A Functional Investigation of Species Richness Relationships in the North American Tree Flora

Asger Svenning
2022-06-11

Aarhus University

Table of Contents

1 Introduction

1.1 Functional analysis & Ecosystem functioning

Global crises are threatening critical ecosystem functioning for peoples and species on a global scale[1]. Understanding the link between community composition and ecosystem services requires mapping functional attribute-space to ecosystem properties and derived downstream services[2]. By measuring the composition-function-service link, functional community ecology and biogeography could guide conservation ecology, by determining the possible changes to ecosystem-provided services as a result of biodiversity decline. Specifically, analyses of the patterns in functional traits are increasingly being used to investigate the interactions between ecology and evolution[3,4], community assemblage patterns[5–8], and drivers of ecosystem functioning[9,10]. Furthermore, functional ecology has been used to explain how species are distributed in niche-space, the drivers of these patterns and their connections to ecosystem services, and resilience to global change[3,9,11–13]. Therefore, functional ecology and biogeography has the potential to shape conservation actions by informing policymakers and other stakeholders of the value and vulnerability of species and ecosystems.

Functional analysis have been be applied at different ecosystem scales; from the local-community to biogeographical scale. At community scale, functional analysis are primarily used in local assessments or experiments, focused on explaining species coexistence patterns[9,14], ecosystem services[15,16] and invasiveness[17]. Importantly, due to the local perspective and scale community centered studies can relatively easily obtain and integrate (relative) abundances[14–16,18]. Additionally, these local efforts are often coupled with both site- and individual-based measurements of functional traits[14], which has been argued as being an important feature in functional analysis[19].

The biogeographical perspective focuses on understanding the ecosystem-level effects of compositional changes due to past[20] and ongoing[21] environmental changes, predicting vegetation types[22], ecosystem services[10,13,19] and investigating the relationship between niche-space and biodiversity patterns[3–5,7,8]. In these approaches range maps or regional floras are often used to define the presence of species[5,8], given the limited number of datasets that summarize relative abundances for complete assemblages in a standardized way (see[22,23] for exceptions). Furthermore, biogeographical literature has, for the most part, intentionally limited the used trait databases to ecological attributes that define different and important functional dimensions[7,8,10,22–24].

A key feature determining the difference between the community and biogeographical perspective is the use of abundances vs presence-only or presence/absence to define the topological attributes (i.e., variability and evenness) of the functional space under evaluation. The literature acknowledges that presence data will introduce biases in describing the functional space topological attributes[7], but these biases are poorly understood. However, as the amount of available occurrence data increases thanks to standardized large scale inventory protocols (e.g. Forest Inventory and Analysis[25] and breeding bird surveys[26,27]) and broad coverage (functional) trait databases (e.g. PLANTS[28], TRY[29] and EltonTraits[30]), we can now assess these biases.

This work investigates how shifting the focus from presences to abundance in functional analyses at biogeographic scales affects topological attributes of the functional space of woody plants in Eastern North America. This work will also determine how the size of the species pool and the evenness in the composition of the evaluated assemblages determine the discrepancies between presences and abundance estimates. The null expectation of this work is that integrating abundances will not result in greater than random increases in the functional space topological attributes that quantify central tendency and/or clustering in functional space. The alternative hypothesis is then, that integrating abundances will result in a greater than random increase in the central tendency and clustering of species in functional space, suggesting that presence data overly emphasizes rare and functionally unique species. The analyses and results presented here are intended to highlight when and how significant the use of abundances is when describing the functional space.

2 Methods and Materials

2.1 Data sources

To obtain an estimate of the relative abundances (quantified by individual counts) of woody plant species in the study region, I utilize the Forest Inventory and Analysis[25] dataset, which includes over 22 million individual trees covering 418 species and varieties of woody plants across the contiguous Unites States from the period 1968-2021. The geographical size of this dataset allows for large-scale analysis (approximately 7.8 million square kilometers), while the large number of study units and individuals permits medium scale assemblage aggregation to a 50×50 km grid using the NAD27 US National Atlas Equal Area coordinate projection. In order to aggregate the individual counts from different study units from different time periods, I have chosen to quantify the abundances as the number of individuals per plot per year using a two-step procedure; (1) individuals for each species in each grid cell for each year are tallied, (2) then I calculate the weighted mean of the tallies, by the number of study units in each grid cell for each year.

The abundance data is paired with the PLANTS trait-database[28], which includes over 80 traits, of which I have deemed a subset (n = 50) to be both non-redundant and functionally relevant (more details in appendix B). The utilized subset of traits are composed of 10 continuous and 40 ordinal/binary traits. These high quality data sources allows for a detailed analysis across a broad ecological range.

However, in the pool of species which are present in the FIA[a] data set only 123 of 372 have complete trait coverage in the PLANTS database. The remaining 249 species are missing trait information for an average of ~23 traits per species. The missing values are imputed following a simple scheme; specifically missing values are imputed as the mean genus values of all woody plants species in the PLANTS dataset. Following this operation the average number of missing traits, for species with missing traits, is lowered to approximately 1.67 traits per species. Traits which still have more than a few missing values[b] are then removed, followed by finally removing remaining species with missing values. This approach was chosen based on an assumption of phylogenetic conservatism and its simplicity, combined with the relatively low amount of values that needed to be imputed, due to the high initial trait coverage. Following the procedure the utilized data consisted of a total of 49 effective traits for 244 species across 68 genera. See appendix B for further details on the traits used.

2.2 Functional indices

I have chosen to focus on the functional indices described in Villéger et al. 2008[11] and Laliberté & Legendre 2010[31] (functional dispersion, divergence, evenness and richness). I have chosen to mirror some of their methodological choices (using Gower dissimilarity as well as principal coordinate analysis for dimensionality reduction). A noteable methodological difference is that I use of Gower dissimilarity as the distance/dissimilarity measure for all indices, not only those that require dimensionality reduction, never Euclidean distance. This resolves a minor issue, which Laliberté & Legendre themselves note:

> *"FDis also satisfies all criteria but the first one (i.e., to be constrained between 0 and 1 for convenience) of Masonet al. (2003) if traits are standardized prior to its computation..."*[31]

Since Gower dissimilarity is itself constrained between 0 and 1 (unlike most other distance/dissimilarity measures, which are mostly constrained between 0 and $+\infty$), all four functional indices which I use in this project, are also restricted between 0 and 1 (see appendix A for further explanation). I interpret the functional indices as follows:

- **Functional dispersion (FDis)**: This metric is the is the average dissimilarity ("distance") to the "average" (community weighted mean) species (i.e. centroid) in functional space. Thus a lower functional

[a] In the contiguous US, not in e.g. Alaska.

[b] 1 out of every 50 species missing. This is the only parameter that was chosen manually in the procedure, and the specific value was chosen to balance the number of species and traits removed.

dispersion corresponds to a stronger central tendency. Functional dispersion can then be understood as the degree to which species/individuals occupy a common niche within a community[31].

- **Functional evenness (FEve)**: This metric is the abundance-weighted evenness of the branch length of the minimum spanning tree (MST) of species in functional space. A high functional evenness is then achieved when the functional space is partitioned evenly[c], that is when species/individuals are uniformly spread *not* clustered in functional space[11].
- **Functional divergence (FDiv)**: This metric is a measure of the right-skewness of the dissimilarity from the un-weighted convex hull centroid distribution, and is strongly related to Rao's Quadratic Entropy (**Q**), however unlike **Q** functional divergence is not associated with functional richness[12]. A community with a high functional divergence then contains relatively many species/individuals which exhibit "extreme" functional traits[11].
- **Functional Richness (FRic)**: This metric is the convex hull hypervolume of the set of species in a community. Functional richness then measures the size of the utilized niche-space in a community[11].

2.3 Statistical modelling

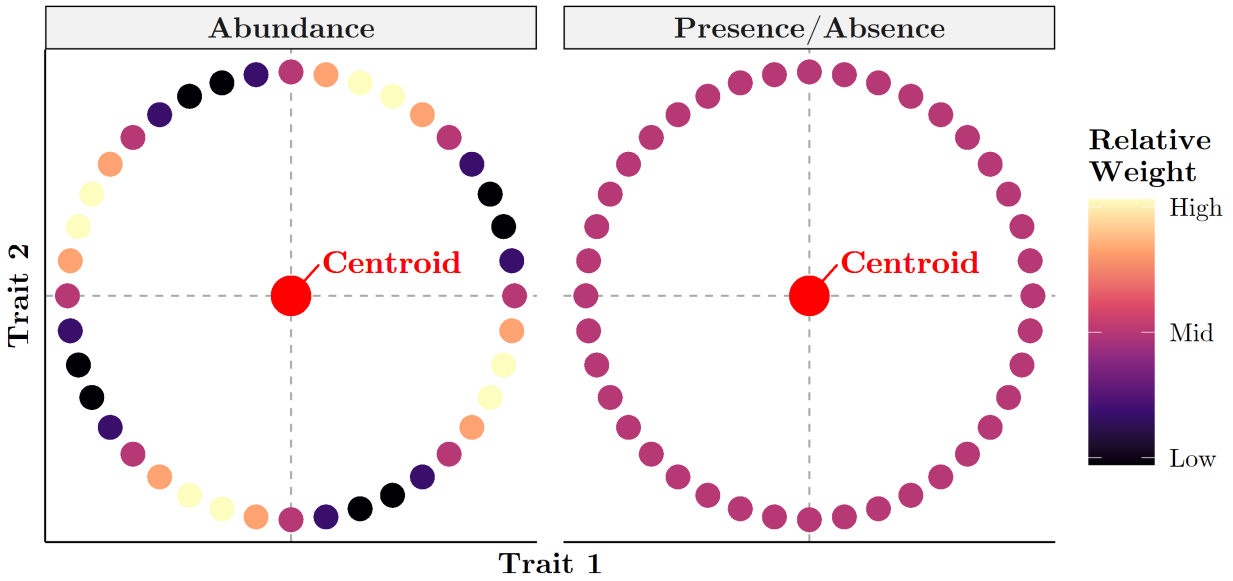

Figure 1

Functional dispersion does not necessarily change based on weights. It is not obvious for which, if any, of the functional indices, a Shannon evenness of less than one, is a sufficient condition for a change in the value of the index. For this purpose I have constructed a set of species, which are distributed in a manner in which all distances to the centroid are equal, and a set of uneven weights, which do not result in a shift in the centroid location.

To capture the environmental gradients in the analyses, I disaggregate the results across different "ecosystems" based on the EPA ecoregion framework[32]. I found that the EPA ecoregions delineate distinct woody plant assemblages or ecosystems at the scale of my analysis (50 × 50 km grid), based on a support vector machines (SVM) classification on the relative abundance vectors. Before classification the abundance vectors are ordinated using principal coordinate analysis[d], and the number of axes are determined based on a broken stick approach (n = 18). The repeated (n=5) k-fold (k=5) cross-validation accuracy is between

[c] Along the minimum spanning tree.
[d] The dissimilarity metric is Bray-Curtis, i.e. manhattan distance divided by two.

[0.819, 0.976] with a mean of 0.935[e]. The confusion matrix for the SVM model can be found in appendix C, also visually confirms the validity of the result.

I calculate the functional indices (FDis, FDiv, FEve & FRic) for all communities (grid cells) twice, once with the relative abundances from FIA and once where all abundances that are not zero are standardized to one, i.e. presence data. Conceptually the standardization of abundances is then treated as a "treatment" and each grid cell as a replicate. I define bias in the context of this procedure as the difference between the functional index value given the true abundances and the standardized "presence" abundances for a grid cell; Bias : $\text{Index}_{\text{Abundance}} - \text{Index}_{\text{Presence}}$. To determine the drivers of bias in the functional indices I have chosen to use generalized additive mixed models (GAMM), due to their great flexibility in mixing parametric, semi-parametric (here thin plate regression splines[33]) and random effects[34]. Semi-parametric terms and random effects are used only to account for spatial autocorrelation, imbalanced sampling and pseudo-replication. Species richness is first log-transformed and then both Shannon Evenness and species richness are scaled and modelled as fixed linear effects. These models were formalized as: $y = X\boldsymbol{\beta} + \boldsymbol{\zeta} + \varepsilon$, where $X\boldsymbol{\beta}$ are the fixed parametric effects and $\boldsymbol{\zeta}$ are the semi-parametric and random effects. $\boldsymbol{\zeta}$ is composed of two terms; $\zeta = \hat{f}(x, y, \log(\text{N}_{\text{samples}})) + Z\mathbf{u}$, where $\hat{f}$ is a smoothing term (thin plate splines) of the geographic coordinates (x, y) and the logarithm of the number of FIA plots in a grid cell ($\log(\text{N}_{\text{Samples}})$), while $Z\mathbf{u}$ is a random effect term of grid cell ID nested in Ecoregion. The smoothing term is included to account for spatial as well as sampling patterns in both the value and bias of the index. The initial model formulation was then:

$$\text{Index} \sim \text{Type} + s(x, y, \text{N}_{\text{samples}}) + (1 \,|\, \text{Ecoregion/Type})$$

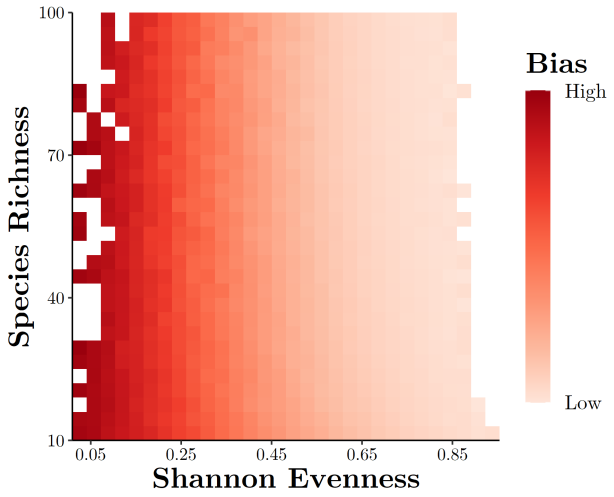

Figure 2

A baseline for the bias-relationships of functional dispersion. The color is based on an aggregated mean bias, based on 100 000 virtual communities, with log-normal abundance distributions, across 50 evenly spaced values of species richness in $[10, 100]$. The log-standard deviation is kept constant at $\sigma_{\log} = 2$ & traits are normally distributed.

This model framework not only detects, but also decompose the functional index biases into three parts (baseline, species richness, Shannon Evenness). Bias decomposition using this model is done by adding the new components as fixed effects with an interaction with 'Type' (the binary parameter indicating the abundance standardization "treatment"). This addition also allows for interpretation of the unbiased effects of species richness and Shannon Evenness on the functional index at hand. The best fitted models were selected based on the minimum AIC from the set of models that can be constructed by removing fixed effects (see subsection 3.2), semi-parametric and random effects were kept the same for all models.

When considering the bias introduced by using presence data instead of abundance data, it is important to consider the conditions under which a bias might arise. For all the functional indices which I analyse in this report, the abundances are used to compute a species-weight vector for each community[11]. As such it must be a necessary, but not obviously sufficient, condition, that the abundance composition of communities must be uneven (see Figure 1). By considering the following expressions for Shannon

[e] This approach is chosen over the more classical linear discriminant analysis, since this has a much lower minimum and mean CV accuracy of [0.161, 0.927] & 0.64.

diversity given presence ($D_{Shannon}(w^\star)$):

$$D_{Shannon}(w^\star) = -\sum_{i=1}^{N} \frac{w_i^\star}{\Sigma w^\star} \cdot \log\left(\frac{w_i^\star}{\Sigma w^\star}\right), \quad w_i^\star = \begin{cases} 1, & w_i > 0 \\ 0, & w_i = 0 \end{cases}$$
$$= -\Sigma w^\star \cdot \frac{1}{\Sigma w^\star} \cdot \log\left(\frac{1}{\Sigma w^\star}\right) = \log(\Sigma w^\star)$$

then recognizing $\Sigma w^\star$ as the number of species present N^*:

$$D_{Shannon}(w^\star) = \log(N^\star)$$

It becomes clear that Shannon evenness which is defined as[35] $E_{Shannon} = \frac{D_{Shannon}(w)}{\log(N)}$, can be reformulated as the ratio between Shannon diversity of a community given abundance data and the same community given presence data:

$$E_{Shannon} = \frac{D_{Shannon}(w)}{D_{Shannon}(w^*)}$$

Thus functional indices for abundance and presence data must converge, when Shannon evenness is equal to 1. For this reason I propose that Shannon evenness can be used as an indicator of possible biases in functional analyses. Simulations also provide a strong expectation that it is indeed Shannon evenness, not species richness, which primarily drives the bias, at least for functional dispersion, see Figure 2. However for the data distribution of the real woody plant communities in my investigation, we do not have such broad coverage of Shannon evenness for the entire range of species richness, especially within specific sampling regimes.

2.4 Software

All analyses and illustrations were carried out in R (R version 4.1.3 (2022-03-10))[36]. Calculation of the functional indices were done using a package I created for the specific purposes and needs of this work, and is accessible on github (https://github.com/asgersvenning/bachelor). The package provides separate functions for calculating the individual indices, as well as support for both euclidean distance and (weighted) Gower dissimilarity, as well as ordination. The package also includes all the data necessary to reproduce my results. It is inspired by the FD[37] package, but has some key changes/improvements that allow more flexibility over FD, as well as being faster in some cases. In the backend the package relies heavily on the use of the tidyverse[38] (for data manipulation) and the geometry[39] package (for computing convex hulls and hypervolumes).

All modelling was done using the GAMM4[34] package, which is an extension to the mgcv[33,40–43] package providing more efficient generalized additive mixed models for large numbers of random effects, while classification using Support Vector Machines was done using the kernlab[44,45] package. Visualizations were done using a multitude of packages including but not limited to ggplot2[46], patchwork[47] and sf[48], while geospatial operations were done using terra[49], sf[48] and stars[50]. I also thank the creators of the packages hash[51] and magrittr[52] for providing cleaner and faster workflows. Scripts for visualizations and analyses will be available on the package github.

3 Results

3.1 Spatio-topological patterns of functional space

The most obvious spatio-topological pattern in the functional space is a clear distinction between the eastern and western United States as can be seen in Figure 3 & Figure 4. The east-west divide is also visible in the distribution of topological patterns across ecoregions, however, several ecoregions diverge from this pattern, notably the Everglades in Florida. It is also striking that FDis and FEve are consistently larger using presence data, while FDiv is basically unchanged across different ecosystems. All three indices (FDis, FDiv, FEve) also show smaller variance under presence than abundance.

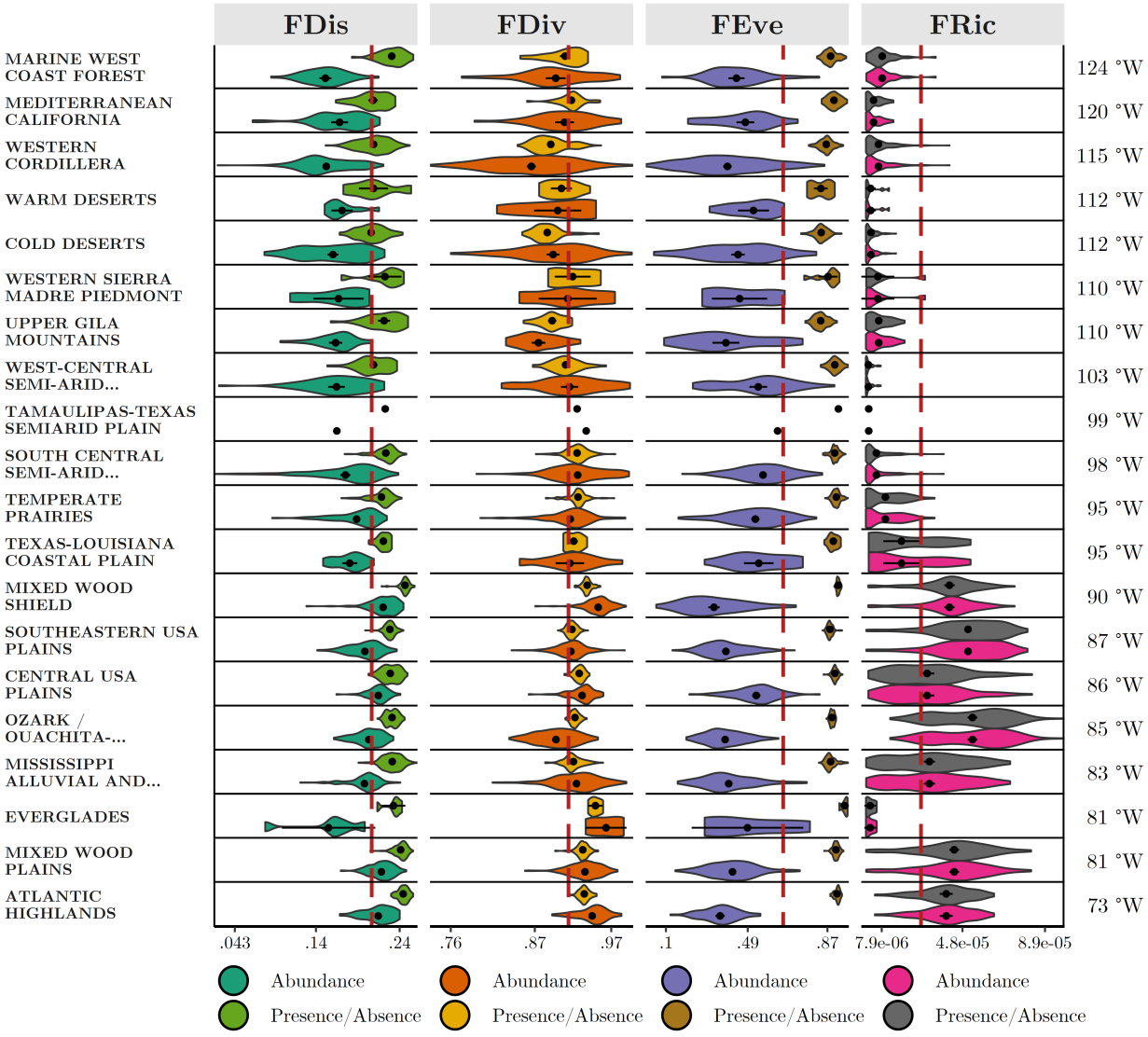

Figure 3

Functional indices disaggregated by ecoregion arranged: West (top) to East (bottom). The dashed red line indicates the overall mean for each index, while the horizontal black error bars indicate 95% confidence intervals. Coordinates on the right vertical axis annotate the approximate centroid longitude of each Ecoregion.

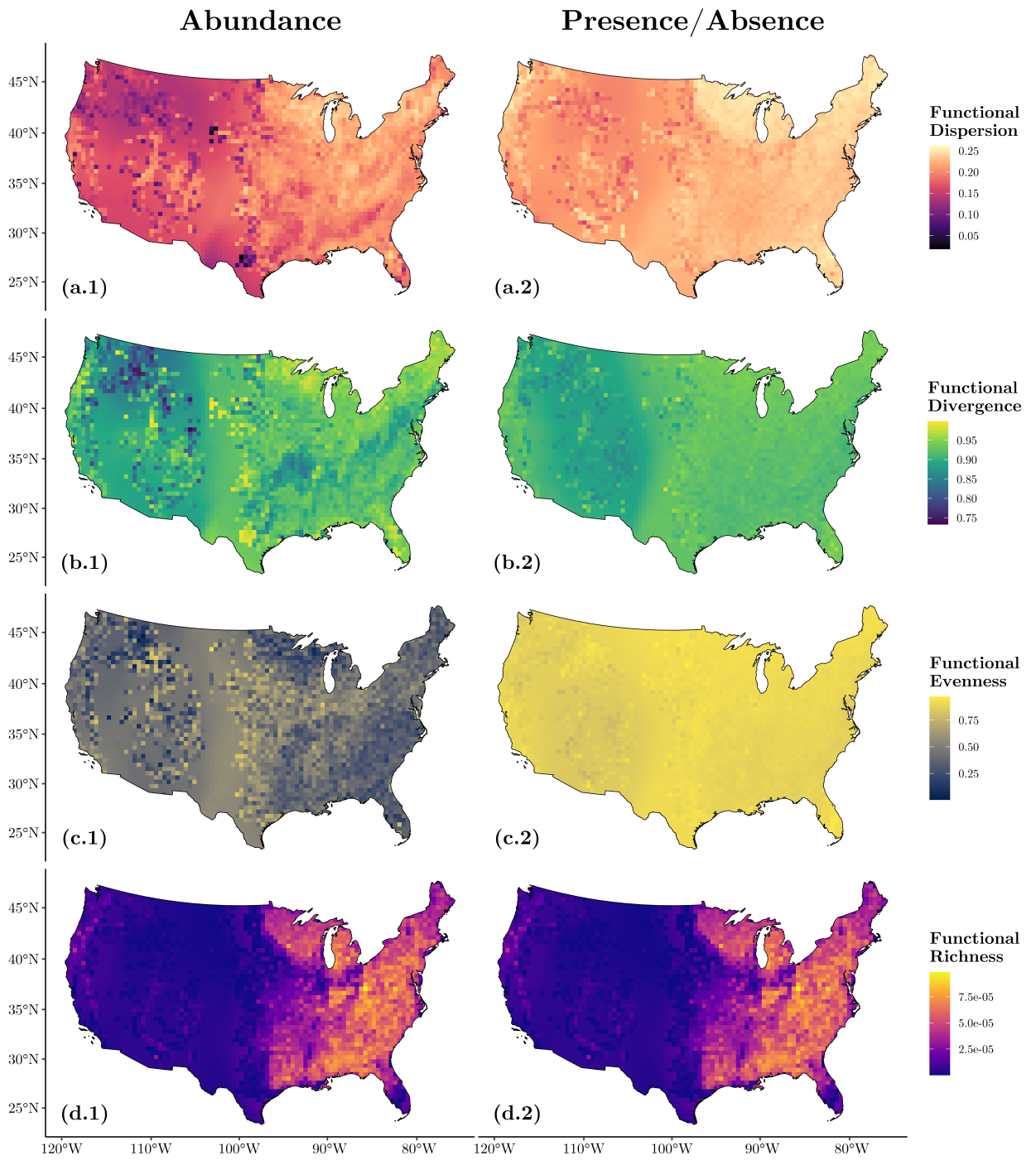

Figure 4

Interpolated spatial patterns in the topological indices of the functional ecological space. The values for the functional indices (FDis: **(a.*)**, FDiv = **(b.*)**, FEve = **(c.*)**, FRic = **(d.*)**) are calculated using the relative weighted mean abundances of species (panel **(*.1)**) and presence values (panel **(*.2)**). Grid cells with too species (n< 7) are interpolated using local distance-weighted geographic regression. The raw maps can be found in appendix D.

3.2 Biases in functional analysis using presence

	Sum Sq	F value	Pr(>F)
FDis			
With	0.045	5.953	0.0150⋆
Without	0.300	36.530	$2.9 \cdot 10^{-09}$ ⋆⋆⋆
FDiv			
With	$1.4 \cdot 10^{-04}$	0.029	0.8636
Without	$3.4 \cdot 10^{-05}$	0.007	0.9316
FEve			
With	$4.2 \cdot 10^{-04}$	0.141	0.7068
Without	0.011	3.666	0.0560.

Table 1

ANOVA type 2 results (simulated data) for the effect of species richness on functional indices with and without Shannon evenness included.

In the woody plant assemblages I investigated, the variance (and to some degree mean) of Shannon Evenness is strongly correlated with species richness, particularly in the low to medium sampling density regime (0-200 plots). To illustrate the issues that might arise from this structure I ran a simulation across the three topological functional indices which integrate abundances, with species richness in the range 10-100 following a log-normal abundance distributed communities ($\mu_{\log} = -4, \quad \sigma_{\log} = 1.25$) with 5 uniformly distributed traits for each virtual species[f]. I then ran a set of type II ANOVA's on linear models of the bias (for each index) using either only species richness, or species richness along with Shannon evenness. This synthetic analysis shows how a correlation structure between species richness and unknown true predictors of bias (Shannon Evenness), can result in potentially misleading findings (Table 1).

	Estimate	Std. Error	t value	Pr(>|t|)
FDis				
β	0.189	0.001	170.005	0 ⋆⋆⋆
β^{Bias}	0.036	$2.6 \cdot 10^{-04}$	140.861	0 ⋆⋆⋆
α_{Richness}	0.020	0.001	19.165	$1.1 \cdot 10^{-78}$ ⋆⋆⋆
$\alpha_{\text{Shannon Evenness}}$	0.020	$3.2 \cdot 10^{-04}$	63.220	0 ⋆⋆⋆
$\alpha^{\text{Bias}}_{\text{Richness}}$	-0.003	0.001	-2.901	0.0037 ⋆⋆
$\alpha^{\text{Bias}}_{\text{Shannon Evenness}}$	-0.021	$3.8 \cdot 10^{-04}$	-57.427	0 ⋆⋆⋆
FDiv				
β	0.911	0.001	859.329	0 ⋆⋆⋆
β^{Bias}	0.005	$5.2 \cdot 10^{-04}$	10.220	$3 \cdot 10^{-24}$ ⋆⋆⋆
FEve				
β	0.429	0.001	251.252	0 ⋆⋆⋆
β^{Bias}	0.459	0.002	190.751	0 ⋆⋆⋆
$\alpha_{\text{Shannon Evenness}}$	0.021	0.002	9.133	$9.8 \cdot 10^{-20}$ ⋆⋆⋆
$\alpha^{\text{Bias}}_{\text{Shannon Evenness}}$	-0.024	0.003	-7.529	$6.1 \cdot 10^{-14}$ ⋆⋆⋆

Table 2

Summary statistics for the best models (real data) selected based on minimizing AIC. Predictors are scaled (after transformation).

Using the abundances from FIA and traits from PLANTS, I found a significant positive bias for all three

[f] The distribution parameters are chosen to mimic the ones in the FIA dataset.

functional indices which can integrate abundance (FDis, FDiv and FEve), using a generalized additive mixed model (GAMM) adjusting for both spatial and sampling patterns, as well as differences between ecosystems. This result falls in line with the visual patterns found in subsection 3.1 (see Figure 3). This bias was decomposed by expanding the fixed effects, $X\boldsymbol{\beta}$, in the GAMM model as follows:

$$X\boldsymbol{\beta} = \beta + \alpha_{\text{Richness}} \cdot \text{Richness} + \alpha_{\text{Shannon Evenness}} \cdot \text{Shannon Evenness} + \mathbf{B}$$
$$\mathbf{B} = T \cdot (\beta^{\text{Bias}} + \alpha^{\text{Bias}}_{\text{Richness}} + \alpha^{\text{Bias}}_{\text{Shannon Evenness}} \cdot \text{Shannon Evenness})$$

Where T is short for data standardization **T**ype and is an indicator variable which is equal to 1, when the data is presence-only. The first three parameters (β, α_{Richness}, $\alpha_{\text{Shannon Evenness}}$) model the patterns in the values of the index given abundance data, while the latter three parameters in $\mathbf{B}$ model the difference (bias) in the value of the index given presence data. The decomposition of $\mathbf{B}$ can be interpreted as the estimated. . . :

- β: functional index value using abundances given the mean of the transformed predictors.
- α_{Richness}: scaled log-linear relationship between species richness and the functional index given abundances.
- $\alpha_{\text{Shannon Evenness}}$: scaled linear relationship between Shannon Evenness and the functional index given abundances.
- β^{Bias}: baseline bias component given presence data.
- $\alpha^{\text{Bias}}_{\text{Richness}}$: scaled log-linear bias component explained by species richness given presence data.
- $\alpha^{\text{Bias}}_{\text{Shannon Evenness}}$: scaled linear bias component explained by Shannon Evenness given presence data.

As expected for the indices with a significant bias (FDis and FEve), the estimated component for Shannon Evenness, $\alpha^{\text{Bias}}_{\text{Shannon Evenness}}$, is highly significant and large in magnitude, while the estimated component for species richness, $\alpha^{\text{Bias}}_{\text{Richness}}$, is smaller in magnitude and/or insignificant, as can be seen in Table 2. For FDiv only the baseline bias component was even slightly significant based on model AIC, however the variance in FDiv was much smaller given presence data (not shown).

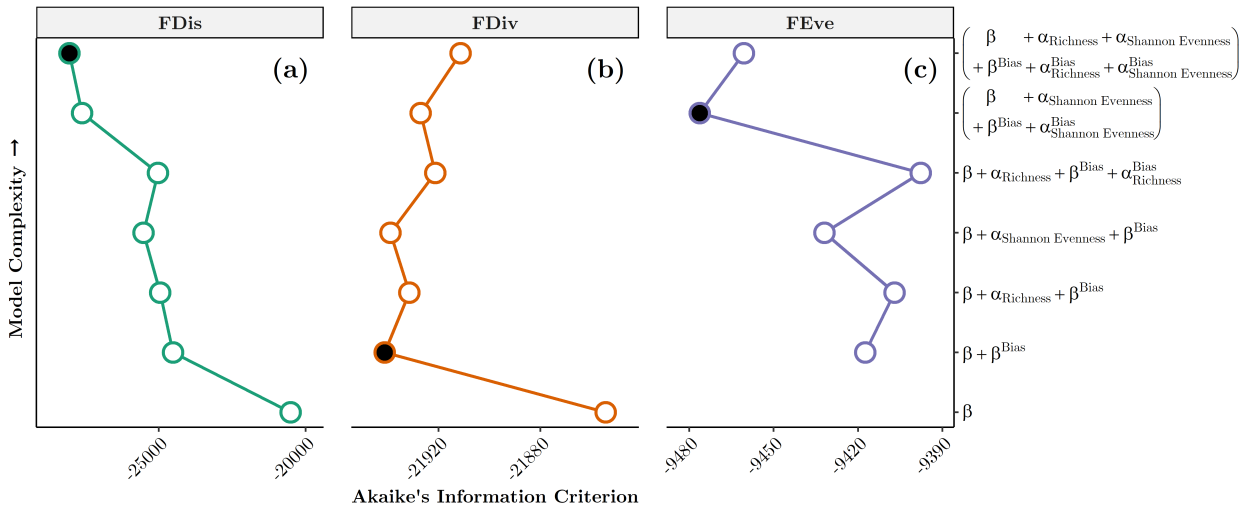

Figure 5

Model selection based on minimizing AIC. The intercept only model for FEve **(c)** is not shown due to extreme AIC (AIC = 335).

4 Discussion

Here I have shown the functional indices FDis and especially FEve, are biased when presence data is used, and that this bias is strongly related to the evenness of the abundance distribution. This is not entirely surprising and makes sense ecologically, given that communities where the abundance is mainly concentrated on a few species, will likely be less functionally diverse than communities where abundance is spread more evenly among species, which is inherently assumed in presence data.

Although the effects of integrating abundances seem to suggest that limiting similarity is less important than what is traditionally thought, the species richness relationships that I find support the theory of carrying capacity of species richness[3]. Both functional divergence and evenness show no significant correlation with species richness. Thus, clustering of species in functional space is also not correlated with species richness, suggesting that niche partitioning occurs equally at all levels of species richness supporting the theory of limiting similarity. However the most significant species richness relationship that I find, is a positive correlation with functional dispersion suggesting that niche partitioning begins at the central niche in a community, which I hypothesize is often most productive. If true, this pattern can be explained by species richness-productivity relationships which have been shown associated with niche packing[53]. My results thus show that species-rich communities are more functionally diverse, even and show increasing partitioning of the niche space. These patterns are initially strongly correlated with the sampling regime, which is especially evident in the spatial patterns of the functional indices, however the reported relationships are adjusted for spatial, ecosystem and sampling effects.

I interpret the strong longitudinal divide as a coupled ecological and sampling effect. There are clearly different sampling regimes in the eastern and western parts of FIA effort[54,55] (also see appendix E), however at the same time the western United States is covered by large swathes of deserts, which might be unsuitable for the FIA, and coniferous forests which show different patterns of functional diversity compared to decidous forests[56].

A caveat to the rigor of this model, is a large degree of residual heteroscedacity between abundances and presence, with abundances generally displaying much larger variance. I was unfortunately not able to resolve this due to the gamm4 package not supporting general model classes (in this case the "gaulss" family, would be more appropriate) while the alternative solution using random effects in gam (from the package mgcv) is exceedingly slow when the number of random effects is large (in this case n = 2291). The effect of this compromise is likely larger standard errors (higher p-values) on the estimated effects given abundances, and likewise smaller standard errors (lower p-values) for the estimated effects given presence.

5 Summary

I find significant biases in certain functional indicies (FDis, FEve) when comparing presence data with abundance data underscoring the need for greater integration of abundances in functional ecology, a trend which I expect will happen naturally, as data availability and statistical/modelling knowledge increases. This work also provides further support for the carrying capacity of species richness[3] and limiting similarity, by showing that species-rich areas are not more functionally clustered. Future work on robust testing of these patterns using abundance permutation tests on multiple spatial scales and extending the work beyond the US and woody plants is needed however, before we can be fully confident in the patterns I have shown here and their ecological generality.

6 References

1. IPBES. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. (2019) doi:10.5281/ZENODO.6417333.
2. Lavorel, S. & Garnier, E. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Functional Ecology* **16**, 545–556 (2002).
3. Storch, D. & Okie, J. G. The carrying capacity for species richness. *Global Ecology and Biogeography* **28**, 1519–1532 (2019).
4. Rabosky, D. L. & Hurlbert, A. H. Species richness at continental scales is dominated by ecological limits. *American Naturalist* **185**, 572–583 (2015).
5. Ricklefs, R. E. Species richness and morphological diversity of passerine birds. *Proceedings of the National Academy of Sciences* **109**, 14482 LP–14487 (2012).
6. Mouillot, D., Graham, N. A. J., Villéger, S., Mason, N. W. H. & Bellwood, D. R. A functional approach reveals community responses to disturbances. *Trends in Ecology & Evolution* **28**, 167–177 (2013).
7. Swenson, N. G. *et al.* Constancy in Functional Space across a Species Richness Anomaly. *The American Naturalist* **187**, E83–E92 (2016).
8. Ordonez, A. & Svenning, J. C. Greater tree species richness in eastern North America compared to Europe is coupled to denser, more clustered functional trait space filling, not to trait space expansion. *Global Ecology and Biogeography* **27**, 1288–1299 (2018).
9. McGill, B. J., Enquist, B. J., Weiher, E. & Westoby, M. Rebuilding community ecology from functional traits. *Trends in Ecology and Evolution* **21**, 178–185 (2006).
10. Paine, C. E. T. *et al.* Globally, functional traits are weak predictors of juvenile tree growth, and we do not know why. *Journal of Ecology* **103**, 978–989 (2015).
11. Villéger, S., Mason, N. W. H. & Mouillot, D. NEW MULTIDIMENSIONAL FUNCTIONAL DIVERSITY INDICES FOR A MULTIFACETED FRAMEWORK IN FUNCTIONAL ECOLOGY. *Ecology* **89**, 2290–2301 (2008).
12. Mouchet, M. A., Villéger, S., Mason, N. W. H. & Mouillot, D. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology* **24**, 867–876 (2010).
13. Violle, C., Reich, P. B., Pacala, S. W., Enquist, B. J. & Kattge, J. The emergence and promise of functional biogeography. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 13690–13696 (2014).
14. Weiher, E., Clarke, G. D. P. & Keddy, P. A. Community Assembly Rules, Morphological Dispersion, and the Coexistence of Plant Species. *Oikos* **81**, (1998).
15. Scherer-Lorenzen, M. Functional diversity affects decomposition processes in experimental grasslands. *Functional Ecology* **22**, 547–555 (2008).
16. Zavaleta, E. S., Pasari, J. R., Hulvey, K. B. & Tilman, G. D. Sustaining multiple ecosystem functions in grassland communities requires higher biodiversity. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 1443–1446 (2010).
17. Hejda, M. & Bello, F. de. Impact of plant invasions on functional diversity in the vegetation of Central Europe. *Journal of Vegetation Science* **24**, 890–897 (2013).
18. Catford, J. A. *et al.* Traits linked with species invasiveness and community invasibility vary with time, stage and indicator of invasion in a long-term grassland experiment. *Ecology Letters* **22**, 593–604 (2019).
19. Yang, J., Cao, M. & Swenson, N. G. Why Functional Traits Do Not Predict Tree Demographic Rates. *Trends in Ecology and Evolution* **33**, 326–336 (2018).

20. Svenning, J.-C., Eiserhardt, W. L., Normand, S., Ordonez, A. & Sandel, B. The Influence of Paleoclimate on Present-Day Patterns in Biodiversity and Ecosystems. *Annu. Rev. Ecol. Evol. Syst* **46**, 551–72 (2015).
21. Culshaw, V., Mairal, M. & Sanmartín, I. Biogeography Meets Niche Modeling: Inferring the Role of Deep Time Climate Change When Data Is Limited. *Frontiers in Ecology and Evolution* **9**, 599 (2021).
22. Swenson, N. G. & Weiser, M. D. Plant geography upon the basis of functional traits: An example from eastern North American trees. *Ecology* **91**, 2234–2241 (2010).
23. Swenson, N. G. *et al.* The biogeography and filtering of woody plant functional diversity in North and South America. *Global Ecology and Biogeography* **21**, 798–808 (2012).
24. Blonder, B. *et al.* Late Quaternary climate legacies in contemporary plant functional composition. *Global Change Biology* **24**, 4827–4840 (2018).
25. U.S. Department of Agriculture, Forest Service, N. R. S. Forest Inventory and Analysis Database. (2022).
26. Sauer, J. R., Hines, J. E. & Fallon, J. The North American breeding bird survey, results and analysis 1966–2007. Version 5.15. 2008. *US geological Survey Patuxent Wildlife research center, laurel, maryland* (2008).
27. Gillings, S. *et al.* Breeding and wintering bird distributions in Britain and Ireland from citizen science bird atlases. *Global Ecology and Biogeography* **28**, 866–874 (2019).
28. USDA, N. *The PLANTS Database*. http://plants.usda.gov (2022).
29. Fraser, L. H. TRY—A plant trait database of databases. *Global Change Biology* **26**, 189–190 (2020).
30. Wilman, H. *et al.* EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology* **95**, 2027–2027 (2014).
31. Laliberté, E. & Legendre, P. A distance-based framework for measuring functional diversity from multiple traits. *Ecology* **91**, 299–305 (2010).
32. Environmental Protection Agency, U. S. Level IV Ecoregions of the Conterminous United States. (2013).
33. Wood, S. N. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* vol. 65 95–114 (2003).
34. Wood, S. & Scheipl, F. gamm4: Generalized additive mixed models using 'mgcv' and 'lme4'. (2020).
35. Gotelli, N. J. & Colwell, R. K. Biological diversity: frontiers in measurement and assessment. 39–54 (2011).
36. Team, R. C. R: A language and environment for statistical computing. (2022).
37. Laliberté, E., Legendre, P. & Shipley, B. FD: Measuring functional diversity from multiple traits and other tools for functional ecology. (2014).
38. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* vol. 4 1686 (2019).
39. Habel, K., Grasman, R., Gramacy, R. B., Mozharovskyi, P. & Sterratt, D. C. Geometry: Mesh generation and surface tessellation. (2022).
40. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society (B)* vol. 73 3–36 (2011).
41. Wood, S. N., N., Pya & Safken, B. Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association* vol. 111 1548–1575 (2016).
42. Wood, S. N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* vol. 99 673–686 (2004).
43. Wood, S. N. Generalized additive models: An introduction with r. (2017).

44. Karatzoglou, A., Smola, A. & Hornik, K. Kernlab: Kernel-based machine learning lab. (2022).

45. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. Kernlab – an S4 package for kernel methods in r. *Journal of Statistical Software* vol. 11 1–20 (2004).

46. Wickham, H. ggplot2: Elegant graphics for data analysis. (2016).

47. Pedersen, T. L. Patchwork: The composer of plots. (2020).

48. Pebesma, E. Simple features for r: Standardized support for spatial vector data. *The R Journal* vol. 10 439–446 (2018).

49. Hijmans, R. J. Terra: Spatial data analysis. (2022).

50. Pebesma, E. Stars: Spatiotemporal arrays raster and vector data cubes. (2021).

51. Brown, C. Hash: Full featured implementation of hash tables/AssociativeArrays/dictionaries. (2022).

52. Bache, S. M. & Wickham, H. Magrittr: A forward-pipe operator for r. (2022).

53. Pellissier, V., Barnagaud, J. Y., Kissling, W. D., Şekercioğlu, Ç. & Svenning, J. C. Niche packing and expansion account for species richness–productivity relationships in global bird assemblages. *Global Ecology and Biogeography* **27**, 604–615 (2018).

54. Mcroberts, R. E., Bechtold, W. A., Patterson, P. L., Scott, C. T. & Reams, G. A. The Enhanced Forest Inventory and Analysis Program of the USDA Forest Service: Historical Perspective and Announcement of Statistical Documentation. *Journal of Forestry* (2005).

55. Wiener, S. S. *et al.* United States Forest Service Use of Forest Inventory Data: Examples and Needs for Small Area Estimation. *Frontiers in Forests and Global Change* **4**, 198 (2021).

56. Łubek, A., Kukwa, M., Jaroszewicz, B. & Czortek, P. Identifying mechanisms shaping lichen functional diversity in a primeval forest. *Forest Ecology and Management* **475**, 118434 (2020).

57. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27**, 857 (1971).

Appendices

A Gower traits & dissimilarity

The definition of Gower dissimilarity is[g] 57:

$$\text{gowdis}(\mathbf{Tr}[i,], \mathbf{Tr}[j,]) = \delta_{i,j}^{\mathrm{T}} \cdot \mathbf{w}, \quad \delta_{i,j} = |\mathbf{Tr}[j,] - \mathbf{Tr}[j,]|$$

Where **Tr** is what I choose to call the Gower trait matrix, which can be calculated from the orginal trait matrix using two operations:

1. Ordinal variables are one-hot encoded i.e. turned into a number of binary dummy columns.
2. Continuous variables are scaled by their range (range-scaled):

$$\mathbf{Tr}^*[,i] = \frac{\mathbf{Tr}[,i] - \min(\mathbf{Tr}[,i])}{\max(\mathbf{Tr}[,i]) - \min(\mathbf{Tr}[,i])}$$

The Gower trait matrix must be accompanied by a weight vector $\mathbf{w}$[h]. The weights can be chosen a priori, but for factor variables, they must be divided by the number of dummy columns associated with the given variable.

Missingness is handled by excluding the trait from δ and $\mathbf{w}$, followed by re-normalizing $\mathbf{w}$ such that $\sum \mathbf{w} = 1$:

$$\delta \to \delta_{\mathrm{V}}, \quad \mathbf{w} \to \frac{\mathbf{w}_{\mathrm{V}}}{\sum \mathbf{w}_{\mathrm{V}}}, \quad \mathrm{V} \ : \ \{i \ \text{ for } \ \delta \in \mathbb{R}\}$$

An interesting not is that if all traits are weighted equally:

$$\mathbf{w}_i = \dim(\mathbf{w})^{-1}$$

then Gower dissimilarity is equal to range scaled Manhattan distance, which when taking into consideration that Bray-Curtis dissimilarity is also equal to Manhattan distance between relative abundances divided by 2, draws a clear connection between Gower and Bray-Curtis dissimilarity.

[g] Exluding ordered factors.
[h] Not strictly necessary if all traits are continous.

B Trait summary

The PLANTS database contains information on 83 traits, descriptions for which (at the time of writing) can all be found in the help document. I have aggregated an overview of which traits that I have and have not used, by their data type (continous, binary or factor) and functional category (morphology, growth, reproduction or suitability), derived from the PLANTS help document.

Type	Variable	Included
MORPHOLOGY/PHYSIOLOGY		
Factor	Bloat	FALSE
Factor	Active Growth Period; C:N Ratio; Flower Color; Foliage Color; Foliage Porosity Summer; Foliage Porosity Winter; Foliage Texture; Fruit/Seed Color; Growth Form; Growth Rate; Lifespan; Nitrogen Fixation; Shape and Orientation; Toxicity	TRUE
Binary	Coppice Potential; Fire Resistant; Low Growing Grass	FALSE
Binary	Fall Conspicuous; Flower Conspicuous; Fruit/Seed Conspicuous; Known Allelopath; Leaf Retention; Resprout Ability	TRUE
Continous	Height at 20 Years, Maximum (feet); Height, Mature (feet)	TRUE
GROWTH REQUIREMENTS		
Factor	Hedge Tolerance; Moisture Use	FALSE
Factor	Anaerobic Tolerance; CaCO3 Tolerance; Drought Tolerance; Fertility Requirement; Fire Tolerance; Salinity Tolerance; Shade Tolerance	TRUE
Binary	Cold Stratification Required	FALSE
Binary	Adapted to Coarse Textured Soils; Adapted to Fine Textured Soils; Adapted to Medium Textured Soils	TRUE
Continous	Planting Density per Acre, Maximum; Planting Density per Acre, Minimum	FALSE
Continous	Frost Free Days, Minimum; pH, Maximum; pH, Minimum; Precipitation, Maximum; Precipitation, Minimum; Root Depth, Minimum (inches); Temperature, Minimum (°F)	TRUE
REPRODUCTION		
Factor	Bloom Period; Fruit/Seed Abundance; Fruit/Seed Period Begin; Fruit/Seed Period End; Seed Spread Rate; Seedling Vigor; Vegetative Spread Rate	TRUE
Binary	Propagated by Bare Root; Propagated by Bulb; Propagated by Container; Propagated by Corm; Propagated by Cuttings; Propagated by Seed; Propagated by Sod; Propagated by Sprigs; Propagated by Tubers; Small Grain	FALSE
Binary	Fruit/Seed Persistence	TRUE
Continous	Seed per Pound	TRUE
SUITABILITY/USE		
Factor	Commercial Availability; Fuelwood Product; Protein Potential	FALSE
Factor	Palatable Browse Animal; Palatable Graze Animal	TRUE
Binary	Berry/Nut/Seed Product; Christmas Tree Product; Fodder Product; Lumber Product; Naval Store Product; Nursery Stock Product; Palatable Human; Post Product; Pulpwood Product; Veneer Product	FALSE

Table 3

Allocation of traits from the USDA PLANTS database

C Confusion matrix

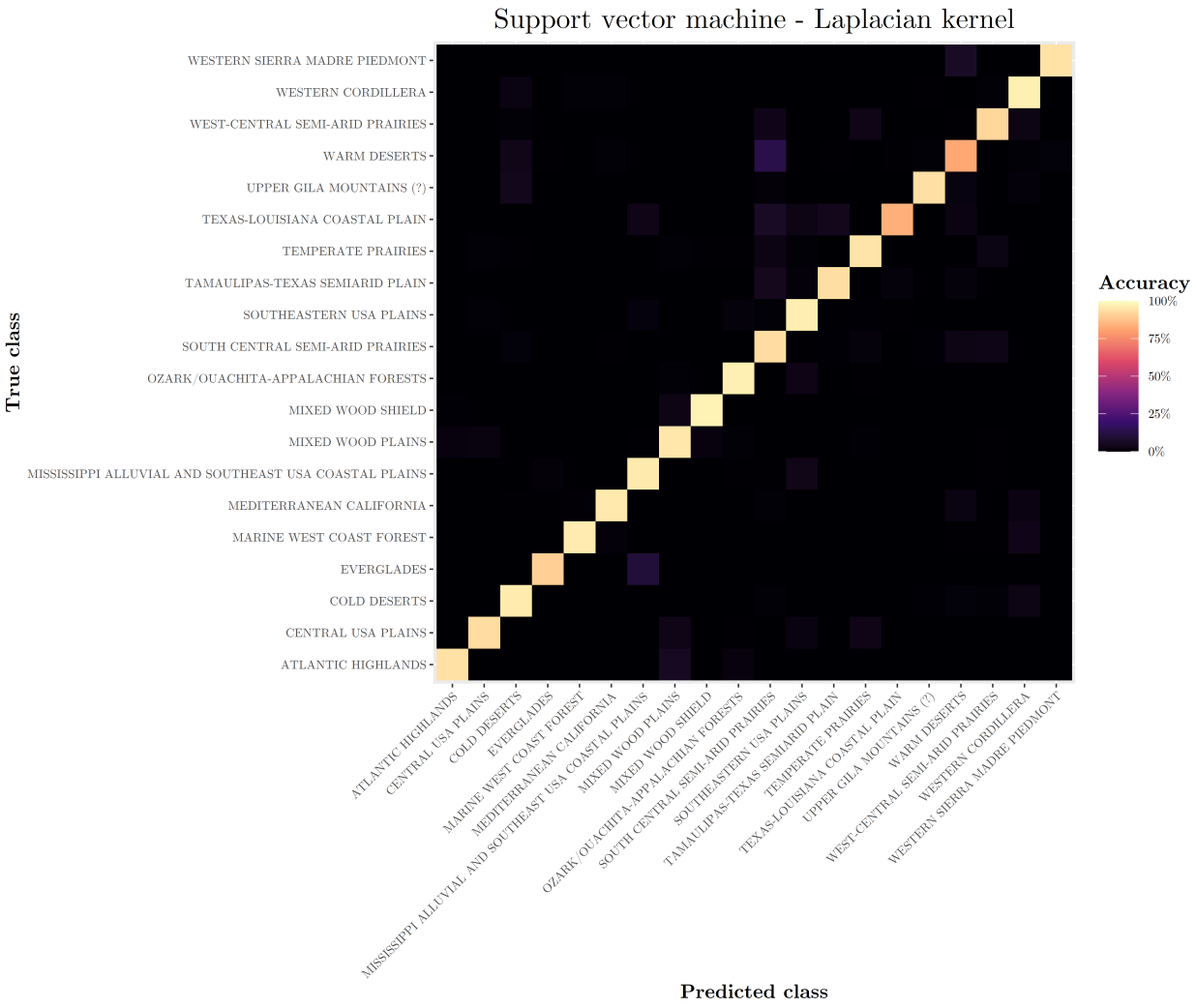

Figure 6

Confusion matrix of the Support Vector Machine model on ecoregion classification. The visualized accuracies are the proportion of the true class being classified as such, the confusion matrix would thus exhibit obvious visible artefacts if the model was overfitting. On the contrary the fact that almost all observations are found on the diagonal shows the voracity of the model and the confirms the validity of using the EPA ecoregions as delineations between distinct assemblages or ecosystems.

D Raw indices maps

For the figures in the report I have interpolated the values of grid cells with missing functional index values using a inverse distance weighted-mean in a circular sliding window. The original maps are provided here:

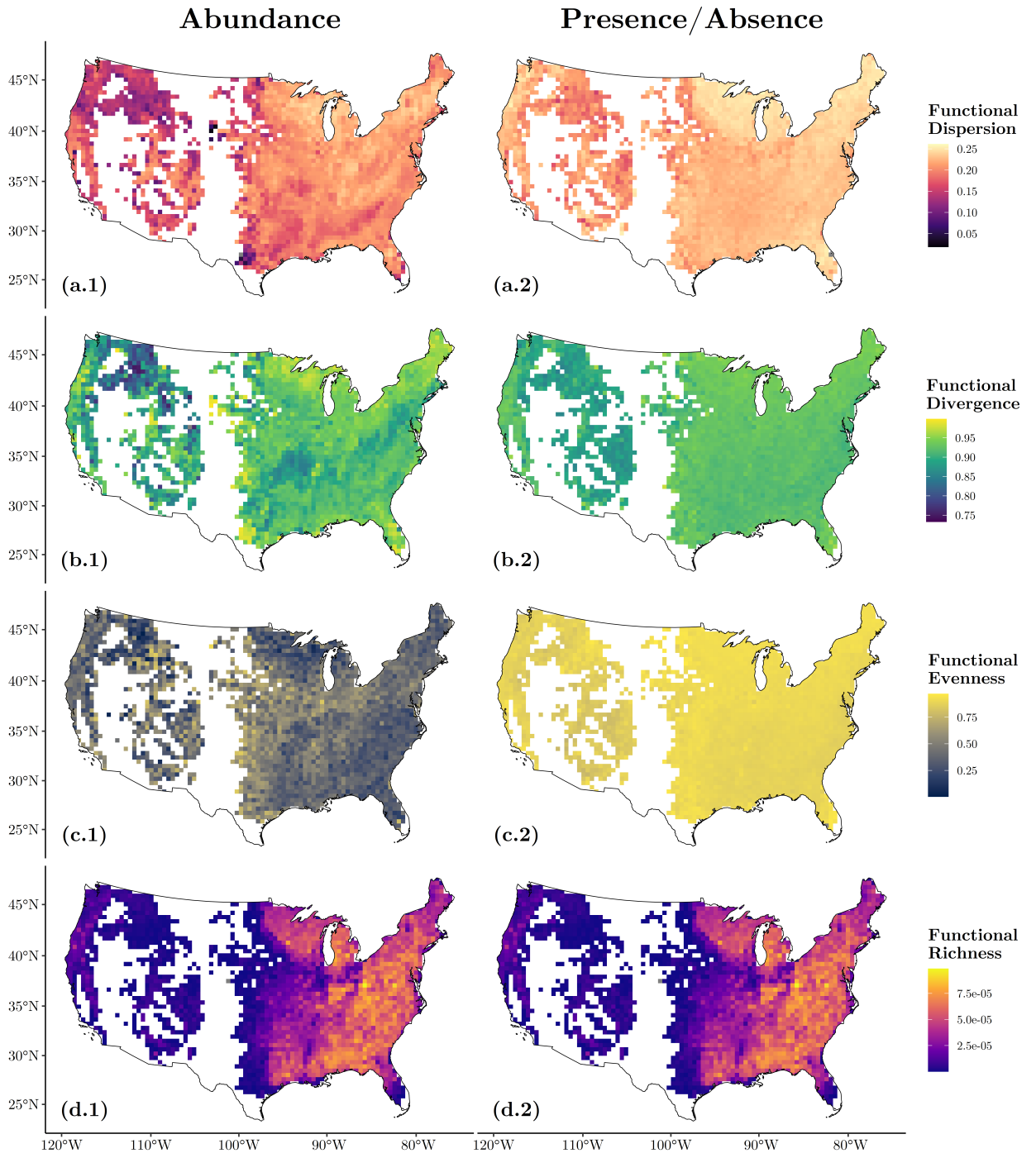

Figure 7

Raw spatial patterns in the topological indices of the functional ecological space. The values for the functional indices (FDis = A, FDiv = B, FEve = C, FRic = D) are calculated using the relative weighted mean abundances of species (panel 1) and presence values (panel 2).

E East-West Sampling Imbalance

There is significant sampling imbalance in the FIA dataset, as I have visualized here:

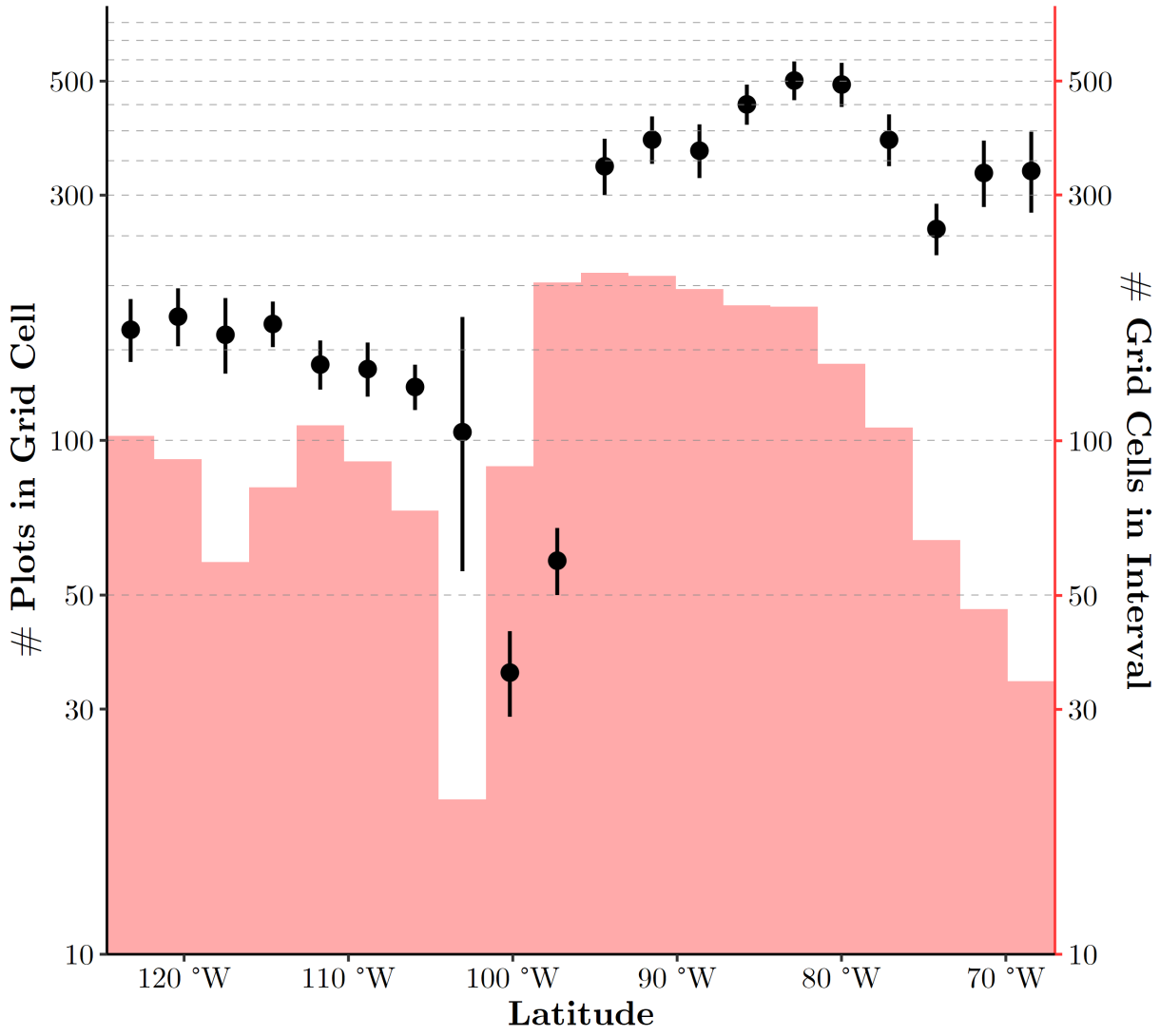

Figure 8

East-West sampling imbalance. The black point-ranges are 95%-confidence intervals of intercept only GLM's on the number of FIA plots per 50 × 50km study unit in each latitude interval. The salmon histogram is the number of 50 × 50km study unit in each equal interval latitude interval.