COENOSES 4 (1): 11-13, 1989
© 1989 by Edizioni LINT Trieste

THE MEASUREMENT OF NON-LINEAR ASSOCIATION IN COMMUNITY DESCRIPTOR SETS

R.C. Jancey, Department of Plant Sciences, the University of Western Ontario, London, Ontario, N6A 5B7 Canada

Keywords: Distance ratio, Minimum distance, Non-linearity, Resemblance, Single linkage

Abstract. A measure of resemblance is described which is independent of the linearity of descriptor relationships. The descriptors are recorded on interval scales, though not necessarily the same scale. The resemblance measure is based on a minimum distance pathway between the data points, related to a similar pathway for an equidistribution of points in the same reference space.

Introduction

The multivariate analysis of ecological or taxonomic data commonly requires, as a first step, the creation of a resemblance matrix. This may take the form of a matrix of similarities, *e.g.*, correlations and distances between descriptors. Taking the Pearson product-moment correlation coefficient as a simple and well known example of a measure, a problem common to the creation of resemblance matrices becomes apparent. Many of the more desirable measures of resemblance assume a linear relationship between descriptor responses. In some, particularly taxonomic cases, a biological rationale can be advanced for data transformations which will result in the approximation of a linear descriptor relationship. It should be noted, parenthetically, that the use of data transformations without a biological rationale may be of questionable validity. Further, a resemblance matrix involves all pairwise comparisons of descriptors (species) or individuals (relevés). Very few of such comparisons will be subject even to valid data transformations. Approaches to this problem have been made via Multidimensional Scaling (*e.g.*, Fewster and Orlóci, 1983; Orlóci, Kenkel and Fewster, 1984). Bradfield and Kenkel (1987), while primarily concerned with the measurement of compositional resemblance between quadrats rather than between descriptors, do employ the concept of shortest pathway. This use of shortest pathway was also employed by Minieka (1978) and Williamson (1978). A more general measure of resemblance is desirable, one which is readily applicable to interspecies, *i.e.*, inter-descriptor resemblance. Such a measure is described here.

Method

I start with the premise that the relationship shown in Fig. 1A is as close as that shown in Fig. 1B, considering the actual data. For each pairwise descriptor comparison, let us create an inter-relevé distance matrix (or inter-O.T.U. distance matrix in the case of taxonomy). Consider now a single-linkage clustering alogorithm, with the added constraint that no individual (relevé) in the study can be joined to more than two other individuals. This constraint applies to n-2 individuals, where n is the number of individuals (relevés) in the study. The remaining two individuals may have no more than one linkage. This algorithm creates a least distance joining of individuals. Summing the inter-individual (relevé) distances provides a measure of association. Knowing the range of both descriptors (simple in the synecological situation, since all species will have been, most probably, recorded on the same scale), it is possible to create a similar distance summation for the "worst case" situation. Thus we have a real distance summation falling between the "worst case" situation and zero. Since the actual distance can be expressed as a fraction of the "worst case", we have a general measure of resemblance with a lower bound of zero and an upper bound one. The one complement of this fraction measures descriptor association.

Fig. 1A (actual data) illustrates a set of 25 points generated by the function $y = x^2 - 6x+9$. Fig. 1B shows a data set with the same inter-point distances, but now in a linear relationship. The two data sets have almost the same descriptor association based on the method described (0.45 and 0.47), but when subjected to a product-moment correlation coefficient, Fig. 1A yields a correlation of 0.0009, while Fig. 1B (actual data), a correlation of 1.000. It may be argued that a relationship such as that shown in Fig. 1A (actual data) may easily be transformed to that shown in Fig. 1B (actual data). This is true, but the effectiveness of the algorithm is independent of such a relationship, *e.g.*, Fig. 1C (actual data).

There are two further aspects to be taken into account. The length of all axes may be standardised prior to computing the distance matrix. Each similarity value is computed from a fraction involving the sum of actual pairwise distances and the sum obtained from the "worst case" grid. This makes the similarity independent of the units used for each descriptor (particularly important for taxonomic descriptors), since the length ratio of the axis pairs would place a constraint

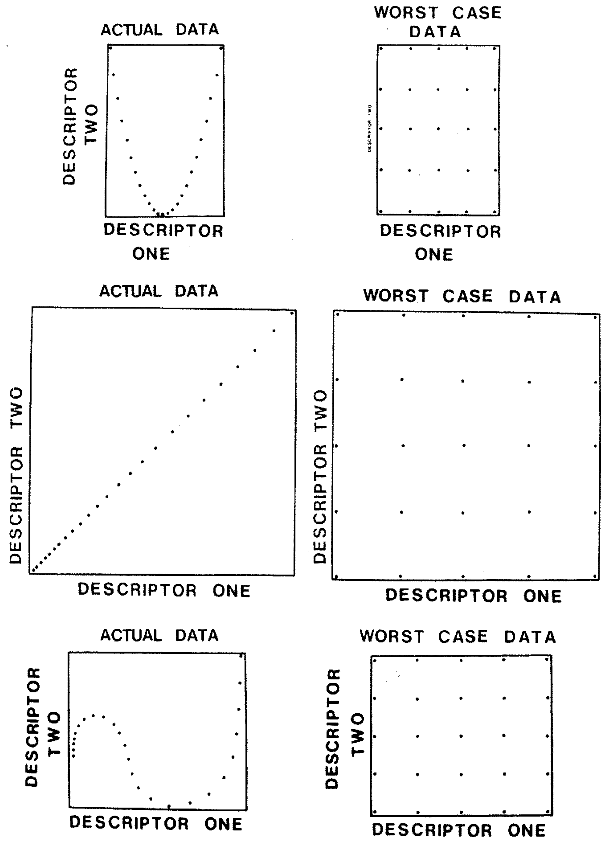

Fig. 1. Three pairs of data sets: 1A shows 25 points based on the relationship $y = x^{2} - 6x + 9$. 1B shows the same inter-point distances, but in a straight line. 1C again shows the same inter-point distances, but in a very non-linear relationship. The corresponding "worst-case" figures show an equidistribution of points for the creation of worst distance values. Units for the axes are not shown since they are not meaningful. The figures are all drawn to the same scale and the areas of each represent descriptor space.

upon the "worst case" sum of distances. For example, a linear relationship described by two axes of equal length will not be compared with the same "worst case" sum as a linear relationship in which one of the descriptor axes is vanishingly small. In the generation of the "worst case" grid, a very small random component is introduced into the grid coordinates. This is done to avoid distance ties and consequent sub-optimal linkage patterns.

A flow diagram is shown in Fig. 2. This diagram represents the computation of association between just one pair of descriptor axes. By iteration it can, of course, generate a matrix of similarities between all pairs of descriptors.

Discussion

The development of this algorithm has raised a number of questions which have their answers in biology rather than in the algorithm itself. Given the basic approach of the algorithm, it can, of course, be pointed out that the "worst case" grid is generated using the upper and lower bounds on each axis which are demonstrated by the data. This should be distinguished from the upper and lower bounds of some recording scale employed by the investigator. This is in accord with the Adansonian principle of allowing each descriptor an equal potential for contribution to the final result.

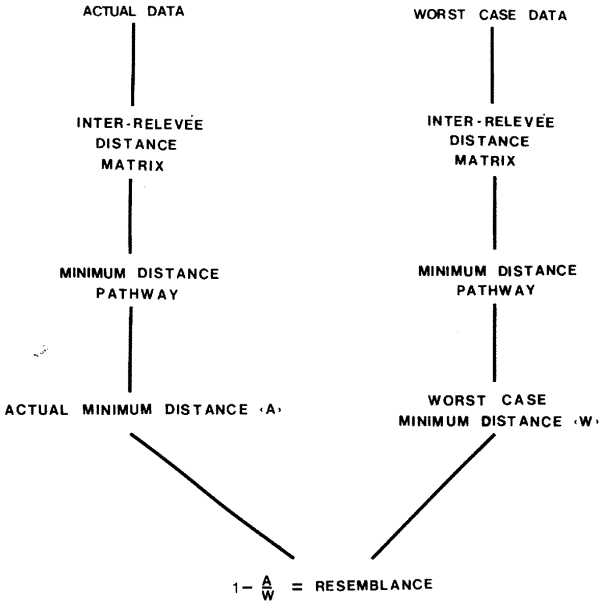
ACTUAL DATA
WORST CASE DATA
INTER-RELEVEE DISTANCE MATRIX
INTER-RELEVEE DISTANCE MATRIX
MINIMUM DISTANCE PATHWAY
MINIMUM DISTANCE PATHWAY
ACTUAL MINIMUM DISTANCE ‹A›
WORST CASE MINIMUM DISTANCE ‹W›
$1-\frac{A}{W}$ = RESEMBLANCE

Fig. 2. Flow chart of minimum distance resemblance calculation.

Notwithstanding the merits of an Adansonian philosophy, a study of extreme situations raised some interesting questions. Consider the case of a pair of descriptors, each of which manifests itself in only two states (despite, from external criteria, a potentially greater number of states). The algorithm will compute a measure of similarity just as it would if each descriptor was represented by twenty or thirty states. The point at question is whether the similarity value based on two states contains as much biological information one based on twenty or thirty for each descriptor. The problem is not unique to this algorithm, but does, in this case, have the potential for solution.

Perhaps a more fundamental question may be asked: Is the regular relationship shown in Fig. 1A indeed reresentative of as close a relationship as that shown in Fig. 1B (actual data)? In the algorithm as described, the relationship is found not to be quite as close, since it is embedded in a more compact space, and thus yields a smaller "worst case" grid and hence a smaller "worst case" total distance. This effect could, of course, be overcome by linearisation, as shown in Fig. 1B (actual data). The author tends to prefer allowing the effect of compactness to be reflected in the resemblance measure. It is felt that this may be more biologically meaningful, especially with non-idealised data. This would not, however, be a particularly easy position to defend.

REFERENCES

BRADFIELD, G.E. and N.C. KENKEL. 1987. Non-linear ordination using flexible shortest path adjustment of ecological distances. Ecology *68*: 750-753.

FEWSTER, P.H. and L. ORLÓCI. 1983. On choosing a resemblance measure for non-linear predictive ordination. Vegetatio *54*: 27-35.

ORLÓCI, L., N.C. KENKEL and P.H. FEWSTER. 1984. Probing simulated vegetation data for complex trends by linear and non-linear ordination methods. Abstracta Botanica *8*: 163-172.

ORLÓCI, L. 1978. Multivariate analysis in vegetation research. Dr. W. Junk Publishers, The Hague, The Netherlands.

MINIEKA, E. 1978. Optimisation algorithms for networks and graphs. M. Dekker, New York, New York, U.S.A.

WILLIAMSON, M.H. 1978. The ordination of incidence data. Journal of Ecology *66*: 911-920.

Manuscript received: January 1988