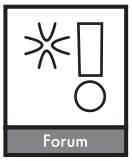

Ecography 36: 1–9, 2013
doi: 10.1111/j.1600-0587.2013.00329.x

Subject Editor: Nathan J. Sanders. Accepted 3 July 2013

Introducing the biogeographic species pool

Daniel Wisbech Carstensen, Jean-Philippe Lessard, Ben G. Holt, Michael Krabbe Borregaard and Carsten Rahbek

D. W. Carstensen (daniel.carstensen@gmail.com), Depto de Botânica, Laboratório de Fenologia, Plant Phenology and Seed Dispersal Group, Inst. de Biociências, Univ. Estadual Paulista (UNESP), Avenida 24-A no. 1515, 13506-900 Rio Claro, São Paulo, Brazil. – J.-P. Lessard, B. G. Holt, M. Krabbe Borregaard and C. Rahbek, Center for Macroecology, Evolution and Climate, Dept of Biology, Univ. of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark. J-PL also at: Quebec Centre for Biodiversity Science, Dept of Biology, McGill Univ., Montreal, QC H3A-1B1, Canada. BGH also at: School of Biological Sciences, Univ. of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK. MKB also at: School of Geography and the Environment, Univ. of Oxford, South Parks Road, Oxford, OX1 3CY, UK.

The species pool concept has played a central role in the development of ecological theory for at least 60 yr. Surprisingly, there is little consensus as to how one should define the species pool, and consequently, no systematic approach exists. Because the definition of the species pool is essential to infer the processes that shape ecological communities, there is a strong incentive to develop an ecologically realistic definition of the species pool based on repeatable and transparent analytical approaches. Recently, several methodological tools have become available to summarize repeated patterns in the geographic distribution of species, phylogenetic clades and taxonomically broad lineages. Here, we present three analytical approaches that can be used to define what we term 'the biogeographic species pool': distance-based clustering analysis, network modularity analysis, and assemblage dispersion fields. The biogeographic species pool defines the pool of potential community members in a broad sense and represents a first step towards a standardized definition of the species pool for the purpose of comparative ecological, evolutionary and biogeographic studies.

Biogeographic species pools to infer historical determinants of community structure

There is an enduring interest in ecology for quantifying the contribution of evolutionary processes and biogeographic history to the structure of local species assemblages (Ricklefs 1987, Harrison and Cornell 2008, Cavender-Bares et al. 2009, Wiens et al. 2011). The contribution of such processes is often evaluated by quantifying the contribution of a predefined species pool to patterns in the structure of local assemblages. The concept of the species pool (also known as the source pool) originates from early studies of ecological communities (Palmgren 1925, Elton 1946, Williams 1947, Patrick 1967) and represents the set of species that could potentially contribute individuals to a local assemblage. In order to relate evolutionary and historical processes to the structure of local communities, however, one needs to define the species pool in a way that explicitly accounts for such processes, rather than using an arbitrary definition. Nevertheless, there is currently no established consensus on what constitutes the most appropriate definition of the species pool.

The review and decision to publish this paper has been taken by the above noted SE. The decision by the handling SE was shared by a second SE.

Traditionally, ecologists have used the concept of the species pool for two distinct purposes: as a way to test whether the structure of communities differ from a random expectation (Connor and Simberloff 1979); and as a way to estimate the influence of the size of the species pool on local species richness (Ricklefs 1987). There is now a long history of research on the attributes of the ecological community (e.g. body size overlap, niche overlap, and phylogenetic dispersion) and how it differs from that of randomly generated communities sampled from a pool of potential colonists (Connor and Simberloff 1979, Strong et al. 1979, Diamond 1982, Gotelli and Graves 1990, Fox and Brown 1993, Graves and Gotelli 1993, Weiher and Keddy 1995, Stone et al. 1996, Gotelli 2000, Gotelli and McCabe 2002). In these studies, the species pool represents the complete set of possibilities, or sampling universe, from which null model algorithms draw species to create 'null communities'. Broadly speaking, these null communities are then used to test whether the observed structure of a given community differs significantly from what would be expected from chance (note that the construction and interpretation of null models can be far more elaborate, see Gotelli and Graves 1996 for a review of the topic). More recently, it has been proposed that an ecologically explicit definition of the species pool in null model analyses can be used to account for and quantify the influence of evolutionary and historical processes

(Swenson et al. 2006, Algar et al. 2011, Lessard et al. 2012a, b). Another common application of the species pool concept is to relate the composition and richness of local communities to that of the regional species pool. For example, many have used a positive and linear relationship between local and regional species richness as evidence for the influence of large-scale evolutionary and historical processes on community structure (Ricklefs 1987, Cornell and Lawton 1992, Ricklefs and Schluter 1993, Srivastava 1999, Shurin et al. 2000, Ricklefs 2007).

If the goal of a study is to use the species pool to infer the role of evolutionary and historical processes on assemblages of species, then biogeographic regions might be viewed as operational species pools. In general, species assemblages can be expected to share much history with other assemblages within a biogeographic region, but relatively little with those in other biogeographic regions. Although the delineation of biogeographic regions based on assemblage similarities does not explicitly account for any particular underlying processes, dispersal between assemblages within a biogeographic region are most likely possible within an historical, if not ecological, timeframe. Biogeographic regions thus define the regional species pool in a broad sense. This is the rationale for the biogeographic species pool concept.

In the following, we briefly summarize methods that have been used to define species pools. We then describe recent methodological advances in delineating biogeographic regions and explain how and why they are useful in defining what we term the 'biogeographic species pool'. Although a multitude of methods exist to delineate biogeographic regions, we here focus on grouping of species assemblages based on distance-based clustering and network modularity analysis. We also describe the use of assemblage dispersion fields, an assemblage-specific tool for defining species pools. Finally, we discuss future perspectives and applications of the biogeographic species pool.

A brief history of the species pool

Early work on community structure used species lists from habitats, geopolitical provinces or roughly defined biogeographic regions to determine the composition of the species pool (Elton 1946, Williams 1947). In later years, the species pool has played a central role in the heated debate over the importance of competition for island community assembly. Following a controversial publication by Diamond (1975), many studies asked whether patterns of community structure could arise by chance alone, by comparing the observed pattern to that of null communities generated by random sampling from a species pool (Connor and Simberloff 1978, 1979, Strong et al. 1979, Grant and Abbott 1980). The species pool was constructed using a wide variety of methodological approaches, for example, from using species lists of the adjacent mainland (Grant 1966, Faaborg 1979) or the archipelago itself (Simberloff and Connor 1978), to using all the species within a predefined area (Simberloff 1970). The problems associated with the definition of the species pool were recognized early (Simberloff 1970) and it was subsequently demonstrated that the definition of the species pool could affect results (Schoener 1988). In sum, species pools were too often subjectively or arbitrarily defined, and suffered from the lack of standardized analytical approaches.

Recognizing the problems of earlier definitions and the importance of realistic species pools against which to test for non-randomness, Graves and Gotelli (1983) proposed a new approach for constructing species pools for the assemblages of birds on land-bridge islands off Panama and northern South America (also see the book cover of Gotelli and Graves 1996). They argued that focal communities in a region (e.g. the avifaunas of each island) will not have identical species pools, but that differently positioned communities will draw species from different pools. To account for such distance effects, they defined a species pool for each assemblage by including the species breeding within an area defined by a circle with a fixed radius centered on the island location nearest to the mainland. In spite of methodological problems associated with it, this method is still in use today (Belmaker and Jetz 2012), because it is convenient, transparent, and makes fewer unrealistic assumptions regarding dispersal probabilities than other classical approaches. However, the radii of concentric circles are more or less arbitrarily fixed because information on the dispersal capabilities of species is usually lacking. Hence, species pools defined using this approach do not necessarily reflect accurately the pool of species capable of colonizing the local community.

Studies of the relationship between the size of the species pool and local species richness have taken a somewhat different approach. Most often, such studies define the species pool by pooling the species obtained by local-scale surveys (Terborgh and Faaborg 1980, Cornell 1985, Ricklefs 1987, Belote et al. 2009, Chase et al. 2011, Kraft et al. 2011, Kristiansen et al. 2011), by using all species known to occur in the region (White and Hurlbert 2010) or a combination of both data sources (Clarke and Lidgard 2000, Ricklefs 2000, Witman et al. 2004). The region, however, is usually more or less arbitrarily defined, e.g. as the island in which the focal assemblage is embedded (Terborgh and Faaborg 1980, Ricklefs 1987), or the geographic scope of the study (Blackburn and Gaston 2001, Sanders et al. 2007).

Today, the increased availability of global distributional and phylogenetic databases provides us with the opportunity to establish a standardized protocol to define the species pool based on occurrence data. Several methodological tools are available to summarize repeated patterns in the geographic distribution of species, phylogenetic clades and taxonomically broad lineages.

Tools for defining the biogeographic species pool: biogeographic regions

The first attempts to delineate biogeographic regions date to the early 19th century (von Humboldt 1806, de Candolle 1820, Sclater 1858, Wallace 1876). Since then, defining ecologically and evolutionarily distinctive regions of the world has been a central aim of biogeography (Simpson 1953, 1977, Darlington 1957, Crowe and Crowe 1982, Cox 2001, Morrone 2002, 2009, Procheş 2005), and new analytical tools have improved our ability to group species assemblages into distinct biogeographic units (Smith 1983, Carstensen and Olesen 2009, Kreft and Jetz 2010, Holt et al. 2013).

Distance-based clustering

An extensive body of literature exists with regard to clustering data points based on observed distances among points (Sokal and Michener 1958, MacQueen 1967) and clustering techniques have been used extensively for the purpose of delineating biogeographic regions (Smith 1983, How and Kitchener 1997, Kreft and Jetz 2010, Linder et al. 2012, Procheş and Ramdhani 2012). The application of clustering algorithms to biogeographic analysis requires the user to choose an algorithm, a distance metric, and the number of clusters to produce (Fig. 1A).

Two general approaches are seen in the biogeographical literature with regard to choosing a clustering algorithm: optimization and consensus. To determine the optimal methodology, the results of different algorithms are generally compared via an evaluation statistic, with the best performing algorithm used in subsequent analyses. For example, the cophenetic correlation statistic (Sokal and Rohlf 1962) is often used to reflect the strength of the correlation between the original distance metrics and distances shown in a hierarchical clustering tree and can be used to identify the best performing algorithm. The choice of an evaluation statistic in itself will have key influence on the final clustering and should be chosen carefully, with due consideration of the overall goal of the study. The alternative approach, consensus, attempts to identify clustering results that are consistent across clustering algorithms. For instance, Penner et al. (2011) used seven different clustering algorithms to build a consensus clustering hierarchy for west African amphibian assemblages. However, since many clustering methods will not be optimal for a particular data set or study goal, combining methods will inevitably produce a sub-optimal and potentially uninformative result.

With regards to appropriate distance metrics, a wide variety of beta-diversity and species turnover metrics have been used to quantify the distance between two or more species assemblages. For biogeographic regionalization, many previous studies have used beta-diversity measures that are heavily influenced by species richness gradients, such as the Sørensen/Bray–Curtis, Jaccard, and Kulczynski metrics (Lennon et al. 2001). These measures may be unsuitable for biogeographic regional analyses as comparisons of nested assemblages (i.e. smaller assemblages contain only a proper subset of species present in larger assemblages) can still return high distance values. Metrics that attempt to purely quantify species turnover, such as βsim, may be more appropriate (Kreft and Jetz 2010).

The decision regarding the final number of clusters is a central issue in cluster analysis. Hierarchical clustering methods produce a complete clustering dendrogram describing the relationships among all objects being analyzed and

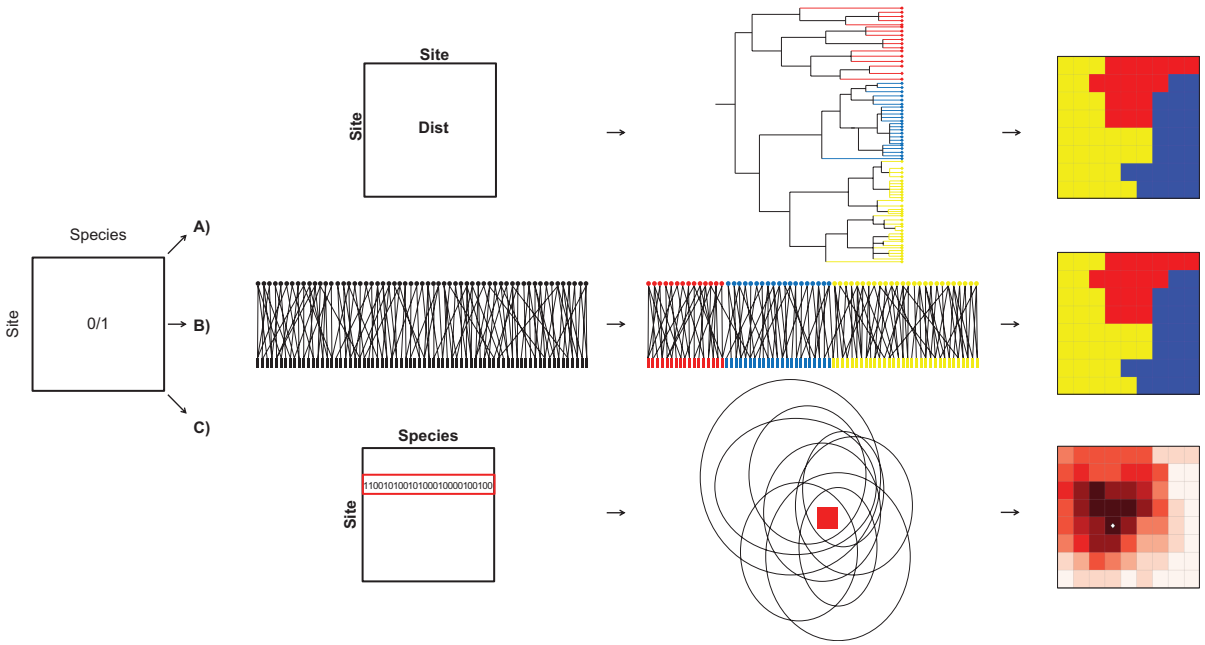

Figure 1. Schematic overview of three approaches to define the biogeographic species pool. (A) Distance-based clustering analysis. The first step of a clustering analysis is calculating the distance between sites (in this case grid cells) according to the similarity of their species assemblages. Similar sites are then classified into groups using a clustering algorithm. The appropriate number of groups is decided upon according to the results of the clustering analysis. The final result is the division of sites into biogeographic regions based on similarities of species assemblages across sites. (B) Network modularity analysis of species assemblages. First, species and sites are arranged as a two-mode (or bipartite) network, with sites (rectangles) and species (circles) sharing a link if the species is present at the site. Sites and species that are strongly interconnected are then grouped using a modularity analysis. The final result is conceptually the same as for clustering analysis; the division of sites into biogeographic regions based on site similarity. (C) Assemblage dispersion fields (ADFs). The ADF approach is site-specific. Sites can thus be included in the regionalization, or species pool, if they share at least one species with the focal assemblage. The ADF of a site is created by overlapping the geographic ranges of all the species occurring in the focal site, creating a site-specific species richness map. The colour intensity of a cell indicates the height of the ADF, that is, darker cells share more species with the focal cell. The focal cell is marked by a white dot.

do not require an a priori decision regarding the number of clusters. Usually, however, biogeographic regionalization analysis attempts to reduce the dimensionality of the original data to produce large scale clusters (of sites, grid cells, etc.). Non-hierarchical clustering approaches require an a priori choice of cluster number. Evaluation methods, such as the *v* folds technique, are used to explore a range of potential cluster sizes to find the optimum number (Rueda et al. 2010, Vasconcelos et al. 2011). Such techniques are also appropriate for hierarchical methods. The less computationally intensive 'finding the knee' method (Salvador and Chan 2004), which attempts to locate the point of maximum curvature on plots of an evaluation statistic (e.g. percentage endemism) against a range of cluster numbers, was suggested by Kreft and Jetz (2010). Alternatively, the maximisation of cluster evaluation statistics, such as mean silhouette values is a commonly used approach (Rousseeuw 1987). More subjective decisions on cluster number, e.g. based on previous regionalization schemes (Heikinheimo et al. 2007) or manual inspection of cluster dendrograms (Procheş and Ramdhani 2012), should be used with caution or avoided altogether if more objective solutions exist. The choice of cluster number is closely related to the choice of clustering algorithm and different algorithms may give hugely different evaluation results for the same number of clusters. Therefore the simultaneous evaluation of many algorithms across an appropriate range of cluster numbers may represent the ideal approach towards making both of these decisions.

Network modularity analysis

Network analysis is widely used in ecology to analyze the organization of interactions among species (May 1973, Paine 1980, Jordano 1987, Cohen et al. 1990, Strogatz 2001, Dunne et al. 2002, Bascompte et al. 2003, Jordano et al. 2003, Olesen et al. 2007). The structural components of a network are nodes (representing species) connected by links (representing interactions). Carstensen and Olesen (2009) proposed the use of a network approach to detect patterns in species distribution data. In the context of island biogeography, islands and species in an archipelago function as network nodes; a link exists between an island and a species if the species is present on the island. Thus, each species is linked to one or more islands and each island is linked to one or more species. The resulting topography (i.e. pattern of connectivity) of this network can then be interpreted in a biogeographic context (Fig. 1B). The network analysis approach allows the grouping of species and sites into distinct compartments, or modules (Girvan and Newman 2002, Newman and Girvan 2004). In analogy to a cluster in a distance-based clustering approach, a module is a compartment of densely connected nodes that are only weakly connected to nodes in other compartments of the network. The overall modularity of the system (described by a modularity index) quantifies how distinctly the system is divided into such compartments. Using species distribution data, modules represent biogeographic regions in which sites share a distinct biota (i.e. as in the distance-based clustering approach).

Delineating biogeographic regions using network analyses requires the choice of a modularity index to assess the modularity of the network and an algorithm to maximize this index. Several indices and optimization algorithms have been proposed over the last decade (Clauset et al. 2004, Newman and Girvan 2004, Duch and Arenas 2005, Guimerà and Amaral 2005b, Newman 2006a, Pons and Latapy 2006, Barber 2007, Guimerà et al. 2007, Rosvall and Bergstrom 2007, Blondel et al. 2008, Ball et al. 2011). The precise definition of a module differs between methods, and no generally accepted definition currently exists. However, the quality of a partition is usually measured by comparing the number of links within and between modules and then measuring whether nodes within modules have more links to each other than expected by chance (Newman and Girvan 2004). The most accurate methods, (Guimerà and Amaral 2005b, Newman 2006b) are computationally slow, whereas faster methods (Clauset et al. 2004) are less accurate (Danon et al. 2005, Costa et al. 2007). In general, a user is faced with a trade-off between accuracy and speed in networking algorithms. However, a recent method, the Louvain method (Blondel et al. 2008), is very fast and has been reported to show good accuracy compared to some much slower methods (Blondel et al. 2008).

The software used by Carstensen and Olesen (2009) to identify biogeographic patterns, Netcarto (Guimerà and Amaral 2005a, b), is based on a simulated annealing algorithm (Kirkpatrick et al. 1983). With this algorithm, nodes are initially placed randomly in a number of modules. Nodes are then stochastically moved between modules, modules are merged and split up at random, and modularity is assessed anew for each system update. If the new modularity value is lower than before, the update can still be accepted with a probability that decreases with the 'system temperature'. The system temperature slowly decreases as the algorithm is running. This allows for a more exhaustive search for the optimal modularity (Guimerà and Amaral 2005b). The algorithm will stop running when the maximum modularity is unchanged for several updates, i.e. the constellation with the maximum value of the modularity function has been found. The method requires the user to choose an iteration factor and a cooling factor, representing another trade-off between speed and accuracy. The number of modules (i.e. biogeographic regions) is an outcome of the algorithm and is unsupervised (Guimerà and Amaral 2005b). Simulated annealing is regarded as the most accurate algorithm for optimizing modularity; however it is also the most computationally demanding and therefore best suited for small datasets (Danon et al. 2005). The Louvain method (Blondel et al. 2008) optimizes modularity in two steps. First it detects small, local modules in the network. These local modules are then used as nodes in a new network. These two steps are repeated iteratively until a maximum modularity is attained (Blondel et al. 2008). This method is fast and could be well suited for large datasets where methods using simulated annealing would be deemed too computationally demanding.

Ultimately, the appropriate choice of algorithm will depend on what is feasible for the size of the dataset. An evaluation of the performance of three different modularity indices in detecting biogeographic regions was performed recently (Thébault 2013); however, further work is needed

to compare the applicability of fast versus accurate algorithms. Compared to a traditional UPGMA cluster analysis, Netcarto, the most accurate algorithm explored thus far, has shown great capabilities for detecting divisions in faunal assemblages (Carstensen and Olesen 2009). However, more comparative studies are needed before one approach can claim superiority for detecting divisions in biotic assemblages.

The network approach holds potential advantages for biogeographic data although these have not been well tested. First, it can provide information on underlying structural patterns of the assemblage delineation. Whereas the distance-based clustering methods group sites according to calculated distances between pairs of sites, the network approach seeks to account for the entire link structure of the network by minimizing links between modules. Because it retains both species and sites during the analysis, the network approach will also assign a region to species, and the regionalization will be based on tight link formations of sites and species, independent of whether some sites have many links (i.e. species) and some few. Whether two sites are grouped together is not just a matter of how many species they share. It is also a matter of how these species themselves are distributed and to which regions they are assigned. Therefore, two sites that are most similar to each other might be assigned to different regions. However, the network approach can report sites that are connected in this manner and identify them for further consideration (Carstensen et al. 2012, 2013). Second, some network algorithms allow for the use of weighted links (i.e. link strengths) when calculating modularity (Newman 2004, Blondel et al. 2008, Dormann and Strauss 2013). Thus, instead of simple presence/absence data, information on species abundance or other quantitative measurements to describe the affinity of a species to a certain site can be used in the site-species matrix when optimizing modularity to delineate biogeographic regions. This could potentially increase the realism of the species pools defined from biogeographic regions. Finally, as the number of modules, or biogeographic regions, is an outcome of the algorithm, the network approach involves fewer choices by the user compared to distance-based clustering and potentially offers a more objective result.

Defining species pools using biogeographic regions

Studies that define species pools simply based on the geographical coverage of the data analysed, fail to address whether there is any biological basis for such a definition. Delineating biogeographic regions provides an opportunity to identify natural groupings within species assemblage data via a transparent, objective framework. Such regions can then be used to define the biogeographic species pool for a particular assemblage by including only those species present at sites within the region where the assemblage is found (Fig. 1). Metrics used to evaluate clustering performance and the optimum number of regions can be used to evaluate individual regions and regions with dubious statistical support can be removed from further analysis. Studies based on biogeographic species pools are focused only on species assemblage data that are grouped in a biologically meaningful way and with strong statistical support. Distance-based clustering and the network modularity approach are conceptually similar and the actual implementation of the two approaches to define the biogeographic species pool is identical; the clusters and modules, produced by the two methods respectively by analysis of the distribution data, can be implemented as operational, broad-scale, species pools (Fig. 1). See Fig. 2 for an example of how biogeographic regions can be used to define the biogeographic species pool.

Tools for defining the biogeographic species pool: assemblage dispersion fields

Assemblage dispersion fields (ADFs) illustrate the spatial structure of species turnover among communities by overlapping the geographic ranges of all species that occur at the site, creating a site-specific species richness map (Fig. 1C) (Graves and Rahbek 2005). The value at any point within the ADF equals the number of species shared with the focal site. Thus, the shape of the ADF reveals the geographic decline of species similarity around the site (Graves and Rahbek 2005, Borregaard and Rahbek 2010).

The geometric shape of ADFs is often highly asymmetrical, and Graves and Rahbek (2005) suggested that the spatial configuration of a dispersion field should reflect the geography of the species pool. This idea seems to be supported by an observed concordance between the shape and extent of global dispersion fields for birds and recognized vegetation biomes (Graves and Rahbek 2005). However, the relationship between dispersion fields and vegetation biomes is still largely conjectural, and awaits further study.

ADFs define a unique source region for each site. In its broadest sense, any site within the ADF (i.e. any site that is inhabited by at least one species that occurs at the focal site) is part of the potential source region for a focal site. Under this very wide definition, the species pool is composed of all species with geographic distributions that overlap the ADF. This species pool definition has an intuitive ecological interpretation: if a site shares any species with the focal site, dispersal between the sites should also potentially be possible for other species. In most cases, however, this approach will lead to a very wide definition of the species pool. For many systems, e.g. for a global analysis of bird communities at large grain size, the broad definition could mean most bird species of the world would be considered part of the species pool. One solution to this problem is to define a threshold level, so that only sites that share at least, e.g. 50% of the species at the focal site are included in the dispersion field. This removes the impact of widespread cosmopolitan species that are not confined within biogeographic regions and thus tend to obscure regional differences. However, the effect of using different threshold levels, which are essentially arbitrary, has yet to be well described.

Lessard et al. (2012b) instead proposed to use ADFs to create probabilistic species pools (i.e. in the context of null model analyses). The height of the dispersion field (z-axis) can be interpreted as a probability distribution, which represents the probability for a site to contribute species to the focal site. Thus, the probability of including a species in a random community is proportional to the dispersion field

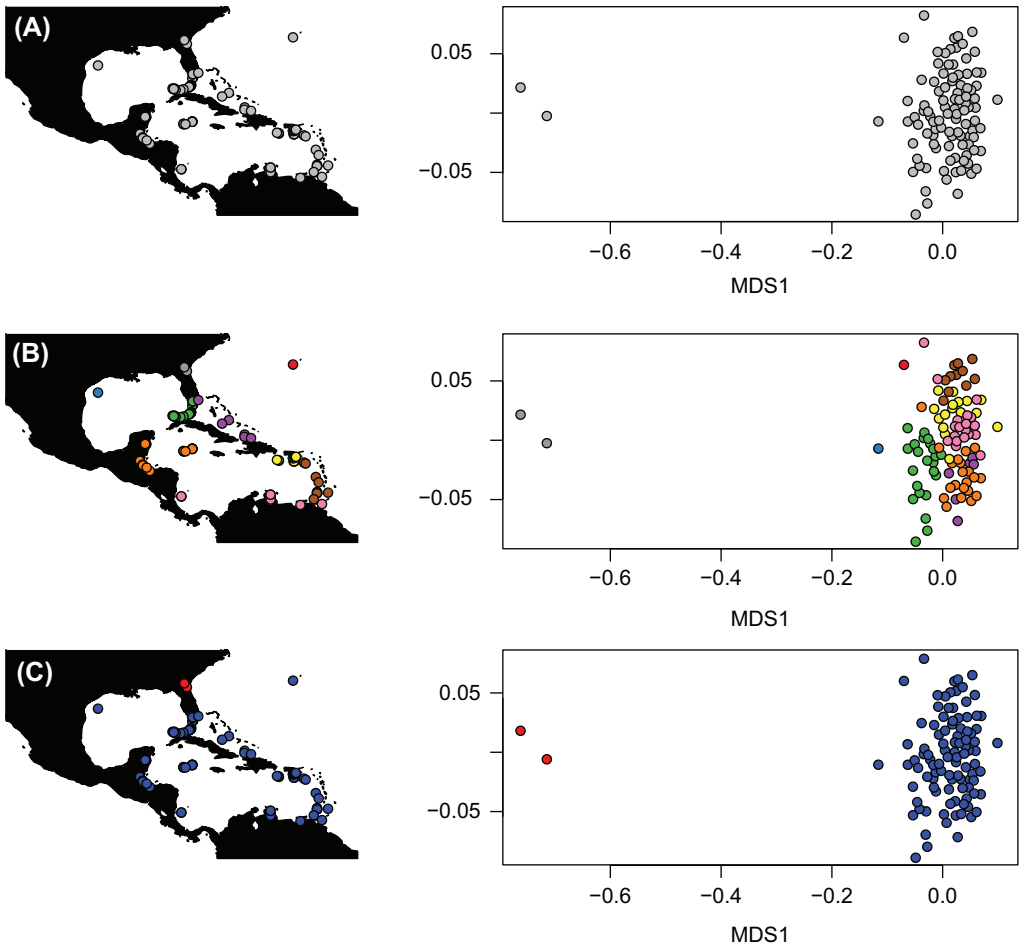

Figure 2. Biogeographic regionalization analysis of fish survey data for 50 randomly selected sites across the Western Atlantic, provided by Reef Environmental Education Foundation (REEF 2011). Plots on the right side show two dimension nMDS ordination of the data set. (A) Full data set. A simple case of using a single species pool and analyzing all data together. Such analysis treats all sampled sites as a single biogeographic region, whereas, the associated ordination plot suggests that at least two sites within the data set are in fact relatively distinct from the rest. (B) Biogeographic regions as defined by REEF. Species pool definitions are here based upon external information, in this case the major regions in the survey area as defined by the organization responsible for collecting the data. This delineation is based primarily on geopolitical boundaries. However, the statistical support for many of the identified regions is questionable and thus these regions do not constitute meaningful species pools based on this community data set. (C) Biogeographic regions identified via UPGMA based analysis and PAM based cluster number evaluation. The results of this analysis suggests that the most legitimate manner to divide the data is in to two separate regions, with two sites that are furthest north along the US east coast being separated from the rest of the data. Since both of the regions show high values for the cluster evaluation statistic (mean silhouette values = 0.70 and 0.79) it seems to be appropriate to split the data in this manner.

values of sites occupied by that species. This weighting can be implemented by first picking a site, with a probability equal to the number of species shared with the focal site (i.e. the ADF value), and then picking one species from that site at random. Thus, species occurring at sites that share many species with the focal site are more likely to be part of the species pool for the focal assemblage. An advantage of this method is that it explicitly accounts for the range size distribution of the study organism. This means that, e.g. species pools for taxa with many small-ranged species will consist mainly of species occurring near to the focal site, whereas species pools containing mainly widespread taxa will contain species from a wider geographic area. In sum, dispersion fields have several advantages as a tool to define the biogeographic species pools, as the method is transparent and the outcome does not depend on a choice among algorithms.

Applications of the biogeographic species pool concept and future perspectives

We highlight three analytical methods for defining the biogeographic species pool. We believe that a standardized protocol for defining the species pool in ecological studies will increase consistency and comparability among studies, and remove the ambiguity caused by arbitrary species pool

definitions. Though the highlighted methods all represent an objective way to define the biogeographic species pool, they differ conceptually and methodologically. Clustering and network methods define geographical regions with distinct species composition, whereas ADFs define a unique species pool for each focal site, based on the set of sites that share a certain proportion of the species at that site.

Each of these methods have strengths and weaknesses, and the best approach to use for a given study will depend on the research question. Biogeographic regionalization methods are more appropriate than ADFs for examining the relationship between local and regional species richness, because these techniques define the species pool in a way that is independent of the local species richness. These methods also offer the possibility of varying the size of the 'species pool', e.g. from biomes to habitats, an approach that has been used, to assess the relative roles of history and contemporary processes on the phylogenetic and functional structure of communities (Algar et al. 2011, Lessard et al. 2012b).

For null model tests of local community structure, on the other hand, ADFs may be more appropriate than biogeographic regions (Lessard et al. 2012a). Because ADFs define a site-specific species pool for each local site, they may be more ecologically realistic. Several methodological approaches exist and can be used in combination with ADFs to create more ecologically realistic species pool for the purpose of detecting the signature of local ecological processes (reviewed by Lessard et al. 2012a).

Because biogeographic species pools are based on large geographic regions, they may include species that would never co-occur in local communities due to their inability to colonize or to the lack of habitat suitability. Thus, they may be too broadly defined for studies of community structure that focus on interspecific interactions. Several authors have suggested to address this by defining a smaller 'habitat pool', which only includes the species that may occur in the habitat of the focal site (Graves and Gotelli 1983, Zobel 1997). Note however, that biogeographic species pools based on local assemblage data (rather than grid cells) already will reflect habitat associations (Graves and Gotelli 1983, Zobel 1997), in addition to dispersal barriers and broad-scale abiotic conditions (Graves and Rahbek 2005), and thus partially alleviate this concern.

Recently, an increasing number of studies focus on testing hypotheses regarding the role of evolutionary processes in shaping the structure of regional assemblages over continental to global scales (Stevens 2011, Kissling et al. 2012). Working at very large temporal and spatial scales requires reconsidering the concept of the species pool to include dispersal over evolutionary time scales and adaptation to environmental conditions (Zobel et al. 2011). Promising advances in the delineation of regional assemblages and definition of the species pool might thus come from the integration of distribution data with information on the phylogenetic relationships and functional traits of species.

Acknowledgements – We greatly appreciate the improvements made to the manuscript by Gary Graves. We also thank Fundacão de Amparo à Pesquisa do Estado de São Paulo (DWC) and the Danish National Research Foundation (J-PL, MKB, BGH, and CR) for financial support. BGH furthermore thanks the Marie Curie Actions under the Seventh Framework Programme (PIEF-GA-2009-252888). MKB is currently supported by a postdoctoral grant from the Danish Councils for Independent Research. We gratefully acknowledge the staff and volunteers at Reef Environmental Education Foundation for collecting and providing the example data for this work.

References

Algar, A. C. et al. 2011. Quantifying the importance of regional and local filters for community trait structure in tropical and temperate zones. – Ecology 92: 903–914.

Ball, B. et al. 2011. An efficient and principled method for detecting communities in networks. – Phys. Rev. E 84: 036103.

Barber, M. J. 2007. Modularity and community detection in bipartite networks. – Phys. Rev. E 76: 066102.

Bascompte, J. P. et al. 2003. The nested assembly of plant–animal mutualistic networks. – Proc. Natl Acad. Sci. USA 100: 9383–9387.

Belmaker, J. and Jetz, W. 2012. Regional pools and environmental controls of vertebrate richness. – Am. Nat. 179: 512–523.

Belote, R. T. et al. 2009. Disturbance alters local–regional richness relationships in Appalachian forests. – Ecology 90: 2940–2947.

Blackburn, T. M. and Gaston, K. J. 2001. Local avian assemblages as random draws from regional pools. – Ecography 24: 50–58.

Blondel, V. D. et al. 2008. Fast unfolding of communities in large networks. – J. Stat. Mech. Theory E 10: P1000.

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

Carstensen, D. W. and Olesen, J. M. 2009. Wallacea and its nectarivorous birds: nestedness and modules. – J. Biogeogr. 36: 1540–1550.

Carstensen, D. W. et al. 2012. Biogeographical modules and island roles: a comparison of Wallacea and the West Indies. – J. Biogeogr. 39: 739–749.

Carstensen, D. W. et al. 2013. The functional biogeography of species: biogeographical species roles of birds in Wallacea and the West Indies. – Ecography doi: 10.1111/j.1600-0587.2012.00223.x

Cavender-Bares, J. et al. 2009. The merging of community ecology and phylogenetic biology. – Ecol. Lett. 12: 693–715.

Chase, J. M. et al. 2011. Using null models to disentangle variation in community dissimilarity from variation in α-diversity. – Ecosphere 2: art24.

Clarke, A. and Lidgard, S. 2000. Spatial patterns of diversity in the sea: bryozoan species richness in the North Atlantic. – J. Anim. Ecol. 69: 799–814.

Clauset, A. et al. 2004. Finding community structure in very large networks. – Phys. Rev. E 70: 066111.

Cohen, J. E. et al. 1990. Community food webs: data and theory. – Springer.

Connor, E. F. and Simberloff, D. 1978. Species number and compositional similarity of the Galápagos flora and avifauna. – Ecol. Monogr. 48: 219–248.

Connor, E. F. and Simberloff, D. 1979. The assembly of species communities – chance or competition? – Ecology 60: 1132–1140.

Cornell, H. V. 1985. Species assemblages of Cynipid gall wasps are not saturated. – Am. Nat. 126: 565–569.

Cornell, H. V. and Lawton, J. H. 1992. Species interactions, local and regional processes, and limits to the richness of ecological communities – a theoretical perspective. – J. Anim. Ecol. 61: 1–12.

Costa, L. da F. et al. 2007. Characterization of complex networks: a survey of measurements. – Adv. Phys. 56: 167–242.
Cox, C. B. 2001. The biogeographic regions reconsidered. – J. Biogeogr. 28: 511–523.
Crowe, T. M. and Crowe, A. A. 1982. Patterns of distribution, diversity and endemism in Afrotropical birds. – J. Zool. 198: 417–442.
Danon, L. et al. 2005. Comparing community structure identification. – J. Stat. Mech. 2005: P09008.
Darlington, P. J. 1957. Zoogeography: the geographical distribution of animals. – Wiley.
de Candolle, A. 1820. Essai elementaire de gèographie botanique. – Dictionaire Des Sciences Naturelles 18: 359–422.
Diamond, J. 1982. Effect of species pool size on species occurrence frequencies – musical chairs on islands. – Proc. Natl Acad. Sci. USA 79: 2420–2424.
Diamond, J. M. 1975. Assembly of species communities. – In: Cody, M. L. and Diamond, J. M. (eds), Ecology and evolution of communities. Harvard Univ. Press, pp. 342–444.
Dormann, C. and Strauss, R. 2013. Detecting modules in quantitative bipartite networks: the QuaBiMo algorithm. – arXiv: 1304.3218 [q-bio.QM].
Duch, J. and Arenas, A. 2005. Community detection in complex networks using extremal optimization. – Phys. Rev. E 72: 027104.
Dunne, J. A. et al. 2002. Network structure and biodiversity loss in food webs: robustness increases with connectance. – Ecol. Lett. 5: 558–567.
Elton, C. 1946. Competition and the structure of ecological communities. – J. Anim. Ecol. 15: 54–68.
Faaborg, J. 1979. Qualitative patterns of avian extinction on neotropical land-bridge islands – lessons for conservation. – J. App. Ecol. 16: 99–107.
Fox, B. J. and Brown, J. H. 1993. Assembly rules for functional-groups in North-American desert rodent communities. – Oikos 67: 358–370.
Girvan, M. and Newman, M. E. 2002. Community structure in social and biological networks. – Proc. Natl Acad. Sci. USA 99: 7821–7826.
Gotelli, N. J. 2000. Null model analysis of species co-occurrence patterns. – Ecology 81: 2606–2621.
Gotelli, N. J. and Graves, G. R. 1990. Body size and the occurrence of avian species on land-bridge islands. – J. Biogeogr. 17: 315–325.
Gotelli, N. J. and Graves, G. R. 1996. Null models in ecology. – Smithsonian Inst. Press.
Gotelli, N. J. and McCabe, D. J. 2002. Species co-occurrence: a meta-analysis of J. M. Diamond's assembly rules model. – Ecology 83: 2091–2096.
Grant, P. R. 1966. Ecological compatibility of bird species on islands. – Am. Nat. 100: 451–462.
Grant, P. R. and Abbott, I. 1980. Interspecific competition, island biogeography and null hypotheses. – Evolution 34: 332–341.
Graves, G. R. and Gotelli, N. J. 1983. Neotropical land-bridge avifaunas – new approaches to null hypotheses in biogeography. – Oikos 41: 322–333.
Graves, G. R. and Gotelli, N. J. 1993. Assembly of avian mixed-species flocks in amazonia. – Proc. Natl Acad. Sci. USA 90: 1388–1391.
Graves, G. R. and Rahbek, C. 2005. Source pool geometry and the assembly of continental avifaunas. – Proc. Natl Acad. Sci. USA 102: 7871–7876.
Guimerà, R. and Amaral, L. A. N. 2005a. Cartography of complex networks: modules and universal roles. – J. Stat. Mech. Theory E 2005: P02001.
Guimerà, R. and Amaral, L. A. N. 2005b. Functional cartography of complex metabolic networks. – Nature 433: 895–900.
Guimerà, R. et al. 2007. Module identification in bipartite and directed networks. – Phys. Rev. E 76: 036102–036108.
Harrison, S. and Cornell, H. 2008. Toward a better understanding of the regional causes of local community richness. – Ecol. Lett. 11: 969–979.
Heikinheimo, H. et al. 2007. Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. – J. Biogeogr. 34: 1053–1064.
Holt, B. G et al. 2013. An update of Wallace's zoogeographic regions of the World. – Science 339: 74–78.
How, R. A. and Kitchener, D. J. 1997. Biogeography of Indonesian snakes. – J. Biogeogr. 24: 725–735.
Jordano, P. 1987. Patterns of mutualistic interactions in pollination and seed dispersal – connectance, dependence assymmetries, and coevolution. – Am. Nat. 129: 657–677.
Jordano, P. et al. 2003. Invariant properties in coevolutionary networks of plant–animal interactions. – Ecol. Lett. 6: 69–81.
Kirkpatrick, S. et al. 1983. Optimization by simulated annealing. – Science 220: 671–680.
Kissling, W. D. et al. 2012. Cenozoic imprints on the phylogenetic structure of palm species assemblages worldwide. – Proc. Natl Acad. Sci. USA 109: 7379–7384.
Kraft, N. J. et al. 2011. Disentangling the drivers of β diversity along latitudinal and elevational gradients. – Science 333: 1755–1758.
Kreft, H. and Jetz, W. 2010. A framework for delineating biogeographical regions based on species distributions. – J. Biogeogr. 38: 2029–2059.
Kristiansen, T. et al. 2011. Local and regional palm (Arecaceae) species richness patterns and their cross-scale determinants in the western Amazon. – J. Ecol. 99: 1001–1015.
Lennon, J. J. et al. 2001. The geographical structure of British bird distributions: diversity, spatial turnover and scale. – J. Anim. Ecol. 70: 966–979.
Lessard, J.-P. et al. 2012a. Inferring local ecological processes amid species pool influences. – Trends Ecol. Evol. 27: 600–607.
Lessard, J.-P. et al. 2012b. Strong influence of regional species pools on continent-wide structuring of local communities. – Proc. R. Soc. B 279: 266–274.
Linder, H. P. et al. 2012. The partitioning of Africa: statistically defined biogeographical regions in sub-Saharan Africa. – J. Biogeogr. 39: 1189–1205.
MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. – In: Le Cam, L. M. and Neyman, J. (eds), Fifth Berkeley Symposium on Mathematical Statistics and Probability. Univ. of Calif. Press, pp. 281–297.
May, R. M. 1973. Stability and complexity in model ecosystems. – Princeton Univ. Press.
Morrone, J. J. 2002. Biogeographical regions under track and cladistic scrutiny. – J. Biogeogr. 29: 149–152.
Morrone, J. J. 2009. Evolutionary biogeography: an integrative approach with case studies. – Columbia Univ. Press.
Newman, M. E. J. 2004. Analysis of weighted networks. – Phys. Rev. E 70: 056131.
Newman, M. E. J. 2006a. Finding community structure in networks using the eigenvectors of matrices. – Phys. Rev. E 74: 036104.
Newman, M. E. J. 2006b. Modularity and community structure in networks. – Proc. Natl Acad. Sci. USA 103: 8577–8582.
Newman, M. E. J. and Girvan, M. 2004. Finding and evaluating community structure in networks. – Phys. Rev. E 69: 026113.
Olesen, J. M. et al. 2007. The modularity of pollination networks. – Proc. Natl Acad. Sci. USA 104: 19891–19896.
Paine, R. T. 1980. Food webs: linkage, interaction strength and community infrastructure. – J. Anim. Ecol. 49: 666–685.

Palmgren, A. 1925. Die Artenzahl als pflanzengeographischer Charakter sowie der Zufall und die sekuläre Landhebung als pflanzengeographische Faktoren. Ein pflanzengeographischer Entwurf, basiert auf Material aus dem ålændischen Schärenarchipel. – Acta Bot. Fenn. 1: 1–143.

Patrick, R. 1967. Effect of invasion rate, species pool, and size of area on the structure of the diatom community. – Proc. Natl Acad. Sci. USA 58: 1335–1342.

Penner, J. et al. 2011. A hotspot revisited – a biogeographical analysis of west African amphibians. – Divers. Distrib. 17: 1077–1088.

Pons, P. and Latapy, M. 2006. Computing communities in large networks using random walks. – J. Graph Algorithms Appl. 10: 191–218.

Procheş, Ş. 2005. The world's biogeographical regions: cluster analyses based on bat distributions. – J. Biogeogr. 32: 607–614.

Procheş, Ş. and Ramdhani, S. 2012. The World's zoogeographic regions confirmed by cross-taxon analyses. – Bioscience 62: 260–270.

REEF 2011. Reef Environmental Education Foundation. – <www.reef.org>.

Ricklefs, R. E. 1987. Community diversity – relative roles of local and regional processes. – Science 235: 167–171.

Ricklefs, R. E. 2000. The relationship between local and regional species richness in birds of the Caribbean Basin. – J. Anim. Ecol. 69: 1111–1116.

Ricklefs, R. E. 2007. History and diversity: explorations at the intersection of ecology and evolution. – Am. Nat. 170: S56–S70.

Ricklefs, R. E. and Schluter, D. 1993. Species diversity in ecological communities: historical and geographical perspectives. – Univ. of Chicago Press.

Rosvall, M. and Bergstrom, C. T. 2007. An information-theoretic framework for resolving community structure in complex networks. – Proc. Natl Acad. Sci. USA 104: 7327–7331.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. – J. Comput. Appl. Math. 20: 53–65.

Rueda, M. et al. 2010. Towards a biogeographic regionalization of the European biota. – J. Biogeogr. 37: 2067–2076.

Salvador, S. and Chan, P. 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. – In: 16th IEEE International Conference on Tools with Artificial Intelligence, pp. 576–584.

Sanders, N. J. et al. 2007. Assembly rules of ground-foraging ant assemblages are contingent on disturbance, habitat and spatial scale. – J. Biogeogr. 34: 1632–1641.

Schoener, T. W. 1988. Testing for non-randomness in sizes and habitats of West Indian lizards: choice of species pool affects conclusions from null models. – Evol. Ecol. 2: 1–26.

Sclater, P. L. 1858. On the general geographical distribution of the members of the class Aves. – J. Proc. Linn. Soc. 2: 130–145.

Shurin, J. B. et al. 2000. Local and regional zooplankton species richness: a scale-independent test for saturation. – Ecology 81: 3062–3073.

Simberloff, D. 1970. Taxonomic diversity of island biotas. – Evolution 24: 23–47.

Simberloff, D. S. and Connor, E. F. 1978. Q-mode and R-mode analyses of biogeographic distributions:null hypotheses based on random colonization. – In: Patil, G. P. and Rosenweig, M. (eds), Contemporary quantitative ecology and related ecometrics. Int. Cooperative Publ. House, pp. 123–138.

Simpson, G. G. 1953. Evolution and geography. – Oregon System of Higher Education, Eugene.

Simpson, G. G. 1977. Too many lines; the limits of the Oriental and Australian zoogeographic regions. – Proc. Am. Phil. Soc. 121: 107–120.

Smith, C. H. 1983. A system of world mammal faunal regions I. Logical and statistical derivation of the regions. – J. Biogeogr. 10: 455–466.

Sokal, R. R. and Michener, C. D. 1958. A statistical method for evaluating systematic relationships. – Univ. Kans. Sci. Bull. 28: 1409–1438.

Sokal, R. R. and Rohlf, J. 1962. The comparison of dendrograms by objective methods. – Taxon 11: 33–40.

Srivastava, D. S. 1999. Using local–regional richness plots to test for species saturation: pitfalls and potentials. – J. Anim. Ecol. 68: 1–16.

Stevens, R. D. 2011. Relative effects of time for speciation and tropical niche conservatism on the latitudinal diversity gradient of phyllostomid bats. – Proc. R. Soc. B 278: 2528–2536.

Stone, L. et al.1996. Community-wide assembly patterns unmasked: the importance of species' differing geographical ranges. – Am. Nat. 148: 997–1015.

Strogatz, S. H. 2001. Exploring complex networks. – Nature 410: 268–276.

Strong, D. R. et al. 1979. Tests of community-wide character displacement against null hypotheses. – Evolution 33: 897–913.

Swenson, N. G. et al. 2006. The problem and promise of scale dependency in community phylogenetics. – Ecology 87: 2418–2424.

Terborgh, J. W. and Faaborg, J. 1980. Saturation of bird communities in the West-Indies. – Am. Nat. 116: 178–195.

Thébault, E. 2013. Identifying compartments in presence–absence matrices and bipartite networks: insights into modularity measures. – J. Biogeogr. 40: 759–768.

Vasconcelos, T. da S. et al. 2011. Biogeographic distribution patterns of south american amphibians: a regionalization based on cluster analysis. – Natureza Conservação 9: 67–72.

von Humboldt, A. 1806. Essai sur la geographie des plantes; accompagné d'un tableau physique des régions équinoxales, accompagné d'un tableau physique des régions e´quinoctiales. – Schoel and Co., Paris.

Wallace, A. R. 1876. The geographical distribution of animals. – Macmillan.

Weiher, E. and Keddy, P. A. 1995. Assembly rules, null models, and trait dispersion – new questions front old patterns. – Oikos 74: 159–164.

White, E. P. and Hurlbert, A. H. 2010. The combined influence of the local environment and regional enrichment on bird species richness. – Am. Nat. 175: E35–E43.

Wiens, J. J. et al. 2011. Phylogenetic origins of local-scale diversity patterns and the causes of Amazonian megadiversity. – Ecol. Lett. 14: 643–652.

Williams, C. B. 1947. The generic relations of species in small ecological communities. – J. Anim. Ecol. 16: 11–18.

Witman, J. D. et al. 2004. The relationship between regional and local species diversity in marine benthic communities: a global perspective. – Proc. Natl Acad. Sci. USA 101: 15664–15669.

Zobel, M. 1997. The relative of species pools in determining plant species richness: an alternative explanation of species coexistence? – Trends Ecol. Evol. 12: 266–269.

Zobel, M. et al. 2011. The formation of species pools: historical habitat abundance affects current local diversity. – Global Ecol. Biogeogr. 20: 251–259.