Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices

Root Gorelick

Gorelick, R. 2006. Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices. – Ecography 29: 525–530.

Shannon's and Simpson's indices have been the most widely accepted measures of ecological diversity for the past fifty years, even though neither statistic accounts for species abundances across geographic locales ("patches"). An abundant species that is endemic to a single patch can be as much of a conservation concern as a rare cosmopolitan species. I extend Shannon's and Simpson's indices to simultaneously account for species richness and relative abundances – i.e. extend them to multispecies metacommunities – by making the inputs to each index a matrix, rather than a vector. The Shannon's index analogue of diversity is mutual entropy of species and patches divided by marginal entropy of the individual geographic patches. The Simpson's index analogue of diversity is a modification of mutual entropy, with the logarithm moved to the outside of the summation, divided by Simpson's index of the patches. Both indices are normalized for number of patches, with the result being inversely proportional to biodiversity. These methods can be extended to account for time-series of such matrices and average age-classes of each species within each patch, as well as provide a measure of spatial coherence of communities.

R. Gorelick (Root_Gorelick@carleton.ca) United States Environmental Protection Agency, National Center for Environmental Assessment, Washington, DC 20460, USA (present address: Dept of Biology, Carleton Univ., Ottawa, ON K1S 5B6, Canada).

Shannon's and Simpson's index have been the most widely used measures of biodiversity for the past half century. Both measures predate the publication of the oft cited papers by Shannon (1948) and Simpson (1949). What biologists call Shannon's index was, in fact, not the subject of Shannon's paper. His paper instead introduced mutual entropy, which is the crucial concept here. What biologists call Shannon's index was first developed by Boltzmann (1872), and is also referred to as entropy, marginal entropy, or marginal information. Simpson was probably the first person to introduce what we now call Simpson's index into ecology, but his work was also predated by others (Gini 1912). Shannon's and Simpson's indices are now regarded as being members of the same family of indices (Routledge 1979, Keylock 2005). Both Shannon's and Simpson's indices have truly stood the test of time and are still generally regarded as the premier measures of ecological diversity (Lande 1996, Magurran 2004, Buckland et al. 2005).

The crux of Shannon's index, Simpson's index, and virtually all other measures of ecological diversity is to construct a vector of probabilities, i.e. all elements of the vector are greater than or equal to zero and they sum to one. That is, the inputs are the abundances of each species (or other taxon; I only use "species" as a convenient handle). Let p_i be the relative abundance (probability) of species i, so that $\Sigma_{i=1}^{m} p_i = 1$. Shannon's index, Simpson's index, and many other diversity indices take the vector $(p_1, p_2, \ldots, p_m)$ as an input and provide a scalar output,

Accepted 25 March 2006

ISSN 0906-7590

i.e. Shannon's index $=-\Sigma_{i=1}^{m} p_i \cdot \log(p_i)$ and Simpson's index $=-\log\Sigma_{i=1}^{m} p_i^2$ (alternatively, sometimes Simpson's index is given as $(\Sigma_{i=1}^{m} p_i^2)^{-1}$or $1-\Sigma_{i=1}^{m} p_i^2)$.

Note that in all formulae herein, I have been ambiguous about the base of the logarithm. Shannon used base 2, which made sense in dealing with binary signals. Many ecologists (Keylock 2005) use the natural logarithm, base e, but it is not always obvious why. Fortunately, it does not matter which base is used, so long as one is consistent.

The above indices only measure α diversity, i.e. diversity within a single geographic locale. The relative abundance p_i is a composite of all individuals of that species throughout whatever geographic area is being studied. I shall refer to the smallest geographic unit of analysis here as a "patch", which is a geographic area in which we have compiled abundance data on multiple species. The amalgamation of all non-overlapping patches and the individuals of all censused species comprises a multispecies metacommunity (Hanski 1999). In order to measure β diversity, i.e. diversity amongst patches, we need to examine a matrix of data that gives the relative abundance of each species in each patch. The data matrix $[p_{ij}]$ contains m rows (one for each species) and n columns (one for each patch). With this formulation, p_i still reflects the cumulative relative abundance of that species throughout the entire geographic range, while p_j reflects the number of patches and also partly reflects relative area of patch j compared with the total area of all patches – insofar as relative abundance of a species in a patch is roughly proportional to area of the patch.

Indices with matrix inputs: Shannon's index

The modification of Shannon's index to handle matrix inputs is called mutual entropy (also known as mutual information, transinformation) and was Shannon's (1948) seminal contribution. Mutual entropy is a special case of Kullback-Leibler information (Kullback and Leibler 1951), which is occasionally encountered in ecology. Although mutual information occasionally appears in measures of ecological diversity (Colwell and Futuyma 1971, Ernoult et al. 2003, Cazelles 2004, Gorelick et al. 2004, Vaughan and Ormerod 2005), it still seems to be poorly known. Almost none of the hundreds of papers that cite Colwell and Futuyma's renowned 1971 paper on measuring niche overlap even mention mutual entropy. Nor does Magurran's (2004) seminal book on quantifying biodiversity mention mutual entropy. Regardless of the cryptic nature of mutual entropy in the ecological literature, the important thing is that it be normalized for the number of patches (Colwell and Futuyma 1971, Gorelick et al. 2004, Vaughan and Ormerod 2005). Thus, extending Shannon's index to biodiversity measured amongst patches, we should measure $\frac{I(i,j)}{H(i)}$, where mutual entropy is given by $I(i,j)=\Sigma_{\substack{i=1\\j=1}}^{m,n} p_{ij}\cdot\log(\frac{p_{ij}}{p_i \cdot p_j})$, marginal entropy of patches is $H(i)=-\Sigma_{i=1}^{m} p_i \cdot \log(p_i)$ and where $p_i=\Sigma_{j=1}^{m} p_{ij}$. Normalized mutual entropy is inversely proportional to biodiversity. Equivalently, $1-\frac{I(i,j)}{H(i)}$ is a directly proportional to biodiversity (Fig. 1).

Mutual entropy can also be written as $I(i,j)=H(i)-H(i|j)=H(j)-H(j|i)$, where i|j is the conditional probability of i given j. H(i|j) and H(j|i) are called conditional entropy (Cover and Thomas 1991). This makes the above proposed biodiversity conceptually simpler in that $1-\frac{I(i,j)}{H(i)}=\frac{H(i|j)}{H(i)}=\frac{H(j|i)}{H(j)}$.

The normalization, $\frac{I(i,j)}{H(i)}$, only works when there are more species than patches in a given multispecies metacommunity. Otherwise $\frac{I(i,j)}{H(i)}$ can be greater than one.

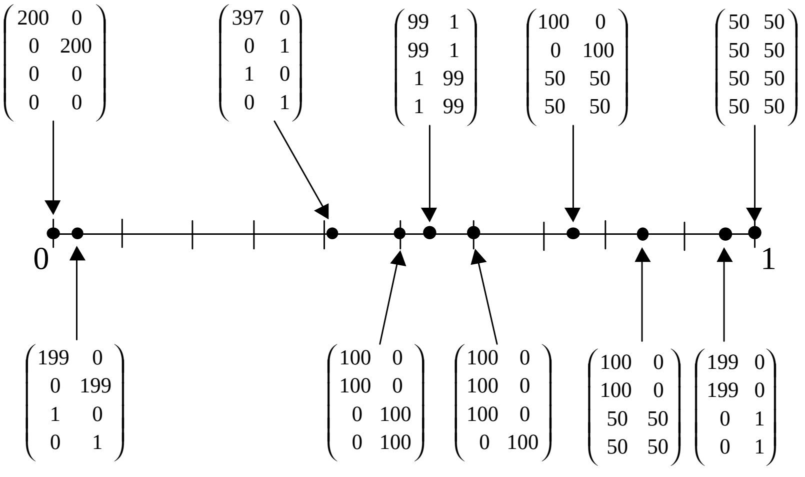

Fig. 1. $1-\frac{I(i,j)}{H(i)}$ for some matrices with four species (rows) in two patches (columns). The only exception is the upper left matrix, which only contains two species. Note the far left and far right matrices on the bottom row, which both have virtually only two species, yet radically different levels of biodiversity.

The biodiversity index $\frac{I(i,j)}{H(i)}$ is normalized for the number of patches. As an illustration, consider n equally common endemic species (assuming more species than patches). The input matrix here is $\frac{1}{n}$ times the identity matrix. This yields maximum possible values for both mutual entropy and marginal entropy of patches, both of which equal log(n). Hence $\frac{I(i,j)}{H(i)}$ equals its maximum possible value of one, which does not depend upon the number of patches.

Furthermore, $\frac{I(i,j)}{H(i)}$ is also roughly normalized for the area of patches because species abundance within a patch is roughly proportional to patch area. Admittedly, this proportionality is rough, but it does help normalize for possible variability in patch area. If this is not a sufficiently systematic normalization, then there exists a way to explicitly normalize for patch area. Weight the data matrix by multiplying each row by the area of its patch and then compute $\frac{I(i,j)}{H(i)}$ for this modified input matrix. This will yield a biodiversity statistic that is still normalized, between zero and one, and asymptotically F-distributed (Gorelick et al. 2004).

Shannon had originally constructed mutual entropy to simultaneously measure the amount of information transmitted from one individual to another and the information transmitted from the second individual back to the first… hence the moniker mutual information. This is exactly what we want in measuring biodiversity using mutual entropy. Here, the first individual consists of the ensemble of species and the second individual consists of the ensemble of patches. Assume that biodiversity is low. Also assume that someone has specified which species we are examining. Then, for each species, we should be able to predict which patch it is in. Low biodiversity means high endemism, which means that information about the species gets transmitted to information about the patch. Likewise, because each patch has so few species, information about which patch you are examining gets transmitted into information about which species you are most likely to find in that patch. This is bidirectional information transmission, exactly as envisioned by Shannon. Conversely, now assume that biodiversity is high. Then knowing which patch was selected from the ensemble transmits no information about which species are most likely to be found there because high biodiversity implies cosmopolitan species. Likewise, knowing which species you encounter transmits no information about which patch you are in because many species occupy each patch. Information transmission is between individuals and patches, and vice versa.

Indices with matrix inputs: Simpson's index

Although Shannon's index is firmly rooted in information theory and possesses many nice statistical properties, Simpson's index is often the diversity index of choice (Lande 1996) and can be re-written as $S(i) = -\log(\Sigma_{i=1}^{m} p_i \cdot p_i)$. In order to extend this formula to matrix inputs in a way that is analogous with Shannon's index, write $\varsigma(i,j) = -\log(\Sigma_{\substack{i=1\\j=1}}^{n,m} \frac{p_{ij}}{p_i} \cdot \frac{p_{ij}}{p_j})$. I have never seen this expression before in the literature. As with Shannon's index, normalize ς(i,j) to account for varying numbers and areas of patches between two different geographic areas. Thus the matrix modification of Simpson's index is $\frac{\varsigma(i,j)}{S(i)}$, which is inversely proportional to biodiversity. Thus, $1 - \frac{\varsigma(i,j)}{S(i)}$ is directly proportional to biodiversity.

Both $\frac{I(i,j)}{H(i)}$ and $\frac{\varsigma(i,j)}{S(i)}$ measure interdependencies amongst all rows and columns of the data matrix, i.e. amongst all patches and species. Both statistics produce numbers between zero and one. Ceteris paribus, the minimal value for biodiversity would therefore be $\frac{I(i,j)}{H(i)} = 0$ or $\frac{\varsigma}{S_j}$ for m evenly distributed and equally common species. Assuming that there are more species than patches, endemism for all species with equal numbers of species per patch yields the maximal value of one for biodiversity.

Both $\frac{I(i,j)}{H(i)}$ and $\frac{\varsigma(i,j)}{S(i)}$ increase as the number of species increases because we did not include H(j) nor S(j) in the denominators.

Including a third (temporal) component

Explicitly including time requires a new data structure: a three-dimensional array of n species, m patches, and τ times. Mutual entropy and its Simpson-like analogue can be extended to such arrays (Cazelles 2004): $I(i,j.t) = \Sigma_{\substack{i=1\\j=1\\t=1}}^{m,n,\tau} p_{ijt} \cdot \log(\frac{p_{ijt}}{p_i \cdot p_j \cdot p_t})$ and $\varsigma(i,j,t) = -\log(\Sigma_{\substack{i=1\\j=1\\t=1}}^{n,m,\tau} \frac{p_{ijt}}{p_i} \cdot \frac{p_{ijt}}{p_j} \cdot \frac{p_{ijt}}{p_t})$.

As before, normalize for the number (and area) of patches by computing $\frac{I(i,j,t)}{H(i)}$ and $\frac{\varsigma(i,j,t)}{S(i)}$. However, if the number of time increments varies, then simultaneously normalize for both patches and time by computing $\frac{I(i,j,t)}{I(i,j)}$ and $\frac{\varsigma(i,j,t)}{\varsigma(i,j)}$. This methodology has a much more solid theoretical foundation than Buckland et al.'s (2005) modified (vector-based) Simpson's index for dealing with a temporal variation in ecological diversity. Buckland et al.'s methodology also only deals with α diversity, and not β diversity.

Although this temporal component can be most readily conceptualized as a time series, it can also be reconceptualized to be the average-age class of a patch. This can be useful for incorporating age-class structure of plants communities into analyses of biodiversity or age polyethism of eusocial insect colonies.

Caveats

One should not immediately jump to the normative conclusion that a larger biodiversity index is better than a smaller one. For some conservation purposes, such as preclusion of pandemics, endemism (or at least spotty distributions) may be preferable. The opposite may be true if the sites of endemism are being destroyed by human activity. There may be instances in between, where intermediate evenness for a species may be optimal, such as seems to be the case with many temperate orchids that are obligately insect-pollinated and hence individuals cannot be too widely separated from one another.

As is the case with conventional Shannon's and Simpson's indices (Lande 1996), it is not obvious which of these two new biodiversity indices is preferable. The modified Simpson's index is more sensitive to slight changes in species richness and evenness if most species are cosmopolitan. On the other hand, the modified Shannon's index is more sensitive to slight changes if most species are endemic. This may indicate that we should let data values determine which index to use, but this is contrary to the maxim that statistics should be chosen based on theory. In that light, my inclination would be to use the ratio of mutual to marginal entropy solely because it is firmly grounded in information theory.

Thus far, I have described using modified Shannon's and Simpson's indices with data on species abundance from each patch. These methods can also be used with species presence/absence data for each patch. However, each column of the matrix (representing each species) should be weighted by the relative rarity of that species, if this data is known (Gorelick et al. 2004). This weighting of columns can be done in conjunction with weighting of rows for patch area (i.e. weight both the rows and columns), and then $\frac{I(i,j)}{H(i)}$ or $\frac{\varsigma(i,j)}{S(i)}$ can be computed based on this new input matrix. This weighting provides an extension of the method advocated by Vaughan and Ormerod (2005).

The normalizations $\frac{I(i,j)}{H(i)}$ and $\frac{\varsigma(i,j)}{S(i)}$ yield maximum values of one if and only if there are at least as many species as patches, i.e. at least as many rows as columns in the matrix. If perchance there are more patches than species, which is unlikely for most analyses, then the largest value that $\frac{I(i,j)}{H(i)}$ or $\frac{\varsigma(i,j)}{S(i)}$ can attain is the logarithm of the ratio of species to patches and consequently $1-\frac{I(i,j)}{H(i)}$ and $1-\frac{\varsigma(i,j)}{S(i)}$ could be negative. For example, say there is one patch with n species, in which case $\frac{I(i,j)}{H(i)}$ and $\frac{\varsigma(i,j)}{S(i)}$ can be as large as log(n). We could instead normalize by dividing by H(j), S(j), $\sqrt{H(i)\cdot H(j)}$, or $\sqrt{S(i)\cdot S(j)}$ (Gorelick et al. 2004). This would be mathematically convenient, but would be biologically unrealistic. We want to normalize for things that are in our experimental design, i.e. the independent variable, such as the number of patches in which data was collected. We do not want to normalize for those things that we care most about, i.e. for the dependent variables, such as the number and relative abundances of species found in each patch. Thus, throughout this paper, I have implicitly assumed that there are more species than patches in any given multispecies metacommunity.

One can never hope to capture all aspects of species richness, evenness, and relative abundances in a single scalar statistic. By going from a matrix to a scalar, we lose some valuable information. For example, the diversity indices for the following data sets are identical and in fact maximal: $\begin{pmatrix}5 & 5\\5 & 5\end{pmatrix}$ and $\begin{pmatrix}10 & 0\\10 & 0\end{pmatrix}$. The first data set is five individuals from each species (row) in each patch (column). The second data set is ten individuals from each species in the first patch and no individuals of either species in the second patch. Intuitively, we would want to ascribe lower diversity to the second data set, but the mathematics force us to effectively ignore the existence of the second patch which has none of the censused species. Another difficulty is exemplified by $\begin{pmatrix}500 & 500\\500 & 500\end{pmatrix}$ and $\begin{pmatrix}500 & 0\\500 & 0\end{pmatrix}$, which have the same diversity index values as the previous two examples because these indices only account for relative abundances, and not absolute abundances. In all four examples $1-\frac{I(i,j)}{H(i)}=1-\frac{\varsigma(i,j)}{S(i)}=0$. The good news is that minimal diversity is ascribed to the data set $\begin{pmatrix}10 & 0\\0 & 10\end{pmatrix}$, in which each species is endemic to a separate patch: $1-\frac{I(i,j)}{H(i)}=1-\frac{\varsigma(i,j)}{S(i)}=0$. Furthermore, the data sets $\begin{pmatrix}500 & 1\\1 & 500\end{pmatrix}$ and $\begin{pmatrix}500 & 0\\0 & 1\end{pmatrix}$ have diversity values only slightly greater than zero, which seems sensible. My goal in constructing the matrix analogues of Shannon's and Simpson's indices was to capture as much detail about species richness and relative abundance as possible in a single number between zero and one, maintaining what most people would consider a reasonable rank order amongst a nominal set of matrices of biodiversity data (Fig. 1).

Another problem with the matrix analogues of Shannon's and Simpson's indices is that they do not take into account spatial distributions of patches nor taxonomic relatedness. We may want to weight isolated island populations more heavily than interconnected mainland populations (Galapagos vs mainland Ecuador). We may want to weight taxonomically isolated taxa more heavily than others (monotremes vs eutherians). Looking towards the future, maybe it will be possible to combine Rao's (1982) quadratic Simpson's index with the modified Shannon's and Simpson's indices, $\frac{I(i,j)}{H(i)}$ and $\frac{\varsigma(i,j)}{S(i)}$, and thereby also include weights regarding geographic and taxonomic isolation. However, using accumulation curves of the modified Shannon's and Simpson's indices, we can at least discern how much geographic isolation matters.

Spatial coherence of communities

Above, when computing the modified Shannon's or Simpson's index, we did so for the entire multispecies metacommunity. In order to measure some facet of spatial coherence of communities, instead we need to compute an accumulation curve of these indices, with the last data point being the modified Shannon's or Simpson's index over the entire multispecies metacommunity. I outline the algorithm in the next paragraph.

Establish a lattice of points across the multispecies metacommunity; these points can be centered on the patches. Consider all patches that are within a neighborhood of diameter x, where the neighborhoods are centered on points of the lattice. Compute the modified Shannon's or Simpson's for each neighborhood of diameter x and then compute the mean over all such neighborhoods of diameter x. This yields a point on the accumulation curve. Increment x from zero to the diameter of the entire multispecies metacommunity. Then plot the mean for each value of x.

I am as befuddled as many by what constitutes an ecological community. However, the accumulation curve allows us to specify when a community grows so large (geographically) that it loses spatial coherence. Where the accumulation curve takes a discreet jump, the community has lost spatial coherence.

This methodology can also be used to determine when relatedness in kin groups becomes so tenuous that division of labor is effectively lost. For this application, normalized mutual entropy measures division of labor of tasks amongst individuals (Gorelick et al. 2004), in which relatedness of individuals replaces spatial distance of patches. A sensible measure of relatedness here would be the reciprocal (r) of the average fraction of genes that two individuals have in common (e.g. for full siblings r = 2; for half-siblings r = 4). Where the accumulation curve takes a discreet jump, the kinship group has lost its coherence.

Concluding remarks

The modifications of Shannon's and Simpson's indices to handle matrix inputs allows us to more systematically quantify diversity when there exists species abundance data from multiple patches. When there are more species than patches, these methods allow us to amalgamate species richness and relative abundance data into a single scalar statistic that is normalized for the number and area of patches. These methods will also allow us to determine which multispecies metacommunity has greater diversity and whether the difference in biodiversity between two metacommunities is statistically significant. For example, we could discern whether a geographic region has incurred a statistically significant change in diversity over time, regardless of changes in the number of patches monitored. Neither the matrix analogues of Shannon's nor Simpson's indices are mentioned in Anne Magurran's seminal text (2004).

The modified Shannon's index (mutual entropy divided by marginal entropy) has much more theoretical grounding than the modified Simpson's index. Furthermore, it gives credit for Shannon's real achievement. Therefore, I would recommend the modified Shannon diversity index. Nonetheless, there are many criteria by which a good diversity index should be judged (Lande 1996) and therefore good reason to remember that a modified Simpson's index for matrix inputs now exists.

The matrix-based diversity indices discussed herein will never answer all questions about diversity. Neither a larger nor a smaller diversity value is necessarily better. The biology still needs to dictate interpretation of these indices. These indices also do not allow us to discern effects of different absolute abundances or lack of all censused species from a given patch. Nonetheless, these matrix analogues of Shannon's and Simpson's indices will go a long way towards testing hypotheses about diversity within and between multispecies metacommunities and towards quantifying what constitutes a coherent community or kin group.

Acknowledgements – Thanks to Denice Shaw, Anne Marsh, Sue Bertram, László Orlóci, Jennifer Fewell, Tom Taylor, and Peter Killeen. The research described in this paper has been partly funded by the United States Environmental Protection Agency (EPA) through cooperative agreement number CR828392 with the American Association for the Advancement of Science (AAAS). The views expressed herein may not necessarily reflect the views of EPA or AAAS, and no official endorsement should be inferred.

References

Boltzmann, L. 1872. Weitere studien über das wärmegleichgewicht unter gasmolekälen. – S. K. Akad. Wiss. Wein 66: 275–370.

Buckland, S. T. et al. 2005. Monitoring change in biodiversity through composite indices. – Phil. Trans. R. Soc. B 360: 243–254.

Cazelles, B. 2004. Symbolic dynamics for identifying similarity between rhythms of ecological time series. – Ecol. Lett. 7: 755–763.

Colwell, R. K. and Futuyma, D. J. 1971. On the measurement of niche breadth and overlap. – Ecology 52: 567–576.

Cover, T. M. and Thomas, J. A. 1991. Elements of information theory. – Wiley.

Ernoult, A. et al. 2003. Patterns of organisation in changing landscapes: implications for the management of biodiversity. – Landscape Ecol. 18: 239–251.

Gini, C. 1912. Variabilité mutabilitá. – Studi Economicoaguridici della facotta di Giurisprudenza dell Univ. di Cagliari III, Parte II [cited by Rao 1982].

Gorelick, R. et al. 2004. Normalized mutual entropy in biology: quantifying division of labor. – Am. Nat. 164: 677–682.

Hanski, I. 1999. Metapopulation ecology. – Oxford Univ. Press.

Keylock, C. J. 2005. Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy. – Oikos 109: 203–207.

Kullback, S. and Leibler, R. A. 1951. On information and sufficiency. – Ann. Math Stat. 22: 79–86.

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

Magurran, A. E. 2004. Measuring biological diversity. – Blackwell.

Rao, C. R. 1982. Diversity and dissimilarity coefficients: a unified approach. – Theor. Popul. Biol. 21: 24–43.

Routledge, R. D. 1979. Diversity indexes: which ones are admissible? – J. Theor. Biol. 76: 503–515.

Shannon, C. E. 1948. A mathematical theory of communication. – Bell Syst. Tech. J. 27: 379–423, 623–656.

Simpson, E. H. 1949. Measurement of diversity. – Nature 163: 688.

Vaughan, I. P. and Ormerod, S. J. 2005. The continuing challenge of testing species distribution models. – J. Appl. Ecol. 42: 720–730.

Subject Editor: Andrew Liebhold.