Community Ecology 20(1): 104-106, 2019
1585-8553 © Akadémiai Kiadó, Budapest
DOI: 10.1556/168.2019.20.1.11

FORUM

Can we trust the chord (and the Hellinger) distance?

C. Ricotta

Department of Environmental Biology, University of Rome 'La Sapienza', Piazzale Aldo Moro 5, 00185 Rome, Italy.
E-mail: carlo.ricotta@uniroma1.it

Keywords: Dalton's principle of transfers, Data transformation, Euclidean distance, Monotonicity to changes in species abundances, Species relative abundances.

Abstract: To avoid the problems associated with the Euclidean distance for the calculation of plot-to-plot dissimilarity, a variety of alternative measures have been proposed. Among them, the chord and the Hellinger distances are both obtained by first transforming separately the species abundances in each plot vector and then by calculating the Euclidean distance on the chord-transformed or the Hellinger-transformed data. However, although both measures are routinely used by ecologists as substitutes for the Euclidean distance, they have very different properties. In this paper, using a modified version of Dalton's principle of transfers, I will show that, unlike the Euclidean distance, the chord and the Hellinger distances are not monotonic to changes in absolute abundances. Therefore, they are not interchangeable with the Euclidean distance. The moral of this story is that although dissimilarity may appear an intuitively simple concept, the properties of even the best-known indices are not fully understood. Therefore, a clear understanding of old and new coefficients is needed to evaluate their ability to highlight relevant aspects of compositional dissimilarity among plots.

Introduction

> *Do not let your left hand know what your right hand does* (Matthew 6:3)

Dissimilarity measures between pairs of plots are routinely used in plant ecology to explore the effects of ecological and evolutionary mechanisms that drive community assembly. In this context, a well-known problem of the Euclidean distance is that rare species may have an excessive impact on the analysis. This effect can lead to an anomalous situation, which is usually known as the Orlóci paradox: two plots which have no species in common may appear more similar than two plots with the same species list. For details, see Orlóci (1978).

To overcome this problem, several authors (e.g., Orlóci 1967, 1978, Legendre and Gallagher 2001) have proposed to replace the Euclidean distance with the chord or the Hellinger distance. Given two plots *U* and *V*, the chord distance (Orlóci 1967), is obtained by scaling separately each plot vector according to

$$x'_{Uj} = x_{Uj} \Big/ \sqrt{\sum_{j=1}^{S} x_{Uj}^2}$$

and then by calculating the Euclidean distance on the chord-transformed data:

$$CH_{UV} = \sqrt{\sum_{j=1}^{S} \left(x'_{Uj} - x'_{Vj}\right)^2} =$$

$$= \sqrt{\sum_{j=1}^{S} \left(\frac{x_{Uj}}{\sqrt{\sum_{j=1}^{S} x_{Uj}^2}} - \frac{x_{Vj}}{\sqrt{\sum_{j=1}^{S} x_{Vj}^2}} \right)^2}$$

where x_{Uj} and x_{Vj} are the (absolute) abundances of species *j* in plots *U* and *V*, and *S* is the total number of species in both plots. From a geometrical viewpoint, this scaling operation implies that the plot vectors are projected onto the surface of a hypersphere of unit radius and then the Euclidean distance of the chord between both plots is measured (Orlóci 1978).

An additional dissimilarity coefficient that can be obtained by calculating the Euclidean distance from the scaled species abundances in each vector is the Hellinger distance (Legendre and Gallagher 2001). In this case, the species abundances x_{Uj} are first Hellinger-transformed according to

$$x''_{Uj} = \sqrt{x_{Uj} \Big/ \sum_{j=1}^{S} x_{Uj}} \; .$$

Then, the Hellinger distance is obtained by calculating the Euclidean distance on the Hellinger-transformed data:

$$HD_{UV} = \sqrt{\sum_{j=1}^{S} \left(x''_{Uj} - x''_{Vj} \right)^2} =$$

$$= \sqrt{\sum_{j=1}^{S} \left(\sqrt{\frac{x_{Uj}}{\sum_{j=1}^{S} x_{Uj}}} - \sqrt{\frac{x_{Vj}}{\sum_{j=1}^{S} x_{Vj}}} \right)^2}$$

Both measures have been extensively used in exploratory data analysis as they allow ecologists to apply Euclidean-based ordination methods, such as principal component analysis on transformed data "while circumventing the problems associated with the Euclidean distance" (Legendre and Gallagher 2001, p. 271).

However, the chord and the Hellinger distances do not measure plot-to-plot dissimilarity in the same data space as the Euclidean distance. Therefore, contrary to common belief, they are not interchangeable. For example, for the Euclidean distance to be zero, the two plots need to have the same species in equal quantity. On the contrary, being based on transformed data, the chord and the Hellinger distance between two plots is zero if the *proportion* of species abundances is the same in both plots. For the same reason, unlike the Euclidean distance, both measures are not monotonic to changes in abundance in raw data space. Therefore, their applicability for summarizing plot-to-plot dissimilarity needs to be carefully evaluated based on the specific problem at hand.

A new principle of transfers for dissimilarity coefficients

Many authors have proposed a variety of properties for dissimilarity coefficients to reasonably behave in ecological research (e.g., Koleff et al. 2003, Legendre and De Cáceres 2013, Ricotta and Podani 2017). One of these properties is monotonicity to changes in abundance. This is a relevant property for dissimilarity coefficients, which was defined by Legendre and De Cáceres (2013, Appendix S3) as "a necessary condition for any coefficient used to study species assemblages". In its very essence, this property requires that decreasing the difference in abundances of one of several species between two sites decreases their dissimilarity. Legendre and De Cáceres (2013) proposed to test the monotonicity property using ordered comparison case series (OCCAS), a method developed by Hajdu (1981), which consists of progressively decreasing the difference in abundances of one of several species between two sites by changing the abundance values in only one of the two sites, with the other site remaining unchanged (see also Gower and Legendre 1986).

Let x_{Uj} and x_{Vj} be the abundances of species j in plots U and V, respectively with $x_{Uj} \leq x_{Vj}$. All else being equal, the dissimilarity among U and V decreases if the quantity δ is added to x_{Uj} so long as this increment does not reverse the ranking of the two abundances $x_{Uj} + \delta \leq x_{Vj}$.

More recently, Ricotta and Podani (2017) proposed to verify the monotonicity to changes in abundance with a modified version of Dalton's (1920) principle of transfers, which has been commonly used in econometrics for summarizing the inequality in the distribution of household incomes and in ecology for summarizing the species evenness within single plots. Given the species abundances x_{Uj} and x_{Vj} with $x_{Uj} \leq x_{Vj}$, all else being equal, the dissimilarity among U and V is decreased if the quantity δ is *transferred* from x_{Vj} to x_{Uj} so long as that transfer does not reverse the ranking of the two abundances $x_{Uj} + \delta \leq x_{Vj} - \delta$ (Ricotta and Podani 2017). Put in simpler terms, the principle of transfers states that dissimilarity is decreased when abundance is transferred from a species in one plot to the same species in another plot in which that species is less abundant.

Although both approaches have the effect of decreasing the difference in the abundances of one or more species between two sites, the underlying ecological mechanisms are slightly different. The first approach would correspond to an increment of resources in one plot, whereas the second approach implies the transfer of organisms from one site to another as it is often the case in metacommunity dynamics and source-sink population processes (Pulliam 1988, Mouquet and Loreau 2003).

Using the principle of transfers, it is easily shown that the Euclidean distance is monotonic to changes in abundance. We can also use the same method to show that a coefficient for which the dissimilarity between two plots is zero if the proportion of species abundances is identical in both plots cannot be monotonic to changes in absolute abundances. A simple algorithm is sufficient to find a counterexample:

1. Given a species assemblage U with abundances x_{Uj} ($j = 1, 2, ..., S$), multiply all species abundances in U by a constant $k > 1$. This operation gives rise to a new assemblage V with species abundances $x_{Vj} = kx_{Uj}$. Since the proportion of all species abundances is identical in both assemblages, we have that the chord (or Hellinger) distance between U and V is equal to $CH_{UV} = 0$.
2. Transfer the quantity $\delta = (kx_{Uj} - x_{Uj})/2$ from species j in V to the same species in U such that $x_{Uj} + \delta = x_{Vj} - \delta$. Because of this transfer, the difference in abundances of species j between U and V is decreased but since the species proportions in both assemblages are no longer the same, we now have $CH_{UV} > 0$.

For example, let U be a three-species assemblage with abundances $x_{U1} = 10$, $x_{U2} = 20$ and $x_{U3} = 30$. By multiplying all species abundances in U by $k = 5$, a new assemblage V is obtained with abundances $x_{V1} = 50$, $x_{V2} = 100$ and $x_{V3} = 150$. Since the species proportions in both assemblages are identical, we have $CH_{UV} = 0$. If the quantity $\delta = 20$ is transferred from species 1 in V to the same species in U, the abundances in U and V become: $x_{U1} = 30$, $x_{U2} = 20$ and $x_{U3} = 30$ and $x_{V1} = 30$, $x_{V2} = 100$ and $x_{V3} = 150$. If we calculate the chord distance from the new abundances we now obtain $CH_{UV} = 0.52$.

The same reasoning can be easily applied to Hajdu's method: given a species assemblage U with abundances x_{Uj} ($j = 1, 2, ..., S$), first multiply all species abundances in U by a constant $k > 1$. Next, add the quantity $\delta = kx_{Uj} - x_{Uj}$ to species j in U such that $x_{Uj} + \delta = kx_{Uj}$. All else being equal, the addition of δ modifies the species proportions in U such that CH_{UV} becomes larger than zero.

In both cases, decreasing the difference in the absolute abundance of one species between two sites does increase their dissimilarity, thus violating the monotonicity criterion. Therefore, the chord and the Hellinger distances are not monotonic to changes in abundance in raw data space. Note, however, that since both measures are obtained by calculating the Euclidean distance on transformed data, the chord and the Hellinger distances are both monotonic to changes in abundance in transformed data space. As a result, the chord and the Hellinger distances are not simply 'improved' versions of the Euclidean distance which do not suffer from the Orlóci paradox; they are alternative measures with different properties that summarize plot-to-plot dissimilarity in a different data space.

Discussion: looking for the perfect index

Ecologists have developed a multitude of indices for summarizing plot-to-plot dissimilarity based on distinct goals and viewpoints. However, the choice of the most adequate index for solving a given ecological problem remains a complex question which does not have a clear and unequivocal answer (Ricotta and Podani 2017).

In this paper, I showed that being based on a separate data transformation of both plots, the chord and the Hellinger distances are not monotonic to changes in abundance in raw data space regardless of the method used for testing this property. This lack of monotonicity is a general attribute of all coefficients for which the dissimilarity between two plots is zero if the proportion of species abundances is identical in both plots. Accordingly, unless our aim is to measure plot-to-plot dissimilarity based on species proportions, these measures should not be used. For example, if we want to calculate the beta diversity of a set of plots from the average dissimilarity between pairs of plots as suggested for example by Koleff et al. (2003) or Legendre and De Cáceres (2013), the chord or the Hellinger distance are not necessarily the best possible choice (see e.g., Ricotta 2017).

Ecology is not a hard science and hypothesis testing is generally more liberal than in mathematics or physics. Therefore, the lack of monotonicity to changes in absolute abundances does not automatically mean that all the work done using the chord or the Hellinger distances as surrogates for the Euclidean distance is necessarily wrong. However, paraphrasing Jaynes (1957), one expects that deductions made from these measures, if carried far enough, will eventually lead to contradictions.

In this framework, I would like to stress once again that by condensing the structure of a multidimensional set of data into a single indicator, information is inevitably lost to the point that we do not even fully understand the properties of the best-known dissimilarity indices. Like any other ecological indicator, dissimilarity coefficients are part of a complex, plural and dynamic approach to ecological studies (Juhász-Nagy 1984). Therefore, a 'magic' measure that is able of unambiguously portraying all facets of plot-to-plot dissimilarity simply does not exist. Rather, a variety of targeted but imperfect measures may be used and their relevance should be evaluated based on their ability to solve the specific ecological question under scrutiny.

Roberts (2017) argued that the concept of dissimilarity can be quantitatively defined only by the algebra of a given index such that selecting a dissimilarity index necessarily imparts a perspective to the data. The practitioner just has to be sure he endorses this perspective. Therefore, contrary to those believing that we do not need any other research on diversity and dissimilarity because we already know everything, an accurate assessment of the properties of old and new dissimilarity coefficients is a necessary requirement for answering Sherwin et al.'s (2006) fundamental question: Does the index measure what we want to measure (assuming we know what we want to measure)?

References

Dalton, H. 1920. Measurement of the inequality of incomes. *Econ. J.* 30:348–361.

Gower, J.C. and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* 3:5–48.

Hajdu, L.J. 1981. Graphical comparison of resemblance measures in phytosociology. *Vegetatio* 48:47–59.

Jaynes, E. 1957. Information theory and statistical mechanics. *Phys. Rev.* 106:620–630.

Juhász-Nagy, P. 1984. Spatial dependence of plant populations. Part 2. A family of new models. *Acta Bot. Hung.* 30:363–402.

Koleff, P., K.J. Gaston and J.J. Lennon. 2003. Measuring beta diversity for presence-absence data. *J. Anim. Ecol.* 72:367–382.

Legendre, P. and M. De Cáceres. 2013. Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecol. Lett.* 16:951–963.

Legendre, P. and E.D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271–280.

Mouquet, N. and M. Loreau. 2003. Community patterns in source-sink metacommunities. *Am. Nat.* 162:544–557.

Orlóci, L. 1967. An agglomerative method for classification of plant communities. *J. Ecol.* 55:193–206.

Orlóci, L. 1978. *Multivariate Analysis in Vegetation Research.* 2nd ed. Junk, The Hague, NL.

Pulliam, H.R. 1988. Sources, sinks, and population regulation. *Am. Nat.* 132:652–661.

Ricotta, C. 2017. Of beta diversity, variance, evenness, and dissimilarity. *Ecol. Evol.* 7:4835–4843.

Ricotta, C. and J. Podani. 2017. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecol. Complex.* 31:201–205.

Roberts, D.W. 2017. Distance, dissimilarity, and mean-variance ratios in ordination. *Methods Ecol. Evol.* 8:1398–1407.

Sherwin, W.B., F. Jabot, R. Rush, and M. Rossetto. 2006. Measurement of biological information with applications from genes to landscapes. *Mol. Ecol.* 15:2857–2869.

Received February 27, 2019
Revised April 17, 2019
Accepted May 13, 2019