PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Wojtek J. Krzanowski
School of Engineering, Computer Science & Mathematics
University of Exeter, U.K.

ANALYSIS OF DISTANCE FOR STRUCTURED MULTIVARIATE DATA: REVIEW AND ILLUSTRATION

1. Introduction

Multivariate data in which the units are partitioned into *a-priori* groups arise frequently in practice. Moreover, it also frequently happens that, whether by accident or design, the groups are subject to further structure such as nesting or cross-classification. Multivariate analysis of variance (MANOVA) is a well-known technique that can be used to analyse such data. Typically, the analysis permits the testing of specific hypotheses such as presence of treatment main effects or interactions, the isolation of single degree-of-freedom contrasts among groups of interest, and the display of group means on a variety of different canonical variates (each highlighting a particular effect found to be significant). Full details of the technique may be found in most standard text books (e.g. [Mardia, Kent and Bibby 1979]). However, many data sets arise for which MANOVA cannot be used. Typical reasons are that the assumptions of normality of observations or homogeneity of dispersion matrices are violated; that some data values are missing; or that some or all of the variables are discrete, categorical or ordinal in form.

A good example is the set of data kindly supplied by Dr K. R. Clarke and Mr S. Widdicombe of the Plymouth Marine Laboratory, collected for an experiment conducted at the NIVA marine research station in Solbergstrand, Oslofjord, Norway [Widdicombe, Austen 2001]. Sediment was first collected from a sheltered bay in Oslofjord, was then placed in 98 plastic buckets, and was allowed to settle for nine weeks in order to permit the organisms to regain their spatial positions in each bucket.

For the period of the experiment (12 weeks), each bucket was subjected to one of seven different intensities of physical disturbance by stirring (no disturbance, once every four weeks, once every two weeks, once a week, twice a week, three times a week, every day) combined with one of seven levels of organic enrichment (0, 12.5, 25, 50, 100, 200, 400 gm/m^2 carbon). We denote the factor "physical disturbance" by A and the factor "organic enrichment" by P, with levels of these factors given by A0 to A6 and P0 to P6 respectively. Each of the 49 treatment combinations was replicated twice, and the two 7×7 sets of buckets were arranged on the experimental area in graeco-latin squares to remove any systematic positioning effects. In all, 80 species were represented across the whole experiment, so the abundance matrix **Z** was of size 98×80.

This matrix exhibited considerable skewness, with many zero incidences and nearly all non-zero entries in the range 1-9 but with three species having incidences predominantly between 10 and 50, one with incidences up to about 100, and one with some incidences in excess of 1000. Moreover, the observations were discrete counts and there were more variables (species) than error degrees of freedom. So MANOVA of this two-replicate 7×7 factorial structure was ruled out. Univariate ANOVAs of summary measures such as number of species, Margalef's species richness, or Pielou's measure of evenness were conducted but gave no insight into the inter-species relationships.

So Widdicombe and Austen [2001] looked at a series of multidimensional scalings (MDS). They used the Bray–Curtis dissimilarity coefficient [Bray, Curtis 1957] on the fourth roots of the incidences to counteract skewness and to increase the influence of the low abundance species. Two-dimensional MDS configurations were produced from:

1) the 49×49 matrix of dissimilarities between the averages over the two replicates for each combination of factor levels;

2) the 14×14 matrix of dissimilarities between each pair of buckets at each level of Disturbance;

3) the 14×14 matrix of dissimilarities between each pair of buckets at each level of Organic Enrichment.

Points were labelled in each diagram by factor levels, to show up any interesting patterns. While these diagrams gave some interesting pointers, they fell short of a satisfactory analysis. Here, and more generally, such displays are influenced by sampling fluctuations that can be volatile with small samples; there are often too many diagrams and it is difficult to detect patterns when looking across several pictures; and the arbitrariness of axis orientation between different diagrams is a confounding issue. We need a much more systematic approach, and this is provided by an analysis of distance. In the remainder of this paper the various steps of such analysis are reviewed and illustrated on the ecological data. Section 2 outlines the basic theory and descriptive techniques, section 3 focuses on the addition of biplots to MDS diagrams in order to incorporate information on the variables into the analy-

ses, and section 4 considers the assessment of stability of results. Some brief concluding remarks are made in section 5.

2. Analysis of distance

2.1. Background

Suppose that a data set consists of p observations made on each of n individuals, and d_{ij} is the dissimilarity between individuals i and j as measured by one of the large number of possible dissimilarity measures (see, e.g. [Gower 1985]. Let **D** be the $n \times n$ matrix having (i,j)th element $\frac{1}{2}d_{ij}^2$. Metric scaling (or Principal Coordinate Analysis; [Gower 1966; Krzanowski, Marriott 1994]) produces a matrix $\mathbf{X} = \mathbf{LE}^{1/2}$ of coordinates of points in q dimensions such that the Euclidean distance between points i and j is d_{ij}. Here **E** is the diagonal matrix of eigenvalues, and columns of **L** contain the corresponding eigenvectors, of

$$-\left(I - \frac{11'}{n}\right)\mathbf{D}\left(I - \frac{11'}{n}\right),$$

where I is the identity matrix, 1 is a vector of ones (both of appropriate dimension), and superscript t denotes transpose. Thus it follows that

$$-\left(I - \frac{11'}{n}\right)\mathbf{D}\left(I - \frac{11'}{n}\right) = \mathbf{LEL}' = \mathbf{XX}'. \qquad (1)$$

If the original dissimilarity measure is only semi-metric then some eigenvalues may be negative. They can either be ignored, or a small adjustment to elements of D will make them all positive (see [Krzanowski 2004)]. One way forward is MANOVA on the MDS coordinates X [Cuadras, Arenas 1990; Anderson, Willis 2003], but partitioning of distance provides an (arguably) richer alternative.

A justification for such a partitioning comes from the connection between squared distance and variance (or corrected sum of squares). First consider n points in one dimension, with coordinates $x_1, x_2, ..., x_n$. Then

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\{(x_i - \bar{x}) - (x_j - \bar{x})\}^2$$

so that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum\sum_{i<j}(x_i - x_j)^2.$$

If now the points are in p dimensions and x_{ik} is the coordinate of point i on dimension k, then repeating the above for each dimension, summing over dimensions and changing the orders of summation gives

$$\sum_{i=1}^{n}\sum_{k=1}^{p}(x_{ik}-\bar{x}_k)^2=\frac{1}{n}\sum\sum_{i<j}\left\{\sum_{k=1}^{p}(x_{ik}-\bar{x}_k)^2\right\}.$$

So the corrected sum of squares, when summed over the p variables, is equal to the sum of squared Euclidean distances between all distinct pairs of points, divided by n. This gives a basis for an analysis of distance breakdown that parallels the standard ANOVA breakdown. Moreover, metric scaling converts any positive semi-definite matrix of dissimilarities into a configuration of points in Euclidean space, so such a breakdown can be applied with any such matrix of dissimilarities as starting point. This idea was first presented in the one-way case by Digby and Gower [1981], and subsequently developed by d'Aubigny [1988], Pillar and Orloci [1996], Gower and Krzanowski [1999], Legendre and Anderson [1999], McArdle and Anderson [2001], Anderson [2001], and Krzanowski [2002; 2004; 2006]. We now trace some of the main features of this development.

2.2. Simple grouping (one-way arrangements)

Suppose the n units are divided into g groups, with n_i in the ith group and $\mathbf{G}$ the $n \times g$ group indicator matrix. Then $\mathbf{D}$ can be partitioned into g^2 submatrices $\mathbf{D}_{rs}$ containing values $\frac{1}{2}d_{ij}^2$ between each individual in group r and each one in group s. Now if we let $\mathbf{n} = (n_1,...,n_g)'$, $\mathbf{N} = \mathrm{diag}(n_1,...,n_g)$ and $\mathbf{F} = \mathbf{N}^{-1}\mathbf{G}'\mathbf{D}\mathbf{G}\mathbf{N}^{-1}$ with (r,s)th element f_{rs}, then simple algebra [Gower, Krzanowski 1999] establishes that:

(i) $\delta_{rs}=(2f_{rs}-f_{rr}-f_{ss})^{1/2}$ is the distance between the means of groups r, s;

(ii) $\frac{1}{n}\mathbf{1}'\mathbf{D}\mathbf{1} = \sum_{r=1}^{g}\frac{1}{n_r}\mathbf{1}'\mathbf{D}_{rr}\mathbf{1} + \frac{1}{n}\mathbf{n}'\mathbf{M}\mathbf{n}$ where $\mathbf{M}$ has (r,s)th element $\frac{1}{2}\delta^2$.

But since $\mathbf{1}'\mathbf{D}\mathbf{1}$ is just the sum of all elements of $\mathbf{D}$, it follows from section 2.1 that (ii) expresses in terms of squared distances the familiar ANOVA breakdown Total = Within + Between, but summed over all variables.

2.3. Multiple classifications

In more complex structures (e.g. factorial experiments), the between-group distance is an amalgam of different effects (e.g. main effects of factors and interactions between them). If there are m such effects, they can be isolated with the help

of matrices of contrasts $\mathbf{H}_i$ and their associated symmetric idempotent projection matrices

$\mathbf{Q}_i = \mathbf{H}_i(\mathbf{H}_i' \mathbf{H}_i)^{-1} \mathbf{H}_i'$ for $i = 1,\ldots,m$. The fundamental result then underlying any ANOVA breakdown is Cochran's Theorem, which says that if $\mathbf{x}$ is any zero-mean (i.e. mean-centered) vector of values and $\sum_{i=1}^{m} \mathbf{Q}_i = I$, then $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{m} \mathbf{x}' \mathbf{Q}_i \mathbf{x}$. In order to translate this result into the context of analysis of distance, we need to sum the relationship over the p vectors of values $\mathbf{x}_i$ of the separate variables. If the mean-centered vectors $\mathbf{x}_i$ are collected as the columns of matrix $\mathbf{X}$, the sum can be written most concisely as $\text{trace}(\mathbf{X}'\mathbf{X}) = \sum_{i=1}^{m} \text{trace}(\mathbf{X}' \mathbf{Q}_i \mathbf{X})$. Simple algebraic manipulation [Gower, Krzanowski 1999] then leads to

$$\mathrm{T} = \frac{1}{n}\mathbf{1}'\mathbf{D}\mathbf{1} = \sum_{i=1}^{m} \text{trace}(-\mathbf{Q}_i \mathbf{D} \mathbf{Q}_i), \qquad (2)$$

and this forms the analysis of distance breakdown with $r_i = \text{rank}(\mathbf{Q}_i)$ giving the number of degrees of freedom for each term on the right-hand side. If significance tests are required then they can be obtained by permutational methods (see, e.g. [Manly 1997]). However, our primary interest here is on descriptive methods; we can follow up any terms with $r_i > 1$ by either a metric scaling of the relevant $-\mathbf{Q}_i \mathbf{D} \mathbf{Q}_i$ matrices or by decomposing the term into single degree of freedom contrasts.

2.4. Oslofjord data analysis

Each factor has 7 levels, so it is possible to choose 6 orthogonal contrasts between the levels in a number of ways. For example, Krzanowski [2002] gives two possibilities: orthogonal polynomial contrasts and Helmert contrasts. Since the levels of both disturbance and organic enrichment are approximately equally spaced on the logarithmic scale, orthogonal polynomials would be a reasonable choice. Taking the 6 orthogonal contrasts (either from [Krzanowski 2002] or from any table of orthogonal polynomials) as columns of a 7×6 matrix $\mathbf{C}$, we build up the $\mathbf{H}$ matrices as follows. For the main effect of factor A, put into the ith row of $\mathbf{H}_A$ the row of $\mathbf{C}$ corresponding to the level of A to which the ith individual is subjected, and likewise for the main effect of factor P in $\mathbf{H}_P$. These two matrices are thus both of size 98×6. For the AP interaction, simply multiply each column of $\mathbf{H}_A$ elementwise by each column of $\mathbf{H}_P$ to give a 98×36 matrix $\mathbf{H}_{AP}$. $\mathbf{Q}$ matrices are formed from each of these $\mathbf{H}$ matrices as given above, and then the terms on the right-hand side of equation (2) provide the analysis of distance breakdown given in Table 1.

Table 1. Analysis of distance of Oslofjord data

Source	Sum of squares	Degrees of freedom
Main Effect A	5441.34	6
Main Effect P	13736.75	6
AP Interaction	14452.27	36
Error	69895.45	49
Total	103525.81	97

Source: own calculations.

Permutational testing shows that both main effects and interaction are significant. To investigate the two main effects further, the sum of squares can be broken down into individual degree of freedom contrasts based on the orthogonal polynomials (by taking each diagonal element rather than trace of the relevant $-\mathbf{Q}_i\mathbf{D}\mathbf{Q}_i$ matrices), yielding the values in Table 2. Inspection of these values suggests that only the linear contrast is important for the main effect of factor A, but both linear and quadratic contrasts may be important for the main effect of factor B. These suppositions were confirmed by permutational tests, with only these named contrasts giving significant results.

Table 2. Single degree of freedom contrasts for Oslofjord Data

Contrast	Main effect A	Main effect P
Linear	3174.3547	8476.4433
Quadratic	437.2924	3216.2308
Cubic	375.5216	906.0794
Quartic	625.8358	532.2458
Quintic	392.3985	351.4672
Sextic	435.9411	254.2839
Total	5441.3441	13736.7504

Source: own calculations.

Reasons for these significances may be investigated by conducting metric scaling of each relevant $-\mathbf{Q}_i\mathbf{D}\mathbf{Q}_i$ matrix, but this is deferred till we have considered biplots.

3. Biplots

3.1. Theory

By converting the original data matrix $\mathbf{Z}$ to distances $\mathbf{D}$, and then focussing on the configuration $\mathbf{X}$ resulting from metric scaling of $\mathbf{D}$, we have lost all information on the variables originally measured. But often we want to know which variables (e.g. species for Oslofjord) contribute most to the effects of interest. So how can we recover such information?

Let: (i) $\mu e_k = (0,...,0,\mu,0,...,0)'$ represent variable k in the space of $\mathbf{Z}$,

(ii) $\mathbf{d}_k(\mu)$ contain 0.5 times the squared distances from $\mu\,\mathbf{e}_k$ to each individual,

(iii) $\mathbf{d}_0$ contain 0.5 times the squared distances from each individual to the origin.

Then, following the initial metric scaling, the coordinates of $\mu\,\mathbf{e}_k$ in the space of $\mathbf{X}$ are given [Gower 1968] by

$$\mathbf{y}_k(\mu) = -(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'(\mathbf{d}_k(\mu) - \mathbf{d}_0).$$

As μ varies, $\mathbf{y}_k(\mu)$ traces out a trajectory, the *biplot vector,* representing the kth variable in the space of $\mathbf{X}$; see Gower and Hand [1996] for a full exposition of biplots.

However, these biplot vectors will in general be non-linear so tricky to plot. We can linearise them by selecting a suitable number m (say 10) of values of μ and finding $\mathbf{u}_k$ such that $\mathbf{y}_{ki} = a_{ki}\,\mathbf{u}_k + \varepsilon_{ki}$ for $i = 1, \ldots, m$ and each k. Least squares gives the equations

$$\mathbf{u}_k = (\sum_{i=1}^{m} a_{ki}\,\mathbf{y}_{ki})/(\sum_{i=1}^{m} a_{ki}^2) \text{ where } a_{ki} = (\mathbf{u}_k'\mathbf{y}_{ki})/(\mathbf{u}_k'\mathbf{u}_k) \text{ for } i = 1, \ldots, m.$$

A simple alternating scheme starting from $a_{ki} = 1$ converges to the solution fairly rapidly. This gives linearised biplots in the metric scaling space $\mathbf{X}$, but where do these vectors lie in the metric scaling space of any of the $-\mathbf{Q}_i\mathbf{D}\,\mathbf{Q}_i$ matrices? The trick is to notice that these spaces can be equivalently obtained by a principal component analysis of the relevant factorial means derived from $\mathbf{X}$. Thus if $\mathbf{L}$ contains the principal component coefficients as columns, then the biplot vectors have coordinates $\mathbf{L}'\mathbf{u}_k$. For full details of the necessary computations, see [Krzanowski 2004].

3.2. Oslofjord data analysis continued

Returning to the analysis of the data, metric scaling was conducted on each of the $-\mathbf{Q}_i\mathbf{D}\,\mathbf{Q}_i$ matrices derived from the two factorial main effects (i.e. the $\mathbf{Q}_i$ matrices obtained from $\mathbf{H}_A$ and $\mathbf{H}_P$ respectively), and linearised biplots of a selection of species were projected into the resulting two-dimensional spaces. Figure 1 shows the space for the main effect of factor A, and Figure 2 shows the space for the main effect of factor B. Factor levels are shown as points, and biplots as lines, in each figure. The projection of the points onto axis 1 in Figure 1 shows a linear ordering of the factor levels from left to right, demonstrating the clear linear effect that was established from the single degree-of-freedom contrasts; no further patterns are evident among the levels on axis 2. In Figure 2, on the other hand, not only is there a clear linear ordering of levels of P on axis 1 (right to left this time) but there is also a very obvious quadratic pattern of the levels on axis 2, supporting the single degree-of-freedom breakdown of Table 2. The effects of the species can then be gauged by noting the positions of the projections of the points onto each

biplot vector. For example, projecting onto the vector for species 12 in Figure 1 shows three groups of virtually coincident levels (A0,A1), (A2,A3,A5), and (A4,A6), and reference back to the original data confirms a decreasing trend in the abundances at these levels with values 41, 26 and 19 respectively. Krzanowski [2004] provides further discussion of these results.

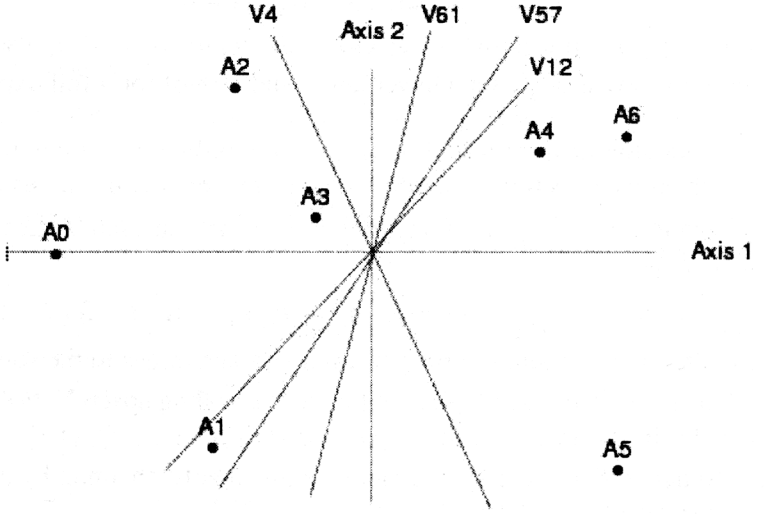

Figure 1. Two-dimensional metric scaling diagram for main effect of Factor A, with selected biplot vectors

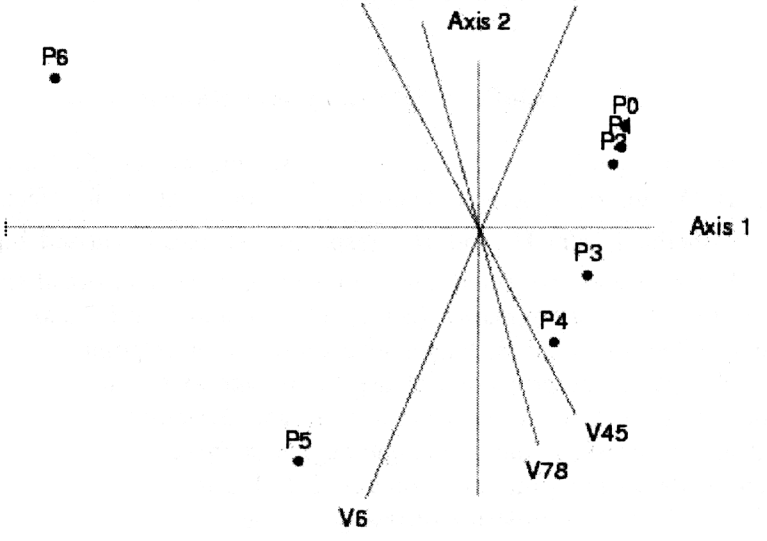

Figure 2. Two-dimensional metric scaling diagram for main effect of factor P, with selected biplot vectors

4. Sensitivity of solutions

A natural question to ask is how stable are the results we have obtained: can we attach any standard errors, or equivalent measures of variability, to the derived quantities? These questions relate to the stability of the underlying metric scaling, but very few results have so far been available in this area. Ramsay [1982] proposed a parametric probability model for the process, which could in principle form the basis of stability calculations, but no-one seems to have followed it up. As far as data-based approaches are concerned, only one possible jackknife scheme has been proposed [DeLeeuw, Meulman 1986], but this involves heavy computation and there have been few instances of its use in practical applications. However, this scheme is a very general one, and if we focus on the specific form of metric scaling used above then there is considerable analytical simplification. Basic steps are as follows.

We assume that we have formed the matrix $\mathbf{X}$ of coordinates from metric scaling of $\mathbf{D}$ via (1). Deleting the ith row and column of $\mathbf{D}$ yields the $(n-1)\times(n-1)$ matrix $\mathbf{D}_{(i)}$ that corresponds to the matrix $\mathbf{D}$ when the ith individual is omitted from the original data. Then finding the eigenvalues and eigenvectors of

$$-\left(I-\frac{11'}{n-1}\right)\mathbf{D}_{(i)}\left(I-\frac{11'}{n-1}\right),$$

produces the matrix of coordinates $\mathbf{X}_{(i)}$ in which the point corresponding to the ith row of $\mathbf{X}$ is missing but points corresponding to all other rows of $\mathbf{X}$ are present. Positioning the missing row in the space of $\mathbf{X}_{(i)}$ using the result of Gower [1968] produces an augmented matrix $\mathbf{X}^{(i)}$ containing n rows. Repeating this whole process for each row/column of $\mathbf{D}$ yields n augmented matrices $\mathbf{X}^{(1)}$, ..., $\mathbf{X}^{(n)}$. Then conducting analyses of distance for each augmented matrix will give an indication of the stability of the original analysis. For example, constructing biplots from each augmented matrix gives a set of "perturbed" biplots, the distribution of whose angles with the original biplots indicates the variability inherent in the original biplots. Full details of the computations are given by Krzanowski [2006].

Table 3: Median angles in degrees over biplot cross-validations

Biplot vector	Full space	Main effect A	Main effect B	AP Interaction
V4	85.12	41.09	61.47	78.19
V6	69.77	87.97	69.91	61.89
V12	83.03	58.80	86.63	9.79
V45	24.05	59.69	10.49	27.65
V57	70.03	83.55	49.51	62.78
V61	55.27	83.79	42.51	37.27
V78	63.49	12.55	61.81	83.65

Source: own calculations.

Table 3 gives median angles in degrees over such cross-validations for the biplot vectors in Figures 1 and 2, within all possible spaces (i.e. full space, subspaces for main effects of A and P, and AP interaction subspace). It can be seen that species V45 is the most stable (lowest median angles) across all these spaces, while species V78 is the most stable with respect to main effect of A and V12 is most stable as regards the AP interaction.

5. Comment

This has been a necessarily brief review, but the hope is that sufficient detail has been provided for the reader to be able to apply the various analyses that have been outlined to other data sets. However, some cautionary points need to be borne in mind at all stages. Most descriptive analyses rely on just two-dimensional graphical displays, and these should always be first checked for adequacy using any of the standard MDS measures (e.g. STRESS, proportion of eigenvalues to trace, etc). The variability inherent in multidimensional scalings should also be remembered, and when comparing two MDS diagrams the arbitriness of signs of axes should be allowed for (see [Krzanowski 2006] for suggestions). In breaking effects down into single degrees of freedom, care should be taken over choosing appropriate H matrices. Finally, if biplots are computed it should be remembered that the linearity is only an approximation; residual sums of squares from the least squares fitting will give an indication as to how close the approximation really is. If these cautionary points are attended to, then the methods surveyed above should provide useful analyses in many situations.

References

Anderson M.J. (2001), *Permutation Tests for Univariate or Multivariate Analysis of Variance and Regression*, „Canadian Journal of Fisheries and Aquatic Sciences" 58, p. 626-639.

Anderson M.J., Willis T.J. (2003), *Canonical Analysis of Principal Coordinates: A Useful Method of Constrained Ordination for Ecology*, „Ecology" 84, p. 511--525.

Bray J.R., Curtis J.T. (1957), *An Ordination of the Upland Forest Communities of Southern Wisconsin*, „Ecological Monographs" 27, p. 325-349.

Cuadras C.M., Arenas C. (1990), *A Distance Based Regression Model for Prediction with Mixed Data*, „Communications in Statistics – Theory and Methods" 19, p. 2261-2279.

d'Aubigny G.D. (1988), *The Additive Decomposition of Some Entropy Functions and (Constrained) Ordination Methods, Proceedings of the XIVth International Biometric Conference*, p. 455-485.

Deleeuw J., Meulman J. (1986), *A Special Jackknife for Multidimensional Scaling*, „Journal of Classification" 3, p. 97-112.

Digby P.G.N., Gower J.C. (1981), *Ordination between- and within-Groups Applied to Soil Classification*, [in:] *Down-to-Earth Statistics: Solutions Looking for Geological Problems*, ed. D.F. Merriam, p. 63-75, Syracuse University Geological Contribution, New York.

Gower J.C. (1966), *Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis*, „Biometrika" 53, p. 325-338.

Gower J.C. (1968), *Adding a Point to Vector Diagrams in Multivariate Analysis*, „Biometrika" 55, p. 582-585.

Gower J.C. (1985), *Measures of Similarity, Dissimilarity and Distance*, [in:] *Encyclopedia of Statistical Sciences*, vol. 5, edp. S. Kotz, N.L. Johnson, C.B. Read, John Wiley and Sons, New York.

Gower J.C., Hand D.J. (1996), *Biplots*, Chapman and Hall, London.

Gower J.C., Krzanowski W.J. (1999), *Analysis of Distance for Structured Multivariate Data and Extensions to Multivariate Analysis of Variance*, „Applied Statistics" 48, p. 505-519.

Krzanowski W.J. (2002), *Multifactorial Analysis of Distance in Studies of Ecological Community Structure*, „Journal of Agricultural, Biological and Environmental Statistics" 7, p. 222-232.

Krzanowski W.J. (2004), *Biplots for Multifactorial Analysis of Distance*, „Biometrics" 60, p. 517-524.

Krzanowski W.J. (2006), *Sensitivity in Metric Scaling and Analysis of Distance*, „Biometrics" 62, in presp.

Krzanowski W.J., Marriott F.H.C. (1994), *Multivariate Analysip. Part 1: Distributions, Ordination and Inference*, Kendall's Library of Statistics, Edward Arnold, London.

Legendre P., Anderson M.J. (1999), *Distance-Based Redundancy Analysis: Testing Multispecies Responses in Multifactorial Ecological Experiments*, „Ecological Monographs" 69, p. 1-24.

Manly B.F.J. (1997), *Randomization and Monte Carlo Methods in Biology*, Chapman and Hall, London.

Mardia K.V., Kent J.T., Bibby J.M. (1979), *Multivariate Analysis*, Academic Press, London.

McArdle B.H., Anderson M.J. (2001), *Fitting Multivariate Models to Community Data: A Comment On Distance-Based Redundancy Analysis*, „Ecology" 82, p. 290-297.

Pillar V. De P., Orloci L. (1996), *On Randomization Testing in Vegetation Science: Multifactor Comparisons of Releve Groups*, „Journal of Vegetation Science" 7, p. 585-592.

Ramsay J.O. (1982), *Some Statistical Approaches to Multidimensional Scaling Data (with Discussion)*, „Journal of the Royal Statistical Society" Series A, 145, p. 285-312.

Widdicombe S., Austen M.C. (2001), *Interactions between Physical Disturbance and Organic Enrichment: An Important Element in Structuring Benthic Communities*, „Limnology and Oceanography" 46, p. 1720-1733.

ANALIZA ODLEGŁOŚCI WIELOWYMIAROWYCH DANYCH STRUKTURALNYCH: PRZEGLĄD I PRZYKŁAD

Streszczenie

Wiele wielowymiarowych zbiorów danych ma strukturę sugerującą wykorzystanie analizy MANOVA, jednakże nie spełniają jej wymaganych założeń. Od niedawna zainteresowanie wzbudza analogiczny typ analizy, przeprowadzanej jednak na macierzy niepodobieństw otrzymanej z nieprzetworzonych danych. Niniejsza praca zawiera przegląd podstaw takiego podejścia, pokazuje główne etapy takiej analizy. Zaprezentowano również, w jaki sposób biploty i analiza wrażliwości mogą być dodane do metody podstawowej, a także ilustruje przedstawione pojęcia za pomocą zbioru danych pochodzących z ekologii.