Journal of Vegetation Science 18: 921-924, 2007

FORUM

Spatial confusion or clarity? Reply to Ricotta & Avena

Podani, János

Department of Plant Taxonomy and Ecology, Eötvös Loránd University, Pázmány P. s. 1/c, H-1117 Budapest, Hungary; Fax +36 13812188; E-mail podani@ludens.elte.hu

Abstract. This communication is a response to a recent suggestion that ordinal data of the Braun-Blanquet type (BB) can be directly or, after conversion to ranks, indirectly analysed by metric methods. I show that the proposals on structure in topological spaces made by Ricotta & Avena in a recent contribution to this Forum are confounding because (1) they use the term 'topological' in the inappropriate way, and (2) the measure they propose is in fact a metric, rather than merely topological. In addition, I illustrate with a few examples how a truly topological measure functions, so that the reader can appreciate the ideas behind their definitions. By reference to axiomatic measurement theory, I argue that whenever vegetation scientists know exactly at the outset what attributes they wish to express by relevé data, what questions they are asking and whenever they are aware of the basic properties of the BB scale, ordinal data analysis is still the most logical choice.

Keywords: Abundance/dominance data; Measurement theory; Multivariate analysis; Ordinal scale; Phytosociology; Scale typology.

The scale problem: misleading suggestions

In two recent papers, I have called the attention of phytosociologists and vegetation ecologists to certain logical and mathematical problems that relate to the treatment of Braun-Blanquetian and similar ordinal data types in multivariate analysis (Podani 2005, 2006). I recommended the application of procedures specifically designed for ordinal scores, as well as a more critical use of conventional ordination and classification methods. An even more recent commentary by Ricotta & Avena (2006) indicates that I was at least able to rock the boat.

Although Ricotta & Avena also recognize the problem, they cut the Gordian knot by suggesting that we can "implicitly shift our attention from the metric space to the topological space" and, as a consequence, ordinal data will then become tractable by conventional multivariate methods. The message is quite clear for all phytosociologists: do not worry about the data types you are using, just simply declare your analytical space as 'topological', and then use the same old methods as before. According to Ricotta & Avena themselves, this move can only be done if the "results are interpreted accordingly" and if we keep in mind that "topological analysis completely ignores metric information". They do not explain, however, how this interpretation differs from that of results obtained by metric procedures. Therefore, without guidance from a specialist their recommendations are potentially misleading and seem to be no more than conjuring with magic words, such as *topological, metric* and *space*, to which most phytosociologists are not accustomed at all. On the contrary, I assumed in my articles that interpretation of results, i.e. ordinations and classifications, may follow the same lines as usual, only the analytical tools require proper adjustment to the data used.

Even more misleading is their use of the words 'structural' and 'topological'. They apparently equate structure and topology by saying that in a topological space "the only available information concerns data structure" which is "usually considered more important than metric information". From these sentences, it is hard to find out how those authors would define structure. As far as I see, there is structure in all kinds of space, i.e. a topological structure in the topological space and a – more detailed – metric structure in the metric space. Actually, in mathematics topological space is defined in set theoretical terms (Munkres 2000) which falls very far from Ricotta & Avena's understanding of the same concept. In addition, every metric space is automatically a topological space (Munkres 2000), showing the ambiguity of Ricotta and Avena's attitude towards these ideas. This explains partially why these authors are mistaken about an answer regarding the topological interpretability of ordinations and classifications. Now, I can provide some answer: in an ordination, for example, a truly topological interpretation should focus only on the neighbourhood of points, rather than their distances. No question that all vegetation scientists implicitly consider interpoint distances when viewing and interpreting ordinations of their data.

Again, it is the adjacency of points that matters in a non-metric topological space. For this reason, Ricotta & Avena's interpretation of differences between values is in general not purely topological. Their suggestion that the difference between 2 and 4 is twice as much as the difference between 1 and 2 is valid only if at least one value of 3 for the given species is present in the data set analyzed. In other words, there *is* a point between the other two in the topological space. If none of the relevés contains this species with a score of 3, then 1 and 2 will be as close topologically as values 2 and 4, because the corresponding points are now neighbors in the topological space. Such 'missing' scores are not uncommon at all in phytosociological data tables. To clarify it further using Ricotta & Avena's staircase metaphor: the topological difference between two persons standing on steps 1 and 4 depends basically on the number of steps occupied in between. Even if we forget about the – otherwise undefined – height of individual steps of the scale and count the number of steps separating these two persons, regardless occupancy, then we find ourselves in the metric domain! Hence, Ricotta & Avena's proposal is actually not a shift from one space to another, but simply relabeling one measurement scale with the name of the other. Paradoxically, this implies a move in a direction opposite to what they meant: Braun-Blanquet scores imply an ordered topological space, whereas their 'staircase measure' is a metric.

Table 1 in Ricotta & Avena's (2006) article gives the impression that conversion of Braun-Blanquet scores into ranks may provide a potential solution of the scale problem, because ranks are suitable to metric analysis. However, the ordinal scales used in phytosociology usually have no more than 10-12 values, so that tied ranks are unavoidable when the number of species in a relevé exceeds 12, which is commonly the case. If, say, relevé *h* has 5 species with the ordinal score of '+', and relevé *i* has 3 such species only, then conversion will lead to the presence of five 3-s in relevé *h* and three 2-s in relevé *i*. That is, we introduce differences between species that had identical scores before, a manipulation obviously unacceptable for phytosociologists. Suppose that one of these species is present in both relevés *h* and *i*, then after transforming them to ranks there will be a 3 – 2 = 1 difference. Amplification of a few-valued scale into ranks cannot help if we insist upon metric measures. We do not have to insist, however, because there have been truly topological solutions to the ordinal scale problem, as shown in the next section.

The question whether a result is meaningful or not for the researcher is a serious one and therefore comparisons between methods are essential. Ricotta & Avena, however, provide neither references nor authors when mentioning that "the use of conventional multivariate methods reproduce (sic) the researcher's intuitive classification/ordination scheme much better than multivariate methods explicitly developed for ordinal data". As far as I am aware, there is no published and detailed account on the performance of metric and ordinal procedures when both are applied to the same set of ordinal data, except for my App. 3 to Podani (2005). In this study, I found that although there are obvious differences between the two approaches, yet the results of ordinal clustering were no less interpretable that those of metric classification. Actually, ordinal clustering was much less prone to group sizes than Ward's method, which is among the most popular metric clustering procedures applied in vegetation studies. I am convinced that one must be more careful with general statements on interpretability and relative meaningfulness without justification from authored comparative evaluations.

Topological measures

When evaluating structure in different abstract spaces, the crucial point is how we express interpoint relationships. If we use Euclidean distances or other metric coefficients, as Ricotta & Avena suggest, then the topological space is a metric one at the same time. The question arises then: how to express topological relationships between points appropriately, so that we remain consistent with the non-metric properties of the data as well? In the literature, we can find several ideas on this problem. Some approaches involve counting the number of points which fall between the two points in question in a hyperspace. The Calhoun 'distance' (Bartels et al. 1970; Orlóci 1978) is a good example. In calculating this formula for the relevé pair *h* - *i*, we have to count the number of relevés, n_1, that are intermediate between *h* and *i* for one species at least. Furthermore, we count the number of relevés, n_2, that take identical values with either *h* or *i* for at least one species. The third value, n_3, is the number of relevés which agree with both *h* and *i* for at least one species. Then, the Calhoun 'distance' is defined as

$$CAL_{hi} = w_1 n_1 + w_2 n_2 + w_3 n_3 \qquad (1)$$

where the weights are arbitrarily specified (according to the proponents of the method, $w_1 = 6$, $w_2 = 3$, and $w_3 = 2$). For Orlóci (1978), a more logical definition is setting $CAL_{hi} = n_1$, since actually only n_1 points fall between *h* and *i* and the arbitrary weights are thus eliminated. Note that *CAL* may be zero even if the two relevés are different, so its use in multivariate analysis has limitations.

An expansion of Gower's formula to ordinal data (Podani 1999) is also based on order topology. In this approach, the number of relevés that are intermediate between *h* and *i* is determined for each species, and then these numbers are summed over all species to give an *interchange measure*. As the name suggests, the relevés are ordered for each species, and the 'distance' is the number of interchanges of neighbouring values in the ordering which are necessary to put relevé *h* into the same position as relevé *i*. The following example clarifies this. Let us have only one species for simplicity, and eight relevés with the following BB scores:

Relevé no.	1 2 3 4 5 6 7 8
Score	1 2 1 4 1 2 2 1

These data lead to the following partial ordering of the relevés:

{1, 3, 5, 8} {2, 6, 7} {4}

(i.e. for items in brackets there is no ordering). The number of moves to put relevé 4 into the position of relevé 5 will be 4 (because we have to skip relevés 2, 6 and 7), for relevés 5 and 6 we have 1, just like for relevés 4 and 6. Thus, this measure violates the triangle inequality ($1 + 1 < 4$), yet it is symmetric, and reflects structure in an ordered topological space. For a pair of identical relevés, the interchange measure yields 0, as desired.

Historical background: two opposing views

Ricotta & Avena's article called my attention to the fact that the scale problem has long been persisting in other fields of science, such as psychology. It was my fault that I did not give an overall picture of this situation in my earlier reports, and I sincerely regret this. Admittedly, a brief historical overview may be illuminating for all of us engaged in the debate on scale typology.

It was the psychologist S.S. Stevens (1946) who suggested first the distinction between nominal, ordinal, interval and ratio scale variables. In a subsequent communication (Stevens 1951), he went further by claiming that the scale types determine automatically which statistical procedures are permissible and which are not. A scale that preserves meaning under a class of transformations cannot be used with statistics whose meaning changes if these transformations are applied to the data. Stevens' ideas were taken over by textbooks in statistical and multivariate analysis (including Anderberg 1973) and became widely accepted in a special branch of mathematics, measurement theory (Luce et al. 1990). The axiomatic approach to measurement problems has proven that if (1) we know what attributes of the real word we wish to measure, (2) we know what questions we shall ask about these attributes and (3) we have assigned numbers to these attributes so that their features and relationships are preserved, then (1) the measurements must be made using one of those scale types, (2) transformations not permissible for the given scale type will alter the answers to the questions, and therefore (3) only those methods can be used that are appropriate for the selected scale type (see Velleman & Wilkinson 1993). My suggestions on the appropriate treatment of ordinal data in ecology were implicitly based on these assumptions.

However, as Velleman & Wilkinson pointed out, these assumptions are not always valid. Opponents of the axiomatic approach provided several, usually non-biological situations when we do not know *a priori* which attributes we wish to measure and what questions we want to ask. A classical example is as follows: people visiting a reception receive a consecutively numbered ticket, starting 1, at the door so that a raffle can be held afterwards. The data, i.e. the values on the tickets represent an ordinal variable if the arrival ordering of people is of interest, but for the procedure to select the winner they are interpreted as a nominal variable. Velleman & Wilkinson therefore conclude that Stevens' typology "is not the attribute of the data, but rather depends upon the questions we intend to ask of the data and upon any additional information we may have".

I am convinced that in vegetation science we usually know exactly at the outset what attributes we wish to express by relevé data, what questions we are asking and most of us are aware of the basic properties of the BB scale. Consequently, assumptions that necessitate the axiomatic approach are satisfied. If we do not know that 1, 2, …, 5 designate BB scores because the data are mined from a database without information on the measurement scale, then I agree with Velleman & Wilkinson that starting from an *a priori* data type is "simply bad science and bad data analysis" and only in such cases can I agree with the propositions put forward by Ricotta & Avena. But again, the results cannot be any better than the quality of input data in any case, no matter how we manipulate them.

Acknowledgements. I am grateful to J. Garay for discussions on some issues of topology. I thank S. Diaz and two anonymous referees for their attention and suggestions. Financial support came from the Hungarian National Scientific Grant (OTKA) no. NI 68218..

References

Anderberg, M.R. 1973. *Cluster analysis for applications*. Wiley, New York, NY, US.

Bartels, P.H., Bahr, G.E., Calhoun, D.W. & Wied, G.L. 1970. Cell recognition by neighbourhood grouping techniques in Ticas. *Acta Cytol*. 14: 313-324.

Luce, L.D., Krantz, D.H., Suppes, P. & Tversky, A. 1990. *Foundations of measurement*. Vol. 3. Academic Press, New York, NY, US.

Munkres, J.R. 2000. *Topology: A first course*. 2nd. ed. Prentice-Hall, Upper Saddle River, NJ, US.

Orlóci, L. 1978. *Multivariate analysis in vegetation research*. 2nd. ed. Junk, The Hague, NL.

Podani, J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. *Taxon* 48: 331-340.

Podani, J. 2005. Multivariate exploratory analysis of ordinal data in ecology: Pitfalls, problems and solutions. *J. Veg. Sci*. 16: 497-510.

Podani, J. 2006. Braun-Blanquet's legacy and data analysis in vegetation science. *J. Veg. Sci*. 17: 113-117.

Ricotta, C. & Avena, G. 2006. On the evaluation of ordinal data with conventional multivariate procedures. *J. Veg. Sci*. 17: 839-842.

Stevens, S.S. 1946. On the theory of scales of measurement. *Science* 103: 677-680.

Stevens, S.S. 1951. Mathematics, measurement and psychophysics. In: Stevens, S.S. (ed.) *Handbook of experimental psychology*, pp. 1-49. Wiley, New York, NY, US.

Velleman, P.F. & Wilkinson, L. 1993. Nominal, ordinal, interval, and ratio typologies are misleading. *Am. Statist*. 47: 65-72.

Received 1 December 2006;
Accepted 13 June 2007;
Co-ordinating Editor: S.M. Díaz.