OCCASIONAL PAPERS

Museum of Texas Tech University

Number 214 15 May 2002

EFFECTS OF SMALL SAMPLE SIZES ON THE SYMMETRY AND RELIABILITY OF DENDROGRAMS

RICHARD E. STRAUSS, ROBERT D. BRADLEY, AND ROBERT D. OWEN

Cluster analysis is an exploratory multivariate technique that continues to have wide application in studies of geographic variation and other kinds of intraspecific population structure. It especially can be valuable when used in conjunction with ordination techniques, such as multidimensional scaling and eigenvector-based methods, and with other kinds of tree-building procedures (Kruskal, 1977; Pruzansky et al., 1982; Lessa, 1990).

During a recent study of geographic variation in the rodent *Peromyscus spicilegus* (Bradley et al., 1996), a peculiar effect of sample-size variation on UPGMA dendrograms was noted, as described below. In hindsight, the effect is an obvious result of sampling theory and can easily be demonstrated by simulation. However, because the effect is likely to arise with other kinds of data and tree-building algorithms (Mooers et al., 1995) and has not been investigated seriously in the literature, it is worth reporting in some detail. It is a topic that was intimated by Sneath and Sokal (1973) but has been unmentioned in methodological texts and monographs in systematics and ecology (Pimentel,1979; Gauch, Jr.,1982; Digby and Kempton,1987; Ludwig and Reynolds,1988). It also has been lighty regarded in the wider literature on hierarchical clustering (Van Ryzin, 1977; Milligan and Cooper, 1987; Bock, 1988; Jain and Dubes, 1988; Diday et al., 1988; Fukunaga, 1990; Kaufman and Rousseeuw, 1990; Everitt, 1993; Diday et al., 1994; Arabie and Hubert, 1995; Arabie et al., 1966), in which data to be clustered generally are taken at face value, although probabilistic aspects are often considered in partition clustering methods (Bock, 1989; Flury, 1995, Puzicha et al., 2000). The goal of our study is to demonstrate that small sample sizes may have an adverse affect on the results of cluster analysis. We have used empirical data obtained from a previous study (Bradley et al., 1989) and computor simulations to illustrate the potential pitfalls.

AN EXAMPLE

A set of morphometric measurements, 18 cranial and 5 external, was taken on 356 adult individuals of *Peromyscus spicilegus* from western Mexico. These data were taken from a previous study (Bradley et al., 1989) that examined the systematic status of *P. spicilegus*. All measurements were taken by a single individual (R. D. Bradley); measurements and localities are detailed in Bradley et al. (1989). Samples were small for some localities and a small number of missing data (1.6% of the total data set) was estimated using the iterative maximum-likelihood estimation-maximization (EM) method as described by Little and Rubin (1987) and Strauss and Atanassov (in press). Euclidean distances were then calculated among locality samples based on character means of standardized data (mean = 0 and standard error = 1), as in Bradley et al. (1989). The resulting UPGMA dendrogram (Fig. 1) displays a rather striking pattern: samples having small sample sizes (1–6, in this case) tend to "chain" individually to the outside of the dendrogram. In addition, within well-defined clusters, the samples having the smallest sample sizes often, but not always, chain singly or in pairs to the outside of the clusters. When samples with few observations are omitted and the dendrogram is reconstructed, other small-sized samples that previously had been embedded within clusters sometimes take their place by chaining to the outside.

In the *Peromyscus* study (Bradley et al., 1989), an analogous effect was noted in principal-component analyses in which the covariance matrix of sample means were used as observations rather than individuals to avoid biasing the results by localities represented by large sample sizes. In the resulting scatterplots of principal-component scores (Fig. 2), small samples tend to lie to the periphery of clusters and therefore tend to control the spatial positions and orientations of the eigenvectors.

The effects of small sample sizes particularly are not only striking in this example, but can be recognized in more subtle variations in published analyses (e.g., Strauss,1980; Bradley et al., 1989; Carleton and Musser, 1995; Gay and Best, 1995; Reichling, 1995; Teixeira and Musick, 1995; Arroyo-Cabrales and Owen, 1996; Engle and Summers, 1999). In these examples, biological explanations usually were proposed to account for the observed patterns. It is perhaps not surprising that small samples should be subject to significant sampling variation and should fall out as "outliers" on dendrograms. However, we demonstrate by simulation that small samples are capable of affecting the clustering patterns of larger samples as well as the overall tree topology.

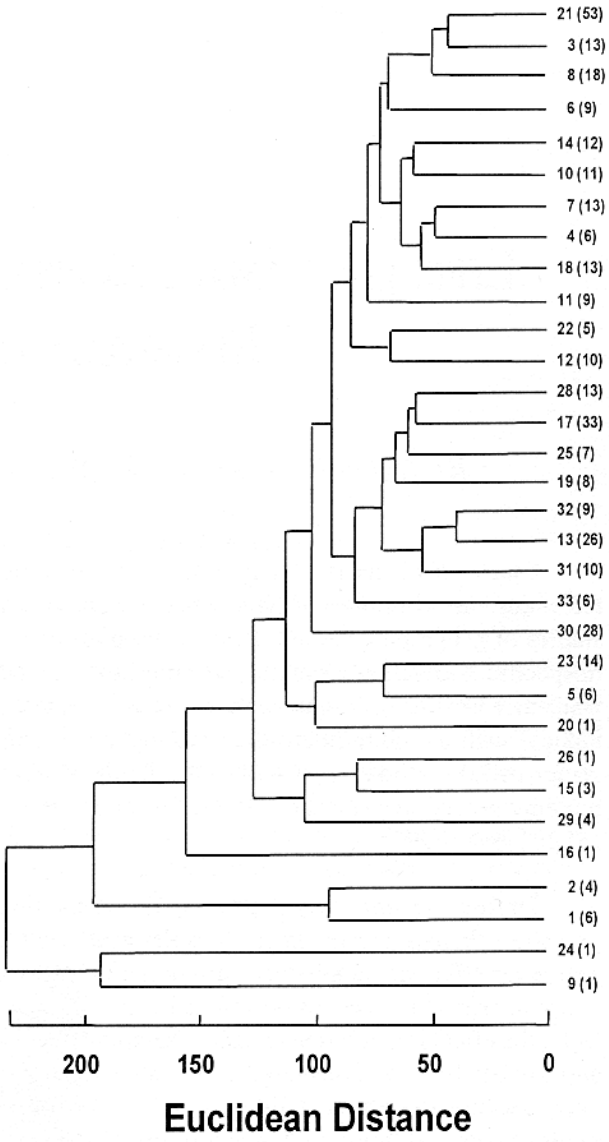

Figure 1. UPGMA dendrogram based on Euclidean distances among 32 geographically distributed samples of *Peromyscus spicilegus*. Distances are based on means of 23 morphometric characters. Numeric labels identify localities; sample sizes are shown in parentheses. Data are from Bradley et al. (1996).

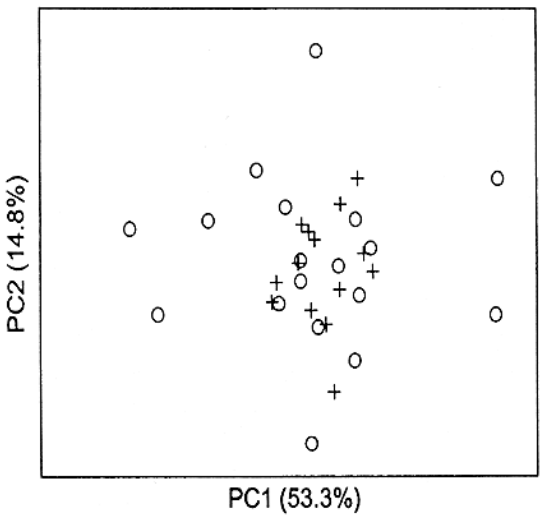

Figure 2. Scattergram of scores of samples of *Peromyscus spicilegus* on the first two principal components of the covariance matrix of sample means. + = samples with ≥10 observations; o = samples with < 10 observations.

SIMULATIONS

Clustering in the Absence of Group Structure

The sample-size effect is most likely to occur in the absence of any other structure in the data, biological or otherwise. To simulate this, we assigned states for five characters and 10 samples by drawing character states randomly from a single, arbitrarily chosen normal distribution, $N(\mu = 10, \sigma = 2)$. Sample sizes for the 10 samples were set at n_i = {1, 1, 2, 2, 5, 5, 50, 50, 60, 60}, and the values for each character were averaged across individuals. This and all other analyses were programmed in Matlab v4.2c (Mathworks, 1997). Typical resulting UPGMA dendrograms of Euclidean distances among sample means are illustrated in Figure 3. There is an obvious and very strong effect of sample size on the clustering structure, with samples having large n_i's clustering together and samples with decreasing n_i's chaining consecutively (not always in strict sample-size sequence) to the main cluster. Repeated simulations varying the numbers of samples, characters per sample, and sample-size distributions display this same effect to varying degrees.

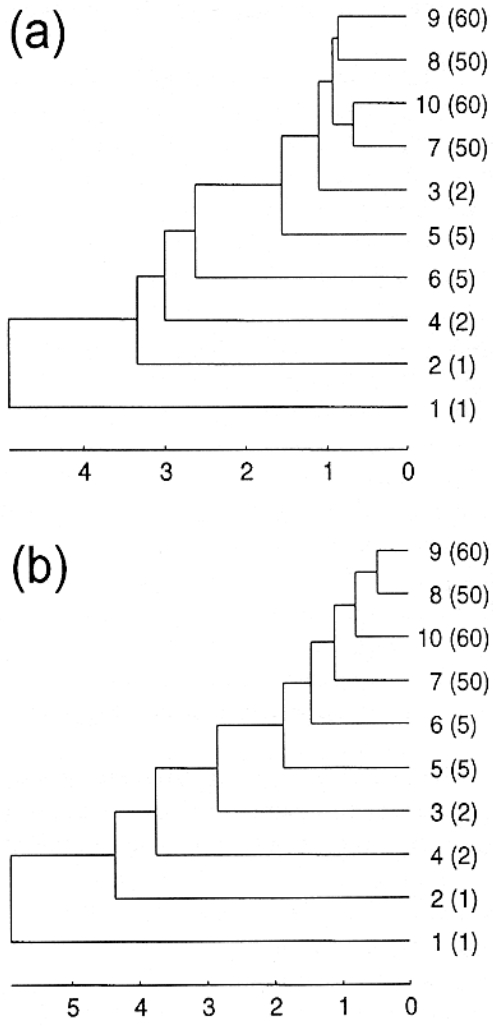

Figure 3. Two replicates (a and b) of randomized cluster analyses of Euclidean distances based on means of five characters for 10 samples. Sample means were calculated from observations drawn from the same normal distribution, N (μ = 10, σ = 2). Sample sizes are shown in parenthesis.

Clustering in the Presence of Group Structure

To simulate biological structure, the 10 samples were divided into two groups (perhaps representing two geographical or ecological domains), the character states from each group being sampled from normal distributions identical in variance but differing by 1.5σ in their means: $N(10,2)$ and $N(13,2)$. Sample sizes for the five samples within each group were set at $n_i = \{1, 2, 5, 50, 60\}$ and, as before, UPGMA dendrograms were computed from Euclidean distances among sample means. It is to be expected that the dendrograms should display two main clusters (Fig. 4) because of the structure built into the sampling regime. However, because the samples within each of the clusters differ only randomly, there is an observable effect of sample-size variation within clusters: the samples having the largest sample sizes group together, with successively smaller samples chaining consecutively. Because character variation is random in this model, the smaller samples occasionally are sufficiently similar to form secondary clusters (Fig. 4*a*).

The presence of group structure was extended by considering 15 samples in three groups, with group means assigned randomly from a uniform distribution over the interval [10, 15]. Character states for five characters were selected from normal distributions having a constant and centered on the group means. Two typical examples are shown in Figure 5. In the first example (Fig. 5*a*), groups A ($\bar{x}_A = 12.6$) and B ($\bar{x}_B = 14.2$) are slightly more similar to one another than either is to group C ($\bar{x}_C = 10.2$), and the dendrogram reflects this by clustering A and B samples which have the largest sample sizes. Some of the smaller A and B samples then chain to this cluster. The two largest C samples also cluster together, followed by two of the smaller C samples. These clusters join and the smallest samples in the data set then chain to the outside.

In the second example (Fig. 5*b*), groups A ($\bar{x}_A = 14.7$) and B ($\bar{x}_B = 14.2$) are very different from group C ($\bar{x}_C = 12.6$). The largest samples of A and B cluster together first, and the smaller A and B samples chain consecutively to this central cord. Group C comprises its own cluster, with the two largest samples joining first. This basic pattern is disturbed only by two small samples of B and C, which join the opposing clusters as outliers.

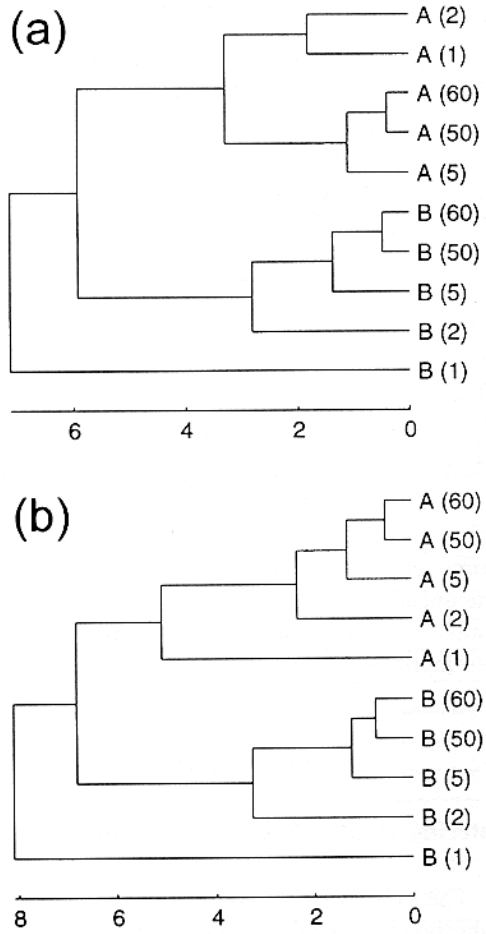

Figure 4. Two replicates (a and b) of randomized cluster analyses of Euclidean distances based on means of five characters for two groups of five samples each. Within each

Systematic Changes in Tree Structure due to Sample-size Effects

These simple examples illustrate the two effects of small samples: the alteration of similarity relationships, and the tendency to increase the degree of asymmetry or imbalance of the dendrogram by consecutively chaining small samples to clusters formed by larger ones. The magnitudes of these effects, and their dependence on number of characters and varying degrees of relationship among samples, were studied with the following simulation design. "True" mean character values for each of 10 samples ($\mu_i, i = 1,\ldots,10$) were assigned randomly from a uni-

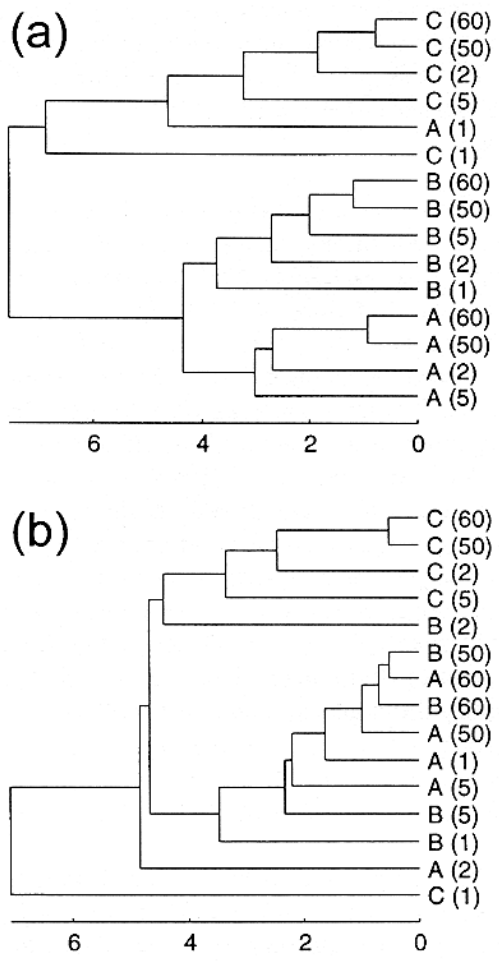

Figure 5. Two replicates (a and b) of randomized cluster analyses of Euclidean distances based on means of five characters for three groups of five samples each. Within each group samples were drawn from the same normal distribution, $N(\bar{X}_i, 2)$, where the three group means $\bar{X}_i$ were drawn from the uniform distribution $U(10, 15)$. Sample sizes are shown in parentheses. (a) $\bar{X}_A = 12.6$, $\bar{X}_B = 14.2$, $\bar{X}_C = 10.2$. (b) $\bar{X}_A = 14.7$, $\bar{X}_B = 14.2$, $\bar{X}_C = 12.6$.

form distribution over the interval [10, 20]; the resulting group differences served as proxies for real biological differences among taxa. Because the sampling error of the mean is inversely proportional to $\sqrt{n}$, a distribution of sample sizes among samples was created by sampling $\sqrt{n_i}, i = 1, \ldots, 10$ from a uniform distribution $U[0, \sqrt{50}]$, squaring the resulting values and rounding them upward to the nearest integers, which thus varied from 1–50 according to a square-root distribution. Heterogeneity among sample sizes was measured as var $(\sqrt{n_i})$. Mean character-state values $\bar{x}_i$ for each sample were then estimated by sampling (with the corresponding sample size) each character from a normal distribution $N(\mu_i, \sigma = 2)$ and averaging across observations; the same distribution was used for all characters of each sample.

From these values two UPGMA dendrograms were produced, T_t based on the "true" means (which, of course, are unknown in actual studies) and T_e on the estimated means (Fig. 6). The differences between these trees must be solely a function of variation in sample size among samples and resulting variation in estimated character means. Two measures were used to describe the differences between corresponding trees that were anticipated based on the preceding observations and simulations. The expected increase in tree asymmetry due to the sample-size effect was measured using the standardized version of Colless's I (Colless, 1980, 1995; Shao and Sokal, 1990), a measure of tree asymmetry. The index was computed separately for the two dendrograms (I_t for the "true" tree and I_e for that based on estimates) and the difference expressed as $\Delta I = I_e - I_t$. The degree of concordance in topology between corresponding trees was measured using the "agreement metric" $d(T_t, T_e)$ of Goddard et al. (1994), standardized to vary between 0–1 for 10 samples.

The procedure of assigning and estimating sample means and of producing and comparing the resulting dendrograms was repeated 500 times for a given number of characters, which was varied from 5–30 in increments of 5. One set of results, those based on 25 characters, is shown in Figure 7, and all results are summarized in Table 1. The patterns are striking: as the number of characters increases the sample-size effect becomes greater, both by increasing the degree of asymmetry of the tree and by increasing the difference between the tree based on true means and that based on estimated means. For approximately 15 or more characters, highly significant correlations with the amount of variability among sample sizes are evident, both for asymmetry increase (ΔI) and for topology incongruence (d). As the variance among sample sizes increases, the amount of tree asymmetry tends

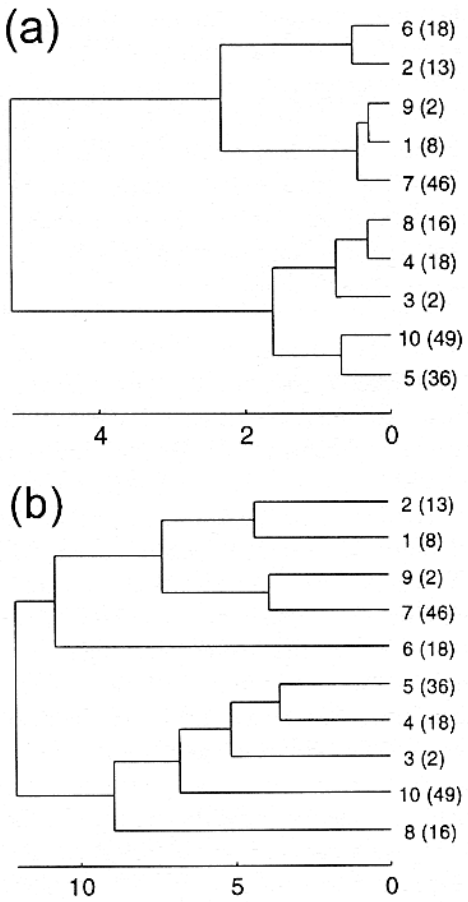

Figure 6. Randomized cluster analysis of Euclidean distances based on means of five characters for 10 samples. (a) Dendrogram based on "true" means μ_i, drawn independently for the 10 samples from the uniform distribution $U(10, 20)$. Sample sizes, although indicated in parentheses, are irrelevant. (b) Corresponding dendogram based on estimated means, drawn separately for each sample from a normal distribution N (μ_i, ó = 2), centered on the sample's true mean. Sample sizes are shown in parentheses.

to increase because of the chaining of small samples, while the level of topological agreement with the "true" tree decreases, primarily due to the chaining of small samples but also due to rearrangements within the main clusters.

These simulations also predict the effect of estimating sample means when all sample sizes are equal, in which case var $\left(\sqrt{n_i}\right) = 0$ and the expected sample size is about 13 ($E\left(\sqrt{n_i}\right) = 12.5$). In Fig. 7*a*, for example, the predicted increase in asymmetry for a balanced design is not significantly different from zero, indicating that the sampling error introduced by estimating group means does not, by itself, lead to a chaining effect. However, a dendrogram produced from estimated means is expected to differ somewhat in topology from that produced from true means, and Fig. 7*b* suggests that the expected level of topological agreement between the two trees is only about $d = 0.75$ even in the absence of sample-size variability, and degrades from there. The value $d = 0.75$ is significantly different from 1.0 at $p < 0.001$.

DISCUSSION

The sample-size effect described here is a result of the well-known Central Limit Theorem, which states that the sampling distribution of any statistic that can be expressed as a function of the sum of n independent observations (which includes the mean as a special case) is asymptotically normal, with an expected standard error of $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, where s is an estimate of the standard deviation of the population. Thus for small samples the sampling variation of the mean can be relatively wide, producing imprecise (though accurate) estimates. An example is shown graphically in Figure 8, which portrays a random sample of observations from a bivariate normal distribution centered on the point (0,0) (Fig. 8*a*), and a corresponding collection of estimates of this true (parametric) mean vector (Fig. 8*b*). The pattern is obvious: means of small samples typically lie relatively far from the population value they represent, while means of large samples tend to be closer to this value. This pattern extrapolates to three or more dimensions, projections of which tend to look like the principal-component space of Figure 2.

Table 1. Asymmetry difference (ΔI) and tree concordance (d) between pairs of "true" and estimated trees, as a function of number of characters. Each row of the table is based on 500 replications.

	ΔI		d	
No. characters	±2SE	Correlation	±2SE	Correlation
5	0.12±0.04	0.04	0.86±0.02	-0.14
10	0.16±0.07	0.07	0.83±0.03	-0.09
15	0.19±0.06	0.26	0.72±0.07	-0.57
20	0.28±0.03	0.42	0.68±0.08	-0.66
25	0.30±0.06	0.53	0.56±0.06	-0.51
30	0.37±0.04	0.51	0.39±0.04	-0.26

[a] Correlation with var $\left(\sqrt{n_i}\right)$

Similar biases may be observed for discrete polymorphic characters as well as continuously varying ones, as well as for other kinds of tree-building methods. For example, Mooers et al. (1995) recently found that "phylogenetic noise" — variation in the states of uninformative (i.e., noisy) cladistic characters — increases the expected magnitude of tree imbalance for cladograms based on maximum parsimony. Thus sampling variation comes in different flavors, but might have similar effects across a wide range of procedures.

Because trees based on large sets of characters show proportionately less uncertainty than trees based on few characters when considering variation *among* characters (Felsenstein, 1985; Sneath, 1986), it might seem counterintuitive that the sample-size effect (which is due to variation within characters) is likely to increase rather than decrease with a larger number of characters. However, the more characters that are sampled, the more likely it is that a small sample will have an outlying mean for at least one of them. In Figure 8*b*, for example, the single-observation sample at the "bottom" of the scatter has a value very near the true mean for character X_1 but is an outlier on X_2. Both of the five-observation samples have means close to the true mean of X_2, but are relative outliers in different directions on X_1.

This effect of character number exists whether or not characters are correlated, although such correlations do reduce the effective number of independent sources of information in the data set. If characters are correlated, circular multivariate distributions become elliptical and it is appropriate to use Mahalanobis rather than Euclidean distances as measures of dissimilarity. However, the principles of sampling underlying the sample-size effect hold for any form of distribution.

These effects are most likely to be encountered in studies of geographic variation of larger animals and plants, because in such studies sizes of locality samples are typically small and dispersed and biological patterns may be weak. However, even in the presence of strong clustering, some secondary effect is likely to be noticed. The overall effect is a tug-of-war between biological signal and statistical artifact. The result depends on the relative strengths of these components in various parts of the tree.

The increase in tree asymmetry is likely to be mitigated by the tendency of UPGMA dendrograms to display only moderate asymmetry (Lance and Williams, 1966). However, this effect must then be compensated for by the alteration of similarity relationships within the tree. This is likely the reason that, in the simulations, the effect of decreasing correlation with "true" structure was greater than the effect of increasing tree asymmetry, although the difference in effects is also undoubtedly a result of differences in sensitivity of the two kinds of tree-comparison metrics. We have yet to study the magnitude of the sample-size effect on the "flexible" clustering algorithm (Lance and Williams,1966; Milligan,1989; Belbin et al., 1992), in which it is possible to explicitly control the degree of tree asymmetry by varying the parameter β.

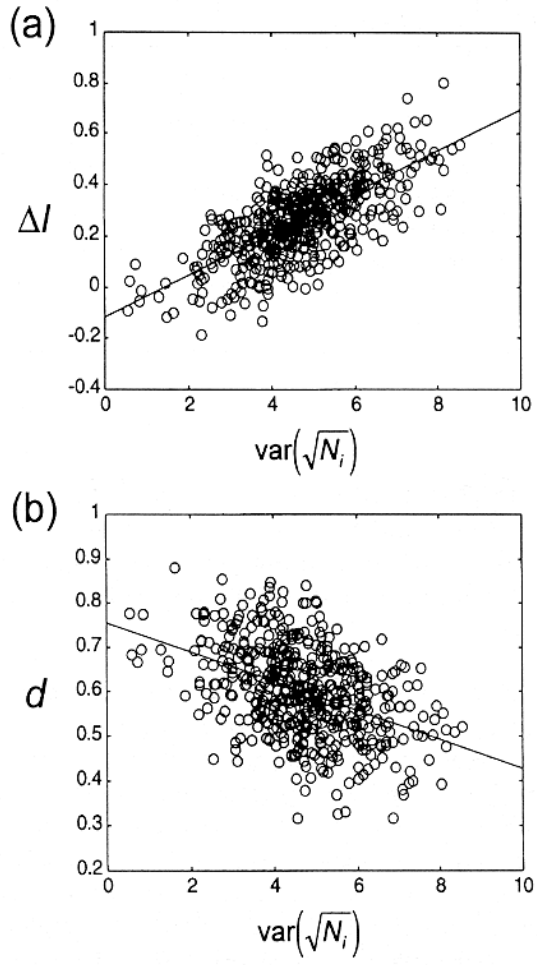

Figure 7. Dependence of asymmetry increase (a) due to sampling effect ΔI, and concordance (b) with the true tree, d, for the case of 25 characters.

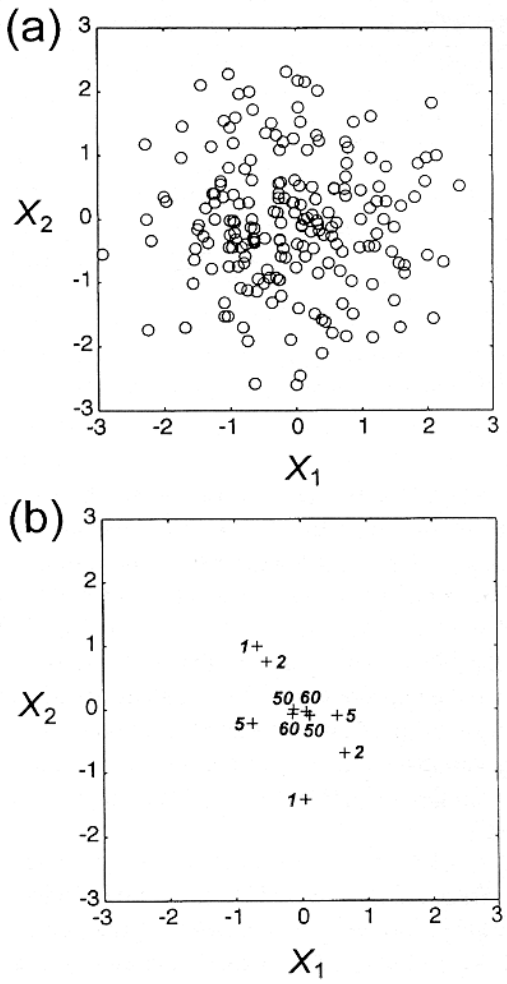

Figure 8. A random sample of (a) 200 observations drawn from a bivariate normal distribution with zero means and unit variances for the two variables X_1 and X_2 and (b) means from the same distribution (corresponding sample sizes are indicated beside each point).

Cluster analyses are designed to produce clusters whether they exist or not. Therefore, a major weakness of cluster analysis as an analytic method is the fact that, except in special cases, it is difficult to assess the statistical "significance" of the results (Sokal,1977; Sneath,1979; Milligan,1981; Bock,1985; Breckenridge,1989). Procedures are available for assessing the significance of the overall clustering structure (Dubes and Jain, 1979; Milligan and Isaac, 1980; Smith and Dubes, 1980; Milligan, 1981, 1996) and, if the sampling distributions are known or can be estimated, it is also possible to randomize the data to obtain confidence intervals around nodal values (Hartigan, 1977; Strauss, 1982; Morey et al., 1983; Murtagh, 1983). It is possible to apply the bootstrap to ultrametric trees (Nemec and Brinkhurst, 1988; Pamilo, 1990) in much the same way that it has been used for additive trees (Felsenstein, 1985; Penny and Penny,

1990), and this should become standard procedure. Alternatively, if one assumes that the data available come from a mixture of several different populations for which the distributions approximate a standard type (e.g., multivariate normal), the clustering problem is then transformed into the problem of estimating, for each of the samples, the parameters of the assumed distribution and the probability that an observation comes from that population (Shieh and Joseph, 1992; Everitt and Dunn, 1992; Banerjee and Rosenfeld, 1993). Strauss (2000) described a similar approach that does not depend on assumptions about underlying structure. These "*k*-means" approaches have the merit of moving the clustering problem away from ad hoc procedures towards the more usual and productive statistical framework of parameter estimation and model testing; they have the disadvantage of not directly providing a hierarchical framework for understanding patterns of biological variation.

All of these alternatives depend critically either upon assumptions about the forms of underlying distributions or on the randomized resampling of empirical distributions. Other reasonable procedures are to carry out analyses both with and without small samples to determine if the basic clustering structure changes, or to superimpose small samples onto the solution provided by larger samples (Perrin et al., 1994). However, the significance of the sample-size effect for very small samples is that it is a problem without a clear solution, except to provoke additional collecting.

ACKNOWLEDGMENTS

We thank B. R. Amman, L. Gordon, J. Juste B., S. Kasper, F. Mendez-Harclerode, and S. Reeder for constructive comments on the manuscript, and M. Houck for valuable discussion.

LITERATURE CITED

Arabie, P., and L. J. Hubert. 1995. Clustering from the perspective of combinatorial data analysis. Pages 1-13 *in* Recent advances in descriptive multivariate analysis (W. J. Krzanowski, ed.), Clarendon Press, Oxford, UK.

Arabie, P., L. J. Hubert, and G. De Soete. 1996. Clustering and classification. World Scientific Publishing, Singapore, 501 pp.

Arroyo-Cabrales, J., and R. D. Owen. 1996. Intraspecific variation and phenetic affinities of *Dermanura hartii*, with reapplication of the specific name *Enchisthenes hartii*. Pages 67-81 *in* Contributions in Mammalogy: A memorial volume honoring Dr. J. Knox Jones, Jr. (H. H. Genoways and R. J. Baker, eds.), Texas Tech University Press, Lubbock.

Banerjee, S., and A. Rosenfeld. 1993. Model-based cluster analysis. Pattern Recognition, 26: 963-974.

Belbin, L., D. P. Faith, and G. W. Milligan. 1992. A comparison of two approaches to beta-flexible clustering. Multivariate Behavioral Research, 27: 417-433.

Bock, H. H. 1985. On some significance tests in cluster analysis. Journal of Classification, 2: 77-107.

Bock, H. H. 1988. Classification and related methods of data analysis. Elsevier Science, Amsterdam, 492 pp.

Bock, H. H. 1989. Probabilistic aspects in cluster analysis. Pages 12-44 *in* Conceptual and numerical analysis of data (O. Opitz, ed.), Springer-Verlag, Berlin.

Bradley, R. D., R. D. Owen, and D. J. Schmidly. 1996. Morphological variation in *Peromyscus spicilegus*. Occasional Papers, Museum of Texas Tech University, 159: 1-23.

Bradley, R. D., D. J. Schmidly, and R. D. Owen. 1989. Variation in the glans penis and bacula among Latin American populations of the *Peromyscus boylii* species complex. Journal of Mammalogy, 70: 712-725.

Breckenridge, J. N. 1989. Replicating cluster analysis: method, consistency, and validity. Multivariate Behavioral Research, 24: 147-161.

Carleton, M. D., and G. G. Musser. 1995. Systematic studies of Oryzomyine rodents (Muridae: Sigmodontinae): Definition and distribution of *Oligoryzomys vegetus* (Bangs, 1902). Proceedings of the Biological Society of Washington, 108: 338-369.

Colless, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: A reappraisal. Systematic Zoology, 29: 288-299.

Colless, D. H. 1995. Relative symmetry of cladograms and phenograms: An experimental study. Systematic Biology, 44: 102-108.

Diday, E., C. Hayashi, M. Jambu, and N. Ohsumi. 1988. Recent developments in clustering and data analysis. Academic Press, New York, 452 pp.

Diday, E., Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy. 1994. New approaches in classification and data analysis. Springer-Verlag, Berlin, 693 pp.

Digby, P. G. N., and R. A. Kempton. 1987. Multivariate analysis of ecological communities. Chapman and Hall, London, 452 pp.

Dubes, R. C., and A. K. Jain. 1979. Validity studies in clustering methodology. Pattern Recognition, 11: 235-254.

Engle, V. D., and J. K. Summers. 1999. Latitudinal gradients in benthic community composition in western Atlantic estuaries. Journal of Biogeography 26: 1007-1023.

Everitt, B. S. 1993. Cluster analysis. Edward Arnold, Sevenoaks, Kent, England, 184 pp.

Everitt, B. S., and G. Dunn. 1992. Applied multivariate data analysis. Oxford University Press, New York, 304 pp.

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

Flury, B. D. 1995. Developments in principal component analysis. Pages 14-33 *in* Recent Advances in Descriptive Multivariate Analysis (W. J. Krzanowski, ed.), Clarendon Press, Oxford, UK.

Fukunaga, K. 1990. Introduction to statistical pattern recognition. Academic Press, New York, 592 pp.

Gauch, H. G., Jr. 1982. Multivariate analysis in community ecology. Cambridge Univ. Press, Cambridge, England, 298 pp.

Gay, S. W., and T. L. Best. 1995. Geographic variation in sexual dimorphism of the puma (*Puma concolor*) in North and South America. Southwestern Naturalist, 40: 148-159.

Goddard, W., E. Kubicka, G. Kubicki, and F. R. McMorris. 1994. The agreement metric for labeled binary trees. Mathematical Biosciences, 123: 215-226.

Hartigan, J. A. 1977. Distribution problems in clustering. Pages 45-71 *in* Classification and clustering (J. Van Ryzin, ed.), Academic Press, New York.

Jain, A. K., and R. C. Dubes. 1988. Algorithms for clustering data. Prentice Hall, Englewood Cliffs, New Jersey, 482 pp.

Kaufman, L., and P. J. Rousseeuw. 1990. Finding groups in data: an introduction to cluster analysis. John Wiley, New York, 342 pp.

Kruskal, J. B. 1977. The relationship between multidimensional scaling and clustering. Pages 17-44 *in* Classification and clustering (J. Van Ryzin, ed.), Academic Press, New York.

Lance, G. N., and W. T. Williams. 1966. A generalized sorting strategy for computer classification. Nature, 212: 218.

Lessa, E. P. 1990. Multidimensional analysis of geographic genetic structure. Systematic Zoology, 39: 242-252.

Little, R. J. A., and D. B. Rubin. 1987. Statistical analysis with missing data. John Wiley and Sons, New York, 267 pp.

Ludwig, J. A., and J. F. Reynolds. 1988. Statistical ecology. John Wiley and Sons, New York, 337 pp.

Mathworks. 1997. Matlab reference guide. Mathworks, Natick, Massachusetts, 403 pp.

Milligan, G. W. 1981. A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 1: 379-407.

Milligan, G. W. 1989. A study of the beta-flexible clustering method. Multivariate Behavioral Research, 24: 163-176.

Milligan, G. W. 1996. Clustering validation: Results and implications for applied analyses. Pages 341-375 *in* Clustering and classification (P. Arabie, L. J. Hubert, and G. De Soete, eds.), World Scientific Press, River Edge, New Jersey.

Milligan, G. W., and M. C. Cooper. 1987. Methodological review: clustering methods. Applied Psychological Measurement, 11: 329-354.

Milligan, G. W., and P. D. Isaac. 1980. The validation of four ultrametric clustering algorithms. Pattern Recognition, 12: 41-50.

Mooers, A. Ø., R. D. M. Page, A. Purvis, and P. H. Harvey. 1995. Phylogenetic noise leads to unbalanced cladistic tree reconstructions. Systematic Biology, 44: 332-342.

Morey, L. C., R. K. Blashfield, and H. A. Skinner. 1983. A comparison of cluster analysis techniques within a sequential validation framework. Multivariate Behavioral Research, 18: 309-329.

Murtagh, F. 1983. A probability theory of hierarchic clustering using random dendrograms. Journal of Statistical Computation and Simulation, 18: 145-157.

Nemec, A. F. L., and R. O. Brinkhurst. 1988. Using the bootstrap to assess statistical significance in the cluster analysis of species abundance data. Canandian Journal of Fisheries and Aquatic Sciences, 45: 965-970.

Pamilo, P. 1990. Statistical tests of phenograms based on genetic distances. Evolution, 44: 689-697.

Penny, D., and P. Penny. 1990. Search, parallelism, comparison, and evaluation: algorithms for evolutionary trees. Pages 75-91 *in* Trees and hierarchical structures, lecture notes in biomathematics 84 (A. Dress and A. von Haeseler, eds.), Springer-Verlag, New York.

Perrin, W. F., G. D. Schnell, D. J. Hough, J. W. Gilpatrick, Jr., and J. V. Kashiwada. 1994. Reexamination of geographic variation in cranial morphology of the pantropical spotted dolphin, *Stenella attenuata*, in the eastern Pacific. Fishery Bulletin 92:324-346.

Pimentel, R. A. 1979. Morphometrics: the multivariate analysis of biological data. Kendall-Hunt, Dubuque, Iowa, 276 pp.

Pruzansky, S., A. Tversky, and J. D. Carroll. 1982. Spatial versus tree representations of proximity data. Psychometrika, 47: 3-24.

Puzicha, J. 2000. A theory of proximity based clustering: Structure detection by optimization. Pattern Recognition, 33: 617-634.

Reichling, S. B. 1995. The taxonomic status of the Louisiana pine snake (*Pituophis melanoleucus ruthveni*) and its relevance to the evolutionary species concept. Journal of Herpetology, 29: 186-198.

Shao, K. T., and R. R. Sokal. 1990. Tree balance. Systematic Zoology, 39: 266-276.

Shieh, D. S. S., and B. Joseph. 1992. Exploratory data analysis using inductive partitioning and regression trees. Industrial Engineering and Chemical Research, 31: 1989-1998.

Smith, S. P., and R. C. Dubes. 1980. Stability of a hierarchical clustering. Pattern Recognition, 12: 177-187.

Sneath, P. H. A. 1979. Basic program for a significance test for clusters in UPGMA dendrograms obtained from squared Euclidean distance. Computers and Geosciences, 5: 127-137.

Sneath, P. H. A. 1986. Estimating uncertainty in evolutionary trees from Manhattan-distance triads. Systematic Zoology, 35: 470-488.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy: The principles and practice of numerical classification. W.H. Freeman, San Francisco, 573 pp.

Sokal, R. R. 1977. Clustering and classification: Background and current directions. Pages 1-15 *in* Classification and clustering (J. Van Ryzin, ed.), Academic Press, New York.

Strauss, R. E. 1980. Genetic and morphometric variation and the systematic relationships of eastern North American sculpins (Pisces: Cottidae). Ph.D. Dissertation, Pennsylvania State University, University Park.

Strauss, R. E. 1982. Statistical significance of species clusters in association analysis. Ecology, 63: 634-639.

Strauss, R. E. 2000. Cluster analysis and the identification of aggregations. Animal Behaviour, 60: 481-488.

Strauss, R. E., and M. N. Atanassov. In press. Evaluation of two methods for estimating missing data in morphometric studies of vertebrate skeletons. Journal of Vertebrate Paleontology.

Teixeira, R. L., and J. A. Musick. 1995. Trophic ecology of two congeneric pipefishes (Syngnathidae) of the lower York River, Virginia. Environmental Biology of Fishes, 43: 95-309.

Van Ryzin, J. 1977. Classification and clustering. Academic Press, New York, 467 pp.

Addresses of authors:

RICHARD E. STRAUSS

Department of Biological Sciences
Texas Tech University
Lubbock, Texas 79409-3131
e-mail: Rich.Strauss@ttacs.ttu.edu

ROBERT D. BRADLEY

Department of Biological Sciences and Museum
Texas Tech University
Lubbock, Texas 79409-3131
e-mail: rbradley@ttacs.ttu.edu

ROBERT D. OWEN

Department of Biological Sciences
Texas Tech University
Lubbock, Texas 79409-3131
e-mail: rowen@ttu.edu

PUBLICATIONS OF THE MUSEUM OF TEXAS TECH UNIVERSITY

It was through the efforts of Horn Professor J Knox Jones, as director of Academic Publications, that Texas Tech University initiated several publications series including the Occasional Papers of the Museum. This and future editions in the series are a memorial to his dedication to excellence in academic publications. Professor Jones enjoyed editing scientific publications and served the scientific community as an editor for the Journal of Mammalogy, Evolution, The Texas Journal of Science, Occasional Papers of the Museum, and Special Publications of the Museum. It is with special fondness that we remember Dr. J Knox Jones.

Institutional subscriptions are available through the Museum of Texas Tech University, attn: NSRL Publications Secretary, Box 43191, Lubbock, TX 79409-3191. Individuals may also purchase separate numbers of the Occasional Papers directly from the Museum of Texas Tech University.

ISSN 0149-175X

Museum of Texas Tech University, Lubbock, TX 79409-3191