Short title: FEASIBLE SETS OF ABUNDANCE DISTRIBUTIONS

Article title: How species richness and total abundance constrain the distribution of abundance

Kenneth J. Locey[‡1] and Ethan P. White[*2]

Department of Biology, Utah State University, Logan, UT, 84322

*Ecology Center, Utah State University, Logan, UT, 84322

‡To whom correspondence should be addressed

[1] ken@weecology.org; phone: (435) 764-5070, fax (435) 797-1575

[2] ethan@weecology.org;

Statement of authorship: KL conceived the idea of using feasible sets and integer partitioning to examine SADs. KL and EW designed the analyses. KL developed the project code, acquired and managed the microbial data, and conducted the analyses. EW acquired and managed the macroorganismal data, and helped with project code development. KL and EW wrote the manuscript.

Keywords: constraints, distribution of wealth, feasible set, hollow-curve, species abundance distribution, macroecology, MaxEnt

Article type: Letters

words in the abstract: 150

words in the main text: 4,999

number of cited references: 50

number of figures: 5

Abstract. The species abundance distribution (SAD) is one of the most intensively studied distributions in ecology and its hollow-curve shape is one of ecology's most general patterns. We examine the SAD in the context of all possible forms having the same richness (S) and total abundance (N), i.e. the feasible set. We find that feasible sets are dominated by similarly-shaped hollow-curves, most of which are highly correlated with empirical SADs (most R^2 values > 75%), revealing a strong influence of N and S on the form of the SAD and an *a priori* explanation for the ubiquitous hollow-curve. Empirical SADs are often more hollow and less variable than the majority of the feasible set, revealing exceptional unevenness and relatively low natural variability among ecological communities. We discuss the importance of the feasible set in understanding how general constraints determine observable variation and influence the forms of predicted and empirical patterns.

INTRODUCTION

The species abundance distribution (SAD) is one of the most widely studied patterns in ecology and exhibits a consistent structure with many rare and few common species; the canonical "hollow curve" (McGill et al. 2007). The form of the SAD has been predicted by a variety of models based on an array of different processes including niche differentiation (e.g. Sugihara 1980), stochastic population dynamics (e.g. Hubbell 2001), and the structure of abundance across a species range (e.g. McGill & Collins 2003). Though SADs are potentially influenced by some or all of these processes, the ability to distinguish between different structuring processes depends on the presence of sufficient variation among the possible shapes of the SAD (Haegeman and Loreau 2008). If most of the possible SADs have similar shapes, it will be difficult to determine what processes generated them.

Haegeman and Loreau (2008) introduced the use of the set of all possible distributions (the feasible set) to examine ecological patterns and theory. They argue that if the feasible set is small then there is little information in the pattern being examined. Likewise, if theoretical predictions do not deviate from the center of the feasible set, then they may provide limited information about process. To explore the implications of these ideas for understanding the species abundance distribution we use McGill et al.'s (2007) definition of the SAD as the "vector of abundances of all species present in a community". This distribution is necessarily influenced by two values: total abundance (N; i.e. the number of individuals in a community) and species richness (S; i.e. the number of species in a community). Though ecological theories often use N and S as inputs to fit or predict the shape of the SAD (McGill 2010), knowing N and S constrains the form of the SAD in ways that ecologists rarely address. Specifically, there are a limited number of ways that the abundances of S species can sum to a total abundance of N, and thus,

there is a limited feasible set of uniquely-shaped SADs for any combination of N and S. For example, there is only one possible SAD form when S = 1 and N = 1 (i.e. {1}) and only two possible forms if S = 2 and N = 5 (i.e. {4,1}, {3,2}). As we show, N and S not only determine the number of possible SADs but also the general form of the possible distributions, making it necessary to understand how the properties of the feasible set constrain the form of the SAD and how constraints such as N and S influence empirical patterns and the predictions of ecological models.

We refer to each uniquely-shaped SAD within the feasible set as a macrostate (i.e. an unordered vector of unlabeled species abundances). This differs from a microstate, which refers to a unique distribution of individuals among species leading to a specific macrostate. The terms feasible set, macrostate, and microstate have been used in recent applications of entropy maximization (MaxEnt) to macroecology (Haegeman and Loreau 2008, McGill & Nekola 2010; Harte 2011). In short, MaxEnt infers the most likely macrostate as that with the most microstates based on sets of state variables (e.g. N, S) and related constraints. Though the framework of MaxEnt implies the existence of a feasible set, MaxEnt does not explicitly consider it. Here, we focus solely on the distribution of macrostates within the feasible set without considering the numbers of ways in which macrostates can arise. As we show, feasible sets have strong central tendencies, meaning that most of the possible macrostates have similar shapes. If empirical SADs have shapes similar to those near the center of the feasible set, then there may be little ecological information in the shape of the SAD beyond that contained in N and S. Since most of the observable forms of the abundance distribution have shapes that are very similar to this central tendency, many different processes will result in distributions of the same general form as will many different models. This observation goes beyond the issue of equivalent models (e.g.,

Pielou 1975, McGill et al. 2007), suggesting that many different models and empirical patterns may be expected to take similar forms because most possible states of the pattern are similar in shape (White et al. 2012). However, if the shape of an SAD is exceptional to the majority of shapes with the same N and S, then this exceptional evenness or unevenness would require an explanation, especially if consistent across communities. Consequently, the feasible set provides a context for understanding whether predicted and empirical patterns are exceptional to or representative of the majority of possible forms. As such, studies of the SAD would benefit from considering the shape of the SAD relative to the feasible set, rather than the shape of the SAD *per se*.

Here, we explore general properties of the feasible set and reveal the strength of the influence of N and S on the shape of the SAD. We use the feasible set as a contextual framework for understanding how richness and abundance necessarily constrain ecological patterns. We show that most of the possible SAD shapes are similarly-shaped hollow-curves, revealing an *a priori* reason for the ubiquitous hollow-curve. Using one of the most taxonomically diverse and geographically expansive data compilations in community ecology, we show that the central tendency of the feasible set is strongly correlated with empirical SAD patterns within and among sites for birds, mammals, trees, and metagenomic datasets of prokaryotes and fungi. Moving beyond single SADs, we use ensembles of SADs with the same values of N and S to assess relationships between the variance predicted by the feasible set and that observed in ecological systems. We discuss the importance of using the feasible set as a context for understanding variation in the forms of empirical patterns and the inference that can be drawn from models that successfully predict them.

METHODS

Finding macrostates of feasible sets

Finding all possible macrostates for a community of a particular total abundance (N) and species richness (S) is equivalent to finding all unordered ways of summing S positive integers to obtain the positive integer N, a combinatorial approach known as integer partitioning (Andrews & Eriksson 2004). For example, the feasible set for N = 10 and S = 3 is {8+1+1, 7+2+1, 6+3+1, 6+2+2, 5+4+1, 5+3+2, 4+4+2, 4+3+3}. Different unordered sets of S integers that sum to N are partitions (i.e. macrostates) of N and S. Hence, sets of the same integers in different order, e.g. {8+1+1, 1+8+1, 1+1+8}, constitute the same partition. Likewise, each would produce the same frequency distribution (i.e. two 1's, one 8) and the same rank distribution (i.e. 8+1+1). Several algorithms are available for integer partitioning problems, such as finding the size of the feasible set for a given N and S (Nijenhuis & Wilf 1978). We used the implementation of these algorithms in the free open-source Python-based Sage computer algebra system (http://www.sagemath.org/).

Generating the feasible set for a community of a particular N and S can require a large amount of time and computational memory. This is because feasible sets become large for communities of realistic size; a result of combinatorial explosion (i.e. large changes in the number of possible outcomes for small changes in the values of inputs). For example, there are nearly 8.8×10^{14} macrostates for N = 1,000 and S = 10 and nearly 6.28×10^{26} macrostates for N = 1,000 and S = 50. While complete enumeration of the feasible set can be untenable for many values of N and S, the form of the feasible set space can be determined by randomly sampling macrostates from the feasible set. We used the random partition algorithm described by

Nijenhuis and Wilf (1978) and implemented in Sage to generate uniform random samples of feasible sets.

The partitioning algorithm we used, and all currently implemented integer partitioning algorithms, generates random partitions (i.e. macrostates) based on N but not S. We randomly drew partitions of N and rejected partitions that did not have S parts; an approach that can be computationally expensive. For example, randomly drawing one macrostate for N = 1,000 and S = 10 requires drawing from a feasible set of nearly 2.4×10^{31} macrostates, one of the roughly 8.9×10^{14} for which S = 10; a probability of nearly 3.7×10^{-17}. Consequently, we used substantial computational resources (one in-house cluster of three dual Quad Core Intel Xeon 3 GHz processors with 16 GB of RAM each, plus 20 High-CPU Extra Large Amazon Web Service instances with 7GB of RAM each and a total of 160 AWS cores) and computational time (> 10,000 compute hours) to generate random macrostates for combinations of N and S. Code for replicating our analyses are available at https://github.com/weecology/feasiblesets. All software required to run our scripts (e.g. Sage, Numpy, Python) is free and open source.

We chose the integer partitioning approach to the feasible set over the random walk method used by Haegeman and Loreau (2008) because it is conceptually simpler and, by definition, yields uniform random samples of the feasible set without requiring decisions regarding burn-in periods and the number of steps between samples. However, this approach can be very slow for some combinations of N and S, and further research comparing the speed and accuracy of these two approaches would be valuable.

Data

We used a subset of previously compiled datasets of site-specific species abundance data (see White et al. 2012). Our subset represents 9,562 different sites of bird, tree, and mammal communities. The data set includes four continental-to-global scale surveys, including the Christmas Bird Count (129 sites) (CBC; National Audubon Society 2002), North American Breeding Bird Survey (1,586 sites) (BBS; Sauer et al. 2011), Gentry's Forest Transect Data Set (182 sites) (GENTRY; Phillips and Miller 2002), Forest Inventory Analysis (7,359 sites) (FIA; U.S. Department of Agriculture 2010), and one global-scale data compilation, the Mammal Community Database (42 sites) (MCDB; Thibault et al. 2011). White et al. (2012) used one year of sampling for each site and only used data for communities with a minimum of 10 species (Ulrich et al. 2010). We included only sites with combinations of N and S for which random macrostates could be generated based on reasonable computational effort. This includes large fractions of all of the datasets except for CBC, which only includes ~6% of the original sites. Additionally, we restricted our analysis of FIA to natural forest stands (e.g. absence of human disturbance, plots without artificial regeneration, plots without silvicuture treatment). More details regarding the data can be found in Appendix S1 of the supporting information for this manuscript and in Appendix A of White et al. (2012).

We also compiled relative abundance data at the species level from five microbial metagenome projects for a total of 264 surveys of geographically distinct bacterial, archaeal, and indoor fungal communities. Metagenomes are produced from genetic material recovered from environmental samples and are the primary means of studying microbial diversity *in situ*. Despite the lack of a universally accepted microbial species definition, there is a well-established convention for demarcating species-level units. Taxonomic levels representing species are commonly delineated at 97% 16S rRNA sequence similarity for prokaryotes and 97% rRNA

sequence and rRNA related ITS (internal transcribed spacer) sequence similarity for fungi (Roselló-Mora and Amann 2001, Schloss and Handelsman 2006, Marshal et al. 2008, Amend et al 2010, Chu et al. 2010, Flores et al. 2011, Fierer et al 2012). This convention was used by studies that generated the metagenomic data used in our study.

We used SAD data from region-to-global scale PCR-targeted projects from the metagenomics server MG-RAST (Meyer et al. 2008). PCR-targeted (i.e. amplicon sequenced) approaches provide better overall coverage of a specific gene (e.g. 16S rRNA) than a random shotgun approach by sequencing an amplified target gene. We used the rRNA library provided by MG-RAST (i.e. M5RNA) to obtain SAD data for each metagenome used in our study. We used common thresholds for sequence comparison and species-level determination (Lazarevic et al. 2009; Lamendella et al. 2011) including a maximum e-value (probability of observing an equal or better match in a database of a given size) cutoff of $1e^{-5}$, a minimum alignment length of 50 base pairs, and a minimum percent identity of 97% to the M5RNA reference sequence. However, because microbial species are sometimes defined below or above 97% (e.g. Webster et al. 2010, Martiny et al. 2011) we also analyzed microbial communities at 95% and 99% species-level cutoffs.

We compiled metagenomic data into datasets representing aquatic prokaryotic communities (48 metagenomes), terrestrial prokaryotic communities (92 metagenomes), and terrestrial fungal communities (124 metagenomes). The aquatic datasets (AQUA) included the Archaeal and Bacterial Diversity of Geographically and Geologically Distinct Deep-Sea Hydrothermal Vent Mineral Deposits project (Flores et al. 2011) and the Catlin Arctic Survey of bacterial and archaeal diversity (www.catlin.com/en/Responsibility/CatlinArcticSurvey). The terrestrial prokaryotic datasets (TERA) included the archaeal and bacterial diversity of the

Lauber 88 Soils project (Fierer et al. 2012), and the Chu Arctic Soils project (Chu et al. 2010). The terrestrial fungi dataset was a global-scale survey of fungal community data sampled from indoor habitats of human cities (Amend et al. 2010). Detailed information about each metagenome project is available on the MG-RAST website (http://metagenomics.anl.gov/) and additional details on our use of microbial metagenomes is available in Appendix S1.

Our compilation of data is taxonomically diverse. As such, there are differences among our datasets that should be recognized. First, despite our use of accepted species level delineations for microbes, these species and communities do not represent the same ecological and evolutionarily meaningful units as our other datasets, i.e., genetically distinct populations of biological species. Whereas our macrobial data represent a few well-known members of one domain (i.e. Eukaryota), our microbial datasets include many poorly understood members from all three. Second, whereas abundances in macrobial datasets were reported as counts of individuals, taxonomic abundance and identification of microbes in natural environments is commonly derived from DNA harvested from environmental samples; individual counts are not practical. Third, among macrobial datasets there are large differences in how communities were sampled (e.g. plot counts of trees, transects for breeding birds, multiple trapping/sampling methods for mammals) (see Appendix S1).

Form of SADs in the Feasible Set

The canonical, hollow-curve, form of the species abundance distribution includes large numbers of rare species and small numbers of abundant species, leading to frequency distributions with the mode at small values of N and long, right-skewed, tails. To determine if this form is common in the set of possible SADs we analyzed the distribution of modal

abundance class, species evenness, and skewness within the feasible set for a variety of N and S combinations and N/S ratios. We avoided extremely large values of S because values of S close to N are uncommon in nature and constrain the SAD to a nearly even vector of singletons. We used uniform random samples of 500 macrostates for each N-S combination in this analysis. These numbers are large enough to characterize the general form of the feasible set and small enough to permit doing so in reasonable time (Fig. 1 of Appendix S2).

Comparing Observed Data to Central Tendencies of Feasible Sets

We determined which SAD represented the center of each feasible set by generating 300 to 500 random macrostates from the feasible set (generating 500 random macrostates for some combinations of N and S was untenable). Random samples of 300, 500, and 700 macrostates produce equivalent results (Fig 1. of Appendix S2). We chose the macrostate that overlapped the most on average with other random macrostates across the S ranked abundances. In the case of a tie, we favored the macrostate having the more evenly distributed overlap across ranked abundances (i.e. the macrostate with the smaller variance in overlap with other macrostates). This yielded SADs that were centered within the densest regions of random samples (Fig. 2 of Appendix S2), and hence, within the central tendency of the feasible set. We compared this central SAD for each community to the observed SAD using rank-abundance distributions (RADs). Specifically we compared the observed value of abundance at each rank (most abundant to least abundant) at each site to the abundance at that same rank from the SAD representing the central tendency of the feasible set. We used log-transformed values of abundance at each rank (not log-transformed bins; see Nekola et al. 2008) to make visual comparisons and calculate R^2 values following Marks & Muller-Landau (2007) to avoid overweighting rare species, to address

heteroscedasticity, and because we are generally more interested in proportional differences in abundance within a rank rather than absolute differences.

RESULTS

The majority of possible SAD shapes exhibit the classic hollow-curve form with modes at low abundance classes and positive skewness, revealing an overall hollow-curve shape for most of the macrostates in the feasible set (Fig. 1, see also Fig 3 of Appendix S2). The specific form of the distribution is influenced by the values of N, S, and average species abundance (i.e. N/S), which are associated with modal abundance, species evenness, and skewness of the SAD (Fig 2). This means that differences in community structure among sites (or directional changes along gradients) could result from the constraining influence of N and S. For realistic values of average abundance, the portion of highly uneven macrostates in the feasible set will increase as N is partitioned across a greater number of species. However, as average abundance approaches 1.0, the SAD must necessarily become highly even.

Observed ranked abundances were often similar to those near the central tendency of the feasible set, both within and across sites for trees, animals, and microorganisms (Fig 3). The SAD at the central tendency of the feasible set consistently explained the majority of variation in observed abundance distributions both within sites and among entire datasets (R^2: BBS = 0.93; CBC = 0.77; FIA = 0.84; GENTRY = 0.81; MCDB = 0.78; TERA = 0.83; AQUA = 0.58; FUNGI = 0.76; R^2 values are with respect to the central tendency, not a fitted relationship). However, clear deviations from the form of the central tendency did occur and were strongest among microbial metagenomes where the central tendency of the feasible set contained lower abundances for dominant species and higher abundances for rare species than the observed

communities. We observed the same pattern for microbes regardless of whether species were defined at 95, 97, or 99% (Fig. 4 of Appendix S2).

Because many of the possible SADs are similar, the similarity between the center of the feasible set and the observed data means that the shapes of observed SADs tend to look very similar to the majority of possible shapes, suggesting a strong influence of the limits of observable variation on natural variation. Evaluating the correlation between the observed SAD and all random macrostates shows that randomly choosing a macrostate will often produce a distribution that is well correlated with observed data (Fig 4). This was most obvious for BBS, where the majority of randomly sampled macrostates explained more than 80% of observed variation in abundance for nearly all sites.

DISCUSSION

The hollow-curve SAD has been referred to as an ecological law and is thought to be universal across taxa (McGill et al. 2007). This pattern is also observed in non-biological systems (Gaston et al. 1993; Nekola & Brown 2007; Warren et al. 2011) suggesting that the unevenness and ubiquity of the hollow-curve SAD might be explained by emergent statistical phenomena rather than specific biological processes (Šizling et al. 2009b; McGill 2010; White et al. 2012, Yen et al. 2012). Here, we have described the first attempt to understand the shape of the SAD in terms of the set of all possible shapes given two general constraints that are commonly used as inputs in ecological theory. The majority of feasible SADs share similar forms that, like observed SADs, resemble a hollow curve frequency distribution. As such, the feasible set provides an *a priori* reason for the ubiquity of the hollow-curve and a reason why many different models tend to produce the same general SAD form. Examination of over 9,000

communities shows that observed SADs are often similar to the central tendency of the feasible set and, because most macrostates are clustered near the central tendency, the majority of possible distributions often explain substantial portions of variation in observed abundances.

While much of the variation in empirical SADs is characterized by the center of the feasible set, SADs are often more uneven than the central tendency. SADs for microbial communities were almost always exceptionally uneven, regardless of whether species were delineated at 95, 97, or 99% sequence similarity. Though hollow-curve SADs have been widely documented for microbes and macrobes, our examination reveals that the structure of microbial communities, with respect to the influence of N and S, may differ from that of macrobes. Indeed, microbial communities are known for their large rare portions (i.e. rare biosphere). However, it has also been suggested that the exceptional unevenness of microbial SADs may result from detection issues related to metagenomic methods that can exaggerate dominance and rarity (Woodcock et al. 2006). Observational/sampling biases are also a potential issue for the macrobial datasets (e.g., MacKenzie & Kendall 2002) and therefore have the potential to play a role in deviations from the feasible set in those analyses as well.

Empirical data can also be compared to the feasible set by comparing distributions of a statistical property (e.g. species evenness) across the feasible set. This allows the values for individual communities to be placed within context. For example, a community with a value of species evenness in the 50^{th} percentile of the feasible set, i.e. near the central tendency and the majority of possible macrostates, would not have an exceptionally even or uneven distribution of abundance, regardless of whether the value of evenness itself is large or small.(Fig 5 Appendix S2). This is particularly important when comparing sites that differ in N and S, since differences in evenness can be expected based purely on differences in the feasible set (Figure 2, Fig 5

Appendix S2). Consequently, comparisons of species evenness that do not account for the feasible set are primarily comparisons of N and S.

In addition to contextualizing single communities, ensembles of sites with shared values of N and S can be used to compare distributions of a property across communities to the distribution of that property in the feasible set. Conducting this analysis using FIA data and species evenness (E_{var}; Smith and Wilson 1996) reveals that, while the modal values of E_{var} for feasible sets and FIA sites were often similar, the distribution of E_{var} across the feasible set was broader than that of empirical distributions (Fig. 5).This relatively low natural variability could indicate that interactions between ecological processes and statistical phenomena prevent the extreme values of evenness that are otherwise possible. Additionally, the tendency for the distribution of empirical E_{var} values to be concentrated at lower or higher values of E_{var} was related to average abundance (i.e. N/S), with higher N/S leading to lower E_{var} (i.e. lower evenness) for both empirical SADs and the feasible set. While the general decrease in species evenness with average abundance can be explained by the feasible set (Fig. 2), the actual change in empirical E_{var} outpaced that of the feasible set (Fig 5), suggesting that mechanisms leading to unevenness may strengthen as N/S increases (e.g. via positive frequency dependence), but not so much that the lowest possible range of species evenness is attained.

While the feasible set reveals that a small number of community-related constraints may explain the general shape of the SAD by limiting observable variation, it also demonstrates that in some cases empirical patterns deviate directionally from the majority of possible states (Figure 3) and are more tightly clustered than expected (Figure 5). Consequently, the ecological interactions of individuals, populations, and species may be needed to explain the specific form of ecological patterns as well as the frequent occurrence of exceptionally uneven SADs and the

rare occurrence of exceptionally even ones. High degrees of competition and dispersal limitation, and low degrees of invasiveness may all lead to the degrees of excessive dominance or unevenness that are commonly observed among microorganisms and macroorganisms and which cannot be attributed to the constraining influences of N and S. However, without the feasible set it would not possible to recognize that this degree of unevenness and its relatively low natural variability are exceptional.

The feasible set approach focuses on the observable variation among the possible forms of a pattern of interest (i.e. macrostates). In a way, it assumes that all possible forms of the SAD are equally likely because it assumes nothing about the ways in which each macrostate may arise. However, by accounting for the ways in which macrostates can arise through microstate configurations, approaches like MaxEnt (Pueyo et al. 2007; Harte et al. 2008; Frank 2011) produce a most likely form that may better explain the general shape of the SAD. Indeed, 4 out of the 5 datasets shared with White et al. (2012) are at least somewhat better fit by the predictions of the MaxEnt model of Harte (2011); the exception being BBS. This comparison is approximate because we worked with subsets of the datasets in White et al. (2012) and because the model of Harte (2011) requires additional assumptions to be made beyond fixing N and S.

The idea that empirical SADs may be more similar to the form with the greatest number of microstates than to the form closest to the center of the feasible set is complicated by the fact that MaxEnt yields different predictions depending on the specific approach to the problem (Haegeman & Etienne 2010). In cases where the number of constraints is small, it is unlikely that the most likely macrostate from one of the several MaxEnt approaches will occur at the center of the feasible set. In fact, Haegeman and Loreau (2008) consider differences between MaxEnt predictions and the center of the feasible set to be a necessary condition for applications of

MaxEnt to be considered non-trivial. This presents an interesting philosophical question: should we try to understand patterns in the context of their distribution of macrostates alone, in the context of these macrostates weighted by the number of microstates, or some combination of the two. Current microstate-based approaches do not explicitly consider the properties of either the feasible set or the full set of microstates, only a single most likely macrostate. This prevents existing MaxEnt approaches from providing a general context for how extreme an abundance distribution is relative to the most likely macrostate, though this can probably be addressed through sampling approaches to randomly select microstates. Further research is needed to compare and understand the relationships between these microstate and macrostate based-approaches, to form a more comprehensive understanding of how to contextualize empirical patterns and theoretical predictions.

Another area for additional research is understanding what functional forms (e.g. log-series, log-normal, etc.; McGill et al. 2007) are most common in the feasible set and whether the most common forms change as a function of S and N. This would provide information useful for comparing the quality of distributional fits to empirical data. It would also provide context for one of the current challenges for theoretical models of macroecological patterns - making predictions that are valid at multiple taxonomic and spatial scales (e.g, Šizling et al. 2009a). In contrast to most theoretical models, it is possible that the form of the central tendency changes with N, S, and N/S, becoming more or less similar to different standard distributions (e.g. log-series, log-normal). Knowing how the feasible set responds to changes in N (e.g. with sample size and area) and S (e.g. different taxonomic levels) could enlighten the discussion of whether universal forms of macroecological patterns exist (Šizling et al. 2009a). It has been suggested that, as N approaches infinity and as S changes as a function of N (i.e. $S = cN^{1/2}$), there is a limit

shape to random integer partitions (Vershik and Yakubovich 2001). Further studies are needed to explore whether this is the case for combinations of N and S observed in natural systems, and whether this limit shape is similar to known distributions.

The feasible set approach is part of an emerging area of ecology that uses constraint or state-variable-based approaches to understand ecological patterns (Shipley et al. 2006; Pueyo 2007; Haegman and Loreau 2008; McGill 2010; Harte 2011). These approaches take a top down perspective on understanding ecological patterns, suggesting that much of the information contained in a distribution can be captured by a small number of constraints. This approach to understanding ecological patterns has received empirical support from observational (Harte 2011; White et al. 2012) and experimental (Supp et al. 2012) studies. However, even when these approaches successfully characterize empirical patterns, they do not indicate whether ecological processes are not operating. Instead ecological processes may influence emergent patterns indirectly through their influence on constraints or state variables (White et al. 2012; Supp et al. in 2012). These constraint-based approaches reinforce the fact that ecological processes operate within but also influence constraints that necessarily determine a set of possible outcomes.

The feasible set represents a new perspective in understanding empirical patterns. This approach is potentially applicable to many other widely-known distributions in ecology and other areas of science. In particular, the SAD is a specific type of distribution of wealth and uneven distributions of wealth are widespread in social, economic, and physical systems (Zipf 1949, Gaston et al. 1993, Reed 2001, Nekola and Brown 2007). The feasible set approach should be applicable to distributions of wealth and abundance that are characterized by the partitioning of a total quantity (e.g. individuals, species, dollars, hectares) among a number of classes (e.g. species, islands, socioeconomic classes, countries). This includes classic ecological patterns,

such as the species-area relationship and species-time relationship, and emerging patterns in microbial ecology such as distribution of functional traits, as well as distributions of wealth, size, and abundance among human populations.

ACKNOWLEDGMENTS

We thank the numerous individuals involved in collecting and providing the data used in this paper including the essential citizen scientists who collect the BBS and CBC data, USGS and CWS scientists and managers, researchers who collected and sequenced the CHU, LAUB, HYDRO, CATLIN, and FUNGI metagenomes, the MG-RAST project, the Ribosome Database Project, the Audubon Society, the U.S. Forest Service, the Missouri Botanical Garden, and Alwyn H. Gentry. We thank X. Xiao, D. McGlinn, B. Burnside, J. Kitzes, J. Parnell, James O'Dwyer and an anonymous reviewer for fruitful discussions and critical comments. This research was supported by a CAREER grant from the U.S. National Science Foundation to EPW (DEB-0953694) and by a research grant from Amazon Web Services to EPW and KJL.

REFERENCES

1.

Amend, A.S., Seifert, K.A., Samson, R., & Bruns, T.D. (2010). Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *P. Natl. Acad. Sci. USA*., 107, 13748-13753.

2.

Andrews, G.E. & Eriksson, K. (2004). *Integer Partitions*. Cambridge Univ. Press, New York.

3.

Brown, J.H. (1995). *Macroecology*. Univ. Chicago Press, Chicago.

4.

Chu, H., Fierer, N., Lauber, C.L., Caporaso, J.G., Knight, R. & Grogan, P. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ. Microbiol.*, 12,2998–3006.

5.

Fierer, N., Lauber, C.L., Ramirez, K.S., Zaneveld, J., Bradford, M.A. & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.*, 6, 1007–17.

6.

Flores, G.E., Campbell, J., Kirshtein, J., Meneghin, J., Podar, M., Steinberg, J.I. *et al*. (2011). Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environ. Microbiol.*, 13, 2158-2171.

7.

Frank, S.A. (2011). Measurement scale in maximum entropy models of species abundance. *J. Evolution. Biol.*, 24, 485-496.

8.

Gaston, K.J., Blackburn, T.M. & Lawton, J.H. (1993). Comparing animals and automobiles: a vehicle for understanding body size and abundance relationships in species assemblages? *Oikos*, 66, 172-179.

9.

Haegeman, B. & Etienne, R.S. (2010). Entropy Maximization and the Spatial Distribution of Species. *Am. Nat.*, 175, E74-E90.

10.

Haegeman, B. & Loreau, M. (2008). Limitations of entropy maximization in ecology. *Oikos*, 117, 1700-1710.

11.

Harte, J. (2011). *Maximum entropy and ecology*. Oxford Univ. Press, Oxford.

12.

Harte, J., Zillio, T., Conlisk, E., & Smith, A.B. (2008). Maximum entropy and the state-variable approach to macroecology. *Ecology*, 89, 2700-2711.

13

Hubbell, S.P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton Univ. Press, Princeton.

14.

Lamendella, R., Domingo, J.W.S., Ghosh, S., Martinson, J. & Oerther, D.B. (2011). Comparative fecal metagenomics unveils unique functional capacity of the swine gut. *BMC microbiol.*, 11, 103.

15.

Lazarevic, V., Whiteson, K., Hernendez, D., François, P. & Schrenzel, J. (2009). Study of inter- and intra-individual variations in the salivary microbiota. *BMC Genomics*, 11, 523.

16.

MacKenzie, D. I., & Kendall, W. L. (2002). How should detection probability be incorporated into estimates of relative abundance?. *Ecology*, *83*, 2387-2393.

17.

Marks, C.O. and Muller-Landau, H.C. (2007). Comment on “from plant traits to plant communities: A statistical mechanistic approach to biodiversity”. Science, 316, 5830.

18.

Marshal, M.M., Amos, R.N., Henrich, V.C., & Rublee, P.A. (2008). Developing SSU rDNA metagenomic profiles of aquatic microbial communities for environmental assessments. *Ecol. Indic.*, 8, 442-453.

19.

Martiny, J.B.H, Eisen, J.A., Penn, K., Allison, S.D., & Horner-Devine M.C. (2011). Drivers of bacterial β-diversiy depend on spatial scale. *P. Natl. Acad. Sci. USA*., 108, 7850-7854.

20.

McGill, B.J. (2010). Towards a unification of unified theories of biodiversity. *Ecol. Lett.*, 13, 627–642.

21.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K. *et al*. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.*, 10, 995–1015.

22.

McGill, B.J. & Nekola, J.C. (2010). Mechanisms in macroecology: AWOL or purloined letter? Towards a pragmatic view of mechanism. *Oikos*, 119, 591–603.

23.

Meyer, F., Paarmann, D., D’Souza, M., Olson, R., Glass, E.M., Kubal, M. *et al*. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9, 386.

24.

National Audubon Society. (2002). The Christmas Bird Count historical results. Retrieved from http://www.audubon.org/bird/cbc.

25.

Nekola, J.C. & Brown, J.H. (2007). The wealth of species: ecological communities, complex systems and the legacy of Frank Preston. *Ecol. Lett.*, 10, 188-196.

26.

Nekola, J.C., Šizling, A.L., Boyer, A.G., & Storch, D. (2008). Artifactions in the log-transformation of species abundance distributions. *Folia Geobot*., 43, 259-268.

27.

Nijenhuis, A. & Wilf, H.S. (1978). *Combinatorial Algorithms for Computers and Calculators*. Academic Press, New York.

28.

North American Butterfly Association. (2009). *NABA Butterfly Counts: 2009 Report*, http://www.naba.org.

29.

Pueyo, S., He, F. & Zillio, T. (2007). The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecol. Lett.*, 10, 1017-1028.

30.

Pielou, E. 1975. Ecological diversity. Wiley, New York, New York, USA.

31.

Reed, W. J. 2001. The Pareto, Zipf, and other power laws. Economic Letters 74:15-19.

32.

Roselló-Morak, R. & Amann R. (2001). The species concept for prokaryotes. *FEMS Microbiol. Rev.*, 2001, 39-67.

33.

Sauer, J.R., Hines, J.E., Fallon, J.E., Parkieck, D.J., Ziolkowski, D.J. Jr. & Link, W.A. (2011). *The North American Breeding Bird Survey 1966-2009*. Version 3.23.2011. USGS Patuxent Wildlife Research Center, Laurel, MD.

34.

Schloss, P.D. & Handelsman, J. (2006). Toward a census of bacteria in soil. *PLoS Comput. Biol.*, 2, e92.

35.

Shipley, B., Vile, D., & Garner, É. (2006). From plant traits to plant communities: A statistical mechanistic approach to biodiversity. *Science*, 314, 812-814.

36.

Šizling, A.L., Storch, D., Reif, J., and Gaston, K.J. (2009a). Invariance in species-abundance distributions. *Theor. Ecol.*, 2009, 89-103.

37.

Šizling, A.L., Storch, D., Šizlingova, E., Reif, J. & Gaston, K.J. (2009b). Species abundance distribution results from a spatial analogy of central limit theorem. *P. Natl. Acad. Sci. USA*.,106, 6691-6695.

38.

Smith, B. & Wilson, J.B. (1996). A consumer's guide to evenness indices. *Oikos*, 76, 70-82.

39.

Sugihara, G. (1980). Minimal Community Structure: An explanation of species abundance patterns. *Am. Nat*., 116, 770–787.

40.

Supp, S.R., Xiao, X., Ernest, S.K.M. & White, E.P. (2012). An experimental test of the response of macroecological patterns to altered species interactions. *Ecology*, 93, 2505-2511.

41.

Thibault, K.M., Supp, S.R., Giffin, M., White, E.P. & Ernest, S.K.M. (2011). Species composition and abundance of mammalian communities. *Ecology*, 92, 2316-2316.

42.

Ulrich, W., Ollik, M. & Ugland, K.I. (2010). A meta-analysis of species–abundance distributions. *Oikos*, 119, 1149–1155.

43.

U.S. Department of Agriculture, F.S. (2010). Forest inventory and analysis national core field guide (Phase 2 and 3), version 4.0. Washington, DC: U.S. Department of Agriculture Forest Service, Forest Inventory and Analysis.

44.

Vershik, A. & Yakubovich, Y. (2001). The limit shape and fluctuations of random partitions of naturals with fixed number of summands. *Mosc. Math. J.*, 1, 457-468.

45.

Warren, R.J. II., Skelly, D.K., Schmitz, O.J. & Bradford, M.A. (2011). Universal Ecological Patterns in College Basketball Communities. *PLoS ONE*, 6, e17342.

46.

Webster, N.S., Taylor, M.W., Behnam, F., Lücker, S., Rattel, T., Whalan, S., *et al*., (2010). Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. *Environ. Microbiol*., 12, 2070-2082.

47.

White, E.P., Thibault, K.M. & Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, 93, 1772–1778.

48.

Woodcock, S., Curtis, T.P., Head, I.M., Lunn, M. & Sloan, W.T. (2006). Taxa–area relationships for microbes: the unsampled and the unseen. *Ecol. Lett.*, 9, 805–812.

49.

Yen, J.D.L, Thompson, J.R. & MacNally, R. (2012). Is there an ecological basis for species abundance distributions? *Oecologia*, 10.1007/s00442-012-2438-1.

50.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press, Oxford.

Figure 1.

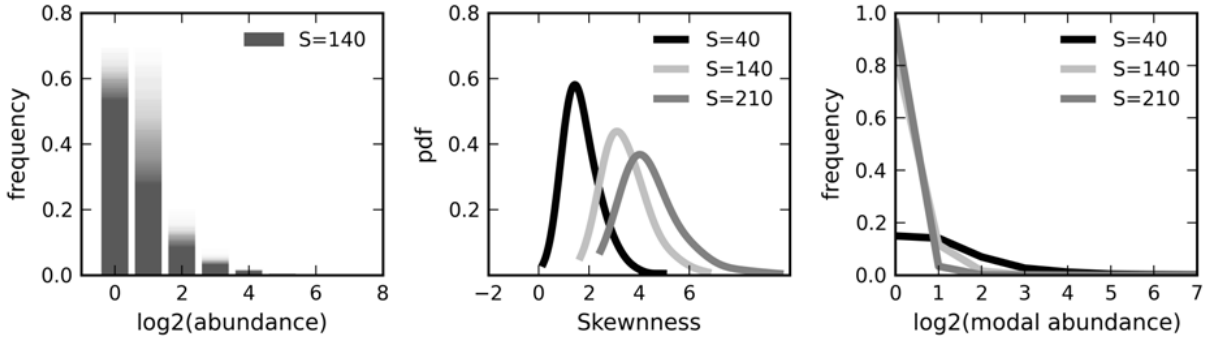
frequency
log2(abundance)
S=140
pdf
Skewnness
S=40
S=140
S=210
frequency
log2(modal abundance)
S=40
S=140
S=210

Figure 2.

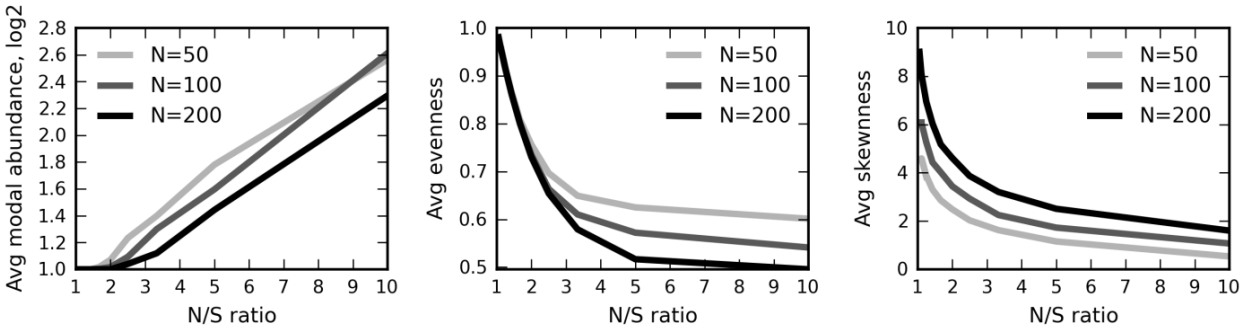
Avg modal abundance, log2
N/S ratio
N=50
N=100
N=200
Avg evenness
N/S ratio
N=50
N=100
N=200
Avg skewnness
N/S ratio
N=50
N=100
N=200

Figure 3.

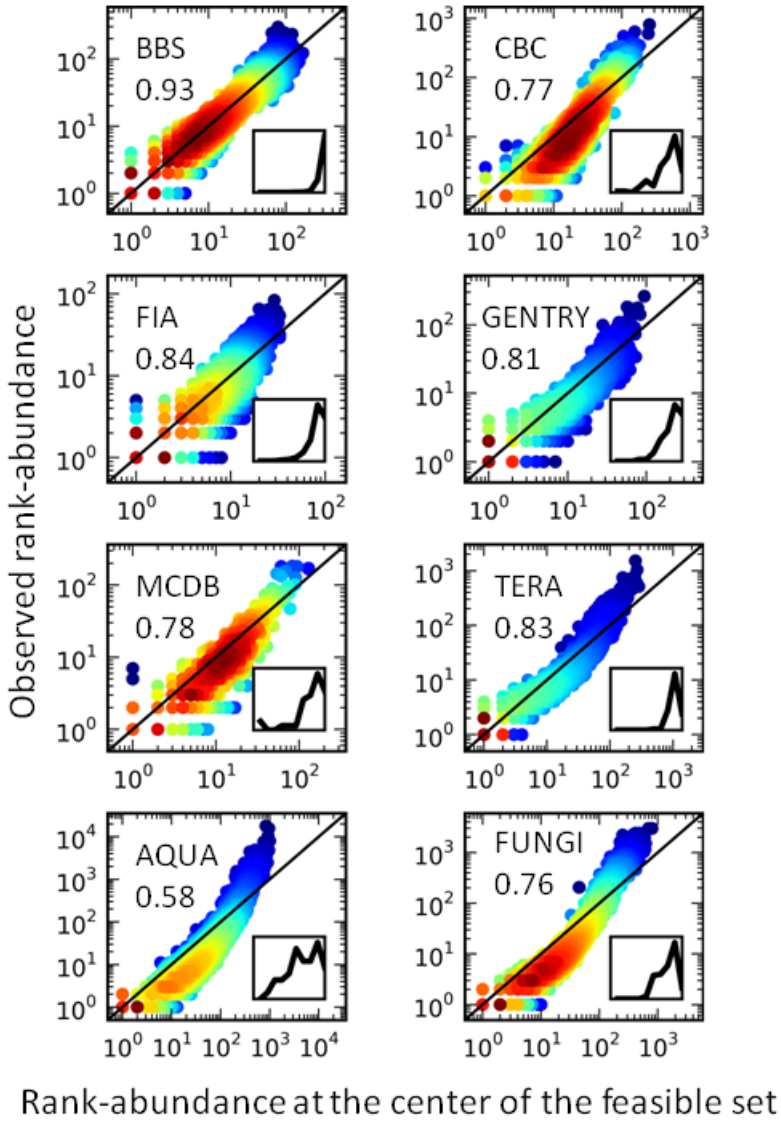
BBS
0.93
CBC
0.77
FIA
0.84
GENTRY
0.81
MCDB
0.78
TERA
0.83
AQUA
0.58
FUNGI
0.76
Observed rank-abundance
Rank-abundance at the center of the feasible set

Figure 4.

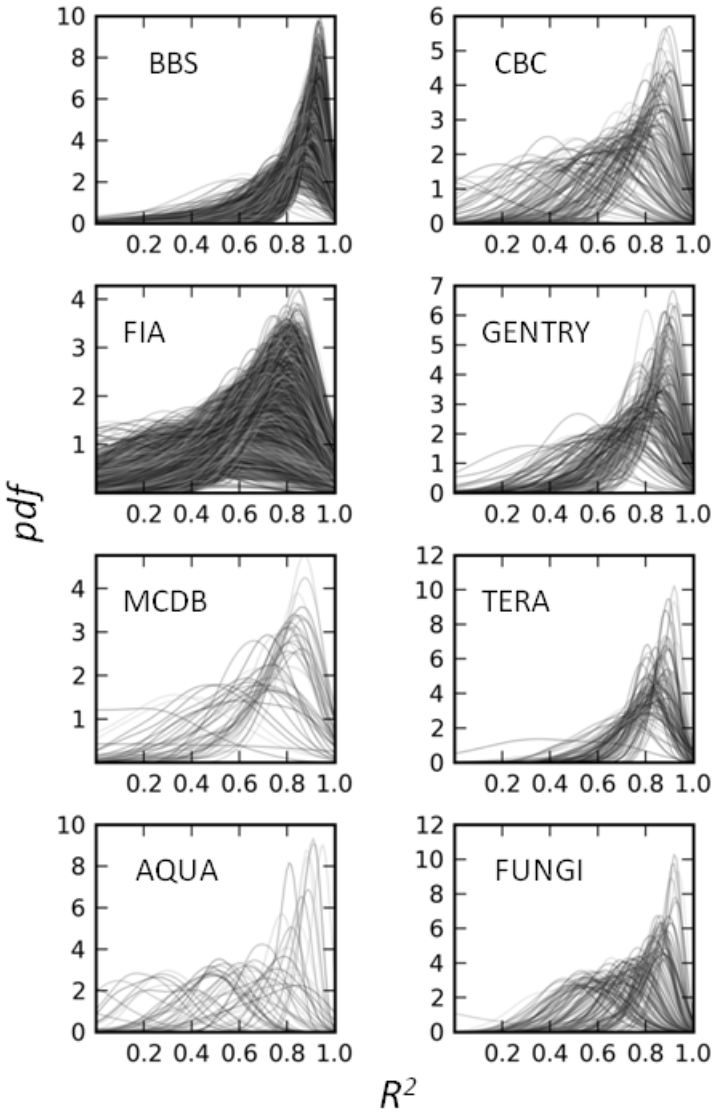

Figure 5.

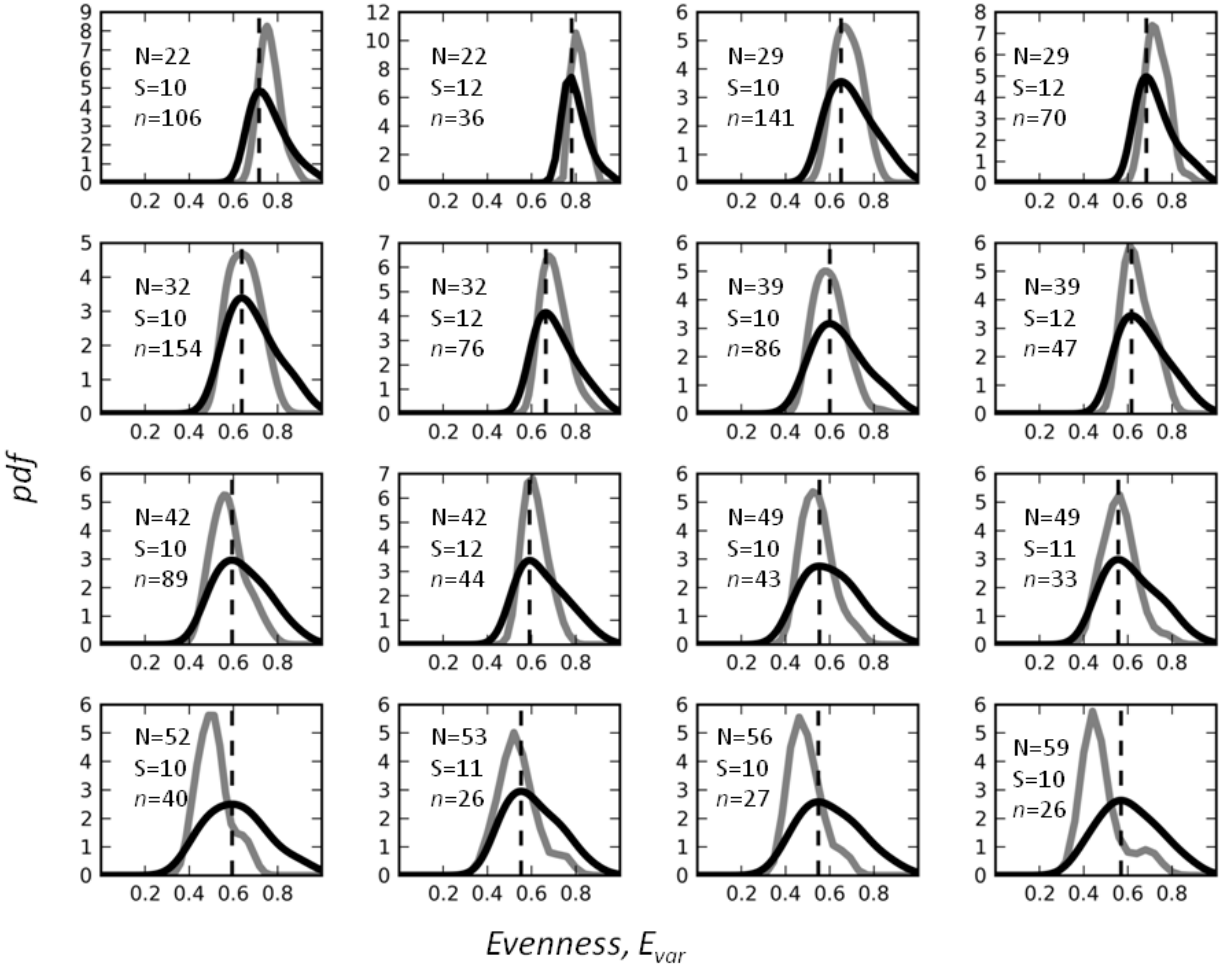
N=22
S=10
n=106
N=22
S=12
n=36
N=29
S=10
n=141
N=29
S=12
n=70
N=32
S=10
n=154
N=32
S=12
n=76
N=39
S=10
n=86
N=39
S=12
n=47
N=42
S=10
n=89
N=42
S=12
n=44
N=49
S=10
n=43
N=49
S=11
n=33
N=52
S=10
n=40
N=53
S=11
n=26
N=56
S=10
n=27
N=59
S=10
n=26
pdf
Evenness, E_var

Figure 1. Left, plots of 500 randomly sampled macrostates in the feasible set for N = 1,000 and S = 140. Each macrostate is plotted as a light grey frequency distribution of $\log_2$ abundance classes. Overlap of these distributions produces the gradual shading to dark grey. Center and Right, plots of skewness and modal abundance across random samples of feasible set for N = 1000 and S = 40, 140, 210 reveal that feasible sets are dominated by right-skewed macrostates and that the modal abundance class tends towards singletons or small abundances, indicating that feasible sets are dominated by similarly-shaped hollow-curves.

Figure 2. Plots of average abundance (N/S) against modal abundance, evenness, and skewness averaged across 500 randomly sampled macrostates for N = 50, 100, 200 and S = {N/10, N/9, …, N}. The monotonic change in these features of the feasible set with increasing N/S across doublings of N suggests predictable changes and constraints on community structure resulting from changes in N and N/S.

Figure 3. Plots of the relationship between observed rank-abundances from all sites in a dataset and the corresponding ranked abundances at the center of the feasible set. Each point represents a rank in a community with the y-coordinate showing the observed abundance at that rank and the x-coordinate showing the abundance at the center of the feasible set. Data are heat mapped to reveal the density of rank-abundance states, which is largely centered around the 1:1 line for some datasets (e.g. BBS, GENTRY) and deviates more greatly for others (e.g. AQUA, FUNGI). Insets are of kernel density curves for site specific R^2 values; the x-axis ranges from 0.0 to 1.0.

Figure 4. Kernel density curves of R^2 values relating random macrostates to the observed RAD as in Figure 3. Each site is represented by a single kernel density curve, revealing that the

majority of a random sample of the feasible set often describes large portions of variation in ranked abundances at a site.

Figure 5. Plots of kernel density curves for E_{var} across entire feasible sets (black curves) and kernel density curves for E_{var} across sites in FIA with the same N and S (grey lines). Sample size, i.e. number of FIA sites, is given as 'n'. Modes of the feasible sets are shown by vertical dashed lines. Each column reveals 1.) a shift in the mode of the feasible set towards lower evenness as average abundance (i.e. N/S) increases as in Figure 2; 2.) a shift in the distribution of empirical E_{var} towards lower evenness that out-paces the changing mode of the feasible set; and 3) a more narrow distribution of observed E_{var} values than expected from sampling from the feasible set.