Research

Ecography 34: 807–813, 2011
doi: 10.1111/j.1600-0587.2010.06599.x

Subject Editor: Pedro Peres-Neto. Accepted 3 December 2010

Evaluating the stability of the classification of community data

Lubomír Tichý, Milan Chytrý and Petr Šmarda

L. Tichý (tichy@sci.muni.cz), M. Chytrý and P. Šmarda, Dept of Botany and Zoology, Masaryk Univ., Kotlářská 2, CZ-611 37 Brno, Czech Republic.

We propose a method for a posteriori evaluation of classification stability which compares the classification of sites in the original data set (a matrix of species by sites) with classifications of subsets of its sites created by without-replacement bootstrap resampling. Site assignments to clusters of the original classification and to clusters of the classification of each subset are compared using Goodman-Kruskal's lambda index. Many resampled subsets are classified and the mean of lambda values calculated for the classifications of these subsets is used as an estimation of classification stability. Furthermore, the mean of the lambda values based on different resampled subsets, calculated for each site of the data set separately, can be used as a measure of the influence of particular sites on classification stability. This method was tested on several artificial data sets classified by commonly used clustering methods and on a real data set of forest vegetation plots. Its strength lies in the ability to distinguish classifications which reflect robust patterns of community differentiation from unstable classifications of more continuous patterns. In addition, it can identify sites within each cluster which have a transitional species composition with respect to other clusters.

Different types of hierarchical or non-hierarchical cluster analysis are routinely applied in the classification of plant and animal communities. Although the existence of multiple gradients in community data may negatively affect classification stability (repeatability, robustness; Cao et al. 1997, Hennig 2008), community ecologists continue to publish clustering results without any information on how stable these results are. Many studies describe and compare the effects of different clustering methods on classification results in ecology and systematics (Gauch and Whittaker 1981, Podani 1989, Belbin and McDonald 1993, Cao et al. 1997, Brown and Martin 1998, Tichý et al. 2010), but they have hardly ever focused on the stability of clustering results produced by different algorithms after data reduction. In contrast, data resampling has become a standard tool for obtaining statistically verified results in other disciplines (Good 2006). Gauch and Whittaker (1981) defined classification stability (robustness) as a resistance against several types of data set modifications: a) simulation of error or noise, b) random division of the data set into subsets, which are classified separately, c) addition, or d) removal of sites. However, some of these options are of limited use for testing the classification stability in community ecology. The distributions of many species are closely related to those of other species because community patterns consist of complex interactions among species and between species and the environment. A simple addition of noise (a) or artificial sites with a random species composition (c) may bias data set structure and introduce unrealistic species combinations. Classifications of subsets (b) can be compared only indirectly (e.g. using similarities of species compositions between clusters; Illyés et al. 2007, Botta-Dukát 2008) and such tests require a high number of sites within each cluster. Therefore, only random reduction of data sets (d) allows direct comparison of the classification of the original data set with classifications of its subsets.

Bootstrap with replacement is the most popular randomization technique in ecological and environmental applications (Manly 2007). This method was also proposed for testing the classification stability of community data (Pillar 1999, McKenna 2003). However, this type of bootstrap cannot be properly applied to multi-dimensional space, where correlations between variables (species) invalidate the assumptions of the independence of variables and the identical distribution of observations (Brocchieri 2000). In addition, replication of sites in the bootstrap samples simulates community data with a high spatial autocorrelation between some sites or the overrepresentation of sites from particular habitat types. However, ecologists tend to avoid the oversampling of some areas or some habitat types within their data sets by sampling within environmentally and/or spatially defined strata (Austin and Heyligers 1989, Goedickemeier et al. 1997) or by the stratified resampling of large databases prior to their analysis (Knollová et al. 2005). These issues can be overcome by using a modification of bootstrap resampling called

without-replacement bootstrap, which does not permit the inclusion of any site into the bootstrap sample more than once (Shao and Tu 1995, Shuangge 2006).

The aim of this paper is to introduce a simple resampling method which 1) enables a posteriori overall testing of classification stability and 2) identifies transitional sites which may destabilize classification results. We demonstrate this method using artificial and real ecological data sets.

Material and methods

Assessment of classification stability

The newly proposed algorithm repeatedly compares the classification of the original data set with classifications of its randomly selected subsets created by the without-replacement bootstrap.

Consider a data set S of n sites (the original data set). From this data set we generate a with-replacement bootstrap sample S^* of n sites and then remove all duplicated sites from this bootstrap sample. In such a scheme, the bootstrap sample size varies from sample to sample, but its average size is 63.2% of the original data set size n. We classify the original data set and the bootstrap sample to a pre-defined number of clusters using the same classification method and the same input options. Each site included in the bootstrap sample is then labelled with two cluster identifiers, one from the original partition of the whole data set and the other from the partition of the bootstrap sample. The cluster identifiers of sites in the original data set and its bootstrap sample are cross-tabulated and compared using Goodman–Kruskal's lambda index (λ; Goodman and Kruskal 1954), a measure of association based on the proportional reduction in error in cross-tabulation analysis, which was proposed as a consistent and reliable measure for comparing classifications (Podani 1986). The λ is defined as:

$$\lambda = \frac{\sum_i \max_j(n_{i,j}) + \sum_j \max_i(n_{i,j}) - \max(n_{i,.}) - \max(n_{.,j})}{2\left(\sum_i \sum_j (n_{i,j})\right) - \max(n_{i,.}) - \max(n_{.,j})} \quad (1)$$

where $n_{i,j}$ denotes the number of sites of the i-th cluster of the original data set partition which appeared in the j-th cluster of the bootstrap sample partition; $n_{i,.}$ and $n_{.,j}$ are marginal totals of the cross-tabulation. Only sites present in the bootstrap sample are considered. The λ ranges from 0 to 1, with the lowest and highest values indicating minimum and maximum agreement between classifications, respectively. The value of λ is computed for many different without-replacement bootstrap samples and the mean of the λ values based on different bootstrap samples is used as an estimation of classification stability.

The lambda index tends to give better results in the case of higher number of clusters. In the extreme situations, when the number of sites is equal to the number of clusters, λ has a value of 1. Hence we suggest adjustment of the mean λ to suppress the effect of different data set sizes or numbers of clusters:

$$\lambda_{adj} = \frac{\lambda - \lambda_{rand}}{1 - \lambda_{rand}} \quad (2)$$

where λ_{rand} is the mean of the λ values calculated from the random distribution of bootstrap sample sites across the tested number of clusters.

Identification of transitional sites

We assume that each site included in the data set may have a different effect on classification stability. If more sites with a transitional species composition between clusters of the partition of the original data set are included in the bootstrap sample, the agreement between the partition of the original data set and the partition of the bootstrap sample would be lower, and so would be the value of λ. This is because the transitional sites would have a higher probability of being assigned to different cluster than they were in the partition of the original data set. If bootstrapping is repeated many times, the transitional sites would have higher probability of occurring in the bootstrap samples with lower λ values than non-transitional sites. Therefore we propose using the mean of the λ values from the bootstrap samples in which the site is included as a measure of the site effect on classification stability: the lower the mean λ value, the stronger the effect of the site on classification instability.

Artificial data sets

We tested the new method using three artificial data sets with species presences/absences, which were divided into three clusters (Fig. 1). Data sets A, B and C consisted of 30 species (rows) and 90 sites (columns). Data set A contained three site clusters of equal size (30, 30 and 30 sites) with species presences in one third of sites. Data set B contained three clusters of unequal size (70, 10 and 10 sites), where the first 10 species occurred in 70 sites, while the other 20 species occurred in 10 sites. Data set C was divided into three clusters (32, 26 and 32 sites) and the frequency of each species was 32. Data sets A and B contained natural groups of sites with species fully concentrated in one of the three clusters, whereas data set C represented a continuous community pattern. To simulate chance species occurrences in data sets A, B and C, we shifted 10%, 20%, . . ., 90% of randomly chosen species occurrences from the original cluster to a randomly chosen site in the data set (Fig. 1). The degree of species occurrence concentration within clusters (fidelity) was quantified using the mean of all positive values of the phi coefficient of association calculated between each cluster and each species (Chytrý et al. 2002, De Cáceres and Legendre 2009). Similarity between sites within and among clusters was illustrated in ordination diagrams of principal coordinates analysis applied to the Sørensen distance matrix (PCoA; Legendre and Legendre 1998), calculated using the R program (R Development Core Team 2010). The partition of the original data set

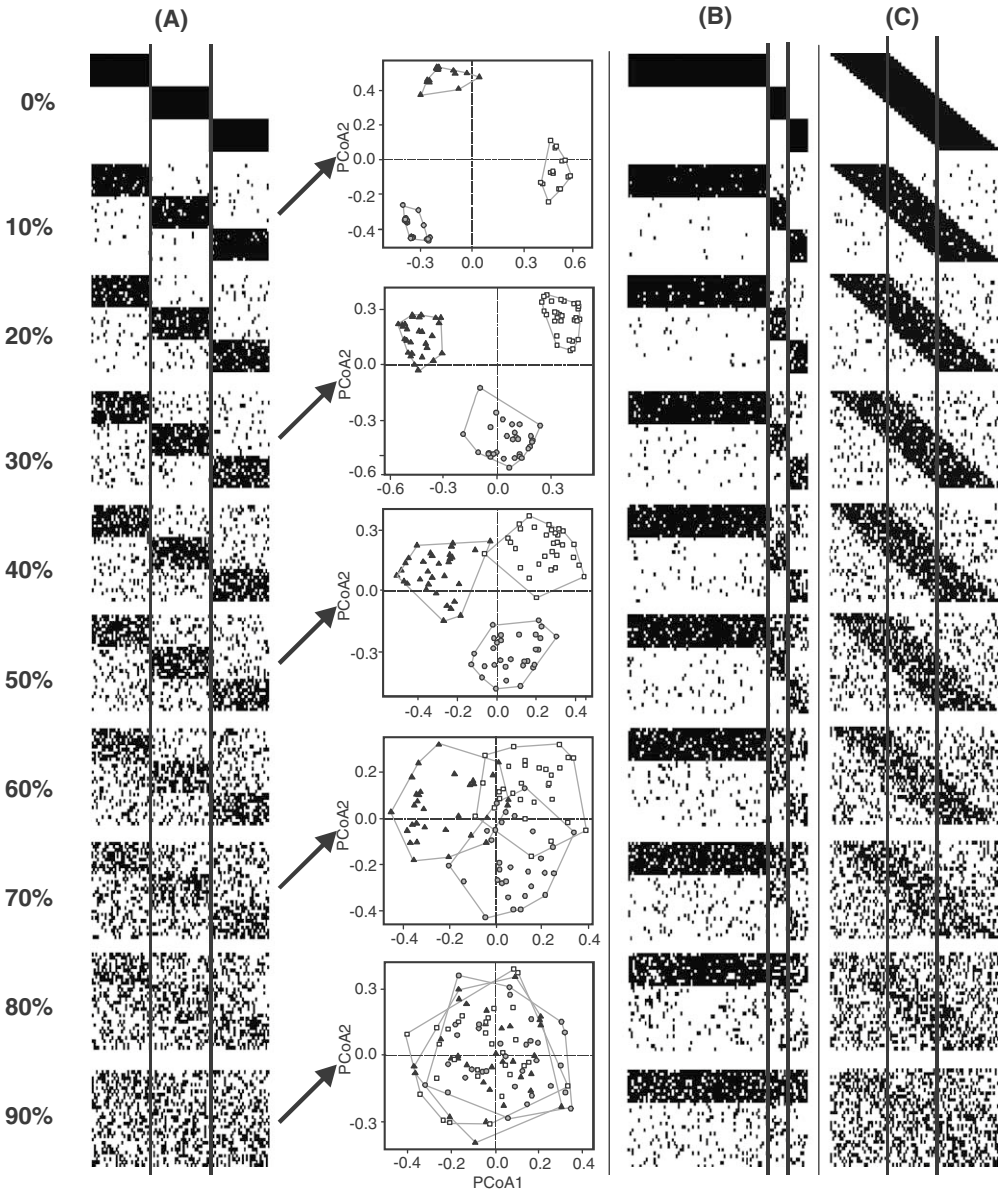

Figure 1. Schemes of the artificial data sets of 90 sites and 30 species used for testing of the new method. Data set A represents sharply separated clusters of equal size, B represents sharply separated clusters of unequal size, and C represents a partition of a continuous gradient in species composition. Percentages of random shifts of species occurrences to another cluster are indicated. Two-dimensional PCoA ordination diagrams (Sørensen distance matrix) for data set A illustrate the differences in data set structure for five selected modifications to the original data set.

with three clusters was compared with the newly established partitions (also with three clusters each) of its bootstrap samples.

Tests with artificial data sets

We applied four clustering algorithms to demonstrate the performance of the new method, including complete linkage, unweighted pair group method with arithmetic averages (UPGMA), beta-flexible clustering (with $\beta = -0.1; -0.25; -0.4$) and k-means. Each of the first three algorithms was combined with Sørensen distance as a measure of the resemblance between sites. Since the k-means algorithm does not allow Sørensen distance to be used as a resemblance measure, we calculated the principal coordinate analysis with Sørensen distance and used the matrix of coordinates from this analysis instead of the raw species-by-plot matrix. In such a way we obtained results for k-means clustering that were comparable with the results of the other clustering methods. For each original data set and clustering method, we repeated the bootstrap resampling, classification of the bootstrap sample and λ calculations 1000 times. The mean value of λ was used as a measure of classification stability. The λ_{rand} was calculated 10 000 times.

Data set C, representing a continuous gradient in species composition, was used to test the ability of the proposed method to identify transitional sites. Each λ value calculated in each iteration was assigned to all sites of the particular bootstrap sample and the process of bootstrap resampling, classification and calculation of λ was repeated 100 000 times to decrease the standard error and to obtain precise mean values for each site.

Real ecological data set

In addition to the tests using artificial data, we demonstrated the performance of the new method on a real ecological data set which included 202 vegetation plots sampled in deciduous forests in the Podyjí National Park, Czech Republic (Chytrý and Vicherek 1995). Each plot included records of all plant species with covers estimated on the Braun-Blanquet scale. The mid-percentage values of each degree on the scale were square-root transformed and classified using Sørensen distance and beta-flexible clustering with $\beta = -0.25$. The mean λ was calculated for 2, 3, 4, ..., 50 clusters.

Results

Differences in λ between the data sets and clustering methods are summarized in Table 1. The classifications of data sets A and B were almost fully repeatable when $<40\%$ of species occurrences were randomly shifted to another cluster, and for most methods they were quite stable until 60% of species occurrences were randomly shifted. However, a higher random noise led to lower classification stability. Data set A with clusters of equal size usually yielded slightly more stable classifications than data set B with clusters of unequal size. K-means clustering gave the most stable results out of all the compared clustering algorithms. Classifications of data set C, representing a community with continuous variation in species composition, showed some degree of instability even when no noise was introduced. Using the mean λ value for the sites from different bootstrap samples, transitional sites were identified near the cluster borders and at the extremes of the gradient (Fig. 2). The vegetation data set showed local peaks of λ value for partitions with four and nine clusters (Fig. 3), indicating that these partitions were more stable than the alternative partitions with different numbers of clusters.

Discussion

For a long time, bootstrap resampling has been routinely used for the evaluation of classifications of population samples based on molecular AFLP or isozyme data (Felsenstein 1985, Efron et al. 1996, Holmes 2003). Like species composition in community ecology, molecular sequences are not samples of a homogeneous statistical population because of high correlations between nucleotides adjacent in the sequences. Therefore, a high number of replicated samples does not yield reliable clusters, but instead emphasizes prevailing biases (Brocchieri 2000). In the bootstrap without replacement method, the weight of each object (i.e. site in community ecology) remains the same and we simply test the classification stability of randomly undersampled data sets. The changing number of sites in each bootstrap sample allows testing the overall classification stability, which may vary with the number of sites selected from the original data set.

The new method proposed here is completely independent of the classification hierarchy and can be applied to partitions with any number of clusters resulting from any clustering method. The results are comparable between different data sets, classification algorithms, resemblance measures and data transformations. However, we also suggest the use of adjusted lambda index in cases when the data set size or cluster number of the compared classifications is considerably different.

The tests with artificial data sets of a known structure, performed in this study, demonstrated that the new method reliably identifies the stability of a partition, which reflects the degree of noise in the input data. We found systematic differences between the clustering algorithms tested. The k-means clustering is perhaps the most popular non-hierarchical classification method due to its simple algorithm and usually easily interpretable results (Li and Chung 2007). This method gave very robust results, even in data sets with a rather high degree of noise. However, it has also some disadvantages. With noisy data, k-means clustering need not yield the same result for each run: since the final partition depends on the initial assignment of sites to clusters, two partitions of the same data set may differ. It is also sensitive to outliers (Ray and Turi 1999, Tran et al. 2003). Generally, non-hierarchical clustering algorithms are less frequently used in community ecology. Researchers usually prefer hierarchical clustering methods, which give the best performance with noisy data (Bowman et al. 2004), and they select an appropriate partition with a certain number of clusters by pruning the classification dendrogram. Of the hierarchical methods, we tested complete linkage, UPGMA and the beta flexible method. The classification stability of beta-flexible clustering was only slightly lower than that of k-means clustering, but complete linkage and UPGMA produced less stable partitions.

The comparison of classification stability of the different numbers of clusters applied to the actual community data set of forest vegetation identified two peaks of the mean λ value, which reasonably reflected the structure of the data set. The first maximum at four clusters distinguished thermophilous oak forests on basic bedrock, dry forests on acidic bedrock, mesophilous forests and alluvial forests. The second maximum at nine clusters mostly reflected phytosociological alliances distinguished using expert knowledge in the original study (Chytrý and Vicherek 1995). This agreement with the expert-based classification indicates that this approach may also identify partitions with an optimal number of clusters, under the assumption that the optimal partition is the most stable one.

The method proposed here, based on the combination of without-replacement bootstrap resampling with the Goodman–Kruskal's lambda index, not only quantifies classification stability but also enables the identification of sites that are transitional between individual clusters. Such a method is applicable a posteriori to any partition.

Table 1. Mean values of Goodman–Kruskal's lambda calculated as a measure of similarity of classification between the partition of the original data set and partitions of its bootstrap samples from data sets A, B and C (Fig. 1). Data set A – sharply separated clusters of equal size (i.e. with an equal number of sites); data set B – sharply separated clusters of unequal size; data set C – continuous gradient. The second column shows the percentage of species occurrences randomly shifted to another cluster. The third column shows the mean of all positive values of the phi coefficient of association, indicating the overall degree of association between species and clusters.

Data set	Percentage of random shifts	Mean phi coefficient	Mean Goodman–Kruskal's lambda					
			Complete linkage	UPGMA	Flexible beta ($\beta = -0.1$) Sørensen	Flexible beta ($\beta = -0.25$) Sørensen	Flexible beta ($\beta = -0.4$) Sørensen	K-means
A	0%	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	10%	0.900	1.000	1.000	1.000	1.000	1.000	1.000
	20%	0.800	1.000	1.000	1.000	1.000	1.000	1.000
	30%	0.700	0.995	1.000	0.995	0.999	0.995	1.000
	40%	0.600	0.871	0.991	0.990	0.991	0.988	0.994
	50%	0.485	0.612	0.904	0.898	0.888	0.884	0.961
	60%	0.379	0.458	0.742	0.712	0.760	0.759	0.880
	70%	0.262	0.234	0.291	0.389	0.435	0.430	0.617
	80%	0.152	0.082	0.051	0.096	0.130	0.149	0.157
	90%	0.084	0.068	0.045	0.073	0.097	0.105	0.048
B	0%	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	10%	0.904	1.000	1.000	1.000	1.000	1.000	1.000
	20%	0.799	1.000	1.000	1.000	1.000	1.000	1.000
	30%	0.679	1.000	1.000	1.000	1.000	1.000	1.000
	40%	0.587	0.992	0.923	0.987	0.998	0.997	0.998
	50%	0.471	0.883	0.541	0.905	0.917	0.903	0.939
	60%	0.378	0.502	0.203	0.508	0.800	0.797	0.800
	70%	0.276	0.256	0.205	0.271	0.314	0.352	0.249
	80%	0.186	0.131	0.131	0.142	0.162	0.169	0.127
	90%	0.123	0.058	0.072	0.075	0.061	0.061	0.000
C	0%	0.507	0.663	0.729	0.794	0.805	0.811	0.938
	10%	0.480	0.547	0.654	0.758	0.767	0.770	0.868
	20%	0.420	0.433	0.595	0.762	0.787	0.802	0.929
	30%	0.366	0.366	0.565	0.664	0.713	0.728	0.816
	40%	0.353	0.299	0.326	0.596	0.686	0.699	0.821
	50%	0.265	0.259	0.195	0.495	0.526	0.534	0.533
	60%	0.210	0.219	0.217	0.355	0.420	0.434	0.521
	70%	0.163	0.108	0.057	0.144	0.289	0.314	0.293
	80%	0.097	0.069	0.031	0.064	0.107	0.129	0.091
	90%	0.119	0.059	0.036	0.051	0.078	0.094	0.052

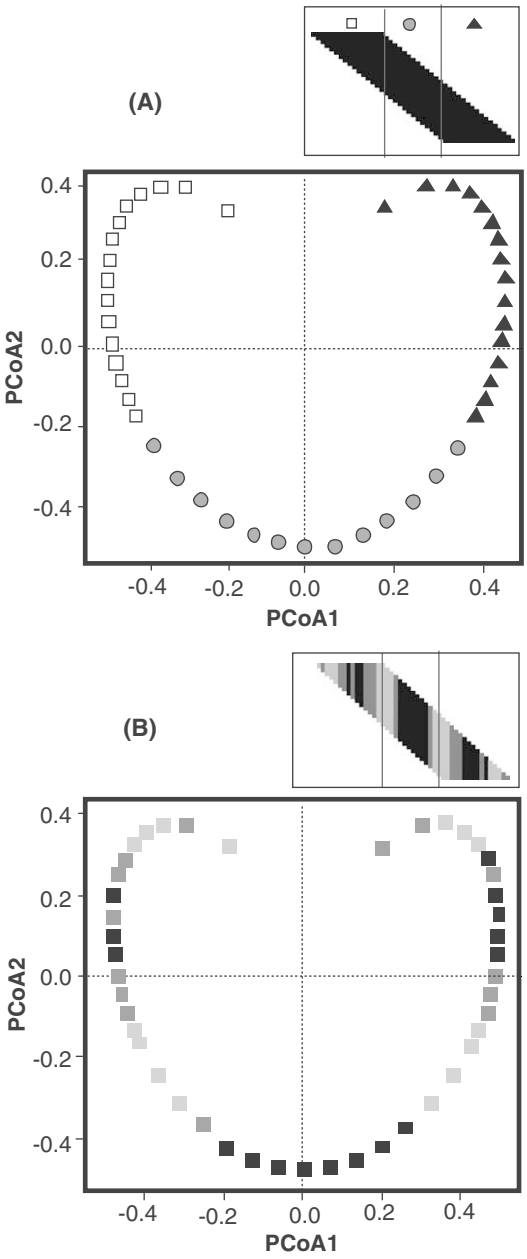

Figure 2. PCoA ordination (Sørensen distance matrix) of artificial data set C containing 90 sites and 30 species. Sites represent a continuous gradient in species composition. They are divided into three groups. The table scheme with species as rows and sites as columns is presented in the upper-left insets of the ordination diagrams. (A) Ordination of sites with symbols representing the three groups. (B) The same ordination with an indication of the degree of site transitionality: black (core) sites mostly appeared in classifications that fully reproduced the original classification, whereas light grey (transitional) sites were sources of classification instability.

It identifies sites within each cluster which have a transitional species composition between clusters and thus contribute to classification instability. Our example of site transitionality presented in the PCoA ordination diagram (Fig. 2) includes an artifact known as the horseshoe effect, in which the pattern along the second axis is curved relative

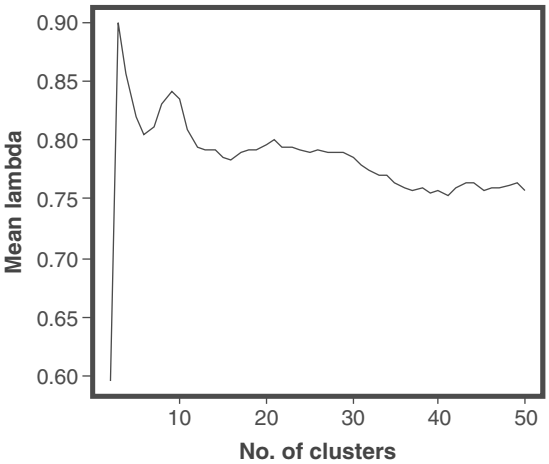

Figure 3. Mean Goodman–Kruskal's lambda index calculated for different numbers of clusters of the vegetation data set established using the beta-flexible clustering method. Distinct local peaks occur at three and nine clusters which correspond to the expert-based classifications of this data set into vegetation units.

to the first axis, and does not represent a true gradient (Podani and Miklós 2002). However, the position of each site along the gradient is still clearly visible. In some applications it can be useful to exclude such sites from classification or to reduce their influence on the classification. De Cáceres et al. (2009) proposed the identification and removal of transitional sites in order to maximize the number of mathematically reproduced clusters. However, their approach is restricted to the use of a non-hierarchical fuzzy algorithm and it cannot be applied to partitions based on pruned dendrograms from hierarchical clustering.

In the new method, we propose measuring classification stability using Goodman–Kruskal's lambda index (Podani 1986). This index gives similar results as the Rand index (Hubert and Arabie 1985), which is frequently used in ecological studies and which can correct for chance effects. However, we refrained from using the Rand index because we found that it overestimated the classification stability for partitions with unequal cluster sizes, like our artificial data set B (not shown).

We conclude that the new method is suitable for the assessment of classification stability, particularly in community ecology. Although in fields such as genetics and systematics the evaluation of classification stability has recently become a standard research tool, its importance has not yet been recognized in community ecology, although classification based on species composition is a common procedure applied to many studies of ecological communities. In the future, community ecologists should pay more attention not only to how good the clusters are (in sense of cluster interpretability in ecological context, their internal variability etc.), but also they should test how stable (robust, repeatable) their classifications are. The algorithm of this method is available in the JUICE program for the classification and analysis of quantitative community data (Tichý 2002), which is freely available on the internet (<www.sci.muni.cz/botany/juice>).

Acknowledgements – This study was supported by the Czech Science Foundation (206/09/0329 and 505/11/0732) and the Ministry of Education of the Czech Republic (MSM0021622416).

References

Austin, M. P. and Heyligers, P. C. 1989. Vegetation survey design for conservation: gradsect sampling of forests in northeast New South Wales. – Biol. Conserv. 50: 13–32.

Belbin, L. and McDonald, C. 1993. Comparing three classification strategies for use in ecology. – J. Veg. Sci. 4: 341–348.

Botta-Dukát, Z. 2008. Validation of hierarchical classifications by splitting dataset. – Acta Bot. Hung. 50: 73–80.

Bowman, F. D. B. et al. 2004. Methods for detecting functional classifications in neuroimaging data. – Hum. Brain Mapp. 23: 109–119.

Brocchieri, L. 2000. Phylogenetic inferences from molecular sequences: review and critique. – Theor. Popul. Biol. 59: 27–40.

Brown, R. D. and Martin, Y. C. 1998. An evaluation of structural descriptors and clustering methods for use in diversity selection. – SAR QSAR Environ. Res. 8: 23–39.

Cao, Y. et al. 1997. A comparison of clustering methods for river benthic community analysis. – Hydrobiologia 347: 25–40.

Chytrý, M. and Vicherek, J. 1995. Lesní vegetace Národního parku Podyjí/Thayatal (Forest vegetation of the Podyjí/Thayatal National Park). – Academia, Praha, CZ.

Chytrý, M. et al. 2002. Determination of diagnostic species with statistical fidelity measures. – J. Veg. Sci. 13: 79–90.

De Cáceres, M. and Legendre, P. 2009. Associations between species and groups of sites: indices and statistical inference. – Ecology 90: 3566–3574.

De Cáceres, M. et al. 2009. Numerical reproduction of traditional classifications and automatic vegetation identification. – J. Veg. Sci. 20: 620–628.

Efron, B. et al. 1996. Bootstrap confidence levels for phylogenetic trees. – Proc. Natl Acad. Sci. USA 93: 13429–13434.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. – Evolution 39: 783–791.

Gauch, H. G. Jr and Whittaker, R. H. 1981. Hierarchical classification of community data. – J. Ecol. 69: 537–557.

Goedickemeier, I. et al. 1997. Sampling for vegetation survey: some properties of a GIS-based stratification compared to other statistical sampling methods. – Coenoses 12: 43–50.

Good, P. I. 2006. Resampling methods, 3rd ed. – Birkhäuser.

Goodman, L. A. and Kruskal, W. H. 1954. Measures of association for cross-classification. – J. Am. Stat. Assoc. 49: 732–764.

Hennig, C. 2008. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. – J. Multivar. Anal. 99: 1154–1176.

Holmes, S. 2003. Bootstrapping phylogenetic trees: theory and methods. – Stat. Sci. 2: 241–255.

Hubert, L. and Arabie, P. 1985. Comparing partitions. – J. Classification 2: 193–218.

Illyés, E. et al. 2007. Semi-dry grasslands along a climatic gradient across central Europe: vegetation classification with validation. – J. Veg. Sci. 18: 835–846.

Knollová, I. et al. 2005. Stratified resampling of phytosociological databases: some strategies for obtaining more representative data sets for classification studies. – J. Veg. Sci. 16: 479–486.

Legendre, P. and Legendre, L. 1998. Numerical ecology, 2nd English ed. – Elsevier.

Li, Y. and Chung, S. M. 2007. Parallel bisecting k-means with prediction clustering algorithm. – J. Supercomput. 39: 19–37.

Manly, B. F. J. 2007. Randomization, bootstrap and Monte Carlo methods in biology, 3rd ed. – Chapman and Hall/CRC.

McKenna, J. E. 2003. An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. – Environ. Model. Softw. 18: 205–220.

Pillar, V. D. 1999. How sharp are classifications? – Ecology 80: 2508–2516.

Podani, J. 1986. Comparison of partitions in vegetation studies. – Abstr. Bot. 10: 235–290.

Podani, J. 1989. New combinatorial clustering methods. – Plant Ecol. 81: 61–77.

Podani, J. and Miklós, I. 2002. Resemblance coefficients and the horseshoe effect in principal coordinate analysis. – Ecology 83: 3331–3343.

R Development Core Team 2010. R: a language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria.

Ray, S. and Turi, R. H. 1999. Determination of number of clusters in k-means clustering and application in color image segmentation. – In: Pal, N. R. et al. (eds), Proceedings of the Fourth International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99). Calcutta, India, pp. 137–143.

Shao, J. and Tu, D. 1995. The jackknife and bootstrap. – Springer.

Shuangge, M. 2006. Empirical study of supervised gene screening. – BMC Bioinfor. 7: 537.

Tichý, L. 2002. JUICE, software for vegetation classification. – J. Veg. Sci. 13: 451–453.

Tichý, L. et al. 2010. OptimClass: using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. – J. Veg. Sci. 21: 287–299.

Tran, T. et al. 2003. SpaRef: a clustering algorithm for multispectral images. – Anal. Chim. Acta 490: 303–312.