Relative performance of non-metric multidimensional scaling in vegetation studies: an application to the Lama Forest Reserve (Benin)

Introduction

In vegetation studies, ordination aims at arranging samples and/or species along a few axes which must represent the main compositional gradients in the data set, using either abundance or presence/absence data (Økland 1996). From Pearson (1901) to today, a huge development has occurred in ordination methods. Although all ordination methods in current use are burdened with defects (Økland 1990) and none of them appropriate under all circumstances (Kenkel & Orloçi 1986), non-metric multidimensional scaling (NMDS) is still undoubtedly the most widely accepted and routinely used ordination technique (Podani 2005).

The first step but extremely important in NMDS is the computation of a matrix of dissimilarities/similarities among a set of items in a multidimensional space (McCune & Grace 2002). In NMDS, several similarity measure indexes can be considered and available in literature (McCune & Grace 2002, Choi 2008). McCune & Grace (2002) recommend using the quantitative Sørensen coefficient whereas with reference to the principle of the method, Palm (2003) considered the similarity index of Sokal & Michener which takes into account the co-presence and co-absence of items. Jaccard index is also one of the most used similarity measure (McCune & Grace 2002). As those indexes express differently the similarity, it is possible for them to result in different ordinations. Standardization, a kind of data transformation which ecologically aims to make distance measures work better, reduce the effect of sample unit totals to put focus on relative quantities, equalize (or otherwise alter) the relative importance of common and rare species or emphasize informative species at the expense of uninformative species (McCune & Grace 2002) also affects results from NMDS (Faith et al. 1987). Standardization should therefore be of importance in ordination, mainly when calculating dissimilarities or similarities (Faith et al. 1987).

As mentioned above, NMDS can be applied either on binary (presence-absence) or abundance data. We hypothesize that as more information is provided by abundance data, they are expected to result in more reliable ordination compared to presence-absence data. In addition, sample size is a key part of sampling design. Findings of several studies reported that an increase in sample size invariably results in improved estimation efficiency (Condit et al. 1996). The effect of sample size on the performance of NMDS is then an important issue in vegetation studies.

The purpose of this study was to analyze the relative performance of NMDS in vegetation studies focusing on the effect of sample size, types of data (binary versus abundance data) and similarity (or dissimilarity) measures.

Methodology

Data collection: Data used in this study come from Bonou et al. (2009) and are linked to the identification of plant vegetation communities in the Lama Forest reserve (LFR), a dense semi-deciduous forest in Benin. Data were based on a matrix of Presence–absence of 31 species recorded in 100 plots of 0.15 ha. We have also been provided with the abundance data.

Simulation design: Four factors including the nature of data, the sample size, the similarity indexes and the type of data standardization were considered in this simulation design. The basic matrix used is the abundance data matrix (let M). Two types of data matrix were considered: binary and abundance data matrices. The binary matrices were drawn from the abundance data matrices by replacing all non-zero values with 1. Four values of the sample size were considered by truncating (or not) the original data set: 25, 50, 75 and 100 plots. This was done using bootstrap, a resampling method (Efron & Tibshirani 1993). For binary data, three similarity indexes were considered: Sokal & Michener, Sorensen and Jaccard indexes. From the plots i and j similarity values (S_{ij}), dissimilarities (d_{ij}) were drawn using the formula (Gower & Legendre 1986): $d_{ij} = \sqrt{(1 - S_{ij})}$. As for abundance data, two dissimilarity indexes were examined: the Sorensen dissimilarity (also known as Bray-Curtis coefficient) and the Jaccard dissimilarity indexes. Four techniques of

standardization (only abundance matrices were concerned) have been used. The first one includes the species adjustment to equal maximum abundances (SPM) i.e. divide abundance of a species *h* in a given plot *i* by the species maximum abundance in the matrix. The second technique was the samples standardization to equal totals (SAT) which is equivalent to the computation of species relative abundance (in %). The two last techniques were the Bray-Curtis successive double standardizations i.e. SPM followed by SAT (let DBL1) and inverse Bray-Curtis successive double standardizations i.e. SAT followed by SPM (let DBL2). For abundance data, each of the 4 sample sizes was combined with the 5 different types of standardization (non standardization was also considered) and each of the two dissimilarities indexes. Forty (4×5×2) combinations were therefore examined for abundance data whereas for binary data, each sample size was only combined with each of the 3 similarity indexes i.e. examination of twelve (4×3) combinations. 500 replications for each combination were generated using the bootstrap technique.

Data analysis: The basic assumption of NMDS is that for a good ordination, there should be a rank-order relationship between inter-sample dissimilarity and inter-sample distance in the ordination space (McCune & Grace 2002). This means that the more similar two samples are, the closer they should be in the ordination space. Provided this assumption, the spearman rank correlation (Rs) was used as criterion of efficiency. We also used the s-stress value (Takane & Young 1977) as criterion of efficiency. It measures the departure from monotonicity in the plot of distance in the original p-dimensional space (dissimilarity) versus distance in the ordination space (k-dimensional space). The closer the points lie to a monotonic line, the better the fit and the lower the stress (Kruskal & Carroll 1969). For each combination of factors considered, the Rs correlation and s-stress values were computed using a group of codes written in MATLAB software (V. R2006a). Boxplots of the Rs and s-stress-values for all combinations of dissimilarity indexes and types of standardization (for abundance data) and similarity indexes (for presence-absence data) were established. A visual analysis of the boxplots helped selecting the best similarity index (binary data) and the best combination of dissimilarity index and standardization (abundance data). This selection was done with respect to the highest values of the Rs and the lowest values of the s-stress. ANOVA was then performed in SAS 9.2 software to test the effect of sample size on efficiency criteria with regard to the best combinations of factors. When the effect of the sample size was significant, contrast analysis (Everitt 2002) was performed to model the relationship between sample size and s-stress values and determine the optimal sample size. In this study, linear (s-stress = $\beta_0 + \beta_1 size + \varepsilon$) and quadratic (s-stress = $\beta_0 + \beta_1 size + \beta_1 size^2 + \varepsilon$) models were tested.

Results

Efficiency of NDMS according to the factors considered for abundance data

Irrespective to sample size and type of standardizations, the two dissimilarity measures (Sorensen and Jaccard) performed equally (Figure 1). The Spearman rank correlation (Rs) became lower with the increase in the sample size and seemed to stabilize from 75 plots (Figure 1). Unlike the dissimilarity index and the sample size, the Rs varied greatly among types of standardization. For most of the combinations of factors considered, the standardization to equal totals (SAT) yielded higher Rs values (> 0.939) than those produced by the others. It was followed respectively by no standardization (0), the inverse of the Bray-Curtis double standardization (DBL2), the Bray-Curtis double standardization (DBL1) and Species adjustment to equal maximum abundances (SPM) standardization (0.893). The same trend was noted for s-stress values (Figure 2): regardless to the sample size, the lower values of the s-stress were obtained for SAT. This standardization therefore performed better than the others. Figures 1 and 2 clearly showed that the s-stress values decreased when the Rs values increased. The former ranged from 0.120 to 0.241. A closer examination of Figure 2 denotes an increase in s-stress value with sample size. From these descriptive analyses, we deduced that SAT was the best standardization and the two dissimilarity indexes (Sorensen and Jaccard) were not distinguishable for both Rs and s-stress. Results from analysis of variance (ANOVA) showed significant (Prob.<0.05) difference only for s-stress, with

regard to each of the two dissimilarity indexes, indicating the significant effect of sample size on s-stress values. The linear and quadratic models of the relationship between sample size and s-stress values were highly significant. The quadratic model was then retained to determine the optimum sample size (Figure 3) which was 90 plots with a s-stress value of 0.167.

Efficiency of NDMS according to the factors considered for binary data

Results revealed a decrease of Rs values, from 0.921 (for the Sorensen index) to 0.907 (for the Jaccard index) when the sample size increased (Figure 4). As with abundance data, the dispersion around the median value also decreased when sample size increased. Furthermore, for a given sample size, the three similarity indexes examined yielded approximately the same value. But, s-stress values increase with sample size. The s-stress values ranged from 0.10 (Sokal & Michener index) to 0.17 (Sorensen or Jaccard index). Whatever is the sample size, Sokal & Michener index yielded the lowest values of s-stress, and therefore was retained for further investigations. The ANOVA performed to test the effect of sample size showed a significant difference (Prob.<0.05) only for s-stress values. The linear and the quadratic models were also shown to be highly significant. The relationship between s-stress values and sample size for the quadratic model was thus plotted (Figure 5) and indicated an optimum of 75 plots with a s-stress value of 0.120.

Discussion and Conclusion

This study is a complement to previous investigations on designing accurate and strong way for vegetation data analysis. Consistently with Faith et al. (1987) and McCune & Grace (2002), results obtained showed that type of standardization greatly affects NMDS efficiency. Some of them improve ordinations in contrary to others. The standardization to sample totals (SAT) was revealed to be the most outperformed. It then appears that NMDS perform better when plots have similar weight (number of trees or species). Species adjustment to equal maximum abundance (SPM) standardization often results in poor ordination. When applied alone, it was the least successful standardization and then is very little recommended. But when used in combination with SAT, the SPM is however preferable to be used before to being used after SAT. These results somewhat contrast those of Faith et al. (1987) who found the SPM standardization to perform better than SAT for some dissimilarity indexes as Canberra metric and Chi-squared. This may suggest that standardization effect varies according to the indexes since many dissimilarity coefficients have in-built standardization (Faith et al. 1987).

The most critical step in selecting the appropriate method of ordination is the choice of dissimilarity index which must be compatible with available data (Podani 2006). Results showed that using quantitative version of either Jaccard or Bray-Curtis dissimilarity coefficient, the NMDS yielded the same result indicating that despite their mathematical difference (in reference to their formulas) both are similar. This was confirmed by the Spearman rank and Pearson linear correlation between the two coefficients which were respectively 1 and 0.997. The same observations can be drawn for binary matrices. Here however, Sokal & Michener similarity index showed the best result. In addition to the co-presence which is common to all the three indexes, this index takes into account the co-absence of species when computing similarity of a couple of plots (Palm 2003). Sokal & Michener similarity index can therefore be suggested for use in NMDS as long as the data matrix is binary. We then conclude that similarity indexes do affect NMDS efficiency. However, since only three indexes were examined in this study, it is possible for others dissimilarity indexes to perform better than Sokal & Michener. Choi (2008) and Podani (2001) have actually described respectively 76 indexes for binary data and 17 indexes for ratio scale data.

The stress-value increased with the number of plots and is consistent with the increase of objects being scaled since stress can be viewed as a variance (McCune & Grace, 2002). Indeed, the increase in sample size means an increase in objects to scale and therefore an increase in the stress-value. However, the lower coefficient of variation obtained with the increase in sample size suggests the higher the sample size the more accurate the scaling.

Binary matrices were shown to be more efficient than abundance matrices from which they derived. Binary matrices yielded high values of Rs and the low stress values, probably because of

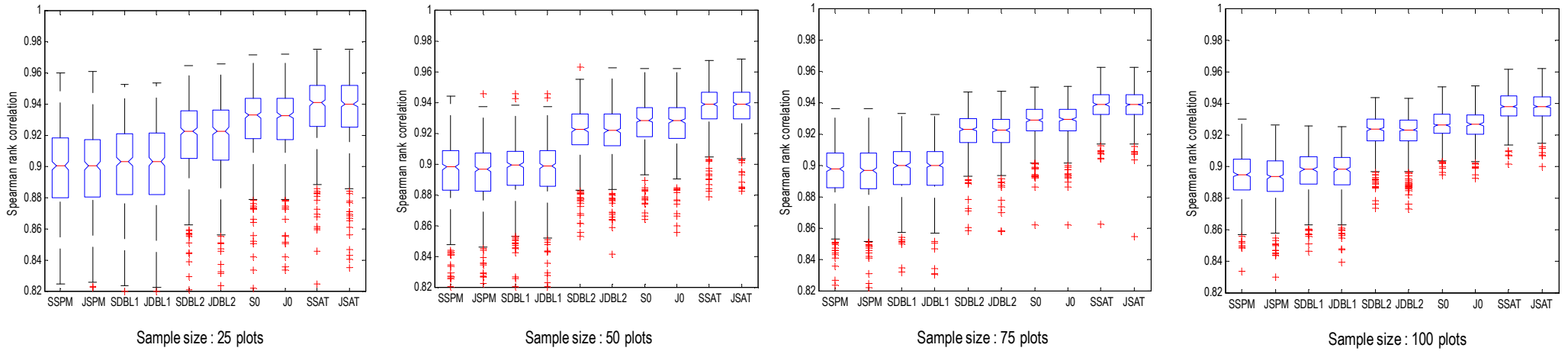

Figure 1. Box plots of Spearman rank correlation for each combination of similarity index, type of standardization and sample size for abundance data.

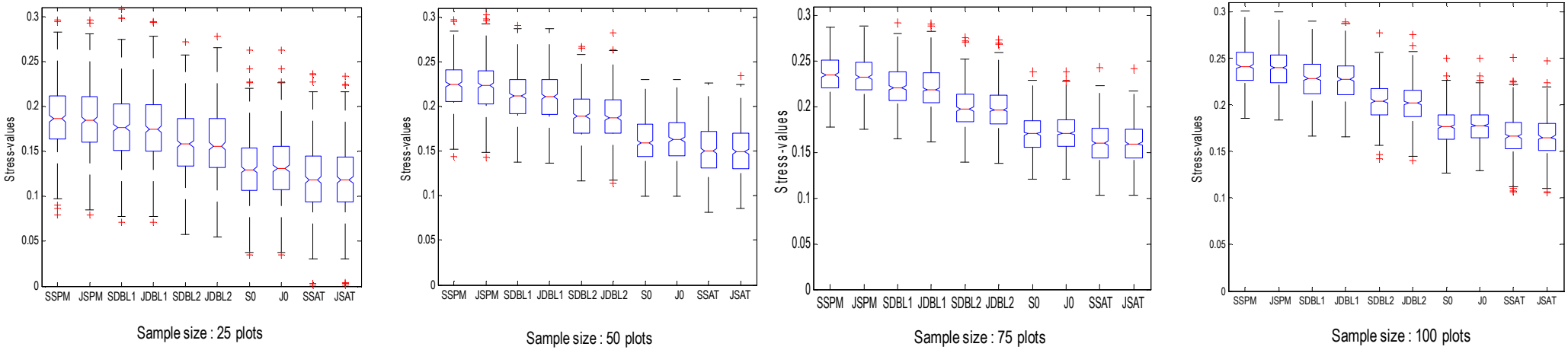

Figure 2. Boxplots of s-stress-values for each combination of similarity index, type of standardization and sample size for abundance data.

Legend: On the x-axis, first letters of variables are initials of the dissimilarity index (S=Sorensen; J=Jaccard); the following are relative to the types of standardization (0 = No standardization; SPM=Species adjustment to equal maximum abundances; SAT=standardization to equal totals; DBL1=SPM followed by SAT; DBL2=SAT followed by SPM). Example: SSPM correspond to combination of Sorensen index and SPM.

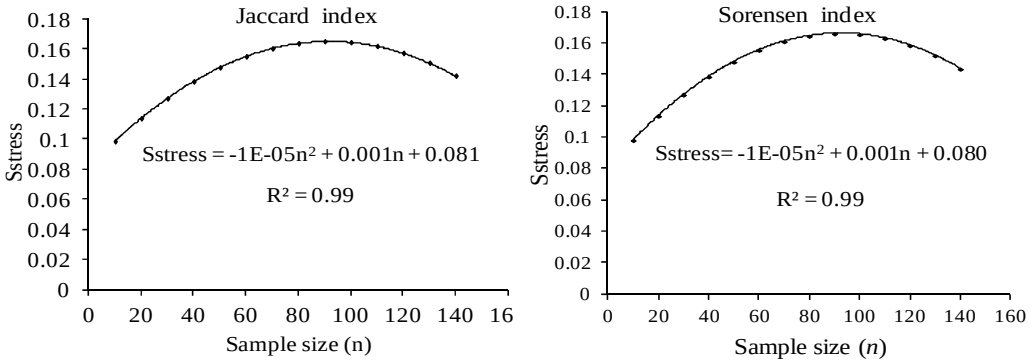

Figure 3. Relationship between s-stress-values and sample size for Jaccard and Sorensen dissimilarities.

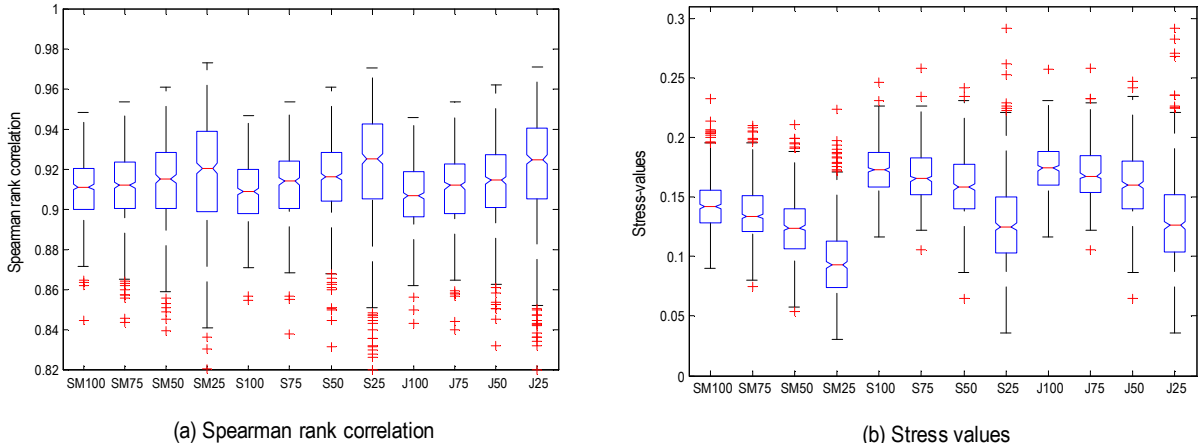

Figure 4: Boxplots of Rs and s-stress-values for each combination of similarity index and sample size for binary data.

Legend: On the x-axis, first letters of variables are initials of the similarity index (S=Sorensen; J=Jaccard; SM=Sokal and Michener) and the following are linked to the sample size. Example: SM25 correspond to combination of Sokal & Michener similarity index and sample size of 25 plots.

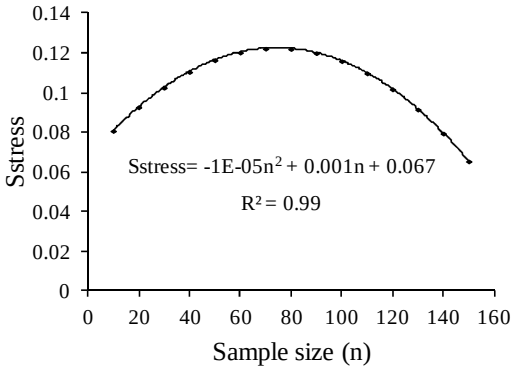

Figure 5. Relationship between the s-stress-values and sample size for Sokal & Michener index.

the small differences between pairs of objects with this type of matrix in comparison to abundance matrices. In fact, two plots containing the same species will be viewed to be much closed with presence-absence similarity indexes. But a slightly difference in species abundance can result in a great distinction with quantitative (ratio scale) dissimilarity coefficient.

The methodology of this study used the 2-dimensions spaces for the initial configuration. The number of dimensions should be determined for each data matrix before choosing the dimension of the initial configuration (Kruskal 1964a&b). In fact, the determination of the number of dimensions to use in the ordination space is an important issue in NMDS. With simulated data, this is known as a priori but for real data, a better method could be to use the dissimilarity matrix to calculate the linkages of the minimum spanning tree (Gower and Ross, 1969). However, the first few dimensions are sufficient to explain most of the variation (Podani 2001). Besides this aspect, Shepard (1962) has strongly argued for solutions in two dimensions as this is more readily interpretable.

Literature cited

Bonou, W., Glèlè Kakaï, R., Assogbadjo, A.E., Fonton, H.N., Sinsin, B., 2009, Characterization of *Afzelia africana* Sm. habitat in the Lama Forest reserve of Benin. Forest ecology and management 258, 1084–1092.

Choi, S.S., 2008, *Correlation Analysis of Binary Similarity Measures and Dissimilarity Measures*, Doctorate dissertation, Pace University.

Condit, R., Hubbell, S.P., Lafrankie, J.V., Sukumar, R., Manokaran, R., Foster, R.B., Ashton, P.S., 1996, Species-area and species-individual relationships for tropical trees: a comparison of three 50-ha plots. Journal of Ecology 84, 549-562.

Efron, B., Tibshirani, R.J., 1993, *An introduction to the bootstrap*. Chapman and Hall, New York, New York, USA.

Everitt, B.S., 2002, *Cambridge Dictionary of Statistics*, CUP, ISBN 0-521-81099-x

Faith, D.P., Minchin, P.R., Belbin, L., 1987, Compositional dissimilarity as a robust measure of ecological distance. Vegetatio, 69: 57-68.

Gower, J.C., Ross G.J.S., 1969, Minimum spanning trees and single linkage cluster analysis. Applied Statistics. 18, 54-64.

Gower, J.C., Legendre P., 1986, Metric and Euclidean properties of Dissimilarity Coefficients. Journal of Classification 3, 5-48.

Kenkel, N.C., Orlóci, L., 1986, Applying Metric and Nonmetric Multidimensional Scaling to Ecological Studies: Some New Results. Ecology, 67(4), 919-928.

Kruskal, J.B., 1964a, Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. Psychometrika, 29, 1-27.

Kruskal, J.B., 1964b, Nonmetric multidimensional scaling: a numerical method. Psychometrika, 29, 115-129.

McCune, B., Grace, B.J., 2002, *Analysis of Ecological Communities*. Oregon, USA. 300p.

Økland, R.H., 1990, Vegetation ecology: theory, methods and applications with reference to Fennoscandia. Sommerifeltia Suppl 1, 1-233.

Økland, R.H., 1996, Are Ordination and Constrained Ordination Alternative or Complementary Strategies in General Ecological Studies? Journal of Vegetation Science, 7(2), 289-292.

Palm, R., 2003, *Notes de statistique et d'informatique. Le Positionnement multidimensionnel: Principes et application*. 33 p.

Podani, J., 2000, *Introduction to the exploration of multi-variate biological data*. Backhuys, Leiden, NL.

Podani, J., 2001, SYN-TAX 2000. *Computer Programs for Data Analysis in Ecology and Systematic. User's Manual*. Budapest, Hungary, 53p.

Podani, J., 2006, Braun-Blanquet's legacy and data analysis in vegetation science. Journal of Vegetation Science 17(1), 113-117.

Shepard, R.N., 1962, The analysis of proximities. Multidimensional scaling with an unknown distance function. I. Psychometrika, 27, 125-140.

Takane, Y., Young, F.W., 1977, Non-metric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika, 42(1), 7-67.