étendue spatiale. De plus, j'ai cherché à vérifier si cette méthode pouvait être appliquée à des prédictions de la répartition des espèces produites à partir de données provenant de grandes bases de données citoyennes. Mon mémoire est donc divisé en trois sections. La première comporte une mise en contexte, ainsi qu'une revue de littérature présentant les concepts pertinents. La seconde partie consiste en un article scientifique présentant les résultats de mes travaux et analyses. Celui-ci sera soumis sous peu à la revue *Global Ecology and Biogeography*. La troisième partie consiste en un retour sur les résultats, en lien avec la mise en contexte présentée dans la première section. Enfin, mes travaux d'analyse nous ont amené, mon directeur et moi, à développer la librairie de fonctions `SimpleSDMLayers.jl`, au sujet de laquelle nous avons écrit un court article scientifique. Celui-ci a été publiée dans la revue *Journal of Open Source Software*. J'ai ajouté cet article en annexe, puisqu'il s'agit de travaux reliés à mon mémoire, mais il n'en constitue toutefois pas l'une des parties principales.

## 0.2. Biodiversité et diversité bêta

La biodiversité peut difficilement être séparée de sa dimension spatiale vu la diversité des espèces en chaque endroit donné sur Terre. La diversité bêta, soit la variation dans la composition en espèces entre les sites d'une région géographique d'intérêt (Legendre et al., 2005), est donc une mesure essentielle de l'organisation de la biodiversité dans l'espace. En écologie des communautés, l'intérêt pour celle-ci est d'autant plus grand, puisque la variation spatiale dans la composition en espèces permet de tester des hypothèses portant sur les processus qui génèrent et maintiennent la biodiversité dans les écosystèmes (Legendre et De Cáceres, 2013).

Dans le cadre du présent mémoire, trois phases importantes sont à retenir du développement du concept de diversité bêta, soit une première phase portant sur la définition de la diversité bêta



même, une deuxième sur le développement de différentes méthodes pour la calculer et une troisième sur son utilisation comme mesure spatialement explicite pour évaluer l'unicité écologique de sites spécifiques. Ainsi, depuis les premières formulations des composantes de la diversité des espèces par Whittaker (1960), l'attention s'est progressivement tournée vers le partitionnement de ces composantes, menant entre autres à la formulation d'une mesure spatialement explicite par Legendre et De Cáceres (2013), puis à l'utilisation de celle-ci pour évaluer l'unicité écologique, notamment pour un très grand nombre de sites (Niskanen et al., 2017) ou même sur des prédictions de la répartition des espèces (Vasconcelos et al., 2018). Dans cette section, j'effectuerai donc une revue du développement de ces trois phases en lien avec l'évaluation de l'unicité écologique à grande échelle spatiale, ce qui constitue l'objectif de mon mémoire.

### 0.2.1. Définition de la diversité bêta

Whittaker (1960) a détaillé trois composantes de la diversité des espèces au sein des communautés écologiques : 1) la diversité alpha, soit la richesse en espèces d'un site ou d'une communauté donnée, 2) la diversité bêta, qui représente le degré de différenciation dans la composition des communautés au sein d'un environnement (ou d'un gradient), et 3) la diversité gamma, soit la richesse en espèces des communautés d'un environnement (ou d'un ensemble de communautés). La diversité gamma est donc le résultat et la conséquence à la fois des diversités alpha et bêta.

Sous cette formulation initiale, la diversité bêta peut être mesurée comme  $\beta = \gamma / \bar{\alpha}$ , soit le ratio entre la diversité gamma et la diversité alpha moyenne des sites d'un échantillon, autrement dit le ratio entre le nombre d'espèces total et le nombre d'espèces moyen (Whittaker, 1960, 1972). Il s'agit donc d'une approche dite multiplicative, puisque  $\alpha$  et  $\beta$  peuvent être multipliées pour obtenir  $\gamma$ . La diversité bêta peut également être mesurée à partir de mesures de similarité d'échantillons, comme le coefficient de communauté, le pourcentage de similarité ou une distance statistique (Whittaker, 1972). De cette façon, la diversité bêta représente une seule valeur pour l'ensemble des sites visés, plutôt qu'une mesure pour chacun d'entre eux. Cette mesure est donc utile pour comparer des ensembles de sites, mais pas pour analyser la répartition de la variation entre les sites eux-mêmes.

Suite à cette formulation par Whittaker, la diversité bêta a été utilisée et mesurée de différentes façons, menant Koleff et al. (2003) à dire qu'une nouvelle mesure est dérivée pour chaque nouvelle utilisation du concept. Selon Vellend (2001) et Anderson et al. (2011), deux concepts ont cependant été associés à la diversité bêta au point d'être confondus et doivent être distingués : la variation dans la composition en espèces indépendamment de la position spatiale et le renouvellement en espèces (*species turnover*) le long de gradients spatiaux ou environnementaux. La première est une mesure non directionnelle portant simplement sur la variation, alors que la deuxième est une mesure directionnelle impliquant l'existence d'une certaine structure entre les parcelles. En révisant les mesures de diversité bêta pouvant être utilisées sur des données de présence-absence, Koleff et al. (2003) ont soulevé deux distinctions fondamentales similaires, soit entre les mesures au sens large (*broad sense measures*), utilisées pour les gradients de richesse, et les mesures au sens étroit (*narrow sense measures*), axées sur les différences de composition indépendantes des gradients (et particulièrement sur les espèces partagées entre les sites sans égard aux espèces qui diffèrent, d'où le qualificatif étroit).

### **0.2.2. Méthodes de calcul**

L'une des méthodes les plus fréquemment utilisées pour le calcul de la diversité bêta est celle du calcul de dissimilarité entre paires de sites, inspirée de la deuxième approche de Whittaker (1972). Or, cette approche fournit plutôt une valeur de comparaison entre deux sites, et non une valeur pour un ensemble de sites au sein d'une région. Pour obtenir une valeur générale représentant la diversité bêta pour la région, une approche commune est de calculer les dissimilarités par paires de sites, puis de calculer la moyenne (Anderson et al., 2011 ; Anderson et al., 2006). Baselga (2013) a cependant critiqué cette approche, montrant qu'elle ne tient pas correctement compte de la co-occurrence entre les sites, et a plutôt suggéré de la remplacer par une mesure de dissimilarité de sites multiples. L'approche de Baselga (2013) offre également un avantage comparativement à l'approche multiplicative de Whittaker (1960), car, en plus de convenir pour mesurer l'hétérogénéité entre plus de deux sites, elle permet également de distinguer le remplacement d'espèces de leur emboîtement. Le remplacement est analogue au renouvellement discuté par Anderson et al.

(2011), alors que l'emboîtement survient lorsque les assemblages les plus pauvres sont des sous-ensembles des sites les plus riches et mesure donc une perte d'espèces entre les sites (Baselga, 2010, 2013)

Une autre approche utilisée pour calculer la diversité bêta est l'approche additive (Lande, 1996 ; Veech et al., 2002), où les diversités alpha et bêta sont reliées à la diversité gamma par un facteur additif, et non multiplicatif comme dans l'approche originale de Whittaker (1960). La formulation de cette approche est donc  $D_T = D_{among} + \bar{D}_{within}$ , où  $D_T$  représente  $\gamma$ , mesuré comme la fréquence moyenne pondérée des espèces au sein d'une communauté,  $D_{among}$  représente  $\beta$  et  $\bar{D}_{within}$  représente  $\alpha$ . Cette formulation permet d'exprimer les trois diversités selon les mêmes unités, permettant des comparaisons directes entre elles (Lande, 1996).

Legendre et al. (2005) ont quant à eux légèrement reformulé la définition originale de Whittaker pour présenter une autre forme de calcul. Selon eux, la diversité bêta est la variation de la composition en espèces entre les sites d'une région géographique d'intérêt. En suivant cette définition, la variance d'une matrice de composition des communautés est une mesure juste de la diversité bêta, dont la variation spatiale peut être partitionnée en composantes environnementales et spatiales par partitionnement canonique (Legendre et al., 2005). Bâtissant sur cette approche, Legendre et De Cáceres (2013) ont ensuite montré que la diversité bêta peut être calculée de deux façons, soit par le calcul de la somme des carrés de la matrice des communautés ou par une matrice de dissimilarité.

Leur formulation se résume comme suit.  $Y$  représente la matrice de communautés contenant les valeurs de présence-absence ou d'abondance.  $Y$  est de dimensions  $n$  rangées par  $p$  colonnes, où  $n$  est le nombre de sites (ou d'unités échantillonnage) et  $p$  est le nombre d'espèces.  $i$  et  $j$  représentent les indices pour les sites et les espèces, respectivement, de sorte que  $y_{ij}$  représente les valeurs individuelles de la matrice  $Y$ .

Sous sa forme basée sur la somme des carrés, la diversité bêta totale peut être calculée comme :

$$s_{ij} = (y_{ij} - \bar{y}_j)^2$$

$$SS_{Total} = \sum_{i=1}^n \sum_{j=1}^p s_{ij}$$

$$BD_{Total} = Var(Y) = SS_{tot}/(n - 1) \quad (0.2.1)$$

$BD_{Total}$  représente donc un estimé de la variance non biaisé en fonction du nombre de sites, comparable entre régions. Avant ce calcul, les données d'abondance ou de présence-absence doivent toutefois être transformées de façon appropriée selon l'une des transformations suggérées par Legendre et De Cáceres (2013), par exemple la transformation de Hellinger. Cette approche remplit un critère important suggéré par Ellison (2010), soit de développer une formulation de la diversité bêta indépendante des diversités alpha et gamma, comme dans la formulation originale de Whittaker. De plus, cette mesure offre l'avantage de pouvoir partitionner la variation pour tester des hypothèses sur l'origine et le maintien de la diversité bêta au sein des écosystèmes (Legendre et De Cáceres, 2013).

### 0.2.3. Utilisation comme mesure spatialement explicite

Un aspect important de la mesure suggérée par Legendre et De Cáceres (2013) est qu'elle permet de dériver une mesure pour évaluer l'unicité écologique pour des sites précis, donc de façon spatialement explicite. Ainsi, la diversité bêta totale au sein d'une communauté peut, lorsque calculée comme la variance de la matrice de communautés, être décomposée en contributions locales à la diversité bêta (*local contributions to beta diversity*, LCBD), où chaque site obtient sa propre valeur. Cela permet d'identifier les sites possédant une composition en espèces exceptionnelle, autrement dit une biodiversité unique. Les LCBD sont calculées comme suit :

$$SS_i = \sum_{j=1}^p s_{ij}$$

$$LCBD_i = SS_i/SS_{Total} \quad (0.2.2)$$

Pour illustrer leur utilité, Legendre et De Cáceres (2013) ont calculé les LCBD pour montrer les sites les plus uniques parmi des communautés de poissons échantillonnées à intervalles le long

d'une rivière. Plusieurs études ont donc repris cette mesure des LCBd pour évaluer l'unicité écologique de sites précis, généralement dans un contexte semblable à l'étude originale. La plupart d'entre elles l'ont utilisée à échelle locale, donc sur des étendues spatiales restreintes, et sur un petit nombre de sites (da Silva et Hernández, 2014 ; Heino et al., 2017 ; Heino et Grönroos, 2017). Quelques études ont utilisé la mesure des LCBd sur de plus grandes étendues spatiales, donc comportant potentiellement une plus forte hétérogénéité spatiale, mais ces études comportaient un nombre de sites encore assez faible (Poisot et al., 2017 ; Taranu et al., 2020 ; Yang et al., 2015). Quelques études récentes l'ont utilisée sur des données arrangées en grille couvrant uniformément le territoire (D'Antracoli et al., 2020 ; Legendre et Condit, 2019 ; Tan et al., 2017 ; Tan et al., 2019). Cependant, celles-ci portaient également sur des échelles spatiales restreintes. Ainsi, dans la plupart des cas, la mesure des LCBd est utilisée sur de petites étendues spatiales et sur un petit nombre de sites.

Un enjeu potentiel pouvant être soulevé pour l'utilisation de la mesure des LCBd est le besoin de données appropriées. Les exemples précédents montrent l'utilité de la mesure des LCBd pour évaluer l'unicité écologique dans différentes situations, y compris sur de grandes étendues spatiales. Par contre, ces études sur de grandes étendues n'ont pas porté sur un grand nombre de sites. Une raison potentielle est le manque de données appropriées à une telle situation, puisque le calcul des LCBd nécessite une matrice de communautés complète, donc que la composition en espèces de chaque site soit connue précisément. Or, il est difficile de connaître la composition précise des communautés en couvrant à la fois une grande étendue spatiale et un grand nombre de sites.

Deux études récentes ont cependant développé de nouvelles approches prédictives qui pourraient ouvrir la voie à de nouvelles utilisations des LCBd sur de plus grandes étendues spatiales. En premier, Niskanen et al. (2017) ont utilisé la mesure sur un très grand nombre de sites (plus de 25 000) et sur des données arrangées en grille. Pour ce faire, ils ont utilisé des modèles prédictifs pour prédire les valeurs de LCBd et de trois autres mesures de diversité (des mesures alternatives pour identifier des sites importants en conservation) directement en fonction des conditions environnementales. Ensuite, Vasconcelos et al. (2018) ont de leur côté utilisé des modèles pour prédire la niche écologique des espèces en fonction des conditions climatiques (actuelles et suivant des

scénarios de changements climatiques), puis ont calculé les LCBD sur les communautés prédites. Cette approche s'apparente à l'utilisation originale de la mesure, puisqu'elle implique une matrice de communautés comportant les informations sur la présence ou l'absence des espèces aux différents sites. Par contre, leur utilisation s'est restreinte à la forêt Atlantique et au Cerrado au Brésil, ainsi qu'à environ 20 000 occurrences disponibles pour leur modèle d'études (les anoures). Or, avec les développements récents des bases de données massives, il existe des espèces et des bases de données pour lesquelles nous possédons beaucoup plus d'occurrences sur des étendues spatiales encore plus grandes, dont il serait intéressant de tirer parti. Je suggère donc de s'inspirer de leur démarche et de passer à un autre niveau, de façon à évaluer l'unicité écologique sur de plus grandes étendues spatiales, tout en cherchant à comprendre ce que la mesure des LCBD indique sur celles-ci.

### **0.3. Modèles prédictifs**

Les approches prédictives utilisées par Niskanen et al. (2017) et Vasconcelos et al. (2018) sont prometteuses et constituent les premières utilisations des LCBD et de l'unicité écologique dans une démarche clairement prédictive et spatialement explicite. Dans la prochaine section, je décrirai les types de modèles prédictifs qui pourraient être utilisés avec les LCBD, particulièrement en vue d'une utilisation sur de grandes étendues spatiales. Deux niveaux de modélisation sont importants pour les LCBD, soit un premier niveau portant sur les espèces seules et un deuxième portant sur les communautés sur lesquelles les LCBD peuvent être calculées.

#### **0.3.1. Modèles de répartition d'espèces**

Le type de modèles utilisés par Vasconcelos et al. (2018) fait partie de la grande famille des modèles de répartition d'espèces (*species distribution models*, ci-après SDM) (Guisan et Thuiller, 2005), qui servent notamment à prédire la répartition des espèces en fonction des conditions environnementales et d'observations déjà réalisées. Les assises théoriques derrière ces types de modèles remontent aux premières formulations de la niche écologique (Grinnell, 1917a, 1917b, 1924). Ils

reposent également sur l'idée de l'hypervolume de Hutchinson (1957, 1959) selon laquelle les tolérances environnementales d'une espèce forment un hypervolume au sein duquel la présence de l'espèce est possible. L'un des premiers SDM, le modèle d'enveloppe climatique BIOCLIM (Nix, 1986), illustre particulièrement bien cette relation avec les concepts de niche et d'hypervolume. Selon le modèle, la répartition potentielle des espèces devrait être contrainte au sein de l'étendue des conditions bioclimatiques où des observations ont été réalisées (Booth et al., 2014 ; Franklin, 2010). Le modèle classe les sites observés selon leur rang centile pour chaque variable bioclimatique fournie, puis attribue le score le plus élevé à la médiane, considérée comme l'endroit où les conditions conviennent le mieux à l'espèce (Hijmans et al., 2017). La valeur minimale parmi toutes les variables environnementales est ensuite interprétée comme la probabilité d'occurrence de l'espèce au site. Bien qu'il illustre assez simplement la relation entre les SDM et les concepts de niche et d'hypervolume, le modèle BIOCLIM est toutefois peu performant en comparaison avec des modèles plus récents (Elith et al., 2006).

L'un des modèles les plus utilisés aujourd'hui dans le domaine des SDM (Booth et al., 2014) est MAXENT (Phillips et al., 2017 ; Phillips et al., 2006 ; Phillips et Dudík, 2008), basé sur le principe d'entropie maximale, qui offre de bonnes performances tout en étant adapté aux données d'occurrence (à présence seulement). Plusieurs méthodes d'intelligence artificielle sont également performantes et très utilisées (Elith et al., 2006), notamment les forêts d'arbres décisionnels (*Random Forests*, RF) (Breiman, 2001) et les arbres de régressions renforcés (*Boosted Regression Trees*, BRT) (Elith et al., 2008). Carlson (2020) ont récemment suggéré d'utiliser les arbres de régression additifs bayésiens (*Bayesian Additive Regression Trees*, BART) (Chipman et al., 2010) pour les SDM, une alternative prometteuse aux RF et BRT permettant d'obtenir de meilleurs résultats en réduisant le surajustement, tout en permettant d'évaluer l'incertitude sous une formulation bayésienne. Les modèles bayésiens sont particulièrement adaptés au contexte actuel des SDM, puisqu'ils pourraient constituer une avenue pertinente considérant les différents appels à propager efficacement l'incertitude associée aux prédictions (Hortal et al., 2015 ; Poisot et al., 2016 ; Pollock et al., 2020).

### 0.3.2. Extension des modèles aux communautés

Plusieurs méthodes ont été suggérées afin de réaliser des prédictions au niveau de la communauté à partir de SDM, qui portent généralement sur des espèces seules. Trois types de stratégies ont été avancés pour la modélisation spatiale des communautés (D'Amen et al., 2017 ; Ferrier et Guisan, 2006) : 1) la stratégie « Assembler en premier, prédire ensuite », 2) la stratégie « Prédire en premier, assembler ensuite », et 3) la stratégie « Assembler et prédire simultanément ». La méthode la plus simple et la plus directement reliée aux SDM est celle des modèles de répartition d'espèces superposés (*stacked species distribution models*, S-SDM), qui consiste à réaliser des prédictions séparées pour chaque espèce présente dans la communauté, puis à superposer les prédictions de façon à connaître la composition en espèces pour chaque site d'une région d'intérêt (Ferrier et al., 2002 ; Ferrier et Guisan, 2006 ; Guisan et Rahbek, 2011). Celle-ci fait partie de la deuxième stratégie, « Prédire en premier, assembler ensuite ». Des mesures de description des communautés, comme la richesse spécifique ou l'unicité écologique, peuvent ensuite être calculées sur les communautés prédites. Cette approche est l'opposé des modèles macro-écologiques (*macroecological models*, MEM), dont le but est de prédire directement les propriétés d'un assemblage d'espèces (Gotelli et al., 2009 ; Guisan et Rahbek, 2011), sans passer par les SDM et l'identité des espèces présentes. Ceux-ci font donc partie de la première stratégie, « Assembler en premier, prédire ensuite ». Des méthodes plus complexes ont également été suggérées pour combiner les SDM et les MEM, ou encore pour intégrer de nouveaux éléments prédictifs, notamment par la modélisation spatialement explicite des assemblages d'espèces (*spatially explicit species assemblage modelling*, SESAM) (Guisan et Rahbek, 2011), les modèles conjoints de répartition d'espèces (*joint species distribution models*, JSJM) (Pollock et al., 2014), les superpositions probabilistes (Calabrese et al., 2014), la modélisation hiérarchique des communautés d'espèces (*hierarchical modelling of species communities*, HMSC) (Ovaskainen et al., 2017) et les réseaux bayésiens (*bayesian networks*, BN) intégrant les interactions biotiques (Staniczenko et al., 2017). Ces modèles ont l'avantage de prendre en compte plusieurs facteurs supplémentaires affectant la répartition des espèces, comme la co-occurrence entre les espèces, mais ils sont cependant plus complexes à réaliser. Malgré leur simplicité, les S-SDM offrent toutefois des résultats comparables aux autres modèles quant aux



prédictions de valeurs concernant les communautés, notamment avec les JSDM (Norberg et al., 2019; Zurell et al., 2020). De plus, ces modèles ont été développés et validés en bâtissant sur un grand nombre d'études prédictives portant sur la richesse spécifique. De leur côté, l'unicité écologique et les LCBD ont peu été utilisées selon des approches prédictives, de sorte qu'il existe peu de comparatifs de prédictions. L'utilisation des S-SDM pourrait donc s'avérer pertinente pour établir une référence de base pour les prédictions d'unicité écologique à grande étendue spatiale, qui pourrait ensuite être suivie par des modèles plus complexes comme les SESAM.

## **0.4. Données**

L'utilisation de méthodes prédictives pour évaluer l'unicité écologique sur de grandes étendues spatiales nécessite bien évidemment des données appropriées. Dans la prochaine section, j'expliquerai comment le contexte actuel est particulièrement favorable en raison de développements récents au niveau des bases de données, qui changent les méthodes pouvant être utilisées, et de la science citoyenne. De plus, j'expliquerai comment les SDM pourraient bénéficier d'une particularité de la base de données eBird.

### **0.4.1. Développement de nouvelles bases de données et de méthodes associées**

Depuis plusieurs années, de grandes bases de données en ligne sur la biodiversité se sont développées et fournissent des informations écologiques pouvant être exploitées lors d'études portant sur de grandes étendues spatiales, notamment GBIF (GBIF, s. d.), eBird (Sullivan et al., 2009) et iNaturalist (iNaturalist, s. d.). En même temps, nous disposons désormais de données de plus en plus précises sur les conditions environnementales partout sur le globe. Par exemple, WorldClim (Fick et Hijmans, 2017; Hijmans et al., 2005) et CHELSA (Karger et al., 2017) fournissent des données climatiques, alors que Copernicus (Buchhorn et al., 2019) et EarthEnv (Tuanmu et Jetz, 2014) fournissent des informations sur l'utilisation du territoire. Dans les deux cas, ces informations sont parfois disponibles à des échelles spatiales très fines.

La situation particulière des LCBD s'apparente à celle décrite par Poisot et al. (2016), selon qui le test d'hypothèse pour des systèmes à grande échelle est limité, de façon inhérente, par la

disponibilité de jeux de données adéquats. Ainsi, nos connaissances sur la biodiversité souffrent de nombreuses lacunes, qui sont cependant en voie d'être améliorées par le développement de nouvelles sources de données et par le développement de nouvelles méthodes. Hortal et al. (2015) a identifié 7 catégories de déficits, notamment sur la répartition des espèces (déficit Wallaceen), leur niche abiotique (déficit Grinnellien) et leurs interactions biotiques (déficit Eltonien). Plusieurs mégaprojets de collecte et assemblage de données ont cours en ce moment et offriront de grandes opportunités d'avancement, mais ceux-ci devront toutefois s'accompagner d'une évaluation critique des déficits et de l'incertitude (Hortal et al., 2015).

Cet essor des données massivement disponibles en ligne survient toutefois en même temps qu'un développement important des méthodes computationnelles. Mouquet et al. (2015) ont parlé de «*Datavalance*» pour décrire la prévalence des données qui modifie la façon de faire de la recherche en écologie. Poisot et al. (2019) ont de leur côté suggéré de passer à des approches dirigées par les données disponibles dans une optique de synthèse, ce qui offre le potentiel de générer de nouvelles informations écologiques à partir de données existantes. Ces méthodes sont particulièrement adaptées aux problèmes complexes, ainsi que pour faire le lien entre les modèles complexes et les données réelles. Devant l'ampleur des jeux de données maintenant disponibles, Pollock et al. (2020) ont mentionné qu'il est nécessaire d'intégrer la modélisation de la biodiversité et la conservation. Selon eux, les déficits de connaissances persistent malgré l'augmentation des collectes de données et de nombreuses initiatives mondiales d'assemblage de données de types variés, ce que les modèles de biodiversité offrent la possibilité de changer. Il existe deux types de modèles : les modèles d'imputation, pour les données manquantes, et les modèles spatiaux prédictifs, qui visent à prédire différentes mesures pour des endroits non échantillonnés (Pollock et al., 2020). Les SDM, que ce soient les modèles à espèces simples, les modèles superposés (S-SDM) ou les modèles conjoints (JSDM), font justement partie de cette catégorie de modèles. Ces modèles peuvent notamment traiter des jeux de données complexes avec de nombreuses dimensions, qui peuvent rapidement s'avérer difficiles à analyser (Poisot et al., 2019; Pollock et al., 2020). D'un point de vue de conservation, Pollock et al. (2020) ont également appelé à utiliser la diversité bêta en tant que mesure indicatrice pouvant capter des patrons non représentés dans des mesures centrées sur les

espèces seules, ce qui fait écho aux recommandations apportées par plusieurs études ayant utilisé la mesure des LCBD (da Silva et al., 2018 ; Landeiro et al., 2018).

#### **0.4.2. Science citoyenne**

Les développements de la science citoyenne sont également très importants, car ils permettent maintenant de disposer d'un grand nombre de données pouvant être utilisées dans des modèles portant sur de grandes étendues spatiales. Le nombre cumulatif de projets de science citoyenne a augmenté exponentiellement de 10 % par année de façon constante entre 1987 et 2015 (Pocock et al., 2017). Le type de projets de science citoyenne a également changé avec le temps de façon directionnelle, passant de surveillance systématique à une participation de masse, puis d'approches élaborées à des approches plus simples. Ce faisant, la diversité des projets en cours a augmenté avec le temps, même si les nouveaux projets démarrés ne sont pas plus diversifiés (Pocock et al., 2017). La science citoyenne implique de très nombreux volontaires et génère énormément de données, notamment sur de grandes étendues spatiales et sur des durées plus longues que la durée moyenne de financement académique (Theobald et al., 2015). Celles-ci sont cependant peu utilisées pour produire des articles scientifiques, malgré leurs données vérifiées, standardisées et accessibles en ligne, ce que Theobald et al. (2015) qualifient d'opportunité manquée pour la science et la société.

Par contre, l'incertitude géographique associée aux observations dans les bases de données publiques peut mener à des évaluations erronées des patrons de diversité et à une surestimation de la richesse spécifique dans les régions pauvres (alors que l'effet des incertitudes taxonomiques est moindre) (Maldonado et al., 2015). De plus, augmenter l'étendue spatiale ne semble pas régler ces problèmes et la précision est variable selon la provenance des données. Il y a cependant quelques avantages, notamment le fait que plus d'institutions fournissent des données à GBIF (il y a donc plus de données), que cela économise du temps et de l'argent et que les données sont plus uniformes (Maldonado et al., 2015). Les données issues de collectes volontaires et citoyennes sont biaisées de différentes façons : l'échantillonnage spatial et temporel, l'effort d'échantillonnage par visite et la détectabilité sont tous inégaux (Isaac et Pocock, 2015). Beck et al. (2014) ont cependant montré que sous-échantillonner et réduire le nombre d'observations pour éliminer les biais spatiaux, sans

toutefois réduire l'étendue spatiale visée, permet d'obtenir de meilleurs modèles prédictifs, même si ceux-ci sont basés sur moins de données.

### 0.4.3. Le problème des données d'absence

Plusieurs de ces méthodes SDM mentionnées plus tôt représentent toutefois des méthodes d'apprentissage supervisé, de sorte qu'elles ont besoin d'être entraînées sur des données déjà étiquetées. La principale conséquence au niveau des SDM est donc le besoin de disposer de données d'absence, en plus de données de présence, afin de pouvoir entraîner les algorithmes. Or, les données d'absence sont plus difficiles à obtenir, notamment en raison du problème du double zéro (Legendre et Legendre, 2012), selon lequel il est plus difficile d'obtenir une information représentant une absence en raison de la foule de facteurs pouvant mener à une non-détection.

La base de données *eBird* comporte toutefois un avantage à ce sujet, puisqu'il s'agit d'une base de données semi-structurée contenant des listes complètes (Johnston et al., 2020). Les données (et donc les observations) y sont structurées par listes d'observations. En rapportant leurs observations, les utilisateurs doivent déclarer si celles-ci constituent une liste complète des espèces détectées lors de leur échantillonnage. Ainsi, cela permet un peu plus justement d'inférer la non-détection d'autres espèces (Johnston et al., 2020). Les enjeux liés aux listes complètes ont été discutés en détail par Isaac et Pocock (2015). Leur utilisation se défend selon l'idée du contenu d'information (*information content*) d'une observation. Ainsi, selon la pyramide d'information croissante, les listes complètes contiennent plus d'informations que les registres d'incidence, puisqu'elles contiennent des informations au sujet de non-détections (Isaac et Pocock, 2015). À l'inverse, elles en contiennent moins que les sondages systématiques (qui impliquent un protocole structuré), que les visites à des sites fixés (plus faciles à comparer dans le temps) et que les visites répétées (qui permettent d'estimer directement la détectabilité). Suivant le principe du contenu d'informations, il est possible de comprendre l'incidence des comportements et des types d'observations, d'améliorer le contenu d'informations total avec des mesures ciblées, ainsi que de les prendre en compte dans les modèles prédictifs. Une autre façon d'améliorer l'information est de prendre en compte

des métadonnées sur le processus d'échantillonnage (Isaac et Pocock, 2015), ce que permet notamment *eBird* (Johnston et al., 2020).

## **0.5. Enjeux spatiaux**

Jusqu'à maintenant, j'ai expliqué comment la mesure des LCBD peut être utilisée pour évaluer l'unicité écologique, j'ai montré que nous disposons de modèles et de données pour effectuer des prédictions plus ambitieuses et j'ai exprimé ma volonté d'utiliser la mesure dans une approche prédictive sur de grandes étendues spatiales. Cependant, les déterminants d'une forte unicité telle que mesurée à grande échelle spatiale restent à évaluer. Bien que plusieurs études aient cherché à comprendre ces déterminants, peu d'entre elles les ont étudiés spécifiquement pour de grandes étendues spatiales et un très grand nombre de sites. Les éléments à examiner en priorité sont la relation entre l'unicité écologique, la richesse spécifique et la présence d'espèces rares, qui pourrait avoir une grande importance quant à l'utilité de la mesure en conservation et pour la gestion des aires protégées.

### **0.5.1. Relation avec la richesse spécifique et la proportion d'espèces rares**

Selon la formulation initiale de Legendre et De Cáceres (2013), les LCBD devraient normalement identifier les sites les plus uniques, que ce soit en raison de leur nombre d'espèces élevé ou faible, d'une composition en espèces particulière dans une région ou en raison de la présence d'espèces rares. Leur exemple initial a montré une relation négative entre la richesse spécifique et la valeur d'unicité (Legendre et De Cáceres, 2013). Ainsi, les sites les plus pauvres ressortent comme les plus uniques. Cette relation négative a également été observée dans plusieurs études ayant repris la mesure des LCBD (da Silva et Hernández, 2014; Heino et al., 2017; Heino et Grönroos, 2017), mais d'autres études ont montré que la relation pouvait également être positive dans certaines circonstances (Kong et al., 2017; Teittinen et al., 2017; Yao et al., 2021). Qiao et al. (2015) ont montré une relation négative avec le nombre d'espèces communes, mais également une relation positive avec le nombre d'espèces rares. Pour expliquer ces différences, da Silva et al. (2018)

ont avancé que la proportion d'espèces rares et communes dans les communautés pourraient déterminer si la relation sera globalement positive, négative ou non significative. Yao et al. (2021) ont confirmé cette hypothèse en montrant que la force et la direction de la relation sont effectivement reliées à la proportion d'espèces rares. Ainsi, les sites ayant une faible proportion d'espèces rares montrent une relation négative entre la richesse et la valeur de LCBD, alors que les sites ayant une proportion élevée montrent plutôt une relation positive.

Par contre, la rareté est complexe à définir, en particulier en lien avec la diversité bêta, car toutes deux dépendent de l'échelle spatiale considérée. Par exemple, sur de grandes étendues spatiales, certaines espèces peuvent être très communes à échelle locale dans une région donnée, mais rares pour l'ensemble de la région, ce qui pourrait influencer la relation richesse-LCBD. Différentes définitions de rareté peuvent donc être utilisées. Yao et al. (2021) ont repris la définition de rareté utilisée par De Cáceres et al. (2012), où les espèces rares sont celles qui se retrouvent dans moins de 40 % des sites, alors que Qiao et al. (2015) ont considéré comme rares les espèces comptant moins de 25 individus. De plus, la diversité bêta totale augmente avec l'étendue spatiale (Barton et al., 2013) et dépend de l'échelle, notamment en raison de l'augmentation de l'hétérogénéité spatiale, ainsi qu'en raison du recoupement de bassins d'espèces locaux différents (Heino et al., 2015). Ainsi, l'effet des espèces rares sur l'unicité écologique devrait également être étudié en fonction de l'échelle spatiale, de même que sur des groupes taxonomiques différents.

### **0.5.2. Utilité en conservation**

La relation entre la richesse spécifique et la valeur d'unicité écologique a également une importance quant à l'utilisation de la mesure des LCBD en conservation. Legendre et De Cáceres (2013) ont indiqué que les LCBD pourraient être utiles pour identifier des sites ayant une grande valeur de conservation, ayant besoin de restauration, ayant subi des invasions écologiques ou ayant besoin d'être étudiés plus amplement. Certaines études ont interprété la relation négative comme une indication de l'importance de conserver les sites pauvres en espèces en plus des sites les plus riches, notamment parce qu'ils pourraient contenir des espèces rares (da Silva et al., 2018; Heino et al., 2017). Landeiro et al. (2018) ont noté que les sites à forte unicité, malgré leur faible richesse,

contribuent au maintien de la diversité bêta dans des régions hyperdiversifiées, et sont importants à conserver de façon à préserver des habitats représentatifs de la diversité actuelle. Considérant les ressources limitées en conservation, un compromis intéressant pourrait être de conserver une combinaison de sites à forte unicité écologique et de sites riches en espèces (Heino et Grönroos, 2017). Yao et al. (2021) ont appuyé cette approche, ajoutant qu'il faut également s'intéresser à la diversité bêta, puisque celle-ci peut varier différemment de l'unicité écologique et est affectée différemment par les facteurs environnementaux. Dubois et al. (2020) ont également appuyé cette approche de conservation, montrant le compromis à conserver les sites les plus uniques, les sites les plus riches et les sites contenant des espèces rares. Selon eux, il faut donc porter attention à la composition en espèces lorsque la mesure des LCBd est utilisée et idéalement la compléter avec d'autres critères de conservation avant de déterminer quels sites protéger. Une hypothèse pour aller plus loin a été avancée par da Silva et al. (2018), qui ont souligné que les notions de renouvellement et d'emboîtement peuvent aider à distinguer comment les sites pauvres diffèrent des profils généraux au sein d'une communauté. Par contre, ces auteurs ont également noté que ces deux notions entraînent des stratégies de conservation différentes.

L'utilisation des LCBd en conservation s'inscrit dans une approche plus générale visant à prendre en compte la diversité bêta. Socolar et al. (2016) ont montré l'importance générale de la diversité bêta en conservation. Les LCBd, et donc le principe d'unicité écologique, peuvent être contrastés avec d'autres approches visant insistant plutôt sur l'équitabilité et sur la dissimilarité entre les sites (Ricotta, 2017). L'équitabilité peut notamment être quantifiée et reliée à la diversité bêta (Chao et Ricotta, 2019). D'autres approches spatialement explicites pouvant potentiellement être utilisées sont l'endémisme pondéré des espèces (Crisp et al., 2001 ; Guerin et al., 2015) et l'unicité (*uniquity*) (Ejrnæs et al., 2018). Celles-ci pourraient s'avérer des alternatives intéressantes à contraster avec l'unicité écologique.

## 0.6. Objectifs de l'étude

Suivant les éléments présentés dans les différentes sections de cette introduction, j'ai donc réalisé une étude où j'ai cherché à évaluer comment l'unicité écologique peut être évaluée à grande

échelle spatiale. Mon objectif était de déterminer si les modèles de répartition d'espèces permettent une évaluation juste de l'unicité écologique en comparaison avec les données d'occurrences disponibles dans les grandes bases de données de science citoyenne. J'ai également cherché à déterminer si la relation entre la richesse et les LCBD est constante entre les régions et entre les échelles spatiales pour un groupe taxonomique donné. Pour ce faire, j'ai utilisé des SDM, des arbres de régression additifs bayésiens (BART), des observations de parulines en Amérique du Nord provenant de la base de données eBird, ainsi que des données environnementales provenant des bases de données WorldClim et Copernicus. Dans un premier temps, j'ai effectué des prédictions pour la répartition des espèces en fonction des conditions environnementales, que j'ai assemblée pour former des prédictions portant sur les communautés. J'ai ensuite calculé les valeurs de LCBD pour les données observées et prédites, puis comparé ces valeurs de trois façons différentes. J'ai également calculé la relation entre la richesse spécifique et l'unicité écologique pour deux sous-régions aux profils de richesse variés, ainsi qu'à trois étendues spatiales différentes. J'ai finalement vérifié l'effet de la proportion d'espèces rares au sein des communautés sur cette relation.



**First Article.**

# **Evaluating ecological uniqueness over broad spatial extents using species distribution modelling**

by

Gabriel Dansereau<sup>1</sup>, Pierre Legendre<sup>1</sup>, and Timothée Poisot<sup>1</sup>

(<sup>1</sup>) Département de sciences biologiques, Université de Montréal  
1375 avenue Thérèse-Lavoie-Roux, Montréal, QC, Canada H2V 0B3

This article will be submitted to Global Ecology and Biogeography.

GD developed and performed the analyses and wrote the first version of the manuscript;

TP developed a preliminary version of the analyses.

PL and TP provided guidance on the analyses and interpretation of the results and revised the manuscript.

All authors read and approved the manuscript.

## RÉSUMÉ.

**Objectif:** La mesure des contributions locales à la diversité bêta (LCBD) permet d'identifier les sites à forte unicité écologique et ayant une composition en espèces exceptionnelle au sein d'une région d'intérêt. L'utilisation de cette mesure est cependant typiquement restreinte aux échelles locales et régionales avec un petit nombre de sites, puisqu'elle requiert des informations sur la composition complète des communautés parfois difficiles à acquérir à grande échelle spatiale. Dans cette étude, nous examinons comment la mesure des LCBD peut être utilisée pour prédire l'unicité écologique sur de grandes étendues spatiales à l'aide de modèles de répartition d'espèces et de données de science citoyenne.

**Lieu:** Amérique du Nord

**Période de temps:** Années 2000

**Taxon étudié:** Parulidés

**Méthodes:** Nous avons utilisé les arbres de régression additifs bayésiens (BARTs) pour prédire la répartition des parulines en Amérique du Nord à l'aide de données provenant de la base de données eBird. Nous avons ensuite calculé les valeurs de LCBD pour les données observées et prédites, puis nous avons examiné la différence par site à l'aide d'une comparaison directe, d'un test d'autocorrélation spatiale et de modèles de régression linéaire généralisés. Nous avons également examiné la variation de la relation entre les valeurs de LCBD et la richesse spécifique entre différentes régions et sur différentes étendues spatiales, ainsi que l'effet de la proportion d'espèces rares sur cette relation.

**Résultats:** Nos résultats ont montré que la relation entre la richesse et les LCBD varie selon la région et l'étendue spatiale où celle-ci est mesurée et qu'elle est influencée par la proportion d'espèces rares dans la communauté. Les modèles de répartition d'espèces ont fourni des estimés fortement corrélés avec les données observées, mais qui montraient également de l'autocorrélation spatiale.

**Principales conclusions:** Les sites identifiés comme uniques sur de grandes étendues spatiales peuvent varier en fonction de la richesse régionale, de l'étendue totale et de la proportion d'espèces rares dans la communauté. Les modèles de répartition d'espèces peuvent être utilisées pour prédire l'unicité écologique sur de grandes étendues spatiales, ce qui pourrait permettre d'identifier des points chauds de diversité bêta et d'importantes cibles de conservation au sein de régions non échantillonnées.

**Mots clés :** diversité bêta, unicité écologique, contributions locales à la diversité bêta, modèles de répartition d'espèces, échelle spatiale étendue, eBird

## ABSTRACT.

**Aim:** Local contributions to beta diversity (LCBD) can be used to identify sites with high ecological uniqueness and exceptional species composition within a region of interest. Yet, these indices are typically used on local or regional scales with relatively few sites, as they require information on complete community compositions difficult to acquire on larger scales. Here, we investigate how LCBD indices can be used to predict ecological uniqueness over broad spatial extents using species distribution modelling and citizen science data.

**Location:** North America.

**Time period:** 2000s.

**Major taxa studied:** Parulidae.

**Methods:** We used Bayesian additive regression trees (BARTs) to predict warbler species distributions in North America based on observations recorded in the eBird database. We then calculated LCBD indices for observed and predicted data and examined the site-wise difference using direct comparison, a spatial autocorrelation test, and generalized linear regression. We also investigated the relationship between LCBD values and species richness in different regions and at various spatial extents and the effect of the proportion of rare species on the relationship.

**Results:** Our results showed that the relationship between richness and LCBD values varies according to the region and the spatial extent at which it is applied. It is also affected by the proportion of rare species in the community. Species distribution models provided highly correlated estimates with observed data, although spatially autocorrelated.

**Main conclusions:** Sites identified as unique over broad spatial extents may vary according to the regional richness, total extent size, and the proportion of rare species. Species distribution modelling can be used to predict ecological uniqueness over broad spatial extents, which could help identify beta diversity hotspots and important targets for conservation purposes in unsampled locations.

**Keywords:** beta diversity, ecological uniqueness, local contributions to beta diversity, species distribution modelling, broad spatial scale, eBird

## 1. Introduction

Beta diversity, defined as the variation in species composition among sites in a geographic region of interest (Legendre et al., 2005), is an essential measure to describe the organization of biodiversity through space. Total beta diversity within a community can be partitioned into local

contributions to beta diversity (LCBD) (Legendre & De Cáceres, 2013), which allows the identification of sites with exceptional species composition, hence unique biodiversity. Such a method, focusing on specific sites, is useful for both community ecology and conservation biology, as it highlights areas that are most important for their research or conservation values. For example, sites with unique community composition often differ from those with high species richness, possibly as they harbour rare species or help maintain beta diversity (da Silva et al., 2018; Heino et al., 2017; Landeiro et al., 2018). Hence, focusing on unique community composition may prove useful as a complementary approach to species richness (da Silva et al., 2018; Dubois et al., 2020; Heino & Grönroos, 2017; Yao et al., 2021). However, the use of LCBD indices is currently limited in two ways. First, LBCD indices are typically used on data collected over local or regional scales with relatively few sites, for example on fish communities at intervals along a river or stream (Legendre & De Cáceres, 2013). Second, LCBD calculation methods require complete information on community composition, such as a community composition matrix  $Y$ ; thus, they are inappropriate for partially sampled sites (e.g., where data for some species is missing or uncertain) and cannot directly provide assessments for unsampled ones. Accordingly, the method is of limited use to identify areas with exceptional biodiversity in regions with sparse sampling. However, predictive approaches are increasingly common given the recent development of computational methods, which often uncover novel ecological insights from existing data (Poisot et al., 2019), including in unsampled or lesser-known locations, as well as larger spatial scales. Thus, biodiversity modelling can fill in knowledge gaps and needs to be better integrated with conservation decisions (Pollock et al., 2020).

Species distribution models (SDMs) (Guisan & Thuiller, 2005) can bring a new perspective to LCBD studies by filling in gaps and performing analyses on much broader scales. In a community matrix  $Y$ , such as required for LCBD calculation, ecological communities are abstracted as assemblages of species present at different sites. Viewing communities as such opens the perspective of predicting community composition from predictions of individual species, which is the aim of SDMs, thus allowing calculating LCBD values. At their core, SDMs aim at predicting the distribution of a species based on information about where the species was previously reported, matched

with environmental data at those locations, and then make predictions at other (unsampled) locations based on their respective environmental conditions. Community-level modelling from SDMs is not an especially novel idea (Ferrier et al., 2002; Ferrier & Guisan, 2006), but it is increasingly relevant with the advent of large-scale, massive, and open data sources on species occurrences, often contributed by citizens, such as GBIF and eBird (Sullivan et al., 2009).

Many approaches allow going from single-species SDMs to a whole community, although these have not been explicitly evaluated for ecological uniqueness and LCBD indices. The simplest approach is stacked distribution models (S-SDMs) (Ferrier & Guisan, 2006; Guisan & Rahbek, 2011). Single species SDMs are first performed separately, then combined to form a community prediction on which community-level analyses can be applied. On the other hand, joint species distribution models (JSDMs) (Pollock et al., 2014) try to improve species distribution predictions by incorporating species co-occurrence into the models. However, these models do not always improve community-level predictions compared to S-SDMs (Norberg et al., 2019; Zurell et al., 2020). S-SDMs also tend to overestimate species richness (D'Amen et al., 2015; Dubuis et al., 2011; Zurell et al., 2020), which could result from thresholding the probabilities into presence-absence data before stacking the species distributions (Calabrese et al., 2014). Summing the occurrence probabilities without applying a threshold is, therefore, another alternative (Calabrese et al., 2014). However, this may limit some analyses as it does not return species identities for every site (Zurell et al., 2020), as is required with LCBD calculations. Spatially explicit species assemblage modelling (SESAM) (Guisan & Rahbek, 2011), hierarchical modelling of species communities (HMSC) (Ovaskainen et al., 2017), and Bayesian networks (BN) (Staniczenko et al., 2017) are other alternatives that could yield better community predictions than S-SDMs. However, they add methodological and computational overload, impeding their use for broad spatial extents, and are often validated against extensive work on species richness. By comparison, ecological uniqueness and LCBD indices have rarely been used in predictive frameworks. Therefore, S-SDMs may prove an appropriate first step to establish some baselines, especially given that calculating LCBD values on SDM predictions will raise some important issues, such as calculating the uniqueness scores on much bigger community matrices and broader scales than what is typically done.

The total number of sites will increase (1) because of the spatially continuous nature of the predictions, as there will be more sites in the region of interest than the number of sampled sites, and (2) because of the larger spatial extent allowed for the SDM predictions. A high number of SDM-predicted sites with a large extent opens up the possibility of capturing a lot of variability of habitats and community composition, but also many very similar ones. This could change the way that exceptional sites contribute to the overall variance in the large-scale community. LCBD scores have typically been used at local or regional scales with relatively few sites (da Silva & Hernández, 2014; Heino et al., 2017; Heino & Grönroos, 2017; Legendre & De Cáceres, 2013). Some studies did use the measure over broader, near-continental extents (Poisot et al., 2017; Taranu et al., 2020; Yang et al., 2015), but the total number of sites in these studies was relatively small. Recent studies also investigated LCBD and beta diversity on sites distributed in contiguous grids or as pixels, hence uniform sampling intervals and no spatial gaps, but these did not cover large extents and a high number of sites (D'Antraccoli et al., 2020; Legendre & Condit, 2019; Tan et al., 2017; Tan et al., 2019). Two recent studies have, however, adopted promising predictive approaches on regional extents. First, Niskanen et al. (2017) predicted LCBD values of plant communities (and three other diversity measures) on a continuous scale and a high number of sites (> 25,000) using Boosted Regression Trees (BRTs). However, they modelled the diversity measures directly after calculating them on a smaller number of sampled sites. They obtained lower predictive accuracy for LCBD than for the other diversity measures, highlighting the challenge to predict the measure. Second, Vasconcelos et al. (2018) used ecological niche models (ENMs) to predict anurans ecological niches according to actual and forecasted environmental conditions, then calculated the LCBD values on the predictions to identify biodiversity hotspots. Using this approach, predicted LCBD values are calculated in a closer way to the original formulation based on community composition. This development of predictive approaches is exciting, especially as it could be pushed a step further to continental extents, a higher number of sites, and more species occurrences using SDMs and massive data sources. Still, it should be accompanied by an investigation of the determinant of ecological uniqueness in such conditions.

Measuring ecological uniqueness from LCBD indices over broad spatial extents and spatially continuous data also raises the question of which sites will be identified as exceptional and for what reason. The method intends that sites should stand out and receive a high LCBD score whenever they display an exceptional community composition, be it a unique assemblage of species that may have a high conservation value or a richer or poorer community than most in the region (Legendre & De Cáceres, 2013). Both the original study and many of the later empirical ones have shown a negative relationship between LCBD scores and species richness (da Silva & Hernández, 2014; Heino et al., 2017; Heino & Grönroos, 2017; Legendre & De Cáceres, 2013), although other studies observed both negative and positive relationships at different sites (Kong et al., 2017) or quadrats (Yao et al., 2021). Some studies showed that the direction of the relationship is related to the percentage of rare species in the community (da Silva et al., 2018; Yao et al., 2021). The LCBD-richness relationship and the effect of the proportion of rare species should therefore be investigated over broad spatial extents, as beta diversity and species rarity are both concepts that depend on scale. For instance, total beta diversity increases with spatial extent (Barton et al., 2013) and is strongly dependent on scale, notably because of higher environmental heterogeneity and sampling of different local species pool (Heino et al., 2015), which could potentially add some variation to the relationship.

Here, we examine whether the LCBD method can assess ecological uniqueness over broad spatial extents based on predictions of species distributions and evaluate whether this reveals novel ecological insights regarding the identification of exceptional biodiversity areas. We first predicted species distributions on continental scales using extended occurrence data from eBird and Bayesian additive regression trees (BARTs). We then quantified uniqueness with the local contributions to beta diversity measure for both predicted and observed data. Next, we examined the site-wise difference using direct comparison, a spatial autocorrelation test, and generalized linear regression. We then investigated the relationship between uniqueness and species richness for different regions and scales and according to the proportion of rare species.

## 2. Methods

### Occurrence data

We used occurrence data from eBird (Sullivan et al., 2009) downloaded through the eBird Basic Data set from June 2019 (eBird Basic Dataset, 2019). We restricted our analyses to the New World warbler family (*Parulidae*) in North America (Canada, United States, Mexico) using the *R* package *auk* (Strimas-Mackey et al., 2018) to extract and process bird sightings records from the eBird database. eBird is a semi-structured citizen science data set, meaning that observations are reported as checklists of species detected in an observation run (Johnston et al., 2020). Observers can explicitly specify that their checklist contains all species they could detect and identify during a sampling event, in which case it is labelled as a “complete checklist.” Using complete checklists instead of regular checklists allows researchers to infer non-detections in locations where detection efforts did occur, which offers performance gains in species distribution models (Johnston et al., 2020). Therefore, we selected the data from the complete checklists only. Our final data set comprised 62 warbler species and nearly 23 million observations from 9 million checklists. Warblers are a diverse group with many species, are popular among birders given their charismatic aspect, are distributed in diverse areas and present relatively everywhere in North America.

### Environmental data

Our environmental data consisted of climatic data from the WorldClim 2.1 database (Fick & Hijmans, 2017) and land cover data from the Copernicus Global Land Service (Buchhorn et al., 2019). We restricted these data to a spatial extent comprised between -145.0 and -50.0 degrees of longitude and between 20.0 and 75.0 degrees of latitude (Fig. 1). First, the WorldClim data consists of spatially interpolated monthly climate data for global land areas. We used the standard *bioClim* variables from WorldClim 2.1, which represent annual trends, ranges, and extremes of temperature and precipitation, but selected only 8 out of the 19 ones to avoid redundancy (bio1, bio2, bio5, bio6, bio12, bio13, bio14, bio15). We downloaded the data at a resolution of 10 arcminutes (around 18 km<sup>2</sup> at the equator), the coarsest resolution available, using the *Julia* package



`SimpleSDMLayers.jl` (Dansereau & Poisot, 2021). The coarse resolution should mitigate potential imprecision in the eBird data regarding the extent of the sampled areas in each observation checklist. Moreover, some studies have argued that coarser resolutions lead to less overestimation of species richness and better identification of bird biodiversity hotspots given the patchiness of observation data (Hurlbert & Jetz, 2007). We acknowledge that using an arcminutes-based resolution means that the surface area of our pixels will not be equal depending on the latitude. Second, the Copernicus data is a set of variables representing ten land cover classes (e.g., crops, trees, urban areas) and measured as a percentage of land cover. The data is only available at a finer resolution of 100 m, which we downloaded directly from the website. We coarsened it to the same ten arcminute resolution as the WorldClim data by averaging the pixels' cover fraction values with GDAL (GDAL/OGR contributors, 2021). We first selected the ten land cover variables but later removed two (moss and snow) from our predictive models. Their cover fraction was 0% on all sites with warbler observations; hence they did not provide any predictive value to our SDM models.

## Species distribution models

We converted the occurrence data to a presence-absence format compatible with community analyses. We considered every pixel from our ten arcminutes environmental layers as a site. We then verified, for each species, if there was a single observation in every site. Finally, we recorded the outcome as a binary value: present (1) if a species was ever recorded in a site and absent (0) if it was not. Complete checklists help ensure that these zeros represent non-detections, rather than the species not being reported; hence we considered them as absence data, similar to Johnston et al. (2020), although we recognize that there subsists a doubt about this.

We predicted species distribution data on continuous scales from our presence-absence data using Bayesian Additive Regression Trees (BARTs) (Chipman et al., 2010), a classification and regression trees method recently suggested for species distribution modelling (Carlson, 2020). BARTs are sum-of-trees models, conceptually similar to Boosted Regression Trees and Random Forest, but following a Bayesian paradigm: trees are constrained as weak learners by priors regarding structure and nodes (Carlson, 2020; Chipman et al., 2010). Then, fitting and inference is made

through an iterative Bayesian backfitting Markov Chain Monte Carlo (MCMC) algorithm generating a posterior distribution of predicted classification probabilities (Carlson, 2020; Chipman et al., 2010). We used the package `embarcadero` (Carlson, 2020) in *R* to compute the BART models. First, we performed BARTs separately for all species and estimated the probability of occurrence for all the sites in the pseudo-rectangular spatial units of 10 arcminutes in the region of interest. We then converted the results to a binary outcome according to the threshold that maximized the True Skill Statistic (TSS) for each species, as suggested by Carlson (2020).

## Quantification of ecological uniqueness

We used the method of Legendre & De Cáceres (2013) to quantify compositional uniqueness from overall beta diversity for both the observed and predicted data. First, we assembled the presence-absence data by site to form two site-by-species community matrices, one from observed data, called  $Y$  (39,091 sites by 62 species), and one from predicted data, called  $\hat{Y}$  (99,609 sites by 62 species). Next, we measured species richness per site as the sum of the presences in each row, i.e., the number of species present. Finally, we removed the sites without any species from the predicted community matrix  $\hat{Y}$ , for a new total of 92,117 sites (this was not necessary for the observed community matrix  $Y$ , as it was, by design, only composed of sites with at least one species present). We applied the Hellinger transformation to both matrices, as recommended by Legendre & De Cáceres (2013) for presence-absence data. We then measured total beta diversity as the variance of the community matrices and calculated the local contributions to beta diversity (LCBD), which quantify how much a specific site (a row in each matrix) contributes to the overall variance in the community (Legendre & De Cáceres, 2013). High LCBD values indicate a unique community composition, while low values indicate a more common species set. Measuring beta diversity as the variance of the community matrices offers a critical advantage in computations in this case, as alternative approaches based on sites' pairwise dissimilarity would require a much higher number of calculations given the high number of sites in our study. We note that our LCBD values, which add up to 1 because the raw LCBD values are divided by the total sum-of-squares of

the data matrix, were very low given the high number of sites in both  $Y$  and  $\hat{Y}$ . However, the relative difference between the scores in one set matters more than the absolute value to differentiate their uniqueness.

## Comparison of observed and predicted values

We performed three verification to compare the species richness and uniqueness estimates obtained from our predicted species distributions to those obtained with the occurrence data from eBird. First, we performed a direct comparison by subtracting the richness and LCBBD estimates obtained from  $Y$  (the observed data) from the estimates obtained from  $\hat{Y}$  (the predicted data). To do so, we used the richness estimates as-is but modified the LCBBD values to achieve a non-biased comparison, given that the original values are calculated for the same sites but on sets of different lengths. Therefore, we recomputed the LCBBD scores only for the sites for which we had occurrences in both  $Y$  and  $\hat{Y}$ , which mostly corresponded to the sites in  $Y$ , minus a few sites where the SDMs predicted no species occurrence. We then plotted the richness and LCBBD differences to examine their spatial distributions. Second, we performed the modified  $t$  test from Clifford et al. (1989) to assess the correlation between the observed and predicted estimates and test for spatial autocorrelation. We performed the test separately for the richness and the LCBBD estimates. We used the `modified.ttest` function from the package `SpatialPack` (Vallejos et al., 2020) in *R*. Third, we performed Generalized Linear Models between the observed and predicted estimates and plotted the deviance residuals to examine their spatial distribution. We used a negative binomial regression with a log link function using the package `MASS` (Venables & Ripley, 2002) in *R* for the richness estimates as our values showed overdispersion (Fletcher & Fortin, 2018). We used a beta regression with a logit link function using the package `betareg` (Cribari-Neto & Zeileis, 2010) for the LCBBD values as they vary between 0 and 1, similar to Heino & Grönroos (2017) and Yao et al. (2021).

## Investigation of regional and scaling variation

We recalculated LCBD values on various subregions at different locations and scales to investigate possible regional and scaling effects. First, we selected two subregions of equivalent sizes (20.0 longitude degrees by 10.0 latitude degrees) with contrasting richness profiles to verify if the relationship between species richness and LCBD values was similar. We selected a Northeast subregion (longitude between -80.0 and -60.0, latitude between 40.0 and 50.0), mostly species-rich, and a Southwest subregion (longitude between -120.0 and 100.0, latitude between 30.0 and 40.0), mostly species-poor (for both the observed and predicted data). Fig. 4 shows the coordinates and spatial extents of both subregions. Second, we recalculated the LCBD indices at three different extents, starting with a focus on the Northeast subregion and progressively extending the extent to encompass the Southwest subregion (Fig. 5). These are conceptually similar to the spatial windows of Barton et al. (2013), which allow one to study the variation of beta diversity according to spatial extent. We did these two verifications with both the observed and predicted data but only illustrate the results with the predicted data as both were qualitatively similar.

## Proportion of rare species

We investigated the effect of the proportion of rare species in the community on the direction of the relationship between species richness and LCBD values in our Northeast and Southwest subregions. Following De Cáceres et al. (2012) and Yao et al. (2021), we classified species as rare when they occurred in less than 40% of the sites in each subregion. We calculated the proportion of rare species for every site. We then grouped the sites for both subregions depending on whether they were part of an ascending or a descending portion in the LCBD-richness relationship. Given that the relationship sometimes displays a curvilinear form with a positive quadratic term (Heino & Grönroos, 2017; Tan et al., 2019), we separated the ascending and descending portions based on the species richness at the site with the lowest LCBD value (we used the median richness if there were multiple sites with the lowest LCBD value), which corresponds to the inflection point of the relationships shown on Fig. 4. For example, the lowest LCBD value was  $7.032e-5$  in the Northeast subregion and the median richness (as there were multiple sites with this LCBD value) was 23. All

the sites with more than 23 species were assigned to the ascending portion, and all the sites with 23 species or fewer were assigned to the descending portion. In the Southwest subregion, the lowest LCBD value and its corresponding (median) richness were  $6.035e-5$  and 12, respectively. We then mapped the ascending and descending groups to view their spatial distribution. We also examined the distribution of the rare species proportions in both groups using a density plot. Similar to our previous verification, we performed this analysis with both observed and predicted data but once again only illustrate the results with the predicted data as both were qualitatively similar.

## Software

We used *Julia v1.6.1* (Bezanson et al., 2017) for most of the project and *R v4.0.2* (R Core Team, 2020) for some specific steps. We used the *Julia* package `SimpleSDMLayers.jl` (Dansereau & Poisot, 2021) as the basic framework for our analyses, to download the WorldClim 2.1 data, and to map our results through the package’s integration of `Plots.jl`. We also used `StatsPlots.jl` to produce the density plots in our rare species analysis. We used the *R* packages `auk` (Strimas-Mackey et al., 2018) to extract and manipulate eBird data, `embarcadero` (Carlson, 2020) to perform the BART models, `vegan` (Oksanen et al., 2019) to perform the Hellinger transformations, `SpatialPack` to perform the modified  $t$  test, as well as `MASS` (Venables & Ripley, 2002) and `betareg` (Cribari-Neto & Zeileis, 2010) to perform the negative binomial and beta regressions, respectively. We used GDAL to coarsen the Copernicus land cover data. All the scripts used for the analyses are available at <https://github.com/gabrieldansereau/betadiversity-hotspots>.

## 3. Results

### Species distribution models generate relevant community predictions

Species richness from observation data (Fig. 1a) was higher on the East Coast and lower on the West Coast, with many unsampled patches in the North, Midwest, and Southwest. Richness results from SDM data (Fig. 1b) displayed higher richness on the East Coast and sites with few or no species up north and in the Midwest. There was no clear latitudinal gradient in richness but rather an East-West one. Landmarks such as the Rockies and croplands in the Midwest (which

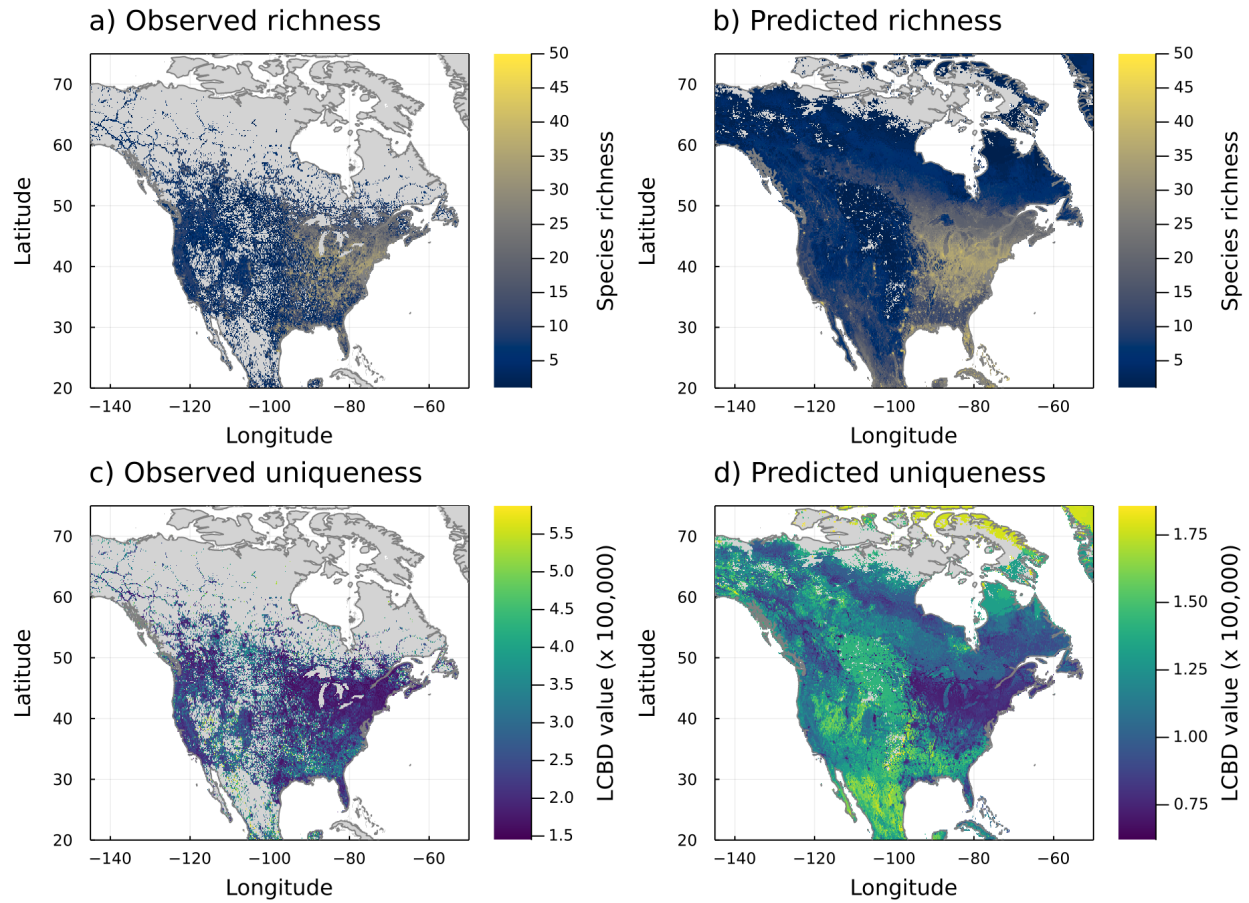
should be species-poor habitats) were notably visible on the maps, separating the East and West. LCBD scores from observation data (Fig. 1c) were low on the East Coast and higher on the border of sampled sites in the Midwest. They were also higher in the North and in the South, where observations were sparser. Results from SDM predictions were similar (Fig. 1d), with lower LCBD values in the East and more unique sites in the Midwest region, Central Mexico, and some Northern regions. There was no clear latitudinal gradient, and the East-West contrast, while present, was less clear than on the richness maps. LCBD values ranged between  $1.444\text{e-}5$  and  $5.860\text{e-}5$  for observation data and between  $5.788\text{e-}6$  and  $1.706\text{e-}5$  for SDM data. The total beta diversity was 0.608 for the observation data and 0.775 for the SDM data.

The modified  $t$  test of Clifford et al. (1989) showed high correlation between the observed and predicted estimates of richness and uniqueness, as well as spatial autocorrelation in the values. For species richness, the correlation coefficient was 0.7802, the  $F$ -statistic was 20,162, and the  $p$ -value was  $6.126\text{e-}04$ . For LCBD scores, the correlation coefficient was 0.5207, the  $F$ -statistic was 40,942, and the  $p$ -value was  $3.936\text{e-}09$ .

The difference between the observed and predicted estimates (predicted richness - observed richness and predicted LCBD - observed LCBD) showed opposite geographic distributions for species richness and ecological uniqueness (Fig. 2). Predicted richness estimates were higher than observed estimates on the East Coast in particular as well as on the West Coast and in Mexico, but were lower than observed estimates in the Midwest (Fig. 2a). Predicted LCBD estimates, on the other hand, were lower than observed estimates on the East Coast and higher in the Midwest (Fig. 2b). Regression residuals showed similar geographic distributions to their corresponding difference values (Fig. 3).

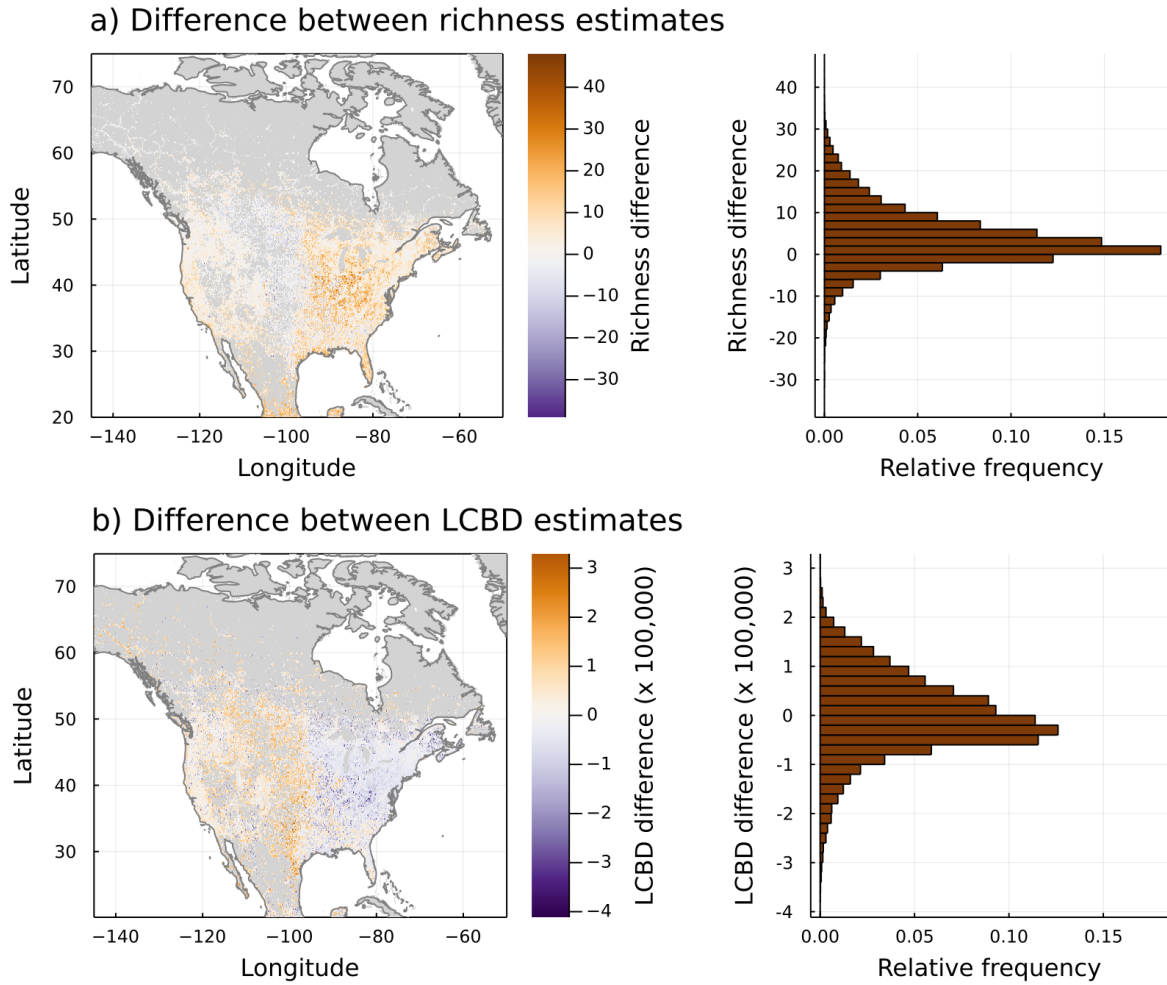
## **Uniqueness displays regional variation as two distinct profiles**

The relationship between LCBD values and species richness displayed contrasting profiles in species-rich and species-poor regions (Fig. 4). In the species-rich northeastern region of our study extent (North America), LCBD scores displayed a mostly decreasing relationship with species richness, with a slightly curvilinear form and increase of values for very rich sites. The sites with the



**Figure 1** – Comparison of species richness and LCBd scores from observed and predicted warbler occurrences in North America. Values were calculated for sites representing ten arcminute pixels. We measured species richness after converting the occurrence data from eBird (a) and the SDM predictions from our single-species BART models (b) to a presence-absence format per species. We applied the Hellinger transformation to the presence-absence data, then calculated the LCBd values from the variance of the community matrices, separately for the occurrence data (c) and the SDM predictions (d). LCBd values ranged between  $1.444\text{e-}5$  and  $5.860\text{e-}5$  for observation data and between  $5.788\text{e-}6$  and  $1.706\text{e-}5$  for SDM data. The total beta diversity was 0.608 for the observation data and 0.775 for the SDM data. Areas in light grey (not on the colour scale) represent mainland sites with environmental data but without any warbler species present.

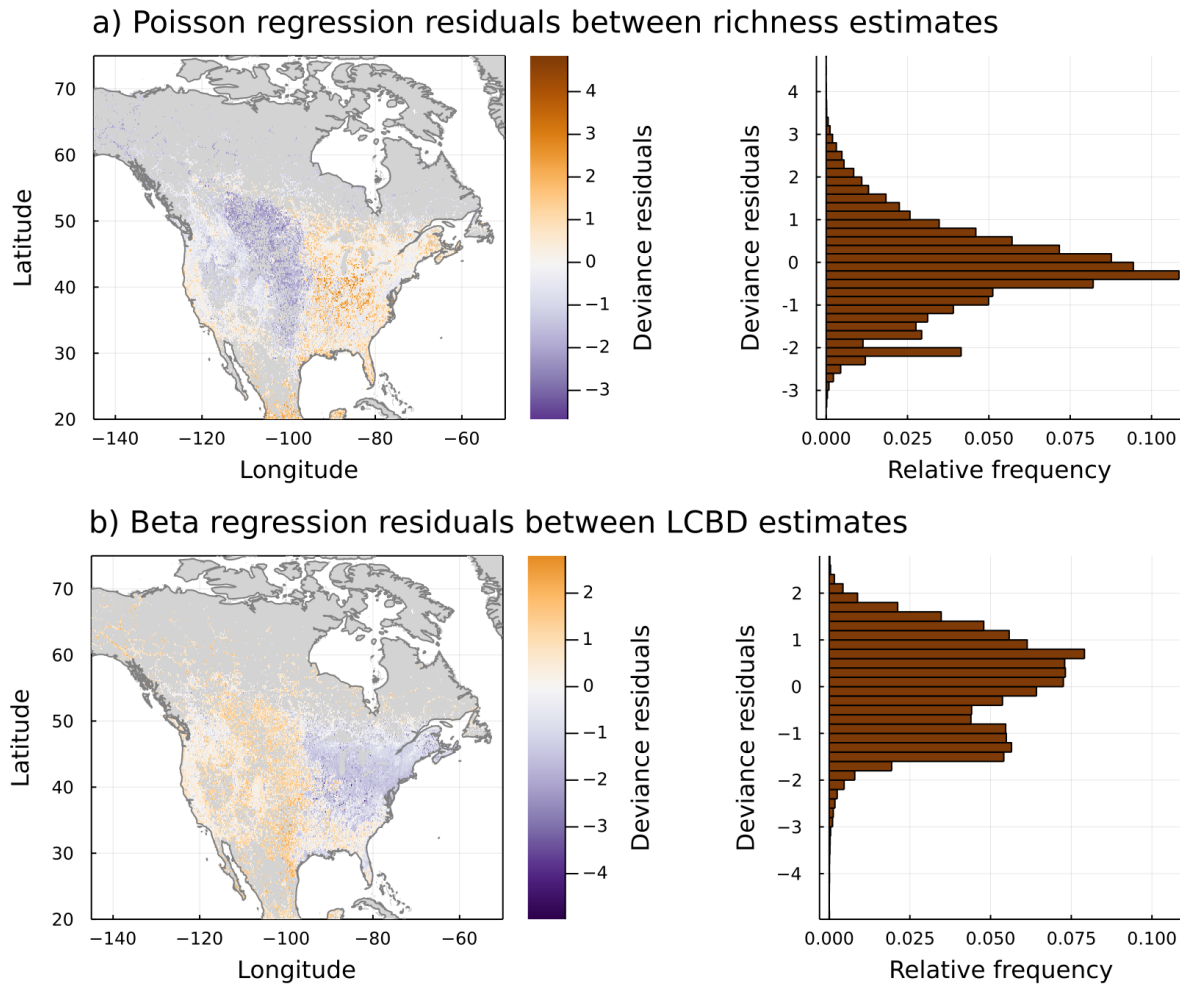
highest LCBd values, i.e., the unique ones in terms of species composition, were the species-poor sites, while the species-rich sites displayed lower LCBd scores. The Southwest subarea showed a different relationship with a much sharper initial decline and a more significant increase as richness reaches 20 species. The sites with the highest LCBd values were the poorest ones in terms of species richness, as in the Northeast region, but the species-rich sites were proportionally more unique in the Southwest region. Total beta diversity was also higher in the Southwest subregion



**Figure 2** – Comparison between observed and predicted estimates of species richness (a) and ecological uniqueness (b). The difference values represent the estimate from the predicted data set minus the estimate from the observed data set. The difference values for richness ranged between -38 and 45 (a). LCBD values were recalculated for the same set of sites with observations in both data sets. Recalculated LCBD ranged between  $1.450\text{e-}5$  and  $5.910\text{e-}5$  for observation data and between  $1.117\text{e-}5$  and  $5.132\text{e-}5$  for SDM data. The difference values for LCBD scores ranged between  $-4.060\text{e-}5$  and  $3.297\text{e-}5$  (b).

(0.441) than in the Northeast one (0.176), indicating higher compositional differences between the sites. LCBD values ranged between  $7.032\text{e-}5$  and  $1.333\text{e-}3$  for the Northeast subregion and between  $6.035\text{e-}5$  and  $5.236\text{e-}4$  for the Southwest one.

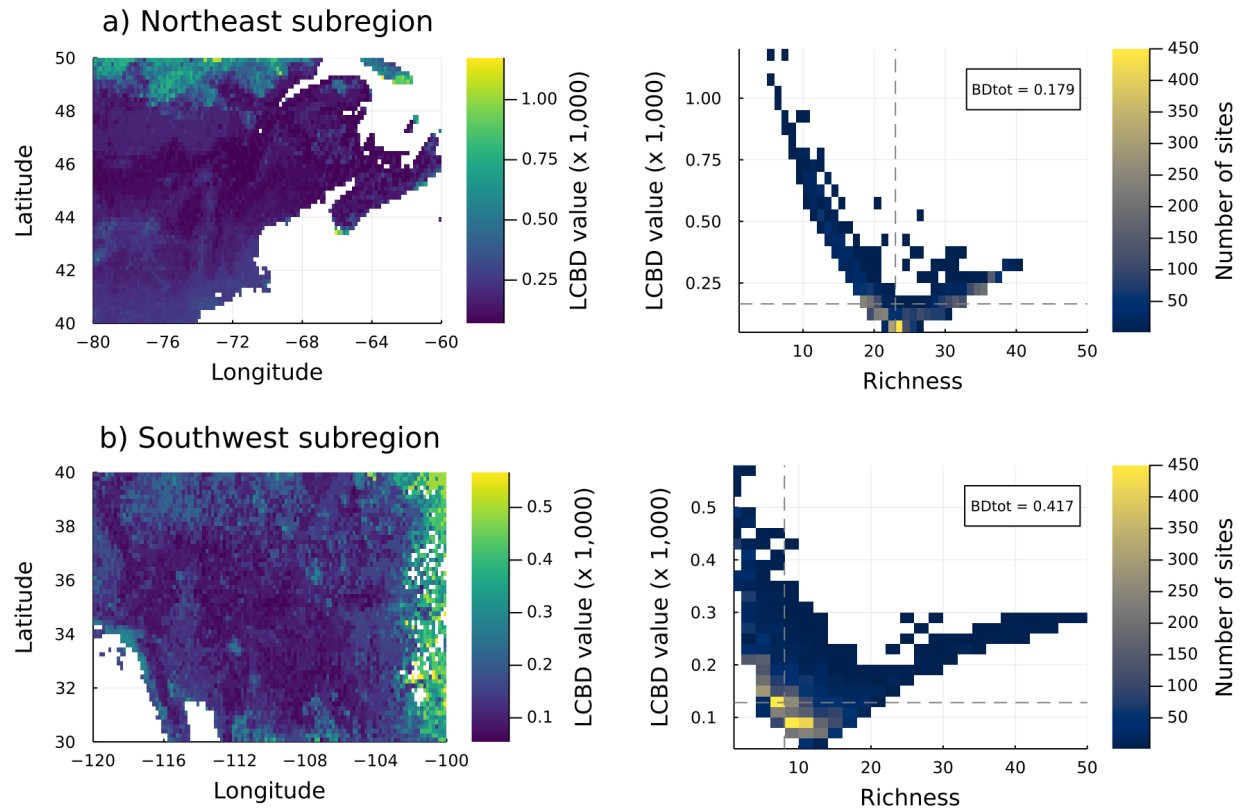




**Figure 3** – Comparison of the regression residuals between the observed and predicted estimates of species richness (a) and ecological uniqueness (b). The estimate from the predicted data set was used as the dependent variable and the estimate from the observed data set as the independent variable. A negative binomial regression with a log link function was used for species richness, and a beta regression with a logit link function was used for uniqueness. The deviance residuals for richness ranged between -3,591 and 4,654 (a). LCBD values were recalculated for the same set of sites with observations in both data sets. The deviance residuals for LCBD scores ranged between -4,866 and 2,799 (b).

### Uniqueness depends on the scale on which it is measured

The LCBD-richness relationship showed important variation when scaling up and changing the region's extent (Fig. 5). For smaller extents, starting with a species-rich region, the relationship is well defined, decreasing, and curvilinear. However, as the extent increases and progressively reaches species-poor regions, the relationship broadens, displays more variance, and loses its curvilinear aspect while keeping a decreasing form. Total beta diversity was higher when increasing the

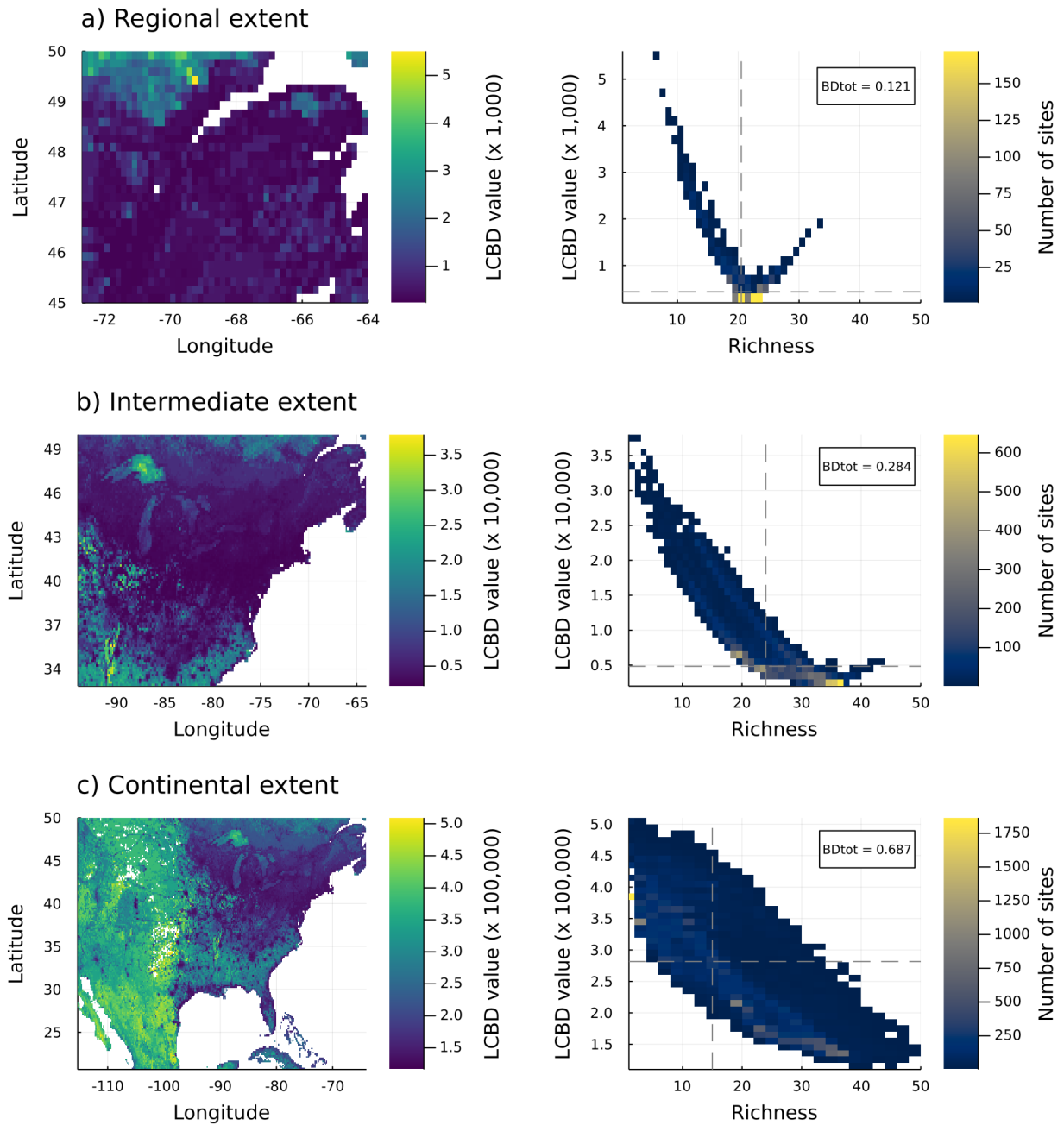


**Figure 4** – Comparison between a species-rich region (Northeast, a) and a species-poor one (Southwest, b) based on the SDM predictions for warbler species in North America. The richness-LCBD relationship displayed contrasting profiles for the subregions according to their general richness. Total beta diversity was higher in the Southwest subregion than in the Northeast one. The left-side figures represent the LCBD scores for the assembled presence-absence predictions, calculated separately in each region after applying the Hellinger transformation. The right-side figures represent the decreasing relationship between LCBD values and species richness, with the number of sites in the bins of the 2-dimensional histogram. The vertical and horizontal dashed lines respectively represent the median richness and LCBD value in each region. LCBD values ranged between  $7.032\text{e-}5$  and  $1.333\text{e-}3$  for the Northeast subregion and between  $6.035\text{e-}5$  and  $5.236\text{e-}4$  for the Southwest one.

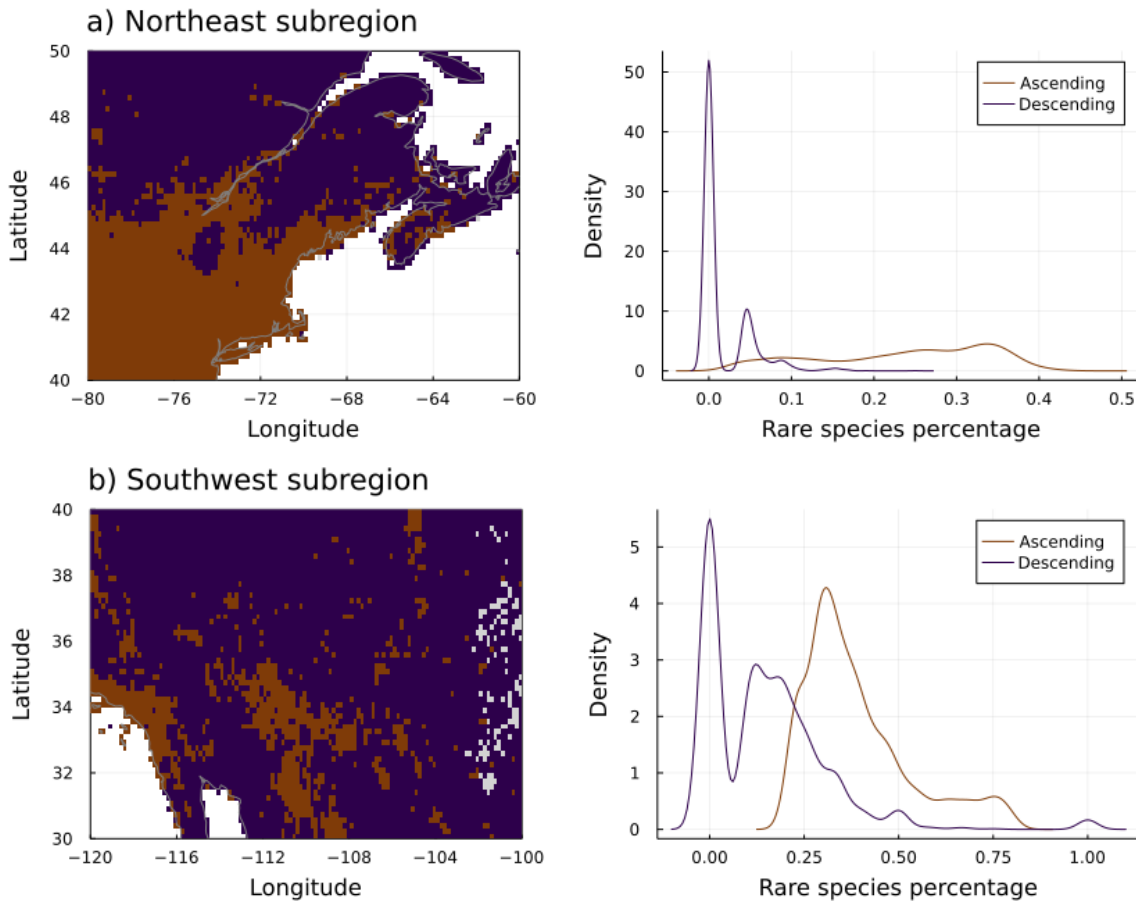
spatial extent, going from 0.116 to 0.279 and 0.682. LCBD values ranged between  $2.195\text{e-}4$  and  $5.209\text{e-}3$  at the finest scale, between  $1.478\text{e-}4$  and  $3.500\text{e-}3$  at the intermediate one, and between  $1.179\text{e-}5$  and  $5.218\text{e-}5$  at the broadest one.

## Uniqueness depends on the proportion of rare species

The proportion of rare species differed depending on the classification of sites in the ascending or descending portions of the LCBD-richness relationship (Fig. 6). The proportion of rare species



**Figure 5** – Effect of extent size on the relationship between site richness and LCB values based on the SDM predictions for warbler species in North America. The relationship progressively broadens and displays more variance when scaling up while total beta diversity increases. The LCB values were recalculated at each scale based on the sites in this region. LCB values ranged between  $2.195 \times 10^{-4}$  and  $5.209 \times 10^{-3}$  at the finest scale, between  $1.478 \times 10^{-4}$  and  $3.500 \times 10^{-3}$  at the intermediate one, and between  $1.179 \times 10^{-5}$  and  $5.218 \times 10^{-5}$  at the broadest one. The vertical and horizontal dashed lines respectively represent the median richness and LCB value in each region.



**Figure 6** – Proportion of rare species in the ascending and descending portions of the LCBD-richness relationship for the Northeast (a) and Southwest (b) subregions. The left side figures show the geographic distribution of the sites from each group. Sites were assigned to the ascending portion if their species richness was higher than the richness of the site with the lowest LCBD value, which corresponds to the inflection point of the right side figures of Fig. 4, and in the descending portion otherwise. The right side figures are density plots of the proportion of rare species in each group. Species were classified as rare when they occurred in fewer than 40% of the sites in the subregion. The proportion of rare species was then calculated for every site.

was higher in the sites corresponding to the ascending portions of the relationships shown in Fig. 4 than in the sites corresponding to the descending portions for both subregions. The classification of the sites in the two portions showed a clear latitudinal gradient in the Northeast subregion, while it was distributed in patches in the Southwest subregion.

## 4. Discussion

Our results showed a decreasing relationship between species richness and LCBD values on broad spatial extents (Fig. 5c) but also highlighted that the exact form of this relationship varies depending on the region and the spatial extent on which it is measured. Our species-rich Northeast subregion (Fig. 4a) showed a decreasing relationship, very similar to previous studies, and slightly curvilinear, as described by Heino & Grönroos (2017) and Tan et al. (2019). This result for warbler species is in line with the original study on fish communities (Legendre & De Cáceres, 2013) and with following ones on insect metacommunities (da Silva & Hernández, 2014; Heino et al., 2017; Heino & Grönroos, 2017), dung beetles (da Silva et al., 2020; da Silva et al., 2018), aquatic beetles (Heino & Alahuhta, 2019), stream macroinvertebrates (Sor et al., 2018), stream diatoms (Vilmi et al., 2017), multi-trophic pelagic food webs (phytoplankton, zooplankton, fish) (Taranu et al., 2020), temperate forest trees (Tan et al., 2019), mammals (medium-to-large, small, volant) (da Silva et al., 2020), wetland birds (de Deus et al., 2020), and a few other phylogenetic groups (plants, lizards, mites, anurans, mesoinvertebrates) (Landeiro et al., 2018). However, it was originally argued that the negative relationship was not general or obligatory (Legendre & De Cáceres, 2013). Different LCBD-richness relationships have also been observed, with both positive and negative relationships for different sites or taxonomic groups in some studies (Kong et al., 2017; Teittinen et al., 2017), as well as a negative relationship with the number of common species but a positive relationship with the number of rare species (Qiao et al., 2015). Our results further show that the relationship may depend on the region's richness profile, as the relationship was different in our species-poor Southwest subregion, with a sharper initial decrease (Fig. 4b). Therefore, the curvilinear form may depend on how big the contrast will be between the region's median richness and its richest ecologically possible sites.

The variation in the LCBD-richness relationship when scaling up and changing the overall study extent showed that the uniqueness patterns highlighted are not necessarily the same depending on the scale on which it is used (Fig. 5). The subregions' uniqueness profiles will merge at broad spatial scales, but this can create a new profile with a lot more variation. When too many poor sites are present, rich sites will almost certainly have lower LCBD values. Aggregating too

many different sites might then possibly mask some patterns of uniqueness. Total beta diversity, on the other hand, showed the variation expected from previous studies, increasing with spatial extent (Fig. 5) (Barton et al., 2013; Heino et al., 2015). Its value was high at the continental scale (0.628) but lower than what has been observed in some studies (e.g. 0.80 in Sor et al. (2018)).

Our results confirm that the proportion of rare species in the community may have an effect on the direction of the relationship between species richness and ecological uniqueness (Fig. 6). da Silva et al. (2018) suggested that the proportion of rare and common species in the communities seems to determine if the relationship will be negative, non-significant, or positive. Yao et al. (2021) supported this idea and showed that the relationship was negative when the proportion of rare species was lower and positive when the relationship was high. Our results further show that sites associated with a positive relationship within a curvilinear one tended to have a higher rare species proportion (Fig. 6). This also implies that the proportion of rare species was higher in species-rich sites than in species-poor ones in both our Northeast and Southwest subregions. Further work should attempt to disentangle the effects of the rare species proportion and of the region's richness profile.

Our results showed that SDM models provide richness and uniqueness predictions highly correlated to the occurrence data while filling gaps in poorly sampled regions (Fig. 1). However, the results showed spatial autocorrelation and tended to underestimate uniqueness in species-rich regions and overestimate it in species-poor ones (Figs. 2, 3). Overprediction of richness using S-SDMs was reported previously (D'Amen et al., 2015; Dubuis et al., 2011; Zurell et al., 2020). Predictive approaches and spatial structures have been less studied for LCBD values compared to species richness, but some studies did show that LCBD distributions were spatially structured across sampling sites (da Silva et al., 2018). Our results over broad spatial extents did not concord exactly with those of Heino & Alahuhta (2019), as LCBD scores increased both in the North and South, hence did not strictly increase with latitude, and also showed a clear East-West gradient. Nonetheless, the predictions for regions with sparse sampling are of interest as they allow a quantitative evaluation, however imperfect, for sites where we would otherwise have no information. Our

SDMs also offered relevant LCBD predictions using eBird, arguably one of the largest presence-absence data sets available (when using its complete checklist system), and showed the measure's potential on such massive data. These two elements open up new opportunities for LCBD analyses on extended spatial scales, as well as for the diversity of taxons to which this type of analysis can be applied. An interesting way forward would be to test these results using more advanced community assembling techniques than S-SDMs. The use SESAM with probabilistic SDMs, probability ranking, and species richness predictions as macroecological constraints returns high site-level prediction accuracy (Zurell et al., 2020) and would be compatible with presence-absence LCBD calculations. The use of probabilistic stacks rather than binary ones (Calabrese et al., 2014) could also constitute a novel way to calculate LCBD indices. Both these procedures should reduce the richness deviation we observed, and it would be interesting to verify if this can also be the case with LCBD values. Overall, our distribution results have implications for conservation, as they confirm that species richness and ecological uniqueness measured from LCBD values may conflict and highlight different potential hotspots (Dubois et al., 2020; Yao et al., 2021), thus reinstating the need to protect both with complementary strategies.

This study showed how ecological uniqueness can be measured on broad spatial extents, including for regions with sparse sampling. First, the negative relationship often observed between species richness and local contributions to beta diversity (LCBD) can take different forms depending on the richness profile of the regions on which it is measured. Therefore, species-rich and species-poor regions may display different ways to be unique. Second, the negative relationship is not constant when varying the spatial study extent and may be less clearly defined at broad scales when contrasting regional relationships are present, even creating a new, distinct one. Finally, species distribution models (SDMs) offer a promising way to generate uniqueness predictions on broad spatial extents and could prove useful to identify beta diversity hotspots in unsampled locations on large spatial scales, which could be important targets for conservation purposes.

## **5. Acknowledgments**

We acknowledge that this study was conducted on land within the traditional unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat, and Omàmiwininiwak nations. We have received financial support from the Fonds de recherche du Québec - Nature et technologie (FRQNT) and the Computational Biodiversity Science and Services (BIOS<sup>2</sup>) NSERC CREATE training program.



## Conclusion

---

Dans l'introduction du présent mémoire, j'ai expliqué comment l'évaluation de l'unicité écologique sur de grandes étendues spatiales passe par une bonne intégration des concepts de biodiversité, des méthodes prédictives et des données à grande échelle, idéalement accompagné d'un examen des facteurs pouvant influencer la diversité en fonction des échelles spatiales. J'ai donc combiné tous ces éléments pour remplir mon objectif, qui était de déterminer si les modèles de répartition d'espèces permettent d'évaluer correctement l'unicité écologique sur de grandes étendues spatiales, de même que de vérifier si la relation entre la richesse spécifique, l'unicité écologique et la présence d'espèces rares est constante lorsqu'évaluée à différentes échelles. Pour ce faire, j'ai utilisé la mesure des contributions locales à la diversité bêta (LCBD), une mesure spatialement explicite permettant d'évaluer l'unicité écologique de sites précis au sein d'une région d'intérêt. J'ai utilisé des données d'occurrences portant sur les parulines en Amérique du Nord provenant de la base de données eBird, ainsi que des données environnementales provenant des bases de données WorldClim et Copernicus. J'ai effectué des modèles de répartition d'espèces selon la méthode des arbres de régression additifs bayésiens (BARTs), puis j'ai assemblé les prédictions par espèces pour former une prédiction des communautés suivant la méthode des modèles de répartition d'espèces superposés (S-SDM). J'ai ensuite calculé les LCBD pour les données observées et pour les prédictions, ainsi que pour différentes sous-régions et différentes étendues spatiales. J'ai finalement comparé les estimations obtenues pour les prédictions avec ceux obtenus pour les données brutes, puis j'ai vérifié la relation entre la richesse spécifique et l'unicité écologique dans les différents ensembles de données, de même que l'effet de la proportion d'espèces rares sur cette même relation.

Dans un premier temps, mes résultats ont montré que la relation entre la richesse spécifique et l'unicité écologique pour un groupe taxonomique donné peut varier selon la région et selon l'étendue spatiale sur laquelle l'unicité est évaluée. Ce résultat est important, car il montre pour une première fois la variation dans cette relation entre différentes étendues spatiales, ainsi que sa forme et sa direction sur une étendue continentale avec un grand nombre de sites. La plupart des études précédentes portant sur la mesure l'ont plutôt utilisée à échelle locale ou régionale (Legendre et De Cáceres, 2013) ou sur de grandes étendues spatiales (Poisot et al., 2017) et des données disposées en grilles uniformes (Legendre et Condit, 2019), mais comportant peu de sites. Considérant le potentiel des LCBD comme mesure spatialement explicite pour identifier des sites exceptionnels précis, mentionné en introduction, il est donc pertinent d'avoir des résultats montrant comment celle-ci se comporte sur de grandes étendues spatiales et un grand nombre de sites. La forme négative de la relation richesse-LCBD à étendue continentale est conforme au résultat observé dans plusieurs études portant sur de plus petites étendues (par exemple da Silva et al., 2018; Heino et al., 2017; Taranu et al., 2020). La forme curvilinéaire à échelle régionale est également conforme à celle observée dans certaines études (Heino et Grönroos, 2017; Tan et al., 2019). Par contre, la variation de cette relation entre les régions selon leur profil général de richesse est une nouvelle information, tout comme le changement dans la forme de la relation et l'augmentation de la variabilité de celle-ci en fonction de l'étendue spatiale. Notre résultat sur l'effet de la proportion d'espèces rares est également conforme à celui montré lors d'études précédentes (da Silva et al., 2018; Qiao et al., 2015; Yao et al., 2021), mais il montre pour une première fois que cette proportion varie au sein même d'une relation curvilinéaire. Ainsi, nos trois résultats sont importants pour l'utilisation future de la mesure des LCBD, car ils montrent que la mesure permet d'identifier des sites exceptionnels selon des caractéristiques différentes entre les régions ou les étendues spatiales. Ce faisant, ils mettent en lumière l'importance de prendre en compte le contexte spatial spécifique lors de l'utilisation de la mesure pour évaluer l'unicité écologique, notamment quant à la richesse et la proportion générale d'espèces rares de la région d'intérêt.

L'approche prédictive que j'ai employée dans cette étude fait suite à celles qui ont été avancées dans les dernières années par Niskanen et al. (2017) et Vasconcelos et al. (2018). Celles-ci constituent les premiers développements d'une approche prédictive plus fréquente pour la mesure des LCBD, qui pourra mener à des évaluations poussées des LCBD à grande échelle spatiale et dans de nouveaux contextes, ainsi que dans une approche de prévision face aux changements climatiques. Les grandes bases de données ouvertes comme eBird, WorldClim et Copernicus permettent des développements impressionnants, dont il faut tirer profit, d'autant plus qu'elles s'accompagnent simultanément du développement de méthodes computationnelles poussées. Notre approche s'inscrit dans la démarche souhaitée par Pollock et al. (2020) visant l'intégration des domaines de la conservation et de la modélisation de la biodiversité, puisque les LCBD peuvent aider à prendre des décisions de conservation ou de gestion d'aires protégées, ou encore permettre d'identifier des endroits à échantillonner en priorité.

Bien que nos modèles prédictifs aient fourni des résultats pertinents et hautement corrélés à ceux des données observées, l'autocorrélation spatiale et la distribution structurée spatialement des valeurs de différences et des résidus indiquent des éléments qui devraient être examinés avec attention dans des travaux futurs. Les prochaines étapes à réaliser pourraient notamment être d'utiliser des méthodes d'assemblage des communautés plus complexes et qui ont montré leur potentiel pour améliorer les prédictions portant sur les communautés. Celles-ci ont généralement été validées en fonction de la richesse spécifique et gagneraient donc à être validées avec d'autres mesures comme les LCBD. Parmi ces modèles plus complexes, la modélisation spatialement explicite des assemblages d'espèces (SESAM) (Guisan et Rahbek, 2011 ; Zurell et al., 2020) pourrait s'avérer une option particulièrement intéressante et plus facile à mettre en place, puisqu'elle permettrait de conserver la structure de la matrice des communautés utilisée dans la plupart des études. D'un autre côté, l'utilisation de superpositions probabilistes (Calabrese et al., 2014) pourrait également constituer une approche novatrice pour le calcul des LCBD, car elle impliquerait de calculer les valeurs sur des probabilités d'occurrence, plutôt que sur des données de présence-absence binaires ou des données d'abondance. Ce faisant, ces nouvelles approches pourraient mener à de meilleures

prédictions de l'unicité écologique sur de grandes étendues spatiales, qui pourraient avoir des retombées importantes pour la conservation et la gestion des aires protégées.

## Bibliographie

---

- Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders, N. J., Cornell, H. V., Comita, L. S., Davies, K. F., Harrison, S. P., Kraft, N. J. B., Stegen, J. C. et Swenson, N. G. (2011). Navigating the Multiple Meanings of  $\beta$  Diversity : A Roadmap for the Practicing Ecologist. *Ecology Letters*, 14(1), 19-28. <https://doi.org/10.1111/j.1461-0248.2010.01552.x>
- Anderson, M. J., Ellingsen, K. E. et McArdle, B. H. (2006). Multivariate Dispersion as a Measure of Beta Diversity. *Ecology Letters*, 9(6), 683-693. <https://doi.org/10.1111/j.1461-0248.2006.00926.x>
- Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E. et Rahbek, C. (2019). Standards for Distribution Models in Biodiversity Assessments. *Science Advances*, 5(1), eaat4858. <https://doi.org/10.1126/sciadv.aat4858>
- Barton, P. S., Cunningham, S. A., Manning, A. D., Gibb, H., Lindenmayer, D. B. et Didham, R. K. (2013). The Spatial Scaling of Beta Diversity. *Global Ecology and Biogeography*, 22(6), 639-647. <https://doi.org/10.1111/geb.12031>
- Baselga, A. (2010). Partitioning the Turnover and Nestedness Components of Beta Diversity. *Global Ecology and Biogeography*, 19(1), 134-143. <https://doi.org/10.1111/j.1466-8238.2009.00490.x>
- Baselga, A. (2013). Multiple Site Dissimilarity Quantifies Compositional Heterogeneity among Several Sites, While Average Pairwise Dissimilarity May Be Misleading. *Ecography*, 36(2), 124-128. <https://doi.org/10.1111/j.1600-0587.2012.00124.x>

- Beck, J., Böller, M., Erhardt, A. et Schwanghart, W. (2014). Spatial Bias in the GBIF Database and Its Effect on Modeling Species' Geographic Distributions. *Ecological Informatics*, 19, 10-15. <https://doi.org/10.1016/j.ecoinf.2013.11.002>
- Bezanson, J., Edelman, A., Karpinski, S. et Shah, V. B. (2017). Julia : A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1), 65-98. <https://doi.org/10.1137/141000671>
- Booth, T. H., Nix, H. A., Busby, J. R. et Hutchinson, M. F. (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>
- Breiman, L. (2001). Random Forests. *Machine learning*, 45(1), 5-32.
- Buchhorn, M., Smets, B., Bertels, L., Lesiv, M., Tsendbazar, N.-E., Herold, M. et Fritz, S. (2019). Copernicus Global Land Service : Land Cover 100m : Epoch 2015 : Globe. <https://doi.org/10.5281/zenodo.3243509>
- Calabrese, J. M., Certain, G., Kraan, C. et Dormann, C. F. (2014). Stacking Species Distribution Models and Adjusting Bias by Linking Them to Macroecological Models. *Global Ecology and Biogeography*, 23(1), 99-112. <https://doi.org/10.1111/geb.12102>
- Carlson, C. J. (2020). Embarcadero : Species Distribution Modelling with Bayesian Additive Regression Trees in R. *Methods in Ecology and Evolution*, 11(7), 850-858. <https://doi.org/10.1111/2041-210X.13389>
- Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L. et Ram, K. (2020). rgbif : Interface to the Global Biodiversity Information Facility API. <https://CRAN.R-project.org/package=rgbif>
- Chao, A. et Ricotta, C. (2019). Quantifying Evenness and Linking It to Diversity, Beta Diversity, and Similarity. *Ecology*, 100(12), e02852. <https://doi.org/10.1002/ecy.2852>
- Chipman, H. A., George, E. I. et McCulloch, R. E. (2010). BART : Bayesian Additive Regression Trees. *Annals of Applied Statistics*, 4(1), 266-298. <https://doi.org/10.1214/09-AOAS285>
- Clifford, P., Richardson, S. et Hemon, D. (1989). Assessing the Significance of the Correlation between Two Spatial Processes. *Biometrics*, 45(1), 123-134. <https://doi.org/10.2307/2532039>

- Cribari-Neto, F. et Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, 34(1), 1-24. <https://doi.org/10.18637/jss.v034.i02>
- Crisp, M. D., Laffan, S., Linder, H. P. et Monro, A. (2001). Endemism in the Australian Flora. *Journal of Biogeography*, 28(2), 183-198. <https://doi.org/10.1046/j.1365-2699.2001.00524.x>
- da Silva, P. G., Bogoni, J. A. et Heino, J. (2020). Can Taxonomic and Functional Metrics Explain Variation in the Ecological Uniqueness of Ecologically-Associated Animal Groups in a Modified Rainforest? *Science of The Total Environment*, 708, 135171. <https://doi.org/10.1016/j.scitotenv.2019.135171>
- da Silva, P. G. et Hernández, M. I. M. (2014). Local and Regional Effects on Community Structure of Dung Beetles in a Mainland-Island Scenario. *PLOS ONE*, 9(10), e111883. <https://doi.org/10.1371/journal.pone.0111883>
- da Silva, P. G., Hernández, M. I. M. et Heino, J. (2018). Disentangling the Correlates of Species and Site Contributions to Beta Diversity in Dung Beetle Assemblages. *Diversity and Distributions*, 24(11), 1674-1686. <https://doi.org/10.1111/ddi.12785>
- D'Amen, M., Dubuis, A., Fernandes, R. F., Pottier, J., Pellissier, L. et Guisan, A. (2015). Using Species Richness and Functional Traits Predictions to Constrain Assemblage Predictions from Stacked Species Distribution Models. *Journal of Biogeography*, 42(7), 1255-1266. <https://doi.org/10.1111/jbi.12485>
- D'Amen, M., Rahbek, C., Zimmermann, N. E. et Guisan, A. (2017). Spatial Predictions at the Community Level : From Current Approaches to Future Frameworks. *Biological Reviews*, 92(1), 169-187. <https://doi.org/10.1111/brv.12222>
- Dansereau, G. et Poisot, T. (2021). SimpleSDMLayers.Jl and GBIF.Jl : A Framework for Species Distribution Modeling in Julia. *Journal of Open Source Software*, 6(57), 2872. <https://doi.org/10.21105/joss.02872>

- D'Antraccoli, M., Bacaro, G., Tordoni, E., Bedini, G. et Peruzzi, L. (2020). More Species, Less Effort : Designing and Comparing Sampling Strategies to Draft Optimised Floristic Inventories. *Perspectives in Plant Ecology, Evolution and Systematics*, 45, 125547. <https://doi.org/10.1016/j.ppees.2020.125547>
- De Cáceres, M., Legendre, P., Valencia, R., Cao, M., Chang, L.-W., Chuyong, G., Condit, R., Hao, Z., Hsieh, C.-F., Hubbell, S., Kenfack, D., Ma, K., Mi, X., Noor, M. N. S., Kassim, A. R., Ren, H., Su, S.-H., Sun, I.-F., Thomas, D., ... Et He, F. (2012). The Variation of Tree Beta Diversity across a Global Network of Forest Plots. *Global Ecology and Biogeography*, 21(12), 1191-1202. <https://doi.org/10.1111/j.1466-8238.2012.00770.x>
- de Deus, F. F., Schuchmann, K.-L., Arieira, J., de Oliveira Tissiani, A. S. et Marques, M. I. (2020). Avian Beta Diversity in a Neotropical Wetland : The Effects of Flooding and Vegetation Structure. *Wetlands*, 40(5), 1513-1527. <https://doi.org/10.1007/s13157-019-01240-0>
- Dubois, R., Proulx, R. et Pellerin, S. (2020). Ecological Uniqueness of Plant Communities as a Conservation Criterion in Lake-Edge Wetlands. *Biological Conservation*, 243, 108491. <https://doi.org/10.1016/j.biocon.2020.108491>
- Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P. et Guisan, A. (2011). Predicting Spatial Patterns of Plant Species Richness : A Comparison of Direct Macroecological and Species Stacking Modelling Approaches. *Diversity and Distributions*, 17(6), 1122-1131. <https://doi.org/10.1111/j.1472-4642.2011.00792.x>
- eBird Basic Dataset. (2019). *Version : EBD\_relJun-2019*. Cornell Lab of Ornithology, Ithaca, NY, USA.
- Ejrnæs, R., Frøslev, T. G., Høye, T. T., Kjøller, R., Oddershede, A., Brunbjerg, A. K., Hansen, A. J. et Bruun, H. H. (2018). Uniquity : A General Metric for Biotic Uniqueness of Sites. *Biological Conservation*, 225, 98-105. <https://doi.org/10.1016/j.biocon.2018.06.034>
- Elith, J., Leathwick, J. R. et Hastie, T. (2008). A Working Guide to Boosted Regression Trees. *Journal of Animal Ecology*, 77(4), 802-813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>



- Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. M., Peterson, A. T., ... Et Zimmermann, N. E. (2006). Novel Methods Improve Prediction of Species' Distributions from Occurrence Data. *Ecography*, 29(2), 129-151. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>
- Ellison, A. M. (2010). Partitioning Diversity. *Ecology*, 91(7), 1962-1963. <https://doi.org/10.1890/09-1692.1>
- Ferrier, S., Drielsma, M., Manion, G. et Watson, G. (2002). Extended Statistical Approaches to Modelling Spatial Pattern in Biodiversity in Northeast New South Wales. II. Community-Level Modelling. *Biodiversity & Conservation*, 11(12), 2309-2338. <https://doi.org/10.1023/A:1021374009951>
- Ferrier, S. et Guisan, A. (2006). Spatial Modelling of Biodiversity at the Community Level. *Journal of Applied Ecology*, 43(3), 393-404. <https://doi.org/10.1111/j.1365-2664.2006.01149.x>
- Fick, S. E. et Hijmans, R. J. (2017). WorldClim 2 : New 1-Km Spatial Resolution Climate Surfaces for Global Land Areas. *International Journal of Climatology*, 37(12), 4302-4315. <https://doi.org/10.1002/joc.5086>
- Fletcher, R. et Fortin, M.-J. (2018). *Spatial Ecology and Conservation Modeling : Applications with R*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-01989-1>
- Franklin, J. (2010). *Mapping Species Distributions : Spatial Inference and Prediction*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511810602>
- GBIF. (s. d.). What Is GBIF? Récupérée 30 avril 2021, à partir de <https://www.gbif.org/what-is-gbif>
- GDAL/OGR contributors. (2021). *GDAL/OGR Geospatial Data Abstraction Software Library*. Manual. Open Source Geospatial Foundation. <https://gdal.org>
- Gotelli, N. J., Anderson, M. J., Arita, H. T., Chao, A., Colwell, R. K., Connolly, S. R., Currie, D. J., Dunn, R. R., Graves, G. R., Green, J. L., Grytnes, J.-A., Jiang, Y.-H., Jetz, W., Lyons, S. K., McCain, C. M., Magurran, A. E., Rahbek, C., Rangel, T. F. L. V. B., Soberón, J., ...

- Et Willig, M. R. (2009). Patterns and Causes of Species Richness : A General Simulation Model for Macroecology. *Ecology Letters*, 12(9), 873-886. <https://doi.org/10.1111/j.1461-0248.2009.01353.x>
- Grinnell, J. (1917a). Field Tests of Theories Concerning Distributional Control. *The American Naturalist*, 51(602), 115-128. <https://doi.org/10.1086/279591>
- Grinnell, J. (1917b). The Niche-Relationships of the California Thrasher. *The Auk*, 34(4), 427-433. <https://doi.org/10.2307/4072271>
- Grinnell, J. (1924). Geography and Evolution. *Ecology*, 5(3), 225-229. <https://doi.org/10.2307/1929447>
- Guerin, G. R., Ruokolainen, L. et Lowe, A. J. (2015). A Georeferenced Implementation of Weighted Endemism. *Methods in Ecology and Evolution*, 6(7), 845-852. <https://doi.org/10.1111/2041-210X.12361>
- Guisan, A. et Rahbek, C. (2011). SESAM – a New Framework Integrating Macroecological and Species Distribution Models for Predicting Spatio-Temporal Patterns of Species Assemblages. *Journal of Biogeography*, 38(8), 1433-1444. <https://doi.org/10.1111/j.1365-2699.2011.02550.x>
- Guisan, A. et Thuiller, W. (2005). Predicting Species Distribution : Offering More than Simple Habitat Models. *Ecology Letters*, 8(9), 993-1009. <https://doi.org/10.1111/j.1461-0248.2005.00792.x>
- Heino, J. et Alahuhta, J. (2019). Knitting Patterns of Biodiversity, Range Size and Body Size in Aquatic Beetle Faunas : Significant Relationships but Slightly Divergent Drivers. *Ecological Entomology*, 44(3), 413-424. <https://doi.org/10.1111/een.12717>
- Heino, J., Bini, L. M., Andersson, J., Bergsten, J., Bjelke, U. et Johansson, F. (2017). Unravelling the Correlates of Species Richness and Ecological Uniqueness in a Metacommunity of Urban Pond Insects. *Ecological Indicators*, 73, 422-431. <https://doi.org/10.1016/j.ecolind.2016.10.006>

- Heino, J. et Grönroos, M. (2017). Exploring Species and Site Contributions to Beta Diversity in Stream Insect Assemblages. *Oecologia*, 183(1), 151-160. <https://doi.org/10.1007/s00442-016-3754-7>
- Heino, J., Melo, A. S., Bini, L. M., Altermatt, F., Al-Shami, S. A., Angeler, D. G., Bonada, N., Brand, C., Callisto, M., Cottenie, K., Dangles, O., Dudgeon, D., Encalada, A., Göthe, E., Grönroos, M., Hamada, N., Jacobsen, D., Landeiro, V. L., Ligeiro, R., ... Et Townsend, C. R. (2015). A Comparative Analysis Reveals Weak Relationships between Ecological Factors and Beta Diversity of Stream Insect Metacommunities at Two Spatial Levels. *Ecology and Evolution*, 5(6), 1235-1248. <https://doi.org/10.1002/ece3.1439>
- Hijmans, R. J. (2020). raster : Geographic Data Analysis and Modeling. <https://CRAN.R-project.org/package=raster>
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. et Jarvis, A. (2005). Very High Resolution Interpolated Climate Surfaces for Global Land Areas. *International Journal of Climatology*, 25(15), 1965-1978. <https://doi.org/10.1002/joc.1276>
- Hijmans, R. J., Phillips, S., Leathwick, J. et Elith, J. (2017). dismo : Species Distribution Modeling. <https://CRAN.R-project.org/package=dismo>
- Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M. et Ladle, R. J. (2015). Seven Shortfalls That Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46(1), 523-549. <https://doi.org/10.1146/annurev-ecolsys-112414-054400>
- Hurlbert, A. H. et Jetz, W. (2007). Species Richness, Hotspots, and the Scale Dependence of Range Maps in Ecology and Conservation. *Proceedings of the National Academy of Sciences*, 104(33), 13384-13389. <https://doi.org/10.1073/pnas.0704469104>
- Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22(0), 415-427. <https://doi.org/10.1101/SQB.1957.022.01.039>
- Hutchinson, G. E. (1959). Homage to Santa Rosalia or Why Are There So Many Kinds of Animals? *The American Naturalist*, 93(870), 145-159. <https://doi.org/10.1086/282070>
- iNaturalist. (s. d.). Récupérée 30 avril 2021, à partir de <https://www.inaturalist.org>

- Isaac, N. J. B. et Pocock, M. J. O. (2015). Bias and Information in Biological Records. *Biological Journal of the Linnean Society*, 115(3), 522-531. <https://doi.org/10.1111/bij.12532>
- Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Gutierrez, V. R., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S. T. et Fink, D. (2020). Analytical Guidelines to Increase the Value of Citizen Science Data : Using eBird Data to Estimate Species Occurrence. *bioRxiv*, 574392. <https://doi.org/10.1101/574392>
- Karger, D. N., Conrad, O., Böhrner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P. et Kessler, M. (2017). Climatologies at High Resolution for the Earth's Land Surface Areas. *Scientific Data*, 4, 170122. <https://doi.org/10.1038/sdata.2017.122>
- Koleff, P., Gaston, K. J. et Lennon, J. J. (2003). Measuring Beta Diversity for Presence–Absence Data. *Journal of Animal Ecology*, 367-382. [https://doi.org/10.1046/j.1365-2656.2003.00710.x@10.1111/\(ISSN\)1365-2656.BIODIV](https://doi.org/10.1046/j.1365-2656.2003.00710.x@10.1111/(ISSN)1365-2656.BIODIV)
- Kong, H., Chevalier, M., Laffaille, P. et Lek, S. (2017). Spatio-Temporal Variation of Fish Taxonomic Composition in a South-East Asian Flood-Pulse System. *PLOS ONE*, 12(3), e0174582. <https://doi.org/10.1371/journal.pone.0174582>
- Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J. et Ma, K. (2019). Evaluating the Popularity of R in Ecology. *Ecosphere*, 10(1), e02567. <https://doi.org/10.1002/ecs2.2567>
- Lande, R. (1996). Statistics and Partitioning of Species Diversity, and Similarity among Multiple Communities. *Oikos*, 76(1), 5-13. <https://doi.org/10.2307/3545743>
- Landeiro, V. L., Franz, B., Heino, J., Siqueira, T. et Bini, L. M. (2018). Species-Poor and Low-Lying Sites Are More Ecologically Unique in a Hyperdiverse Amazon Region : Evidence from Multiple Taxonomic Groups. *Diversity and Distributions*, 24(7), 966-977. <https://doi.org/10.1111/ddi.12734>
- Legendre, P., Borcard, D. et Peres-Neto, P. R. (2005). Analyzing Beta Diversity : Partitioning the Spatial Variation of Community Composition Data. *Ecological Monographs*, 75(4), 435-450. <https://doi.org/10.1890/05-0549>

- Legendre, P. et Condit, R. (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, P. et De Cáceres, M. (2013). Beta Diversity as the Variance of Community Data : Dissimilarity Coefficients and Partitioning. *Ecology Letters*, 16(8), 951-963. <https://doi.org/10.1111/ele.12141>
- Legendre, P. et Legendre, L. (2012). *Numerical Ecology* (Third English edition). Elsevier. <http://www.sciencedirect.com/science/bookseries/01678892/24>
- Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., Chilquillo, E., Rønsted, N. et Antonelli, A. (2015). Estimating Species Diversity and Distribution in the Era of Big Data : To What Extent Can We Trust Public Databases ? *Global Ecology and Biogeography*, 24(8), 973-984. <https://doi.org/10.1111/geb.12326>
- Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., Faure, D., Garnier, E., Gimenez, O., Huneman, P., Jabot, F., Jarne, P., Joly, D., Julliard, R., Kéfi, S., Kergoat, G. J., Lavorel, S., Gall, L. L., Meslin, L., ... Et Loreau, M. (2015). REVIEW : Predictive Ecology in a Changing World. *Journal of Applied Ecology*, 52(5), 1293-1310. <https://doi.org/10.1111/1365-2664.12482>
- Niskanen, A. K. J., Heikkinen, R. K., Väre, H. et Luoto, M. (2017). Drivers of High-Latitude Plant Diversity Hotspots and Their Congruence. *Biological Conservation*, 212, 288-299. <https://doi.org/10.1016/j.biocon.2017.06.019>
- Nix, H. A. (1986). A Biogeographic Analysis of Australian Elapid Snakes. *Atlas of elapid snakes of Australia*, 7, 4-15.
- Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., ... Et Ovaskainen, O. (2019). A Comprehensive Evaluation of Predictive Performance of 33 Species Distribution Models at Species and Community Levels. *Ecological Monographs*, 89(3), e01370. <https://doi.org/10.1002/ecm.1370>

- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. et Wagner, H. (2019). *Vegan : Community Ecology Package*. <https://CRAN.R-project.org/package=vegan>
- Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T. et Abrego, N. (2017). How to Make More out of Community Data ? A Conceptual Framework and Its Implementation as Models and Software. *Ecology Letters*, 20(5), 561-576. <https://doi.org/10.1111/ele.12757>
- Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. et Blair, M. E. (2017). Opening the Black Box : An Open-Source Release of Maxent. *Ecography*, 40(7), 887-893. <https://doi.org/10.1111/ecog.03049>
- Phillips, S. J., Anderson, R. P. et Schapire, R. E. (2006). Maximum Entropy Modeling of Species Geographic Distributions. *Ecological Modelling*, 190(3), 231-259. <https://doi.org/10.1016/j.ecolmodel.2005.03.026>
- Phillips, S. J. et Dudík, M. (2008). Modeling of Species Distributions with Maxent : New Extensions and a Comprehensive Evaluation. *Ecography*, 31(2), 161-175. <https://doi.org/10.1111/j.0906-7590.2008.5203.x>
- Pocock, M. J. O., Tweddle, J. C., Savage, J., Robinson, L. D. et Roy, H. E. (2017). The Diversity and Evolution of Ecological and Environmental Citizen Science. *PLOS ONE*, 12(4), e0172579. <https://doi.org/10.1371/journal.pone.0172579>
- Poisot, T., Gravel, D., Leroux, S., Wood, S. A., Fortin, M.-J., Baiser, B., Cirtwill, A. R., Araújo, M. B. et Stouffer, D. B. (2016). Synthetic Datasets and Community Tools for the Rapid Testing of Ecological Hypotheses. *Ecography*, 39(4), 402-408. <https://doi.org/10.1111/ecog.01941>
- Poisot, T., Guéveneux-Julien, C., Fortin, M.-J., Gravel, D. et Legendre, P. (2017). Hosts, Parasites and Their Interactions Respond to Different Climatic Variables. *Global Ecology and Biogeography*, 26(8), 942-951. <https://doi.org/10.1111/geb.12602>

- Poisot, T., LaBrie, R., Larson, E., Rahlin, A. et Simmons, B. I. (2019). Data-Based, Synthesis-Driven : Setting the Agenda for Computational Ecology. *Ideas in Ecology and Evolution*, 12. <https://doi.org/10.24908/iee.2019.12.2.e>
- Pollock, L. J., O'Connor, L. M. J., Mokany, K., Rosauer, D. F., Talluto, M. V. et Thuiller, W. (2020). Protecting Biodiversity (in All Its Complexity) : New Models and Methods. *Trends in Ecology & Evolution*, 35(12), 1119-1128. <https://doi.org/10.1016/j.tree.2020.08.015>
- Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A. et McCarthy, M. A. (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>
- Qiao, X., Li, Q., Jiang, Q., Lu, J., Franklin, S., Tang, Z., Wang, Q., Zhang, J., Lu, Z., Bao, D., Guo, Y., Liu, H., Xu, Y. et Jiang, M. (2015). Beta Diversity Determinants in Badagongshan, a Subtropical Forest in Central China. *Scientific Reports*, 5(1), 17043. <https://doi.org/10.1038/srep17043>
- R Core Team. (2020). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ricotta, C. (2017). Of Beta Diversity, Variance, Evenness, and Dissimilarity. *Ecology and Evolution*, 7(13), 4835-4843. <https://doi.org/10.1002/ece3.2980>
- Socolar, J. B., Gilroy, J. J., Kunin, W. E. et Edwards, D. P. (2016). How Should Beta-Diversity Inform Biodiversity Conservation? *Trends in Ecology & Evolution*, 31(1), 67-80. <https://doi.org/10.1016/j.tree.2015.11.005>
- Sor, R., Legendre, P. et Lek, S. (2018). Uniqueness of Sampling Site Contributions to the Total Variance of Macroinvertebrate Communities in the Lower Mekong Basin. *Ecological Indicators*, 84, 425-432. <https://doi.org/10.1016/j.ecolind.2017.08.038>
- Staniczenko, P. P. A., Sivasubramaniam, P., Suttle, K. B. et Pearson, R. G. (2017). Linking Macroecology and Community Ecology : Refining Predictions of Species Distributions Using Biotic Interaction Networks. *Ecology Letters*, 20(6), 693-707. <https://doi.org/10.1111/ele.12770>

- Strimas-Mackey, M., Miller, E. et Hochachka, W. (2018). auk : eBird Data Extraction and Processing with AWK. <https://cornelllabofornithology.github.io/auk/>
- Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D. et Kelling, S. (2009). eBird : A Citizen-Based Bird Observation Network in the Biological Sciences. *Biological Conservation*, 142(10), 2282-2292. <https://doi.org/10.1016/j.biocon.2009.05.006>
- Tan, L., Fan, C., Zhang, C., von Gadow, K. et Fan, X. (2017). How Beta Diversity and the Underlying Causes Vary with Sampling Scales in the Changbai Mountain Forests. *Ecology and Evolution*, 7(23), 10116-10123. <https://doi.org/10.1002/ece3.3493>
- Tan, L., Fan, C., Zhang, C. et Zhao, X. (2019). Understanding and Protecting Forest Biodiversity in Relation to Species and Local Contributions to Beta Diversity. *European Journal of Forest Research*, 138(6), 1005-1013. <https://doi.org/10.1007/s10342-019-01220-3>
- Taranu, Z. E., Pinel-Alloul, B. et Legendre, P. (2020). Large-Scale Multi-Trophic Co-Response Models and Environmental Control of Pelagic Food Webs in Québec Lakes. *Oikos*, n/a(n/a). <https://doi.org/10.1111/oik.07685>
- Teittinen, A., Wang, J., Strömgård, S. et Soininen, J. (2017). Local and Geographical Factors Jointly Drive Elevational Patterns in Three Microbial Groups across Subarctic Ponds. *Global Ecology and Biogeography*, 26(8), 973-982. <https://doi.org/10.1111/geb.12607>
- Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M. A. et Parrish, J. K. (2015). Global Change and Local Solutions : Tapping the Unrealized Potential of Citizen Science for Biodiversity Research. *Biological Conservation*, 181, 236-244. <https://doi.org/10.1016/j.biocon.2014.10.021>
- Tuanmu, M.-N. et Jetz, W. (2014). A Global 1-Km Consensus Land-Cover Product for Biodiversity and Ecosystem Modelling. *Global Ecology and Biogeography*, 23(9), 1031-1045. <https://doi.org/10.1111/geb.12182>
- Vallejos, R., Osorio, F. et Bevilacqua, M. (2020). *Spatial Relationships between Two Georeferenced Variables : With Applications in r*. Springer. <http://srb2gv.mat.utfsm.cl/>



- Vasconcelos, T. S., do Nascimento, B. T. M. et Prado, V. H. M. (2018). Expected Impacts of Climate Change Threaten the Anuran Diversity in the Brazilian Hotspots. *Ecology and Evolution*, 8(16), 7894-7906. <https://doi.org/10.1002/ece3.4357>
- Veech, J. A., Summerville, K. S., Crist, T. O. et Gering, J. C. (2002). The Additive Partitioning of Species Diversity : Recent Revival of an Old Idea. *Oikos*, 99(1), 3-9. <https://doi.org/10.1034/j.1600-0706.2002.990101.x>
- Vellend, M. (2001). Do Commonly Used Indices of  $\beta$ -Diversity Measure Species Turnover? *Journal of Vegetation Science*, 12(4), 545-552. <https://doi.org/10.2307/3237006>
- Venables, W. N. et Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. <http://www.stats.ox.ac.uk/pub/MASS4/>
- Vilmi, A., Karjalainen, S. M. et Heino, J. (2017). Ecological Uniqueness of Stream and Lake Diatom Communities Shows Different Macroecological Patterns. *Diversity and Distributions*, 23(9), 1042-1053. <https://doi.org/10.1111/ddi.12594>
- Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, 30(3), 279-338. <https://doi.org/10.2307/1943563>
- Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon*, 21(2/3), 213-251. <https://doi.org/10.2307/1218190>
- Yang, J., Sorte, F. A. L., Pyšek, P., Yan, P., Nowak, D. et McBride, J. (2015). The Compositional Similarity of Urban Forests among the World's Cities Is Scale Dependent. *Global Ecology and Biogeography*, 24(12), 1413-1423. <https://doi.org/10.1111/geb.12376>
- Yao, J., Huang, J., Ding, Y., Xu, Y., Xu, H. et Zang, R. (2021). Ecological Uniqueness of Species Assemblages and Their Determinants in Forest Communities. *Diversity and Distributions*, 27(3), 454-462. <https://doi.org/10.1111/ddi.13205>
- Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T. et Wüest, R. O. (2020). Testing Species Assemblage Predictions from Stacked and Joint Species Distribution Models. *Journal of Biogeography*, 47(1), 101-113. <https://doi.org/10.1111/jbi.13608>



**Second Article.**

# **SimpleSDMLayers.jl and GBIF.jl: A Framework for Species Distribution Modelling in Julia**

by

Gabriel Dansereau<sup>1</sup>, and Timothée Poisot<sup>1</sup>

(<sup>1</sup>) Département de sciences biologiques, Université de Montréal  
1375 avenue Thérèse-Lavoie-Roux, Montréal, QC, Canada H2V 0B3

This article was published in Journal of Open Source Software.

GD and TP developed the software. GD wrote the manuscript.

## **Summary**

Predicting where species should be found in space is a common question in ecology and biogeography. Species distribution models (SDMs), for instance, aim to predict where environmental conditions are suitable for a given species, often on continuous geographic scales. Such analyses require the use of geo-referenced data on species distributions coupled with climate or land cover

information, hence a tight integration between environmental data, species occurrence data, and spatial coordinates. Thus, it requires an efficient way to access these different data types within the same software, as well as a flexible framework on which to build various analysis workflows. Here we present `SimpleSDMLayers.jl` and `GBIF.jl`, two packages in the *Julia* language implementing a simple framework and type-system on which to build SDM analyses, as well as providing access to popular data sources for species occurrences and environmental conditions.

## Statement of need

Species distribution modeling (SDM) is an increasingly growing field in ecology and biogeography, with many applications in biodiversity assessment, management, and conservation (Araújo et al., 2019). Most SDM models aim at predicting a species distribution in space based on environmental data and information on where the species was previously seen. Hence, SDM studies require a tight and efficient integration between geo-referenced environmental and species occurrence data. However, such data are complex to handle and often require different software: climate and land use data are stored as layers in raster files, then visualized and manipulated in specialized GIS (geographic information systems) software, while occurrence data are stored in tables and spreadsheets, then manipulated in data analysis and statistics-oriented tools or programming languages. Therefore, there is a need for efficient tools to manipulate bioclimatic data, specifically oriented towards species distribution modeling.

In recent years, *R* (R Core Team, 2020) has become the most widely used programming language in ecology, especially in spatial ecology studies (Lai et al., 2019). Hence, many efficient packages and tools for species distribution modeling have been developed in *R*. For instance, the package `raster` (Hijmans, 2020) can be used to manipulate raster format data (for example climatic or land use data), `dismo` (Hijmans et al., 2017) implements many SDM models and provides access to common climatic data sources, and `rgbif` (Chamberlain et al., 2020) provides access to the GBIF database, a common source of species occurrence data in SDM studies. In comparison, few specific SDM resources currently exist for the *Julia* language (Bezanson et al., 2017), although SDM models could benefit from its speed, efficiency and ease of writing (removing the need to

rewrite functions in other languages for higher performance, as in *R*). There are currently packages such as `GDAL.jl` and `ArchGDAL.jl` to manipulate raster data; however, these are lower-level implementations than what is typically used by most ecologists, and they lack support for common layer manipulation. generalized linear models (`GLM.jl`), random forests (`DecisionTrees.jl`), neural networks (`Flux.jl`), and other commonly used models have excellent implementations in *Julia*, although not oriented towards species distribution modeling and raster format data.

`SimpleSDMLayers.jl` is a package to facilitate manipulation of geo-referenced raster data in *Julia*, specifically aimed at species distribution modeling. It is a higher-level implementation, building on `ArchGDAL.jl`, that is easier to use and provides support for common SDM operations (see *Feature Overview* section below). The package implements simple type structures to manipulate the input and output data of SDM models, and is meant to be a flexible framework on which to build more complex analyses. While it does not implement SDM models in itself, we believe the package is a step that will make *Julia* more popular for species distribution modeling, and lead to the development of more complete implementations. `SimpleSDMLayers.jl` also offers built-in access to some of the most common data sources in SDM studies, such as the WorldClim 2.1 climatic data, which is the most common source of climate data in SDM studies (Booth et al., 2014). The package is also tightly integrated with `GBIF.jl`, which allows easy access to the GBIF database, a common data source for species occurrences. Both `SimpleSDMLayers.jl` and `GBIF.jl` are part of the *EcoJulia* organization, whose aim is to integrate a variety of packages for ecological analyses in *Julia*.

## Basic structure

The core structure implemented in `SimpleSDMLayers.jl` is the `SimpleSDMLayer` type, with two variants, `SimpleSDMPredictor` and `SimpleSDMResponse`, depending if the layer is meant to be mutable or not.

A `SimpleSDMLayer` element is made of a grid field, which contains the raster data as a simple `Array` (matrix) of any type, easily manipulable. It also contains the fields `left`, `right`, `bottom`, and `top`, representing the bounding coordinates of the layer.

To illustrate this structure, the following code loads a layer of WorldClim 2.1 climate data, which also shows how easily this can be done in a single call. By default, this will return a layer with the values for the whole world if no bounding coordinates are specified.

```
# Load package
using SimpleSDMLayers

# Get world temperature data
temperature = worldclim(1)

SimpleSDMLayers.SimpleSDMPredictor{Float32} (Union{Nothing, Float32} [-31.017
105f0 -31.62153f0      -32.81253f0 -31.620333f0; -30.391916f0 -31.63478f0
-32.81005f0 -30.995281f0;      ; nothing nothing      nothing nothing; nothing
nothing      nothing nothing], -180.0, 180.0, -90.0, 90.0)
```

The raster values can be displayed by calling the `grid` field.

```
# Display data grid
temperature.grid

1080 2160 Array{Union{Nothing, Float32},2}:
-31.0171  -31.6215  -31.6227      -32.8129  -32.8125  -31.6203
-30.3919  -31.6348  -31.6341      -32.8092  -32.8101  -30.9953
-33.4822  -34.1494  -34.1493      -35.4658  -35.4633  -34.1374
-33.6104  -34.2875  -34.2865      -35.596   -35.5931  -34.2528
-33.7199  -34.4041  -34.4014      -35.6932  -35.691   -34.3311
-33.8224  -34.5184  -34.5162      -35.8037  -35.7996  -34.4165
-31.6613  -32.3194  -32.3184      -33.5133  -33.5101  -32.2032
-31.7635  -32.4307  -33.7036      -34.9522  -33.6282  -32.3038
-33.7063  -36.0738  -39.2075      -40.6438  -37.3938  -34.3026
```

-33.9768	-34.7016	-35.8662	-37.2408	-36.0364	-34.5988
nothing	nothing	nothing	nothing	nothing	nothing
nothing	nothing	nothing	nothing	nothing	nothing
nothing	nothing	nothing	nothing	nothing	nothing
nothing	nothing	nothing	nothing	nothing	nothing
nothing	nothing	nothing	nothing	nothing	nothing
nothing	nothing	nothing	nothing	nothing	nothing
nothing	nothing	nothing	nothing	nothing	nothing
nothing	nothing	nothing	nothing	nothing	nothing
nothing	nothing	nothing	nothing	nothing	nothing

`SimpleSDMLayers.jl` then makes it very simple to plot and visualize the layer as a map using `Plots.jl` (Figure A1).

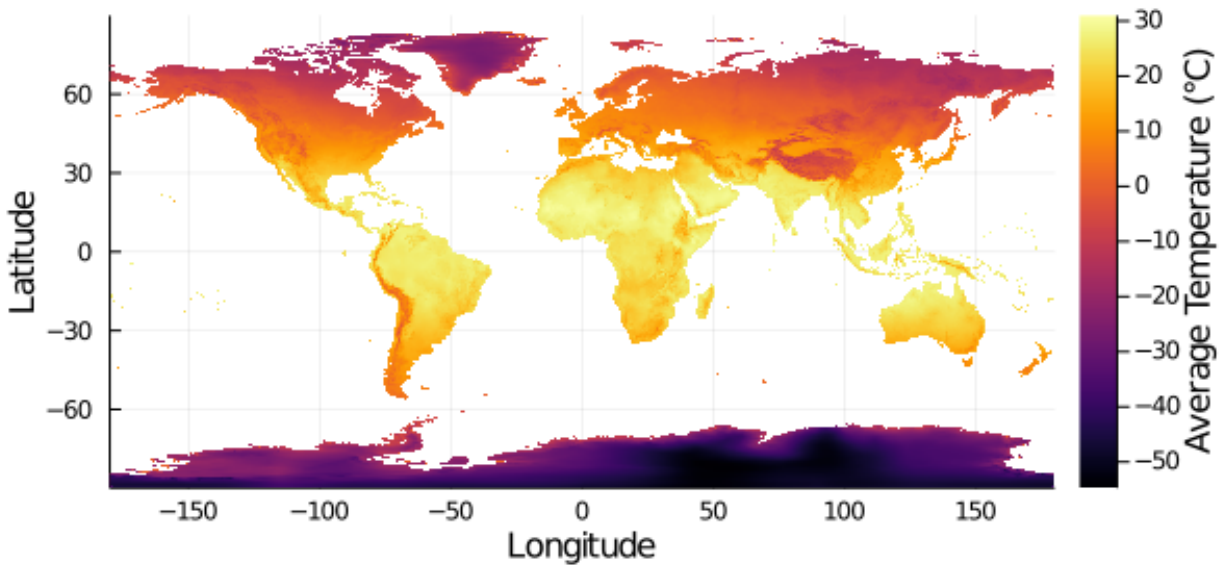
```
using Plots
```

```
plot(temperature)
```

## Feature overview

`SimpleSDMLayers.jl` implements the following features:

- **Overloads for common functions:** The `SimpleSDMLayer` types are implemented along with overloads for many common functions and operations, such as subsetting, changing values, copying, and iterating. Therefore, the layers and the raster values stored in the `grid` field can be manipulated as easily as any `Array`, without losing their spatial aspect.
- **Statistical operations on layer values:** Common operations can be performed directly on the layer values without worrying about the underlying structure (for example, `sum`, `minimum`, `maximum`, `mean`, `median`).



**Figure A1** – Map of the average annual temperature data from WorldClim 2.1, accessed as a layer through SimpleSDMLayers.jl

- **Statistical operations on multiple layers:** Operations can also be performed between layers to produce a new layer, just as `Arrays`, as long as they share the same coordinates. For instance, two layers can be added or subtracted, and calling `mean()` will produce a new layer with the mean value per pixel.
- **Spatial operations:** `SimpleSDMLayers.jl` implements spatial operations such as clipping a layer to given coordinates, coarsening the resolution by grouping values, and performing sliding window operations given a certain radius.
- **Datasets supported:** The package provides access to climate data at different resolutions from WorldClim 2.1 and CHELSA, as well as land cover data from EarthEnv. Custom raster data can be loaded as well.



- **Plotting recipes:** Default recipes are implemented for the `SimpleSDMLayer` types, allowing to directly map them, view the grid data as histograms and density plots, or compare layers as 2-dimensional histograms.
- **Integration with GBIF.jl (and DataFrames.jl):** `SimpleSDMLayer.jl` is well integrated with `GBIF.jl`, allowing to clip layers based on the occurrence data, as well as to map occurrences by displaying them over the layers. Both packages also offer an integration with `DataFrames.jl` to easily convert environmental and occurrence data to a table format.

## Examples

Spatial operations. To illustrate a few of the spatial operations supported by `SimpleSDMLayers.jl`, the following code reuses the previous temperature layer, and shows how it is possible to : 1) clip the layer to a region of interest (Europe for instance); 2) coarsen the resolution by averaging groups of cells for large scale analyses; and 3) perform sliding window operations to aggregate values for each site based on a certain radius. Each of these operations can be performed in a single command and returns new layers, which can then be plotted as shown previously.

```
using Statistics

# Clip to Europe
temperature_europe = temperature[left = -11.2, right = 30.6, bottom = 29.1,
    top = 71.0]

# Coarsen resolution
temperature_coarse = coarsen(temperature_europe, Statistics.mean, (4, 4))

# Sliding window averaging
temperature_slided = slidingwindow(temperature_europe, Statistics.mean, 100.0)
```

GBIF integration. The following example shows how the integration between `SimpleSDMLayers.jl` and `GBIF.jl` allows to easily map the occurrences of any species in GBIF. The species represented in this example is the belted kingfisher (*Megaceryle alcyon*).

`GBIF.jl` first allows us to retrieve the latest occurrences from the GBIF database. Note that the element returned here uses the `GBIFRecords` format, which contains the metadata associated to each GBIF occurrence.

```
using GBIF

kingfisher = GBIF.taxon("Megaceryle alcyon", strict=true)
kf_occurrences = occurrences(kingfisher)

# Get at least 200 occurrences
while length(kf_occurrences) < 200
    occurrences!(kf_occurrences)
    @info "$(length(kf_occurrences)) occurrences"
end

kf_occurrences
```

```
GBIF records: downloaded 200 out of 100000
```

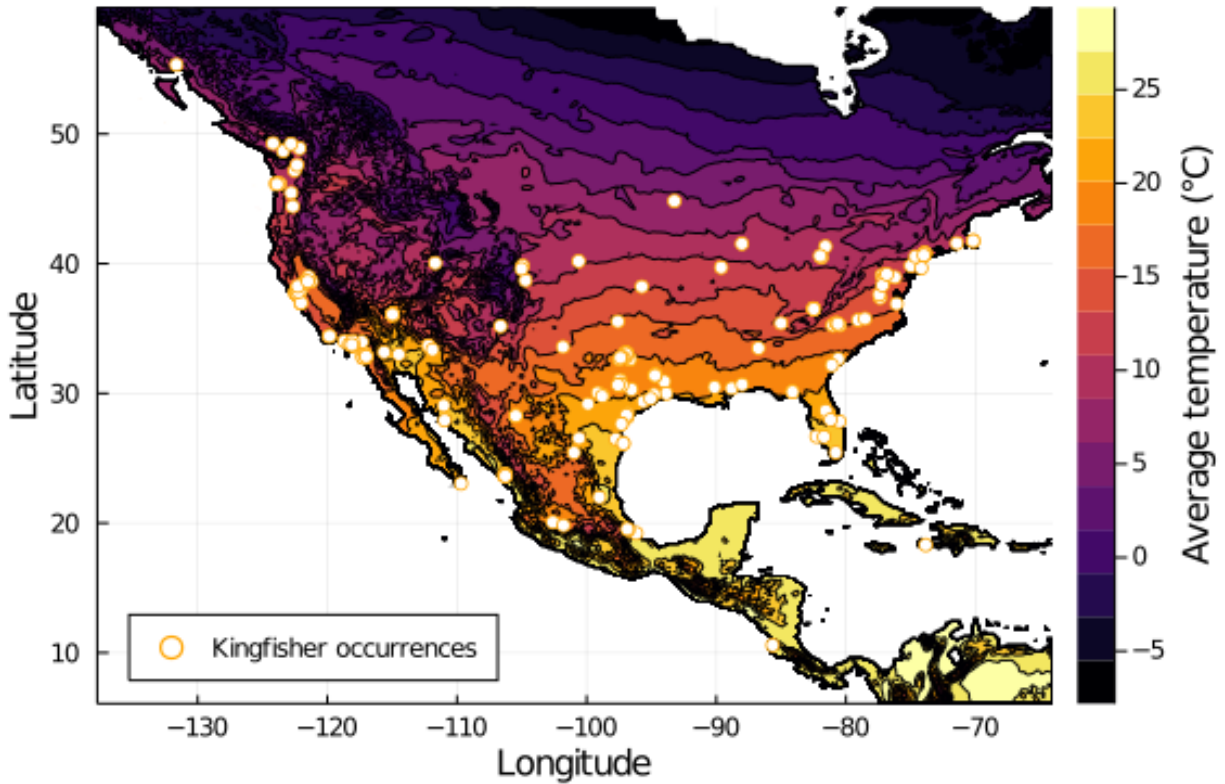
`SimpleSDMLayers.jl` then provides a simple integration between the occurrence data and the environmental layers. The layers can be clipped to the spatial extent of the occurrences in a single call using the `clip()` function. The occurrences' coordinates can also be extracted with `longitudes()` and `latitudes()`. Using these functions, we can easily create a map of the occurrences by overlaying them on top of the clipped environmental layer (Figure A2).

```
# Clip layer to occurrences

temperature_clip = clip(temperature, kf_occurrences)


# Plot occurrences

contour(temperature_clip, fill = true)
```



**Figure A2** – Latest belted kingfisher occurrences from the GBIF database displayed over the temperature data through the integration between SimpleSDMLayers.jl and GBIF.jl

```
scatter!(longitudes(kf_occurrences), latitudes(kf_occurrences))
```

## Acknowledgements

We would like to thank all contributors to the *EcoJulia* organization for their help in developing this series of packages for ecological research. Funding was provided by Fonds de recherche du Québec - Nature et technologies (FRQNT) and the Computational Biodiversity Science and Services (BIOS<sup>2</sup>) training program.