EUROPEAN VEGETATION SURVEY, 16TH WORKSHOP, ROME (ITALY), MARCH 22 - 26, 2007

New stopping rules for TWINSPAN

Lubomír Tichý, Jan Roleček, David Zelený & Milan Chytrý

Department of Botany and Zoology, Masaryk University, Kotlářská 2, CZ-611 37 Brno, Czech Republic

VEGETATION SCIENCE GROUP
MASARYK UNIVERSITY BRNO

One of the most popular hierarchical divisive clustering techniques in community ecology is the Two-way Indicator Species Analysis (TWINSPAN). It was developed by Hill (1979) and until nowadays it is widely distributed in the stand-alone DOS or WINDOWS versions or included in software packages for the analysis of ecological data.

A major technical limitation of TWINSPAN is that the number of clusters of the final classification cannot be set manually, but it is determined by the simple divisive rule that each initial cluster is divided into two smaller clusters and each of these is again divided into two smaller clusters. Thus the number of clusters in the hierarchy increases in the sequence 2 › 4 › 8 › 16 › 32 etc.

The new method uses **the standard TWINSPAN algorithm** and calculates the **heterogeneity of each cluster** which is created when going down the classification hierarchy. First, TWINSPAN divides the data set into two clusters. Then TWINSPAN is stopped and a 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, whose heterogeneity is quantified again and only the cluster with the highest heterogeneity is divided into two clusters, i.e. four clusters are obtained, etc. The process is repeated until the number of clusters specified by the user is reached.

Measures of cluster heterogeneity

1. **Whittaker's beta**, proposed for measuring beta diversity:

$$\beta = \alpha / \gamma - 1$$

where α is the mean number of species per site, calculated from all sites belonging to the cluster, and γ is the total number of species across all sites within the cluster.

2. **Total Inertia**, which is the sum of all eigenvalues and reflects the spread of sites around the centroid of correspondence analysis.

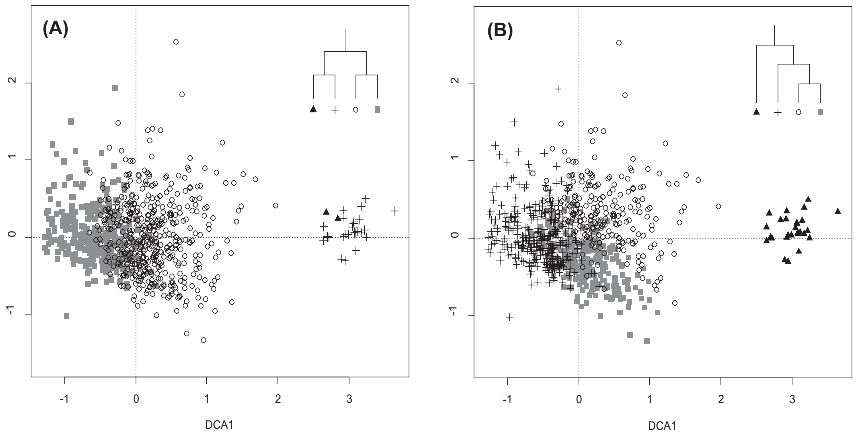
3. **Average Jaccard Dissimilarity** (AJD) between all pairs of sites in a cluster. For one pair of sites, it is calculated as:

$$AJD = 1 - a / (a + b + c)$$

where a is the number of species occurring in both sites, b is the number of species occurring only in the first site and c is the number of species occurring only in the second site.

Test of the proposed method

For illustration of the basic principle of the method we selected a data set of 823 relevés of oak and oak-hornbeam forest vegetation (*Melampyro nemorosi-Carpinetum, Carici pilosae-Carpinetum, Genisto pilosae-Quercetum petraeae*). The first DCA ordination diagram (A) shows unbalanced divisive classification of standard TWINSPAN algorithm while the second diagram (B) represents modified TWINSPAN classification, where the order of divisions is weighted by cluster heterogeneity.



Conclusion

Proposed new stopping rules for TWINSPAN do not alter the logic of the divisive classification, but they modify the hierarchy of divisions in the final classification tree. Divisive classification with these stopping rules tends to avoid imposed divisions of homogeneous clusters at the higher levels of classification hierarchy. At the same time, it is able to provide any number of clusters, which remarkably increases flexibility of TWINSPAN. Our preliminary tests showed that Whittaker's beta and Total Inertia measures of cluster heterogeneity produce slightly more interpretable results without outliers, but the use of Average Jaccard Dissimilarity is also possible.

Fig. 1. Divisive TWINSPAN classification tree with imperfect hierarchy of clusters.

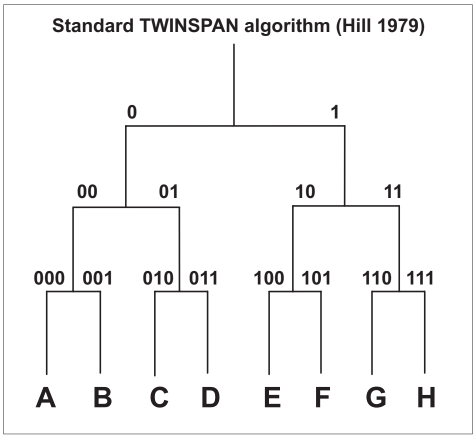


Fig. 2a. The same tree as in Fig. 1 with an identification of internal cluster heterogeneity.

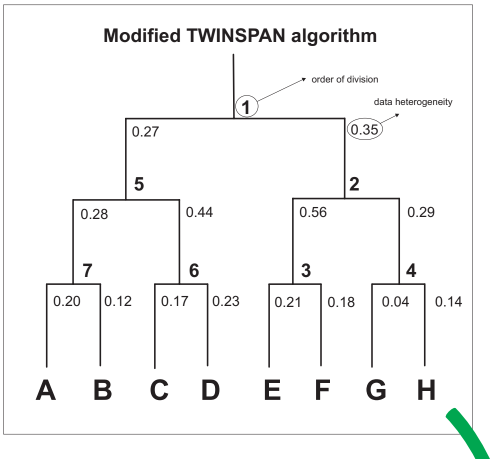
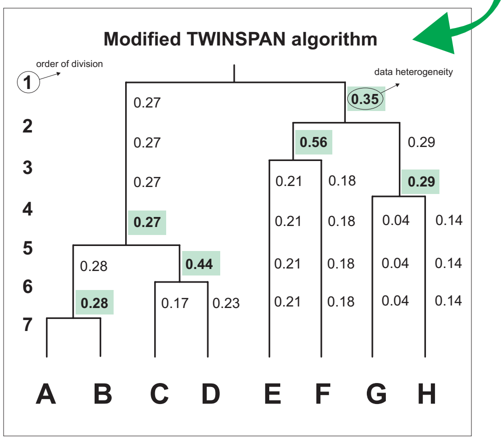


Fig. 2b. Fully hierarchically organized TWINSPAN tree using a comparison of cluster data heterogeneity in each level of division.


Modified TWINSPAN algorithm
order of division
data heterogeneity
1
0.27
0.35
2
0.27
0.56
0.29
3
0.27
0.21
0.18
0.29
4
0.27
0.21
0.18
0.04
0.14
5
0.28
0.44
0.21
0.18
0.04
0.14
6
0.28
0.17
0.23
0.21
0.18
0.04
0.14
7
A
B
C
D
E
F
G
H