**Short title**: SAMPLING FEASIBLE SETS OF ECOLOGICAL PATTERNS

**Article title**: Sampling feasible sets of ecological patterns of abundance

Kenneth J. Locey‡1 and Daniel J. McGlinn\*1,2

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 daniel.mcglinn@usu.edu;

**Statement of authorship**: KL derived the integer partitioning algorithms, coded them in Python, conducted the analyses, and served as primary author. DM coded algorithms into R, and served as secondary author. Both KL and DM contributed substantially to the organization of the paper, its ideas, and discussion.

**Keywords**: combinatorics, constraints, feasible set, integer partitions, macroecology, sampling algorithms, species abundance distribution, SAD, spatial abundance distribution

**Words in the abstract**: 269

**Words in the document**:6,396

**Number of cited references**: 29

**Number of figures**: 5

**Number of tables**: 1

SUMMARY

**1.** Ecological variables such as species richness (*S*) and total abundance (*N*) can strongly influence the forms of ecological patterns. For example, the shape of the species abundance distribution (SAD) can often be explained by the majority of possible shapes having the same *N* and *S*, i.e., the feasible set. The feasible set reveals whether the variation that is possible in the forms empirical patterns is sufficient for interpreting them in terms of either statistical or ecological mechanisms, whether random or non-random. However, the feasible set can be immense and randomly sampling from it is a non-trivial problem with a non-arbitrary solution. **2.**  The feasible set approach has only been recently developed and has been limited to the SAD and to using inefficient random sampling algorithms. We expand the feasible set approach to another primary ecological pattern, i.e., the spatially implicit distribution of species abundance also known as the intraspecific spatial abundance distribution (SSAD) and develop algorithms that greatly increase the efficiency of sampling from the SAD and SSAD feasible sets. We provide Python and R based implementations of these algorithms along with testing software.

**3.** The algorithms we develop are often several orders of magnitude faster than the most recently used approach for generating random samples of the SAD feasible set. This greatly increases the size and diversity of communities that can be examined. Defining and examining the SSAD feasible set reveals how the forms of the SAD and SSAD are coupled.

**4.** The algorithms we developed greatly improve the practical use of the feasible set approach, and hence, the ability of ecologists to understand and quantify how distributions of abundance among species and across space are constrained by two general variables.

INTRODUCTION

**1. Patterns are important and many processes have been suggested  
-- different processes produce similar patterns  
-- often not clear why this is the case**

Understanding patterns of abundance, distribution, and diversity is a primary goal in ecology (Brown 1995; Rosenzweig 2002; Blackburn & Gaston 2003). Such patterns are used to characterize community structure and provide insights into mechanisms that shape communities, while the ability to predict the forms of such patterns serves as a primary basis for evaluating ecologies theories (e.g. MacArthur & Wilson 1967; Scudo & Ziegler 1978; May 1981; Hubbell 2001; Harte 2011). Though ecological theories are underpinned by a host of different processes and mechanisms, e.g., colonization and dispersal limitation, niche differentiation, stochastic population dynamics, and random placement, it is increasingly been shown how general ecological variables such as total community abundance (*N*) and species richness (*S*) also have strong constraining influences on the forms of ecological patterns (Harte *et al*. 2008; McGlinn & Hurlbert 2012; Supp et al. 2012; White *et al*. 2012; Locey & White 2013).

**3. The feasible set provides a means of examining observable variation  
-- define macro/micro states and explain why we focus on unique curves  
-- If observable variation is highly aggregated then one should expect  
many different explanations to converge on the same curve - suggesting  
that curve fitting is a weak test of ecological inference**

More than 90% of observed variation in the species abundance distribution (SAD), i.e. the vector of abundances of species in a community (McGill *et al*. 2007), can often be explained using models constrained by *N* and *S* (Harte 2011; White *et al*. 2012). This means that despite the manifold processes that may lead to an uneven SAD characterized by many rare species and few abundant ones, an accurate prediction for the shape of the SAD can often be obtained by modeling how *N* can be distributed among *S* species, without more explicit information. In fact, the majority of possible SAD shapes having the same *N* and *S* (i.e. the feasible set) can often explain the majority of variation in abundance among species (Locey & White 2013). This is important when considering that only the shape of the SAD is generally predicted and, if most of the possible SAD shapes are similar, then the prediction of an SAD model might only reflect the constraining influence of *N* and *S*, and not the importance of whatever random or non-random ecological or statistical process or mechanism underpins a model.

**4. The feasible set has done a good job on the SAD but other patterns**

**should be possible**

Locey & White (2013) used the SAD feasible set to reveal how *N* and *S* constrain observable variation in the shape of the SAD and that the degree of variation that is possible is not always sufficient for interpreting the SAD in terms of statistical and ecological mechanisms, whether random or non-random; a point initially introduced by (Haegeman & Loreau 2008). In addition to revealing how general variables like *N* and *S* constrain observable variation in the SAD, general properties of the feasible set reveal simple mechanism-free explanations for common patterns. For example, the ubiquity of right-skewed SADs, where many species have low abundances and few species have high abundances, is predicted by the vast majority of possible SAD shapes, regardless of whether the possible shapes would arise through random or non-random ecological or statistical processes. Though only the SAD has been examined using a feasible set approach, Locey & White (2013) suggest that other ecological patterns should also be examined in the context of their feasible sets. However, there are several obstacles to extending the feasible set concept to other patterns and to using the feasible set approach without great computational resources.

Feasible sets can be immense and enumerating them can be untenable. For example, there are 8.9x1014 unique shapes of the SAD for N = 1000 and S = 10, i.e. when the abundances of ten species sum to 1000. However, small random samples can be used to characterize the properties of the feasible such as the distribution of statistical features (e.g. species evenness) within it. Locey & White (2013) took a conceptually simple and unbiased approach to sampling the feasible set of SAD shapes, an approach known as integer partitioning. This approach is based on the fact that there are a limited number of unordered ways that *n* integers can sum to a total *q*, and hence, a limited number of ways that the abundances of *S* unlabeled species can sum to a total abundance of *N*. These unordered configurations of integers are called integer partitions (Bóna 2006). For example, the feasible set for *q* = 6 and *n* = 3 is: {(4, 1, 1), (3, 2, 1), (2, 2, 2)}, where differently ordered configurations having the same integer values (e.g. (4, 1, 1), (1, 1, 4), (1, 4, 1)) represent the same integer partition, i.e. (4, 1, 1). In the same way, they represent the same frequency distribution (one 4 and two 1’s) and the same rank distribution (4, 1, 1). Consequently, each integer partition represents a unique shape of the SAD for a given *N* and *S*.

Sampling from the set of all possible SAD shapes means sampling from the set of all distinguishable SAD outcomes, i.e. the observable variation. Again, this is because SAD models and theories only predict the shape of the SAD and not which species has which abundance. In sampling from the feasible set of SAD shapes, we are effectively accounting for how variables such as *N* and *S* influence the shape of the SAD apart from how any particular shape might arise (e.g. [4, 1, 1] versus [1, 4, 1]) through random or non-random ecological or statistical sampling processes and mechanisms. In this way, the feasible set is the only way to address the following question: Regardless of what mechanisms or processes could produce a highly uneven SAD, be they ecological or statistical, random or non-random, are nearly all possible SAD outcomes highly uneven and very similar?

Though, use of integer partitioning to randomly sample the feasible set allows the feasible space to be characterized without generating all possible forms, the current sampling algorithms are computationally inefficient. All published partitioning algorithms sample the feasible set only with regards to the total, *q*. In this way, all partitions of *q* have the same probability of being drawn, regardless of the number of elements *n*. This means that randomly sampling the feasible set for a given *q* and *n*, requires generating partitions according to *q* and then rejecting those not having *n* elements, often resulting in impractically high rejection rates. For example, randomly generating one partition for *q* = 1000 and *n* = 10 requires drawing from a feasible set of nearly 2.4x1031 partitions, one of the roughly 8.9x1014 having 10 elements; a probability of nearly 3.7x10-17.

Another challenge in applying integer partitioning to the study of feasible sets is that some ecological patterns of abundance include parts with zero values. One example is the species spatial abundance distribution (SSAD) describing the frequency with which individuals of a single species occupy areas within a landscape (Brown *et al*. 1995; Harte *et al*. 2008; Haegeman & Etienne 2010; Harte 2011). The SSAD reflects the spatially implicit pattern of aggregation across a landscape and is mechanistically linked to other ecological patterns (Brown *et al*. 1995; Harte 2011). In the SSAD, individuals can be absent from a number of areas, meaning that there are some areas with zero individuals of the particular species. Because integer partitions *per se* do not include zeros, integer partitioning methods need to be modified to examine the SSAD feasible set.

Here, we present algorithms that greatly increase the efficiency of sampling the SAD and SSAD feasible sets. We explain each algorithm and develop Python and R based implementations of them. We test each algorithm for sampling bias and for speed against the method of Locey & White (2013). To reveal the practical gains of these new algorithms, we reanalyze the SAD datasets of Locey & White (2013), one of the largest and most diverse compilations of species abundance data, wherein it took more than 10000 compute hours to examine 60% of the available data (9562 of 15950 SADs). Our algorithms will allow us to examine a larger portion of the data in less time. Finally, we examine general characteristics of the SSAD feasible set, which provides a constraint-based explanation for previous connections made between the distribution of abundances among species (i.e. SAD) and the distribution of individuals of a species across the landscape (i.e. SSAD) and suggest a constraint-based explanation for Taylor’s Law, i.e., the increase in variance with increase in average abundance. Our work expands the feasible set approach to an additional ecological pattern, to values of *q* (e.g., total community abundance, total species abundance) and *n* (e.g. species richness, number of samples or quadrats) that were previously untenable. This work greatly advances the front of numerical constraint-based methods in ecology (Pueyo *et al*. 2007; Haegeman & Loreau 2008; Harte *et al*. 2008; Haegeman & Etienne 2010; Harte 2011; Locey & White 2013).

METHODS

Here, we develop efficient and unbiased integer partitioning algorithms to generate random samples of feasible sets for discrete ecological patterns such as the SAD and SSAD that are defined by a total *q* and composed of *n* elements, where both *q* and *n* are positive integers. Integer partitioning is a mature field of mathematics and algorithms for generating random partitions of *q* (e.g. Nijenhuis & Wilf 1978) are often implemented in mathematical environments (e.g. Sage, Maple, Mathematica). However, these approaches do not allow the random partitioning of *q* into exactly *n* elements. Here, we use two well-established theorems and a partitioning identity to develop a general method for generating a random partition of *q* having *n* elements. We then modify our method to allow elements having zero values. We begin by using two theorems to generate a random partition of *q* (see Appendix 1 for generating functions and recurrence relations).

1. For every integer *q* there are *p*(*q*) partitions having *q* or less elements.
2. For every integer *q* there are *pk*(*q*) partitions having *k* or less as the first element.

For example, there are 5 partitions of 4 and each has 4 or less as the first element; i.e. *p*(4) = 5, {(4), (3, 1), (2, 2), (2, 1, 1), (1, 1, 1, 1)}. Likewise, there are 4 partitions of 4 having 3 or less as the first element; *p*3(4) = 4, {(3, 1), (2, 2), (2, 1, 1), (1, 1, 1, 1)}.

A random partition of *q* can be built, element by element, by iteratively applying these two theorems (Fig. 1). Specifically, if we choose a random number *x* from 1 to *p*(*q*), we can say there are at least *x* partitions of *q* having some value *k* or less as the first element, i.e. *x* ≤ *pk*(*q*). Likewise, there must also be some value of *k* – 1, for which, there are less than *x* partitions of *q* having *k* – 1 or less as the first element, i.e. *pk*-1(*q*) < *x*. Putting these statements together, there must be a value *k* for which *pk*-1(*q*) < *x* ≤ *pk*(*q*). In this way, we can find the value of the first element in one of the partitions by finding the value of *k* that satisfies *pk*-1(*q*) < *x* ≤ *pk*(*q*). Having found the value of the first element, we can decrease *x* by *pk*-1(*q*) and *q* by *k*, and then find the first element for this combination of smaller values. Repeating this process will sequentially build the partition until *q* = 0 and the sum of the partition is equal to the original value of *q* (Fig 1).

The above approach is similar to well-established methods of generating random partitions of *q* (e.g. Nijenhuis & Wilf 1978, Stojmenovic 2008). However, our goal is to generate random partitions of *q* having *n* elements. For this, we use a well-known integer partitioning identity to restrict the number of elements in a randomly chosen partition to *n*. Specifically, the number of partitions of *q* having *n* elements equals the number of partitions of *q* having *n* as the first element (Bóna 2006). This is because each partition of *q* having *n* elements corresponds to one unique partition of *q* having *n* as the first element (Bóna 2006). For example, consider the partition (3, 1), which can be illustrated with rows of dots, called a Ferrer’s diagram (Fig 1). In the Ferrer’s diagram for (3, 1) there are two rows, the largest having three dots. Flipping the diagram on its diagonal produces its conjugate (2, 1, 1), which has three rows, the largest row having two dots (Fig 1). So, the conjugate of (3, 1) is (2, 1, 1) and vice versa. Consequently, the first part of an integer partition determines the number of parts in its conjugate (Bóna 2006). This allows us to extend the problem of generating random partitions of *q* to random partitions of *q* having exactly *n* elements. That is, knowing the first element must be *n* so that its conjugate has *n* elements, we can decrease *q* by *n* and then generate a random partition for this decreased value of *q* having *n* or less as the first element. Once the partition is generated, we append *n* to the beginning of the partition and conjugate it to produce a random partition of the original *q* having exactly *n* elements (Fig 1).

The approach outlined above begins with a randomly chosen number *x* between 1 and *p*(*q*). It then finds the value of *k* that satisfies the inequality *pk*-1(*q*) < *x* ≤ *pk*(*q*), which is the value of the first element of the partition*.* However, the question remains as to which value of *k* to start with and how to proceed to different values. Indeed, we could start with the smallest possible value of *k* (*k* = 1) and take a ‘bottom-up’ approach, or the largest possible value and take a ‘top-down’ approach, or even choose *k* at random and use a ‘divide-and-conquer’ method. These approaches differ only in how *k* is chosen and each builds the partition one element at a time. However, in the event that *q* is much larger than *n*, e.g. all trees in the 50 ha Barro-Colorado Island mapped forest plot where *q* ≈ 200,000 individuals and *n* ≈ 300 species, the three algorithms above will still be inefficient. This is because they would first generate a partition of, say, 200000 having 300 as the first element, but having as many as 199701 elements, and then conjugate it to produce a partition of 200000 having 300 elements. Clearly building the partition one element at a time would be inefficient in this case.

An alternative approach is to build a partition using multiples of integers – the ‘multiplicity’ approach. Instead of finding the value of *k* corresponding to the first element, appending it, and moving on, we can instead ask how many times must *k* occur, i.e. the partitions of *q* having some multiple *m* of *k*. We can start with the smallest possible multiple (i.e. *m* = 1) and ask whether *x* is less than or equal to the number of partitions of *q* – *k*\**m* having less than *k* as the first part. This is because the set of partitions of *q* having a number of *k*’s equal to *m* actually contains the set of partitions of *q* – *k*\**m* having less than *k* as the first part (Appendix 1). We can increase *m* by one until *x* ≤ *pk*(*q* – *k*\**m*), at which point we will have found the corresponding multiple of *k.* One drawback to this method is the time taken for the extra computation.

*Random partitions for q and n, with some parts having zero values*

The above algorithms address distributions having positive values, such as the distribution of abundance among species (SAD). In contrast, some ecological patterns include zero values (e.g. absences). One example is the species spatial abundance distribution (SSAD), a frequency distribution that characterizes the number of quadrats, cells, or areas containing a given abundance of a species (Brown *et al*. 1995; Haegeman & Etienne 2010; Harte 2011). However, only small changes are needed to adapt the above approaches to cases allowing zero-valued parts. For example, let 10 unlabeled individuals occupy a landscape sectioned into quarters. The most aggregated distribution would be for all 10 to occupy the same quarter, [10, 0, 0, 0]. The least aggregated would be for 3 to occupy two quarters while 2 occupy the other two quarters, [3, 3, 2, 2]. In fact, the number of configurations for 10 unlabeled individuals distributed across 4 unlabeled sections equals the number of partitions of 10 having 4 or less parts, i.e. *p*4(10 + 4) = 23. Consequently, if *n* ≤ *q*, a random partition for *q* and *n* allowing for zero-valued parts, is simply a random partition for *q* having *n* or less parts, with zeros appended to ensure the final form of the partition has *n* parts.

On the other hand if *n* > *q* a different approach is needed. To see this let 4 unlabeled individuals occupy a landscape sectioned into tenths. The most aggregated distribution would be for all 4 to occupy the same subsection, [4, 0, 0, 0, 0, 0, 0, 0, 0, 0] and the least aggregated configuration would be for 4 sections to have one individual and for 6 sections to have zero, i.e. [1, 1, 1, 1, 0, 0, 0, 0, 0, 0]. In this way, the number of possible configurations for 4 unlabeled individuals distributed across 10 unlabeled sections is *p*(*q*). Consequently, if *q* < *n*, a random partition for *q* and *n*, allowing for zero-valued parts, is simply a random partition for *q* having *q* or less parts, with zeros appended to ensure the partition has *n* parts.

*Examining for bias and speed*

We implemented the above algorithms in Python and R and made them freely available using a public Github repository (https://github.com/klocey/partitions). We also developed these algorithms into an R package (“rpartitions”) to be distributed on The Comprehensive R Archive Network (CRAN; http://cran.us.r-project.org/). We used kernel density curves to visually compare the results of the above algorithms to full feasible sets and random samples generated with the function implemented in the Sage mathematical environment (www.sagemath.org) that is based on the algorithm of Nijenhuis and Wilf (1978) and is the method used in Locey & White (2013). If our algorithms are unbiased, then their distributions will not differ in any systematic way from full feasible sets and random samples generated using the proven function implemented in Sage.

Additionally, our Python and R packages include test files that conduct, among several other tests of our source code and partitioning functions, 2-sample t-tests and 2-sample Kolmogorov Smirnov tests on kernel density curves of the variance of logarithmically transformed abundances from random samples generated by the Sage software and random samples generated using the algorithms derived here. These tests are important to regularly run because additional code developments can corrupt source code.

We compare the computational speed of our algorithms to that of the approach used in Locey & White (2013) (i.e., using Sage to generate random partitions for a given *q* and rejecting those not having *n* elements) across a range of values of *q*, *n*, and *q*-*n* ratios for which the latter method was likely to return random samples within reasonable time (one hour). Because Sage is coded in Python, our comparisons are made using the Python versions of our algorithms.

*Empirical Demonstration of the New Algorithms*

Locey & White (2013) analyzed the species abundance distributions (SADs) of 9562 sites of trees, bird, mammal, fungi, and prokaryote communities using a partitioning algorithm that sampled the feasible set according to total abundance *N* but not with respect to species richness *S* (i.e. the number of elements). Those data consisted, in part, of a subset of previously compiled datasets of site-specific species abundance data (see White *et al*. 2012), and included four continental-to-global scale surveys, including the Christmas Bird Count (129 sites) (National Audubon Society 2002), North American Breeding Bird Survey (1,586 sites) (Sauer *et al*. 2011), Gentry’s Forest Transect Data Set (182 sites) (Phillips & Miller 2002), Forest Inventory Analysis (7,359 sites) (U.S. Department of Agriculture 2010), and one global-scale data compilation, the Mammal Community Database (42 sites) (Thibault *et al*. 2011). Locey & White (2013) also compiled abundance data at the species level from five microbial metagenome projects for a total of 264 SADs. Those data were obtained from the metagenomics server MG-RAST (Meyer et al. 2008). Metagenomic data were compiled into datasets representing aquatic prokaryotic communities (48 metagenomes) (Flores *et al*. 2011, [www.catlin.com/en/Responsibility/CatlinArcticSurvey](http://www.catlin.com/en/Responsibility/CatlinArcticSurvey)), terrestrial prokaryotic communities (92 metagenomes) (Chu et al. 2010; Fierer *et al*. 2012), and terrestrial fungal communities (124 metagenomes) (Amend *et al*. 2010). We refer the reader to Locey & White (2013) and White *et al*. (2012) for more thorough descriptions of those datasets.

The inefficiency of the partitioning method used by Locey & White (2013) restricted their analyses to combinations of abundance *N* and species richness *S*, for which, there was a reasonable probability of generating a random integer partition of *N* with exactly *S* elements. This restriction allowed for only 60% of the available data to be examined despite more than 10000 compute hours worth of effort. We reanalyze those datasets using the algorithms developed here, which should allow for random samples of a greater number of SADs to be produced in less time.

*General characteristics of the SSAD feasible set*

Brown *et al*. (1995) revealed evidence that the general form of the SSAD, like that of the SAD, is characterized by a hollow-curve. In the sense of the SSAD, a hollow-curve implies that many areas are occupied by few or no individuals and that relatively few areas are occupied by many individuals. Both Brown *et al*. (1995) and Harte (2011) mechanistically couple the SSAD and SAD using arguments based on local niche differences and the constraint-based approach of entropy maximization, respectively. We generated random samples of the feasible set of the SSAD for ecologically realistic combinations of *q* and *n*, and examined their general features.

RESULTS

Statistical properties of entire feasible sets are indistinguishable from random samples generated with our sampling algorithms, demonstrating that the implementations of our algorithms were unbiased (Fig 2 and Figs 1-2 of Appendix). When generating 300 random partitions, i.e. enough to safely characterize the feasible space (Locey & White 2013), these implementations were, at worst, 10 to 100 times faster than the method used by Locey & White (2013) and were, at best, 10000 to 100000 times faster for the combinations of *q* and *n* we tested (Fig 3). These combinations were limited to values for which the algorithm used in Sage could generate random samples in reasonable time. Consequently, the algorithms we developed quickly produce random samples for values of *q* and *n* that are impractical with algorithms that sample only according to *q*. Each algorithm was best suited for particular values of *q* and *n* (Fig 4). For cases where all parts have positive values, the multiplicity algorithm is the fastest for combinations where *q* is partitioned among a relatively small number of elements (Fig 3 Appendix).

The greater efficiency of the algorithms developed here allowed us to generate between 300 and 500 random partitions for 92.7% of the SADs (14786/15950) from the compilation of SAD data used by Locey & White (2013), in less than 1000 compute hours. In contrast, the method used by Locey & White (2013) required more than 10000 compute hours to generate between 300 and 500 random partitions for 60% of the available data (9562/15950 SADs).

Our examination of the SSAD feasible set supports the observation of Brown *et al*. (1995) that SSADs are characterized by hollow-curves (few cells with many individuals and many cells with few individuals) (Fig 5). The hollow-curve nature of the SSAD feasible set increases as the total *q* is distributed across a greater number of elements (e.g. quadrats, subplots, cells). This reveals an expectation for decreased variation in the spatially implicit distribution of a species (i.e., most samples are going to have few individuals and few samples will have many) as *q* is distributed across an increasing number of elements. According to the feasible set, variance increases with increasing average cell abundance.

DISCUSSION

The feasible set approach based on integer partitioning is a contextual framework for understanding how constraints influence the forms of discrete ecological patterns and in discrete distributions of wealth, size, and abundance (Locey & White 2013). We used integer partitioning to understand how the feasible set is ordered, to find the size and general features of the feasible set, and to generate unbiased random samples of the feasible set for a given a total (e.g. total community abundance, total species abundance) and number of entities (e.g. species, quadrats). The algorithms we derived greatly increase the practical use of feasible set by decreasing computing time. In addition to examining the distribution of abundance among species (i.e. SAD), we expanded the feasible set approach to discrete distributions with zero values, such as the spatially implicit distribution of individuals of a species within a landscape (i.e. SSAD). We also provided the algorithms in two computing languages frequently used by ecologists, R and Python, and have taken steps to ensure our implementations are unbiased and have provided a script to detect computational errors that can result from future code development (e.g. when a user modifies the code).

Integer partitioning is only one way to examine and randomly sample the feasible set of possible SAD and SSAD shapes. Other possibilities include constraint-based programming (see http://cran.r-project.org/web/views/Optimization.html) and iterative random walks, such as that used by Haegeman & Loreau (2008). Those approaches may not require combinatorial problems to be solved and so may not suffer from the problem of combinatorial explosion (large increases in the size of the feasible set for small changes in the total q and number of elements). However, as stated by Locey & White (2013) one benefit to the integer partitioning approach is that the random sampling algorithms are inherently unbiased and do not require ‘burn-in’ periods to produce effectively independent samples. The combinatorial approach also reveals properties such as the size of the feasible set, the distribution of statistical characteristics (e.g. species evenness, diversity, modal abundance class) across the feasible set, and connections between different ecological patterns, e.g. SAD and SSAD. However, we suggest that constraint-based programming and iterative random walks should also be examined and compared to the combinatorial approach.

Our examination of the SSAD feasible set (Fig. 5) reveals that the central tendency of the set is characterized by a hollow-curve which is consistent with the empirical SSADs found by Brown *et al*. (1995). In that study, the authors state that the highly ‘clumped’ and hollow-curve nature of the SSAD resembles distributions used to predict the form of the SAD. The authors offer an ecological interpretation for the similarity between the patterns in terms of niche requirements. Likewise, the Maximum Entropy Theory of Ecology (Harte 2011) also argues for a fundamental but purely constraint-based relationship between the forms of the SAD and SSAD. However, in terms of their feasible sets determined by the total and the number of entities, the SSAD differs from the SAD only in that zero values are allowed. Consequently, the forms of the two patterns are not only coupled by ecological and statistical mechanisms, but are more simply coupled by the mathematical properties of their feasible sets.

The results of examining general properties of the SSAD feasible set suggest the potential for a simple constraint-based explanation for the increase in variance with average abundance, i.e. Taylor's Law (Taylor 1961). Taylor’s Law also generally predicts a linear slope between 1 and 2 for the log-log relationship of the variance versus mean abundance (Kilpatrick & Ives 2003). If a feasible set based examination of this relationship consistently yields slopes between 1 and 2, then this would provide a second feasible set based explanation for a pattern that has been intensively studied in ecology but commonly observed in other disciplines; the ubiquity of the hollow-curve frequency distribution being the other.

The feasible set approach taken here and in Haegeman & Loreau (2008) and in Locey & White (2013) ignores biological and statistical mechanism and focuses entirely on observable variation in the shape of empirical patterns. Consistency of empirical patterns with the center of the feasible set suggests that the shapes of those patterns contain little information beyond that encoded by the constraints used to characterize the feasible set (Haegeman & Loreau 2008; Locey & White 2013). However, consistency with the feasible set does not mean that biological processes are not operating but rather that they may indirectly influence empirical patterns through their effects on constraints (Supp *et al*. 2012; White *et al*. 2012). Indeed, if the majority of variation in an ecological pattern can be explained/predicted by a few general variables, then it is probably important to the prediction of the pattern to understand the forces, processes, or mechanisms driving the values of the variables (McGill 2010). Alternatively if empirical patterns occupy an uncommon portion of the feasible set (e.g., in being exceptionally uneven) biological processes or additional constraints beyond those used to characterize the feasible set may be relevant.

Our work greatly advances the ability of ecologists to characterize and explore observable variation in ecological patterns of abundance by greatly decreasing computational time and by defining the feasible set of another ecological pattern of abundance, i.e., the SSAD. These advances allow combinations of constraint values to be examined that were previously out of reach, and hence, will help provide a greater understanding of the degree to which small combinations of general variables can explain the forms of ecological distributions (e.g. SAD, SSAD) and common ecological patterns (e.g. hollow-curve frequency distributions, Taylor’s Law). The algorithms we developed apply to frequency distributions such as the SAD and SSAD. However, many ecological patterns are also cumulative, describing the rates at which species are encountered with increasing area (species-area relationship) or time (species-time relationship) or both (species-time-area relationship), as well as the spatially implicit distribution of occupancy among species within a landscape (occupancy-frequency distribution). Characterizing and randomly sampling the feasible sets of these and other patterns may require modification of the algorithms we developed, approaches more similar to that of Haegeman & Loreau (2008), or altogether new approaches.

ACKNOWLEDGMENTS

We thank X. Xiao and E. P. White for critical discussions and friendly reviews. We thank the numerous individuals involved in collecting and providing the data used in this paper including the essential citizen scientists who collect the North American Breeding Bird Survey and Christmas Bird Count data, USGS and CWS scientists and managers, researchers who collected and sequenced the microbial metagenomic data, the MG-RAST project, the Ribosome Database Project, the Audubon Society, the U.S. Forest Service, the Missouri Botanical Garden, and Alwyn H. Gentry.

LITERATURE CITED

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. *Proceedings of the National Academy of Sciences USA*, 107, 13748-13753.

Bóna, M. (2006) *A walk through combinatorics: An introduction to enumeration and graph*

*theory*. 2nd Edition. World Scientific Publishing Co. Singapore.

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

Brown, J.H., Mehlman, D.W. & Stevens, G.C. (1995) Spatial variation in abundance. *Ecology*, 76, 2028-2043.

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.

*Environmental Microbiology*, 12,2998–3006.

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. *The* *ISME Journal*, 6, 1007–17.

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. *Environmental Microbiology*, 13, 2158-2171.

Haegeman, B. & Etienne, R.S. (2010) Entropy Maximization and the Spatial Distribution of

Species. *The* *American Naturalist*, 175, E74–E90.

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

117, 1700–1710.

Haegeman, B. & Loreau, M. (2009) Trivial and non‐trivial applications of entropy maximization in ecology: a reply to Shipley. *Oikos*, 118, 1270-1278.

Harte, J., Zillio, T., Conlisk, E. & Smith, A.B. (2008) Maximum entropy and the state-variable

approach to macroecology. *Ecology*, 89, 2700–2711.

Locey,K*.*J. & White, E.P. (2013) How species richness and total abundance constrain the

distribution of abundance. *Ecology Letters*. 16, 1177-1185.

MacArthur, R.H. & Wilson, E.O. (1967) *The theory of island biogeography* (Vol. 1). Princeton University Press.

May, R.M. (1981) Theoretical ecology. Principles and applications. *Theoretical ecology. Principles and applications.*, (Ed. 2).

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

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. *Ecology Letters*, 10, 995–1015.

McGlinn, D.J. & Hurlbert, A.H. (2012) Scale dependence in species turnover reflects variance in species occupancy. *Ecology*, 93, 294–302.

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.

National Audubon Society. (2002) *The Christmas Bird Count historical results*. Retrieved from

<http://www.audubon.org/bird/cbc.>

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

Academic Press, New York.

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

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.

Scudo, F.M. & Ziegler, J.R. (1978) *The Golden age of theoretical ecology, 1923-1940: a collection of works by V. Volterra, VA Kostitzin, AJ Lotka, and AN Kolmogoroff*. Springer-Verlag.

Stojmenovic, I. (2008) Generating all and random instances of a combinatorial object. In

*Handbook of Applied Algorithms: Solving scientific, Engineering and Practical Problems*. John Wiley & Sons, Inc., New Jersey, pp. 1-38.

Storch, D., Šizling, A.L., Reif, J., Polechová, J., Šizlingová, E. &Gaston, K.J. (2008) The quest

for a null model for macroecological patterns: geometry of species distributions at multiple spatial scales. *Ecology Letters*, 11,771–784.

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.

Taylor, L.R. (1961) Aggregation, variance and the mean. *Nature*, 189, 732–735

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.

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.

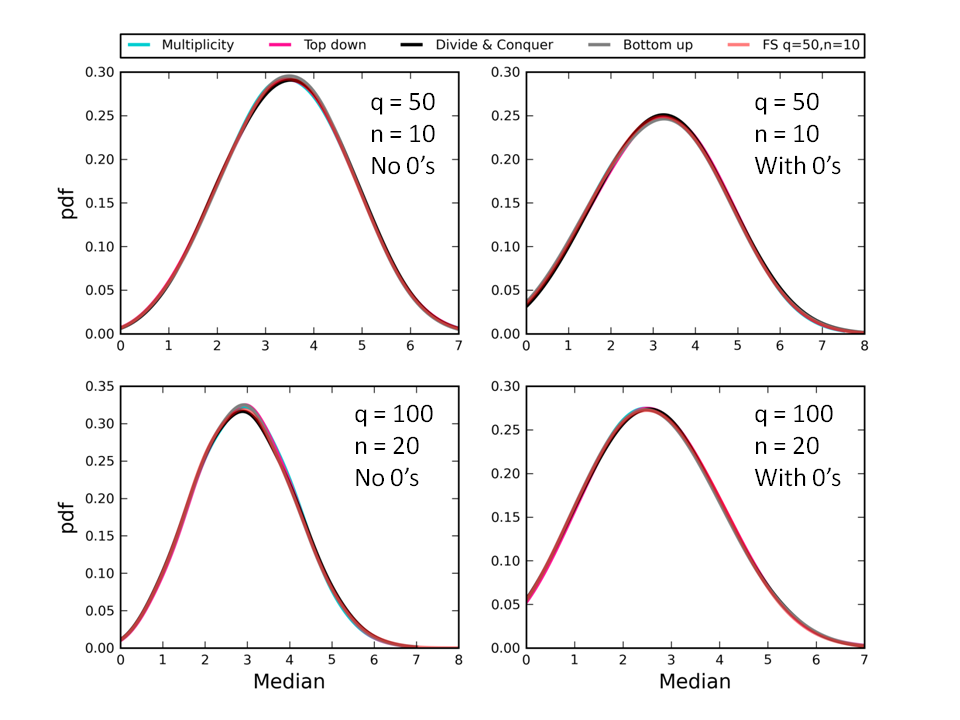
White, E.P.,Thibault, K*.*M. & Xiao, X. (2012) Characterizing species abundance distributions

across taxa and ecosystems using a simple maximum entropy model. *Ecology*, 93, 1772–

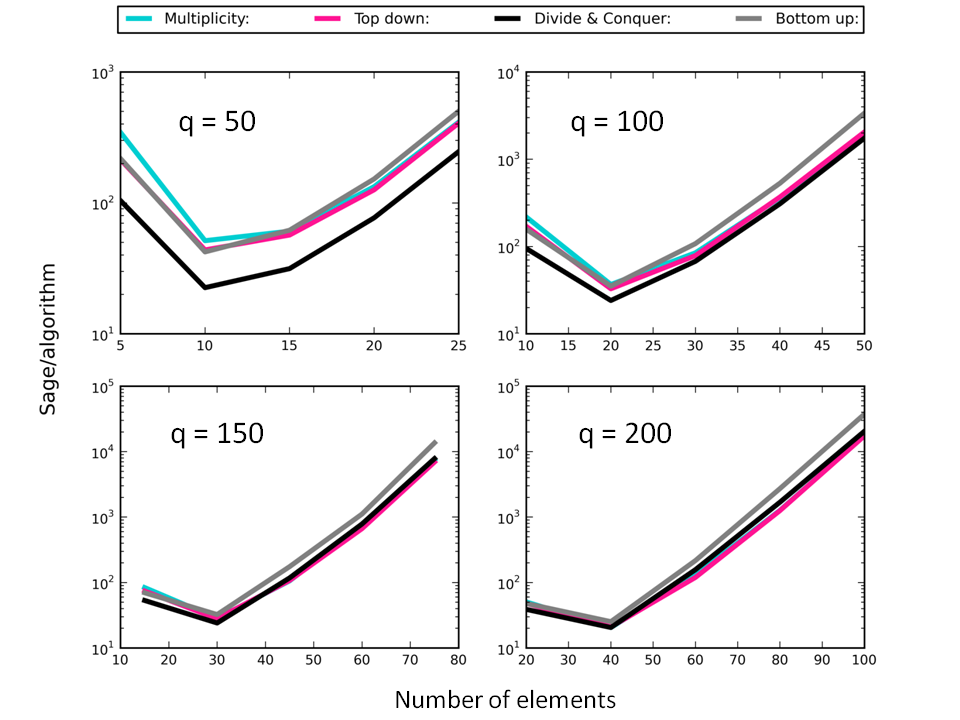
1778.

Figure 1.emf

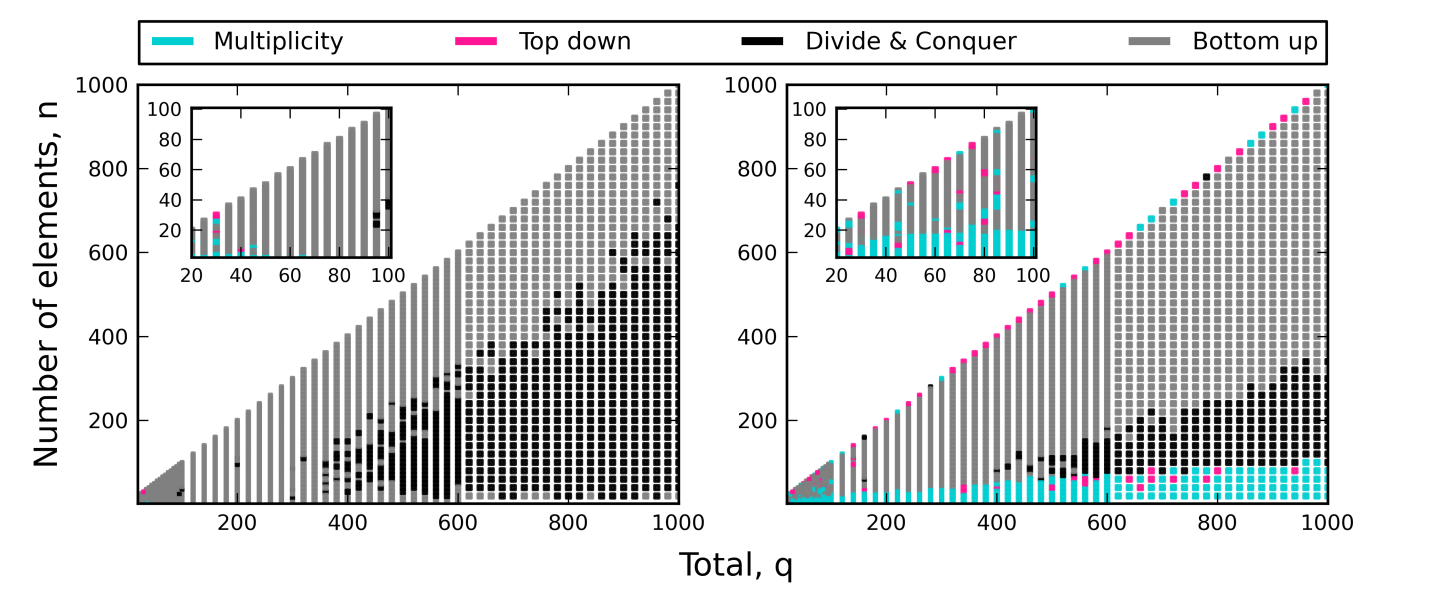
**Figure 1**. *Top:* Two theorems and one identity used to generate random integer partitions. *Center:* General method for generating a random partition of *q* having 1 to *q* elements using the two theorems. *Bottom:* General method for generating a random partition of *q* into exactly *n* elements, using the two theorems and the partitioning identity.



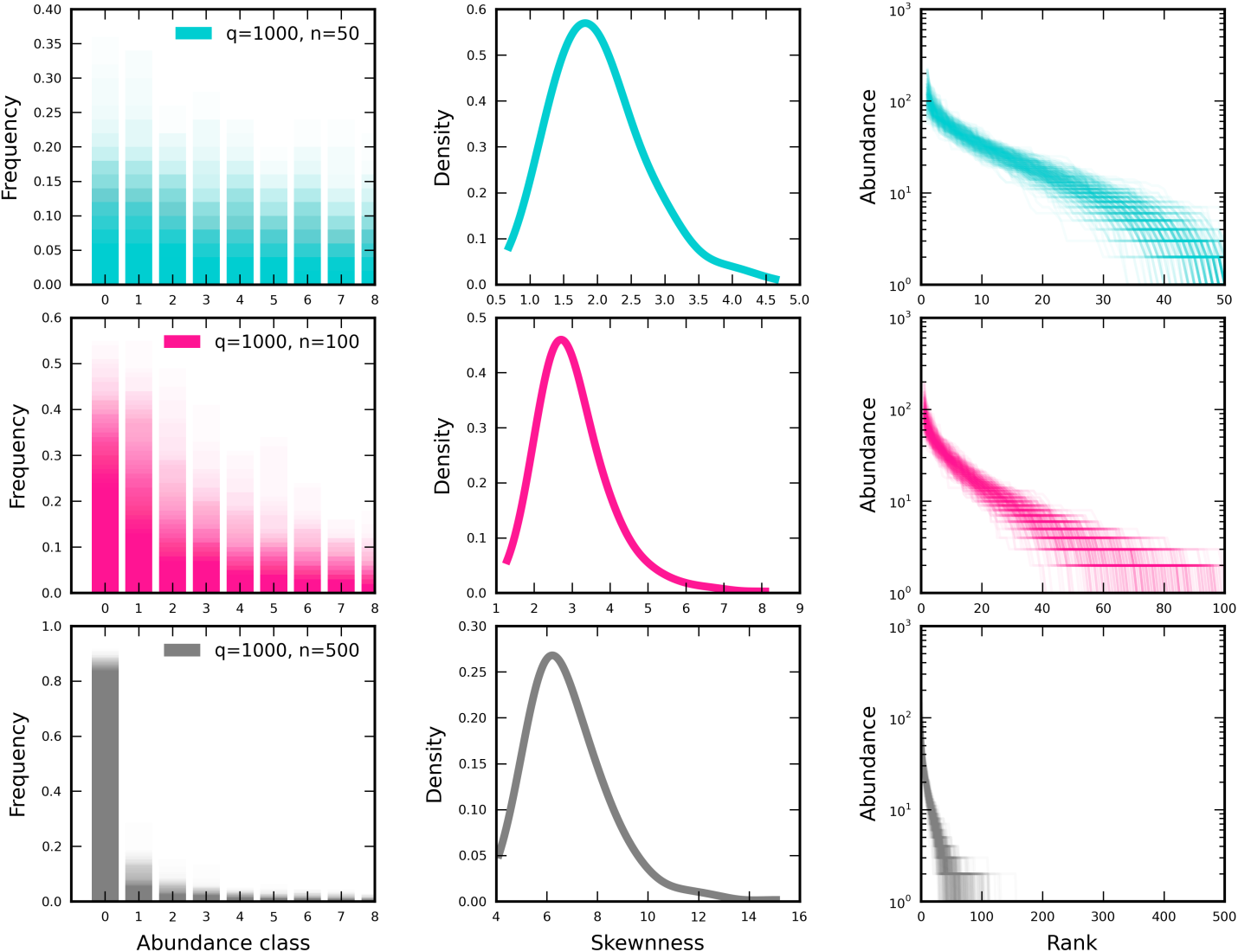
**Figure 2**. A comparison of the full feasible set and kernel density curves for the median derived from 1000 random samples for different combinations of *q* and *n* using our four new algorithms for parts without zeros (left column) and for parts with zeros (right column). The similarity between the results derived using our algorithms and the full feasible set reveals that the algorithms produce unbiased random samples of the feasible set. We used Sage to generate the entire feasible set for *q* = 50 and *n* = 10 (16928 partitions) and used the random partitioning function in Sage to generate 1000 partitions for *q* = 100 and *n* = 20, which is too large to enumerate in full in reasonable time (10474462 partitions).



**Figure 3**. Plots of the ratio of the computational time for Sage to generate 300 random integer partitions (no zeros) to the time taken for the new algorithms (‘Multiplicity’, ‘Top down’, ‘Divide and Conquer’, ‘Bottom up’) to do the same across a range of *n* values at four different values of *q*. The ratio is plotted on a log-transformed axis.



**Figure 4**. Color map revealing the fastest algorithm for specific combinations of *q* ≤ 1000 and *n* ≤ *q*. Comparisons were based on the time taken to generate 300 random partitions for each combination of *q* and *n*, both for cases where parts were allowed to have zero values (left) and when parts had positive values only (right). Insets reveal the small corner of the main graph where *q* ≤ 100.



**Figure 5**. The SSAD feasible sets for *q* = 1000 and *n* = {50, 100, 500} for 300 random samples displayed as frequency distributions (left), kernel density curve of skewness (center), and ranked distributions of abundance (i.e. areas or subplots ranked from most-to-least occupied) (right). Each row applies to one value of *n*. For n = 500, ranks with values of 0 are plotted but difficult to discern.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Number of SADs\* | Locey & White  **>10K hours** | Present Study  **~1K hours** |
| North American Breeding Bird Survey | 2769 | 1586, 57% | 2769, 100% |
| Christmas Bird Count | 1992 | 129, 6.5% | 1231, 62% |
| Gentry’s forest transects | 222 | 182, 82% | 221, 99.5% |
| Forest Inventory and Analysis | 10356 | 7359, 71% | 10101, 98% |
| Mammal Community Database | 103 | 42, 41% | 103, 100% |
| Aquatic metagenomes | 252 | 48, 19% | 120, 48% |
| Terrestrial metagenomes | 128 | 92, 72% | 113, 88% |
| Fungi metagenomes | 128 | 124, 97% | 128, 100% |
| Total | 15950 | 9562, **60%** | 14786, **92.7%** |

\* The number of SADs in the dataset matching the sampling criteria (see methods).

**Table 1**. The number and proportion of the SAD datasets that we analyzed in the present study compared to the past analysis of Locey & White (2013). Note, sampling effort per dataset was not controlled or accounted for.