Vol. 372: 303–306, 2008
doi: 10.3354/meps07841

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published December 9

OPEN ACCESS

COMMENT

Identification of the Bray-Curtis similarity index: Comment on Yoshioka (2008)

Paul J. Somerfield*

Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK

ABSTRACT: In a recent As I See It, Yoshioka (2008, Mar Ecol Prog Ser 368:309–310) stated that the term Bray-Curtis should only be used for doubly-standardised data, and when data counts are used to calculate similarity, the term Czekanowski is correct; by using the wrong name, it is claimed, marine ecologists demonstrate a lack of knowledge and introduce confusion. This needs to be rectified. It is incorrect: (1) to confuse pretreatment of data with calculation of similarity (although both are part of a process, few ecologists would accept that it is logical for the name of a coefficient to alter depending on the data used to calculate it); (2) that the name for the coefficient calculated from counts is Czekanowski; (3) that terrestrial ecologists reserve the name Bray-Curtis for doubly-standardised data; (4) to assume that alternative names for coefficients cause confusion when it is clear which coefficient is used; (5) that this is a problem with software, or the way software is used. Yoshioka (2008) is right that pretreatments alter the rank order of similarities between samples, but this is not a novel observation, and it is why ecologists who know what they are doing use different pre-treatments. What ecologists mean by Bray-Curtis is well understood, and they do not mean Czekanowski's coefficient as originally described, although some textbooks confuse the 2 indices. We should maintain use of the term Bray-Curtis in community ecology.

KEY WORDS: Bray-Curtis · Czekanowski · Similarity · Dissimilarity · Resemblance

Resale or republication not permitted without written consent of the publisher

Introduction

In a recent As I See It, Yoshioka (2008) contends that marine ecologists (and specifically, users of the PRIMER software) often misidentify a coefficient which they term Bray-Curtis (Bray & Curtis 1957), used as a measure of resemblance between samples in analyses of multivariate data. Yoshioka (2008) cites Goodall (1978) as evidence that the measure usually attributed to Bray & Curtis (1957) was actually described previously by Czekanowski (1909). As Yoshioka (2008) concedes, the measures attributed to both Bray & Curtis (1957) and to Czekanowski (1909) by Goodall (1978) are the same. The formula for the similarity measure being considered is:

$$S_{jk} = 100\frac{\sum_{i=1}^{p} 2\min(y_{ij}, y_{ik})}{\sum_{i=1}^{p}(y_{ij} + y_{ik})} \quad (1)$$

where y_{ij} represents the entry in the ith row and jth column of the data matrix, i.e. the abundance (or biomass, or cover) for the ith species in the jth sample (i = 1, 2, ..., p; j = 1, 2, ..., n). Similarly, y_{ik} is the count for the ith species in the kth sample. The min(.,.) term is the minimum of the 2 counts. The separate sums in the numerator and denominator are both over all rows (species) in the matrix.

The difference, according to Yoshioka (2008), is that because of the nature of their data and the aim of their study, Bray & Curtis (1957) did not use counts to calculate their intersample similarities. Instead they used a form of double standardisation prior to calculating similarity ('standardisation' is a widely used term in ecology for converting counts to percentages or proportions of a total, although it has a different meaning in statistics). Therefore, Yoshioka (2008) states that the term Bray-Curtis should be reserved for similarity

*Email: pjso@pml.ac.uk

© Inter-Research 2008 · www.int-res.com

calculated from doubly-standardised data; if actual counts are used, one should refer to the coefficient as Czekanowski's. The specific charge is that by using the term Bray-Curtis for the coefficient calculated from counts, marine ecologists have been committing an error, as among terrestrial ecologists it is well understood that Bray-Curtis similarity refers to similarities calculated from doubly-standardised data. Yoshioka further contends that software that uses the term Bray-Curtis for similarity that is based on counts instead of doubly-standardised data makes the situation worse because it confuses our terrestrial colleagues about what we have actually done in our analyses.

Treatment of the data

Yoshioka (2008) states that double-standardisation is understood in terrestrial ecology to be inherent to Bray-Curtis similarity. I disagree. Although double-standardisation is common in terrestrial ecology, as this is part of the data manipulation implicit in analytical techniques such as correspondence analysis and derivatives thereof (e.g. ter Braak 1986, though the form of double-standardisation differs in that case), the calculation of Bray-Curtis similarity based on doubly-standardised data is not the norm. Indeed, I think it is done rarely. I suggest that, rather than stating 'the true BC index is widely used in terrestrial ecology, and the Cz index is used widely in marine ecology' (Yoshioka 2008, p 310), Yoshioka actually should give real data from a statistically random selection of the 2000+ papers quoting Bray & Curtis (1957).

Yoshioka (2008) is confusing 2 different things: (1) data pretreatment; and (2) calculation of similarity. It is well understood by the vast majority of those who use non-parametric multivariate analyses, based on calculating similarities between samples, that there are 2 steps in the process, each of which is equally important (Clarke et al. 2006a, Olsgard et al. 1997, 1998). This was understood by Bray & Curtis (1957, p. 327) who used separate subheadings within a section entitled 'Treatment of the data' to describe the 2 steps. The first is an appropriate pretreatment of the data matrix (Clarke 1993, Clarke & Warwick 2001), described under 'Use of score sheets' by Bray & Curtis (1957, p. 328). This may take many forms, depending on the precise hypothesis that the investigator wishes to address, but in its simplest form it requires a manipulation of the data (or a decision not to manipulate) which takes into account the nature of the counts and the aspect of community structure that the investigator is interested in. There is a range of possible pretreatments (Clarke & Warwick 2001), including ones that take into account spatial clustering of organisms (Clarke et al. 2006b). Basically, a standardisation (conversion to percentages) is appropriate if sampling effort is not controlled, and increasing strengths of power transformations (none, square root, 4th root, presence/absence) are used to downweight the contribution of abundance to intersample similarities. This is well understood, not just by marine ecologists, but by ecologists in general, and it is implemented in PRIMER (Clarke 1993, Clarke & Warwick 2001, Clarke & Gorley 2006), as well as in other software packages which allow calculation of Bray-Curtis (dis-)similarity and analysis of ecological data without requiring a double standardisation.

Index of similarity

Once an appropriate transformation has been applied, a relevant coefficient must be used to calculate intersample similarities (Clarke et al. 2006a), discussed in Bray & Curtis (1957) under the subheading 'Index of Similarity', p. 328. One formula that empirically matches ecologists' perceptions of similarity of 2 assemblages is given in Eq. (1), and can also be calculated as:

$$S_{jk} = 100\left\{1 - \frac{\sum_{i=1}^{p}|y_{ij} - y_{ik}|}{\sum_{i=1}^{p}(y_{ij} + y_{ik})}\right\} \quad (2)$$

where $|...|$ represents the absolute value of the difference (the sign is ignored). In a widely cited paper, a team of terrestrial ecologists (Faith et al. 1987) also demonstrated that this coefficient and a very closely related one (termed Kulczynski; see Clarke et al. 2006a for discussion of their relationship) outcompete any others in the accurate reconstruction of simulated ecological gradients. For presence/absence data, Eqs. (1) or (2) become the Sørensen or Dice coefficient.

All of the measures mentioned in this text have alternative names, and there is no evidence that this has led to serious confusion among ecologists when it is clear what is meant by a particular name. These formulations encapsulate the same idea, namely that dissimilarity between samples is well described as the sum (over species) of the modulus of the difference between counts (Manhattan distance, $d_{jk} = \sum_{i=1}^{p}|y_{ij} - y_{ik}|$) divided by the sum of the counts in the samples being compared. As an aside, calculating either Bray-Curtis dissimilarity after double standardisation, which Yoshioka (2008) considers to be the 'true' Bray-Curtis, or Kulczynski dissimilarity, simply returns (Manhattan distance)/2 between samples in the standardised data matrix (since the sample totals are identical and the denominator term in Eq. 2 is always 200). Although they do not use the name Manhattan (nor any of its alternatives), it is clear from their paper that Bray & Curtis (1957) were well aware of this.

Manhattan distance, also known as the taxicab metric or the city-block metric (Legendre & Legendre 1998), is a universal and long-established distance measure, so there seems little need to name a coefficient Bray-Curtis when it is simply Manhattan, calculated following some particular pretreatment. Returning to the general case of Eqs. (1) & (2), the reason why this is such an appropriate measure for biological data is that (unlike many distance measures) it gives zero a special role, successfully combining the structural information on presence/absence with quantitative counts of species that are present. As this is a very sensible formula, it has been rediscovered (or redescribed) many times.

Identification of the measure known as Bray-Curtis

Although resemblance measures are not controlled by an equivalent of e.g. the International Code for Zoological Nomenclature (ICZN), there may be a feeling that priority should play some part. So, what should we call such a measure? According to Yoshioka (2008) we should call it Czekanowski. I have been unable to obtain a copy of Czekanowski (1909); apparently, Yoshioka has not read it either (he cites Goodall 1978 as his authority). Failing to access rare original material, where can we turn for guidance? In my opinion, the best source is Legendre & Legendre (1998), who took the trouble to look at a wide range of similarity, dissimilarity and distance (collectively referred to as resemblance) measures in order to make some sense of the plethora of measures available. According to them, the measure was attributed to a Polish mathematician, H. Steinhaus, by Motyka (1947). It was subsequently 'rediscovered' a number of times, and its one-complement is known as the Odum (1950) or Bray-Curtis coefficient. Ignoring the bit about the one-complement, as it is axiomatic that dissimilarity = (1 – similarity) — or (100 – similarity), depending on the scale — the coefficient could, therefore, be attributed to Steinhaus (undated, but working during World War II; P. Legendre pers. comm.), Odum (1950) or Bray & Curtis (1957).

More importantly, Legendre & Legendre (1998), having checked the original publications (P. Legendre pers. comm.), state that the coefficient known to us as Bray-Curtis is sometimes *incorrectly* attributed to the anthropologist Czekanowski (1909, 1913), and that Czekanowski's coefficient, mean character difference (*durchschnittliche Differenz* in German), is actually

$$d_{jk} = \frac{1}{p}\sum_{i=1}^{p}|y_{ij} - y_{ik}| \qquad (3)$$

where p is the number of variables. To use it with species abundances, one should modify it to exclude joint absences (double zeros) from the computation and replace p with (p – number of double zeros). Therefore, according to Legendre & Legendre (1998), Czekanowski's original coefficient is *not* the same as the one we refer to as Bray-Curtis.

Sources of confusion

Yoshioka (2008) is based on Goodall (1978). Various recent textbooks and many documents researched via the Internet give Czekanowski as an alternative name for the measure that marine ecologists (and many others) refer to as Bray-Curtis. Assuming that Legendre & Legendre (1998) are right in their assessment that Czekanowski's coefficient is different from Bray-Curtis, why is there confusion? I am reminded of Stephen Jay Gould's (1988) essay 'The case of the creeping fox terrier clone', in which he questions why the size of *Hyracotherium* (an early ancestor of the horse) is equated to that of a fox terrier in textbooks — despite the facts that the animals are not the same size and that the vast majority of readers of those textbooks would have neither familiarity with fox terriers nor knowledge of their size. The answer, it transpires, is that the comparison was made in 1904 by Henry Fairfield Osborne (a keen rider and hunter of foxes) in an article for a popular magazine. No one ever bothered to check its validity or utility. Once the comparison took hold it was simply copied by generation after generation of textbook writers. This practice of unquestioning copying makes science texts virtual clones of each other on this, and many other points, and places science textbook writers among the most egregious purveyors of myth and inaccuracy (McComas 1998).

I do not know who originally equated Czekanowski's coefficient with that which we choose to call Bray-Curtis, but the fact that textbooks continue to state that they are the same does not make it true. As a matter of fact, Bray & Curtis (1957, p. 346) did not claim to have invented the measure, given in Eq. (1) above, either. They referred to it as 'Gleason's coefficient of community' and attributed it to Gleason (1920). Apparently there is nothing in Gleason (1920) that may be equated to the coefficient given in Eq. (1) (P. Legendre pers. comm.), despite Bray & Curtis's (1957) assertion that this is where they got the measure. They were involved in a debate about Gleasonian vs Clementsian views of communities at the time, with Bray and Curtis supporting Gleason (McCune & Beals 1993), and perhaps this played a part in their writing. A charge of 'Fox terrier clone syndrome' could therefore be levelled at all those who have used the term Bray-Curtis for the measure. The difference, however, is that Bray & Curtis (1957) made it very clear what the measure is,

despite attributing it, perhaps erroneously, to someone else, and so quoting them as an authority does not confuse others. In addition to being explicit about the formula, reasons that Bray & Curtis (1957) are considered as the authority for the measure, in preference to others who described it previously, may be that they described it within a framework of multivariate analysis, considering issues such as data treatment, choice of measures, methods of ordination and classification, and ecological theory.

Concluding remarks

The true Czekanowski coefficient can be calculated in PRIMER, which offers a choice of 50 resemblance measures and identifies them in all outputs with the near-definitive numbering scheme of Legendre & Legendre (1998) to avoid confusion of terminology. Yoshioka's assertion that he has discovered an apparent misidentification of Czekanowski's coefficient as Bray-Curtis, that has gone unrecognised for decades, has forced me to clarify several points in my own mind, and hopefully those of others. The coefficient could be termed Steinhaus, or Odum, but it is generally known as Bray-Curtis, and the vast majority of ecologists, whether studying marine, freshwater or terrestrial ecosystems, are very clear about what is meant. For animal species names, the ICZN allows the principle of priority to be modified in the interests of stability and universality. For similar reasons, we should maintain use of the term Bray-Curtis in community ecology, acknowledging that Bray & Curtis (1957) themselves attributed it (perhaps erroneously) to Gleason (1920).

Acknowledgements. I am indebted to P. Legendre for revisiting many of the original publications mentioned in this work and describing what he saw there. This work is a contribution to the Plymouth Marine Laboratory's core strategic research programme. Work underpinning it was supported in part by the UK Natural Environment Research Council (NERC), and by the UK Department for Environment, Food and Rural Affairs (Defra) through project ME3109.

LITERATURE CITED

Bray RJ, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr 27: 325–349

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. Aust J Ecol 18:117–143

Clarke KR, Gorley RN (2006) Primer v6: user manual/tutorial. PRIMER-E, Plymouth

Clarke KR, Warwick RM (2001) Change in marine communities: an approach to statistical analyses and interpretation, 2nd edn. PRIMER-E, Plymouth

Clarke KR, Somerfield PJ, Chapman MG (2006a) On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. J Exp Mar Biol Ecol 330:55–80

Clarke KR, Chapman MG, Somerfield PJ, Needham HR (2006b) Dispersion-based weighting of species counts in assemblage analyses. Mar Ecol Prog Ser 320:11–27

Czekanowski J (1909) Zur differential Diagnose der Neandertalgruppe. Korrespbl dt Ges Anthrop 40:44–47

Czekanowski J (1913) Zarys metod statystcznych w zastosowaniu do anthropologii. Travaux de la Société des Sciences de Varsovie. III. Classe des sciences mathématiques et naturelles, no. 5. Société des Sciences de Varsovie, Warsaw

Faith DP, Minchin PR, Belbin L (1987) Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57–68

Gleason HA (1920) Some applications of the quadrat method. Bull Torrey Bot Club 47:21–33

Goodall DW (1978) Sample similarity and species correlation. In: Whittaker RH (ed) Ordination of plant communities. W Junk, Boston, MA, p 101–149

Gould SJ (1988) The case for the creeping fox terrier clone. Natural History 97:16–24

Legendre P, Legendre L (1998) Numerical ecology, 2nd Engl edn. Elsevier, Amsterdam

McComas WF (1998) The principal elements of the nature of science: dispelling the myths. In: McComas WF (ed) The nature of science in science education: rationales and strategies. Kluwer Academic Publishers, Dordrecht, p 53–70

McCune B, Beals EW (1993) History of the development of Bray-Curtis ordination. In: Fralish JS, McIntosh, Loucks OL (eds) John T. Curtis: fifty years of Wisconsin plant ecology. Wisconsin Academy of Science, Arts, and Letters, Madison, WI, p 67–80

Motyka J (1947) O zadaniach i metodach badan geobotanicznych. Sur les buts et les methods des recherches géobotaniques. Annales Universitatis Mariae Curie-Sklodowska (Lublin, Polonia), Sectio C, Supplementum I. Universitatis Mariae Curie-Sklodowska, Lublin

Odum EP (1950) Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. Ecology 31:587–605

Olsgard F, Somerfield PJ, Carr MR (1997) Relationships between taxonomic resolution and data transformations in analyses of a macrobenthic community along an established pollution gradient. Mar Ecol Prog Ser 149:173–181

Olsgard F, Somerfield PJ, Carr MR (1998) Relationships between taxonomic resolution, macrobenthic community patterns and disturbance. Mar Ecol Prog Ser 172:25–36

ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179

Yoshioka PM (2008) Misidentification of the Bray-Curtis similarity index. Mar Ecol Prog Ser 368:309–310

Editorial responsibility: Matthias Seaman, Oldendorf/Luhe, Germany

Submitted: October 28, 2008; Accepted: November 24, 2008
Proofs received from author(s): November 26, 2008