Biodiversity Informatics, 10, 2015, pp. 22-34

INDICES OF BIODIVERSITY PATTERN BASED ON PRESENCE-ABSENCE MATRICES: A GIS IMPLEMENTATION

JORGE SOBERÓN[1, 2] AND JEFF CAVNER[1]
[1]*Biodiversity Institute, University of Kansas, Lawrence, KS 66045, USA*
[2]*Department of Ecology & Evolutionary Biology, University of Kansas, Lawrence, KS 66045, USA*

Abstract.—In this paper we present mathematical notation and formulae relating a number of indices of the biodiversity pattern of an aggregate of species, one index of phylogenetic similarity, and an implementation of them as linked maps or phylogenies in a plug-in for the increasingly popular open source geographic information system Quantum GIS. We provide detailed formulae relating three indices of beta diversity, two of pattern of nestedness, one of checkerboard pattern, and two of ratios of variances. The above synthesis is achieved by deriving six vectors from the full presence-absence matrix. Our GIS implementation is done via web services, tapping the LifeMapper platform for estimating potential distributions of species.

Key words.—Biodiversity patterns, beta diversity, presence-absence matrix, phylogeny, LifeMapper, ARBOR.

Ecologists, macroecologists and biogeographers use a variety of indices and statistics to describe biodiversity patterns. Data used to describe a pattern may be continuous or discrete, and among the simplest are presence-absence data. If members of a set of species are present or absent in a set of localities (islands, countries, reserves, or cells in a grid, for example), the presence-absence matrix (PAM), is defined as a N sites by S species binary matrix $\mathbf{X} = \left[\delta_{i,j}\right]$ with elements equal to 1 if species j is present in cell i, and 0 otherwise.

The PAM contains the data from which one can estimate a variety of metrics of biodiversity pattern. There are metrics that describe the local or alpha species numbers, and their sum, average and spatial variance and covariance (Borregaard and Rahbek 2010; Graves and Rahbek 2005; Lande 1996; Legendre et al. 2005; Routledge 1977; Schluter and Ricklefs 1993; Whittaker 1972). There are other numbers that have been used to describe the degree of commonality in the composition of communities, or the overlap between the ranges of distribution of different species (Gotelli 2000; Schluter 1984; Stone and Roberts 1990). Finally, there are indices related to what is called "nestedness" (Almeida-Neto et al. 2007; Rodríguez-Gironés and Santamaría 2006; Ulrich et al. 2009) which is the degree to which species compositions of smaller communities are proper subsets of larger ones, or equivalently, the degree to which larger distributional ranges contain smaller distributions in a nested way.

All of the above indices can be derived from manipulations of the PAM and most of the calculations are straightforward. However, these metrics were first presented in the literature for small-sized PAMs, of the kind often found in ecological problems, or in biogeography when considering just a handful of regions or islands (Simberloff and Connor 1984). For instance, a typical problem could use tens of species in a few dozen sites, for PAMs of in the order of 10^2 to 10^4 elements. The same metrics, however, can be used to describe patterns at biogeographic extents and much higher resolutions, for instance, in grids of thousands of cells (Arita et al. 2008; Borregaard and Rahbek 2010; Orme et al. 2006; Rahbek and Graves 2000; Soberón and Ceballos 2011; Villalobos et al. 2013). Unfortunately, the size of the PAMs that are encountered when describing whole faunas over gridded regions is often very significant, not unusually of ~10^6 to 10^8 elements. The amount of calculations required to estimate some of these indices is large, and testing hypotheses about difference of observed PAMs using null-models requires randomizations that are beyond the capacity of existing software.

Moreover, to assemble such large extent, high-resolution PAMs for hundreds or thousands of species one has to resort to very cumbersome manipulations of one of two basic sets of data. In

the first case, it is possible to use "extent of occurrence" (Hurlbert and Jetz 2007) maps as presented in many biodiversity information initiatives; an excellent example is NatureServe[1]. Typically these data come in the form of vector-formatted files ("shapefiles"), one per species, that must be overlaid on a given grid to create a PAM of N = number of cells in the grid, and S = number of different species. Manipulating and operating large sets of such files is conceptually simple, but practically cumbersome.

Another possible source of distributional data is species distribution modeling (SDM, Franklin 2010), where occurrence data and ecological features are modeled together to get predicted ranges. Generally speaking, it is unadvisable to use occurrence data directly due to the biases inherent to this type of data (Lira-Noriega et al. 2007), and therefore SDM is needed. The typical outputs are raster files that after suitable post processing, including thresholding (Liu et al. 2005), can be used to create the presence-absence maps. Then individual species maps are overlaid, or "stacked," to create a PAM. Stacked maps (whether from SDM or from extent of occurrence datasets) are interpreted as PAMs under the very obvious hypothesis that interactions between species do not matter. Although this assumption is probably false in many cases (Araujo and Luoto 2007), there are several others in which it may be valid to use it, at least as a first order approximation (Peterson et al. 2011) to more realistic procedures.

In this work we describe the operations required to estimate the main indices, showing some of their relationships, and then present a practical implementation of these operations, developed as web services. We demonstrate this implementation using the open source GIS platform Quantum GIS.

MATHEMATICAL RELATIONS

We will show how a number of indices of geographic biodiversity pattern, already published, can be obtained from operations on six objects derived from a PAM, denoted by X (see summary in Table 1). As well known (Gower 1966), the post-multiplication and pre-multiplication of $\mathbf{X}$ by its transpose $\mathbf{X}^T$ yield, respectively, a matrix of shared species in sites, and a matrix of co-occurrence of usage of sites by species. The N x N matrix containing the number of shared species between sites i and h is $\mathbf{A} = \mathbf{X}\mathbf{X}^T$ (in the diagonal it contains the number of species in site i= 1, 2 … N). The S x S matrix containing the number of sites shared by species j and k is $\mathbf{\Omega} = \mathbf{X}^T\mathbf{X}$ (the diagonal contains the incidence of species j=1, 2 … S, which for large extents and high resolution of cells approaches the size of the area of distribution).

The vectors $\boldsymbol{\alpha}^T = (\alpha_1, \alpha_2, ...\alpha_N)$ and $\boldsymbol{\omega}^T = (\omega_1, \omega_2, ...\omega_S)$ contain the species richness of each site, and the range size of every species. By defining vectors of k-ones: $\mathbf{1}_k^T = (1,1,...1)_{1\times k}$, which are used to obtain sums, and using the symbol *diag* to denote the diagonal of a matrix, we get:

$$\begin{aligned} \boldsymbol{\alpha} &= diag(\mathbf{A}) = \mathbf{X}\mathbf{1}_S \\ \boldsymbol{\omega} &= diag(\mathbf{\Omega}) = \mathbf{X}^T\mathbf{1}_N \end{aligned} \quad (1)$$

Since $trace(\mathbf{X}^T\mathbf{X}) = trace(\mathbf{X}\mathbf{X}^T)$, then the sum of the elements in $\mathbf{A}$ is equal to the sum of the elements in $\mathbf{\Omega}$ and equal to the total number of ones in the $\mathbf{X}$ matrix, also called the fill, f.

Two other vectors (of size S and N) contain the total sum of shared range sizes, including the species with itself, and the total number of shared community compositions, including the species number of each site. These are called the mean proportional range-size, and the mean proportional species-diversity (Arita et al. 2008; Christen and Soberón 2009; Graves and Rahbek 2005) and Borregaard and Rahbek (2010) refer to the first as the dispersion field. The vectors are defined as:

$$\begin{aligned} \boldsymbol{\varphi} &= \mathbf{X}\boldsymbol{\omega} = \mathbf{A}\mathbf{1}_N \\ \boldsymbol{\psi} &= \mathbf{X}^T\boldsymbol{\alpha} = \mathbf{\Omega}\mathbf{1}_S \end{aligned} \quad (2)$$

Their averages are $\bar{\varphi} = \frac{1}{N}\boldsymbol{\varphi}^T\mathbf{1}_N$ and $\bar{\psi} = \frac{1}{S}\boldsymbol{\psi}^T\mathbf{1}_S$.

We use an asterisk to define proportionality with respect to S and N, for example: $\boldsymbol{\alpha}^{*T} = (\alpha_1/S, \alpha_2/S, ...\alpha_N/S)$ and $\boldsymbol{\omega}^{*T} = (\omega_1/N, \omega_2/N, ...\omega_S/N)$.

[1] http://www.natureserve.org/conservation-tools/data-maps-tools/digital-distribution-maps-mammals-western-hemisphere.

Table 1. Summary of relationships among different indices of biodiversity pattern.

	Name	Algebraic definition	Linear algebra
1	Whittaker's multiplicative beta	$\beta_W = \frac{1}{\bar{\omega}^*}$	$\beta_W = \frac{SN}{Trace(\mathbf{\Omega})}$
2	Lande's additive beta	$\beta_A = S(1-1/\beta_W)$	$\beta_A = S[1-\frac{Trace(\mathbf{\Omega})}{SN}]$
3	Legendre's beta	$\beta_L = SS(\mathbf{X}) = SN/\beta_W - \left(\sum_{j=1}^{s}\omega_j^{\,2}\right)/N$	$\beta_L = Trace(\mathbf{\Omega}) - \boldsymbol{\varphi}^T\mathbf{1}_N$
4	Richness-field of a species	$\psi_j = \sum_{i=1}^{N}\delta_{i,j}\alpha_i$	$\boldsymbol{\psi} = \mathbf{X}^T\boldsymbol{\alpha} = \mathbf{\Omega}\mathbf{1}_S$
5	Dispersion-field of a locality	$\varphi_i = \sum_{j=1}^{S}\delta_{i,j}\omega_j$	$\boldsymbol{\varphi} = \mathbf{X}\boldsymbol{\omega} = \mathbf{A}\mathbf{1}_N$
6	Matrix of covariance of composition of sites	$\mathbf{\Sigma}_{sites}(j,k) = \frac{1}{S}\sum_{l=1}^{S}\delta_{j,l}\delta_{k,l} - \frac{\alpha_j\alpha_k}{S^2}$	$\mathbf{\Sigma}_{sites} = \frac{1}{S}\mathbf{A} - \boldsymbol{\alpha}^*(\boldsymbol{\alpha}^*)^T$
7	Matrix of covariance of ranges of species	$\mathbf{\Sigma}_{sps}(h,i) = \frac{1}{N}\sum_{j=i}^{N}\delta_{i,j}\delta_{h,j} - \frac{\omega_i\omega_h}{N^2}$	$\mathbf{\Sigma}_{species} = \frac{1}{N}\mathbf{\Omega} - \boldsymbol{\omega}^*(\boldsymbol{\omega}^*)^T$
8	Mean composition covariance	$\alpha_j^* = \frac{\tau_j}{\bar{\varphi}_j^* - \beta_W^{-1}}$	$\bar{\boldsymbol{\tau}} = \frac{1}{NS}\boldsymbol{\varphi} - \beta_W^{-1}\boldsymbol{\alpha}^*$
9	Mean range covariance	$\omega_i^* = \frac{\bar{\rho}_i}{\bar{\psi}_i^* - \beta_W^{-1}}$	$\bar{\boldsymbol{\rho}} = \frac{1}{NS}\boldsymbol{\psi} - \beta_W^{-1}\boldsymbol{\omega}^*$
10	Schluter sites-composition covariance	$V_{sites} = \frac{\bar{\varphi}^* - S/\beta_W^2}{1/\beta_W - \bar{\psi}^*/N}$	$V_{sites} = \frac{\mathbf{1}^T\mathbf{\Sigma}_{sites}\mathbf{1}}{Trace(\mathbf{\Sigma}_{sites})}$
11	Schluter species-ranges covariance	$V_{sps} = \frac{\bar{\psi}^* - N/\beta_W^2}{1/\beta_W - \bar{\varphi}^*/S}$	$Vsps = \frac{\mathbf{1}\mathbf{\Sigma}_{sps}\mathbf{1}}{Trace(\mathbf{\Sigma}_{sps})}$
12	Wright & Reeves' nestedness	$N_C = \frac{1}{2}\sum_{j=1}^{S}\omega_j(\omega_j - 1) = \frac{N}{2}(\bar{\varphi} - \frac{S}{\beta_W})$	$N_C = \frac{1}{2}(\boldsymbol{\varphi}^T\mathbf{1} - \frac{NS}{\beta_W})$
13	Stone & Roberts C-score	$C = \frac{2}{S(S-1)}\left[\sum_{i=1}^{N}\sum_{h<i}(\omega_i - \omega_{i,h})(\omega_h - \omega_{i,h})\right]$	$\frac{\mathbf{1}_S^T(\mathbf{\Omega}\otimes\mathbf{\Omega})\mathbf{1}_S - N\bar{\varphi}}{2}$

Finally, the covariance matrix between the species inhabiting sites *j* and *k*, and the covariance matrix between ranges of distribution of species *h* and *i* are, respectively:

$$\begin{aligned}\boldsymbol{\Sigma}_{sites} &= \frac{1}{S}\mathbf{A} - \boldsymbol{\alpha}^*(\boldsymbol{\alpha}^*)^T = \frac{1}{S}\left(\mathbf{A} - \frac{1}{S}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\right)\\ \boldsymbol{\Sigma}_{species} &= \frac{1}{N}\boldsymbol{\Omega} - \boldsymbol{\omega}^*(\boldsymbol{\omega}^*)^T = \frac{1}{N}\left(\boldsymbol{\Omega} - \frac{1}{N}\boldsymbol{\omega}\boldsymbol{\omega}^T\right)\end{aligned} \quad (3)$$

Notice the use of the symbol small bold sigma ($\boldsymbol{\Sigma}$) to denote covariance matrices. There are two matrices in each of the formulae (3), and their traces are useful. The trace of **A** and of $\boldsymbol{\Omega}$ equal the fill, as we saw, and $trace(\boldsymbol{\alpha}\boldsymbol{\alpha}^T) = \sum_{i=1}^{N}\alpha_i^2, trace(\boldsymbol{\omega}\boldsymbol{\omega}^T) = \sum_{j=1}^{S}\omega_j^2$. These sums of squares are:

$$\begin{aligned}\sum_{i=1}^{N}\alpha_i^2 &= \boldsymbol{\alpha}^T\boldsymbol{\alpha} = \mathbf{1}_S^T\mathbf{X}^T\mathbf{X}\mathbf{1}_S = \mathbf{1}_S^T\boldsymbol{\Omega}\mathbf{1}_S = \mathbf{1}_S^T\boldsymbol{\psi} = S\bar{\psi}\\ \sum_{j=1}^{S}\omega_i^2 &= \boldsymbol{\omega}^T\boldsymbol{\omega} = \mathbf{1}_N^T\mathbf{X}\mathbf{X}^T\mathbf{1}_N = \mathbf{1}_N^T\mathbf{A}\mathbf{1}_N = \mathbf{1}_N^T\boldsymbol{\varphi} = N\bar{\varphi}\end{aligned} \quad (4)$$

It is a matter of a few lines of algebra and substitutions of the above to show that the traces of the two covariance matrices are:

$$\begin{aligned}trace(\boldsymbol{\Sigma}_{sites}) &= \frac{N}{\beta_W} - \bar{\psi}^*\\ trace(\boldsymbol{\Sigma}_{species}) &= \frac{S}{\beta_W} - \bar{\varphi}^*\end{aligned} \quad (5)$$

and that the vectors containing the averages of all their entries are:

$$\begin{aligned}\bar{\boldsymbol{\tau}} &= \frac{1}{N}\boldsymbol{\Sigma}_{sites}\mathbf{1}_N\\ &= \frac{1}{N}\left(\frac{1}{S}\mathbf{A}\mathbf{1}_N - \frac{1}{S^2}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\mathbf{1}_N\right)\\ SN\bar{\boldsymbol{\tau}} &= \boldsymbol{\varphi} - \frac{f}{S}\boldsymbol{\alpha}\\ S\bar{\boldsymbol{\tau}} &= \frac{1}{N}\boldsymbol{\varphi} - \frac{1}{\beta_W}\boldsymbol{\alpha}\\ &= \boldsymbol{\varphi}^* - \frac{1}{\beta_W}\boldsymbol{\alpha}\end{aligned} \quad (6)$$

and

$$\begin{aligned}\bar{\boldsymbol{\rho}} &= \frac{1}{S}\boldsymbol{\Sigma}_{species}\mathbf{1}_S\\ &= \frac{1}{S}\left(\frac{1}{N}\boldsymbol{\Omega}\mathbf{1}_S - \frac{1}{N^2}\boldsymbol{\omega}\boldsymbol{\omega}^T\mathbf{1}_S\right)\\ SN\bar{\boldsymbol{\rho}} &= \boldsymbol{\psi} - \frac{f}{N}\boldsymbol{\omega}\\ N\bar{\boldsymbol{\rho}} &= \frac{1}{S}\boldsymbol{\psi} - \frac{1}{\beta_W}\boldsymbol{\omega}\\ &= \boldsymbol{\psi}^* - \frac{1}{\beta_W}\boldsymbol{\omega}\end{aligned} \quad (7)$$

The above orgy of Greek letters and symbols will be used now to find a common thread among a variety of indices of biodiversity pattern, based on the matrices **A** and $\boldsymbol{\Omega}$, and the vectors $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$, $\boldsymbol{\psi}$, $\boldsymbol{\varphi}$, $\bar{\boldsymbol{\rho}}$ and $\bar{\boldsymbol{\tau}}$.

Relations to Biodiversity Pattern Indices

Whittaker's Beta Diversity

The two traces in equations (1) are equal, by definition of trace. Therefore:

$$N\bar{\alpha} = trace(\mathbf{A}) = trace(\boldsymbol{\Omega}) = S\bar{\omega} = f \quad (8),$$

which immediately implies $S = \bar{\alpha}\frac{N}{\bar{\omega}}$. Whittaker, (1972) defined his first measure of beta diversity as $\beta_W = S/\bar{\alpha} = NS/f$, which means that Whittaker's beta is mathematically equivalent to the reciprocal of the average proportion of the total region (*N*) occupied by the species inhabiting the region (Arita and Rodríguez 2002; Routledge 1977; Schluter and Ricklefs 1993; Vellend 2001). Both $\bar{\alpha}$ and β_W are invariant to randomization of the matrix values subject to the condition that the total count of presences and the dimensions of the matrix remain constant, as is obvious from (1), because both traces are simply the fill of the **X** matrix (the total number of ones). For this reason β_W is not a measure of turnover (in which case it would be sensitive to the actual physical positions of presences and absences), but rather it is a simple measure of how much more diverse is an entire collection than the average of its subsets (Vellend 2001).

Lande's Beta Diversity

Another measure of beta diversity, the so-called additive beta diversity, (Lande 1996) is being used again (Veech et al. 2002): $\beta_A = S - \bar{\alpha}$. For presence-absence matrices this measure is related very simply to Whittaker's multiplicative measure by

$$\beta_A = S(1 - 1/\beta_W) \quad (6)$$

This equation is obtained by substitution of the value of $\bar{\alpha}$ in the additive beta formula (Lande 1996; Veech et al. 2002; Ricotta 2005), as β_W, and for the same reason, if the dimensions of the PAM and its fill remain constant, β_A are insensitive to permutations of ones and zeroes.

Legendre's Beta Diversity

Legendre et al. (2005), proposed as a new measure of beta diversity: the total sum of squares, or SS(X) = β_L, of the species-composition matrix. This definition is inspired from equating the concept of beta diversity to "variation in species composition among sites…" (Legendre et al. 2005) and is made formal by use of equation (1) of Legendre et al. (2005) which shows that SS(X) is mathematically equivalent to the sum of squared Euclidean distances among sites, divided by N, the number of sites.

The total sum of squares mentioned in Legendre et al. (2005) is, by definition, the trace of the species variance-covariance matrix:

$$\beta_L = trace(\mathbf{\Sigma}_{species}) = \frac{S}{\beta_W} - \bar{\varphi}^* \quad (7)$$

Schluter's Variance Ratios

Schluter (1984) introduced a ratio of variances test to assess simultaneously whether species in a group are associated. The method compares the observed variance in the total number of species in samples, with the variance expected if occurrence of each species is independent of the others. For presence-absence data there are two such ratios, one to test the association of species in communities, and another to test overlap of ranges of distribution. In the notation of Schluter, the row-sums, which we denote as α_j are called T_j. The column sums (ω_j), in Schluter (1984) are denoted by n_i. Making these substitutions and using the identities (4) yields:

$$V_{com} = \frac{(1/N)\sum_{i=1}^{N}(\alpha_i - \bar{\alpha})^2}{\sum_{j=1}^{S}\omega_j^*(1-\omega_j^*)} = \frac{\bar{\varphi}^* - S/\beta_W^2}{1/\beta_W - \bar{\psi}^*/N}$$

$$V_{range} = \frac{(1/S)\sum_{j=1}^{S}(\omega_j - \bar{\omega})^2}{\sum_{i=1}^{N}\alpha_i^*(1-\alpha_j^*)} = \frac{\bar{\psi}^* - N/\beta_W^2}{1/\beta_W - \bar{\varphi}^*/S}$$

Nestedness

One of the earliest indices of nestedness of a PAM is that of Wright and Reeves (1992) which depends only on the marginal values of the PAM, and is defined, with the notation of this paper, as

$$\begin{aligned} N_C &= \frac{1}{2}\sum_{j=1}^{S}\omega_j(\omega_j - 1) \\ &= \frac{1}{2}\left(\sum_{j=1}^{S}\omega_j^2 - \sum_{j=1}^{S}\omega_j\right) \\ &= \frac{1}{2}\left(N\bar{\varphi} - f\right) \\ &= \frac{N}{2}\left(\bar{\varphi} - \frac{S}{\beta_W}\right) \end{aligned}$$

A more recent index proposed by Almeida-Neto et al. (2008) is based on the idea that both nestedness of community-composition and of distributional ranges should be taken into account. They propose ordering a matrix by column and row marginals, and then comparing every pair *i, h* of rows such that $\alpha_i > \alpha_h$, and every pair of columns such that $\omega_j > \omega_k$, adding all the corresponding values of **A** and **Ω** and dividing by the number of compared pairs. This index is more an algorithm than a formula, but it can easily be expressed in terms of the mathematical objects defined above.

C-scores

In 1990, Stone and Roberts (1990) addressed the problem of whether the distributions of pairs of species are "random" in the sense of not being different of what they would be if the species did not interact. Their *C* index is defined using the

number of co-occurrences (entries in the **Ω** matrix above). First they looked for "checkerboard units," which are patterns of the form {1, 0}, {0, 1}, i.e., for two species, pairs of sites where they do not co-occur. Using the notation of this paper, they define the number of checkerboard units as (Stone and Roberts 1990):

$$C_{j,k} = (\omega_j - \omega_{j,k})(\omega_k - \omega_{j,k})$$

and their C index is then

$$C = \frac{2}{S(S-1)} \sum_{j=1}^{S} \sum_{k>j} C_{j,k}$$

The above expands to three sums that can be shown to be, using all the above definitions and relationships:

$$\sum_{j=1}^{S} \sum_{k>j} \omega_j \omega_k = \frac{(S\bar{\omega})^2 - N\bar{\varphi}}{2};$$

$$\sum_{j=1}^{S} \sum_{k>j} \omega_{j,k}(\omega_j + \omega_k) = \boldsymbol{\omega}^T \boldsymbol{\psi} - N\bar{\varphi}$$

$$\sum_{j=1}^{S} \sum_{k>j} (\omega_{j,k})^2 = \frac{\mathbf{1}_S^T (\boldsymbol{\Omega} \otimes \boldsymbol{\Omega}) \mathbf{1}_S - N\bar{\varphi}}{2}$$

where the symbol ⊗ represents the Hadamard product (element by element) of two compatible matrices.

When the only available information is presence-absence, probably most other indices of biodiversity pattern can be reduced to operations between the **A** and **Ω** matrices, or the vectors describing marginals (**α** and **ω**), covariances ($\bar{\boldsymbol{\rho}}$ and $\bar{\boldsymbol{\tau}}$) or the closely related fields (**ψ** and **φ**).

Mathematical Constraints

The relationships among the above different measures of pattern are constrained by their mathematical properties (Soberón and Ceballos 2011), and the constrained graphs are very useful to display in an aggregated way the properties of biodiversity patterns (Arita et al. 2008). Moreover, when these graphs are linked to the corresponding cartographic information in such a way that "brushing" (in GIS terminology) different parts of the graph highlights the corresponding parts areas in a map and vice versa, very interesting relationships are revealed, with biogeographic or conservation implications (Soberón and Ceballos 2011; Villalobos et al. 2013).

In the remainder of the paper, we describe a software implementation that allows inspection of scatterplots among pairs of indices. Our software allows highlighting jointly the locations in maps where given combinations occur, and positions in the scatterplot. Moreover, although we do not yet provide a theoretical background, our software also include indices of phylogenetic proximity, enabling in this way a linked perspective of geographic and phylogenetic structure and covariance among the corresponding indices. Statistical testing is enabled by providing fast bootstrapping for large matrices. Our software is the first implementation that makes full use of the theory described above.

Software Implementation

To implement the above, we extended the platform Lifemapper [2] to include multi-species range and diversity experiment construction, calculations and hypothesis testing for PAMs in a module called LmRAD (Lifemapper Range and Diversity). The platform's services are made available through a customized LmRAD plug-in for QGIS that includes point and click functionality for building and analyzing PAMs in user-defined regions. This plug-in is already implemented and available in QGIS. Linked custom data visualization spaces are also implemented inside QGIS. We will describe the details of using the plugin in a forthcoming paper.

To see how the linkages work consider that the statistics derived from the dispersal field (**φ**) and richness (**α**) vectors are measures attached to each geographic locality and thus can be linked to a map using the geographic coordinates of the centroids of the localities (corresponding to the PAM rows). On the other hand, statistics based on the diversity field **ψ** and the range size vector **ω** result in measures attached to each species in a PAM, so that the column space of the PAM can be linked to a phylogenetic tree, using the names as common field. Scatter plots can be species-based (**ω**, **ψ**) or site-based (**α** and **ω**), (Arita et al. 2008). In both cases, the dispersion of points in the plots is

[2] http://www.lifemapper.org.

determined by patterns of species' co-occurrence, and the key idea is to link plots of site-based, or species-based statistics to maps or phylogenetic trees respect-tively. We plot such associations in QGIS through the plug-in and link to a custom phylogenetic viewer built in QGIS, thus allowing a display of site-based and species-based statistics in plots that are dynamically linked to the map or the tree.

For instance, in Figure 1, we plot the proportional number of species vs. the mean proportional range-size (both site-based numbers). Brushing points in this plot highlights sites in the PAM maps that show proportional range size similarity for the species in those sites and species-numbers similarity among the sites. The Lifemapper QGIS tool also provides tools for mapping the entire site-based statistics from the PAM so that biogeographic patterns in the maps can guide brushing. Layering the statistics with any number of GIS layers for the areas of interest can add to the visual analysis. The phylogenetic statistics from the tree for the species are also aggregated spatially in the tool and can show interesting patterns. The tool also links the shared community composition of species from the PAM to taxon distance statistics derived from the tree in a correlation coefficient of taxon distance to sites shared between species. The site brushing is multi-directional and can go from map-to-plot and plot-to-map (see Figure 1).

The site-based indices that we described above have a natural correspondence to maps because essentially every cell in the grid has a value for all the indices. For species-based analysis the natural corresponding structure would be a tree. The phylogenetic data structure that drives the tree visualization is used to calculate dynamically the mean nearest taxon distance for selected species in the species association plot. In this way the spatially derived statistics for diversity can be compared to the degree of phylogenetic relatedness within species communities. Individual species from those communities can also be sub-selected and their ranges from the PAM shown in the PAM based maps. Selections by clade directly in the tree will vice-versa select those points in the plot (see Figure 2).

Linkages and visual analysis of range-diversity relationships derived from very large PAMs are achieved through custom visualization spaces inside of QGIS, but the construction and the outputs from calculations on the PAMs (exceeding 10^8 elements) are achieved by compute modules that interact with the visualization environment in a client-server relationship through web-services architecture. Compute services for these very large jobs are exposed as Open Geospatial Consortium (OGC)[3] Web Processing Services, and RESTful web-services. This scheme permits to spread the computational load for working with thousands of PAM inputs and calculations using node parallelization and data parallelization across remote distributed computing environments so that dealing with large matrix operations is not dependent on desktop resources, thus freeing the QGIS tool to do what it does best, preparing spatial inputs and visualizing the outputs spatially to discern biogeographic patterns. The advantages of using QGIS was to leverage an entry level but powerful and extensible GIS tool for users to be able to work with spatial data, prepare them, do spatial analyses on outputs, and the tools to define and upload LmRAD experiments through the tool.

The Lifemapper infrastructure is composed of a central management component, LmDbServer, which manages data and analysis operations with a "data pipeline" written in Python [4] and a PostgreSQL/PostGIS database; multiple instances of LmCompute that can be co-located across institutions (currently this is deployed at compute clusters at University of Kansas, the University of Florida, and San Diego Supercomputer Center); and LmWebServer which manages all communications between the components and client applications. (See Figure 3.)

LmRAD is a job-based infrastructure that is environmentally agnostic and its algorithms are portable across compute environments through configurable instances of LmCompute. LmWebServer contains a Job Server tier that feeds compute jobs to any compute environment that can sponsor an instance of LmCompute. The compute plug-in for a specific resource receives compute jobs for PAMs through a job controller which determines which of several plugins are appropriate for the type of calculation. The pipeline and LmDbServer are responsible for presenting jobs to the Job Server and moving jobs through the

[3] http://www.opengeospatial.org.
[4] http://www.python.org.

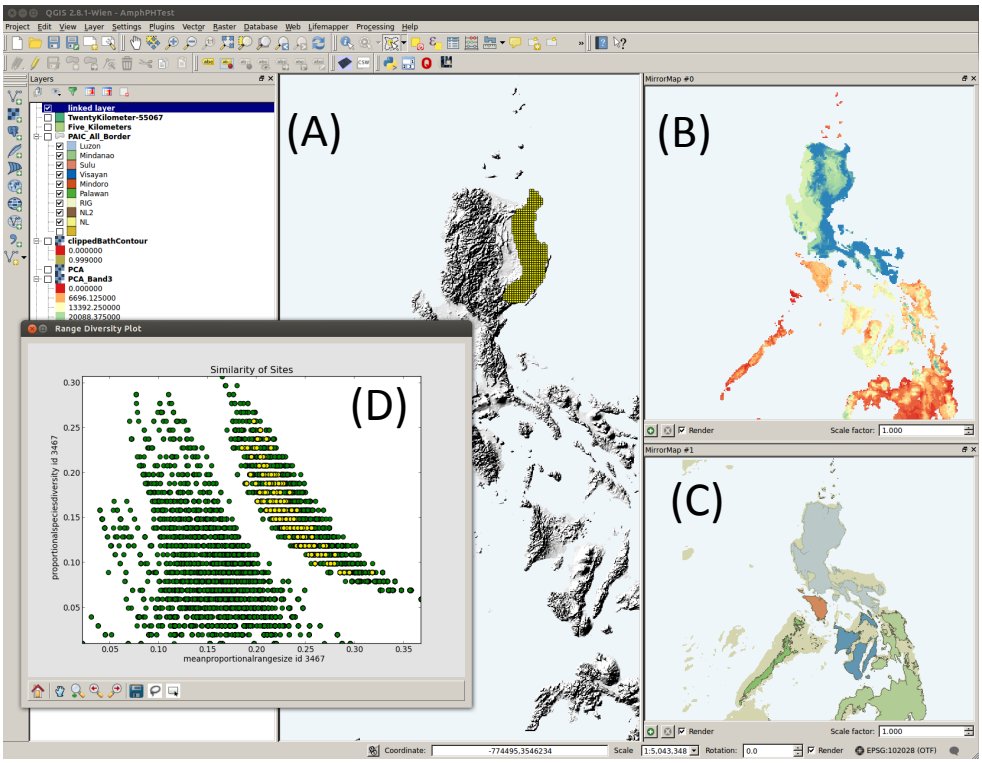

Figure 1. A screen-capture of the QGIS Lifemapper plugin showing a sites-based output. Site similarity plot for amphibians of the Philippines, showing highlands in Luzon (A, in yellow). The value of the mean proportional range size (B), the emerged relief during the glacial maximum (C), and the "brushed" Luzon cells in the scatterplot of richness vs mean proportional range-size (D).

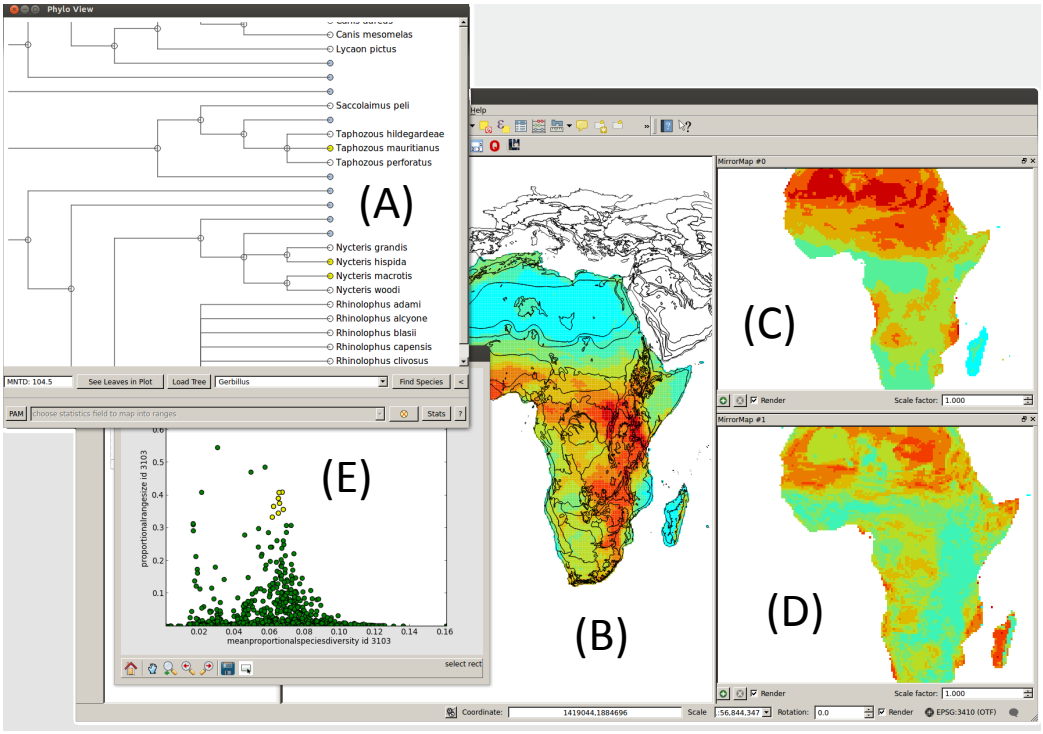

Figure 2. A species-based screen-capture of the QGIS plugin showing part of a mammal phylogeny (A) connected to a 800+ mammals of Africa PAM, with a map of species richness (B), mean proportional species-diversity (C), mean nearest taxon distance (D) and a scatterplot of range-size vs. mean phylogenetic distance (E) with the "brushed" species highlighted (yellow species in (A) and (E)).

system. At different stages in a LmRAD experiment, dependencies and statuses are updated by the compute environment which posts back to the Job Server. PAM construction has been parallelized across processors on any compute environment that receives a PAM job. Data products for large PAMs (extents $>10^6$ km^2, resolutions 10 km or less, $>10^3$ species) can be constructed and analyzed in this way with reasonable response times. Results from the experiment are then posted back to the Job Server from the compute environment and are written to the database and file system shared by the LmDbServer and LmWebserver.

Data parallelization across multi-core architectures on each of the nodes in a compute cluster helps to speed large PAM construction jobs. PAM construction uses a combination of Rtree[5] and matplotlib's nxutils and GDAL[6] for vector and raster based intersections, respectively. Calculations on the matrices use NumPy[7] built with the Basic Linear Algebra Subprograms (BLAS). Permutations on the PAM matrices for hypothesis testing against null models use methods that are specific to binary matrices, where row and column totals can both be kept intact while changing the mix of species in sites and the range size of each species. Data parallelization is not suited to these computations since the entire matrix needs to be taken into account. However, since several hundred permutations may be required per experiment, the current job based parallelization across compute nodes works well for computing these models *in toto*. Another method in LmRAD for permuting the matrix is perfectly suited for both types of parallelization. It uses a dye dispersion algorithm which is a 2-dimensional geometric-constraints model that assumes range continuity (Jetz and Rahbek 2001). Since range allocations are reassembled individually for each species, those data can be split across cores on a single machine or across nodes.

The Lifemapper plugin allows QGIS to operate as a web service client to the architecture described above, edit and submit data, parameterize inputs and request computations. Experiment results can be pulled down as statistical and geospatial outputs and linked to phylogenetic trees and range-diversity plots that depict the 6 vectors described above, species richness and range size vectors, Whitakker's Beta, Legendre's Beta, and Lande additive beta. Range-diversity plots are produced in the plug-in that summarize these fields as indexes of site similarity and the degree of association of species, allowing the user to experiment across scale, geographic extent and PAM grid resolution. The range-diversity plots and the tree viewer are custom visualization spaces built for the plug-in inside of QGIS something made possible because the user community for QGIS is free to customize QGIS through plug-in development to do a wide variety of analysis with access to all of the QGIS functionality through a Python API.

An example of a unique environment for QGIS is the tree viewer for LmRAD. The tree viewer presents the phylogenetic data as interactive SVG built dynamically from incrementally loaded JavaScript Object Notation (JSON) data using web-based techniques in a document driven JavaScript framework. Using advances in web-based JavaScript visualization libraries alleviates the need for the user to install external libraries when installing the plugin. There are several tree formats, e.g. phyloXML[8], Newick[9], Nexus (Maddison et al. 1997), NeXML[10] and NexSON, used by the Open Tree of Life[11] for connections to web-services. Since these allow construction of tree databases that are either JSON or are easily translated into JSON, they can be directly mapped to Python dictionaries, and are easily transported back and forth from LmCompute for analysis and they are ideal for a document driven visualization framework. Visualization then is made possible with the JavaScript library for Data Driven Documents (D3) D3.js[12]. D3 allows the JSON document to be dynamically bound to the Document Object Model so that data-driven transformations can be applied to the document with smooth transitions and fluid interaction. Such smooth transitions are especially useful when navigating large trees with many nodes and edges.

The D3 based interactive tree is rendered in the plug-in through a Qt dialog using QtWebKit.

[5] https://pypi.python.org/pypi/Rtree/.
[6] http://www.gdal.org/.
[7] http://www.numpy.org/.
[8] www.phyloxml.org.
[9] http://bit.ly/1n6ELcZ.
[10] http://nexml.org.
[11] http://blog.opentreeoflife.org/.
[12] http://d3js.org/.

Communication between the tree and the rest of the plug-in is effected by QtWebKit Bridge. The bridge allows the JavaScript and PyQt objects to communicate with one another. The tree viewer is linked to the interactive range-diversity plots in matplotlib (Hunter 2007) by simple PyQt signals and slots. A similar method connects the range-diversity plots for site-based statistics to the maps in QGIS based on the PAM. Using JavaScript in PyQt dialogs for QGIS allowed us to achieve fluid visual representations of trees for large clades, e.g. one tree used in testing is the entire phylogeny for the Phylum Mollusca with over 85,000 nodes.

DISCUSSION

Availability of biodiversity data is increasing very rapidly allowing researchers, in principle, to perform a large number of analyses. However, as long as the different perspectives (phylogenetic, ecologic, morphologic, functional and others, see MacLaurin & Sterelny (2008)) remain unlinked, the analyses are mostly disconnected from one another. In fact, it is possible to say that the fundamental task of biodiversity informatics should be to enable "integration" of different perspectives of biodiversity (Harmon et al. 2013; Miller and Jolley-Rogers 2014; Peterson et al. 2010). Integration is not a well-defined term, but one possible meaning may be the capacity to display simultaneously different perspectives of the data (Laffan et al. 2010), preferably in a linked way. For instance, in Figure 2, a simultaneous and linked display of phylogenetic, ecological and geographical data is presented.

Software has been developed that links geographic display of data with a variety of mathematical graphs, but very few integrated systems exist that address biogeography, community assembly, ecological niche and phylogeny. Web-based solutions for viewing phylogenies are popular but are limiting in that geospatial tools for the web that allow *ad hoc* analysis of range data as character traits for species in trees struggle to keep pace with desktop GIS implementations. Most of these implementations choose to focus on simple clade-area relationships. A few of these contain similar analysis to LmRAD and have spatial components that can be compared with our tools.

Spatial Analysis in Macroecology (Rangel et al. 2010) is a software alternative for biodiversity experiments that directly influenced LmRAD. SAM offers a comprehensive set of tools for spatial statistics, some simple mapping tools and advanced spatial autoregression models. It uses extremely optimized linear algebra libraries for large matrix operations. The data table in SAM accommodates PAM data, and can be formatted as

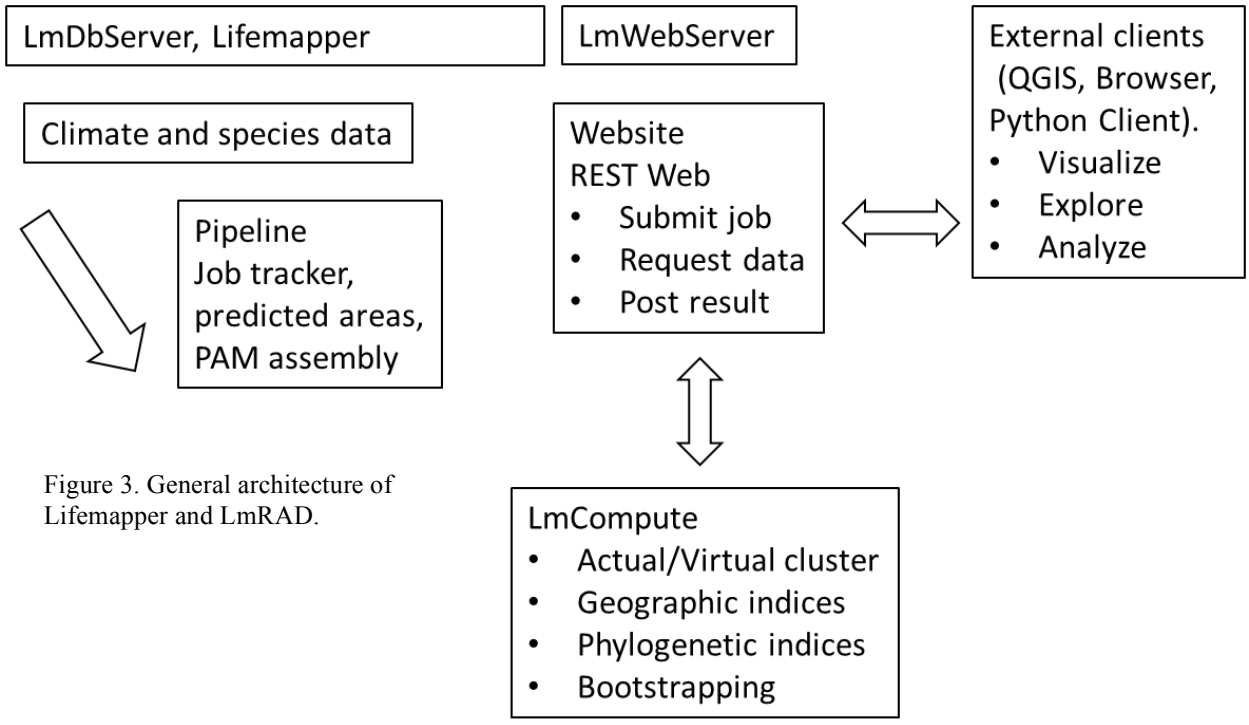

Figure 3. General architecture of Lifemapper and LmRAD.

ESRI shapefiles. Just as in LmRAD, data grids can be prepared in SAM at any extent and resolution as shapefiles and PAMs can be generated directly from the shapefile. Addition-ally, as in LmRAD, SAM allows a user to link scatterplots and maps geographically, where grid cells can be selected in a map and then correspondingly highlighted in a scatter plot or vice versa, allowing a user to detect outliers. SAM also has sophisticated tools for evaluating the changes in spatial correlation as they are affected by changes in scale. On the other hand, SAM does not incorporate phylogenetic analysis or visualization into its data linkages and it is a Windows-dependent desktop software reliant on the processing capacity of the desktop, where the LmRAD plugin is cross-platform and interfaces with remote WPS services so that little processing occurs on the user's machine.

EcoSim (Gotelli and Entsminger 2011) is another Windows-based macroecology software built specifically for dealing with null model hypotheses testing and PAM data. It provides a variety of randomization routines for each of its modules. The co-occurence module randomization routine allows row and column constraints, including fixed-sum similar to LmRAD. EcoSim also allows equiprobable, proportional, and weighted constraints (Gotelli and Entsminger 2011). LmRAD currently has two randomization algorithms. EcoSim has four different ways of dealing with sparse or degenerate matrices (with empty rows and columns; (Ellison 2000). LmRAD currently has a compression algorithm for compressing and re-expanding such matrices which speeds processing time. The limiting factor for EcoSim seems to be the size of the PAM; 240,000 cells, or approximately 800 by 300 rows and columns is an absolute limit. One of the core requirements first addressed by LmRAD was being able to work with much larger matrices. Initial tests in LmRAD for randomization algorithms were done for matrices on the order of 6.0×10^8 cells. Data products for large PAMs at high resolutions (10 km) with upwards of 800 species can be constructed and analyzed with reasonable response times in LmRAD.

Biodiverse (Laffan et al. 2010), an open-source project similar to the Lifemapper plug-in, provides linked visualization across different dataspaces. Biodiverse links species distributions in geographic, phylogenetic, taxonomic and environmental spaces. One advantage of Biodiverse is that scale comparisons are achieved through a window analysis for endemism, phylogenetic diversity, and beta diversity. By varying the size of the windows one can start to understand the effects of scale on those statistics. Currently the Lifemapper plug-in uses a multi-grid approach where several subsets at different cell resolution can be built out within the same experiment, allowing comparisons across scale for the range and diversity statistics including beta diversity.

OpenGeoDa is a free and open source crosss-platform package for exploratory spatial analysis. Its strengths are techniques for dynamic linking and brushing data across multiple data spaces (Anselin et al. 2005). It is like LmRAD in this respect and strives for integration across different measures of spatial association rather than the specific biodiversity focus in LmRAD. OpenGeoDA then represents a very powerful set of tools found in mature desktop GIS applications that are integrated beyond what most GIS applications provide.

Unlike most of the mentioned software, LmRAD is natively oriented to assembling and organizing very large datasets, and, finally, because we provide a simple but robust set of mathematical formulae that uncover the relationships and constraints among many of the main biodiversity indices LmRAD is unique in its integration of species relatedness and spatial components for range, diversity fields and dispersion fields.

The next step in developing the tool is to allow for linking other perspectives of biodiversity as high-end Web Services, a task already under way (Harmon et al. 2013). Nevertheless, the future development of “integrating” views of biodiversity goes well beyond linking displays. Ideally, integration should also strive to express theoretical relationships among different views of biodiversity, preferably in a statistical or mathematical way. It is already possible to analyze statistically multiple views of biodiversity, as it has been demonstrated by a number of authors (Doledec et al. 2000; Doledec et al. 1996; Leibold et al. 2010). In this way a fuller understanding of how different aspects of biodiversity are related, and how, can be achieved.

ACKNOWLEDGMENTS

NSF/BIO/AVAToL award 1208472 supported both authors. We are grateful to our colleagues H. Arita, P. Rodríguez, F. Villalobos and A. Lira for endless conversations on the issue of biodiversity patterns, and to Town Peterson for encouragement and advice, and for his time helping us with editorial issues.

LITERATURE CITED

Almeida-Neto, M., P. Guimarães, P. R. Guimarães, R. D. Loyola, and W. Ulrich. 2008. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. Oikos 117:1227-1239.

Almeida-Neto, M., P. R. Guimarães, and T. M. Lewinsohn. 2007. On nestedness analyses: rethinking matrix temperature and anti-nestedness. Oikos 116:716-722.

Anselin, L., I. Syabri, and Y. Kho. 2005. GeoDa: an introduction to spatial data analysis. Geogr. Anal. 38:5-22.

Araujo, M., and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. Glob. Ecol. Biogeogr. 16:743-753.

Arita, H., and P. Rodríguez. 2002. Geographic range, turnover rate and scaling of species diversity. Ecography 25:541-550.

Arita, H. T., J. A. Christen, P. Rodríguez, and J. Soberón. 2008. Species diversity and distribution in presence-absence matrices: mathematical relationships and biological implications. Amer. Nat. 172:519-532.

Borregaard, M. K., and C. Rahbek. 2010. Dispersion fields, diversity fields and null models: uniting range sizes and species richness. Ecography 33:402-407.

Christen, A., and J. Soberón. 2009. Anidamiento y los análisis Rq y Qr en PAMs. Misc. Mat. 49:51-61.

Doledec, S., D. Chessel, and C. Gimaret-Carpentier. 2000. Niche separation in community analysis: a new method. Ecology 81:2914-2927.

Doledec, S., D. Chessel, C. J. F. ter Braak, and S. Champeley. 1996. Matching species traits to environmental variables: a new three-table ordination method. Environ. Ecol. Stat. 3:143-166.

Ellison, A. M. 2000. EcoSim: null models software for Ecology. Bull. Ecol. Soc. Amer. 81:125-127.

Gotelli, N. J. 2000. Null models analysis of species co-occurrence patterns. Ecology 81:2602-2621.

Gotelli, N. J., and G. L. Entsminger. 2011. EcoSim: null models software for ecology. Version 7.0. Acquired Intelligence Inc. & Kessey-Bear, http://homepages.together.net/~gentsmin/ecosim.htm.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325-338.

Graves, G., and C. Rahbek. 2005. Source pool geometry and the assembly of continental avifaunas. Proc. Nat. Acad. Sci. USA 102:7871-7876.

Harmon, L., J. Baumes, C. Hughes, J. Soberón, C. D. Specht, W. Turner, C. Lisle, and R. W. Thacker. 2013. Arbor: Comparative Analysis Workflows for the Tree of Life. PLoS Currents 5

Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment. Compu. Sci. Eng. 9:90-95.

Hurlbert, A. H., and W. Jetz. 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proc. Nat. Acad. Sci. USA 104:13384-13389.

Jetz, W., and C. Rahbek. 2001. Geometric constraints explain much of the species richness pattern in African birds. Proc. Nat. Acad. Sci. USA 98:5661-5666.

Laffan, S. W., E. Lubarsky, and D. F. Rosauer. 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. Ecography 33:643-647.

Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. Oikos 76 5-13.

Legendre, P., D. Borcard, and P. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. Ecol. Monogr. 75:435-450.

Leibold, M. A., E. P. Economo, and P. Peres-Neto. 2010. Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. Ecol. Lett. 13:1290-1299.

Lira-Noriega, A., J. Soberon, A. G. Navarro-Siguenza, Y. Nakazawa, and A. T. Peterson. 2007. Scale dependency of diversity components estimated from primary biodiversity data and distribution maps. Diversity Distrib. 13:185-195.

Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28:385-393.

MacLaurin, J., and K. Sterelny. 2008. What is Biodiversity? University of Chicago Press, Chicago.

Maddison, D. R., D. L. Swofford, and W. P. Maddison. 1997. Nexus: an extensible file format for systematic information. Syst. Biol. 46:590-621.

Miller, J. T., and G. Jolley-Rogers. 2014. Correcting the disconnect between phylogenetics and biodiversity informatics. Zootaxa 3754:195-200.

Orme, D. L., R. Davies, V. A. Olson, G. H. Thomas, T.-T. Ding, P. C. Rasmussen, R. S. Ridgeley, A. Stattersfield, P. M. Bennett, I. P. F. Owens, T. M. Blackburn, and J. K. Gaston. 2006. Global patterns of geographic range size. PLoS Biol. 4:1276-1283.

Peterson, A. T., S. Knapp, R. Guralnick, J. Soberon, and M. T. Holder. 2010. The big questions for biodiversity informatics. Syst. Biodivers. 8:159-168.

Peterson, A. T., J. Soberón, R. G. Pearson, R. Anderson, E. Martínez-Meyer, M. Nakamura, and M. Araújo. 2011. Ecological Niches and Geographic Distributions. Princeton University Press, Princeton.

Rahbek, C., and G. Graves. 2000. Detection of macro-ecological patterns in South American hummingbirds is affected by spatial scale. Proc. Roy. Soc. Lond. B 267:2259-2265.

Rangel, T. F., J. A. Diniz-Filho, and M. Bini. 2010. SAM: a comprehensive application for Spatial Analysis in Macroecology. Ecography 33:46-50.

Ricotta, C. 2005. On hierarchical diversity decomposition. J. Veg. Sci. 16:223-226.

Rodríguez-Gironés, M. A., and L. Santamaría. 2006. A new algorithm to calculate the nestedness temperature of presence-absence matrices. J. Biogeogr. 33:924-935.

Routledge, R. D. 1977. On Whittaker's components of diversity. Ecology 58 1120-1127.

Schluter, D. 1984. A variance test for detecting species associations, with some example applications. Ecology 65:998-1005.

Schluter, D., and R. E. Ricklefs. 1993. Species diversity: an introduction to the problem. Pp. 1-12 *in* R. E. Ricklefs and D. Schluter, eds. Species Diversity in Ecological Communities: Historical and Geographical Perspectives. University of Chicago Press, Chicago.

Simberloff, D., and E. F. Connor. 1984. Inferring competition from biogeographic data: a reply to Wright and Biehl. Amer. Nat. 124:429-436.

Soberón, J., and G. Ceballos. 2011. Species richness and range size of the terrestrial mammals of the world: biological signal within mathematical constraints. PLoS ONE 6:e19359.

Stone, L., and A. Roberts. 1990. The checkerboard score and species distributions. Oecologia 85:74-79.

Ulrich, W., M. Almeida-Neto, and N. J. Gotelli. 2009. A consumer's guide to nestedness analysis. Oikos 118:3-17.

Veech, J. A., K. Summerville, T. Crist, and J. Gering. 2002. The additive partitioning of species diversity: recent revival of an old idea. Oikos 99:3-9.

Vellend, M. 2001. Do commonly used indices of beta-diversity measure species turnover? J. Veg. Sci. 12 545-552.

Villalobos, F., A. Lira-Noriega, J. Soberón, and H. T. Arita. 2013. Range–diversity plots for conservation assessments: Using richness and rarity in priority setting. Biol. Cons. 158:313-320.

Whittaker, R. H. 1972. Evolution and measurement of species diversity. Taxon 21:213-251.

Wright, D. H., and J. Reeves. 1992. On the meaning and measurment of nestedness of species aseemblages. Oecologia 92:1432-1939.