Lengyel A, Landucci F, Mucina L, Tsakalos JL, Botta-Dukát Z. Joint optimization of cluster number and abundance transformation for obtaining effective vegetation classifications. J Veg Sci. 2018;29:336–347. https://doi.org/10.1111/jvs.12604

Joint optimization of cluster number and abundance transformation for obtaining effective vegetation classifications

Attila Lengyel [1,2,*] (lengyel.attila@okologia.mta.hu)

Flavia Landucci [3] (flavia.landucci@gmail.com)

Ladislav Mucina [4, 5] (laco.mucina@uwa.edu.au)

James Tsakalos [4] (james.tsakalos@research.uwa.edu.au)

Zoltán Botta-Dukát [1,6] (botta-dukat.zoltan@okologia.mta.hu)

[1] MTA Centre for Ecological Research, Institute of Ecology and Botany, Alkotmány u. 2-4, H-2163 Vácrátót, Hungary

[2] Department of Vegetation Ecology, University of Wrocław, ul. Przybyszewskiego 63, 51-148 Wrocław, Poland

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

[4] School of Biological Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley WA 6009, Perth, Australia

[5] Department of Geography & Environmental Studies, Stellenbosch University, Private Bag X1, Matieland 7602, Stellenbosch, South Africa

[6]MTA Centre for Ecological Research, GINOP Sustainable Ecosystems Group, Klebelsberg Kuno u. 3, H-8237 Tihany, Hungary

*Corresponding author

Abstract

Question: Is it possible to determine which combination of cluster number and taxon abundance transformation would produce the most effective classification of vegetation data? What is the effect of changing cluster number and taxon abundance weighting (applied simultaneously) on the stability and biological interpretation of vegetation classifications?

Locality: Europe, Western Australia, simulated data

Methods: Real data sets representing Hungarian submontane grasslands, European wetlands, and Western Australian kwongan vegetation, as well as simulated data sets were used. The data sets were classified using the partitioning around medoids method. We generated classification solutions by gradually changing the transformation exponent applied to the species projected covers and the number of clusters. The effectiveness of each classification was assessed by a stability index. This index is based on bootstrap resampling of the original data set with subsequent elimination of duplicates. The vegetation types delimited by the most stable classification were compared with other classifications obtained at local maxima of the stability values. The effect of changing the transformation power exponent on the number of clusters, indexed according to their stability, was evaluated.

Results: The optimal number of clusters varied with the power exponent in all cases, both with real and simulated data sets. With the real data sets, optimal cluster numbers obtained with different data transformations recovered interpretable biological patterns. Using the simulated data, the optima of stability values identified the simulated number of clusters correctly in most cases.

Conclusions: With changing the settings of data transformation and the number of clusters, classifications of different stability can be produced. Highly stable classifications can be obtained from different settings for cluster number and data transformation. Despite similarly high stability, such classifications may reveal contrasting biological patterns, thus suggesting different interpretations. We suggest testing a wide range of available combinations to find the parameters resulting in the most effective classifications.

Keywords

Clustering; Cluster validation; Community similarity; Cover scale; Data type; Multivariate data analysis; Numerical classification; Stability of classification

Abreviations

MSL = mean standardized lambda; PAM = partitioning around medoids; PCoA = principal coordinate analysis

Nomenclature

The names of high-rank European syntaxa follow Mucina et al. (2016).

Introduction

Numerical methods are applied in vegetation classification studies to reduce the dimensionality of the data in seeking patterns, to increase objectivity in the analyses, and thus to enhance the reproducibility of results. Still, classification protocols often rely on subjective decisions that can significantly influence the results (De Cáceres et al. 2015). Subjective choices can hardly be avoided, yet they should be well-informed and logical to make the analytical procedures reliable and repeatable. In numerical classifications, according to Lengyel & Podani (2015), the choice of the number of clusters and the weight attributed to abundant species relative to scarce species (hence the data transformation), are among the most influential decisions that have to be considered carefully. If the aim of the classification is to delimit a pre-set number of vegetation types within the data set, then the choice of the resulting clusters should be guided by practical considerations. In certain cases there is reasonable external information available for selecting a transformation function as well. For instance, if the abundance estimations are deemed inaccurate, only presence/absence data should be used. Equally, if the purpose of the study is to analyse vegetation types characterised by dominant species, it is more logical to apply a transformation giving high emphasis to differences in species abundance. However, if the aim of the classification is to explore variation by separating and differentiating vegetation types, classifications using a suite of contrasting parameters should be produced. These should be evaluated *a posteriori* in order to identify the optimal parameter values yielding in the 'best' (according to the set criteria) classification.

The optimal number of clusters can be sought for by calculating *cluster effectiveness* (or *validity*) *index* for classifications with increasing number of clusters. Thus, the optimal number of clusters is the one where the effectiveness index reaches maximum or minimum, depending on scaling. This procedure is widely known and regularly applied in classification studies (e.g. Botta-Dukát et al. 2005; Tichý et al. 2010, 2011). However, we are aware of only a few examples when authors evaluated different data transformations for finding the optimal weighting of abundances that would reveal biological patters most effectively or would lead to the most stable results. Jensen (1978) evaluated the effect of several data transformations on classifications and ordinations of a lake vegetation data, and concluded that 'extreme transformations' (i.e. those giving high weight either to high abundance values or, in reverse, to presence/absence data) can yield significantly different results. This finding was corroborated by Campbell (1978) and van der Maarel (1979). Wilson (2012) compared the stability of ordination analyses performed on various vegetation samples using different transformations of abundance and concluded that the ‘optimal’ transformations depend on context, such as geographical extent, environmental heterogeneity, disturbance status of the study area, and quality of abundance estimations. Although, any ‘optimal’ parameterization supposed to produce a robust classification is specific for the actual data set, the low interest of researchers in finding them, or at least in assessing the performance of methods they apply, is surprising, given that vastly different results can be achieved by application of different abundance scales in multivariate analyses – a fact well known for long time (Austin & Greig-Smith 1968; Noy-Meir et al. 1975; van der Maarel 1979).

In this paper, we introduce a procedure for choosing the combination of two factors, namely (1) the number of clusters and (2) varying scale of transformation power, assisting in identification of the most effective classification outcome. Like other approaches aimed at determination of the optimal number of clusters (e.g. Aho et al. 2008), a general guideline for finding the optimal transformation would be to find the function that leads to the most stable of several possible classifications produced by differently parameterized transformation functions. We show that changing one of these two factors has an impact on the optimal values of the other, which influences the biological interpretation of the classification result, and therefore we promote their joint optimization. We test this approach using real and simulated data sets.

Materials and methods

Grasslands data set

The Grasslands data set consists of phytosociological plots collected in the colline and montane belts of northern Hungary. This data set represents different types of mesic, unproductive to moderately productive, grazed, mown, and recently abandoned grasslands on neutral to acidic soils. Several types can be recognized by their dominant species, e.g. *Agrostis capillaris*, *Arrhenatherum elatius*, *Danthonia decumbens*, *Festuca rubra* and *Nardus stricta*. However, these types are not floristically distinctly separated, and stands with different dominant species can be similar in the overall species composition.

Wetlands data set

The Wetlands data set was extracted from the WetVegEurope database (Landucci et al. 2015). It contains plots from Austria, Czech Republic, Germany, Hungary, Poland, Slovakia, and the Netherlands. In these plots the diagnostic species of the class *Phragmito-Magnocaricetea* (according to Mucina et al. 2016) should have dominance of at least 25% of the total cover. Only plots having at least five species and plot sizes between 15 and 50 m^2 were included. The data set was subject to geographical stratification and to heterogeneity-constrained random resampling (Lengyel et al. 2011) as modified by Wiser & De Cáceres (2013) in order to avoid pseudo-replications and maximally diversify the dataset. In this data set, several types can be distinguished on basis of dominant species, however many of these communities share similar species pool. Therefore, classifications are expected to vary with changing power of the data transformation.

Kwongan data set

The Kwongan data set is composed of 375 plots of natural shrubland (heath-like) vegetation of the Geraldton Sandplains (surrounds of the Eneabba township), Western Australia. This unique, endemic-rich vegetation is supported by sandy soils extremely depleted in phosphorus (and also nitrogen) – a product of prolonged tectonic quiescence of the Western Australian landscapes spanning hundreds of millions of years, resulting in lack of soil rejuvenation and progressive nutrient leaching, combined with relatively stable and predictable climatic seasonality, and predictable natural fire disturbance (Lambers 2014). This data set exemplifies an unusual, yet real situation: both alpha and beta diversity are high, resulting in high regional species pool (gamma diversity). Species dominance (in terms of biomass and projected cover) in this vegetation is supressed. We expect that the classification outcomes would be quite resistant to changes of the magnitude of the data transformation.

Characteristics of the three data sets are summarized in Table 1. A more in-depth analysis of the Grasslands data set is presented, while we focused on the relationship between the examined methodological decisions and classification stability in the Wetlands and the Kwongan data sets.

Simulated data

Simulated data matrices consist of N plots (in the rows) and S species (in the columns). Plots belong to K clusters of equal size, thus the number of plots is $N/K = n$ in each cluster, and n is a pre-defined integer. Ten species occur in each cluster and each species occurs in two clusters, thus $S = 10 \times K/2$. Each species has constant abundance across plots within a cluster, while the abundances may differ among clusters. The abundances of species within one of the two clusters where they occur, are drawn from a Poisson-lognormal distribution (Bulmer 1974) where the mean and the standard deviation (SD) of the lognormal distribution are (2; 1) on log scale. For the other cluster, the order of abundances is reversed, thus if a species was the most abundant in one of the clusters where it occurs, then this species will be the least abundant in the other one (considering only species occurring in this cluster). These matrices, therefore, consist of plots of K clusters according to raw abundances of species, but $K/2$ clusters according to presence/absence data because pairs of clusters share the same species occurring with different abundances. We expect the optimal number of clusters to be $K/2$ with low exponents, while with high exponents optimal solution should comprise K clusters. Notably, abundance-based clusters are nested within clusters based on presence/absence data. Within each cluster, plots are identical, thus the clustered structure is initially perfect. An exemplary matrix is shown in Appendix S1. Then, noise was added to this initial matrix following the method of Gotelli (2000) used for 'noise test', but applied to abundances instead of presence/absence data. This procedure applies a swapping algorithm to introduce noise. In a single swap, the rows and columns of the original matrix are permuted, and a 2 × 2 submatrix with positive values in the diagonal is chosen randomly. Then the two diagonal cells are decreased by 1, while abundances in the two off-diagonal cells are increased by 1 individual, thus the sum and the marginal totals of the submatrix do not change. Finally, the original order of rows and columns is restored. A single swap would affect a sparse matrix more than one with high fill. Also, large matrices are more 'resistant' to the same number of swaps than small ones. Therefore, noise is added to the matrices in discrete levels, one level consisting of as many swaps as the number of non-zero elements in the matrix. Our preliminary analyses suggested that in this way a comparable amount of stochasticity can be added to matrices of different size and fill.

Five simulation series were performed, each of them with five different set-ups. In these series, one or two parameters were changed systematically in order to generate simulated matrices that would differ in: i) noise level; ii) size of clusters with number of clusters fixed; iii) number of clusters with cluster sizes fixed; iv) number and size of clusters with total number of plots fixed; v) dominance of species. The dominance was changed by modifying the SD of the lognormal distribution used as input for the Poisson process of species abundances. When SD is high, there is one or a few highly dominant species within a plot and many very scarce species, while with lower SD species abundances should be balanced.

Classification method

For classifying the data sets, we used the partitioning around medoids method (PAM; Kaufman & Rousseeuw 1990) using Marczewski-Steinhaus index as the measure of dissimilarity (Appendix S2). For the Grasslands and Kwongan data set covers of species were directly estimated on percentage scale in the field, while for the Wetlands data set, abundances were mostly recorded on Braun-Blanquet or finer ordinal scales. These ordinal categories were replaced by their midpoint percentages. Cover percentages were power transformed using the function $x' = x^a$, where x is the original cover value on percentage scale, a is the power exponent, and x' is the transformed cover value. The power exponent was gradually changed from 0 to 1, with 21 steps by 0.05 in between in case of real data, and with steps of 0.1 in case of simulations where simpler patterns were expected. Low values of the exponent reduce the effect of differences between species abundances, thus giving more weight to rare species, while values near 1 give more weight to abundant species. The lowest number of clusters examined was 2. The highest number of examined clusters was 10 for the Grasslands data, 40 for the Wetlands and for the Kwongan data, and it varied in simulations according to the pre-defined number of clusters and sample size. The maximal number of clusters was arbitrarily determined to balance between computation time and the number of practically distinguishable vegetation types.

Evaluation of classifications

Several approaches for evaluating classifications exist, and each of them involves numerous indices (e.g. Milligan & Cooper 1985; Vendramin et al. 2010). These approaches include correlating the original distances between objects and their representations in the classification (e.g. Rohlf 1974), measuring compactness, connectedness, and separation of clusters (e.g. Popma et al. 1983), assessing the robustness of the results to changes in methodological decisions and choice of variables (e.g. Chiang & Mirkin 2010), repetitiveness (e.g. McIntyre & Blashfield 1980), stability (e.g. Hennig 2007), interpretability (e.g. Tichý et al. 2010), and predictive power (e.g. Lyons et al. 2016) of the classification, and degree of divergence from a random classification (e.g. Hunter & McCoy 2004).

A family of classification effectiveness (or validity) measures called geometric indices (Aho et al. 2008) rely on dissimilarities between plots which involve a decision on the weighting of species abundances. For example, if an effectiveness index uses resemblances calculated by the Jaccard index (Podani 2000) using presence/absence data, then the classifications produced on the basis of binary occurrences of species are likely to seem to be ‘better’ than classifications based on cover percentages. However, not only geometric indices need decisions on data transformation. The non-geometric OptimClass indices (Tichý et al. 2010), which use the number of characteristic species of clusters as the measure of effectiveness, can be calculated from both presence/absence and cover percentage data. As the form of cover transformation is known to strongly affect the fidelity values of species (Willner et al. 2009), it is expected that classifications based on presence/absence data would have more character species, if only binary occurrences are considered for fidelity calculations, while classifications using cover data would seem less effective.

For an unbiased comparison of effectiveness among classifications based on different data transformations and cluster numbers, it is necessary to compare all classifications to a standardized reference. The stability index, introduced by Tichý et al. (2011), meets this criterion. It compares the classification of plots in the original data set with classifications of its subsets selected by bootstrap resampling with subsequent elimination of duplicates (Tichý et al. 2011). The similarity between the cluster assignments of resampled plots in the original classification and in the classification of the subset is calculated using the *mean standardized lambda* (hereafter called MSL), the standardized version of Goodman & Kruskal’s lambda index (Goodman & Kruskal 1954; Appendix S2). In our analysis, we used 50 without-replacement bootstrap samples for each classification produced by different cluster numbers and data transformations. MSL was plotted on a so-called *heat map*, in which the colour of the respective segment of the space defined by two explanatory variables (i.e. the power exponent and cluster number) refers to the magnitude of the dependent variable (i.e. MSL).

The marginal distribution of the heat map can also be examined for determining those parameter values which are likely to provide the most effective classification ourcomes, or the lowest or highest variation in classification stability. If one of the parameters, e.g. the exponent, is fixed to an actual value, the mean of the MSL values obtained with changing the other parameter, that is the number of clusters, gives how stable the classifications obtained with the actual exponent are on average. By using the SD instead of the mean, the variation of stability can be expressed, too. Therefore, the SD is a measure of how important the decision is about one of the two parameters if the other one is fixed to an actual value. The use of marginal distributions is showed only for the Grasslands data set.

The most stable classification of a real data set (i.e. the classification with settings resulting in the absolute maximum of MSL and the darkest segment on the heat map) was evaluated by creating a synoptic table containing frequency, average percentage cover, and fidelity of species. The fidelity of species to clusters was calculated using the phi coefficient on 0 to 100 scale (Chytrý et al. 2002). Species with phi value over 20 were considered 'characteristic', and only species with Fisher exact test p<0.001 were considered. Classifications at the optimal cluster level obtained by different exponents, with special attention to the commonly used values (a = 0, 0.5 or 1) and local peaks in stability, were compared on basis of the group memberships of plots using cross-tabulations, as well as by contrasting their biological interpretation with the help of characteristic species.

Data analyses were performed in the R software environment (version 3.1.2, www.r-project.org) using the *vegan* (Oksanen et al., http://cran.r-project.org/package=vegan), *cluster* (Maechler et al., http://cran.r-project.org/package=cluster), *rapport* (Blagotić & Daróczi, http://cran.r-project.org/package=rapport), and *fields* (Nychka et al., http://cran.r-project.org/package=fields) packages. R scripts for data simulation, swapping and the optimization procedure are available in the Appendix S3. We used Juice (Tichý 2002) for data management and construction of synoptic tables.

Results

Grasslands data set

The heat map (Fig. 1) showed that the MSL values varied considerably across cluster number and power exponent. With presence/absence data (a = 0), stability was the highest at the five-cluster solution. From a = 0.05 to a = 0.25, the three-cluster level was the most stable, including a = 0.15 where the second highest stability value was obtained (MSL = 0.804). Between a = 0.3 and a = 0.4, the stability peaked at two clusters, then from a = 0.45 the four-cluster solution was optimal until a = 0.90, while for the higher exponent values again three clusters were shown to be the best. The absolute maximum value was found with a = 0.55 and the four-cluster solution, where the stability of the classification was MSL = 0.824. Exponents between a = 0.25 and 0.50 resulted in the highest stability values on average, and the SD of stability was also the lowest in this interval (Fig. 2). Nevertheless, a second local optimum was found at a = 0.8, although the SD was much bigger here. Across the cluster levels, the three- and four-cluster solutions were the most stable on average, while stability values did not vary much, except for 2 clusters where SD was the highest.

We used the most stable classification (i.e. four clusters and exponent 0.55; hereafter called 'Partition A') as the baseline for the interpretation of all clusters and classifications (Appendix S4). This classification was identical with what was obtained by a = 0.50, that is, square-root transformation. Clusters A1, A2, A3, and A4 are the elements of the Partition A. Cluster A1 represents grasslands of the alliance *Violion caninae*, but some species of the mesic meadows of the order *Arrhenatheretalia* are also frequent. Cluster A2 contains plots of the *Arrhenatherion*. This type was recently described as the *Diantho-Arrhenatheretum* association by Lengyel et al. (2016); it represents nutrient-poor, acidic grasslands overgrown by taller grasses (e.g. *Helictotrichon pubescens*, *Arrhenatherum elatius*) after abandonment or changing management to mowing. Cluster A3 comprises unproductive meadows and pastures dominated by *Agrostis capillatis*, *Festuca rubra*, and *Galium verum*. These stands are similar in species composition to the *Anthoxantho-Agrostietum,* known also from Slovakia and the Czech Republic. Cluster A3 is also intermediate between *Arrhenatheretalia* and *Violion caninae*. Cluster A4 contains grasslands dominated by *Nardus stricta*, in which species of waterlogged soils are also present. This type is traditionally also called '*Hygro-Nardetum*' (e.g. Borhidi et al. 2012).

In the presence/absence case (a = 0), five clusters were differentiated. Hereafter, this classification is called 'Partition B'. Cluster B1 included many plots of Cluster A1 and A3, thus representing mesic meadows with some species of the *Violion caninae*, and matching the species composition of *Anthoxantho-Agrostietum*. Cluster B2 and B3 contained mostly plots previously classified to A2, thus differentiating between two subtypes of *Diantho-Arrhenatheretum*: one with more hygrophilous, and one with more forest-steppe species, respectively. Cluster B4 represents the '*Hygro-Nardetum*' type, thus is similar to Cluster A4. Cluster B5 contains only two plots similar to the *Anthoxantho-Agrostietum*.

With a = 0.15 and three clusters a local peak was detected, to be referred to as Partition C. Cluster C1 contains many plots representing the types mediating between the *Arrhenatheretalia* and *Violion caninae*, formerly classified to Clusters A1 and A3. Cluster C2 represents the *Diantho-Arrhenatheretum,* and it is very similar to Cluster A2. Cluster C3 represents the *'Hygro-Nardetum'* and matches with Cluster A4.

With a = 1 (= no data transformation), three clusters provided the most stable resolution. This classification was called Partition D. Cluster D1 represents grasslands on nutrient-poor soils, including the *'Hygro-Nardetum'* and other types related to the *Violion caninae* and containing *Nardus stricta*. It contains plots of Cluster A1 and A4. Cluster D2 represents mesic hay meadows with *Arrhenatherum elatius*, and it shares many plots with Cluster A2. Cluster D3 represents unproductive meadows and pastures with the dominance of *Agrostis capillaris, Briza media* and *Festuca rubra*. Most of its plots were assigned to Cluster A3 and C2. Therefore, the Partitions C and D similarly separated the *Diantho-Arrhenatheretum* from other types, but differed in how they delimited two other clusters in the rest of the data set.

The cross-tabulation of Partition A against Partitions B, C and D, as well as Partition C against Partition D are shown in Appendix S5.

Wetlands data set

The optimal number of clusters ranged between 3 and 7 when the exponent ranged between 0 and 0.20 (Fig. 3). With higher exponents, the optimal cluster levels increased, too; from a = 0.35 the most stable classifications were found at levels of more than 30 clusters. In the binary case (a = 0), the optimal cluster level was 6, with the square-root transformation (a = 0.5) it was 30, with no transformation (a = 1) it was 39. The most stable classification was the one with a = 0.80 and 40 clusters where MSL was 0.933. At this level clusters were distinguished according to dominant species that were both constant and character species in many cases. Using other high exponents (e.g. a = 0.50 or a = 1) resulted in very similar classifications, thus only the comparison of solutions with a = 0 (hereafter called 'Partition W') and a = 0.80 ('Partition Z') are presented using synoptic tables (Appendix S6 and S7, respectively). Since many phytosociological associations and alliances of wetland vegetation are defined by dominant species, classifications with high exponents (Partition Z) showed a good correspondence with low-rank syntaxa. With low exponents, the most stable classifications revealed markedly different patterns that were difficult to interpret, yet these local optima possessed much lower stability. With a = 0 (Partition W) differences in species pools offered some, although not fully satisfactory explanation for the distinction of clusters. Cluster W1 contained many plots of tall-sedge vegetation with short submerged periods and eutrophic soils (supporting mostly *Magnocaricion gracilis* vegetation). Cluster W2 included mostly plots of tall-sedge vegetation on sites with poorer nutrient supply (mostly *Magnocaricion gracilis* and *Magnocaricion elatae*). Cluster W3 is characterised, to a large part, by reed vegetation belonging to the *Phragmition* and *Phalaridion*. Clusters W4 and W5 contained many plots sampled in wetlands characterised by fluctuating shallow waters (mostly *Eleocharito-Sagittario*, *Phramition*, *Glycerio-Sparganion*), however no clear ecological difference could be recognized between them. Cluster W6 included plots from nutrient-poor mire vegetation often classified as the *Scheuchzerio-Caricetea*. Obviously, Partition W showed very low congruence with the syntaxonomical system and Parition Z (Appendix S8).

Classifications with a = 0 and a = 0.80 do not differ only in the resolution. As it is shown in Appendix S8, clusters of the latter are not nested within the former, instead, it is very common that plots classified to the same cluster at a = 0.80 are assigned to different clusters at a = 0.

Kwongan data set

MSL values varied much at low levels of cluster numbers (up to 6 clusters) and showed much less (and also less predictable) variability at cluster levels above 6 (Fig. 4). The highest MSL values occurred at the cluster levels 2 and 4. The highest classification stability was detected at the 4-cluster level (for exponents spanning 0.0 and 0.75) or the 2-cluster level (for exponents spanning 0.8 and 1.0). The most stable classification was obtained with a = 0.95, cluster number = 2, with stability MSL = 0.843.

At a = 0, four clusters were distinguished (Partition K; Appendix S9). Cluster K1 represented a community with typical species *Hakea candolleana* and *Allocasuarina humilis* found on free-draining soils. Cluster K2 was identified as *Xylomelum angustifolium-Banskia menziesii* community thriving on sandy soils on dune swells. Cluster K3 included plots from *Ecdeiocolea monostachya-Scholtzia laxiflora* community occurring on sandy soils with slightly elevated clay content in inter-dune depressions, while Cluster K4 represented *Banksia shuttleworthiana-Cristonia biloba* confined to regolith composed of depositional lateritic scree and sand. Therefore, these clusters represented an edaphic gradient spanning Cluster K2 (deep sandy soils from the sand dune swells) and Cluster K3 (depressions showing elevated clay content), with Clusters K1 and K4 occupying intermediate position along the gradient. At a = 0.95, the 2-cluster solution was the most stable one (Partition L; Appendix S10). The cross-tabulation tables (Appendix S11) showed that all plots of the Cluster K3 were assigned to the Cluster L1 - the only cluster whose plots were assigned to the same cluster in Partitions K and L. The Cluster K1 was concentrated in Cluster L1, while most plots of the Clusters K2 and K4 belonged to L2. Partitions K and L similarly recovered the gradient between vegetation types supported by soils having elevated clay content (represented by Clusters K1 & K3, as well as L1) and sandy soils (as Clusters K2 & K4, and L2) on the basis of characteristic species of the clusters. The relative position of the clusters in a PCoA ordination also supports the notion that the main compositional patterns are similarly revealed by different abundance weighting (Appendix S12).

Simulations

At the noise level 1, where abundances were strongly down-weighted (a = 0 or a = 0.1), the stability was highest at the pre-defined number of four species-pool based clusters (Fig. 5). From a = 0.2 to a = 0.7, two peaks were found, namely at the 4- and 8-cluster levels, the latter being of higher stability, and with one intermediate peak at a = 0.3 and seven clusters. Where abundance differences were not or only slightly reduced (a > 0.7), only the 8-cluster peak was obvious. From the noise level 2 and higher, the stability peaked at the 8-cluster level. As more levels of noise were added, classifications with low exponent were becoming less and less stable.

Two optimal cluster levels were found where the number of plots in each cluster was 5 (Fig. 6). From a = 0 to a = 0.4, the 4-cluster peak (corresponding the species-pool-based number of clusters) was higher, but from a = 0.5 to a = 1 the 8-cluster solution (i.e. the abundance based optimum) was the most stable one. The pattern of stability was similar, although, less distinct, with clusters of 10 and 25 plots. However, with 50 plots per cluster, the locations of the optima were more irregular, with several peaks between four and eight clusters. With 100 plots per cluster, the optima were detected at four clusters for most of the exponent values, except for a = 0.3 and a = 0.4.

When the number of clusters increased from four with constant cluster sizes, the typical pattern of lower optima at low exponents and higher optima at high exponents were found in most cases, yet with some exceptions (Fig. 7). Where the species-pool based cluster number was two and the abundance-based cluster number was four, three clusters were the most stable with low exponent and four with high exponent. With higher number of true clusters, the most stable classification identified the pre-defined cluster numbers correctly: 8, 12, 16, and 24 clusters with higher exponents, and 4, 6, 8, and 12 clusters with lowers exponents, respectively. The point of inflection, when the observed optima shifted from the species-pool-based level to the abundance-based level, was variable. Yet a broad interval with at least two local peaks of stability was detectable in all heat maps at intermediate exponent values. Cluster numbers between the species-pool-based and the abundance-based optima also came out as optimal in some cases, especially with exponents near the inflection value.

A very similar pattern was found when the number of clusters and cluster sizes were changed with constant sample size (Appendix S13). The species-pool-based and the abundance-based cluster numbers were recovered correctly as local or global peaks. Between them, intermediate levels also gained high stability values, but they were identified as optimal only in a few cases.

With SD = 0.1 the optimal cluster level was four clusters irrespective of the exponent value (Appendix S13). Using a > 0.5 classifications of 7 and 8 groups showed local peaks. With increasing SD, the stability of classifications with eight clusters and high exponent also increased. With SD = 4, the 8-cluster solutions appeared the most stable, except for when a = 0, that is, in the binary case.

Discussion

Evaluation of the real data

The choice of data transformation and cluster number influences the delimitation of vegetation types, as concluded in several other studies (e.g. Jensen 1978; Lengyel & Podani 2015). Certain types (e.g. *Diantho-Arrhenatheretum* in the Grasslands data set) are relatively robust to changes in the examined parameters, while others (e.g. transitional types between *Arrhenatheretalia* and *Violion caninae*) are more sensitive. When it comes to making an unambiguous distinction between vegetation types for practical (such as management) purposes or syntaxonomical revision, it is crucial to consider that different weighting of abundant species may have implications for the delimitation of vegetation units, and thus for the future applicability of the classification.

The Wetlands data set showed that the optimal cluster level can markedly differ if different data transformations are used. While presence/absence data yielded six stable clusters that represented types with more or less different species pools, accounting for differences in abundances raised the optimal levels over 30, where each cluster is separated according to the dominant species. The fact that the high number of stable clusters obtained using high exponent were not nested within the few stable clusters based on presence/absence data, is a clear indication that different data transformations can reveal different types of biological patterns. With low exponents, classifications were best explained by patterns generated by habitat-specific species-pools, while with high exponents, community types differing in fine-scale environmental variation, temporal variability and site history were revealed. It is of interest, that in our study, 40 clusters was the finest classification level examined due to a compromise between practical and scientific reasons, but in reality the optimal number of clusters in the Wetlands data set could have been even higher.

The Kwongan data provided a special insight into the interaction of data transformation and cluster number. Changing the exponent changed the optimal number of clusters as well, and the resulting stable classifications were moderately congruent. However, even these, seemingly less similar classifications revealed the most important ecological pattern on the basis of faithful species — the soil gradient, although fine patterns of transitional subtypes between the extremes were not detected equally well. The Kwongan data set, due to its high beta diversity and balanced within-plot abundance distribution, was less sensitive to changes in data transformation and cluster number in terms of biological interpretation, even though the assignment of plots showed some variation.

Lessons from the simulations

In the simulations, we generated data structure with contrasting patterns with respect to occurrence information. If abundance information were emphasized, the true number of clusters (vegetation types) was twice as high as in cases where only presence/absence data were considered, hence we differentiated a ‘species-pool-based’ and an ‘abundance-based’ number of clusters. In reality, however, also an opposite can be observed, where a few species can be dominant in habitats with different species pools. In such a case the number of abundance-based clusters could be lower than those based on species-pools, as it was seen with the Kwongan data set.

We expected that *weak* data transformations (the exponent being close to 1) which preserve the differences in original abundance patterns, would yield a higher cluster number, while *strong* transformations (the exponent approaching 0) which significantly reduce abundance differences would find the half of this number of clusters optimal. Our results confirmed this expectation.

We introduced stochasticity to artificial data using a similar method as that by Gotelli (2000) called ‘noise test’. This type of noise made classifications with stronger transformations less stable than those involving weak transformations. This result can be understood by recalling how we generated species abundances and noise. The species abundances had been drawn from a Poisson-lognormal distribution, which resulted in many scarce and few abundant species. Considering that the artificial matrices are designed in a way that their matrix fill is low, swapping individuals can moderately reduce the abundance of species in a plot, or it can slightly increase less abundant species, or make absent species present with low abundance. However, it is unlikely to make an abundant species absent in a plot, or to make an absent species very abundant. As a result, the applied noise affected binary information more than the proportions of abundances which determine classifications involving weak data transformations. We believe that this type of noise simulates a common form of stochasticity in nature that is caused by random death of individuals followed by random colonization.

The simulations have revealed several tendencies in classification stability as related to cluster number, data transformation, and sample properties. With increasing size of clusters, the number of abundance-based clusters was underestimated, while the number of clusters based on species pools was detected correctly. Despite this observation with both fixed and changing total sample size, we cannot offer a clear explanation for this finding.

Based on the tests with modified pre-defined number of clusters with fixed cluster sizes, the stability as optimality criterion seems to track the changes correctly in most cases. However, when the number of clusters based on presence/absence data was two, the most stable classifications were obtained at the three-cluster level with strong transformation. (With weak transformations, the abundance-based number of clusters was correctly found at the level of four clusters.) Moreover, in a few cases, optima were indicated between the species-pool-based and the abundance-based levels. When the total sample size was fixed, but number and size of clusters changed, stability performed similarly well. Some inconsistency was found at four abundance-based clusters, where the most stable level was found at two clusters for all but one value of the exponent. Surprisingly, the exception was the binary case (a = 0) where all classifications were generally less stable and the optimum was at the pre-defined number of clusters based on abundance, i.e. four clusters. This contradicts our expectation and we have no clear explanation for this. Despite the above mentioned spurious exceptions, the stability seemed rather robust and accurate across a wide range of cluster numbers with PAM. In real situations, mapping a goodness of classification measure as a function of data transformation and cluster number would help avoiding less effective parameter combinations.

Testing the effect of community dominance on stability by changing the logarithm of SD of species abundances revealed that at the lowest dominance (i.e. low SD), the number of clusters based on species pool was optimal regardless of data transformation. As dominance increased, abundance-based cluster number became more stable and was identified as optimal. This is in line with the common experience that in monodominant vegetation types (e.g. aquatic and marsh vegetation) classifications based on abundance data are more effective and can markedly differ from presence/absence-based classifications, while when the species abundances are more balanced, accounting for abundance differences does not give significantly different or more effective classification than what is obtained by species composition.

Concluding remarks

Classification stability depends both on cluster number and data transformation. The trend of stability along increasing power exponent varies across cluster numbers, and vice versa, the number of clusters resulting in the most stable classifications depends on data transformation. Slight changes in any of these two factors may change the stability of a classification, hence different biological conclusions can be reached. At the same time, similarly effective classifications can be produced using different combinations of parameters. Finding such local optima contributes to the thorough understanding of biological patterns in the sample.

Stability, as proposed by Tichý et al. (2011), is a standardized measure of classification effectiveness because every single classification is compared to classifications of its without-replacement bootstrap subsamples obtained with exactly the same methods. We have chosen this index in our study because of this advantage. However, there are many other measures of effectiveness, but we have chosen not to evaluate them experimentally in this paper. For answering specific research questions, other indices may be more appropriate than stability. In such cases the workflow of testing the effect of data transformation and cluster number on classification effectiveness, and the visualization of results should be the same as we presented, only the measure of effectiveness should be replaced by an alternative. Moreover, it is also possible to perform the optimization analysis using several different effectiveness measures, and then combine the results in order to identify the classification which is the most effective on average across the applied indices.

Apart from the cluster number and the power exponent, we see no obstacles to test the effect of other types of methodological decisions using our approach. For example, an effectiveness measure might be calculated for classifications obtained by different values for the β parameter of the flexible clustering method by Lance & Williams (1967), and the β value providing the most stable classification might be determined. Moreover, our optimization approach can easily be adapted to ordinations, too. If the cluster effectiveness index applied here is substituted by a measure of stability of ordinations (as done by Wilson 2012), the effect of data transformation on the stability of ordinations can be evaluated systematically. The extension of the optimization procedure presented here beyond data transformation and cluster number is a future direction of our research.

Acknowledgements

The Authors are grateful to Miquel De Cáceres, David W. Roberts, Lubomír Tichý, and Ákos Bede-Fazekas for their helpful comments. A.L. and Z.B.D. were supported by the GINOP 2.3.3-15-2016-00019 project. The research stay of A.L. at the University of Wrocław was supported by the POLONEZ programme (grant 2016/23/P/NZ8/04260). L.M. thanks the Iluka Chair in Vegetation Science and Biogeography for logistic support. The work of L.M. and J.T. was also supported by ARC Linkage grant LP150100339.

Authors contributions

A.L. outlined the main idea, performed data analysis and wrote the initial manuscript, Z.B.D. contributed with discussion in all stages of the work, F.L. helped in preparation of the

Wetlands data set and the evaluation of the analysis, L.M. and J.T. contributed by providing the Kwongan data set and evaluating the results, L.M. and J.T. performed linguistic revisions of early versions of the text. All authors critically commented on the manuscript and the supplementary materials.

References

Aho, K., Roberts, D.W. & Weaver, T. 2008. Using geometric and non-geometric internal evaluators to compare eight vegetation classification methods. *Journal of Vegetation Science* 19: 549–562.

Austin, M.P. & Greig-Smith, P. 1968. The application of quantitative methods to vegetation survey: II. Some methodological problems of data from rain forest. *Journal of Ecology* 56: 827–844.

Borhidi, A., Kevey, B. & Lendvai, G. 2012. *Plant communities of Hungary.* Akadémiai Kiadó, Budapest, HU.

Botta-Dukát, Z., Chytrý, M., Hájková, P. & Havlová, M. 2005. Vegetation of lowland wet meadows along a climatic continentality gradient in Central Europe. *Preslia* 77: 89–111.

Bulmer, M.G. 1974. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics* 30: 101–110.

Campbell, B.M. 1978. Similarity coefficients for classifying plots. *Vegetatio* 37: 101–108.

Chytrý, M., Tichý, L., Holt, J. & Botta-Dukát, Z. 2002. Determination of diagnostic species with statistical fidelity measures. *Journal of Vegetation Science* 13: 79–90.

Chiang, M. & Mirkin, B. 2010. Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads. *Journal of Classification* 27: 3–40.

De Cáceres, M., Chytrý, M., Agrillo, E., Attorre, F., Botta-Dukát, Z., Capelo, J., Czúcz, B., Dengler, J., Ewald, J., (…) & Wiser, S.K. 2015. A comparative framework for broad-scale plot-based vegetation classification. *Applied Vegetation Science* 18: 543–560.

Goodman, L. & Kruskal, W. 1954. Measures of association for cross classifications. *Journal of the American Statistical Association* 49: 732–764.

Gotelli, N.J. 2000. Null model analysis of species co-occurrence patterns. *Ecology* 81: 2606–2621.

Hennig, C. 2007. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis* 52: 258–271.

Hill, M.O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54: 427–432.

Hunter, J.C. & McCoy, R.A. 2004. Applying randomization tests to cluster analyses. *Journal of Vegetation Science* 15: 135–138.

Jensen, S. 1978. Influences of transformation of cover values on classification and ordination of lake vegetation. *Vegetatio* 37: 19–31.

Kaufman, L. & Rousseeuw, P.J. 1990. *Finding groups in data: An introduction to cluster analysis.* John Wiley & Sons, New York, US.

Király, G. (ed.) 2009. *New Hungarian Herbal. The vascular plants of Hungary.* Identification key. Aggteleki Nemzeti Park Igazgatóság, Jósvafő, HU. (in Hungarian)

Lance, G.N. & Williams, W.T. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal* 9: 373–380.

Landucci, F., Řezníčková, M., Šumberová, K., Chytrý, M., Aunina L., Biţă-Nicolae, C., Bobrov, A., Borsukevych, L., Brisse, H., (…) & Willner W. 2015. WetVegEurope: a database of aquatic and wetland vegetation of Europe. *Phytocoenologia* 45: 187–194.

Lambers, H. (ed.) 2014. *Plant life on the sandplains in Southwest Australia: A global biodiversity hotspot.* UWA Publishing, Crawley, AU.

Lengyel, A., Chytrý, M. & Tichý, L. 2011. Heterogeneity-constrained random resampling of phytosociological databases. *Journal of Vegetation Science* 22: 175–183.

Lengyel, A. & Podani, J. 2015. Assessing the relative importance of methodological decisions in classifications of vegetation data. *Journal of Vegetation Science* 26: 804–815.

Lengyel, A., Illyés, E., Bauer, N., Csiky, J., Király, G., Purger, D. & Botta-Dukát, Z. 2016. Classification and syntaxonomical revision of mesic and semi-dry grasslands in Hungary. *Preslia* 88: 201–228.

Lötter, M.C., Mucina, L. & Witkowski, E. 2013. The classification conundrum: species fidelity as leading criterion in search of a rigorous method to classify a complex forest data set. *Community Ecology* 14: 121–132.

Lyons, M.B., Keith, D.A., Warton, D.I., Somerville, M. & Kingsford, R.T. 2016. Model-based assessment of ecological community classifications. *Journal of Vegetation Science* 27: 704–715.

McIntyre, R.M. & Blashfield, R.K. 1980. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research* 15: 225–238.

Milligan, G.W. & Cooper, M.C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50: 159–179.

Mucina, L., Bültmann, H., Dierßen, K., Theurillat, J.-P., Raus, T., Čarni, A., Šumberová, K., Willner, W., Dengler, J., (…) & Tichý, L. 2016. Vegetation of Europe: hierarchical floristic classification system of vascular plant, bryophyte, lichen, and algal communities. *Applied Vegetation Science* 19: 3–264.

Noy-Meir, I., Walker, D. & Williams, W.T. 1975. Data transformations in ecological ordination: II. On the meaning of data standardization. *Journal of Ecology* 63: 779–800.

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

Podani, J. & Feoli, E. 1991. A general strategy for the simultaneous classification of variables and objects in ecological data tables. *Journal Vegetation Science* 2: 435–444.

Popma, J., Mucina, L., van Tongeren, O. & van der Maarel, E. 1983. On the determinants of optimal levels in phytosociological classification. *Vegetatio* 52: 65–75.

Roberts, D.W. 2015. Vegetation classification by two new iterative reallocation optimization algorithms. *Plant Ecology* 216: 741–758.

Rohlf, F.J. 1974. Methods of comparing classifications. *Annual Review of Ecology & Systematics* 5: 101–113.

Rozbrojová, Z., Hájek, M. & Hájek, O. 2010. Vegetation diversity of mesic meadows and pastures in the West Carpathians. *Preslia* 82: 307–332.

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

Tichý, L., Chytrý, M. & Šmarda, P. 2011. Evaluating the stability of the classification of community data. *Ecography* 34: 807–813.

Tichý, L., Chytrý, M., Hájek, M., Talbot, S.S. & Botta-Dukát, Z. 2010. OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. *Journal of Vegetation Science* 21: 287–299.

van der Maarel, E. 1979. Transformation of cover-abundance values in phytosociology and its effects on community similarity. *Vegetatio* 39: 97–114.

Vendramin, L., Campello, R.J.G.B. & Hruschka, E.R. 2010. Relative clustering validity criteria: A comparative overview. *Statistical Analysis & Data Mining* 3: 209–235.

Willner, W., Tichý, L. & Chytrý, M. 2009. Effects of different fidelity measures and contexts on the determination of diagnostic species. *Journal of Vegetation Science* 20: 130–137.

Wilson, J.B. 2012. Species presence/absence sometimes represents a plant community as well as species abundances do, or better. *Journal of Vegetation Science* 23: 1013–1023.

Wiser, S.K. & De Cáceres, M. 2013. Updating vegetation classifications: an example with New Zealand's woody vegetation. *Journal of Vegetation Science* 24: 80–93.

List of Appendices

Appendix S1: Simulation data example

Appendix S2: Mathematical formulae

Appendix S3: R scripts

Appendix S4: Grasslands synoptic table (Partition A)

Appendix S5: Cross-tabulations of partitions of the Grasslands data set

Appendix S6: Wetlands synoptic table (Partition W)

Appendix S7: Wetlands synoptic table (Partition Z)

Appendix S8: Wetlands cross-tabulations

Appendix S9: Kwongan synoptic table (Partition K)

Appendix S10: Kwongan synoptic table (Partition L)

Appendix S11: Kwongan cross-tabulation

Appendix S12: Kwongan ordination

Appendix S13: Additional heat maps of the simulated data sets

Tables

Table 1. Characteristics of the real vegetation data sets

	Grasslands	Wetlands	Kwongan
Vegetation type	mesic grasslands	reeds and sedge beds	sclerophyllous scrub
Geographical location	Northern Hungary	Central and Western Europe	Geraldton Sandplains, Western Australia
Nr. of plots	55	2725	379
Plot size (m2)	25	15 to 50	100
Number of species			
total	269	844	645
mean per plot	37.78	12.52	49.33
minimum per plot	18	5	20
maximum per plot	54	43	85
Mean diversity of order 1*	12.22	4.8	37
Mean evenness per plot**	0.32	0.38	0.75
Mean SD of species covers	8.77	20.60	1.79
Mean 25–75% quantiles of species covers	0.51–2.52	2.05–6.60	1.00–1.15

*according to Hill (1973)

**mean of diversity of order 1 divided by diversity of order 0, the latter being species richness

Figures

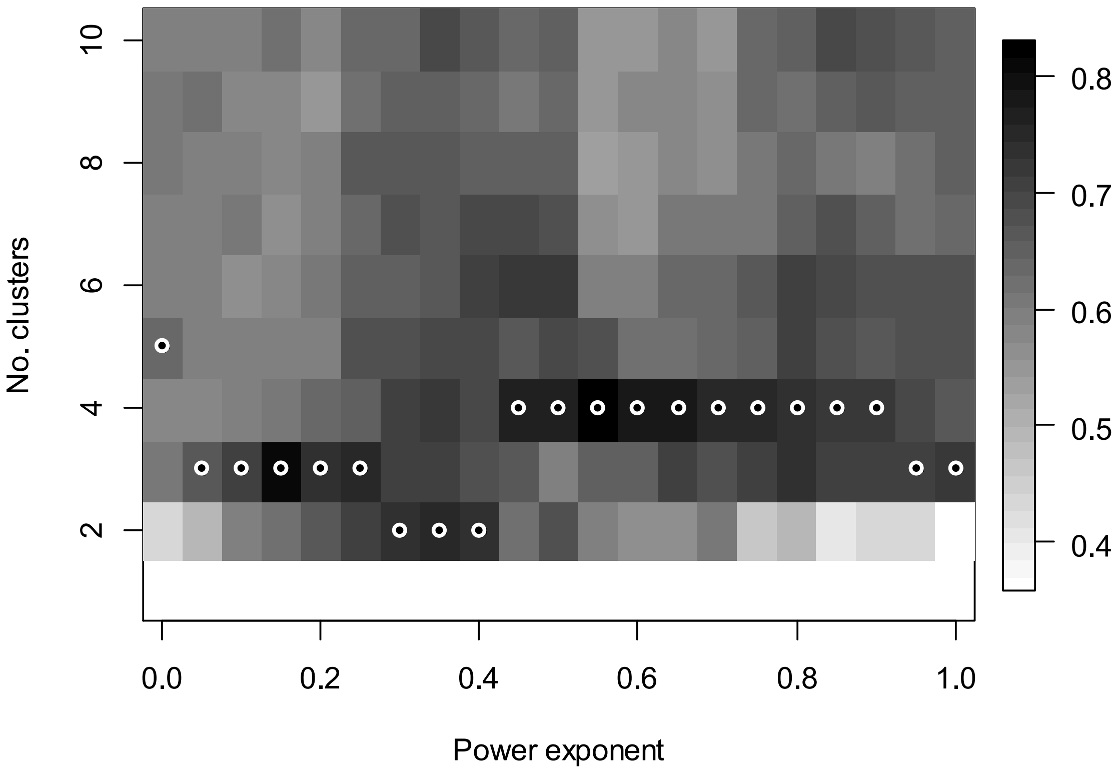

Fig. 1. Analysis of the Grasslands data set showing the heat map of classification stability obtained using different parameters for number of clusters and power exponent. Darkness of the segments correlate with the value of the mean standardized Goodman & Kruskal's lambda (MSL), where the darkest segments marking the combinations of parameters leading to the most stable classifications. White circles with black dots indicate the optimal number of clusters for a given exponent.

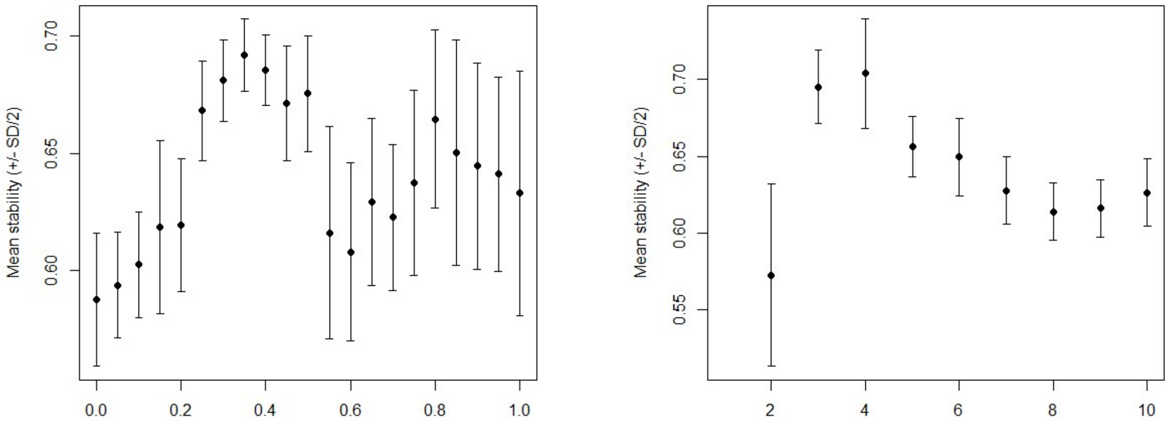

Fig. 2. Mean and standard deviation as error bars of the marginal of the heat map of the Grasslands data set.

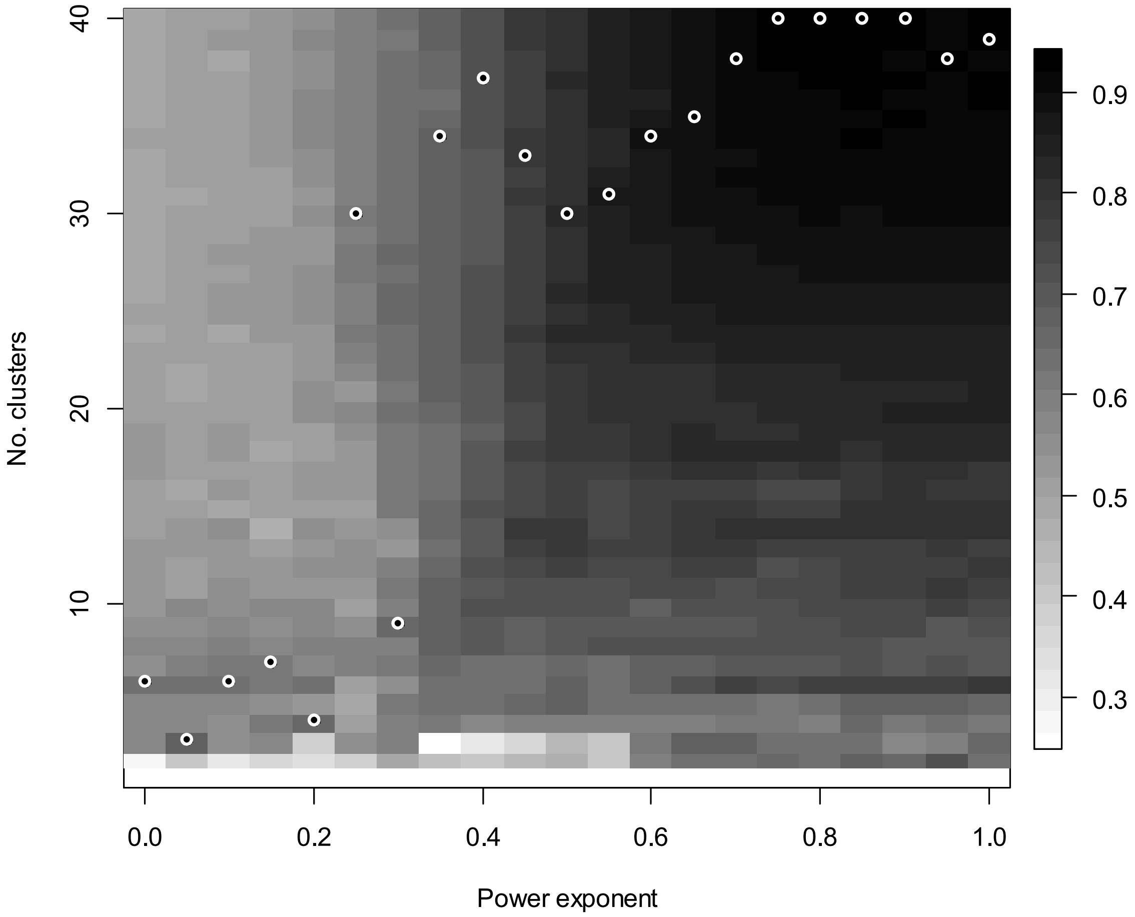

Fig. 3. Analysis of the Wetlands data set showing the heat map of classification stability obtained using different parameters for number of clusters and power exponent. For the meaning of shading and other symbols see Fig. 1.

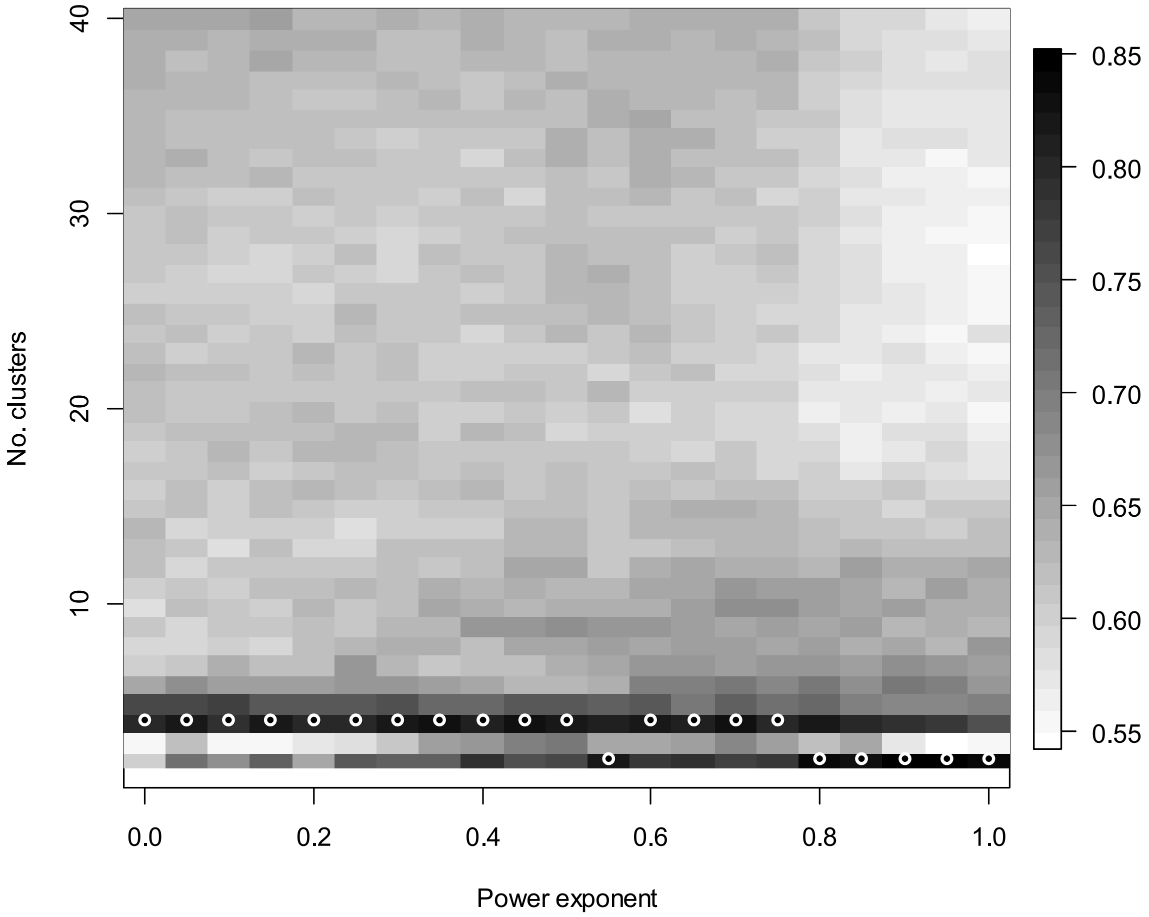

Fig. 4. Analysis of the Kwongan data set showing the heat map of the classification stability obtained using different parameters for number of clusters and power exponent. For the meaning of shading and other symbols see Fig. 1.

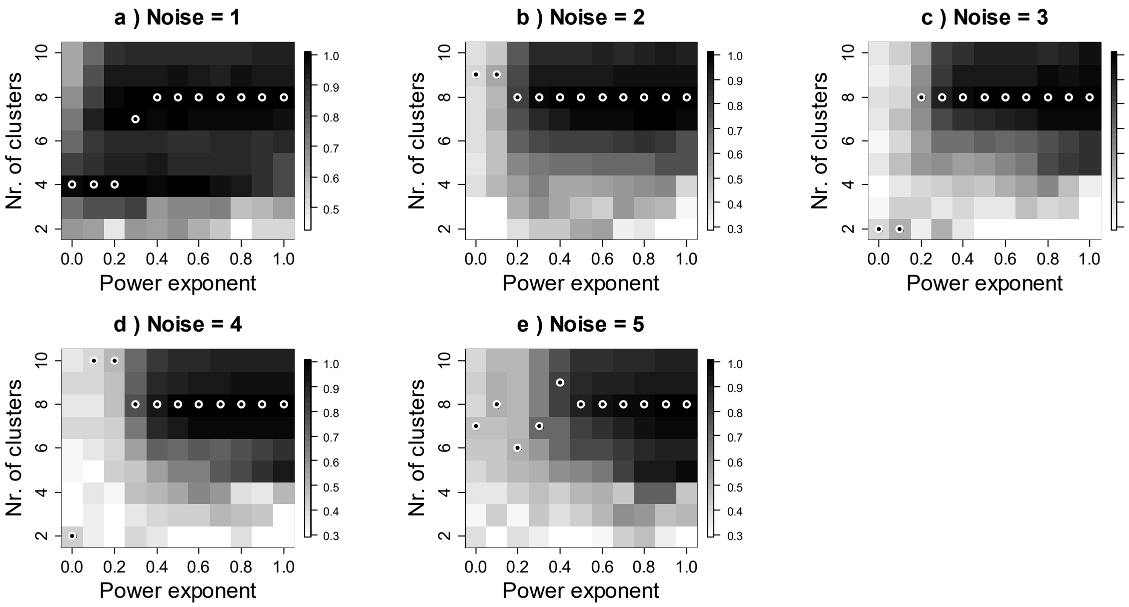

Fig. 5. Simulated data with different noise levels showing the heat maps of classification stability obtained with different parameters for number of clusters and power exponent. For the meaning of shading and other symbols see Fig. 1. The abundance-based numbers of clusters is eight, and the species-pool-based number of clusters is four.

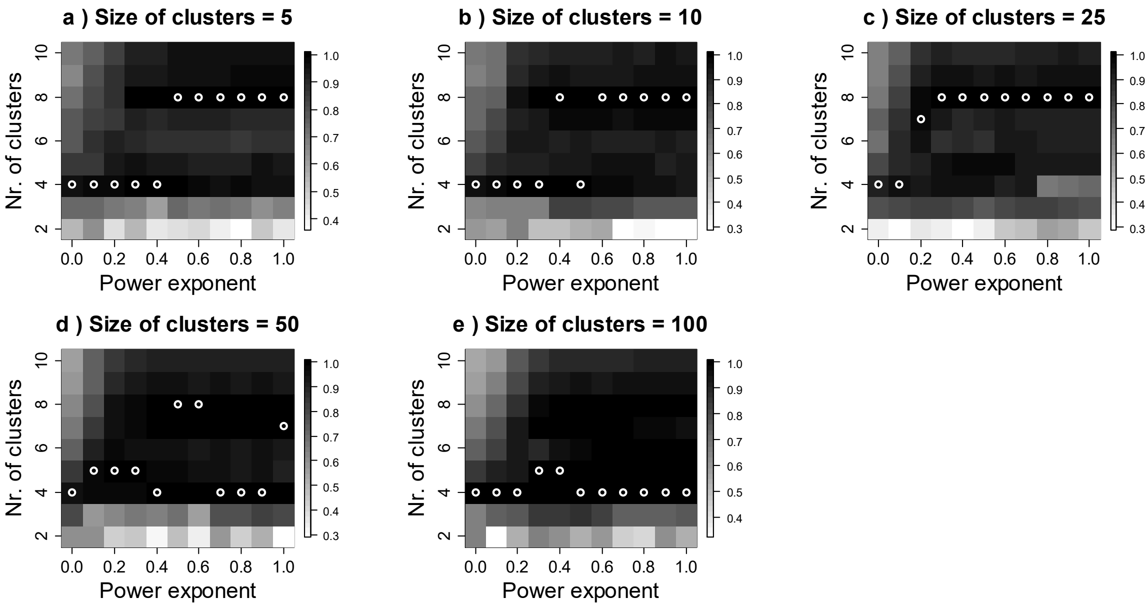

Fig. 6. Simulated data with different cluster sizes and fixed number of clusters showing the heat maps of the classification stability obtained with different parameters for number of clusters and power exponent. For the meaning of shading and other symbols see Fig. 1. The abundance-based numbers of clusters is eight, and the species-pool-based number of clusters is four.

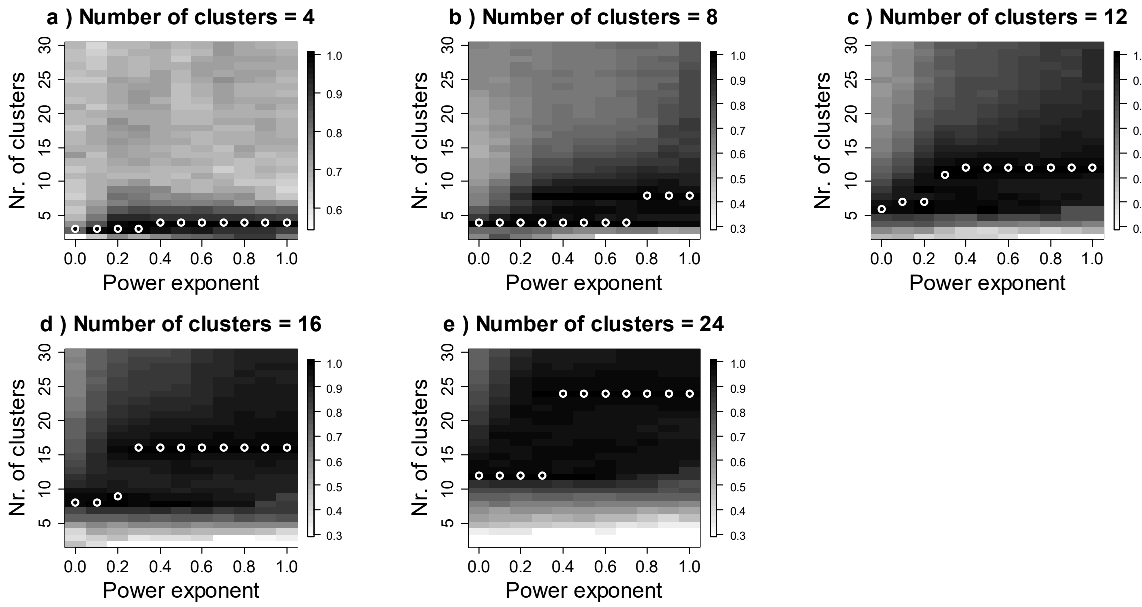

Fig. 7. Simulated data with different numbers and fixed size of clusters showing the heat maps of classification stability obtained with different parameters for number of clusters and power exponents. For the meaning of shading and other symbols see Fig. 1.