Journal of Vegetation Science 20: 596–602, 2009
© 2009 International Association for Vegetation Science

Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity

Roleček, Jan[1,2*]; Tichý, Lubomír[1,3]; Zelený, David[1,4] & Chytrý, Milan[1,5]

[1]*Department of Botany and Zoology, Masaryk University, Kotlářská 2, CZ-611 37 Brno, Czech Republic;* [2]*Department of Vegetation Ecology, Institute of Botany, Academy of Sciences of the Czech Republic, Poříčí 3b, CZ-603 00 Brno, Czech Republic;* [3]*E-mail tichy@sci.muni.cz;* [4]*E-mail zeleny@sci.muni.cz;* [5]*E-mail chytry@sci.muni.cz;* [*]*Corresponding author; Fax +420 532146213; E-mail honza.rolecek@centrum.cz*

Abstract

Aim: To propose a modification of the TWINSPAN algorithm that enables production of divisive classifications that better respect the structure of the data.

Methods: The proposed modification combines the classical TWINSPAN algorithm with analysis of heterogeneity of the clusters prior to each division. Four different heterogeneity measures are involved: Whittaker's beta, total inertia, average Sørensen dissimilarity and average Jaccard dissimilarity. Their performance was evaluated using empirical vegetation datasets with different numbers of plots and different levels of heterogeneity.

Results: While the classical TWINSPAN algorithm divides each cluster coming from the previous division step, the modified algorithm divides only the most heterogeneous cluster in each step. The four tested heterogeneity measures may produce identical or very similar results. However, average Jaccard and Sørensen dissimilarities may reach extreme values in clusters of small size and may produce classifications with a highly unbalanced cluster size.

Conclusions: The proposed modification does not alter the logic of the TWINSPAN classification, but it may change the hierarchy of divisions in the final classification. Thus, unsubstantiated divisions of homogeneous clusters are prevented, and classifications with any number of terminal clusters can be created, which increases the flexibility of TWINSPAN.

Keywords: Compositional data analysis; Correspondence analysis; Dataset heterogeneity; Dissimilarity measures; Divisive classification; Heterogeneity measures; Hierarchical classification; Numerical classification; Plant community; Vegetation classification.

Abbreviations: AJD = Average Jaccard dissimilarity; ASD = Average Sørensen dissimilarity; CA = Correspondence analysis.

Introduction

Numerical classification remains a widespread tool for simplification of complex multivariate relationships in ecological datasets (Orlóci 1978; Legendre & Legendre 1998; McCune & Grace 2002). Hierarchical classifications are particularly popular in community ecology, as they describe the structure of the datasets in a way that corresponds to the traditional view of hierarchical relationships among communities. They also allow choice of the appropriate number of clusters by choosing an appropriate level of division after inspecting the whole hierarchy. The prize paid for knowing the hierarchy is that the assignment of sites to clusters at a given level is, to some degree, determined by the higher (divisive techniques) or lower (agglomerative techniques) levels of the hierarchy (Gauch & Whittaker 1981).

Agglomerative techniques construct the classification from the bottom up, i.e. they begin with clustering of the most similar sites and aggregate these into still larger clusters until there is a single cluster containing all sites. Divisive techniques construct the classification from the top down, i.e. they begin with all sites in a single cluster that is successively divided until individual sites are separated. Agglomerative and divisive techniques differ in emphasis on the higher or lower levels of the hierarchy; while the former reflect well the relationships between individual sites at lower levels of the hierarchy, the latter perform particularly well in the topmost divisions of the datasets (Lambert et al. 1973; Gauch & Whittaker 1981).

One of the most popular hierarchical divisive classification techniques in community ecology is Two-way Indicator Species Analysis (TWINSPAN). This was developed by M. O. Hill and incorporated in a FORTRAN program of the same

name (Hill 1979). It is currently widely used, being distributed in stand-alone DOS and WINDOWS versions (Hill & Šmilauer 2005) and included in software packages for ecological data analysis, e.g. PC-ORD (McCune & Mefford 1999).

The TWINSPAN algorithm starts with primary ordination of sites along the first axis of correspondence analysis (CA; Hill 1973, 1974). In the next step, sites are divided into two clusters by splitting the first CA axis near its middle. Then, site classification is refined using a discriminant function that emphasizes species preferential to one or the other half of the dichotomy (see Hill & Šmilauer 2005 for details). The process can be repeated, each of the two clusters being repeatedly divided in the same way.

TWINSPAN performed well in several comparisons of classification techniques (Gauch & Whittaker 1981; Cao et al. 1997; Moss et al. 1999), although substantial criticisms have also been raised (van Groenewoud 1992; Belbin & McDonald 1993; Dufrêne & Legendre 1997). Interestingly, the crucial reservations of some authors against the method were considered advantages by others: TWINSPAN does not extract "natural" groups of the most similar sites, but partitions the dissimilarity space determined by the main gradients in the data, disregarding possible discontinuities. This ability determines the usefulness of TWINSPAN in particular classifications. When the sites are naturally clustered, i.e. grouped in dissimilarity space, and we want to recover clusters of the most similar sites, some agglomerative clustering techniques will probably be more successful. In contrast, when the sites are dispersed rather regularly in dissimilarity space and we seek a classification reflecting the main gradients of variability of the whole dataset, TWINSPAN may be very useful (Gauch & Whittaker 1981; Dale 1995).

In vegetation classifications, TWINSPAN remains a popular method despite the above-mentioned criticisms. Probably, it is not just because of the tradition and availability of the program, but also because most vegetation datasets show rather continuous variability, and vegetation scientists tend to look for the classifications reflecting main gradients in species composition that follow key ecological factors. However, a major technical limitation of TWINSPAN, which restricts its usability in many cases, is that the number of clusters of the final classification cannot be set manually, but is determined by a simple divisive rule: in each division step, each cluster is divided into two smaller clusters, and each of these is divided again in the next step. Thus the number of clusters increases in powers of two ($2 \rightarrow 4 \rightarrow 8 \rightarrow 16 \rightarrow 32 \rightarrow$ etc.) going down the hierarchy. The only rule incorporated in TWINSPAN that is capable of changing this sequence is the "minimum group size for division." This is specified by the user of the program and determines the minimum number of sites in a group, below which the group is not divided any further. In some cases, this rule creates hierarchies that do not respect the structure of the data, and in any case it is not able to produce partitions with a pre-selected number of clusters.

To improve the method, we propose here a modification of the TWINSPAN algorithm that helps to produce divisive classifications that better respect the structure of ecological data; at the same time, the number of clusters is not constrained to be powers of two, as in the classical TWINSPAN.

Methods

The principle of the TWINSPAN modification

The proposed modification combines the classical TWINSPAN algorithm with analysis of heterogeneity of the clusters prior to each division. First, TWINSPAN divides the dataset into two clusters; then TWINSPAN is stopped and a preselected measure of heterogeneity is calculated for both clusters. In the next step, only the more heterogeneous cluster is divided by TWINSPAN; then there are three clusters, heterogeneity of each of them is again quantified and only the most heterogeneous cluster is divided into two clusters by TWINSPAN, etc. The process is repeated until the number of clusters specified by the user is reached. The differences between classical and modified TWINSPAN algorithms are illustrated in Fig. 1.

Measures of cluster heterogeneity

Cluster heterogeneity can be measured in different ways. We applied four measures that are commonly used in ecology: Whittaker's beta, i.e. the ratio of the total number of species over all sites in the cluster to the average number of species per site (Whittaker 1960); average Jaccard and Sørensen dissimilarity of all pairs of sites (Jaccard 1912; Sørensen 1948; Koleff et al. 2003); and total inertia, as used in CA (Greenacre 2000).

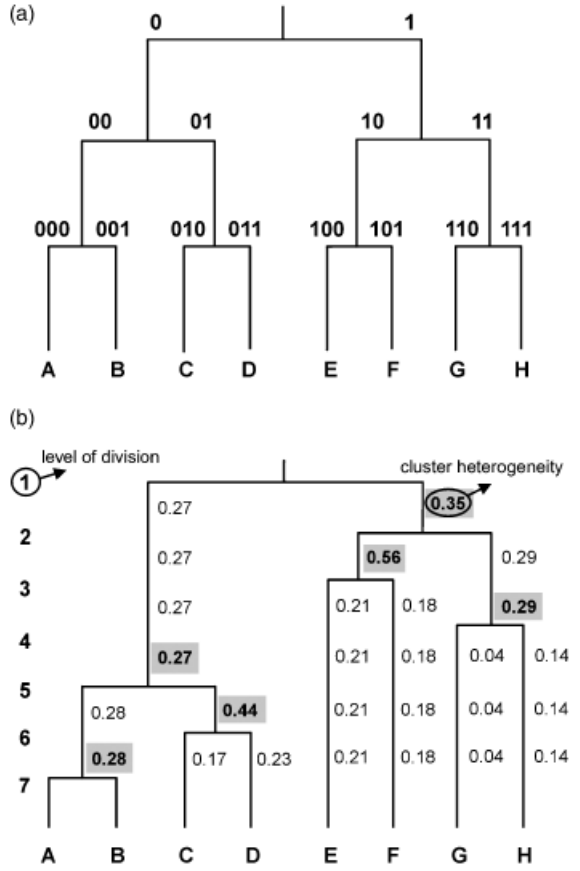

Fig. 1. The differences between classification hierarchies produced by (a) classical and (b) modified TWINSPAN algorithms. Classical TWINSPAN produces classifications, where each initial cluster is divided into two smaller clusters on each hierarchical level. Thus, the number of clusters increases in powers of two, and divisions of some fairly homogeneous clusters may be imposed by this simple divisive rule. Modified TWINSPAN involves the analysis of heterogeneity of the clusters prior to each division. The imposed divisions of homogeneous clusters are thus prevented and the classification is built step-by-step.

Empirical evaluation of the new method

We tested the performance of the new method with several datasets from vegetation plots. For illustration of the basic principle of the method, we selected a set of 823 vegetation plots from the Czech National Phytosociological Database (Chytrý & Rafajová 2003), which included vegetation of mesic oak-hornbeam forests (*Carpinion betuli* alliance) and an outlying group of dry oak forests (*Genisto pilosae-Quercetum petraeae* association). To explore the behaviour of different measures of cluster heterogeneity, we compared their performance using real datasets of vegetation plots. In particular, we investigated the dependence of heterogeneity measures on the number of sites in the dataset and on increasing homogeneity of the dataset of the same size.

We randomly selected 10 000 vegetation plots of deciduous forest vegetation from the Czech National Phytosociological Database as the basic dataset. Species cover values were transformed to presence/absence. From this dataset, we prepared five subsets of different heterogeneity using CA. In the first step, 500 plots were selected randomly from the basic dataset, forming the most heterogeneous subset (H1). After that, the initial dataset of 10 000 plots was subjected to CA, and plots were sorted according to their scores on the first axis. In the next step, 5000 plots with the lowest scores were taken, and of these 500 plots were selected randomly, forming a subset with lower heterogeneity (H2). The above subset of 5000 plots was again subjected to CA, and 2500 plots with the lowest score on the first axis were taken, and, of these, 500 plots were selected randomly (H3). Following this procedure, we created subsets of 1250 and 625 plots, and from each we randomly selected 500 plots (H4, H5). In such a way, we obtained five datasets of equal size whose heterogeneity decreased from H1 to H5.

We then took each subset (H1–H5) and randomly selected groups of increasing number of plots, starting with a group of ten plots and adding ten plots in each step, up to group size of 250 plots. Selection of plots per each group size was repeated five times. In total, 625 groups of plots were created, 125 from each subset. We calculated different dissimilarity measures for each of these groups and then compared values of one dissimilarity measure among the groups derived from subsets of different heterogeneity.

To determine the response shape of Jaccard or Sørensen dissimilarities to different numbers of plots in a group, we again used subset H5, and randomly selected groups of increasing number of plots using a finer stepping. We started with a group of five plots and added two plots in each step up to 250 plots per group, and each selection from each group size was repeated five times (total 615 groups of plots).

Classical TWINSPAN was calculated using the stand-alone TWINSPAN program (Hill 1979), a version compiled by S. M. Hennekens in 2003 for the MEGATAB program, part of the TURBOVEG software (Hennekens & Schaminée 2001). The modified TWINSPAN was operated from the JUICE program (Tichý 2002). All other calculations were performed in R (R Development Core Team

2005), including the detrended correspondence analysis (DCA, Hill & Gauch 1980). All analyses were done on presence/absence data, and all our conclusions refer to this.

Results and Discussion

Performance of the modified algorithm

The performance of classical and modified TWINSPAN algorithms is demonstrated in Fig. 2, where four clusters resulting from each classification are plotted on the ordination diagram of DCA. This figure illustrates the major advantage of the modified TWINSPAN: while the classical algorithm divides all the clusters coming from the previous division step, the modified algorithm divides only the most heterogeneous cluster in each step. Thus, unsubstantiated divisions of homogeneous clusters are prevented, and the number of clusters of the resulting classification is not constrained to be the powers of two, as in the classical TWINSPAN.

Behaviour of different heterogeneity measures

Although the idea of involving some data heterogeneity measure in the TWINSPAN algorithm is quite simple and straightforward, the selection of a particular heterogeneity measure may alter the structure of the resulting classification and its meaning. This is because different measures treat differently two aspects of data heterogeneity, which we call *range* and *complexity*. By range, we mean the extent of dissimilarity of all sites in a group along the main gradient in species composition, as determined by the length of the first CA axis. By complexity, we mean variability in dissimilarity dimensions other than the one that defines the range. Complexity increases with the dataset size, or more specifically, with the total number of species, i.e. the gamma diversity. While Whittaker's beta and total inertia reflect both range and complexity, the average Jaccard and Sørensen dissimilarities (AJD, ASD) consider only range, and thus they are independent of the dataset size until the range of its heterogeneity is fixed. This is shown in Fig. 3: the values of Whittaker's beta and total inertia increase both with range and complexity; by contrast, AJD and ASD do not change until the range of heterogeneity of the dataset is fixed. It should be noted that the similar behaviour of AJD and ASD is a trivial result of the fact that these two measures are very similar in computation (Koleff et al. 2003).

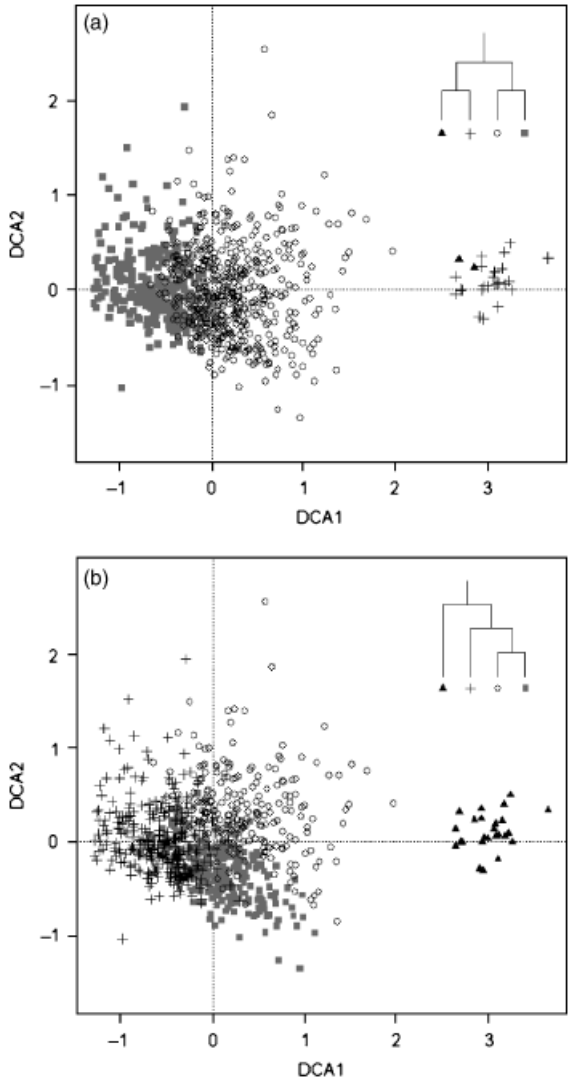

Fig. 2. Detrended correspondence analysis ordination diagrams showing the results of (a) classical TWINSPAN classification and (b) modified TWINSPAN classification. The dendrogram in the upper right corner of each figure indicates the hierarchy of divisions. An identical result is obtained with all measures of cluster heterogeneity discussed in the paper. This rather extreme example was chosen for the purpose of illustration; many ecologists would prefer to remove the outlying cluster for separate analysis.

Our examinations indicate that the four tested measures of cluster heterogeneity may produce identical or very similar results. However, in some cases, they do not, and then it is necessary to choose one of them. The results of simulations in Fig. 3 show that although both measures of complexity (Whittaker's beta and total inertia) are dataset size-dependent, they also vary substantially with changing heterogeneity of the dataset, while keeping the

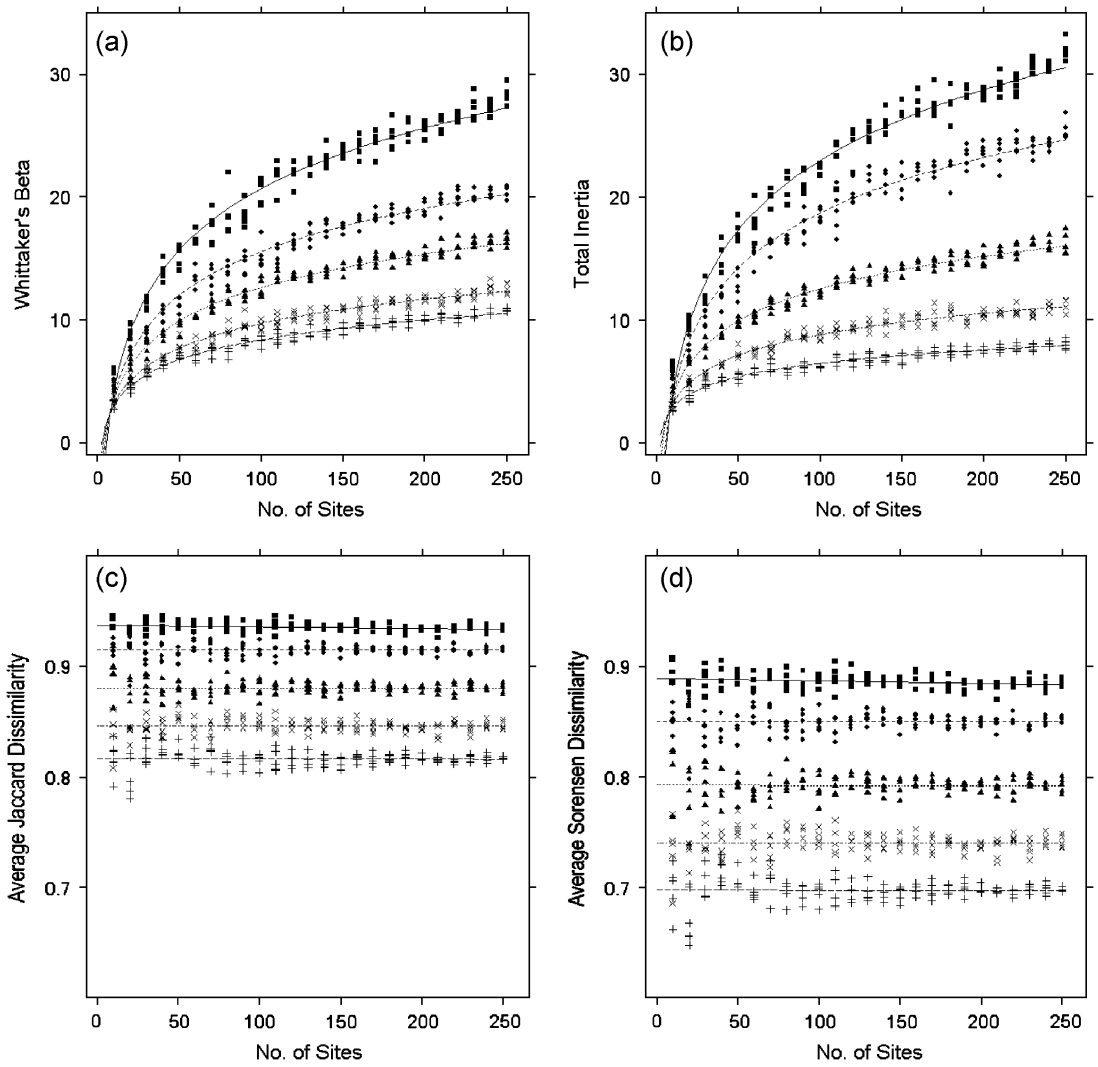

Fig. 3. Relationships between cluster size and heterogeneity calculated from datasets of decreasing heterogeneity using different heterogeneity measures: (a) Whittaker's beta, (b) total inertia, (c) average Jaccard dissimilarity and (d) average Sørensen dissimilarity. Symbols for datasets, ranked by decreasing heterogeneity, are: ■ H1, ● H2, ▲ H3, × H4, + H5.

dataset size constant. Therefore, we consider them to be suitable measures of heterogeneity for the purpose of identification of the candidate cluster for further division in TWINSPAN classifications. Their tendency to measure higher heterogeneity in larger clusters fits the intuitive approach to classification of ecological communities: if there are very large clusters in the dataset, users tend to divide them, even though they are more homogeneous than small clusters, whose representativeness of the vegetation type may be questioned.

The measures of range (average Jaccard or Sørensen dissimilarities), independent of the dataset size, often also give reasonable results. Unfortunately, these measures may occasionally measure higher heterogeneity in smaller groups, which may result in chaining of the dendrogram of modified TWINSPAN classifications. This behaviour can be explained as a simple result of poor representativeness of small groups: with decreasing group size, the estimate of the average becomes instable and may reach extreme values. This is illustrated in Fig. 4 as a funnel-shaped response of dissimilarities to group size.

If chaining occurs with AJD or ASD, it is advisable to use some of the measures of complexity (Whittaker's beta or total inertia), which are more robust in this respect and provide balanced classifications, respecting both heterogeneity and group size. Alternatively, the minimum group size option in TWINSPAN can be set to a higher value to prevent divisions of small clusters.

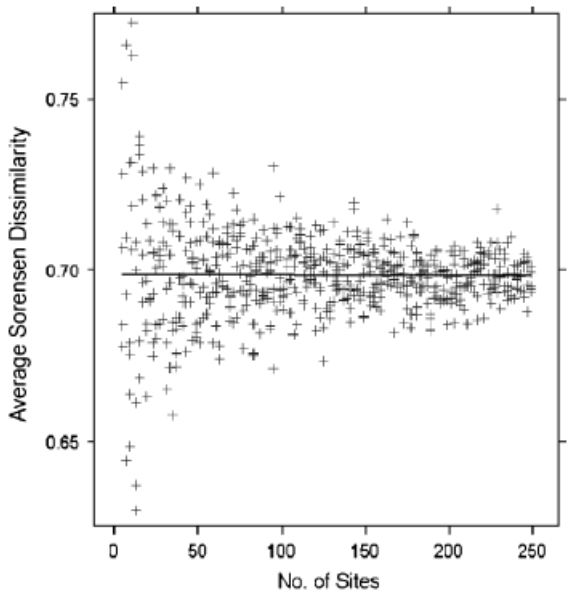

Fig. 4. Funnel-shaped response of average Sørensen dissimilarity to increasing cluster size. Only the least heterogeneous dataset H5 was used for calculation.

The relation to TWINDEND

The idea of calculating heterogeneity of TWINSPAN groups to identify groups for further division was proposed earlier by Vandvik & Birks (2004). They used a proportion of mean dispersion (heterogeneity) within each group relative to the mean dispersion within the whole dataset, based on the procedure developed by Orlóci (1967). Although the authors did not provide any detailed description of their algorithm (it is contained in the TWINDEND program of J. M. Line & H. J. B. Birks, unpublished data), they probably measured heterogeneity using chord distance (Orlóci 1967). The approach of Vandvik & Birks (2004) is analogous to ours, but the heterogeneity measure involved may be unsatisfactory in some cases. Chord distance has the same tendency to divide small groups as average Jaccard or Sørensen dissimilarities; an option to choose other measures is necessary in such a case. Moreover, chord distance is designed for quantitative data, while TWINSPAN typically works with presence/absence data. Although there is no conceptual reason not to use quantitative information for the calculation of cluster heterogeneity, measures for presence/absence data must also be available.

Possible alternative solutions

An alternative strategy for modifying the hierarchy of TWINSPAN classifications could be based on a bottom-up approach consisting of: (1) building a classical TWINSPAN tree, (2) calculating the heterogeneity for each cluster, and (3) pruning the tree, i.e. merging the clusters with low heterogeneity up to the desired level. Also, similarity of the terminal clusters could be calculated and the most similar clusters could be merged, regardless of the position in the dendrogram.

Although we consider both of these alternatives as promising ways of improving ecological classifications, we consider our solution as most in the spirit of classical TWINSPAN. It neither alters the logic of the original divisive classification, nor does it introduce new criteria for evaluation of cluster quality. Our motivation for involving a heterogeneity measure in the TWINSPAN algorithm was not to *produce* homogeneous clusters, but to find some ecologically meaningful criterion for *choosing* the most appropriate clusters for division. We believe that TWINSPAN classifications make sense (and became popular) not because their clusters are homogeneous, but because they are ecologically meaningful, and we consider it necessary to retain this quality.

Conclusions

In conclusion, we would like to emphasize that the modification proposed here does not alter the logic of the TWINSPAN classification, which we consider to be intuitive and useful. However, it may change the hierarchy of divisions in the final classification. Thus, modified TWINSPAN classifications avoid imposed divisions of homogeneous clusters at the higher levels of the classification hierarchy. At the same time, they are able to provide any number of terminal clusters, which increases the flexibility of TWINSPAN. The method is available in JUICE, a program for ecological data analysis (Tichý 2002), which is freeware available at http://www.sci.muni.cz/botany/juice.

Acknowledgements. This study was supported by the institutional long-term research plans AV0Z60050516 and MSM 0021622416 and GA AV ČR grant KJB601630504.

References

Belbin, L. & McDonald, C. 1993. Comparing three classification strategies for use in ecology. *Journal of Vegetation Science* 4: 341–348.

Cao, Y., Bark, A.W. & Williams, W.P. 1997. A comparison of clustering methods for river benthic community analysis. *Hydrobiologia* 347: 25–40.

Chytrý, M. & Rafajová, M. 2003. Czech National phytosociological database: basic statistics of the available vegetation-plot data. *Preslia* 75: 1–15.

Dale, M.B. 1995. Evaluating classification strategies. *Journal of Vegetation Science* 6: 437–440.

Dufrêne, M. & Legendre, P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs* 67: 345–366.

Gauch, H.G. & Whittaker, R.H. 1981. Hierarchical classification of community data. *Journal of Ecology* 69: 537–557.

Greenacre, M. 2000. Correspondence analysis of square asymmetric matrices. *Applied Statistics* 49: 297–310.

Hennekens, S.M. & Schaminée, J.H.J. 2001. TURBOVEG, a comprehensive database management system for vegetation data. *Journal of Vegetation Science* 12: 589–591.

Hill, M.O. 1973. Reciprocal averaging: an eigenvector method of ordination. *Journal of Ecology* 61: 237–249.

Hill, M.O. 1974. Correspondence analysis: a neglected multivariate method. *Applied Statistics* 23: 340–354.

Hill, M.O. 1979. *TWINSPAN – A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes*. Ecology and Systematics, Cornell University, Ithaca, NY, US.

Hill, M.O. & Gauch, H.G. 1980. Detrended correspondence analysis, an improved ordination technique. *Vegetatio* 42: 47–58.

Hill, M.O. & Šmilauer, P. 2005. *TWINSPAN for Windows Version 2.3*. Centre for Ecology and Hydrology & University of South Bohemia, České Budéjovice, CZ.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. *New Phytologist* 11: 37–50.

Koleff, P., Gaston, K.J. & Lennon, J.J. 2003. Measuring beta diversity for presence–absence data. *Journal of Animal Ecology* 72: 367–382.

Lambert, J.M., Meacock, S.E., Barrs, J. & Smartt, P.F.M. 1973. AXOR and MONIT: two new polythetic-divisive strategies for hierarchical classification. *Taxon* 22: 173–176.

Legendre, P. & Legendre, L. 1998. *Numerical ecology*. 2nd edition. Elsevier, Amsterdam, NL.

McCune, B. & Grace, J.B. 2002. *Analysis of ecological communities*. MjM Software Design, Gleneden Beach, OR, US.

McCune, B. & Mefford, M.J. 1999. PC-ORD. Multivariate analysis of ecological data, version 4.0. MjM Software Design, Gleneden Beach, OR, US.

Moss, D., Wright, J.F., Furse, M.T. & Clarke, R.T. 1999. A comparison of alternative techniques for prediction of the fauna of running-water sites in Great Britain. *Freshwater Biology* 41: 167–181.

Orlóci, L. 1967. An agglomerate method for classification of plant communities. *Journal of Ecology* 55: 193–206.

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

R Development Core Team 2005. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. http://www.r-project.org. Accessed 24 February 2007.

Sørensen, T.A. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter* 5: 1–34.

Tichý, L. 2002. JUICE, software for vegetation classification. *Journal of Vegetation Science* 13: 451–453.

Vandvik, V. & Birks, H.J.B. 2004. Mountain summer farms in Røldal, western Norway – vegetation classification and patterns in species turnover and richness. *Plant Ecology* 170: 203–222.

van Groenewoud, H. 1992. The robustness of Correspondence, Detrended Correspondence, and TWINSPAN analysis. *Journal of Vegetation Science* 3: 239–246.

Whittaker, R.H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* 30: 279–338.

Received 8 June 2007;
Accepted 4 March 2008.
Co-ordinating editor: J. Oksanen.