Western North American Naturalist

Volume 66 | Number 3 Article 3

8-10-2006

RCLUS, a new program for clustering associated species: a demonstration using a Mojave Desert plant community dataset

Stewart C. Sanderson
Shrub Sciences Laboratory, Provo, Utah

Jeffrey E. Ott
University of North Carolina at Chapel Hill

E. Durant McArthur
Shrub Sciences Laboratory, Provo, Utah

Kimball T. Harper
Utah Valley State College

Western North American Naturalist 66(3), © 2006, pp. 285–297

RCLUS, A NEW PROGRAM FOR CLUSTERING ASSOCIATED SPECIES: A DEMONSTRATION USING A MOJAVE DESERT PLANT COMMUNITY DATASET

Stewart C. Sanderson[1,4], Jeffrey E. Ott[2],
E. Durant McArthur[1,5], and Kimball T. Harper[3]

ABSTRACT.—This paper presents a new clustering program named RCLUS that was developed for species (R-mode) analysis of plant community data. RCLUS identifies clusters of co-occurring species that meet a user-specified cutoff level of positive association with each other. The "strict affinity" clustering algorithm in RCLUS builds clusters of species whose pairwise associations all exceed the cutoff level, whereas the "coalition" clustering algorithm only requires that the mean pairwise association of the cluster exceeds the cutoff level. Both algorithms allow species to belong to multiple clusters, thus accommodating both generalist and specialist species. Using a 60-plot dataset of perennial plants occurring on the Beaver Dam Slope in southwestern Utah, we carried out RCLUS analyses and compared the results with 2 widely used clustering techniques: UPGMA and PAM. We found that many of the RCLUS clusters were subsets of the UPGMA and PAM clusters, although novel species combinations were also generated by RCLUS. An advantage of RCLUS over these methods is its ability to exclude species that are poorly represented in a dataset as well as species lacking strong association patterns. The RCLUS program also includes modules that assess the affinity of a given species, plot, or environmental variable to a given cluster. We found statistically significant correlations between some of the RCLUS species clusters and certain environmental variables of the study area (elevation and topographical position). We also noted differences in clustering behavior when different association coefficients were used in RCLUS and found that those incorporating joint absences (e.g., the phi coefficient) produced more clusters and more even numbers of species per cluster than those not incorporating joint absences (e.g., the Jaccard index). In addition to the species association application described in this paper, the RCLUS algorithms could be used for preliminary data stratification in sample (Q-mode) analysis. The indirect link between sample plots and RCLUS species clusters could also be exploited to yield a form of "fuzzy" classification of plots or to characterize species pools of plots.

Key words: species association, cluster analysis, k-*means, hierarchical clustering, TWINSPAN,* Ambrosia, Larrea, Chrysothamnus, Senecio, Eriogonum, Prunus.

Plant community data consisting of species recorded in sample plots can be analyzed using a variety of numerical techniques. These may aim to uncover relationships among samples (normal or Q-mode analysis) or among species (inverse or R-mode analysis; Williams and Lambert 1961, Wilson et al. 1990, Legendre and Legendre 1998, McCune and Grace 2002). A species-based approach yields information on the degree to which species co-occur or show correlated abundance in a given set of samples. Co-occurrence is most easily measured using indices of association between species pairs, which can then be used to extract patterns of association among multiple species (Ludwig and Reynolds 1988, Bartha 1992, Turner et al. 2004). Species association patterns can be depicted using traditional matrix sorting and plexus diagram techniques (McIntosh 1978, Ludwig and Reynolds 1988), although these are impractical for large or complex datasets. Quantitative extensions of these techniques, such as TWINSPAN (Hill et al. 1975, Gauch and Whittaker 1981) and nonmetric multidimensional scaling (Kruskal 1964), have also been used to identify and display patterns of species association (e.g., Matthews 1978, Tueller and Eckert 1987, Mwasumbi et al. 1994, Exner et al. 2002).

Groups of positively associated species are essentially equivalent to indicator species of plant community types (associations, syntaxa, etc.), except that the former may not correspond to species composition at any specific location, while the latter are defined by sample

[1]Shrub Sciences Laboratory, 735 N. 500 E., Provo, UT 84606.
[2]Department of Biology, CB #3280, 212 Coker Hall, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3280.
[3]Department of Biology, Utah Valley State College, Orem, UT 84058.
[4]Contact first author to obtain copies of RCLUS. E-mail: ssanderson@fs.fed.us
[5]Corresponding author. E-mail: dmcarthur@fs.fed.us

plots or releves (Ewald 2003, Biondi et al. 2004). The sample-based approach lends itself well to correlation analyses between communities and environmental variables, but these can also be obtained through a species-based approach, as we demonstrate in this paper.

The technique presented in this paper was developed partly in response to dissatisfaction with the results of TWINSPAN analysis of plant community data from topographically complex areas. Although TWINSPAN is appealing for classifying samples and species groups simultaneously, it often gives inadequate results for plant communities responding to more than 1 environmental gradient (van Groenewoud 1992, Lee and McDonald 1993, McCune and Grace 2002). We developed an alternative set of algorithms for identifying species groups in a program called RCLUS (R-mode Cluster Analysis), with versions written in Microsoft Visual Basic 6.0 (installable version) and Visual Basic for Applications (spreadsheet version). The RCLUS algorithms are closely related to agglomerative and nonhierarchical clustering techniques (Gauch 1982, Kaufman and Rousseeuw 1990, McCune and Grace 2002); hence, in this paper, we compare RCLUS with these methods rather than with TWINSPAN.

Two key features of the RCLUS technique are that it (1) extracts clusters of species that meet a user-specified level of association with each other and (2) allows species to be placed into more than 1 cluster. The 1st feature aligns RCLUS with statistical tests of species association in which a threshold significance level is designated, e.g., $P = 0.05$ when $\chi^2 = 3.841$; Ludwig and Reynolds 1988). However, RCLUS is more broadly construed to show patterns of species association at any level using any association index. Species not positively associated with others at a given level are excluded from RCLUS clusters, unlike conventional clustering methods that force all species into clusters.

The 2nd feature—the allowance for species to occur in more than a single cluster—also differs from conventional clustering methods. This feature accommodates the possibility that some species are strongly habitat specific while others are ecological generalists, occurring with many species in many communities (Fridley et al. 2003). Specialist and generalist species can also be identified using measures of fidelity obtained through algorithms such as indicator species analysis (Dufrêne and Legendre 1997) or COCKTAIL (Bruelheide 2000, Bruelheide and Chytry 2000). However, these algorithms require predefined clusters or classes, whereas RCLUS is a method for generating clusters.

METHODS

Clustering

The RCLUS methodology is related to methods we previously carried out using matrices of pairwise species association and ratios of observed to expected values (McArthur and Sanderson 1992b). RCLUS uses measures of species association based on 2×2 (pairwise) contingency tables. The program is currently configured to calculate the phi coefficient, chi square, Jaccard index, and Sorensen index (Ludwig and Reynolds 1988, Jackson et al. 1989). It can also accept similarity or distance matrices provided by the program user, allowing the use of other measures including those that incorporate species abundance. The applications described in this paper are based on species presence or absence only.

RCLUS forms clusters of positively associated species referred to as core groups and then calculates the degree to which the remaining species of the study are associated with each core group. We provide 2 clustering algorithms for the formation of core groups: "strict affinity" and "coalition." Both require that the user specify a cutoff affinity level that determines which species can be included in a core group. There is also an option for excluding outright those species with low occurrence in the dataset (<3 by default). Note that we use the term "affinity" interchangeably with "positive association" or "positive correlation" in the text that follows.

The strict affinity clustering algorithm identifies species that meet a cutoff level of pairwise association with each and every other species within a core group. A randomly selected species is used to initiate the 1st core group, and species are added by proceeding through each of the other species of the dataset in random order. If a species is not used within the existing group, it is used to begin a new group. On each of several passes through the dataset, each species is tested to determine if it fits within any of the existing core groups; each species is added to as many of the core groups as it fits. By specifying that only species not yet incorporated in any

existing cluster can be used to initiate a new cluster, the algorithm limits proliferation of clusters. The algorithm is repeated from the beginning as many times as chosen by the user (50 by default) using a random species order each time. A table is generated containing all the resulting clusters and the frequency of each.

The strict affinity algorithm is useful for producing clusters at a fine level of resolution but has the disadvantage that excessive numbers of variant clusters are produced when incompletely separated species groups are present. The user is left with the task of deciding which of the clusters are most informative. Clusters may be sorted according to their frequency of occurrence (in multiple runs of the program) or their cohesiveness (strength of association of core species), and the user may opt to screen clusters according to such criteria.

The coalition clustering algorithm is named for its relation to the "multi-species coalition" concept discussed by Bartha (1992), referring to groups of species indirectly linked through pairwise associations (see also McIntosh 1978). Similar to the strict affinity algorithm, the coalition algorithm evaluates species in a random order, adding them to clusters when the cutoff association criterion is met. A new species is added to a core group if the mean of its pairwise associations with existing core species meets the cutoff level of association for the group as a whole, even though it may not meet this level with every individual species in the group. Because addition of species to a group causes the character of the group to change to some degree, species may come to fit within the group that did not fit at an earlier time; or species may need to be removed if they no longer fit within the constraints of the selected cutoff level. For this reason, the procedure is repeated a sufficient number of times so that the composition of all of the clusters will have stabilized.

The coalition clustering algorithm also contains a module for testing the clusters against each other to determine if any of them have become alike enough that they need to be combined. The algorithm constructs a matrix of pairwise associations of the species in one cluster with the species in another cluster (including instances in which the same species occurs in both clusters). If the mean of all pairwise associations exceeds the cutoff value as previously defined, the clusters are merged. This merging step was originally done after each cycle through the species list, but consistency was improved by combining clusters only near the middle of the procedure, after clusters had time to stabilize. Several more cycles through the species list are carried out afterwards, to stabilize any combined clusters. We also found it best to limit formation of new core groups to the 1st few passes through the dataset so that the groups would have time to mature before other operations were carried out. The algorithm is currently set to add new clusters during the first 4 cycles through the species list, to combine similar clusters on the 9th and 14th cycles, and to continue for a total of 18 cycles. The program provides a display of the number of clusters resulting from each cycle. Those outputs can be used to judge whether clusters have stabilized.

Once core groups are formed by either of the clustering algorithms in RCLUS, the affinity (degree of positive association) between core groups and other species is assessed. This assessment is made by taking the mean of all pairwise associations between a given species and each of the species in a core group. Even core group species themselves can be assessed for affinity to their own core group, because they may differ in mean association with other members of the group. RCLUS is configured to display in its output a list of species in descending order of affinity to each core group. This list can give the user an indication of which species would likely be added to core group clusters were the cutoff criteria to be relaxed.

Another feature in RCLUS calculates and displays the affinity of each sample plot to each core group cluster. This is done by taking the mean affinity to a given cluster of the species present in the plot. Because plots may differ in species composition and species may have multiple affinities, each plot may have a unique set of affinities to each of the core group clusters. RCLUS can also display the number, percentage, or cover of cluster core species present in each plot. These data offer alternative ways of assessing the affinity between plots and clusters. RCLUS also includes a rudimentary statistical technique for assessing relationships between species clusters and environmental variables. For each core group species in a cluster, environmental values of

plots where the species is present are tabulated; values are also recorded for plots where the species is absent. The combined data for all species of a cluster are then analyzed with a chi-square test for discrete variables or a Kruskal-Wallace test for continuous variables (Pollard 1977).

Data

To demonstrate RCLUS, we performed analyses on a plant community dataset collected on the Beaver Dam Slope of southwestern Utah (McArthur and Sanderson 1992a). The dataset contains lists of perennial vascular plant species occurring in 60 circular 0.01-ha plots. Plots were located at or near cadastral survey section corners at 1.62-km (1-mi) intervals. Nomenclature of the 56 species in this dataset follows Welsh et al. (1993).

The study area covered 50 km^2 of alluvial slope between the Beaver Dam Mountains and Beaver Dam Wash in extreme southwestern Utah. Elevation across the study area ranged from 768 m to 988 m. Vegetation was characteristic of the Mojave Desert, with Joshua trees (*Yucca brevifolia*) and blackbrush (*Coleogyne ramosissima*) at higher elevations. Creosote bush (*Larrea tridentata*) and bur-sage (*Ambrosia dumosa*) dominated at lower elevations and on flatter terrain. The study site was dissected by dry washes that harbored distinctive vegetation including desert almond (*Prunus fasciculata*), wooly bursage (*Ambrosia eriocentra*), and Mojave rabbitbrush (*Chrysothamnus paniculatus*). Water is currently piped to stations throughout the area to support springtime cattle grazing. Based on our familiarity with the vegetation of this area, we expected species in the dataset to cluster into groups characteristic of contrasting topographical position and elevation.

RESULTS AND DISCUSSION

Comparison with Hierarchical Clustering

Species clusters resulting from RCLUS were compared with results of R-mode UPGMA (Unweighted Pair Group Method with Arithmetic Mean) hierarchical clustering of the Beaver Dam Slope dataset (Figs. 1, 2). UPGMA is a hierarchical clustering method commonly used in plant community studies (often for clustering samples rather than species). Like RCLUS, UPGMA builds clusters through agglomeration using a similarity or distance matrix (McCune and Grace 2002). UPGMA clustering was carried out in SAS using PROC CLUSTER (SAS Institute 1989). The Jaccard index was used for both the UPGMA and RCLUS analyses shown in Figs. 1 and 2. Figure 1 shows the composition of 15 clusters obtained by running the strict affinity algorithm at a cutoff affinity value of 0.25 (Fig. 1, left side). A total of 44 clusters were generated by this analysis, but only those containing more than 3 species and meeting other criteria (see Fig. 1 footnote) are shown. At this cutoff level, no cluster contained more than 13 of the 57 species in the dataset. The clusters showed considerable overlap in composition, e.g., a set of 8 clusters all containing *Ambrosia dumosa*, desert rue (*Thamnosma montana*), desert marigold (*Baileya multiradiata*) and wire lettuce (*Stephanomeria pauciflora*). Many of the RCLUS clusters showed a general correspondence to UPGMA clusters (Fig. 1, right side) but few were exactly the same. We expected hierarchical clustering to yield different results than strict affinity clustering because, in the former, (1) every species must be used in the analysis regardless of the strength of its association with other species and (2) once a species is placed into a cluster, it cannot be removed or placed elsewhere. Strict affinity clustering revealed species associations that hierarchical clustering did not, such as Brigham tea (*Ephedra nevadensis*)–Anderson's wolfberry (*Lycium andersonii*)–winterfat (*Ceratoides lanata*) [Cluster K]; goldenhead (*Acamptopappus sphaerocephalus*)–*Baileya multiradiata*–silver cholla (*Opuntia echinocarpa*)–range ratany (*Krameria parvifolia*)–rattlesnake weed (*Euphorbia albomarginata*) [Cluster M]; and bush encelia (*Encelia frutescens*)–paper bag bush (*Salzaria mexicana*)–*Prunus fasciculata*–burrobush (*Hymenoclea salsola*) [Cluster O].

The coalition clustering method proved effective at producing a more manageable number of clusters, with limited overlap in species composition, for a given cutoff affinity value. Figure 2 illustrates this result as well as showing changes in cluster composition at different cutoff affinity values, again compared against UPGMA cluster composition. At Jaccard = 0.65 (a stringent affinity criterion) only 2 clusters of 2 species each formed: *Ambrosia dumosa–Larrea tridentata* in the upper part of

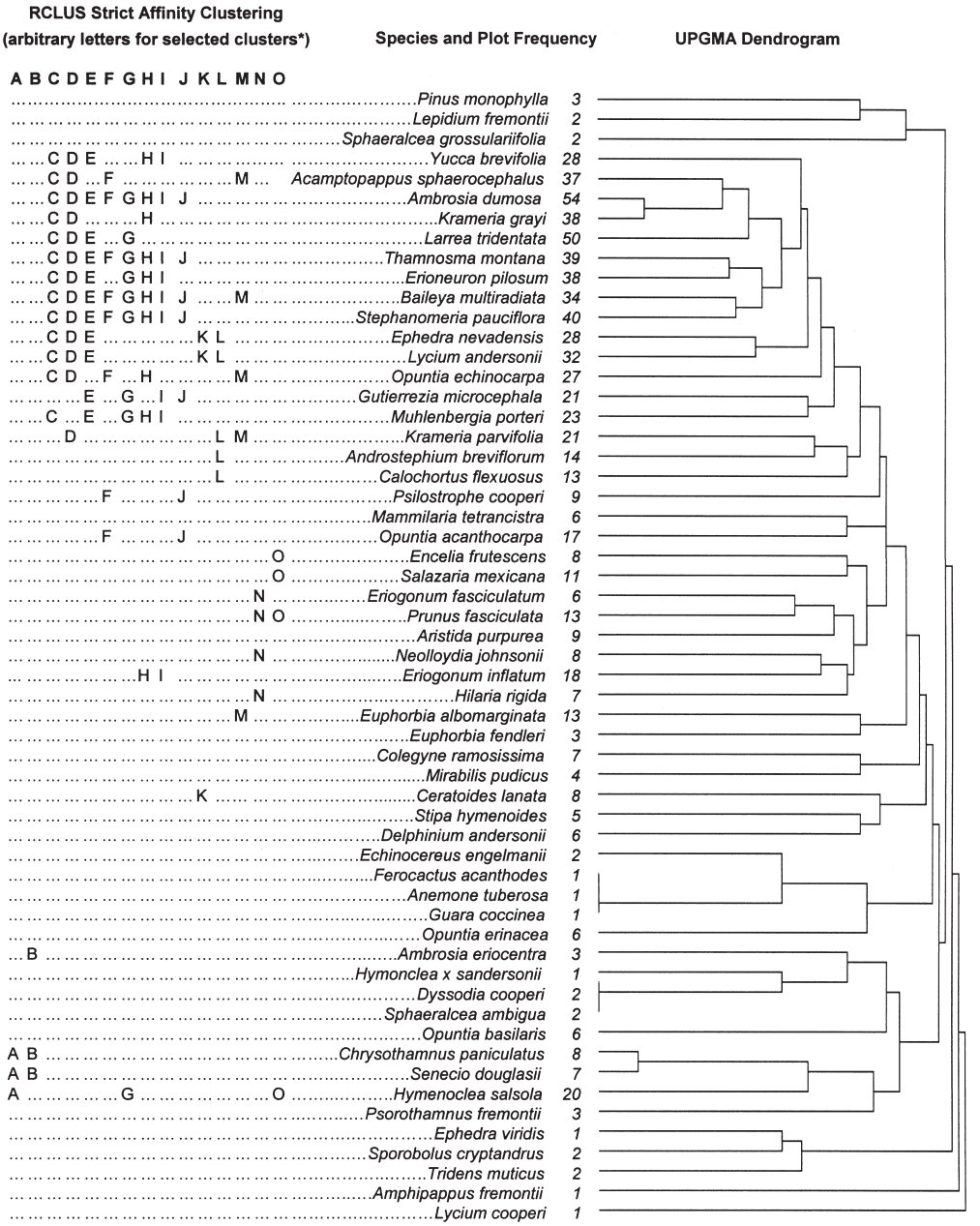

Fig. 1. Comparison of RCLUS strict affinity clustering (left) and UPGMA hierarchical clustering (right) of perennial plant species of 60 plots at Beaver Dam Slope, Utah. Species are ordered according to position on the UPGMA dendrogram to the right. Numbers to the right of species names show the number of plots in which the species occurred. Each letter in the left columns corresponds to a different cluster and indicates the core species of the cluster.

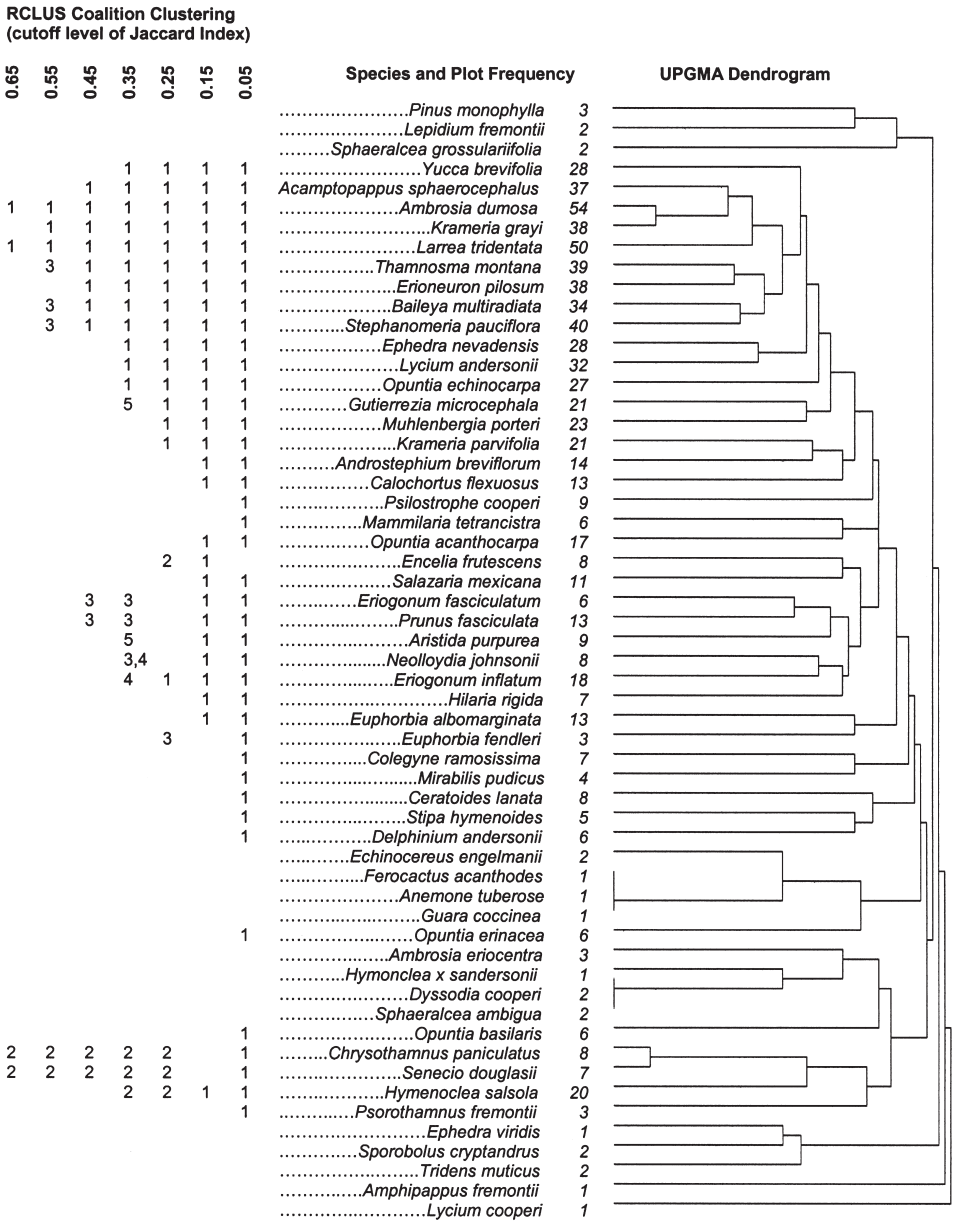

Fig. 2. Comparison of RCLUS coalition clustering (left) and UPGMA hierarchical clustering (right) of perennial plant species of 60 plots at Beaver Dam Slope, Utah. Species are ordered according to position on the UPGMA dendrogram to the right. Numbers to the right of species names show the number of plots in which the species occurred. Numbers within each column on the left indicate coalition clusters at different cutoff affinity levels and their corresponding core species (species without a number at a given cutoff level did not meet the criteria for core species).

the cluster dendrogram and *Chrysothamnus paniculatus*–Douglas groundsel (*Senecio douglasii*) in the lower part. At Jaccard = 0.45, the upper cluster had expanded to include 8 species, and an additional cluster (Mojave buckwheat [*Eriogonum fasciculatum*]–*Prunus fasciculata*) formed in the middle. Other small clusters were added at Jaccard = 0.35 and Jaccard = 0.25, but as the affinity criteria became less stringent these clusters joined together to become a single large cluster containing 41 species (nearly three-fourths of the dataset) at Jaccard = 0.05. Note that species failing to cluster at this level included those that are excluded by default because they occur with low frequency (1–2 occurrences) in the dataset. The increasing size of species clusters with decreasing affinity cutoff values noticeably mirrored the agglomerative pattern of the UPGMA dendrogram, especially for the upper cluster (Fig. 2). In the middle and lower clusters, the cluster joining pattern was disrupted at Jaccard = 0.35 and Jaccard = 0.25, respectively, before re-emerging at lower cutoff values. These appear to be transitional affinity values in which some, but not all, of the species in these clusters were able to join up with the emerging single large cluster (note bottlebush [*Eriogonum inflatum*] at Jaccard = 0.35 and *Hymenoclea salsola* at Jaccard = 0.25). This example highlights the program user's need to explore different cutoff levels in order to find cluster variants that are most stable and useful.

Comparison with *k*-Means Clustering

The class of clustering algorithms known as "*k*-means" identifies a set of nonnested clusters with optimal statistical properties. The number of clusters *k* is designated by the user prior to analysis (Kaufman and Rousseeuw 1990, McCune and Grace 2002). For comparison with RCLUS we used the *k*-means variant algorithm PAM (Partitioning Around Medoids), which identifies *k* representative objects (medoids) and clusters other objects with their closest medoid such that dissimilarity within clusters is minimized and dissimilarity between clusters is maximized (Kaufman and Rousseeuw 1990).

We carried out PAM on the Beaver Dam Slope dataset using the PAM procedure in the CLUSTER package of the R Project for Statistical Computing, version 2.0.1 (R Development Core Team 2004) at values of *k* between 2 and 10. Figure 3 shows results at $k = 6$ compared against the same RCLUS coalition clustering results presented in Fig. 2 (based on the Jaccard index in both cases). The right side of Fig. 3 shows not only the PAM cluster to which each species was assigned, but also the 2nd closest "neighbor" cluster and a "silhouette plot" which depicts the affinity of each species to its assigned cluster relative to its neighbor cluster. Silhouette bars approaching +1 indicate species that fit well in their assigned cluster, while bars near –1 are poorly classified and might actually fit better in a different cluster (Kaufman and Rousseeuw 1990).

Some of the RCLUS coalition cluster core species on the left side of Fig. 3 are subsets of the PAM clusters on the right side (e.g., the 3 coalition clusters resulting from a cutoff value of Jaccard = 0.45 are subsets of PAM clusters 1–3). Furthermore, these coalition cluster core species include the medoids of the PAM clusters as well as other species with high PAM silhouette values. This correspondence between coalition clusters and well-classified PAM objects supports our assertion that RCLUS includes only the best supported species groups in its output. However, note that this correspondence between RCLUS and PAM clusters breaks down at cutoff values less than Jaccard = 0.45, showing novel clusters that PAM analyses (including others not shown, at values of *k* between 2 and 10) did not reveal.

Some of the clusters identified by PAM (e.g., clusters 4–6 on the right side of Fig. 3) did not appear in coalition clusters because their species occurred with low frequency in the dataset. The high silhouette values assigned to these species by PAM draw unwarranted attention to clusters with low statistical support, a problem which RCLUS takes into account.

Comparison of Association Indices

An important step in R-mode community analysis is the selection of an appropriate measure of species association or correlation. Although many association indices are identical to sample similarity indices used in Q-mode analysis, the issues surrounding their use differ for each mode. In the Q-mode, dissimilarity or distance measures are often used

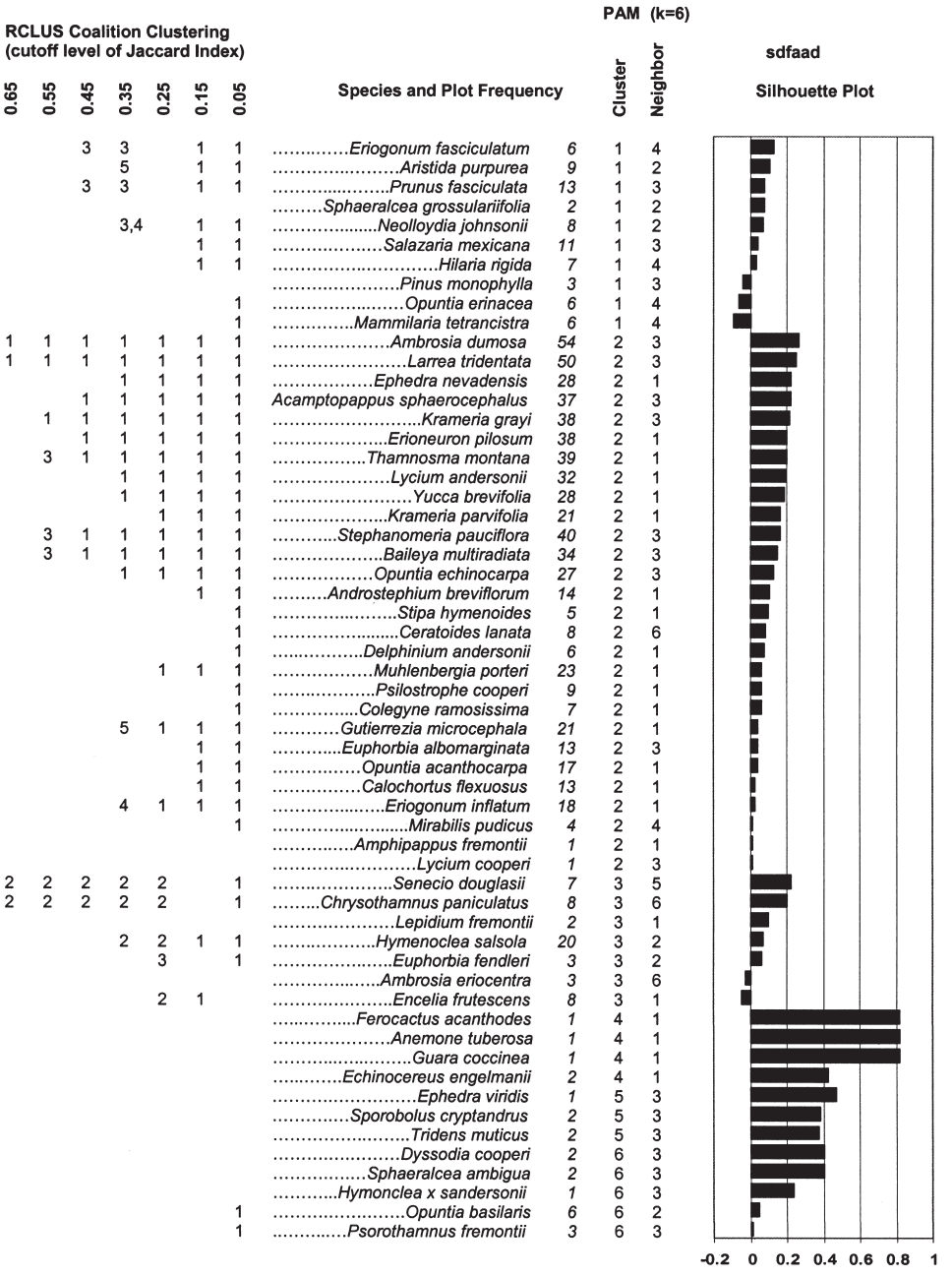

Fig. 3. Comparison of RCLUS coalition clustering (left) and PAM nonhierarchical clustering (right) of perennial plant species of 60 plots at Beaver Dam Slope, Utah. Species are ordered according to their PAM cluster and the size of their silhouette value (bars at right) within their assigned cluster. Numbers to the right of species names show the number of plots in which the species occurred. Numbers within each column on the left indicate coalition clusters at different cutoff affinity levels and their corresponding core species (species without a number at a given cutoff level did not meet the criteria for core species). See text for further discussion.

in lieu of similarity, and a primary consideration is the degree to which compositional distance corresponds with environmental distance (Faith et al. 1987, De'ath 1999, McCune and Grace 2002). In the R-mode, distance measures are nonintuitive and have been avoided in the literature, whereas contingency table coefficients are much more applicable despite debate over the ecological meaning of joint species absences (Janson and Vegelius 1981, Hubalek 1982, Ludwig and Reynolds 1988, Jackson et al. 1989, Legendre and Legendre 1998, Turner et al. 2004).

In RCLUS, we noted that contingency table coefficients incorporating joint absence information (phi, chi-square) yielded notably different results from association indices that do not include such information (Jaccard, Sorensen). For our dataset, the phi coefficient tended to yield a larger number of clusters, with more even numbers of core species per cluster, than the Jaccard index (Table 1). This appears to be related to the tendency of the Jaccard index (and related indices) to yield clusters in which species with similar numbers of occurrences are grouped together (Jackson et al. 1989). Upon examination, the larger clusters obtained with the Jaccard index were found to consist of species of frequent occurrence and to represent vegetation types somewhat similar to those obtained with the phi coefficient. On the other hand, smaller clusters obtained with the Jaccard index, usually 1–3 species each, consisted of species of few occurrences and did not resemble clusters from phi. We concluded that the species were grouped primarily according to number of occurrences and only secondarily according to association. Although this property of the Jaccard index may be useful for some purposes, we found the more direct association information provided by the phi coefficient to be preferable for our dataset. We do not consider joint absences to be ecologically meaningless in this instance because our study area covered a limited range of environmental conditions at a spatial scale within which all species could presumably disperse.

TABLE 1. Number of clusters and cluster size evenness generated by the RCLUS coalition clustering algorithm at different cutoff affinity values of the Jaccard index and phi coefficient for a dataset of perennial plant species in 60 plots at the Beaver Dam Slope, Utah. Evenness was calculated using the Simpson index (Ludwig and Reynolds 1988).

Cutoff affinity	Clusters	Cluster evenness
JACCARD INDEX		
0.15	1	1.00
0.20	3	0.46
0.25	3	0.54
0.30	6	0.42
0.35	5	0.57
0.40	5	0.53
0.45	3	0.67
0.50	2	0.76
0.55	3	0.93
0.60	3	0.96
0.65	2	1.00
0.70	2	1.00
PHI COEFFICIENT		
0.05	1	1.00
0.10	2	0.96
0.15	4	0.80
0.20	6	0.75
0.25	10	0.73
0.30	13	0.72
0.35	15	0.80
0.40	13	0.91
0.45	8	0.89
0.50	4	0.82
0.55	4	0.90
0.60	2	1.00

Environmental Correlates of Clusters

Our preferred set of clusters for the Beaver Dam Slope dataset was obtained using the coalition clustering algorithm with the phi coefficient at a cutoff affinity value of 0.18. The 5 clusters resulting from this analysis (Table 2, left column) agreed with our observations of species occurring at differing elevations and topographic settings within the study area. Correlation tests confirmed that 4 of the 5 clusters could be differentiated based on a combination of topographic position (upland or wash affinity) and elevation (positive or negative correlation) at a significance level of $\alpha = 0.1$ (Table 3, upper); hence, we assigned names to the clusters presented in Table 2. The 5th cluster, containing 2 species in the genus *Opuntia,* was correlated with elevation but not topographic position. Values of unmeasured variables such as grazing history might be needed to fully explain this cluster.

The right side of Table 2 shows 2 clusters obtained using a less stringent cutoff affinity value (phi = 0.08); they could be differentiated by topographic position but not by elevation

TABLE 2. Coalition cluster core species at 2 cutoff association levels of the phi coefficient for data collected at the Beaver Dam Slope, Utah. See Table 3 for environmental correlates of clusters.

Coalition cluster core species; Cutoff association value: Phi = 0.18	
CLUSTER 1: HIGHER ELEVATION UPLANDS (MEAN PAIRWISE PHI = 0.2544)	
Brigham tea	*Ephedra nevadensis*
Anderson's wolfberry	*Lycium andersonii*
Indian ricegrass	*Stipa hymenoides*
Funnel lily	*Androstephium breviflorum*
Winterfat	*Ceratoides lanata*
Range ratany	*Krameria parvifolia*
Sinuous mariposa	*Calochortus flexuosus*
Anderson larkspur	*Delphinium andersonii*
CLUSTER 2: LOWER ELEVATION UPLANDS (MEAN PAIRWISE PHI = 0.3258)	
Bur-sage	*Ambrosia dumosa*
Creosote bush	*Larrea tridentata*
Goldenhead	*Acamptopappus sphaerocephalus*
CLUSTER 3: HIGHER ELEVATION WASHES (MEAN PAIRWISE PHI = 0.2544)	
Desert peach	*Prunus fasciculata*
Mojave buckwheat	*Eriogonum fasciculatum*
Three awn grass	*Aristida purpurea*
Thread snakeweed	*Gutierrezia microcephala*
Bush muhly (grass)	*Muhlenbergia porteri*
Paper bag bush	*Salazaria mexicana*
Neolloydia cactus	*Neolloydia johnstonii*
Fluff grass	*Erioneuron pilosum*
Big galleta grass	*Hilaria rigida*
Bush encelia	*Encelia frutescens*
Bottlebush	*Eriogonum inflatum*
Desert marigold	*Baileya multiradiata*
Wire lettuce	*Stephanomeria pauciflora*
Burrobush	*Hymenoclea salsola*
CLUSTER 4: LOWER ELEVATION WASHES (MEAN PAIRWISE PHI = 0.2865)	
Burrobush	*Hymenoclea salsola*
Mojave rabbitbrush	*Chrysothamnus paniculatus*
Douglas groundsel	*Senecio douglasii*
Bush encelia	*Encelia frutescens*
Wire lettuce	*Stephanomeria pauciflora*
Desert peach	*Prunus fasciculata*
Indigobush	*Psorothamnus fremontii*
Desert marigold	*Baileya multiradiata*
Paper bag bush	*Salazaria mexicana*
CLUSTER 5: LOWER ELEVATIONS—CACTI (MEAN PAIRWISE PHI = 0.2568)	
Beavertail cactus	*Opuntia basilaris*
Silver cholla	*Opuntia echinocarpa*

(Table 3, lower). The coalition algorithm thus recovered the expected contrast between upland and wash species composition in 2 clusters of core species. Interestingly, there was no overlap in core species composition between these 2 clusters, suggesting an absence of generalist species common to both upland and wash environments. In contrast, the 2 wash clusters at phi = 0.18 overlap considerably in core species composition (Table 2, clusters 3 and 4), suggesting that wash-adapted species differ in their degree of specialization along the elevational gradient of the washes. Although these patterns may be partially an artifact of the sampling strategy of this study, they illustrate insights made possible by allowing species membership in multiple clusters. Because of this property, RCLUS may be particularly useful for characterizing fine compositional distinctions among nondiscrete communities. RCLUS clustering algorithms have the potential to detect environmental sorting of species even in settings where species respond differently to multiple environmental gradients.

TABLE 2. Continued.

Coalition cluster core species; Cutoff association value: Phi = 0.08	
CLUSTER 1: UPLANDS (MEAN PAIRWISE PHI = 0.1536)	
Brigham tea	*Ephedra nevadensis*
Anderson's wolfberry	*Lycium andersonii*
Range ratany	*Krameria parvifolia*
Indian ricegrass	*Stipa hymenoides*
Funnel lily	*Androstephium breviflorum*
Winterfat	*Ceratoides lanata*
Creosote bush	*Larrea tridentata*
Bur-sage	*Ambrosia dumosa*
Anderson larkspur	*Delphinium andersonii*
Joshua tree	*Yucca brevifolia*
Sinuous mariposa	*Calochortus flexuosus*
Whitestem paperflower	*Psilostrophe cooperi*
Goldenhead	*Acamptopappus sphaerocephalus*
CLUSTER 2: WASHES (MEAN PAIRWISE PHI = 0.1837)	
Mojave buckwheat	*Eriogonum fasciculatum*
Desert peach	*Prunus fasciculata*
Three awn grass	*Aristida purpurea*
Neolloydia cactus	*Neolloydia johnstonii*
Thread snakeweed	*Gutierrezia microcephala*
Bush encelia	*Encelia frutescens*
Fluff grass	*Erioneuron pilosum*
Bush muhly (grass)	*Muhlenbergia porteri*
Big galleta grass	*Hilaria rigida*
Paper bag bush	*Salazaria mexicana*
Bottlebush	*Eriogonum inflatum*
California pincushion	*Mammillaria tetrancistra*
Desert marigold	*Baileya multiradiata*
Wire lettuce	*Stephanomeria pauciflora*
Desert rue	*Thamnosma montana*
Grizzlybear pricklypear	*Opuntia erinacea*
Burrobush	*Hymenoclea salsola*
Sinuous mariposa	*Calochortus flexuosus*
Buckhorn cholla	*Opuntia acanthocarpa*
Rattlesnake-weed	*Euphorbia albomarginata*

We consider the clustering algorithms to be the primary strength of RCLUS, whereas the RCLUS methods for correlating clusters with environmental variables have more limited value. We expect that more sophisticated environmental correlation methods such as those used in habitat modeling (Guisan and Zimmermann 2000) could be used profitably in conjunction with RCLUS clustering methods.

Other Applications

The RCLUS algorithms resemble methods that were developed in the 1970s for preliminary clustering of plant community samples. Janssen (1975) described an algorithm for placing sample plots in clusters meeting a predefined threshold level of similarity. This is essentially equivalent to a single run of the RCLUS strict affinity algorithm, except that it clusters samples rather than species and allows each sample to occur in only 1 cluster. Gauch (1980, 1982) included a similar algorithm, with additional steps for merging small clusters into large ones, in his composite clustering routines. These methods were promoted as tools for removing outliers and simplifying large datasets prior to applying other analyses such as hierarchical classification or ordination. The latter analyses could be applied either (1) to clusters in lieu of individual samples or (2) to samples within individual clusters. RCLUS, applied to samples rather than species, could potentially be used for these same purposes, especially the 2nd (analysis within clusters) in which the possibility of overlapping clusters would likely be an asset rather than a drawback. The computational restraints that limited these early methods are now less serious;

TABLE 3. Environmental correlates of coalition cluster core species for data collected at the Beaver Dam Slope, Utah, dataset at 2 cutoff affinity levels of the phi coefficient. Topographic position was tested using a chi-square test, and elevation using the Kruskal-Wallis test. See Table 2 for species composition of clusters.

(Environmental correlates)		Position		Elevation	
		Category	*P*-value	Correlation	*P*-value
CUTOFF ASSOCIATION VALUE: PHI = 0.18					
Cluster 1	Higher elevation uplands	Upland	0.05	Positive	0.005
Cluster 2	Lower elevation uplands	Upland	0.025	Negative	0.1
Cluster 3	Higher elevation washes	Wash	0.005	Positive	0.005
Cluster 4	Lower elevation washes	Wash	0.005	Negative	0.01
Cluster 5	Lower elevations—cacti	Wash	0.9	Negative	0.05
CUTOFF ASSOCIATION VALUE: PHI = 0.08					
Cluster 1	Upland slopes	Upland	0.025	Positive	0.005
Cluster 2	Washes	Washes	0.005	Positive	0.01

hence, RCLUS can run its clustering algorithms multiple times in a matter of seconds (for datasets of the size we have presented), yielding a better picture of possible clusters and allowing selection of more optimal ones.

Other potential applications of RCLUS could employ the species clustering approach that has been presented here, taking advantage of the indirect link between sample plots and species clusters. If species clusters are viewed as the equivalent of indicator species of plant communities, then sample plots can be classified into communities according to their affinity to species clusters. Because a sample plot may have affinity to more than 1 species cluster, this type of classification would be "soft" rather than "hard," akin to fuzzy classification methods (Equihua 1990, Nicholls and Tudorancea 2001). RCLUS could also find application in studies seeking to characterize species pools or the set of species that could potentially occur at a site (Grace 2001, Ewald 2002). In this case, species clusters with high affinity to 1 or more plots would indicate the species pool of the plot(s). As in other types of community analysis, spatial scale and sampling strategy must be considered because they affect patterns that will be detected by RCLUS.

The RCLUS program is still in development and will likely gain new features and improvements in the near future. More rigorous tests of the strict affinity and coalition clustering algorithms are needed to determine the full extent of their strengths and weaknesses. Comparisons with additional methods of community analysis would also be valuable. Based on results of our preliminary analyses, we believe that RCLUS offers unique features with promising applications for community studies.

LITERATURE CITED

BARTHA, S. 1992. Preliminary scaling for multi-species coalitions in primary succession. Abstracta Botanica 16:31–41.

BIONDI, E., E. FEOLI, AND V. ZUCCARELLO. 2004. Modeling environmental responses of plant associations: a review of some critical concepts in vegetation study. Critical Reviews in Plant Sciences 23:149–156.

BRUELHEIDE, H., 2000. A new measure of fidelity and its application to defining species groups. Journal of Vegetation Science 11:167–178.

BRUELHEIDE, H AND M. CHYTRY. 2000. Towards unification of national vegetation classifications: a comparison of two methods for analysis of large data sets. Journal of Vegetation Science 11:295–306.

DE'ATH, G. 1999. Extended dissimilarity: a method of robust estimation of ecological distances from high beta diversity data. Plant Ecology 144:191–199.

DUFRÊNE, M., AND P. LEGENDRE. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67: 345–366.

EQUIHUA, M. 1990. Fuzzy clustering of ecological data. Journal of Ecology 78:519–534.

EWALD, J. 2002. A probabilistic approach to estimating species pools from large compositional matrices. Journal of Vegetation Science 13:191–198.

_____. 2003. A critique for phytosociology. Journal of Vegetation Science 14:291–296.

EXNER, A., W. WILLNER, AND G. GRABHERR. 2002. *Picea abies* and *Abies alba* forests of the Austrian Alps: numerical classification and ordination. Folia Geobotanica 37:383–402.

FAITH, D.P., P.R. MINCHIN, AND L. BELBIN. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57–68.

FRIDLEY, J.D., D.B. VANDERMAST, AND D. KUPPINGER. 2003. Assessment of habitat specialization of southeastern trees using large-extent co-occurrence data. Ecological Society of America Annual Meeting Abstracts 88:342.

GAUCH, H.G. 1980. Rapid initial clustering of large datasets. Vegetatio 42:103–111.

_____. 1982. Multivariate analysis in community ecology. Cambridge University Press.

GAUCH, H.G., AND R.H. WHITTAKER. 1981. Hierarchical classification of community data. Journal of Ecology 69:537–557.

GRACE, J.B. 2001. Difficulties with estimating and interpreting species pools and the implications for understanding patterns of diversity. Folia Geobotanica 36: 71–83.

GUISAN, A., AND N.E. ZIMMERMANN. 2000. Predictive habitat distribution models in ecology. Ecological Modeling 135:147–186.

HILL, M.O., R.G.H. BUNCE, AND M.W. SHAW. 1975. Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in Scotland. Journal of Ecology 63:597–613.

HUBALEK, Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biological Reviews 57:669–689.

JACKSON, D.A., K.M. SOMERS, AND H.H. HARVEY. 1989. Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? American Naturalist 133:436–453.

JANSON, S., AND J. VEGELIUS. 1981. Measures of ecological association. Oecologia 49:371–376.

JANSSEN, J.G.M. 1975. A simple clustering procedure for preliminary classification of very large sets of phytosociological results. Vegetatio 30:67–71.

KAUFMAN, L., AND P.J. ROUSSEEUW. 1990. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York. 342 pp.

KRUSKAL, J.B. 1964. Non-metric multidimensional scaling: a numerical method. Psychometrika 29:115–129.

LEE, B., AND C. MCDONALD. 1993. Comparing three classification strategies for use in ecology. Journal of Vegetation Science 4:341–348.

LEGENDRE, P., AND L. LEGENDRE. 1998. Numerical ecology. 2nd English edition. Elsevier Science B.V., Amsterdam, The Netherlands. 853 pp.

LUDWIG, J.A., AND J.F. REYNOLDS. 1988. Statistical ecology: a primer on methods and computing. John Wiley & Sons, New York. 337 pp.

MATTHEWS, J.A. 1978. An application of non-metric multidimensional scaling to the construction of an improved species plexus. Journal of Ecology 66:157–173.

MCARTHUR, E.D., AND S.C. SANDERSON. 1992a. A comparison between xeroriparian and upland vegetation of Beaver Dam Slope, Utah, as desert tortoise habitat. Pages 25–31 *in* W.P. Clary, E.D. McArthur, D. Bedunah, and C.L. Wamboldt, compilers, Proceedings—symposium on ecology and management of riparian shrub communities. Gen Tech. Rep. INT-289, U.S. Department of Agriculture, USDA Forest Service Intermountain Research Station, Ogden, UT. 232 pp.

_____. 1992b. Great Sand Dunes National Monument vegetation patterns. Pages 185–189 *in* University of Wyoming National Park Service Research Center, 15th annual report, 1991.

MCCUNE, B., AND J.B. GRACE. 2002. Analysis of ecological communities. MjM Software Design, Gleneden Beach, OR. 300 pp.

MCINTOSH, R.P. 1978. Matrix and plexus techniques. Pages 151–184 *in* R.H. Whittaker, editor, Ordination of plant communities. Junk, The Hague.

MWASUMBI, L.B., N.D. BURGESS, AND G.P. CLARKE. 1994. Vegetation of Pande and Kiono coastal forests, Tanzania. Vegetatio 113:71–81.

NICHOLLS, K.H., AND C. TUDORANCEA. 2001. Application of fuzzy cluster analysis to Lake Simcoe crustacean zooplankton community structure. Canadian Journal of Fisheries and Aquatic Sciences 58:231–240.

POLLARD, J.H. 1977. A handbook of numerical and statistical techniques. Cambridge University Press. 349 pp.

R DEVELOPMENT CORE TEAM. 2004. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3. Available from: http://www.R-project.org

SAS INSTITUTE. 1989. SAS/STAT user guide, version 6. 4th edition. Volume 1. Cary, NC. 943 pp.

TUELLER, P.T., AND R.E. ECKERT, JR. 1987. Big sagebrush (*Artemisia tridentata vaseyana*) and longleaf snowberry (*Symphoricarpos oreophilus*) plant associations in northeastern Nevada. Great Basin Naturalist 47: 117–131.

TURNER, S.J., A.R. JOHNSON, AND W.G. WHITFORD. 2004. Pairwise species associations in the perennial vegetation of the Northern Chihuahuan Desert. Southwestern Naturalist 49:1–10.

VAN GROENEWOUD, H. 1992. The robustness of correspondence, detrended correspondence, and TWINSPAN analysis. Journal of Vegetation Science 3:239–246.

WELSH, S.L., N.D. ATWOOD, S. GOODRICH, AND L.C. HIGGINS. 1993. A Utah flora. 2nd edition. Brigham Young University, Provo, UT.

WILLIAMS, W.T., AND J.M. LAMBERT. 1961. Multivariate methods in plant ecology. III. Inverse association-analysis. Journal of Ecology 49:717–729.

WILSON, J.B., T.R. PARTRIDGE, AND M.T. SYKES. 1990. The use of the Cole/Hurlbert C_8 association coefficient in inverse ecological classification. Journal of Vegetation Science 1:367–374.

Received 24 April 2005
Accepted 17 March 2006