Phytocoenologia Vol. 50 (2020), Issue 2, 163–172
Published online March 2020

Research Paper

Optimizing the classification of species composition data by combining multiple objective evaluators toward selecting the best method and optimum number of clusters

Samaneh Tavakoli, Hamid Ejtehadi* & Omid Esmailzadeh*

Abstract

Aims: Classification is an appropriate tool for the summarizing of species data in community ecology. Researchers need to select the effective classification method(s) and the optimum number of clusters to perform a reasonable classification. The aims of the present research are to assess the efficacy of various classification algorithms and to select the optimum number of clusters. **Study area:** We used a dataset of 197 400 m² relevés recorded from Tarbiat Modares University research forest located in the north of Iran. **Methods:** For each relevé, a species list and the canopy cover were recorded by using Braun-Blanquet cover-abundance scale modified by van der Maarel. We considered seven classification methods: flexible-β linkage (β = -0.25), Ward's linkage, complete linkage, average linkage, Modified TWINSPAN, k-means, and PAM. Using each of these algorithms, data were classified into 2-21 cluster levels. Then, values of eight internal evaluators viz. ASW, 1-C.index, PARTANA, PBC, 1-ISA.pval, ISA.sig.inds, ISAMIC, 1-Morisita as well as mean lambda index were calculated for each classification level resulted from algorithms. These values were applied in three methods to select the appropriate classification algorithm(s). Also, we used those values to choose the optimum number of clusters in the selected algorithm(s). A discriminate analysis opted for the verification of the selected optimums. **Results:** Our results revealed that, for our data, flexible-β linkage was the proper classification algorithm with 12 the optimum number of clusters. Despite the vast number of available classification algorithms, there is no ultimate best one for all vegetation datasets. Therefore scientists need using multiple criteria to choose their specific appropriate method. With respect to this, our methods and findings could provide a generalized framework for choosing the effective method(s) for the subsequent classification analyses.

Keywords: classification algorithms; evaluators; mean lambda; median scores; outlier data; Hyrcanian forests.

Submitted: 6 October 2019; first decision: 19 November 2019; accepted: 10 February 2020

Introduction

Vegetation classification is an important tool for environmental management that describes critical units for the monitoring of natural communities (De Caceres & Wiser 2012). Management planning for successful forestry and conservation programs depend on vegetation classification (Peet & Roberts 2013). This analysis can be based on various criteria including species composition, physiognomy, plant functional traits, climatic factors and soil conditions (De Cáceres & Wiser 2012).

Methods of vegetation classification are generally divided into two groups: qualitative expert-based and quantitative numerical-based ones (Aho 2006; Peet & Roberts 2013). Cluster analysis, a quantitative method, is the most common approach in classification studies (Aho 2006). This method creates homogeneous groups (Gan et al. 2007) and provides acceptable results (Aho 2006). Choosing an appropriate clustering algorithm and selecting the optimum number of clusters can be considered as important parts of the cluster analysis (Aho 2006).

There are numerous algorithms for a numerical classification of the species data. These algorithms are mainly different in terms of their linkage methods (Tichý et al. 2010). Although different clustering algorithms have been proposed in the literature (Gan et al. 2007; Legendre & Legendre 2012; Peet & Roberts 2013), selecting the effective one is often very tricky.

*Corresponding authors addresses: Esmailzadeh, O., Department of Forest Sciences and Engineering, Faculty of Natural Resources, Tarbiat Modares, Tehran, Iran; oesmailzadeh@modares.ac.ir; Ejtehadi, H., Quantitative Plant Ecology and Biodiversity Research Lab., Department of Biology, Faculty of Science, Ferdowsi University of Mashhad, Mashhad, Iran; hejtehadi@um.ac.ir Complete addresses of all authors can be found at the bottom of the paper.

DOI: 10.1127/phyto/2020/0360

www.borntraeger-cramer.de
0340-269X/2020/0360 $ 4.50

Classification stability and the quality of results are two general ways to assess the effectiveness of a classification algorithm. Furthermore, reproducible results along with the consistency in the assignment of new samples to previously defined clusters are considered as important features in determination of an effective classification method (De Cáceres & Wiser 2012).

Goodman-Kruskal's lambda index (Goodman & Kruskal 1954) can be used to evaluate the stability of classification results against the removal of the site from the original dataset (Tichý et al. 2011). The quality of classification results can be assessed by using external and internal evaluators (indices). Internal indices use the characteristics of the classification algorithms to evaluate their results (Gauch & Whittaker 1981). These evaluators are divided into geometric and non-geometric evaluators. Geometric evaluators (e.g., average silhouette width (ASW) (Rousseeuw 1987)) assess the quality of classification algorithms based on the relation of samples within and between clusters. Non-geometric evaluators, like Morisita's index of niche overlap (Horn 1966), investigate the quality of classification algorithms with respect to species distributions.

Determining the optimum number of clusters is another important part of the cluster analysis (Tibshirani et al. 2001). Various methods including Gap statistic (Tibshirani et al. 2001), Goodman-Kruskal's lambda index (Tichý et al. 2011), the number of diagnostic species (Tichý et al. 2010), and internal evaluators (Aho 2006; Aho et al. 2008; Roberts 2015) are proposed to determine the optimum number of clusters.

Previous researches focused on Kruskal's lambda (e.g., Tichý et al. 2011 and Lengyel et al. 2017) or one or a few internal evaluators like C-index and Goodman-Kruskal index (e.g., Bolshakova & Azuaje 2006); Indicator Species Analysis (ISA) (e.g., Dufrêne & Legendre 1997); Point Biserial Correlation (PBC) (e.g., Milligan 1980); Indicator Species Analysis Minimizing Intermediate Constancies (ISAMIC(and uniqueness index (e.g., Lotter et al. 2013); crispness index (e.g., Botta-Dukat et al. 2005); ISAMIC, Partition Analysis (PARTANA(and Silhouette Width (e.g., Roberts 2015) in selecting the best algorithm or (and) determining the optimum number of clusters. Aho et al. (2008) compared different classification methods by using multiple internal evaluators.

Here, a combination of eight internal evaluators and Kruskal's lambda is used to compare the effectiveness of seven classification methods. In addition to previous methods (e.g., Aho 2006 and Aho et al. 2008), we graphically evaluated the algorithm responses. This step also is used to determine the optimum number of clusters.

We hypothesize that the results of statistical and graphical analysis of algorithm response confirm each other. Also, using environmental data can help us to verify the selected optimum number of clusters. The specific objectives of our study are to (a) compute responses of seven classification algorithms to lambda and eight internal evaluators at 20 cluster levels; (b) use these responses to draw boxplots for each algorithm (graphical representation) and investigate differences between algorithms based on outliers and also to calculate quartiles and medians of responses and compare them with each other; (c) select the appropriate classification methods based on results of part b; (d) determine optimum number of clusters in the selected algorithms and consequently verify these optimums by using environmental factors.

Materials and methods

Study area

The Hyrcanian forests cover an area of 1.9 million hectares along the southern coast of the Caspian Sea in northern Iran (Esmailzadeh et al. 2011; Gholizadeh et al. 2019). These forests have a high growth capacity due to fertile soil and a humid temperate climate (Marvie Mohadjer 2004). The dominant tree species of Hyrcanian forests are oriental beech (*Fagus orientalis* Lipsky), oak (*Quercus castaneifolia* C.A. May.) and hornbeam (*Carpinus betulus* L.) (Sagheb-Talebi et al. 2014). The study area with latitude of 36°29'23" to 36°29'36" N and longitude of 51°43'20" to 51°47'39" E is a part of the Hyrcanian forests. It is located in the south-east part of Sissangan Forest Park and covers about 1721 ha. The elevation ranges from 100 to 1800 m a.s.l. The mean annual precipitation and temperature of the area are 1300 mm and 15 °C, respectively (Iran Meteorological Organization-Mazandaran portal, 2018).

Sampling design

Vegetation sampling was conducted following the Braun-Blanquet approach (Jennings et al. 2009). Data from 197 relevés, each with an area of 400 m^2 (Dengler et al. 2008), were recorded from all types of the plant communities in the area. Sampling was performed along an elevational gradient at the time of the vegetation cover peak of 2017. For each relevé cover of species was recorded by using Braun-Blanquet cover-abundance scale modified by van der Maarel in which absence ranked as 0 and 1: 0–1%, 2: 1–2.5%, 3: 2.5–5%, 4: 5–12.5%, 5: 12.5–25%, 6: 25–50%, 7: 50–75%, 8: 75–100%. The estimated canopy cover values were transformed to mid-point percentage values for each degree (Slezák et al. 2016). Environmental data including elevation, aspect and slope inclination were recorded in each relevé. Slope aspect was recorded as a circular variable, then, transformed into the northness and eastness of aspect by using Equation 1 and 2 (Dobrovic et al. 2007) as follows:

Northness = Cos(A) + 1 (Equation 1)
Eastness = Sin(A) + 1 (Equation 2)

Where A is the Azimuth of the slope (Dobrovic et al. 2007).

In each relevé, a soil sample up to 20 cm depth was collected. These samples were dried, sieved using a 2 mm mesh, and analysed to determine the soil texture as well as N%, C% and pH.

Data analysis

Classification methods

To classify the collected data, first, the Hellinger distance measure was calculated. Then, we used four hierarchical agglomerative algorithms viz. flexible-β linkage (β = -0.25) (Lance & Williams 1967; McCune et al. 2002), Ward's linkage (Ward 1963), complete linkage (Sørensen 1948) and average linkage (Sokal & Michener 1958). Also, a hierarchical divisive algorithm (i.e., modified TWINSPAN) (Roleček et al. 2009), and two non-hierarchical methods, k-means clustering (Hartigan & Wong 1979) and partitioning around medoids (PAM) (Kaufman & Rousseeuw 1990) were used.

Ward's linkage, complete linkage and average linkage were calculated using the vegan package (Oksanen et al. 2017), flexible-β linkage, k-means, and PAM were performed by using the cluster package (Maechler et al. 2016) and modified TWINSPAN by using the TWINSPAN package (Zeleny et al. 2016) in the R ver.3.3 (R Core Team, 2018).

Assessing the effectiveness of the classification algorithms

We used the mean lambda (as an index to evaluate the classification stability) and internal evaluators (as indices to evaluate the quality of classification results) to assess the effectiveness of the seven selected algorithms. At first, data were classified into 20 classification levels (2–21 clusters) by using the seven classification algorithms. The values of mean lambda and internal evaluators were calculated for each of these resulted levels in each algorithm.

To compute the mean lambda for each level in every classification algorithm, 50 subsamples were generated from the original dataset. Each subsample contained about 63% of relevés without replacements. These subsamples were classified to the same number of clusters by using the same classification method (e.g., to compute mean lambda for third classification level in complete linkage method, 50 generated subsamples were classified to 4 clusters using complete linkage method). Then, the lambda value was calculated for each of the classified subsamples. Therefore, 50 lambda values were computed and the mean of these 50 values was considered as a stability criterion of the classification levels (Lengyel et al. 2017).

To calculate internal evaluators, some geometric and non-geometric evaluators were considered. Geometric evaluators were: ASW, C-index (Hubert & Levin 1976), PARTANA (Roberts 2016b), and PBC (Brogden 1949). Non-geometric evaluators were: ISA average p-value α = 0.05 (ISA.pval) and ISA number of significant indicators (ISA.sig.inds) (Dufrene & Legendre 1997), ISAMIC (Roberts 2016a), and Morisita's index of niche overlap.

The calculated mean lambda and evaluator's scores for each algorithm were used by three methods to determine the effective classification algorithm(s).These methods are:

A) Ranking the responses of the algorithms
In this approach, 20 calculated values for mean lambda and evaluators for each classification methods were separately ranked. To rank these values, the lowest score was ranked as 1 and the highest score was ranked as 20 and the average was used for the tied values. Then, for each algorithm by using these ranks, boxplots were drawn. The number of outliers was considered as a criterion to interpret the boxplots. Algorithm(s) with the highest number of outliers were considered as poor functioning algorithm(s).

B) Comparing the statistical quartiles
The second step in decision-making to select the effective classification algorithm(s) was comparing 25, 50, and 75 quartiles. At first, means of 20 calculated values for mean lambda and eight evaluators for each algorithm was measured. Then, statistical quartiles of these values of means were computed. Algorithm(s), in which data are presumed to be distributed systematically not haphazardly in quartiles, was considered as an effective algorithm.

C) Comparing the medians
Medians of those 20 calculated values for mean lambda and every evaluator from each algorithm were determined. Interquartile range (IQR) was employed to determine the confidence interval of medians (McGill et al. 1978). The Kruskal-Wallis multiple comparison tests (Kruskal & Wallis 1952) and Dunn's post hoc test (Dunn 1964) were performed on the resulted scores to assess the statistically significant differences among the seven classification algorithms. The best scores of nine criteria in the seven classification algorithms were determined and the number of times that each algorithm had the highest (or tied for highest) median scores were counted. The classification algorithm(s) with the highest median scores was selected as the more effective classification algorithm. To perform these analyses the packages viz. ASW: cluster (Maechler et al. 2016), C-Index and Morisita: plant.ecol (Aho 2015), PARTANA: optpart (Roberts 2016b), PBC: asbio (Aho 2016), ISA: indicspecies (De Caceres & Leg-

endre 2009), ISAMIC: labdsv (Roberts 2016a), mean lambda: vegan (Oksanen et al. 2017), isopam (Schmidtlein et al. 2010) and tcltk (R core team, 2018) were used. Computations of the quartiles were performed by using SPSS 24 (IBM Corp 2016). The boxplots were drawn using the ggplot2 (Wickham 2016) and reshape2 (Wickham 2007) packages were used in the R software.

Optimum number of clusters

As suggested by Aho (2006), the responses of the evaluators and mean lambda in all classification levels (20 levels) in the selected classification methods were used to determine the optimum number of clusters. By combining different evaluators, their diverse properties can help us to achieve better results in determining the optimal clusters (Bolshakova & Azuaje 2006).

To eliminate the bias caused by the variability of the numerical values of the evaluators and mean lambda, the data were standardized by using relativization by the maximum (Equation 3). Relativization usually has a small effect on the outcome of the analysis, if variability (Coefficient of Variation: CV%) among rows (or columns) be smaller than 50 (McCune et al. 2002).

$$b_{ij} = {}^{x_{ij}}/x\max_j \qquad \text{(Equation 3)}$$

Where rows (i) are the number of clusters and columns (j) are evaluators, $x\max_j$ is the largest value in the matrix for evaluator *j*.

Totally, nine standardized values (one for mean lambda index and eight for evaluators) were calculated for each classification level. Then, the average of these nine standardized values was computed. Classification level showing the highest average value indicated the optimum number of clusters.

Verification of the selected optimum number of clusters

Environmental features can be used as the external attributes when the classification is based on species composition. Therefore, by using discriminant analysis, the degree of differentiation among the clusters is evaluated using those external attributes (Tichý et al. 2010). So, these results can be used to evaluate the percentage of correct classification and/or the accuracy of the optimum number of clusters determined in previous steps. The higher percentage of the samples matching between assigning plots to groups using discriminant analysis and those obtained from optimal groups (generated by selected algorithm(s)) indicate a higher accuracy of classification.

Discriminant analysis was performed by using SPSS 24 (IBM Corp 2016). Environmental data were included in the analysis using a stepwise approach. Wilks' lambda and kappa statistics were used as evaluators of the analysis.

Results

The effectiveness of the classification methods

Rank the algorithms

The boxplots of showing the performances of various classification algorithms are shown in Figure 1. The average linkage, flexible-β linkage (β = -0.25), and k-means algorithms were selected as the best methods because they have a lesser number of outliers compared to the other methods.

Comparing the quartiles

Considering the 25, 50, and 75 quartiles (Table 1), k-means, flexible-β and average linkage were selected as the most effective algorithms. The distribution of data in their quartiles was systematic. In PAM, Modified TWINSPAN, and complete linkage the rationale between the quartiles were not justified. Data distribution in quartiles in Ward's linkage showed an intermediate trend.

Comparing the medians

K-means and flexible-β linkage (β = -0.25) tied for best performance in five of eight Kruskal-Wallis tests comparing geometric and non-geometric evaluator scores. The median scores of three geometric (ASW, 1-C-index, PARTANA) and two non-geometric (1-ISA p-val, ISA sig inds) evaluators were the highest for these algorithms. K-means was the only algorithm chosen by the mean lambda index.

Average linkage had a weaker performance compared to that of the flexible-β linkage and k-means algorithm. It was the best in the two non-geometric evaluators and tied for best in two geometric evaluators. PAM tied for best in two of four Kruskal-Wallis tests comparing non-geometric evaluator scores. Moreover, none of the geometric evaluators selected PAM. Modified TWINSPAN showed the weakest performance and was not selected by evaluators. Ward's algorithm tied for the best in one geometric and two non-geometric evaluators. Complete linkage tied for best in two of four Kruskal-Wallis tests comparing the scores of geometric evaluators. None of the non-geometric evaluators selected this method (Table 2).

Accordingly, across 20 cut levels, k-means produced results with more stability than the others. In terms of quality of results (evaluated by the eight internal evalua-

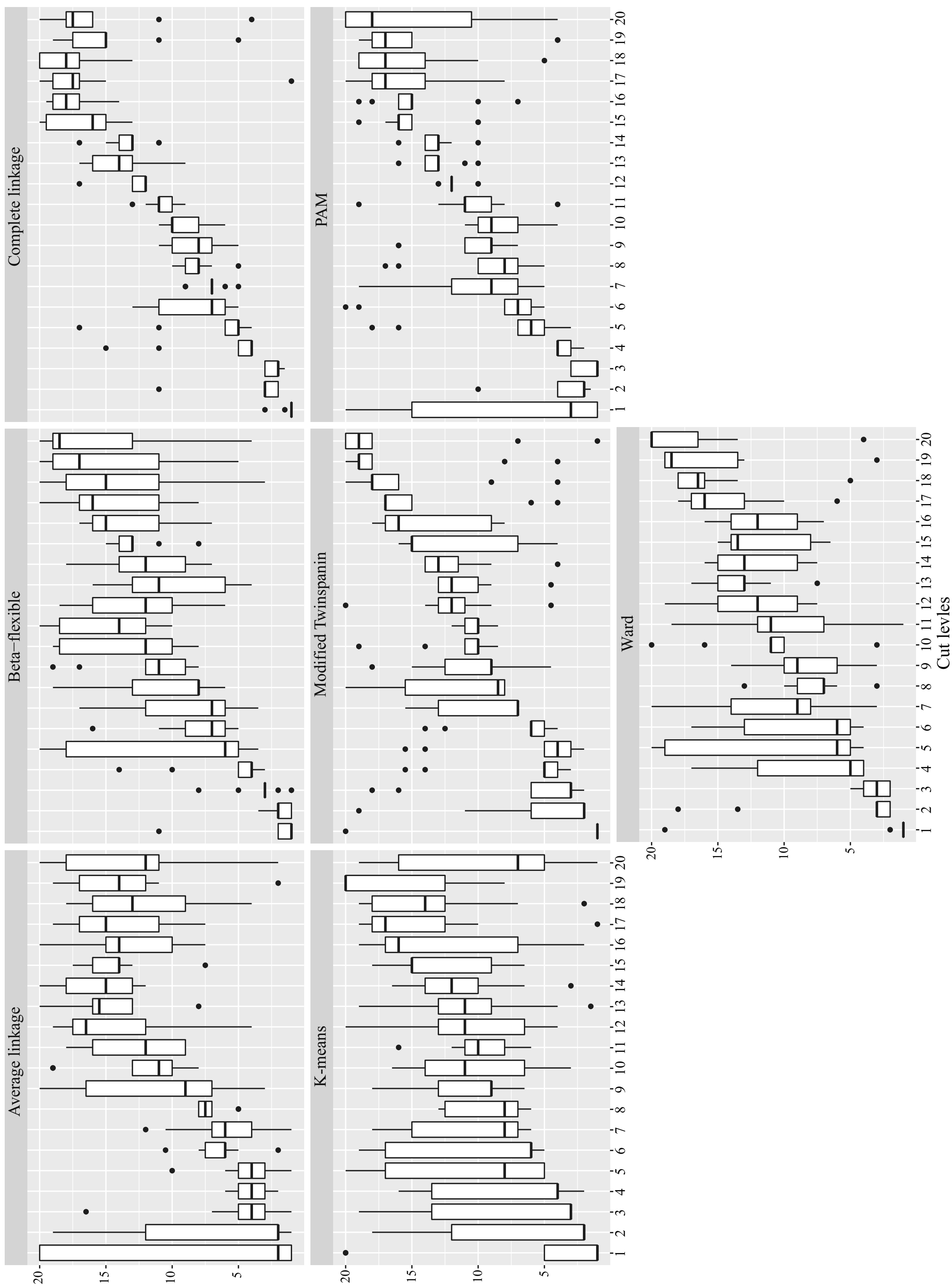

Fig. 1. Boxplot of algorithm ranks by evaluators and mean lambda for each cut level. The box extends from the first quartile to the third quartile. The dark line in the center of the box is the median. Whiskers extend to 1.5 the interquartile distance and dots represent outliers.

Table 1. Statistical quartiles calculated for algorithms based on evaluators and mean lambda values.

Classification algorithms	Quartiles		
	25	50	75
Average Linkage	5.47	14.84	23.69
Flexible β-Linkage	4.83	14.92	23.45
K-means	5.13	14.98	23.93
Wards Linkage	5.05	14.52	23.18
PAM	4.66	15	22.85
Modified TWINSPAN	4.62	15.71	22.82
Complete Linkage	4.63	15.14	22.77

tors), k-means and flexible-β linkage showed similar performance (across 20 cut levels). They were best or tied for the best in the five evaluators. In the rest of algorithms, average linkage had a better quality of results because it was best or tied for best in four evaluators. The performance of others was weak.

In total, comparing the boxplots and quartiles showed similar findings with Kruskal-Wallis tests. Therefore, among the chosen algorithms in this study, k-means, flexible-β linkage and average linkage showed a better performance comparing to the other methods. Considering the number of selection by evaluators as the best or tied for best algorithms, these algorithms did not differ remarkably. They were selected as appropriate algorithms to perform the next step analyses.

Optimum number of clusters

The results of selecting the optimum number of clusters are presented in Table 3. The optimum number of clusters for average linkage, flexible-β linkage and k-means methods were 13, 12 and 20, respectively.

Verification of the selected optimum number of clusters

The accuracy of the generated groups by each selected algorithm was evaluated using discriminant analysis (complete results are presented in the supplementary materials: tables S1 to S18). The percentage of samples matching between assigning plots to groups by using discriminant analysis and those obtained from optimal groups generated by the average linkage, flexible-β linkage and k-means algorithms were 48.2, 56.9 and 50.3, respectively.

The kappa coefficient also showed that the accuracy of the optimal clusters obtained from the average linkage,

Table 2. Medians for evaluator responses and mean lambda calculated across 20 classification levels (2-21 clusters) for each algorithm. 95% confidence intervals for medians were determined using interquartile range (IQR). Algorithm with the same letter are not significantly different at α = 0.05 using Kruskal-Wallis multiple comparisons. Best responses in each column are bolded. Abbreviations: A = ASW, B = 1- C.index, C = PARTANA, D = PBC, E = 1-ISA.pval, F = ISA.sig.inds, G = ISAMIC, H = 1- Morisita, I = "mean lambda".

	Geometric Evaluators				Nongeometric Evaluators				Lambda	
	A	B	C	D	E	F	G	H	I	Total
Average Linkage	0.07±0.01D	**0.86±0.014AB**	1.46±0.028B	**0.50±0.024A**	0.80±0.024D	49±1.93C	**0.80±0.017A**	**0.50±0.014A**	0.67±0.028CD	4
Flexible β-Linkage	**0.09±0.007AB**	**0.85±0.017AB**	**1.49±0.042A**	0.37±0.017CD	**0. 90±0.007A**	**68.5±1.76AB**	0.76±0.017B	0.36±0.031C	0.65±0.007CD	5
K-means	**0.10±0.003A**	**0.86±0.01A**	**1.51±0.035A**	0.40±0.017BC	**0.90±0.01A**	**71±1.14A**	0.76±0.017B	0.34±0.02C	**0.75±0.02A**	6
PAM	0.07±0.007DE	0.82±0.014CD	1.43±0.031B	0.33±0.010E	**0.89±0.01ABC**	**68±1.93AB**	0.76±0.014B	0.33±0.01C	0.70±0.017B	2
Modified TWINSPAN	0.04±0.01E	0.79±0.014D	1.43±0.045B	0.40±0.017CD	0.88±0.01BC	64.5±1.14B	0.77±0.017B	0.34±0.02C	0.16±0.037E	0
Wards Linkage	0.09±0.003BC	0.84±0.021BC	**1.47±0.042AB**	0.38±0.017DE	**0.89±0.003AB**	**69.5±0.96AB**	0.75±0.017B	0.35±0.01C	0.68±0.014BC	3
Complete Linkage	0.08±0.003CD	**0.84±0.024AB**	1.44±0.045B	**0.43±0.017AB**	0.88±0.017C	66±2.90B	0.77±0.017B	0.41±0.03B	0.60±0.042D	2

Table 3. The optimum number of clusters predicted by using Standardized evaluators and "mean lambda" index in selected classification algorithms. Max Clusters (optimum number of clusters) = the number of clusters at the cut level of classification on which the evaluators, "mean lambda" and the average of them show the highest value. Min cluster = the number of clusters at the cut level of classification on which the evaluators, "mean lambda" and the average of them show the lowest value. Avg = Average of standardized values of evaluators and mean lambda. CV%: coefficient of variance.

	A	B	C	D	E	F	G	H	I	Avg
					Average linkage					
CV%	45.77	5.91	4.84	22.38	19.53	28.62	3.35	9.07	23.11	
Max Clusters	2	17	2	15	10	10	21	2	14	13
Min Clusters	21	3	6	2	2	2	4	8	2	2
					Flexible-β linkage					
CV%	8.85	6.36	8.20	15.87	3.83	7.47	4.74	18.42	10.71	
Max Clusters	11	21	19	12	6	10	21	18	20	12
Min Clusters	4	2	3	3	2	3	2	2	2	2
					K-means					
CV%	12.10	4.8	5.6	8.01	4.08	7.80	4.62	16.16	7.98	
Max Clusters	2	20	20	6	20	13	20	20	2	20
Min Clusters	21	2	2	21	2	2	2	2	18	2

Table 4. Summary result of discriminant analysis.

Algorithm	Optimum number of clusters	Total variable	Variable used in the analysis	Number of significant canonical function	% of adaption	Kappa	Sig.
Average linkage	13	10	3	3	48.2	0.415	.000
flexible-β linkage	12	10	5	5	56.9	0.523	.000
K-means	20	10	5	5	50.3	0.475	.000

flexible-β, and the k-means method, 0.41, 0.52, and 0.47, respectively (Table 4). Accordingly, 12 groups presented as optimal numbers in the flexible-β linkage (β = -0.25), had more adaptation with groups created using environmental factors.

Discussion

Choosing the right criteria for classification analysis is a critical step toward gaining an interpretable classification (Roberts 2015). Here, we aimed to extend and improve the previous knowledge about this process. Compared to the previously published methods (e.g., Bolshakova & Azuaje 2006; Aho et al. 2008; Tichý et al. 2011; Lotter et al. 2013; Roberts 2015), we used comprehensive evaluators and visualized the pattern of the performances of the classification methods.

Evaluation of the efficiency of the classification algorithms is a complicated process to find the most suitable algorithm (Roberts 2015). Using a single evaluation method is not advisable and simultaneous use of various evaluators can lead to more acceptable results (Bolshakova & Azuaje 2006). In the present study, internal evaluators and lambda index were simultaneously used to evaluate the performance of the algorithms as well as the optimal number of clusters. Although flexible-β linkage and k-means showed a similar performance considering internal evaluators only their stability results were different in the lambda statistic.

In the previous studies, based on the evaluators, the differences among the responses of the algorithms were statistically investigated and visualization of the behaviour of each algorithm was not done in response to the evaluators at each cut-off level.

In comparing the boxplots, the number of outliers was used as a criterion to compare the performance of different algorithms at different cut-off levels. The resulted outliers in this study (Fig. 1) were not computational errors and they were generated due to the variability in the responsiveness of the algorithms. So, they were capable of causing problems in further statistical analyses (Maddala 1992).

Using the quartiles was another method to assess the effectiveness of the algorithms. Understanding how data are distributed in quartiles, is the importance of this method. Algorithm(s) with a more systematic distribution of data in the quartiles was considered as best method (Table 1). Systematic data distribution in quartiles indicates a smaller number of outlier data from quartile 25 to 50 and 50 to75. In fact, systematic distribution is related to outliers (Pallant 2001). As mentioned before, we used the average of evaluators and mean lambda to determine the optimum number of clusters in effective algorithm(s). Since the mean is generally sensitive to outlier data, therefore, to achieve more reasonable means and acceptable results, the appropriate algorithm(s) must generate the smallest number of outliers.

As suggested by Aho et al. (2008), the statistical analyses were performed using the medians. The results of the boxplots along with those of the medians (Table 2) showed a better performance of flexible-β linkage, k-means, and average linkage than the other algorithms. Furthermore, the results of both statistical and graphical methods were similar. So, our study revealed that boxplots also can be used to evaluate the efficiency of classification algorithms. Previous studies also reported k-means, flexible-β, and average linkage algorithms as the most efficient methods (e.g., Lance & Williams 1967; Milligan 1980; Kaufman & Rousseeuw 1990; McCune & Grace 2002; Aho et al. 2008; Lotter et al. 2013).

The k-means and flexible-β linkage are space-conserving algorithms that create spherical clusters (McCune et al. 2002; Kaufman & Rousseeuw 1990). This property may affect the selection of these algorithms by geometric evaluators which prefer spherical clusters. The average linkage is also space-conserving and generates spherical clusters, but may create chaining state comparing to the other space-conserving methods (McCune & Grace 2002). High chaining state can lead to individual clusters with relatively specific species composition that lacks species with an intermediate intra-group constancy. This phenomenon may contribute to the selection of this algorithm by non-ISA (i.e., Morisita and ISAMIC) evaluators (Table 2). ISA evaluators penalize species with low constancy, however non-ISA evaluators penalize those with intermediate intra-group constancy. In fact, the opposite performance of two groups of non-geometric evaluators (ISA and non-ISA evaluators) in the selection of average linkage is due to differences in the optimality criteria of them (Aho et al. 2008).

In determining the optimum number of clusters, the results of all the statistics (e.g., geometric, non-geometric evaluators, and the mean lambda) were combined because, in each of these, the attributes considered to select the optimal cut-off level are not the same. The geometric evaluators examine the performance of a classification based on the sampled plots and select the cut-off levels with high intra-cluster consistency. However, non-geometric evaluators acted on the basis of species, preferring a level of classification in which the number of indicator species (those most often present in one cluster and not in other clusters) was higher (Aho et al. 2008).

In the mean lambda statistic at each cut level of the algorithm, the classification of all sampling units is compared to the created subgroups (Lengyel et al. 2017). This statistic also selects the cut level(s) in which the most similarity between the original data and its subgroups is observed. Higher similarity indicates greater consistency of classification against casual changes in the dataset (Tichý et al. 2011). In fact, the mean lambda statistic is also an internal evaluator because it uses the algorithms' features to evaluate their stability.

Our results showed that the values of the ISA.sig.inds evaluator were different from those of the others (Table 2). Since the optimal number of clusters was evaluated based on the averaging of all the numerical values at each cut level, this difference has a significant effect on the resulted average and led to bias in the estimation. We standardized the results to eliminate such and for further studies, we advise to do so.

We determined the optimum number of clusters using the three selected algorithms. By using all three algorithms, we aimed not only to show how the optimum number of clusters varies based on the type of algorithm but also to compare them to estimate the optimal number of clusters that are interpretable in the study area. Since there is a close relationship between plant communities and the environmental conditions, environmental factors (as external evaluators) were used to compare and verify the optimal classification level specified in each algorithm. In the discriminant analysis based on environmental factors in addition to evaluating the accuracy of each of the optimum number of clusters, those affecting their separation were identified.

Accordingly, 12 clusters were found as the optimal numbers in flexible-β linkage (β = -0.25). In comparison with other numbers (i.e., 13 and 20), it showed more adaptation with groups created using the environmental factors. This adaptation indicates that these 12 clusters were more interpretable or have a higher justification in the study area. Because in terms of environmental characteristics, these clusters show the higher intra group and lower inter group differences.

Conclusion

Our findings imply that (a) selecting the effective algorithm(s) before an objective classification of any data, is not avoidable; (b) there is no absolute classification algorithm for all vegetation types and combination of evaluation indices used in this study can help the researchers to find the best classification methods for each dataset; (c) the results of different classification methods in determining the optimum number of clusters are disparate, and comparing these results is valuable and highlights dissimilarities of the various classifications.

Author contributions

H.E., O.E., and S.T. had designed the early framework of the research. O.E. and S.T. conducted field sampling and also identified the collected plants. S.T. analyzed the data and wrote the first draft. The final manuscript is revised and approved by all of the authors.

Acknowledgements

We wish to thank the office of the Vice President for Research and Technology, Ferdowsi University of Mashhad (FUM) for the financial support (under grant number 41791). Special thanks go to the "Department of Forest sciences and engineering" of Tarbiat Modares University for giving us the field sampling facilities. We wish to thank H. Asadi, S. Sekhavat and Z. Atashgahi for their help in the field trips, M. Aghabagheri and S. Zahraeifard for their assistance in conducting certain analyses in SPSS and R, and M.B. Erfanian for his comments on the earlier drafts of this manuscript. We acknowledge K. Aho for suggesting useful strategies for performing some analyses.

References

Aho, K. 2006. *Multivariate clustering for objective classification of vegetation data*. Paper was presented at the Billings Land Reclamation Symposium. Proceedings America Society of Mining and Reclamation Lexington, KY.

Aho, K. 2015. *plant.ecol: Quantitative tools for plant ecology*. R package version 0.4-1.

Aho, K. 2016. *asbio: A Collection of Statistical Tools for Biologists*. R package version 1.3-4. URL: https://CRAN.R- project.org/package=asbio

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

Bolshakova, N. & Azuaje, F. 2006. Estimating the number of clusters in DNA microarray data. *Methods of Information in Medicine* 43: 153–158.

Brogden, H.E. 1949. A new coefficient: application to biserial correlation and to estimation of selective efficiencys. *Psychometrika* 14: 169–182.

Botta-Dukat, Z., Chytry, M., Hajkova, P. & Havlova, M. 2005. Vegetation of lowland wet meadows along a climatic continentality gradient in Central Europe. *Preslia* 77: 89–111.

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

De Cáceres, M. & Wiser, S.K. 2012. Towards consistency in vegetation classification. *Journal of Vegetation Science* 23: 387–393.

Dengler, J., Chytrý, M. & Ewald, J. 2008. Phytosociology. In: Jörgensen, S.E. & Fath, B.D. (eds.), *Encyclopedia of ecology*, pp. 2767–2779, Elsevier, Oxford.

Dobrovic, I., Safner, T., Jelaska, S.D. & Nikolic, T. 2007. Ecological and phytosociological characteristics of the association Abieti-Fagetum «pannonicum» prov. on Mt. Edvednica (NW Croatia). *Acta BotanicaCroatica* 65: 41–55.

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

Dunn, O.J. 1964. Multiple Comparisons Using Rank Sums. *Technometrics* 6: 241–252.

Esmailzadeh, O., Hosseini, S.M., Tabari, M., Baskin, C.C. & Asadi, H. 2011. Persistent soil seed banks and floristic diversity in Fagus orientalis forest communities in the Hyrcanian vegetation region of Iran. *Flora* 206: 365–372.

Gan, G., Ma, C. & Wu, J. (2007). *Data Clustering: theory, algorithms, and applications* (*ASA-SIAM series on statistics and applied probability*). SIAM , Philadelphia (PA).

Gauch, H.G. & Whittaker, R.H. 1981. Hierarchical Classification of Community Data. *Journal of Ecology* 69: 537–557.

Gholizadeh, H., Naqinezhad, A. & Chytry, M. 2019. Classification of the Hyrcanian forest vegetation, northern Iran. *Applied Vegetation Sience*, doi:10.1111/avsc.12469.

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

Hartigan, J.A. & Wong, M.A. 1979. A k-means clustering algorithm. *Journal of the Royal Statistical Society* 28: 100–108.

Horn, H.S. 1966. Measurement of "overlap" in comparative ecological studies. *The American Naturalist* 100: 419–424.

Hubert, L.J. & Levin, J.R. 1976. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin* 83: 1072–1080.

IBM Corp. Released 2016. *IBM SPSS Statistics for Windows, Version 24.0*. Armonk, NY: IBM Corp.

Jennings, M.D., Faber-Langendoen, D., Loucks, O.L., Peet, R.K. & Roberts, D. 2009. Standards for associations and alliances of the US National Vegetation Classification. *Ecological Monographs* 79: 173–199.

Kaufman, L. & Rousseeuw, J.P. 1990. *Finding groups in data: an introduction to cluster analysis*. New York (NY), Wiley, (chapter 2, 5).

Kruskal, W.H. & Wallis, W.A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47: 583–621.

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

Legendre, P. & Legendre, L.F.J. 2012. *Numerical ecology*. 3th ed. Elsevier Science, Netherlands, Amsterdam.

Lengyel, A., Landucci, F., Mucina, L., Tsakalos, J. & Botta-Dukát, Z. 2017. Joint optimization of cluster number and abundance transformation for obtaining effective vegetation classifications. *Journal of Vegetation Science* 29: 336–347.

Lötter, M.C., Mucina, L. & Witkowski, E.T.F. 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.

Marvie Mohadjer, M.R. 2005. *Silviculture*. University of Tehran, Tehran.

McCune, B., Grace, J.B. & Urban, D.L. 2002. *Analysis of ecological communities*. Gelenden Beach OR, MjM Software Design.

McGill, R., Tukey, J.W. & Larsen, W.A. 1978. Variations of box plots. *American Statistician* 32: 12–16.

Maddala, G.S. 1992. "*Outliers". Introduction to Econometrics* (2nd ed.). New York, MacMillan.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik K. 2016. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.5.

Milligan, G.W. 1980. An examination of six types of error perturbation on fifteen Clustering algorithms. *Psychometrika* 45: 325–342.

Oksanen, A.J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., Hara, R.B.O., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E. & Wagner, H. 2017. *vegan: community ecology package. R package version 2.4.4*. URL: https://CRAN.R-project.org/package=vegan.

Pallant, J. 2001. *SPSS Survival Manual*. Open University Press, Philadelphia (PA).

Peet, R.K. & Roberts, D.W. 2013. Classification of natural and semi - natural vegetation. In: van der Maarel, E. & Franklin, J. (eds.), *Vegetation ecology*, pp. 28–71. Wiley, [place unknown].

R Development Team. 2018. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna. URL: http://www.Rproject.org.

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

Roberts, D.W. 2016a. *labdsv: Ordination and Multivariate Analysis for Ecology. R package version 1.8.0*. URL: https://CRAN.R-project.org/package=labdsv

Roberts, D.W. 2016b. *Optpart: Optimal part partitioning of similarity relations. R package version 2.3-0*. URL. URL: https://CRAN.R- project.org/package=optpart

Roleček, J., Tichý, L., Zelený, D. & Chytrý, M. 2009. Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. *Journal of Vegetation Science* 20: 596–602.

Rousseeuw, P.J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53–65.

Sagheb-Talebi, K., Sajedi, T. & Pourhashemi, M. 2014. *Forests of Iran: A treasure from the Past, a hope for the future*. Springer,

Schmidtlein, S., Tichý, L., Hannes, F. & Faude, U. 2010. A brute force approach to vegetation classification. *Journal of Vegetation Science* 21: 1162–1171.

Slezák, M., Hrivnák, R., Ujhazy, K., Ujházyova, M., Máliš, F. & Petrasova, A. 2016. Syntaxonomy and ecology of acidophilous beech forest vegetation in Slovakia. *Phytocoenologia* 46: 69–87.

Sokal, R.R. & Michener, C.D. 1958. A statistical method for evaluating systematic relationships. *The University of Kansas Science Bulletin* 38: 1409–1438.

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

Tibshirani, R., Walther, G. & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of Royal Statistical Society. B.* 63: 411–423.

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.

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

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

Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association* 58: 236–244.

Wickham, H. 2007. Reshaping Data with the reshape Package. *Journal of Statistical Software* 21: 1–20.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Zeleny, D., Smilauer P., Hennekens S.M. & Hill, M.O. 2016. *twinspanR: two-way indicator species analysis (and its modified version) in R. R package version 0.16*. URL: https://github.com/zdealveindy/twinspanR.

Author addresses

Tavakoli, S. (s.tavakoli@mail.um.ac.ir)[1], **Ejtehadi, H.** (Corresponding author, hejtehadi@um.ac.ir)[1], **Esmailzadeh, O.** (Corresponding author, oesmailzadeh@modares.ac.ir)[2]

[1] Quantitative Plant Ecology and Biodiversity Research Lab., Department of Biology, Faculty of Science, Ferdowsi University of Mashhad, Mashhad, Iran

[2] Department of Forest Sciences and Engineering, Faculty of Natural Resources, Tarbiat Modares, Tehran, Iran

Electronic supplements

Supplementary material associated with this article is embedded in the article's pdf. The online version of *Phytocoenologia* is hosted at the journal's website www.schweizerbart.com/journals/phyto. The publisher does not bear any liability for the lack of usability or correctness of supplementary material.

Supplement S1–18: Discriminant analysis.

Please save the electronic supplement contained in this pdf-file by clicking the blue frame above. After saving rename the file extension to .zip (for security reasons Adobe does not allow to embed .exe, .zip, .rar etc. files).