biology letters
Community ecology

Biol. Lett. (2007) **3**, 20–22
doi:10.1098/rsbl.2006.0553
Published online 17 October 2006

A multiple-site similarity measure

Ola H. Diserud* and Frode Ødegaard

Norwegian Institute of Nature Research (NINA), 7485 Trondheim, Norway
**Author for correspondence (ola.diserud@nina.no).*

Similarity measures are among the most intuitive and common measures for comparing two or more sites, or samples, with respect to their species overlap. A restriction of similarity measures is that they are limited to pairwise comparisons even in a multiple-site study. This work presents a multiple-site similarity measure that makes use of information on species shared by more than two sites and avoids the problem of covariance between pairwise similarities in a multiple-site study. Further, we show that our multiple-site similarity measure is related to β-diversity measures such as Whittaker's β-diversity. Similarity measures can also be used as descriptors of effective specialization of insects to host species by measuring similarity from host observations. Finally, we show that multiple-site similarity and host specificity are two sides of the same coin.

Keywords: β-diversity; host specificity; similarity; species composition; spatial distribution

1. INTRODUCTION

Understanding spatial patterns of species diversity is a crucial topic in ecology and conservation biology, for instance, when predicting species richness from local to regional scales (MacArthur 1965; Cornell 1985; Ricklefs 1987; Thomas 1990; Gering & Crist 2002). One approach to partition species diversity is to define α-diversity as within-habitat diversity, β-diversity as a measure of between-habitat diversity (within landscape) and γ-diversity as within-landscape diversity (Magurran 2004). The α- and γ-diversities measure inventory diversity (e.g. number of species), whereas the β-diversity describes differentiation diversity (the change in species composition between two or more habitats; Whittaker 1960; Koleff *et al.* 2003; Magurran 2004). There exists a wide variety of methods for measuring β-diversity, among which similarity measures are the simplest and the most commonly used for calculating β-diversity from abundance or presence/absence data (Wolda 1981; Koleff *et al.* 2003).

Most evaluations of similarity between multiple sites are based on the average, or plots, of pairwise similarities (e.g. Lennon *et al.* 2001; Vellend 2001; Condit *et al.* 2002; Basset *et al.* 2004). Information on the identity of species shared across more than two sites is not preserved, so average similarity across all sites does not tell us to what extent there is a change in shared species between pairs. This approach also ignores the problem of covariance between similarities, since some pairs must share the same site (Ødegaard *et al.* 2005). Pairwise comparison of neighbouring sites will suffice if the goal is to look at how species composition changes along a physical or environmental gradient, but if we view our sites as a random collection of samples from a larger region, such as an island or a landscape, a multiple-site similarity measure is required.

2. MATERIAL AND METHODS

(a) *Multiple-site similarity*

All similarity indices represent variations over three parameters: species composition in each of two sites and the species shared between the two sites (Novotny & Weiblen 2005). The widely used Sørensen similarity index (Magurran 2004) measures similarity in species composition for two sites, A and B, by the equation

$$C_S = \frac{2ab}{a+b}, \tag{2.1}$$

where a is the number of species found in site A; b is the number of species in site B and ab is the number of species shared by the two sites.

For studies where more than two sites are evaluated, the overall similarity is calculated as the average of the pairwise similarities. As an illustration of the shortcomings of such an approach, we can look at two hypothetical cases. Let case 1 have three sites with four species in each: $[(s_1,s_2,s_3,s_4),\ (s_1,s_2,s_5,s_6),\ (s_1,s_2,s_7,s_8)]$, where s_i is species number i. The similarity is the same for all pairs of sites, $C_S = 4/8 = 1/2$, with average similarity also equal to 1/2. Case 2 also has three sites with four species in each, but with a different distribution: $[(s_1,s_2,s_3,s_4),\ (s_1,s_2,s_5,s_6),\ (s_3,s_4,s_5,s_6)]$. The similarity is still $C_S = 1/2$ for all pairs, so the Sørensen similarity index does not 'see' the difference in species composition between the two cases. Using traditional similarity measures on assemblages with more than two sites, we will never do more than compare two sites at a time and thereby ignore 'higher order similarities'.

We will now suggest a multiple-site similarity measure and start with the situation where we have three sites in a study. We follow the notation from equation (2.1), with a, b and c the numbers of species found in sites A, B and C, respectively, and ab the number of species shared by sites A and B, etc., until abc which is the number of species found in all three sites. Extending the approach of the Sørensen similarity index, a foundation for a three-site similarity measure can be

$$\frac{ab+ac+bc-abc}{a+b+c}. \tag{2.2}$$

The numerator gives the number of species counts exceeding the first; and the denominator gives the sum of species counts over all the sites. This expression will equal 2/3 if all species are shared by all sites, since a species can contribute at most two times in the numerator and three times in the denominator. The three-site similarity measure should therefore be $(3/2)((ab+ac+bc-abc)/(a+b+c))$ in order to be in the range 0–1, with 1 indicating complete similarity. The general multiple-site similarity measure for T sites can be formulated in the same manner

$$C_S^T = \frac{T}{T-1}\left(\frac{\sum_{i<j} a_{ij} - \sum_{i<j<k} a_{ijk} + \sum_{i<j<k<l} a_{ijkl} - \cdots}{\sum_i a_i}\right), \tag{2.3}$$

where a_i is the number of species in site A_i, $i=1,\ldots,T$; a_{ij} is the number of species shared by sites A_i and A_j; and a_{ikj} is the number of species shared by sites A_i, A_j and A_k, etc. With $T=2$, we are back at the definition of the Sørensen similarity index (equation (2.1)). The total number of species in the T sites can, by the inclusion–exclusion principle, be written as $S_T = \sum_i a_i - \sum_{i<j} a_{ij} + \sum_{i<j<k} a_{ijk} - \sum_{i<j<k<l} a_{ijkl} + \cdots$, simplifying the notation of our multiple-site similarity measure to

$$C_S^T = \frac{T}{T-1}\left(1 - \frac{S_T}{\sum_i a_i}\right). \tag{2.4}$$

For the two hypothetical cases discussed earlier, we get $C_S^T = 1/2$ for case 1 and $C_S^T = 3/4$ for case 2. Our multiple-site similarity measure evaluates the sites in case 2 as more similar than the sites in case 1, which is in agreement with the assumption that evenness in the number of site observations for the species should be valued more, i.e. the similarity measure increases with a more even distribution of site observations. For case 2, we also obtain a lower total number of species (γ-diversity), indicating a lower species turnover, hence a higher similarity.

Received 7 September 2006
Accepted 27 September 2006

Both cases 1 and 2 have covariance 0 between pairwise similarities, since all similarities are equal to 1/2. With $T=3$, all pairs of similarities must necessarily be dependent, since they all share one site. The effect of covariance between pairwise similarities on average similarity will depend on the sign and magnitude of the covariance, as well as the proportion of independent pairwise similarities (Ødegaard *et al.* 2005). To illustrate one possible effect of covariance, let case 3 also have three sites with four species in each: $[(s_1,s_2,s_3,s_4), (s_1,s_2,s_3,s_4), (s_4,s_5,s_6,s_7)]$. Here, the covariance between pairwise similarities is negative. The average similarity is still 1/2, but now $C_S^T = 5/8$.

(b) *Multiple-site similarity versus β-diversity and host specificity*

β-diversity is essentially also a measure of how similar sites are in terms of the variety of species found in them. A high similarity indicates that there are few species differences between sites, yielding low β-diversity values. One of the most straightforward measures of β-diversity is Whittaker's (1972) measure, $\beta_W = S_T/\bar{S}_{within}$, where S_T is the total number of species; and $\bar{S}_{within}$ is the average species richness for the T sites. The link between Sørensen's similarity measure for two sites and β-diversity measures is well known (Koleff *et al.* 2003). The relation between our multiple-site similarity and Whittaker's β_W is simply

$$C_S^T = \frac{T-\beta_W}{T-1}. \tag{2.5}$$

If all sites contain the same species, both C_S^T and β_W will equal 1. If no sites share species, $C_S^T = 0$ and $\beta_W = T$, indicating that the total number of species S_T is just the product $T \times \bar{S}_{within} = \sum_i a_i$.

If, instead of species-sites data, we are studying host observations of, for example, phytophagous insect species on host plant species, the comparison of species compositions on different host plants can be performed by both similarity and host-specificity measures. The host specificity calculated from trophic interactions is defined as $F_T = S_T/(\bar{S}_T \times T)$ (Ødegaard *et al.* 2000; Novotny *et al.* 2002), where S_T is the total number of insect species found on T host plant species; $\bar{S}_T$ is the average number of insect species associated with each host plant species; and T is the number of host plant species in the study. The product $\bar{S}_T \times T$ is thereby the total number of host observations. Host specificity views all host plant species simultaneously and can be considered a 'multiple host dissimilarity measure'. The link between our multiple-site similarity measure and host specificity is

$$C_S^T = \frac{T}{T-1}(1-F_T). \tag{2.6}$$

Note also that $F_T = \beta_W/T$. If all species are shared by all hosts, the host specificity is $1/T$ and the multiple-site similarity equals 1. With no species overlap, host specificity equals 1 and similarity becomes 0. If we regard our first two hypothetical cases as host observations of insect species on three different host species, we get host specificities 2/3 and 1/2, respectively. Case 1 has more monophagous species; therefore, it should also have higher host specificity.

3. DISCUSSION

The proposed similarity measure for multiple sites C_S^T (equation (2.4)) provides a more relevant index for species' spatial distribution. Instead of calculating the average over a set of dependent pairwise similarities, we make use of information on the identity of species shared across more than two sites. For a given number of sites T, C_S^T decreases with increasing number of 'rare' species, i.e. species observed in only one or a few sites. Conversely, C_S^T increases with increasing number of species observed in several sites. The multiple-site similarity measure can be regarded as a linear function of Whitaker's β-diversity and host specificity, thereby inheriting their statistical properties.

Similarity measures are generally believed to be underestimates (e.g. Lande 1996), i.e. true similarity between sites is biased downwards when estimated from random samples. This is often illustrated by simulating random samples from the same community with true similarity equal to 1. But for situations where true similarity is less than 1, the direction of the possible bias will depend on species abundance distributions within sites as well as species turnover between sites, especially whether rare species are shared between sites or not. For example, assume a larger region where some species are abundant at all sites, and for each site, some additional rare species are present at this site only. Small sample sizes from each site may include the dominant species only, thereby overestimating similarity between sites even if species richness at each site and in total are underestimated. In tropical forests, calculations of similarity tend to be underestimated owing to the dominance of rare species in the species pool (Mawdsley 1996; Stork 1997; Ødegaard 2006) and small sample sizes (Chao *et al.* 2000).

The multiple-site similarity measure has been based on presence/absence data, but this approach can be modified to handle abundance data as well. The abundance-based similarity measure used as the foundation for the modification should be chosen so that the effects of small sample sizes and varying sampling efforts are minimized (e.g. Chao *et al.* 2006).

In many studies, applying pairwise similarities between multiple sites may be an appropriate approach, typically when we want to evaluate species turnover along an environmental gradient. However, when the sites are viewed as a random set of observations from a region, evaluating overall similarity from the average of the pairwise similarities can be misleading. The pairwise similarities are not all independent, since each site is included in $T-1$ pairs, and they ignore information on species shared among more than two sites. As our multiple-site similarity measure C_S^T solves these matters, it is more consistent with multiple-site β-diversity. We have shown that the parameters multiple-site similarity C_S^T, Whittaker's β-diversity (Whittaker 1972) and host specificity F_T (considering the hosts as sites) measure the same characteristics of community structure, with simple transformations based on the number of sites, or hosts, T.

Basset, Y., Mavoungou, J. F., Mikissa, J. B., Missa, O., Miller, S. E., Kitching, R. L. & Alonso, A. 2004 Discriminatory power of different arthropod data sets for the biological monitoring of anthropogenic disturbance in tropical forests. *Biodivers. Conserv.* **13**, 709–732. (doi:10.1023/B:BIOC.0000011722.44714.a4)

Chao, A., Hwang, W.-H., Chen, Y.-C. & Kuo, C. Y. 2000 Estimating the number of shared species in two communities. *Stat. Sinica.* **10**, 227–246.

Chao, A., Chazdon, R. L., Colwell, R. K. & Shen, T.-J. 2006 Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62**, 361–371. (doi:10.1111/j.1541-0420.2005.00489.x)

Condit, R. *et al.* 2002 Beta-diversity in tropical forest trees. *Science* **295**, 666–669. (doi:10.1126/science.1066854)

Cornell, H. V. 1985 Local and regional species richness of cynipine gall wasps on California oaks. *Ecology* **66**, 1247–1260. (doi:10.2307/1939178)

Gering, J. C. & Crist, T. O. 2002 The alpha–beta–regional relationship: providing new insight into local–regional patterns of species richness and scale dependence of diversity components. *Ecol. Lett.* **5**, 433–444. (doi:10.1046/j.1461-0248.2002.00335.x)

Koleff, P., Gaston, K. J. & Lennon, J. J. 2003 Measuring beta diversity for presence–absence data. *J. Anim. Ecol.* **72**, 367–382. (doi:10.1046/j.1365-2656.2003.00710.x)

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

Lennon, J. J., Koleff, P., Greenwood, J. J. D. & Gaston, K. J. 2001 The geographical structure of British bird distributions: diversity, spatial turnover and scale. *J. Anim. Ecol.* **70**, 966–979. (doi:10.1046/j.0021-8790.2001.00563.x)

MacArthur, R. H. 1965 Patterns of species diversity. *Biol. Rev.* **40**, 510–533.

Magurran, A. 2004 *Measuring biological diversity*. Oxford, UK: Blackwell Publishing.

Mawdsley, N. 1996 The theory and practice of estimating regional species richness from local samples. In *Tropical rainforest research—current issues* (ed. D. S. Edwards, W. E. Booth & S. C. Choy), pp. 193–213. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Novotny, V. & Weiblen, G. D. 2005 From communities to continents: beta diversity of herbivorous insects. *Ann. Zool. Fennici* **42**, 463–475.

Novotny, V., Basset, Y., Miller, S. E., Weiblen, G. D., Bremer, B., Cizek, L. & Drozd, P. 2002 Low host specificity of herbivorous insects in a tropical forest. *Nature* **416**, 841–844. (doi:10.1038/416841a)

Ødegaard, F. 2006 Host specificity, alpha- and beta-diversity of phytophagous beetles in two tropical forests in Panama. *Biodivers. Conserv.* **15**, 83–105. (doi:10.1007/s10531-004-3106-5)

Ødegaard, F., Diserud, O. H., Engen, S. & Aagaard, K. 2000 The magnitude of local host specificity for phytophagous insects and its implications for estimates of global species richness. *Conserv. Biol.* **14**, 1182–1186. (doi:10.1046/j.1523-1739.2000.99393.x)

Ødegaard, F., Diserud, O. H. & Østbye, K. 2005 The importance of plant relatedness for host utilization among phytophagous insects. *Ecol. Lett.* **8**, 612–617. (doi:10.1111/j.1461-0248.2005.00758.x)

Ricklefs, R. E. 1987 Community diversity: relative roles of local and regional processes. *Science* **235**, 167–171. (doi:10.1126/science.235.4785.167)

Stork, N. E. 1997 Measuring global biodiversity and its decline. In *Biodiversity II* (ed. M. L. Reaka-Kudla, D. E. Wilson & E. O. Wilson), pp. 41–68. Washington, DC: Joseph Henry Press.

Thomas, C. D. 1990 Fewer species. *Nature* **347**, 237. (doi:10.1038/347237a0)

Vellend, M. 2001 Do commonly used indices of beta-diversity measure species turnover? *J. Veg. Sci.* **12**, 545–552.

Whittaker, R. H. 1960 Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol. Monogr.* **30**, 279–338. (doi:10.2307/1943563)

Whittaker, R. H. 1972 Evolution and measurement of species diversity. *Taxon* **21**, 213–251. (doi:10.2307/1218190)

Wolda, H. 1981 Similarity indices, sample size and diversity. *Oecologia* **50**, 296–302. (doi:10.1007/BF00344966)