A new twist on a very old binary similarity coefficient

JOHN ALROY¹

Department of Biological Sciences, Macquarie University, New South Wales 2109 Australia

Abstract. Pairwise similarity coefficients are downward biased when samples only record presences and sampling is partial. A simple but forgotten index proposed by Stephen Forbes in 1907 can help solve this problem. His original equation requires knowing the number of species absent in both samples that could have been present. It is proposed that this count should simply be ignored and that the coefficient should be adjusted using a simple heuristic correction. Four analyses show that the corrected equation outperforms the Dice and Simpson indices, which are highly correlated with many others. In two-sample simulations, similarity is almost always closer to the assumed value when the species pool size and sampling intensity are varied, regardless of whether the underlying abundance distribution is uniform, log-normal, or geometric. The index is also much more robust when sampling is unequal. An analysis of bat samples from peninsular Malaysia buttresses these conclusions. The corrected coefficient also indicates that local assemblages of North American mammals are random subsamples of larger species pools by returning similarity of values of around 1, and it suggests a more consistent relationship between biome-scale comparisons and local-scale comparisons. Finally, it yields a better-dispersed pattern when the biome-scale inventories are ordinated. If these results are generalizable, then the new and old equation should see wide application, potentially taking the place of the two most commonly used alternatives (the interrelated Dice and Jaccard indices) whenever sampling is incomplete.

Key words: abundance distributions; biogeography; community ecology; Forbes index; North American mammals; principal coordinates analysis; sampling bias; similarity coefficients.

INTRODUCTION

The entomologist, ornithologist, and limnologist Stephen Alfred Forbes was one of the most important, if perhaps the most important, of all 19th-century ecologists. A holistic Darwinian, he “anticipated many ecological concepts not formalized until later” such as competitive exclusion, population cycles and limits, and predator–prey interactions (McIntosh 1985:32). He was broadly influential, and his ideas about the stability of ecosystems influenced a generation of younger workers such as Frederick Clements (McIntosh 1985, Kingsland 1995).

Forbes is also of note because he published a measure of similarity (Forbes 1907) that was predated by only a few others (Choi et al. 2010). His metric has good general properties, has an elegant derivation, and uniquely indicates the degree to which overlap in species lists corresponds with a chance expectation given the null hypothesis that both samples are drawn from the same community (Forbes 1907, Michael 1920, Dice 1945). For example, if two samples have a Forbes similarity of 0.8 then the number of species found in both samples is 80% as great as one would expect of two random draws.

In this paper, I note a simple way to make Forbes’ now very obscure index more widely applicable by adjusting the counts it uses to compensate for its potential upward bias. The thrust of the argument in favor of this corrected index is that it is relatively robust to variation in the number of observations (i.e., sample size) and/or species richness, even when sampling is unequal. Although the adjustment is heuristic it is mathematically justified by the fact that it compensates directly for strong bias in the count of species shared by two samples. Its strength lies not only in this property but in its simplicity and elegance combined with its performance. As shown later, other measures are only ever more accurate if sampling is excellent. In such cases, the choice of an index is relatively unimportant.

Despite these facts, readers may be categorically uncomfortable with heuristic methods or with strictly pairwise presence–absence data. However, hundreds of ecologists continue to employ binary presence–absence measures and existing alternatives are simply not robust when sampling is partial. Insisting that these metrics should not be used would be counterproductive because the questions and data often demand it.

The reason that sampling intensity is a particular problem for binary data is that most measures are systematically depressed when one sample is much larger than the other (Wolda 1981) or when sampling is simply poor. The problem cannot be avoided when observational species lists are used because sample sizes are
unknown and cannot be controlled. Historical, biogeographic, and paleontological analyses routinely suffer from this handicap. The sampling problem motivated the development of an early index by Simpson (1943) and has received occasional attention from methodologists (e.g., Fallaw 1979, Raup and Crick 1981, Wolda 1981, Chao et al. 2005), but it remains unsolved.

After presenting the revised equation, I show that it helps to ameliorate sampling biases and I test it against two disparate but well-known alternative indices: the ones proposed by Dice (1943) and Simpson (1943). The Simpson index (misattributed by Wolda [1981] to Dice [1945]) is simple and algebraically very distinct, and unlike most others it does attempt to control for variation in sample size (Simpson 1943, Fallaw 1979). Meanwhile, the choice of the Dice coefficient is obvious given its very broad use and substantial theoretical support despite its lack of any particular justification when sampling is partial (Hubálek 1982, Todeschini et al. 2012). The equally popular coefficient of community (Jaccard 1901) is essentially a transform of it but is even more downward-biased (Hubálek 1982). Other metrics such as the Ochiai and second Kulczynski coefficients that have good properties (Hubálek 1982) are in practice very strongly correlated with the Dice index, so they do not merit detailed attention.

**Methods**

*Sample-size-corrected equation.*—Forbes index $F$ is the expected number of species found in two random samples divided into the observed number (Forbes 1907, Michael 1920). The derivation is simple. Following standard notation throughout most of the literature in this area, let $a$, $b$, $c$, and $d$, respectively, be the number of species found in both samples, the first only, the second only, or neither one; and let $N = a + b + c + d$. Forbes (1907) reasoned that a species has an $(a+b)/N$ chance of being found in the first sample and likewise an $(a + c)/N$ of being found in the second. Thus, $[(a + b)(a + c)]/N^2$ is the chance of being found in both and $[(a + b)(a + c)]/N$ should be the number of species found in both, i.e., the expected value of $a$. The observed-to-expected ratio $F$ is therefore

$$
(a \times N)/[(a + b)(a + c)].
$$

Equation 1

Meanwhile, Dice’s $D$ is $2a/(2a + b + c)$ and Simpson’s $S$ is the maximum of the two ratios is $a/(a + b)$ and $a/(a + c)$. The Jaccard index $a/(a + b + c)$ is only one algebraic step away from either $D$ or $S$: the first is obtained by multiplying both $a$ terms by 2 and the second by discounting either $b$ or $c$.

Forbes’ $F$ stands out both because of its very different formulation and because its term $N$ includes $d$ (the number of shared absences). This property is a handicap because there is often no sensible way to define which entirely unsampled species should be considered potentially present in an observational (meaning nonexperimental) data set. For example, if two quadrats are 1 km apart from each other and a species absent in both of them is encountered just 100 m away from the first, surely anyone would agree that it should count toward $d$. But if the distance is 100 km, then perhaps no. One could counter-argue that $d$ is easily calculable if there are numerous samples instead of just two (e.g., Chase et al. 2011), but this argument is off topic because the whole point of a pairwise index is to examine two samples in isolation. Furthermore, any number of additional samples could be used to inflate $d$ to any arbitrary level with the only restraint being the global number of species in the taxonomic group of concern.

Regardless, one approach to the problem of not knowing $d$, or at least not wishing to define it from arbitrary additional data, would be to simplify $F$ by setting $d$ to 0 and thereby exclude it from the calculation of $N$. For clarity, let the remaining sum $a + b + c$ be notated as $n$:

$$
(a \times n)/[(a + b)(a + c)].
$$

Equation 2

Using $n$ instead of $N$ immediately makes Forbes’ index comparable to the others, and as will be shown by simulation it produces reasonably accurate values.

There is, however, a second problem: Regardless of whether $d$ is set to zero, $F$ scales convexly to the underlying probability of overlap and is therefore upward biased (Hubálek 1982). For example, if there are 20 species in each of two samples, 10 are shared, and all species are sampled then $D$, $K$, and $S$ are all 0.5 but the redefined $F = 0.75$. Likewise, it is 0.19 when the others are 0.1 and it is 0.99 when they are 0.9.

A solution to this problem is hinted at by rearranging Eq. 2 as follows:

$$
(a \times n)/[(a \times n) + (b \times c)].
$$

Equation 3

It can now be seen that all of the information about similarity is coming from the relative size of the product $b \times c$. If it is 0, then similarity is 1 and, if it is much greater than $a$, then similarity is near zero. Thus, the crux of the scaling problem is that $b \times c$ is too small, causing $F$ to be too large. I propose increasing it by a simple constant

$$
(a \times n)/\left[\frac{(a \times n) + \frac{3}{2}(b \times c)}{2}\right].
$$

Equation 4

The range of this ratio is still 0–1. If $a = b = c$ then it is 2/3, which is better than 0.75 but still higher than the more intuitive value of 0.5. Given $a = b = c = 10$, a constant of 3 instead of 2/3 would actually generate this value in the limit, but because it would render the index identical to the Dice coefficient under a wide array of circumstances, it would obliterate the advantages of $F$. As can be shown by simulation, adopting a more modest correction yields much better results when sampling is incomplete. The basic reason is that $b$ and $c$ are much less affected by undersampling than $a$ because species have to be sampled twice to appear in the latter count. Thus, allowing $b$ and $c$ to dominate the equation too...
much by using a higher constant would lead to a negative bias.

I furthermore assert that the $a \times n$ terms are problematic when sampling is very poor because the value $a$ limits too quickly on zero, creating a negative bias. This property is also related to the excessive influence of $b$ and $c$. A reasonable solution would be to shift the ratio closer to 1 when $n$, $b$, and/or $c$ are large by increasing both $a \times n$ terms. I therefore propose adding three simple constants instead of one to produce what turns out to be a robust estimator under a broad array of conditions

$$a(n + \sqrt{n})\left[a(n + \sqrt{n}) + \frac{3}{2}(b \times c)\right]. \quad (5)$$

The resulting equation can be called the corrected Forbes index or $F'$. For clarity, I note that Eq. 5 can be rearranged to resemble Eqs. 1 and 2

$$F' = a(n + \sqrt{n})\left[(a + b)(a + c) + a\sqrt{n} + \frac{3}{2}(b \times c)\right]. \quad (6)$$

The corrections have no effect on limit properties. If $a = 0$, the original and new equations both produce zero (the numerator of the equation being 0). If $a = n$ and therefore $b = c = d = 0$ (meaning the samples are identical), then the basic equation yields 1 (regardless of whether $n$ or $N$ is used) and so does the corrected version.

As mentioned, in the hypothetical 50% similarity case in which $a = b = c$ and all three are large, the new correction pushes $F$ down from 3/4 to 2/3. Likewise, if $a = b/2 = c/2$, the uncorrected Eq. 2 limits on 0.56 while Eq. 6 limits on a more reasonable 0.455 and $D$, $K$, and $S$ all yield 0.333. As mentioned, if $a = b = c$ and the scaling term were to be 3 instead of 2/3, then $F'$ and $D$ would both limit on 0.5. Likewise, if the scaling term were to be 2.5, then the limit would be 0.333. In other words, Eq. 6 causes $F$ to shift specifically in the direction of $D$ but the point of convergence does depend on the circumstances. In any case, I will later show that the tendency of $F'$ to exceed the traditional measures when using the 3/2 constant is generally a good thing given partial sampling.

Although exact expressions for the error in $F$ and $F'$ do not exist, a simple bootstrap analysis can be used to compute confidence limits. For example, given $a = b = c = 20$, the 50% confidence limits on $F$, $F'$, $S$, and $D$ are, respectively, 0.710–0.801, 0.648–0.752, 0.486–0.583, and 0.442–0.537. These ranges are nearly equivalent, and the similar variation in these four indices is emphasized later on by illustrating empirical data.

Sensitivity tests.—Five analyses of three data sets are used to establish the $F'$ equation's relative performance.

1. Simulations.—The first analysis is a simulation showing how variation in true overlap between two sampling pools influences its value. It assumes, by default, that each pool includes 100 species with log-normal frequencies and that 100 individuals are drawn from each one during a given trial. The frequencies have a mean of 2 on a natural log scale and a standard deviation of 2.5. There are 100 trials per run. Alternative simulations have pool sizes of 500 species (to generate poor sampling) or 20 species (to generate almost complete sampling).

As in the preceding discussion, true overlap is defined as the expected value of $a/b$ or $a/c$, given complete sampling of the underlying species pool (the ratios being on average equal). This definition is intuitive because it means that when half the species are shared, similarity is one-half. It is also fair, because $S$, $D$, and many others all equal these simple ratios when sampling is complete. In other words, the simulations define true similarity in a way that tilts the outcome in favor of $D$.

Additional trials employed either the uniform or truncated geometric series frequency distributions instead of the log-normal distribution. The results based on the uniform and log-normal models were almost identical but the former is less realistic, so the log-normal distribution results are presented first to keep things simple. The geometric series parameter $k$ was set at either 0.01 or 0.10, respectively producing either high or low diversity. One important earlier study (Wolda 1981) employed the log-series distribution, which, like the log-normal, is a good fit to many data sets (e.g., Dolan et al. 2009) except when communities are environmentally disturbed (e.g., Bazzaz 1975, Qu et al. 2008). However, the distributions used here effectively bracket its shape and so are deemed sufficient.

2. Similarity of samples from the same pool.—Data for bats from the same small region of peninsular Malaysia, and therefore the same species pool, were used to show whether precise values close to 1 are returned by $F'$ and whether $D$ is instead biased by sampled richness. The data set consisted of 41 published species lists including 62 species. The sources were Luruthusamy et al. (2011, 2014), Francis (1990), Hasan et al. (2012), Kingston et al. (2003), Liat et al. (2010), Lim et al. (2014), Mariana et al. (2005, 2011), Strubeg et al. (2009), and Tingga et al. (2012). Megachiropterans were excluded because they are usually left out of such studies for logistical reasons. Records of *Hipposideros “bicolor”* sensu lato were excluded because this taxon consists of two cryptic species that are often listed separately (*H. “bicolor”* 131-kHz and 142-kHz). Specifically indeterminate records were also omitted. Comparisons were made between the largest single samples and all others. This sample consists of 1733 individuals captured by T. Kingston at Kuala Lompat Research Station in Pahang State (Kingston et al. 2003).

3. Similarity between local and regional samples.—A second empirical analysis was intended to demonstrate potential downward bias by looking at similarities between local- and biome-scale species that were generated by the various indices. The data are invento-
ries of all terrestrial, non-volant mammal species found in 21 North American biomes and in local habitats tied to these biomes (Brown and Nicoletto 1991). Only nine of the biome-scale lists are matched to a local list, i.e., biome numbers 4, 5, 6, 8, 9, 14, 16, 17, and 19 in Brown and Nicoletto (1991). There are also multiple local lists for all but one of the biomes. To minimize pseudoreplication only the first-presented local list tied to a biome was employed.

4. Similarity at different scales.—An additional analysis looked at the offsets between local- and biome-scale similarity values. Given that the local-scale lists should be more or less random subsamples, we would expect similarities between these lists to have the same order and magnitude as similarities between their matched biomes. Alternatively, an index that is not robust to partial sampling should return lower similarities between local lists.

5. Ordination patterns.—The biome-scale mammal species lists are of interest exactly because sampling is effectively complete, meaning that levels of similarity should have a strictly biotic basis, and because biotic similarity is used in the first place to define the biomes. Therefore, we should expect an ordination to space out the points more evenly than might be expected of arbitrarily sampled data, making it easy to see patterns. A robust metric should also cancel out the well-known arch effect (Gauch 1982; and see Results). Given that $F$ assumes sampling is actually partial and is expected to be upward-biased when it is not, this test is particularly stringent. Similarities cannot be ordinated and most metric ordination methods assume the use of a particular distance equation. So, in this analysis, the data are first converted to dissimilarities and then subjected to principal coordinates analysis (PCoA; Gower 1966), which can employ any distance metric.

**Results**

**Simulated data.**—Simulation trials in which richness and sample size are both fixed but the degree of species pool overlap varies show that Dice and Simpson indices tend to underestimate similarity levels when sampling is moderate (Fig. 1A) to poor (Fig. 1B). The simple reason is that the proportion of species recovered in each sample puts a de facto cap on these two indices. For example, if $s$ is the proportion of species sampled and both samples come from the same pool, then $a + b = s \times N$ and $a = s^2 \times N$, so Simpson’s ratio $a/(a + b) = s$. These capping values are 0.65, 0.49, and 0.80 in Fig. 1A, B, and

![Fig. 1. Response of similarity indices to variation in richness and sample size. The underlying data consist of two randomly generated samples that each include the same number of species and are drawn from distinct but partially overlapping pools with log-normally distributed abundances. The x-axis gives the proportion of species in each pool that is also found in the other pool (i.e., the value of many coefficients given complete sampling and equal richness). The sample size is 200 individuals by default. (A) Each sample is drawn from a pool of 100 species. (B) Pool size is 500 species. (C) Pool size is 20 species. (D) Pool size is 100 species and sample size is 500 individuals in one sample and 50 in the other.](image-url)
C, respectively. The same limitation does not apply to F and F₀. As a result, Forbes’ index greatly outperforms the others in either its raw or corrected form. When the number of specimens is close to the number of species (Fig. 1A), both versions are accurate through to an overlap level of perhaps 0.7 and they are both always substantially better than the others. However, F is slightly better than F₀ at higher similarity levels. F₀ is also generally superior to the other metrics when richness and therefore relative sampling are varied (Fig. 1B, C). The only exceptions are that F is even better when sampling is very poor because richness is high (Fig. 1B) and that F is better than F when sampling is nearly complete (Fig. 1C). The routine indices are fairly accurate in this one case but downward-biased and never as good as either F or F₀ in all other circumstances. Importantly, given uneven sampling F, F₀, and S are all highly accurate, but D shows an even stronger downward bias (Fig. 1D). This pattern will also be seen in the later empirical analyses.

F and F₀ are generally insensitive to the choice of underlying abundance distributions: log-normal, uniform, or geometric (Fig. 2). Therefore, F and F₀ are consistently superior to S and D, regardless of what is assumed. D is always even worse than S in these simulations because sampling is partial and S does, weakly, compensate for the problem.

To sum up, because D and S only ever match F when sampling is excellent and similarity is low (Fig. 1C), because the metric with a lower value is almost always more biased (Figs. 1 and 2), and because sampling regimes are unknown when only presence-absence data are available, the choice of metrics is obvious: the highest value should always be favored.

Peninsular Malaysian bat data.—The bat data set is an optimal test case because the geographic region is small and bat communities tend to be spatially homogeneous (for the obvious reason that bats fly). Furthermore, many of the individual samples were taken from nearby sites in the same field areas. It therefore comes as no surprise that both versions of Forbes’ index return very high similarity values, often quite close to 1, when the 40 smaller samples are compared to the single largest one (Fig. 3A, B). As should also be expected of a robust metric, F and F₀ are uncorrelated with the number of species in the smaller sample (P > 0.05 in both cases).

Simpson’s index also performs reasonably well (Fig. 3C), as might be predicted based on the simulation results. However, it is negatively correlated with richness (ρ = −0.494, P = 0.001). This pattern is not entirely intuitive but might relate to random sampling error in the b and c counts (as defined in Methods). All other

**FIG. 2.** Response of indices to the shape of the underlying frequency distribution given poor and uneven sampling. Each analysis assumes that the pool size is 100 species and sample size is 100 individuals in one sample and 50 in the other (compare with Fig. 1D). (A) Log-normal distribution (as in Fig. 1). (B) Uniform distribution. (C) Shallow geometric distribution (geometric series parameter k = 0.01). (D) Steep geometric distribution (k = 0.1).
things being equal, when these two counts are both high due to random chance but $a$ is unchanged, $S$ is low.

By contrast, the Dice coefficient is unambiguously biased (Fig. 3D). Its values are far too low in absolute terms (median = 0.516, as opposed to 0.900, 0.874, and 0.845 for $F$, $F'$, and $S$) and they drift upward with sampled richness ($r = 0.613, P < 0.001$). This trend is predictable given the simulation results, which imply that all indices will rise as sampling improves and that $D$ is particularly biased in this regard (Fig. 1). Much more unexpectedly, the individual $D$ values for comparisons involving the largest sample are higher than the mathematical expectation for comparisons with the entire regional species pool (solid line in Fig. 3D). To put this another way, $D$ suggests that some pairs of samples are more similar to each other than to the pool. This pattern could be explained by the fact that common species are overrepresented in local samples.

Fig. 3 also indicates that $F$, $F'$, and $D$ exhibit comparable amounts of sampling variance. One fair way to draw comparisons is to examine the residual standard error after computing a linear regression of index values on sampled richness. The respective values are 0.059, 0.071, and 0.059. Although much of this variation could relate to real factors such as censusing methods and local environmental effects, these figures suggest that $F'$ is somewhat noisier than $D$ but that the difference is not very large.

North American mammal data.—Comparisons based on data for multiple biomes are also illuminating (Fig. 4). $F$, $F'$, and $S$ all sensibly imply that the local assemblages are random samples of the biome-scale species pools because similarity is very close to 1 (Fig. 4A, B, C), whereas $D$ implies that they are really quite different (Fig. 4D). This difference is dramatic and hard to explain as anything other than an artifact of the way $D$ gives equal weight to quantities $b$ and $c$, which is the reason it has so much trouble with unequal sample sizes (Figs. 1D and 2).

The $D$ values are also a negative function of biome-scale richness, which, likewise, is hard to explain as a biologically meaningful pattern. One explanation could be that $D$ is fooled by the common species that predominate in isolated samples. As a thought experiment, suppose there were 100 species, 10 species were found everywhere, and each of the other 90 was only ever found in one of two local communities. As mentioned, the respective values of $F$, $F'$, $S$, and $D$ for a pair of samples each including 20 species would be 0.75, 0.703, 0.5, and 0.5. If either sample were compared to the joint species pool, as in Fig. 4, the first three indices would rise to 1 but $D$ would fall to 0.333. If a second biome included all of the species found in the first plus an additional 100 rare species, $D$ would fall even further to 0.182. Hence the negative correlation.
In sum, the only fair interpretation is that $F$, $F'$, and $S$ are unbiased and $D$ is strongly downward-biased when comparisons are made between local samples (points in Fig. 3), between local samples and a single species pool (lines in Fig. 3), or between local samples and separate pools (Fig. 4). The reason is that $D$, like almost all other indices, is not designed to handle sampling biases, and local-scale lists are subsamples of regional lists both statistically (because local inventories are usually incomplete) and biologically (because communities are assembled from regional pools). Regardless of simulations, this empirical observation suffices to make the point that $D$ is unacceptable and a better alternative must be found.

However, $D$ has further problems. Similarity varies not only when local- or biome-scale data are compared to each other but when similarities between biomes are computed at different scales (Fig. 5). Local-scale similarities are a steep linear function of biome-scale similarities when $F$ is used (Fig. 5A). The same is true of $F'$ (Fig. 5B), and in both cases the intercept is insignificantly different from zero. There are also multiple instances in which fairly high similarities are identical (meaning points fall on the 1:1 line). In marked contrast, the $S$ and $D$ lines are shallow thanks to a breakdown in the relationship when similarity is low (Fig. 5C, D). The reason is that the lowest local-scale values are clustered around 0.1 to 0.2 whereas the biome-scale values still show some spread.

Principal coordinates analysis of the biome-scale species lists makes a final key point (Fig. 6): $D$ produces what appears to be a classic arch effect (Gauch 1982), which results from compression of distance values (complements of similarities). As a result, it inflates the amount of variance apparently explained by the second axis (Fig. 6D). A simple way to test for an arch is to compute a quadratic regression of coordinate 2 scores on coordinate 1 scores. The adjusted $r^2$ values are 0.657 for $D$ and but only 0.309, 0.400, and 0.327 for $F$, $F'$, and $S$ (the quadratic term is significant at $P < 0.02$ for all four metrics).

It is worth noting that the arch in this data set does not relate to pairs of samples being completely dissimilar, a pattern that will routinely create this problem (Gauch 1982). The reason is simply that there are no zero similarities in the matrix. However, low, non-zero similarities can also create arches. Thus, if a researcher really wants to use similarity indices in an ordination then $F$ might be another useful tool for ameliorating the arch effect, in addition to using such methods as step-across corrections (Williamson 1978, De’ath 1999).

The arch effect in the empirical analysis creates the impression that there are large gaps between biomes that are actually from similar climates (i.e., triangles vs. solid
circles and squares vs. open circles in Fig. 6). These gaps are not as pronounced when \( F \) and \( F' \) are used (Fig. 6A, B). The \( D \) plot (Fig. 6D) also suggests that the tundra/taiga biomes are only weakly related to the western North American temperate biomes. These are the kind of implications that might be touted in a publication focusing on such an analysis.

All of the classic methods assume that if true similarity is invariant then when sampled richness is higher the raw underlying counts \( a, b, \) and \( c \) are all higher (either because sampling has improved or the pool sizes are larger). However, given that real-world frequency distributions are highly skewed we might expect \( a \) to be relatively insensitive to richness. Such a pattern could in theory cause even more trouble with measuring similarity. In this regard, the biome-scale mammal data do send up a red flag: the correlation between \( a \) and the sum \( b + c \) is not only weak but negative (Spearman’s \( p = -0.298 \)). So, at least in this case \( a \) seems to capture very little species pool size signal in the way that it should.

**DISCUSSION**

Alternative corrections.—Although the correction terms used to improve Forbes’ index are elegant they are also heuristic, so it is important to explain why this approach is superior to various other adjustments that could have been employed instead. As it happens, none of the obvious adjustments are sensible. For example, altering the stand-alone \( a \) term in the numerator and denominator of Eq. 5 could have catastrophic effects, such as causing the index to sometimes exceed 1 or else prevent it from ever reaching that value. Adding a fixed constant such as 1 or a summation such as \( b + c \) anywhere would make little difference in a large data set and would therefore fail to address Eq. 3’s upward bias.

![Fig. 5. Comparisons between similarity values for pairs of biome-scale species lists and for pairs of local-scale lists within the same biomes. Quadratic fits are shown only to guide the eye; none of the second-order regression terms are significant. (A) Based on the original Forbes coefficient (\( p = 0.626 \)). (B) Corrected Forbes coefficient (\( p = 0.624 \)). (C) Simpson coefficient (\( p = 0.613 \)). (D) Dice coefficient (\( p = 0.571 \)).]
Taking the logarithm or square root of any term would be arbitrary and would have limited effects. In sum, then, only adding constants to the \(a + b + c = n\) terms and adding multiplicative corrections such as \((b \times c)^{3/2}\) to the denominator seems sensible in the first place. Regardless, no other approach produces simulation results as robust as those shown in Figs. 1 and 2.

There is nonetheless one very elegant way to force \(F\) downward: rescale it as \(1 / \sqrt{C_0 F}\). Doing so would, unfortunately, reduce \(F\) to a close correlate of the other measures. This transform also has no direct connection to sample size, whereas the whole point of the proposed correction is to model sampling bias by looking at the balance of \(a\) on the one hand and \(b\) and \(c\) on the other (the former being more biased).

If there are any other simple ways to substantially improve \(F\) while not making it parrot the other measures, I am not aware of them.

Finally, we might also ask whether a correction factor of some kind might improve other metrics. For \(D\) and \(S\), subtracting an adjustment term from the numerators or adding one to the denominators exacerbates existing downward biases while doing the reverse creates unacceptable upward bias when sampling is heavy. I am also unable to identify any correction that causes \(D\) and \(S\) to perform better in the face of unequal sampling (see Figs. 1D and 2–4). It is possible that some better correction exists, but the generally poor performance of these measures given incomplete sampling combined with the very good performance of \(F\) (Figs. 1–6) would seem render this speculation moot.

Alternative methods.—No clearly better method has been suggested for removing sample size effects from similarity indices. For example, subsampling the data only brings the problem out (Chao et al. 2005 and Figs. 1 and 2). One exception is the Simpson index, which was originally proposed exactly to deal with the case in which two samples are of very different sizes (Simpson 1943). Simpson’s index has numerous sound theoretical properties (Gower and Legendre 1986) and should help when pool sizes vary less than sample sizes. However, it has been criticized on the grounds that it is asymmetrical, taking on different values depending on the richness of each sample (Gower and Legendre 1986). Although this point is somewhat moot because it cannot be calculated for the richer one by definition, the index has very high binomial error and is outperformed by the
two versions of Forbes’ index in this paper’s simulations and empirical analyses (Figs. 1–6).

Raup and Crick (1981) developed a similarity measure that was later improved by Chase et al. (2011). It is conceptually related to the Forbes index, so one might think that it somehow compensates for sampling biases. The gist of the idea is to compute the probability of obtaining some amount of overlap \( a \) or less given that the two samples could have been generated at random from a large species pool. However, sampling probabilities bear no necessary relation to count ratios and are not measures of similarity per se. They are also sample size dependent (a problem not discussed by Chase et al. [2011]). Worse still, the Raup and Crick family of metrics assumes that numerous related samples can be used to constitute a null model universe. By contrast, \( F' \) is computed without knowing anything about the larger species pool. It thereby sidesteps the practically and philosophically very difficult problem of deciding exactly which samples should constitute the sampling universe.

Related literature.—Despite all of its advantages, Forbes’ index has been almost completely forgotten. It is routinely omitted from even the more detailed introductory discussions of similarity indices (e.g., Cheetham and Hazel 1969, Fallaw 1979, Wolda 1981, Gower and Legendre 1986, Digby and Kempton 1987, Tulloss 1997, Cox and Cox 2001, Anderson et al. 2011) or else mentioned in a cursory way (e.g., Hayek 1994, Deza and Deza 2009, Choi et al. 2010, Anderson et al. 2011, Todeschini et al. 2012). It is sometimes misattributed to Mozley (1936) or Margalef (1958); see for example Cox and Cox (2001). Few or perhaps no authors have recognized any advantage to it, early workers were skeptical of it (Michael 1920), I have failed to find any published application of it over the last one-half century, and Hubálek (1982) dismissed it for unfair reasons (as discussed in Objections to the Forbes index).

This lack of attention may come as no surprise because there are more than 70 published binary similarity measures (Choi et al. 2010), a large fraction are known by multiple names, and surveys routinely list dozens of them. Worse still, these overviews most often lack substantive discussion (e.g., Cheetham and Hazel 1969), voice no strong preferences (Gower and Legendre 1986), or argue that multiple measures should be used (Hayek 1994). Instead of making clear-cut recommendations, many researchers have instead focused on more abstract concerns (Hubálek 1982, Hayek 1994, Tulloss 1997, Choi et al. 2010, Todeschini et al. 2012).

The Jaccard and Dice indices.—Despite all of this chaos, much of the literature (e.g., van Tongeren 1995, Chao et al. 2005) focuses on two almost interchangeable measures (Jaccard 1901, Dice 1943): the Jaccard coefficient (also known as the Tanimoto coefficient or coefficient of community, the latter term being favored by Jaccard [1901]) and the Dice coefficient (also known as the Czekanowski, Gleason, or Sorenson coefficient). Note that the reason for the terminological confusion is in part the fact that Jaccard and Dice’s measures have been attributed to different papers and therefore assigned different publication dates (e.g., compare Deza and Deza 2009, Choi et al. 2010, and Todeschini et al. 2012).

Simulation analyses and general methodological arguments have provided good justification for the use of the Jaccard and Dice coefficients (e.g., Hubálek 1982, Todeschini et al. 2012), and they are closely related in algebraic terms to other well-known equations such as the Braun and Blanquet, first and second Kulczynski, first Ochiai (also known as the Otsuka), first Sokal and Sneath, and Tanimoto indices. In particular, Hubálek (1982) argued in great detail for preferring the Jaccard and Dice indices (in addition to the first Ochiai and second Kulczynski indices; both are put aside here because they correlate very strongly with the Dice coefficient). These facts may explain why Anderson et al. (2011) gave the two measures a blanket recommendation.

The fatal problem with all of these related indices is their vulnerability to sample size variation. For example, Chao et al. (2005) performed an analysis somewhat similar to the two-sample simulations of Figs. 1 and 2. They employed a sensible empirical approach (comparing two subsamples of the same data set) to test the Jaccard and Dice coefficients, and they indeed demonstrated very strong sample size effects. Their solution was to recommend a new abundance-based estimator, which is not very helpful in the current context. Chao et al. (2005) also explained how replicated presence–absence data could be used to compute their estimators instead of abundances, but this requirement is only reasonably met when true replicate samples are produced in an experimental field study. By contrast, the purpose of the new index and of others such as the Jaccard and Dice is to handle cases where there are no replicates, sample sizes are unknown, and abundances are unavailable. These cases may be frustrating, but they are still frequent.

Objections to the Forbes index.—In a recent study (Todeschini et al. 2012), Forbes’ index performed well in an empirical test, albeit not in the top tier, and the Forbes and Jaccard indices exhibited very similar behavior in a simulation. Thus, there is prior reason to think \( F' \) might be useful even if we put aside the matter of correcting it. However, Hubálek (1982) dismissed it on the grounds that it is weakly correlated with chi-square, has no firm upper limit, and is nonlinear (meaning that it changes at a varying rate as species shift between the counts one step at a time). The first argument is not relevant here because the count \( d \) is excluded from the corrected equation but is needed to compute chi-square. The same thing applies to all the other metrics under discussion. Although Forbes’ original index can exceed 1
when \( d \) is very large (Michael 1920). Hubálek (1982) was for all intents and purposes wrong on the second point: when \( a \) is zero, the original Forbes equation equals zero, and when all species are found in both samples (\( a = b = c \) and \( d = 0 \)) it equals one. This criticism also categorically does not apply to the corrected index, which has firm bounds of 0 and 1 because it ignores \( d \). Finally, Hubálek (1982) himself considered the linearity criterion to be “non-obligatory.” It also seems like a minor issue because the correction renders \( F^\prime \) more linear (Figs. 1, 2) and because \( F^\prime \) has so many other strong properties (Figs. 1–6).

In the one important case where the Forbes metric shows strong downward bias, the others perform just as badly or worse, i.e., when sampling is very poor relative to the size of the species pool (Figs. 1B, 2B, 2C). Indeed, \( F^\prime \) is always more accurate than the Dice coefficient unless sampling is extraordinarily good (Fig. 1C). Near-complete sampling could mean either that the pool is very small and therefore easily sampled or that the sample size is large, frequencies are even, and the pool is modest (say, well under 100 species). In both cases, almost any traditional index will do. The bottom line is that Forbes’ index outperforms all the others in any realistic situation where many of the species in a community are sampled but far from all (Figs. 1–5). Cases in which sampling is substantial but incomplete are exactly the ones in which there is still much to be discovered about patterns of species richness.

Acknowledgments

This is Ecological Register official publication 1. Research was supported under an Australian Research Council Future Fellowship funding scheme (project number FT0992161). I thank D. Barneche, M. Kowalewski, J. Martinelli, and an anonymous reviewer for helpful comments and D. Raup, J. Sepkoski, and L. Van Valen for spurring my nearly fruitless decades-long interest in this topic.

Literature Cited


Dice, L. R. 1943. The biotic provinces of North America. University of Michigan Press, Ann Arbor, Michigan, USA.


Ecology

ISSN: 0012-9658
Title: Ecology
Publishing Body: Ecological Society of America
Country: United States
Status: Active
Start Year: 1897
Frequency: Monthly
Volume Ends: Dec
Document Type: Journal; Academic/Scholarly
Refereed: Yes
Abstracted/Indexed: Yes
Media: Print
Alternate Edition ISSN: 1939-9170
Language: English
Price: USD 866 combined subscription per year domestic to institutions (Print & Online Eds.)
USD 955 combined subscription per year foreign to institutions (Print & Online Eds.)
Free to members (effective 2010)
Subject: ENVIRONMENTAL STUDIES
Dewey #: 577
LC#: QH540
CODEN: ECOLAR
Special Features: Includes Advertising, Bibliographies, Charts, Illustrations, Book Reviews
Editor(s): Donald R Strong (Editor-in-Chief), J David Baldwin (Managing Editor)
E-Mail: eicesa@ucdavis.edu
URL: http://www.esajournals.org/loi/ecol
Description: Deals with all aspects of ecology, as well as discussions and essays that develop ecological theory.