COENOSES 12(2-3):145-148, 1997

Software abstract

MULTIVARIATE EXPLORATORY ANALYSIS AND RANDOMIZATION TESTING WITH MULTIV

Valério DePatta Pillar

Department of Plant Sciences, The University of Western Ontario, London, Ontario, N6A 5B7, Canada.
Permanent address and address for correspondence: Departamento de Ecologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 91540-000, Brazil; Fax +55 51 3166936; E-mail pillar@vortex.ufrgs.br

MULTIV is a computer application for flexible exploratory analysis and randomization testing with multivariate data. It can handle qualitative, quantitative and mixed data types, offering several options for data transformation, resemblance measures, ordination and clustering techniques, some of which are also available in other analytical packages (e.g., Podani 1994, Wildi & Orlóci 1996). The results are presented in text files and graphs.

Its novelty is that integrated to the exploratory tools are options for randomization testing (Fig. 1). The tests may involve univariate or multivariate comparisons between groups of sampling units defined by one or more factors (Pillar & Orlóci 1996). This is especially useful in the analysis of variance due to factors and interactions in experimental and survey data. Randomization tests are also available for pairwise comparisons of variables (see Manly 1991).

The program is written in C++ and is available as a Mac OS application, in optimized versions for the different CPUs, including native PowerPC. A limited version (**MultivMinor**) is available for demonstration or for handling small data sets. This version and the user's guide can be retrieved from ftp://ftp.ucs.ubc.ca/pub/mac/info-mac/sci/multiv-102.hqx. The application is not yet available for Windows OS.

The user interface is interactive, menu driven on text based screens. The options available are in general the ones that are applicable to the type of data informed by the user and analytical step that has been reached. Major data entry is from text files, containing observations of one or more vari-

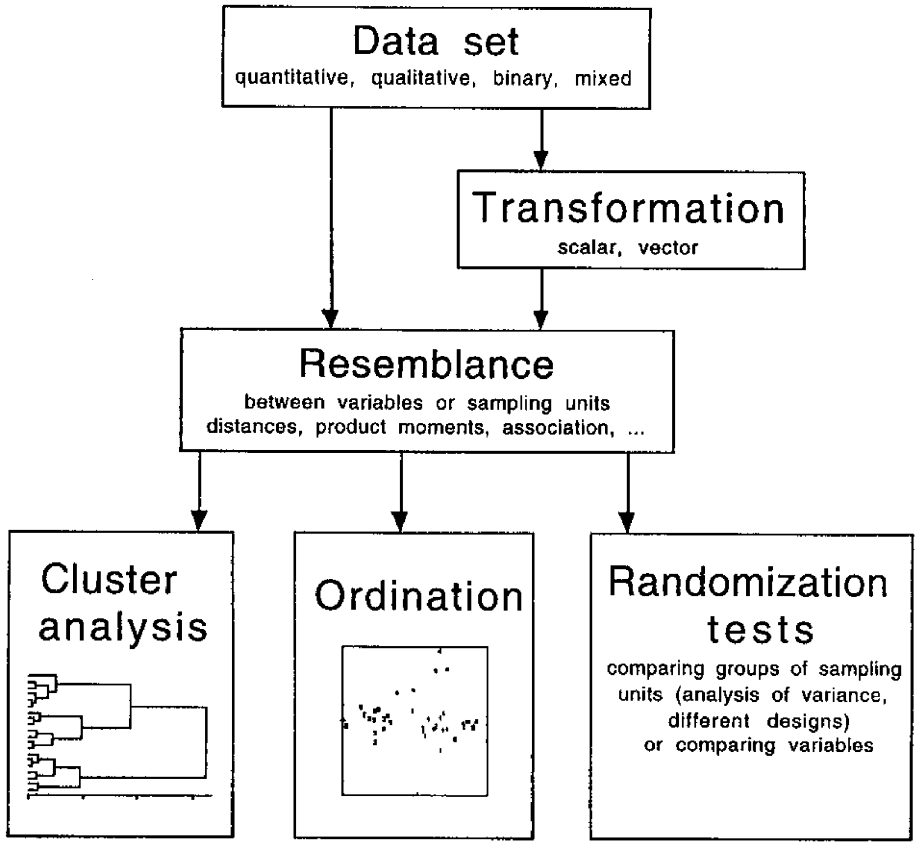

Figure 1. Flow diagram for exploratory analysis and randomization testing with MULTIV, showing options available at different analytical steps.

Table 1. Resemblance measures offered by MULTIV for different data types. The description of the resemblance measures are in the user's guide and references therein (Pillar 1997). The first yes (y) or no (n) refers to comparison between sampling units, the second to comparison between variables.

	Quant. Same units	Quant. Different units	Qualit.	Binary	Mixed Quant.+ qualit.	Mixed Quant.+ binary	Mixed Qual.+ binary	All mixed
Product moment	y y	n n	n n	y y	n n	n n	n n	n n
Absolute value function	y y	n n	n n	y y	n n	n n	n n	n n
Euclidean distance	y y	n n	n n	y y	n n	n n	n n	n n
Mutual information	y y	n n	n n	n n	n n	n n	n n	n n
Gower index	y n	y n	y n	y n	y n	y n	y n	y n
MDIS	n n	n n	n y	n y	n n	n n	n y	n n
Mutual entropy	n n	n n	n y	n y	n n	n n	n y	n n
Rajski's metric	n n	n n	n y	n y	n n	n n	n y	n n
Coherence coefficient	n n	n n	n y	n y	n n	n n	n y	n n
Chi-square	n n	n n	n y	n y	n n	n n	n y	n n
Jaccard	n n	n n	n n	y y	n n	n n	n n	n n
Simple matching	n n	n n	y n	y y	n n	n n	y n	n n
Sokal-Michener	n n	n n	n n	y y	n n	n n	n n	n n
Ochiai	n n	n n	n n	y y	n n	n n	n n	n n
Sorensen	n n	n n	n n	y y	n n	n n	n n	n n
Mean square contingency	n n	n n	n n	y y	n n	n n	n n	n n
Correlation	y y	n y	n n	y y	n n	n y	n n	n n
Chord distance	y y	n y	n n	y y	n n	n y	n n	n n

ables in a set of sampling units. The data can be arranged in matrix format or in a free format. The numerical results are stored on a text file, which can be open with any text editor. Scatter diagrams and dendrograms can be stored as picture files.

The application can run in the background while the user operates on other applications that do not use the CPU as intensively as **MULTIV**, which is especially useful in randomization testing with large data sets. Though randomization testing is very demanding in computations, runs can be very fast in present day microcomputers. As an example, in a Macintosh with a 200 MHz 603e PowerPC CPU, the program took less than one minute to generate 10,000 iterations to test main factors and interactions in a design with two factors and 60 experimental units.

Data transformation

Data transformation is available only when data are quantitative, binary or mixed with both. In scalar transformations the transformed value depends on the non-transformed value only, while in vector transformation it depends on the set of observations in the same variable (within variables), or in the same sampling unit (within sampling units) or in both (double adjustment). Vector transformations within sampling units are only applicable when the variables are measured in the same scale (or standardized).

Resemblance measures

The resemblance measure options available are the ones applicable to the type of data in hand (see Table 1). Note that some measures carry implicit transformations.

Table 2. Output by MULTIV of the test for independence of floristic composition from landscape and grazing intensity in Campos grassland (Pillar & Orlóci 1996).

Elapsed time: 50.9000 seconds
Dimensions: 60 sampling units, 60 variables
Resemblance measure: chord distance between sampling units
Number of iterations: 10000
Group partition of sampling units:

Sampling units:	1	2	3	4	5	6	7	8	9	10	11	12	13	... 60
Landscape:	1	1	1	1	1	1	1	2	2	2	3	3	4	... 4
Grazing:	1	2	1	1	1	1	1	1	1	2	1	1	1	... 1

Source of variation	Sum of squares (Qb)	$P(Qb^{\circ} \geq Qb)$
Between groups by landscape/grazing	14.30	0.0001
Between groups by landscape	9.521	0.0001
Contrasts:		
1 1 1 -3	6.270	0.0001
-1 -1 2 0	2.419	0.0003
1 -1 0 0	0.8326	0.1841
Between groups by grazing	3.271	0.0001
Interaction landscape x grazing	1.506	0.8744
Within groups by landscape/grazing	24.18	
Total	38.48	

Ordination

Ordination is offered if a resemblance matrix is available. Three methods based on eigenanalysis, i.e., principal coordinates analysis, principal components analysis and correspondence analysis (see Podani 1994) are available. Two dimensional scatter diagrams can be produced on screen using as axes the available ordination scores.

Cluster analysis

The methods that are offered, i.e., single linkage, complete linkage and minimum variance, are based on the resemblance matrix available (between variables or between sampling units) and follow an agglomerative algorithm (see Pielou 1984, Podani 1994). The dendrogram is presented on the screen.

Randomization tests comparing variables

This option is offered when the available resemblance is between variables. The test criterion is any pair-wise resemblance function already defined and pertinent to the type of variables, e.g., correlation coefficient, chi-square, coherence coefficient, Euclidean distance. If the Null Hypothesis (H_o) of no association (correlation) between the variables pair-wise is true, the state observed in a variable in a given sampling unit is independent from the states observed in the other variables (see Manly 1991, for details). At each iteration, a random data set is generated by shuffling the observations in each variable among the sampling units, the resemblance matrix is computed and each resemblance value for a variable pair (r_{ik}°) is compared to the corresponding value found in the observed data set (r_{ik}). After many iterations, if $P(|r_{ik}^{\circ}| \geq |r_{ik}|)$ is small (smaller than α), H_o is rejected and we conclude that there is significant association (correlation) between variables i and k. The results present probabilities for all variable pairs.

Randomization tests comparing groups of sampling units

This option is offered when the available resemblance is between sampling units, provided groups of sampling units are also specified. The method is described in Pillar & Orlóci (1996). Complete randomized, uni- or multifactorial and block designs are allowed. Specific contrasts between groups

can be tested as in Scheffé's test. Contrast coefficients indicate which groups are involved in the contrast. The test criterion is a sum of squares between classes Q_b. The randomization follows the Null Hypothesis (H_o) that there is no difference between the groups. If the H_o is true, the observation vector in a given sampling unit is independent from the group to which the unit belongs. Based on this, at each iteration a random data set is generated and a Q_b^o is computed. After a large number of iterations, the result is a probability $P(Q_b^o \geq Q_b)$. These probabilities are interpreted similarly to the ones in an analysis of variance table. A typical result is in Table 2.

Acknowledgments: The author was recipient of a CAPES (Brazil) fellowship during his leave at UWO, Canada.

References

Manly, B. F. J. 1991. Randomization and Monte Carlo Methods in Biology. London, Chapman & Hall. 281 p.

Pielou, E. C. 1984. The Interpretation of Ecological Data; a Primer on Classification and Ordination. New York, J. Wiley. 263 p.

Pillar, V. D. & L. Orlóci. 1996. On randomization testing in vegetation science: multifactor comparisons of relevé groups. Journal of Vegetation Science 7: 585-592.

Pillar, V. D. 1997. MULTIV: Multivariate Exploratory Analysis and Randomization Testing. User's Guide v. 1.1. Universidade Federal do Rio Grande do Sul, Brazil / University of Western Ontario, Canada.

Podani, J. 1994. Multivariate Data Analysis in Ecology and Systematics. The Hague, SPB Academic Publishing. 316 p.

Wildi, O. & L. Orlóci. 1996. Numerical Exploration of Community Patterns; a guide to the use of Mulva-5, 2nd ed. SPB Academic Publishing, The Hague.