Journal of Classification 23: 3-30 (2006)
DOI: 10.1007/s00357-006-0002-6

SPECIAL INVITED ARTICLE

The Practice of Cluster Analysis

Jon R. Kettenring

Drew University, Madison, New Jersey

Abstract*:* Cluster analysis is one of the main methodologies for analyzing multivariate data. Its use is widespread and growing rapidly. The goal of this article is to document this growth, characterize current usage, illustrate the breadth of applications via examples, highlight both good and risky practices, and suggest some research priorities.

Keywords: Principal components analysis; Discriminant analysis; Usage trends; Applications; Process improvement; Research needs.

1. Introduction

The desire to organize data into homogeneous groups is common and natural. The results can provide either immediate insights or a foundation upon which to construct other analyses. This is what cluster analysis is all about – finding useful groupings that are tightly knit (in a statistical sense) and distinct (preferably) from each other. Sometimes the distinction is made (see, e.g., Hand, Mannila, and Smyth 2001) between such naturally occurring statistical clusters and other groupings that are obtained for

My thanks to Phipps Arabie for providing reference materials, Ram Gnanadesikan for suggestions and years of collaboration on cluster analysis research, Jim Miller for schooling in chromatography, and especially to all the authors of the ten featured examples for contributing their stimulating applications to the literature.

Author's Address: The Charles A. Dana Research Institute for Scientists Emeriti, Drew University, 36 Madison Avenue, Madison, NJ 07940, USA; e-mail: jkettenr@drew.edu

convenience, as in a partitioning a homogeneous set of data into contiguous pieces. The primary focus here is on the former.

Because of its utility, clustering has emerged as one of the leading methods of multivariate analysis. Early books on the subject (e.g., Anderson 1958; Rao 1965) did not treat clustering at all, but for at least twenty five years now it has been considered mainstream (e.g., Gnanadesikan 1977; 1997; Seber 1984; Johnson and Wichern 1982, 2002). Excellent specialty books on clustering are also easy to find (e.g., Hartigan 1975; Everitt, Landau, and Leese 2001; Kaufman and Rousseeuw 1990; Gordon 1999). Moreover, many multivariate books aimed at specific areas of application cover clustering as well: e.g., Legendre and Legendre (1998) and Shaw (2003) in ecology; Brown (1998) in geohydrology; and Baxter (2003) in archaeology. Hastie, Tibshirani, and Friedman (2001) treat cluster analysis in the context of unsupervised learning and data mining.

Cluster analysis (CA), principal components analysis (PCA), and discriminant analysis (DA) are three of the primary methods of modern multivariate analysis. In 2003, there were over 1,100 papers published involving CA and over 1,600 utilizing PCA. About 700 dealt with DA. PCA and DA are, of course, longstanding core parts of the field. Their statistical theories are rich and supported by clear and crisp mathematics. The same can't be said for CA, notwithstanding many noble and important efforts.

One way to contrast CA and DA is to think of them as at the opposite ends of a spectrum. At the CA end there is no information at all about the number of groups or their content. At the DA end, both the number of groups and their content are known. The focus is on characterizing group differences and assigning "unknowns" to one of the known groups. In the real world, there are many occasions where the actual problem lies somewhere along this spectrum rather than exactly at either end. Perhaps there is information about the number of groups (clusters) or group membership. Or maybe there is uncertainty about the accuracy of pre-assigned group labels. In other words, there is really a continuum of problems to consider, only some of which have been fully explored.

Sometimes PCA is used as a method to find clusters directly, bypassing any of the usual CA algorithms. The logic for doing so is fuzzy but goes something like this: PCA is a vehicle for reducing dimensionality and visualizing data in a reduced number of dimensions corresponding to the leading PCs. Insofar as these PCs—and more often than not the number is taken to be two or three—capture the directions of greatest variability in the data, and this variability is largely "between" as opposed to "within group" in nature, one may be able to get away with this crude approach to clustering. However, pitfalls abound (see, e.g., Chang 1983).

In fact, there is no easy and rigorous way to quickly extract clusters from complex data. In particular, there is no straightforward eigenanalysis that can be counted on to reveal cluster structure in the same manner that such an approach yields directions of greatest variability in PCA and greatest group separation in DA. This is one reason that literally hundreds of different algorithms have been proposed to get the job done. Each has its own pluses and minuses. Care needs to be exercised at all phases: the form in which the data are analyzed, the choice of algorithm and any associated parameters, and the manner in which outputs are checked for validity.

This challenging situation is one of the motivations behind this paper. Effective clustering is very much an imprecise art. With usage proliferating, making sure that it is practiced effectively is a more important objective than ever.

2. Usage Trends

To assess usage trends in CA, extensive use was made of three Web of Science® databases: the Science Citation Index Expanded^{TM}, the Social Sciences Citation Index®, and the Arts & Humanities Citation Index®.[1] These databases index, respectively, 5,900 scientific journals in 150 scientific disciplines, 150 journals across 50 social sciences disciplines, and over 1,100 arts and humanities journals. The databases can be searched on a yearly basis, where "year" refers to the year in which an entry was made to the database. This is most often the year of publication. The total number of records in the three databases ranges from roughly one million in 1995 to 1.3 million in 2003.

The period, 1995-2003, was chosen in order to concentrate on the last decade and those years for which full results were available at the time the study began. As 2004 results became available, they were added in selectively.

Titles, key words, and abstracts of documents in the three databases were searched for the phrase, cluster analysis. It should be noted that the AHCI did not contain searchable abstracts until 2000, but it also turned up only thirteen distinct hits during the entire period of the search.

Searching on "cluster analysis" by itself, and not other names by which CA often goes by, was a conscious decision to minimize "false positive" counts of papers that might be talking about clusters or clustering in a

[1] The Web of Science, the Social Sciences Citation Index, and the Arts & Humanities Citation Index are registered trademarks and the Science Citation Index Expanded is a trademark of The Thomson Corporation.

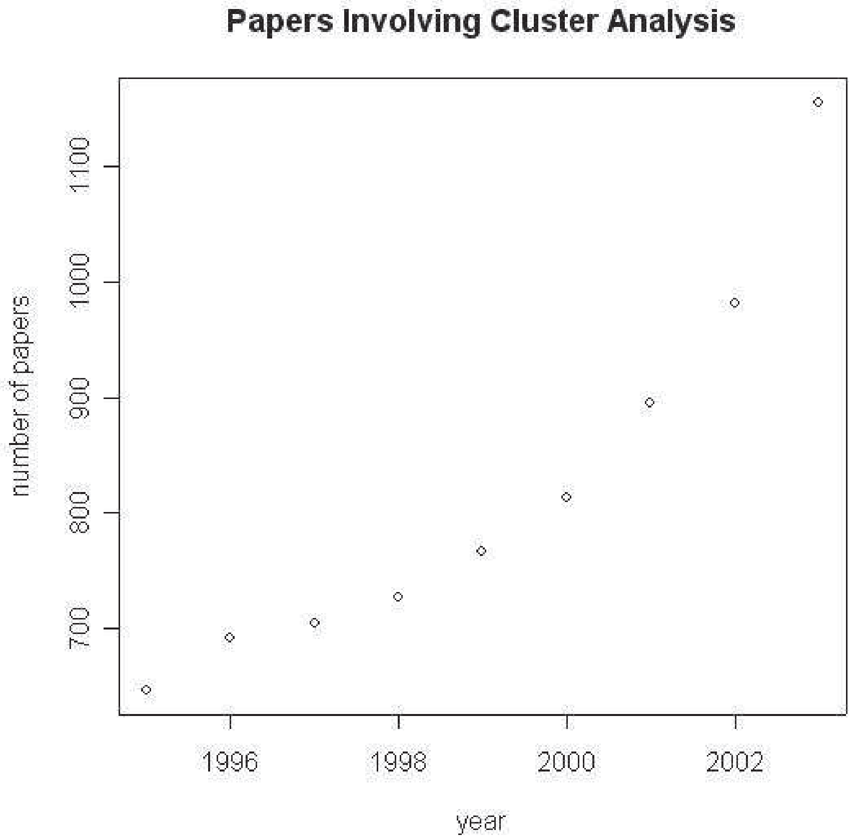

Figure 1. Number of papers involving cluster analysis, 1995-2003.

nontechnical or nonstatistical sense (more on this later).

The search results are shown in Figure 1. The trend is clearly upward and accelerating, with counts ranging from 646 in 1995 to 1,156 in 2003. In Figure 2, the raw counts are normalized by the number of records in each year. The pattern is similar but the growth rate is sharper in the more recent years. (Later evidence for 2004 suggests that the growth may have abated: the raw CA count dropped to 1,118 while the ratio of CA records to the total number of records increased less than one percent.) During this same period, 1995-2003, the count for DA increased as well, but at a slower rate, from 531 to 698. The numbers for PCA (with or without the “s” on “component”) more than doubled from 743 to 1,621.

As mentioned, restricting the CA search to “cluster analysis” paints a very conservative picture of the activity level. To illustrate, during the 1995-

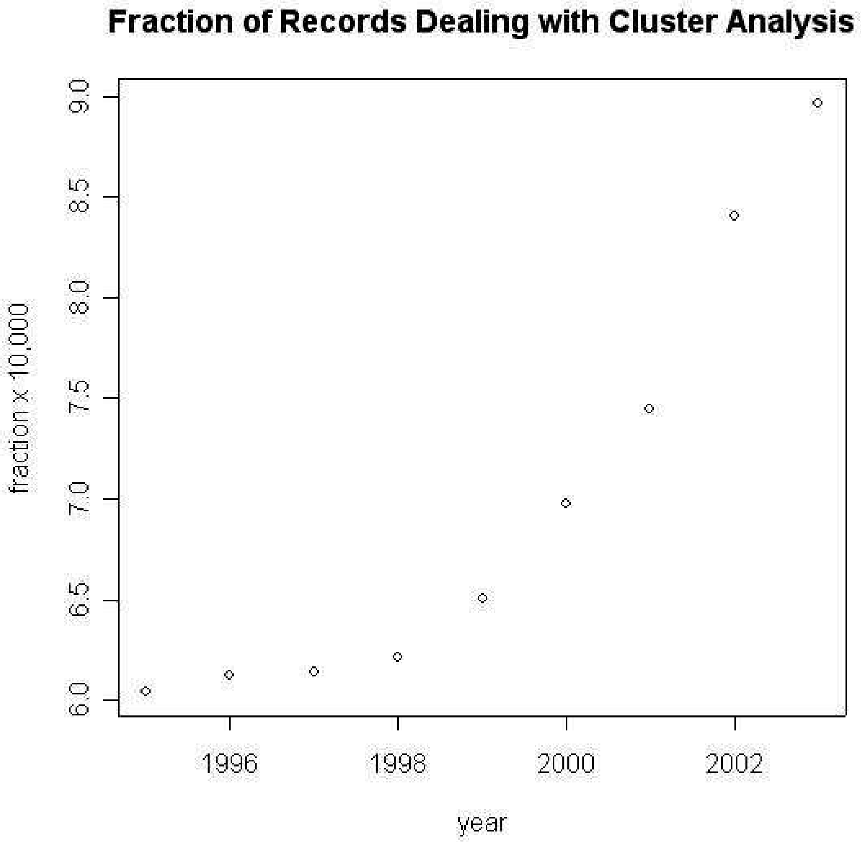

Figure 2. Ratio of number of articles involving cluster analysis to total number in databases, multiplied by 10,000, for 1995-2003.

2003 period, there were 7,380 CA papers picked up by this search. Adding additional terms in succession via an OR-operation pushes the total much higher: “hierarchical clustering” (8,131), “dendrogram” (8,923), “numerical taxonomy” (9,156), and “unsupervised learning” (9,720). However one chooses to look at the data, it appears that there are at present well over 1,000 papers appearing annually for which CA has a prominent role.

The 1,156 records for 2003 were broken down by field, as defined by the Web of Science databases. Fifty percent of them fell into the ten disciplinary categories shown in Figure 3. All are part of the life sciences, interpreted broadly. In fact, the pattern is similar for all the years studied (including 2004) with plant sciences the top runner in each case representing between 8.7 and 12.4% of the CA papers.

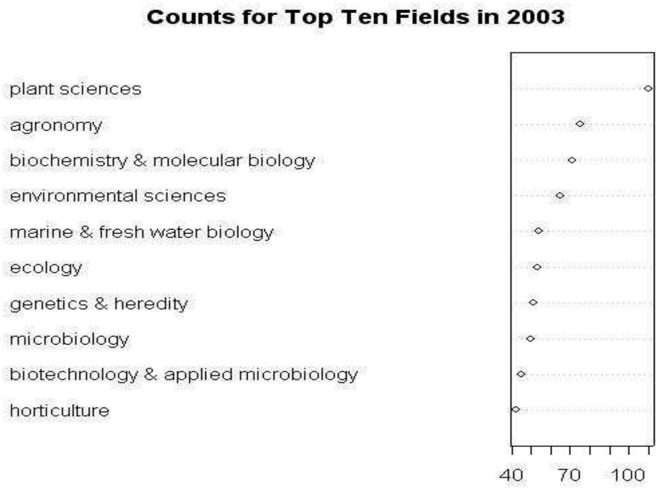

Figure 3. Top ten areas of application, 2003.

To probe deeper into the nature of the papers, a random sample of 100 of the 1,156 records from 2003 was drawn and their abstracts reviewed in detail. Five of the sampled papers were on general aspects of CA methodology. Another five appeared to be false positives in that they were either not full papers or their use of CA was suspect. Yet another eight involved a tight coupling of methodology development and application. Counting this latter group as only "half application", there were 14 records in the sample of 100 that were not applications of CA. Using 14%, then, as the estimate of the fraction of non-applications papers, and adjusting the 2003 total count of 1,156, there are still roughly 1,000 applications papers for this year as picked up by the search on "cluster analysis" alone.

In 24 cases the type of CA employed was clear. In 17 of them, a form of hierarchical cluster analysis (HCA) was chosen. Five relied on k-means CA, and one involved both methods. No other types were mentioned. In addition, 23 employed both CA and PCA in the analysis, affirming the close allegiance of these two methods.

Many of the papers in the sample were identified for study of the full text. The patterns observed in the just cited summary statistics, such as heavy reliance on HCA, were confirmed in the collection of complete papers.

3. Examples

Drawing from the roughly 2,000 papers from 2003 and 2004 that were screened, ten were chosen to illustrate the range of current applications and practices found in the literature. (This is by no means a complete characterization!)

Example 1 (aquatic plants; Kim, Shin, and Choi 2003). This is a typical, modest-sized application of HCA and PCA. The study involves taxonomic relationships among 77 aquatic plants coming from four "currently recognized species." Twenty five quantitative variables—various length, width, thickness, angular, gap, and ratio measurements—were used to characterize the plants. Apparently each was standardized to zero mean and unit variance as a preprocessing step. (Standardizing variables as a prelude to CA is often called *autoscaling*.) The HCA was carried out using the rule that intercluster distance is the average of the pairwise (Euclidean) distances between plants in one group and those in the other. Sneath and Sokal (1973, p. 230) describe this average-distance approach to HCA as "probably the most frequently used clustering strategy" and evidence from the current literature would seem to confirm that this is still the case.

The four plant species appear as four fairly distinct clusters in a scatter plot of the first two PCs and also as four major branches arising at different levels in the dendrogram from the HCA. Thus the authors used their knowledge of the existing species to decide how to extract a partition from the hierarchical representation, rather than the usual practice of making a straight line cut of the tree. In that sense, this is not a "zero knowledge" clustering application.

The primary finding from this research was the matching of the data-based clusters with the currently recognized species, which helped clarify their taxonomic relationships.

Example 2 (disease co-occurrence; John, Kerby, and Hennessy 2003). The authors develop a fresh approach to identifying patterns of co-morbidity that can be used to predict adverse health outcomes. Their study is based on a random sample of over 1,000 rural American Indian elders aged 60 or over from one tribe. Information was collected on the presence or absence of 11 chronic conditions, the variables in the study. Relations among the conditions were studied via several approaches. These included HCA of the conditions based on their pairwise correlations. In the process, different measures of correlation and different inter-cluster similarity rules were employed. Four clusters emerged consistently. The largest of them was labeled *cardiopulmonary*. Various statistical models, some involving the clusters, others not, and all including gender, age, educational attainment, and marital status, were tested for their ability to explain four health outcome indices. Bottom line: CA improved the modeling effort and "identifie[d] specific health problems that have to be addressed to alter American Indian elders' health-related quality of life." Moreover, compared to other approaches considered, CA "appears to better target particular health problems for prevention or remediation."

Example 3 (cultures; Allik and McCrae 2004). This study considers the geographic distribution of personality traits across a broad range of 36 cultures from around the world. The analysis was based on 28,000 responses to a 240 item questionnaire. The questions pertain to 30 specific traits, referred to as *facets*, of personality. The facets in turn define five basic factors—neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. The factor structure has been replicated in the different cultures and is "well suited to an investigation of personality and geography." Scores on each facet were adjusted to avoid confounding of age and sex with culture. Several other issues of within-culture variability were also addressed.

HCA, using Ward's method for minimizing the within cluster error sum of squares, was used "to summarize similarities between cultures across a range of variables." The 30 facet scores were first standardized across the 36 cultures prior to computing inter-culture Euclidean distances for the CA. The analysis was repeated using the five factors in place of the 30 facets and choosing different distance metrics in place of Euclidean distance. All approaches yielded "similar solutions."

The dendrogram from the 30 facet solution is shown in Figure 4. The authors interpret the entire structure at different levels. Most of the early joins make intuitive geographic sense, e.g., Austrians, Germans, and German-speaking Swiss are merged together. At a slightly larger linkage distance, there is a branch containing Canadians, Americans, and Turks, which is harder to explain. Notwithstanding such specific anomalies, the authors conclude that "cluster analysis showed that geographically proximate cultures often have similar profiles."

Example 4 (media usage; van Rees and van Eijck 2003). The authors point out that in the West, "the majority of the population spends more leisure time on media-related activities than on any alternative leisure pursuit." Hence the purpose of their study: "to gain greater insight into the nature of media [usage patterns] and their corresponding audiences."

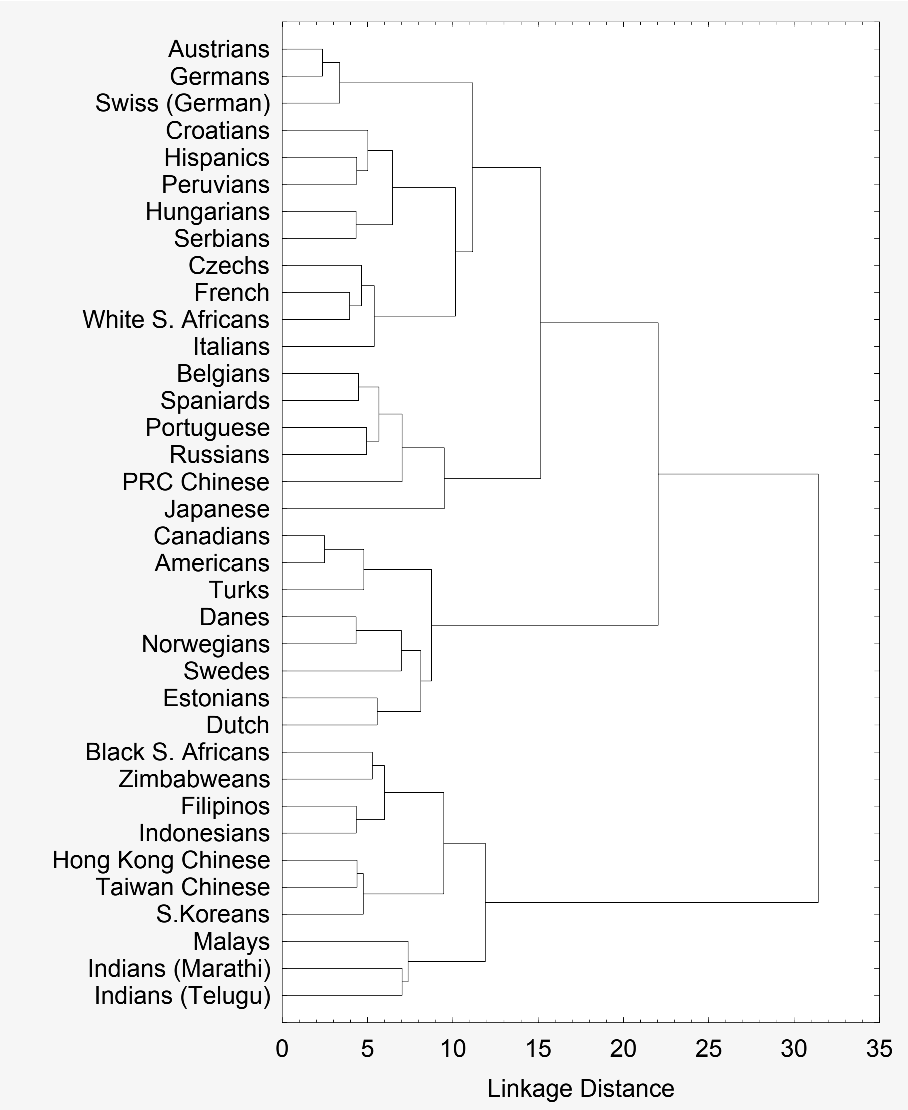

Figure 4. Dendrogram of clusters of cultures.[2]

[2] Reproduced from Allik and McCrae, Journal of Cross-Cultural Psychology (35/1) pp. 1-28, copyright 2004 by Sage Publications, Reprinted by Permission of Sage Publications Inc.

In particular, van Rees and van Eijck study the use of 19 kinds of media, including the Internet, in the Dutch population. A diary of time budget data was compiled for a sample of nearly 1,800 subjects. For each one, a score was developed representing the time spent on each media category over the course of a week. The authors applied Ward's method of HCA to the 19 media types. The resulting nested structure lent itself to natural interpretations of how the media audience split itself. For example, one half of it corresponds to subjects who prefer the Internet, quality newspapers, videos, etc. while the other half is connected with those who prefer regional newspapers, radio and television, etc. The authors go on to interpret the finer grained structure, moving up from the trunk of the tree.

Example 5 (juvenile offenders; Stefurak, Calhoun, and Glaser 2004). This is a study of a sample of 103 male juvenile offenders. It's based on a well-developed 165-item adolescent assessment instrument that measures 12 personality traits, as well as other scales of interest. A separate instrument was used as well "as a source of external validity of groupings derived from a cluster analysis of [the 12] Personality Patterns scales." Ward's method of HCA was applied, as in Example 4, and four clusters of juveniles were identified (presumably by cutting the hierarchical tree but details are not given) because they "yielded an optimal balance between within cluster homogeneity and between-cluster heterogeneity." Each cluster was assigned a name to underscore its perceived clinical relevance. The largest one, *the reactive depressives*, "suggests the importance of considering the role of internalizing problems as a conduit to delinquency in addition to antisocial personality."

Example 6 (chromatography; Le Mapihan, Vial, and Jardy 2004). Chromatography is a physical method used to separate complex mixtures such as pharmaceutical products into their basic components. The separation occurs inside a column that is either packed or coated in a special manner. The mixture to be separated is passed through the column with the aid of a moving gas or liquid solvent. The components of the mixture are separated in the process and emerge at different times called retention times. These data from the output of a detector downstream from the column are used to analyze the mixture. (See Miller 2004 for more background.)

Numerous approaches have been developed for preparing the columns. The paper by Le Mapihan et al. investigates ways of characterizing "columns representative of those commonly used in the pharmaceutical industry." They consider 12 column types and two different solvents for a total of 24 column-solvent combinations, which they refer to as the “individuals” in their analysis.

The data consist of a matrix of 210 chromatographic variables by 24 individuals. After autoscaling, extensive use was made of PCA to reduce dimensionality. Euclidean distances among the individuals were computed from “autoscaled PC scores” and fed into a HCA (centroid method, Euclidean distance) to find clusters. The goal was “to provide interpreted classifications and to reduce drastically the test [time and effort] by eliminating redundant information.”

In one featured case, the dendrogram was cut to yield a partition of nine clusters, and these, along with their subclusters, were overlaid on the scatter plot of the first two PCs to aid interpretation. Various other scenarios were considered, which in statistical terms amount to setting aside subsets of the variables in search of a minimal set that would replicate the information in the full PC-HCA analysis.

Le Mapihan et al. conclude that the “combination of PCA and HCA proved to be an invaluable asset both for understanding classifications and selecting objectively the best [chromatographic test] conditions.” By optimizing the conditions, the amount of test time needed was reduced by a factor of six or down to roughly one day.

Example 7 (archaeology; Hall 2004). Multivariate methods, including CA, have been widely applied to the analysis of archaeological data. In this paper Hall studies data on the geochemical characteristics of 175 sherds of a special type of pottery from six sites in the Tokyo Bay region of Japan. The motivation for using CA was to determine if sherds from different sites belong to the same clusters. This would suggest that “potters utilized raw materials that were geochemically similar, and prepared the paste in similar fashion” and “people were moving pottery between sites.”

The data consist of 16 minor and trace element composition measurements for each sherd. All of the measurements were transformed to a log scale. Hall points out that it is common in pottery studies to assume that such transformations normalize geochemical data, but there are no guarantees. He then uses PCA to reduce the number of variables to five, accounting for 92% of the total variation in the PCs. Thirteen outliers were identified via box plots of the PCs and (apparently) set aside for the CA. Hall applied the multivariate normal model-based clustering methodology of Banfield and Raftery (1993), which has the appealing feature of allowing for different within cluster covariance matrices. In this case a two-cluster model with equal covariance matrices was selected. Hall interpreted the clusters in terms of the six sites from which the sherds originated and noted that they "correspond to the local sedimentary raw materials near [them]." He goes on to observe several types of "significant differences" between the clusters using standard statistical techniques. As additional support for the two-cluster solution, he returns to the 16 log transformed variables and applies DA with cross-validation to correctly classify 97% of the 162 non-outlier cases.

Example 8 (stem cells; Sperger, Chen, Draper, Antosiewicz, Chon, Jones, Brooks, Andrews, Brown, and Thomson 2003). This is one of numerous examples of the use of CA to study gene expression data. The infusion of such papers probably accounts for much of the growth in usage of CA described in Section 2. Several articles provide excellent overviews of the area, e.g., Jiang, Tang, and Zhang (2004), Domany (2003), and Sebastiani, Gussoni, Kohane, and Ramoni (2003).

Sperger et al. compare the gene expression profiles of human embryonic stem cells using average similarity HCA and Pearson correlation coefficients as described in Eisen, Spellman, Brown, and Botstein (1998). Among other findings, the HCA "showed that the five independent cell lines clustered tightly together, reflecting highly similar expression profiles." Detailed interpretations of other meaningful expression patterns in the dendrogram are also provided. Nearly 10,000 profiles were involved in the analysis. The results are presented in a heat map (also called Eisen plot or clustergram), which provides a striking visual display of color-coded summaries of the data juxtaposed against the dendrogram. It effectively illuminates the clustered data in spite of its volume.

Example 9 (purebred domestic dogs; Parker, Kim, Sutter, Carlson, Lorentzen, Malek, Johnson, DeFrance, Ostrander, and Kruglyak 2004). This work explores genetic relationships among 85 domestic dog breeds, using four or five unrelated dogs per breed. A Bayesian clustering algorithm, *structure* (Falush, Stephens, and Pritchard 2003), tailored to inferring population structure from genotype data, was used to test the hypothesis that "breed membership could be determined from individual dog genotypes." Numerous analyses of subsets of 20 to 22 breeds at a time produced clusters of dogs that most often all came from the same breed. Six closely related breeds tended to cluster together in pairs, for instance, Alaskan Malamute and Siberian Husky.

Several more analyses were done to substantiate the genetic distinctiveness and to explore the relationships of the breeds. Experiments were run with varying values of k, the assumed number of sub-populations. An especially attractive summary is the use of colored profile plots to indicate an individual dog's estimated proportion of membership in a cluster, based on the model.

The authors conclude that their work both supports traditional groupings and reveals new ones. It helps lay "the foundation for studies aimed at uncovering the complex genetic basis of breed differences in morphology, behavior, and disease susceptibility."

Example 10 (remote sensing; Braverman, Fetzer, Eldering, Nittel, and Leung 2003). This problem differs from the others in several respects. First, the data—sensing measurements from NASA's Earth Observation System—are truly massive in nature. Second, the data arrive in streams giving rise to the challenge of summarizing them without sacrificing their distributional character. A third difference is that CA is used purely for data reduction rather than to yield directly interpretable clusters. Indeed, the end result could be a mixture of "natural" and "convenience" clusters as discussed in Section 1.

The approach is to break the stream of data down into chunks of only a few days worth so that it all can be stored in memory and summarized. The summary is based on a k-means CA of each chunk on a cell-by-cell basis within a spatial grid. For each cell, the algorithm is run with a fixed value for k and a number of random starting points (20 and 30 in a case presented). The run yielding cluster centroids with the minimum mean squared error with respect to the original data is chosen, and these centroids, plus the corresponding cluster sizes and their mean squared errors are used to represent the original data. A typical real-life application might entail 200 variables measured on different scales (Braverman 2005). Accordingly, one must face up to the same standardization and reduction challenges here as in several of the previous examples. Based on their experimental data, the authors conclude that the summaries "capture important distributional features of the data related to physical processes" but anticipate that "modifications may be necessary" for a data production environment.

4. Commercial Applications

Commercial use of CA is hardly new. In particular, market segmentation services, based on cluster analysis, have been around a long time. See, for example, www.claritas.com for description of a system that develops clusters of people with like characteristics or preferences and then attaches catchy names to them such as "Money & Brains". The basic approach, which involves k-means clustering, has been described in several publications, including *The New Yorker* (2/1/82).

To identify more recent applications, magazines, newspapers, and other electronic sources appearing in 2003-2004 were scoured, again via online searching of databases. Details about how the clustering was carried out are rarely provided and are no doubt regarded as proprietary information in most cases. Here are a few examples illustrating the variety:

The Wall Street Journal (11/24/04) ran an article with the headline “Clustering Can Diversify a Real-Estate Portfolio”. It’s based on a report from Prudential Real Estate Investors that argues for diversifying investments across clusters of major metropolitan areas as opposed to simple geographic diversification.

Forbes (5/24/04) described how an online bank, ING Direct, uses regression analysis to determine variables that contribute the most to profitability. It then uses those variables to develop “clusters of neighbors” with similar attractive profitability profiles as potential new customers. The strategy can be thought of as clustering to find a particular group or groups of interest without bothering to find all of them. This is an instance of what Friedman and Meulman (2004) call targeted clustering.

Business Week (5/3/04) reported that several startup companies are developing clustering technology to organize the results of online searches into folders with computer-generated names. One of these companies is Vivisimo, which operates clusty.com. To try out the clusty system, a search was made of “cluster analysis” that returned 214(!) primary folders, including seemingly sensible ones labeled “hierarchical cluster”, “tools”, and “gene expression” as well one labeled “U.S.”, which included a paper on cluster bombing in Afghanistan. Clustering the results of searches, based on their verbal content, and automatically labeling the clusters, presents its own set of challenges! (See also *BusinessWeek Online*, 1/4/05.) *FORTUNE.COM* (12/16/03) described a new software tool, called Grokker, for organizing search results into a hierarchically-structured visual map summary. Apparently, some form of HCA is involved in the software engine that produces the visual map. (See also *The New York Times*, 5/9/05.)

The New York Times (3/21/04) contained a story on how a clustering algorithm can be used to assist the creation of new song hits. The process was glamorized as “Hit Song Science.” It’s based on acoustic similarities between songs and the ability to tell if a new song is near a cluster of old hit songs, even though it might not sound like any of them. The system was developed by Polyphonic HMI, a company in Barcelona. (See also *Time Magazine*, 10/24/05.)

Other commercial applications spotted on various newswires include: segmenting users of mobile technology, mining microarray data, analyzing hydrocarbon quality in petroleum fields, designing a new hedge fund index, modeling magazine circulation, managing software defects, and detecting fraud in insurance claims.

5. Discussion

Sections 2-4 document and illustrate the upsurge of CA applications. Reflecting on the practices that were observed in the literature review, several stand out as excellent and some as dangerous. It is also apparent that various gaps ought to be filled to provide better methodology for practitioners. The main purpose of this discussion section is to highlight some of these specific practices and research opportunities.

Overall, though, practitioners would benefit from a deeper understanding of the properties of all the methods and processes involved. Good practice typically involves looking at the data in different forms, considering alternative metrics and distance functions, comparing the results from different clustering algorithms, and checking the stability and validity of findings. Generally, it is wise to stick as close to the data as possible and not, for example, to become overly enamored with a nice looking dendrogram.

Over twenty years ago a study similar in spirit to this one, but covering a wider range of multivariate methodologies, recommended that practitioners take a more critical view of multivariate techniques, including CA, and try to avoid canned analyses (Gnanadesikan and Kettenring 1984). While there are plenty of examples to the contrary, the typical application of CA is still dominated by faith-based, fixed processes and unquestioned acceptance of results. With applications on the rise, it is even more critical today to raise the standards.

A convenient high-level view of CA is as a three-step process that involves preprocessing of the data, invoking algorithms to assist in identification of clusters, and assessing the results. The subsections that follow can be mapped to these three stages except the last one, Section 5.8, deals with the process as a whole. The discussion is tightly keyed to the examples in Section 3. These are listed in Table 1 for easy reference.

Table 1.Ten Featured Examples from Section 3

1	Aquatic plants
2	Disease co-occurrence
3	Cultures
4	Media usage
5	Juvenile offenders
6	Chromatography
7	Archaeology
8	Stem cells
9	Purebred domestic dogs
10	Remote sensing

5.1 Autoscaling

It has long been understood that scales of the variables can have a huge impact on the outcome of a CA. Often the nature of the variables is intrinsically different, as in Example 1, making it awkward and unappealing to work with them in their original forms. The most popular tactic for getting out of this mess is to autoscale each of them separately. This seemingly innocent initial step can obscure clusters in the data and render them undetectable in the output of a clustering algorithm.

If one could only standardize or transform the raw data so that any clusters present would appear as homogeneous spherical point clouds, then most of the popular CA algorithms would be able to extract them easily. Of course, only some data sets will lend themselves to such treatment. Even then, while procedures are available for finding sphericizing linear transformations (see Art, Gnanadesikan, and Kettenring 1982 and the SAS procedure called ACECLUS), their success depends in part on selecting suitable starting conditions for an iterative algorithm.

Sometimes simple transformations of variables, such as taking logs as in Example 7, can be very helpful for ameliorating scaling problems. Evidence in the current literature suggests they are underutilized.

Scaling to put the variables on the same footing is one consideration. Another is differential weighting to intentionally overemphasize those which are more likely to help the CA. In the extreme, some variables may merit zero weighting. The comment in Gnanadesikan, Kettenring, and Tsao (1995a) that “worry-free approaches do not yet exist” for any of these scaling challenges still holds.

Scaling of variables is a sensitive issue for most applications and algorithms. Effective alternatives to autoscaling should be a top research priority.

5.2 PCA and CA

Confusion about the role of PCA for reducing dimensionality and the number of variables entering the CA is another huge problem. As a possibly interesting orthogonal projection of the data, it can be very useful. But what is really going on? Assume there are g groups or clusters of n data points and they are known. Proceeding as in analysis of variance, the total sums of squares and cross-products matrix, **T**, can be decomposed into within and between matrices, i.e., **T** = **W** + **B**. If the groups are reasonably homogeneous, **W**/(n – g) provides a statistically sensible estimate of the common group covariance structure, and **B**/(g – 1) captures the variation among the group means. Standard practice is to base PCA on eigenanalysis of **W** and DA on eigenanalysis of $\mathbf{W}^{-1}\mathbf{B}$. The latter computation yields so-called discriminant variables that are designed to pull the groups apart, in contrast to the eigenanalysis of **T** (which is convenient) or **W** (which we don't know in the CA context). In situations where the groups are so separated that the role of **W** is sufficiently diminished or the groups have nearly spherical covariance structure so that **W** is roughly proportional to the identity matrix, **I**, then working directly on **T** without knowing the group structure should work fine. This is the reason that PCA often yields "interesting results" in practice—Example 6 may be one such case—and not because it is an optimal transformation to statistically uncorrelated components independent of the cluster structure. Many of these points can be found in the literature (e.g., Jolliffe 2002; Gnanadesikan, Kettenring, and Tsao 1995b). Yeung and Ruzzo (2001) investigated empirically the effectiveness of using the first few PCs for CA of both gene expression and synthetic data. Their conclusion: "Overall, we would not recommend PCA before clustering except in special circumstances."

Another challenge is to develop viable alternatives to PCA for reducing dimensionality in CA problems.

5.3 Variable Clustering

Numerous applications in the literature boil down, in statistical parlance, to clustering variables instead of, or in addition to, the objects or observations. Examples 2, 4, and 8 are of this kind.

In Example 4, the clustering of variables (media types) is used indirectly to discover how the objects (the audience) are segmented. A more direct approach of clustering the 1,800 subjects would seem preferable, notwithstanding the additional computational effort that this would entail. Indeed, one can envision situations where clustering media types would reveal very little about audience segments. Another strategy would be to carry out a two-way clustering so that any detailed block structure of segments and audience would be revealed all at once. Kaufman and Rousseeuw (1990) provide several references to such methods. Friedman and Meulman (2004) present a new procedure for clustering objects when their structure is defined by possibly different subsets of the variables.

Example 2 is closer to a pure variable clustering problem. In this case, there is a random sample of data from a single population of Indian elders. Correlation coefficients, appropriate for binary data, nicely summarize similarity relationships among the variables, and there is no confounding of them with group effects.

In Example 8, HCA is used to cluster gene expression profiles (the variables) across cell lines (the samples). As is typical in such studies, the samples are not the usual statistical random samples and may cover "all kinds of experimental conditions," according to Jiang et al. (2004). This raises concerns about the impact of unusual observations because of the well known sensitivity of Pearson correlations to outliers. Jiang et al. mention jackknife-based and Spearman's rank-order correlation coefficients as alternatives that also have their own limitations. Other robust estimators, such as one based on standardized sums and differences, developed by Devlin, Gnanadesikan, and Kettenring (1975), provide viable alternatives. Cherepinsky, Feng, Rejali, and Mishra (2002) describe a shrinkage-based correlation metric that improves accuracy for CA of microarray data. Basing the CA on Fisher's z-transform of Pearson-type correlations would reduce the problem to one of location differences only.

More generally, while hardly a new idea, the use of CA either to group variables that are similar, as an end in itself, or to reduce the number of variables to be used to cluster observations is an appealing, relatively underdeveloped, and underappreciated strategy. The trick is to do execute the variable clustering so that it is not thrown off if the observations themselves are clustered.

A deeper understanding of how best to cluster variables for such purposes would be very beneficial to practitioners.

5.4 DA and CA

As pointed out in Section 1, DA and CA are at the opposite ends of a spectrum of problems ranging from zero knowledge to complete knowledge about cluster structure. Examples 1 and 9 involve situations where there is suspected or partial knowledge of this kind.

Portions of this continuum are reflected in the rapidly growing literature on semi-supervised learning in general (see Zhu 2005 for an online updated survey) and semi-supervised clustering in particular (see Grira, Crucianu, and Boujemaa 2004 for a brief survey and Basu, Bilenko, and Mooney 2004 for a concrete example of a new algorithm of this type—a natural generalization of k-means). These methods utilize extra information, such as group labels on some of the data or constraints designed to keep certain pairs of points in the same or different clusters.

New methods for dealing with problems along the CA-DA spectrum should provide data analysts a richer set of CA algorithms to choose from.

5.5 Tree Cutting

The most popular way of obtaining a partition of data into clusters is to perform a straight line cut of the dendrogram at an "appropriate" level and then to treat each separate branch as a cluster. Several software systems, e.g., R and Clustan, conveniently assist in such surgery. Examples 5 and 6 illustrate the approach.

However, tree cutting, if done mindlessly, can be perilous. Consider the data in Figure 5, consisting of three spherical clusters each of size 25. The left one was generated from a standard bivariate normal distribution (means = 0, correlation = 0, and variances =1). The two tighter clusters were generated similarly with means = 2 and variances = 0.1 in one case, and means = 3 and variances = 0.1 in the other. Any sensible approach to clustering such data should be able to detect such clear structure. Applying HCA (average method, Euclidean distance) results in the dendrogram shown in Figure 6. While the three-cluster structure is easily spotted, there is no single horizontal cut of the tree that will reproduce it. Fair warning to tree cutters! (See Stuetzle 2003 for related discussion.)

Because of the popularity of HCA, more sophisticated tools for extracting clusters from the dendrogram would be very beneficial.

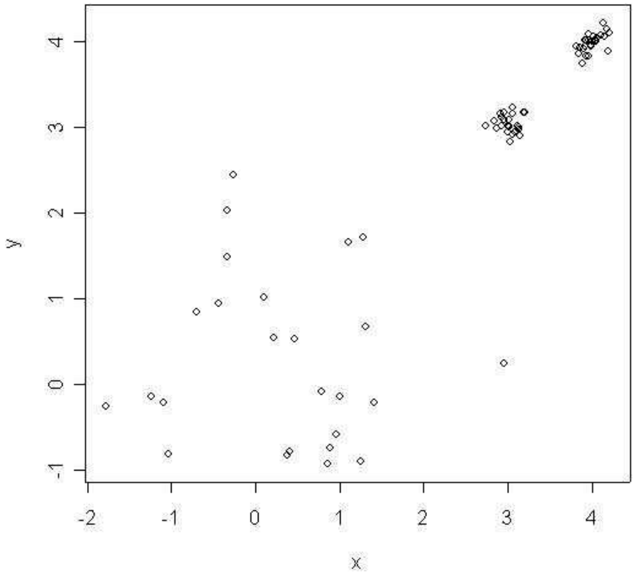

Figure 5. Scatter plot of three simulated spherical clusters each of size 25. The more dispersed cluster is centered at (0, 0) with variances of (1,1). The other two each have variances of (0.1, 0.1).

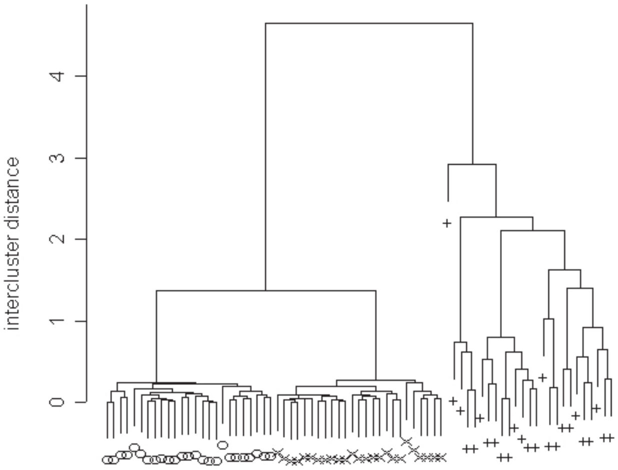

Figure 6. Dendrogram from HCA of data in Figure 5 (average method, Euclidean distance). The symbols "o" and "x" represent points in the two tight clusters, and "+" denotes those from the more dispersed cluster. There are three clearly defined branches that completely capture the three-cluster structure. However, no horizontal line cut of the tree will capture them.

5.6 Very Large Problems

Perhaps the single most striking fact behind the usage data is that most published applications are of moderate size in terms of number of observations (n) and dimensions (p). Example 1, with n = 77 and p = 25, and Example 3, with n = 36 and p = 30, are typical. Nevertheless, the current literature contains a sprinkling of instances where the number of objects being clustered is truly huge. Examples 8 (large p) and 10 (large n) are such cases. In addition, the literature on data mining includes much discussion of large CA problems.

Large streaming data sets, as in Example 10, are increasingly common in many contexts, such as network performance analysis and message monitoring for intelligence purposes. As this trend continues, there will be more demand for methods that cope well with such situations. Murtagh (2002) discusses many issues that arise in these very large problems. Xu and Wunsch (2005) review a number of approaches that have been developed for “big n” and “big p” problems. Already there are software modules for clustering “big n” data sets in several of the standard packages.

Sampling is often a sensible strategy when n is large, especially if it can be done repeatedly, so as not to miss small clusters completely. The results can be compared across samples and integrated as appropriate. Some special purpose algorithms (e.g., Maitra 2001; Rocke and Dai 2003; Tseng and Wong 2005) are built around the idea of sampling.

Data sharpening (Tukey and Tukey 1981) is a tactic for reducing the number of objects to be clustered. The idea is to move or remove individual points depending upon the density of those around them. Stanberry, Nandy, and Cordes (2003) apply a dendrogram sharpening technique to fMRI data. Its effect is to “discard all small-sized children-nodes with a large-sized parent node.” Sharpening techniques involve distance metrics, and hence issues of scale and metric loom large.

There is an increasing need for CA methods that can effectively handle problems with “big n” and/or “big p.”

5.7 Validation and Interpretation

The need for solid validation and careful interpretation of CA results is clear and has been recognized by many researchers. To cite one recent instance, Bottomley and Nairn (2004) report on the results of experiments with marketing managers and found that “random data devoid of meaningful structure were perceived as equally useful for purposes of market segmentation as real data." There are a variety of approaches that can be applied ranging from simple graphical displays to formal inferential procedures to gain confidence that the clusters are more than artifacts of the CA algorithm or process. Many of these are illustrated by the examples in Section 3.

In wrapping up their discussion of graphical displays for CA, Kaufman and Rousseeuw (1990) commented that the topic "has so far received too little attention in the literature." Sadly, the situation has not improved very much over the last 15 years. The literature search uncovered many cases of PC scatter plots (Examples 1, 6, and 7), but few other techniques such as scatter plot matrices of the raw data, with or without brushing, or of DA-based projection plots designed to give the best view of clusters in lower dimensions. Interactive graphical tools, based on grand tours (Buja, Cook, and Swayne 1996) or parallel coordinate plots (Wegman 1990) can also be effective in searching for cluster structure in moderate-sized multivariate data.

Example 1 is a situation that might have benefited from DA projection plots, utilizing either the discriminant variables or simpler ones obtained using the eigenvectors of **B,** with **B** based on either the four pre-specified or the recovered groupings. Several versions of such plots are described in Gnanadesikan, Kettenring, and Landwehr (1982).

The tantalizing dendrogram from Example 3, shown in Figure 4, cries out for a return to the raw data. Is Turkey coupled with the United States and Canada because of a process anomaly or because of something interesting in the data? Heat maps would be helpful in answering such a question. This display technique, which has blossomed in microarray studies, deserves a prominent spot in the CA toolkit.

Sensitivity analyses are relatively easy to do and can be very useful in ensuring robust results. In Example 2, the co-morbidity study, two different similarity metrics and HCA algorithms (average similarity and complete linkage) were applied to check on the sensitivity of findings to these choices. When using an iterative procedure, such as k-means CA, experimenting with different starting points, as in Example 10, can help to avoid sub-optimal results. As Steinley (2003) has shown, implementations in commercial packages usually yield solutions that are only locally optimal. A referee recommends a series of random restarts: "This enables a probability statement about the chance that the next restart will discover a previously unobserved local maximum, and the chance that it will be better than any [one] previously found."

There are many effective ways to work with subsets of the data to check results and aggregate findings. In Example 2, the sample was randomly split into two so that results from one data set could be cross-validated against the other. In Example 9, many analyses were run with different subsets of the data leading to consensus results that go well beyond what one might learn from a single application of CA. Example 7 illustrates how DA cross-validation techniques can be used as a check on CA. In addition to providing an overall figure of merit, such calculations, which treat the tentative clusters as known groups, can help distinguish ones that are well determined from those that are not. A different twist is to work with subsets of the variables to find a minimal one that will reproduce the clusters found using all of them, as done in Example 6.

Many of the examples compare or combine the results of applying different methods to the same or augmented data. For instance, in Example 3, the HCA results were augmented with a multidimensional scaling representation of the cultures based on dissimilarity inputs equivalent to those used to obtain Figure 4. A two-dimensional solution showed all the cultures but two distributed in a circular pattern that could be interpreted in terms of geography and religion. These results nicely complemented the HCA findings. In one major analysis from the chromatography study of Example 6, the dendrogram was cut to yield a partition of nine clusters of the 24 column-solvent combinations and these, along with their subclusters, were overlaid on the scatter plot of the first two PCs to aid interpretation. In Example 4, strong use was made of factor and regression analyses to support and explain the clusters of media repertoires.

There are various ways to gain insights from familiar statistics that summarize goodness of fit and quantify cluster separations. Kruskal’s gamma statistic was applied in Example 2 to conclude that the “four-cluster model has recovered the structure fairly well.” In Example 5, heavy use was made of ANOVA results and Tukey pairwise multiple comparisons to assess the differences among the cluster means on the 12 individual variables. As a way of indicating which comparisons are relatively more different, this is an excellent pragmatic procedure (and many variations on this idea are available). However, the fact that the clusters are data-based, rather than prespecified, renders formal statements of statistical significance invalid. This misleading extra step is fairly prevalent in the literature.

Model-based approaches to clustering have undergone vigorous development in recent years (Fraley and Raftery 2002). Powerful methods and supporting software (Fraley and Raftery 2003) are available to support applications. With such approaches one can capitalize on the added structure to make a variety of inferences about the model such as the number of clusters present and their shapes. Typically the modeling is centered on the assumption that the data are adequately represented by a mixture of multivariate normal distributions, as in Banfield and Raftery (1993) and Yeung, Fraley, Murua, Raftery, and Ruzzo (2001). Examples 7 and 9 both utilize model-based approaches.

Clearly, there are many dimensions to consider under the heading of validation and interpretation. This discussion has only scratched the surface. Fortunately, many useful aids already exist, and their use is evident in the examples and broader literature. Yet, there is nothing close to agreement on exactly how one should proceed.

The research community should push to establish a set of best practices that users can draw on for interpretation and validation of their results.

5.8 Circularity

Circularity is used here to refer to the risk of obtaining CA results that are more due to the vagaries of the process than to the strength of the cluster structure in the data. Because CA is filled with so many opportunities to be fooled, this is a not a trivial issue. The risk is especially high in applications relying on “one shot” computations without any validation.

Autoscaling provides an example of such a risk. It’s a process that invites trouble: variables are standardized to put them on “equal footing”, because it seems the right thing to do, and unwittingly the ability to detect clusters that would otherwise have been apparent is diminished. Invoking Mahalanobis distances for HCA, as discussed in the paper described in Example 6, amounts to generalized autoscaling and is even harder to justify and more likely to mislead. It also illustrates the fallacy of thinking that methods that are invariant to linear transformations of the data are always desirable.

The use of PCA to reduce dimensionality prior to CA introduces its own distortions. Apart from the flaws mentioned already in Section 5.2, different PCs (and hence different CAs) are obtained usually, depending upon whether the analysis is done on **T** or an autoscaled version of it.

Methods of sharpening that are based on distances have a similar problem. The points that are moved or removed depend heavily on the metric involved.

Another version of circularity stems from the choice of clustering algorithm. A k-means algorithm favors finding spherical clusters. A complete-linkage HCA will have the same tendency. All algorithms have their tendencies for finding certain types of cluster structure. Some of them are well understood, others not. In a sense, the choice of a particular algorithm biases the analysis towards solutions that play to its strengths, such as uncovering spherical clusters. (Fisher and Van Ness 1971 offer a principled approach for determining a "best" algorithm, which they describe as a "perplexing problem.")

Improving the process by protecting against intrinsic biases due to circularity should significantly improve the usefulness of CA for practitioners.

6. Conclusions

The use of CA has grown at an astounding rate during the last decade. Applications can be found across a wide spectrum of fields, literally "a" to "z", ranging from archaeology to zoology, and especially in the life sciences. Advances in reliable methods and flexible software have come at a slower pace. Packaged algorithms are often used too routinely, without regard to their limitations. The impact can be serious: progress on important problems is slowed, because informative cluster structures remain hidden, or even reversed, because invalid ones are "discovered". The best hope for improving matters is for continued progress on the research front, including deeper understanding of the properties of different methods, and effective communication of best practices to the wider research community.

References

ALLIK, J., and MCCRAE, R.R. (2004), "Toward a Geography of Personality Traits – Patterns of Profiles Across 36 Cultures," *Journal of Cross-Cultural Psychology*, *35*, 13-28.

ANDERSON, T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley. (3rd ed., 2003)

ART, D.A., GNANADESIKAN, R., and KETTENRING, J.R. (1982), "Data-based Metrics for Cluster Analysis," *Utilitas Mathematica*, *21A*, 75-79.

BANFIELD, J., and RAFTERY, A. (1993), "Model-based Gaussian and Non-Gaussian Clustering," *Biometrics*, *49*, 803-821.

BASU, S., BILENKO, M., and MOONEY, R.J. (2004), "A Probabilistic Framework for Semi-Supervised Clustering," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004),* pp. 59-68.

BAXTER, M. (2003), *Statistics in Archaeology*, London: Hodder Arnold.

BOTTOMLEY, P., and NAIRN, A. (2004), "Blinded by Science: the Managerial Consequences of Inadequately Validated Cluster Analysis Solutions," *International Journal of Market Research*, *46***,** 171-187.

BRAVERMAN, A. (2005). Personal communication.

BRAVERMAN, A., FETZER, E., ELDERING, A., NITTEL, S., and LEUNG, K. (2003), "Semi-streaming Quantization for Remote Sensing Data," *Journal of Computational and Graphical Statistics*, *12*, 759-780.

BROWN, C.E. (1998). *Applied Multivariate Statistics in Geohydrology and Related Sciences*. Berlin: Springer-Verlag.

BUJA, A., COOK, D., and SWAYNE, D.F. (1996), "Interactive High-Dimensional Data Visualization," *Journal of Computational and Graphical Statistics*, *5*, 78-97.

CHANG, W-C. (1983). "On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions," *Applied Statistics*, *32,* 267-275.

CHEREPINSKY, V., FENG, J., REJALI, M., and MISHRA, B. (2002), "Shrinkage-based Similarity Metric for Cluster Analysis of Microarray Data," *Proceedings of the National Academy of Sciences*, *100*, 9668-9673.

DEVLIN, S.J., GNANADESIKAN, R., and KETTENRING, J.R. (1976), "Robust Estimation and Outlier Detection With Correlation Coefficients," *Biometrika*, *62*, 531-545.

DOMANY, E. (2003), "Cluster Analysis of Gene Expression Data," *Journal of Statistical Physics*, *110*, 1117-1139.

EISEN, M.B., SPELLMAN, P.T., BROWN, P.O., and BOTSTEIN, D. (1998), "Cluster Analysis and Display of Genome-wide Expression Patterns," *Proceedings of the National Academy of Sciences*, *95,* 14863-14868.

EVERITT, B.S., LANDAU, S., and LEESE, M. (2001), *Cluster Analysis*, (4th^{t}ed.), New York: Arnold Publishers.

FALUSH, D., STEPHENS, M., and PRITCHARD, J.K. (2003), "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies," *Genetics*, *164,* 1567-1587.

FISHER, L., and VAN NESS, J.W. (1971), "Admissible Clustering Procedures," *Biometrika*, *58*, 91-104.

FRALEY, C. (1998), "Algorithms for Model-based Gaussian Hierarchical Clustering," *SIAM Journal of Scientific Computing*, *20*, 270-281.

FRALEY, C., and RAFTERY, A.E. (2002), "Model-based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, *97*, 611-631.

FRALEY, C., and RAFTERY, A.E. (2003), "Enhanced Model-based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST," *Journal of Classification*, *20*, 263-286.

FRIEDMAN, J., and MEULMAN, J. (2004), "Clustering Objects on Subsets of Attributes (with discussion)," *Journal of the Royal Statistical Society B*, *Ser. B*, 815-849.

GNANADESIKAN, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*, New York: Wiley. (2nd ed., 1997)

GNANADESIKAN, R., and KETTENRING, J R. (1984), "A Pragmatic Review of Multivariate Methods in Applications," in *Statistics: An Appraisal*, Eds. H.A. David and H.T. David, Ames, IA: The Iowa State University Press, pp. 309-337.

GNANADESIKAN, R. KETTENRING, J.R., and LANDWEHR, J.M. (1982), "Projection Plots for Displaying Clusters," in *Statistics and Probability: Essays in Honor of C. R. Rao*, Eds. G. Kallianpur, P.R. Krishnaiah, and J.K. Ghosh, Amsterdam: North-Holland, pp. 269-290.

GNANADESIKAN, R, KETTENRING, J.R., and TSAO, S.L. (1995A), "Weighting and Selection of Variables for Cluster Analysis," *Journal of Classification*, *12*, 113-136.

GNANADESIKAN, R., KETTENRING, J.R., and TSAO, S.L. (1995B), "Some Practical Issues in Using Cluster Analysis," *Bulletin of the International Statistical Institute, Contributed Papers*, *1,* 412-413.

GORDON, A.D. (1999), *Classification* (2nd ed.), Boca Raton: Chapman & Hall/CRC.

GRIRA, N, CRUCIANU, M., and BOUJEMAA, M.,"Unsupervised and Semi-supervised Clustering: A Brief Survey," in *A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (6th Framework Programme).*

HALL, M.E. (2004), “Pottery Production During the Late Jomon Period: Insights from the Chemical Analyses of Kasori B Pottery,” *Journal of Archaeological Science*, *31*, 1439-1450.

HAND, D., MANNILA, H., and SMYTH, P. (2001), *Principles of Data Mining*, Cambridge, MA: The MIT Press.

HARTIGAN, J.A. (1975), *Clustering Algorithms*, New York: Wiley-Interscience.

HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag.

JIANG, D.X., TANG, C. and ZHANG, AD. (2004), “Cluster Analysis for Gene Expression Data: a Survey,” *IEEE Transactions on Knowledge and Data Engineering*, *16*, 1370-1386.

JOHN, R., KERBY, D.S. and HENNESSY, C.H. (2003), “Patterns and Impact of Comorbidity and Multimorbidity Among Community-resident American Indian Elders,” *The Gerontologist*, *43*, 49-660.

JOHNSON, R. A. and WICHERN, D. W. (1982), *Applied Multivariate Statistical Analysis*, Englewood Cliffs, NJ: Prentice-Hall. (5th ed., 2002)

JOLLIFFE, I.T. (2002), *Principal Component Analysis* (2nd ed.), New York: Springer.

KAUFMAN, L., and ROUSSEEUW, P.J. (1990), *Finding Groups in Data: an Introduction to Cluster Analysis*, New York: Wiley-Interscience.

KIM, C., SHIN, H., and CHOI, H. (2003), “A Phenetic Analysis of *Typha* in Korea and Far East Russia,” *Aquatic Botany*, *75*, 33-43.

LE MAPIHAN, K., VIAL, J., and JARDY, A. (2004), “Testing of ‘Special Base’ Columns in Reversed-phase Liquid Chromatography—A Rational Approach Considering Solvent Effects,” *Journal of Chromatography A*, *1030*,135-147.

LEGENDRE, P., and LEGENDRE, L. (1998), *Numerical Ecology*, Amsterdam: Elsevier.

MAITRA, R. (2001). “Clustering Massive Data Sets with Applications in Software Metrics and Tomography,” *Technometrics*, *43*, 336-346.

MILLER, J.M. (2004), *Chromatography: Concepts and Contrasts* (2nd ed.), New York: Wiley-Interscience.

MURTAGH, F. (2002), “Clustering in Massive Data Sets,” in *Handbook of Massive Data Sets*, Eds. J. Abello, P.M. Pardalos, and M.G. Resende, New York: Kluwer, pp. 501-543.

PARKER, H.G., KIM, L.V., SUTTER, N.B., CARLSON, S., LORENTZEN, T.D., MALEK, T.B., JOHNSON, G.S., DEFRANCE, H.B., OSTRANDER, E.A., and KRUGLYAK, L. (2004), “Genetic Structure of the Purebred Domestic Dog,” *Science*, *304*, 1160-1164.

RAO, C.R. (1965), *Linear Statistical Inference and Its Applications*, New York: Wiley. (2nd ed., 1973)

ROCKE, D., and DAI, J. (2003), “Sampling and Subsampling for Cluster Analysis in Data Mining: With Applications to Sky Survey Data,” *Data Mining and Knowledge Discovery*, *7*, 215-232.

SEBASTIANI, P., GUSSONI, E., KOHANE, I.S. and RAMONI, M.F. (2003), “Statistical Challenges in Functional Genomics,” *Statistical Science*, *18*, 33-70.

SEBER, G.A.F. (1984), *Multivariate Observations*, New York: Wiley.

SHAW, P.J.A. (2003), *Multivariate Statistics for the Environmental Sciences*, London: Hodder Arnold.

SNEATH, P.H.A., and SOKAL, R.R. (1973), *Numerical Taxonomy*, San Francisco: Freeman.

SPERGER, J.M., CHEN, X., DRAPER, J.S., ANTOSIEWICZ, J.E., CHON, C H., JONES, S.B., BROOKS, J.D., ANDREWS, P.W., BROWN, P.O., and THOMSON, J.A. (2003), “Gene Expression Patterns in Human Embryonic Stem Cells and Human Pluripotent Germ Cell Tumors,” *Proceedings of the National Academy of Sciences*, *100*, 13350-13355.

STANBERRY, L., NANDY, R., and CORDES, D (2003), "Cluster Analysis of fMRI Data Using Dendrogram Sharpening," *Human Brain Mapping*, *20*, 201-219.

STEFURAK, T., CALHOUN, G.G., and GLASER, B. A. (2004), "Personality Typologies of Male Juvenile Offenders Using a Cluster Analysis of the Millon Adolescent Clinical Inventory Introduction," *International Journal of Offender Therapy and Comparative Criminology, 48,* 96-110.

STEINLEY, D. (2003), "Local Optima in K-Means Clustering: What You Don't Know May Hurt You," *Psychological Methods, 8*, 294-304.

STUETZLE, W. (2003), "Estimating the Cluster Tree of a Density by Analyzing the Minimal Spanning Tree of a Sample," *Journal of Classification*, *20*, 25-47.

TSENG, G.C., and WONG, W.H. (2005), "Tight Clustering: A Resampling-based Approach for Identifying Stable and Tight Patterns in Data," *Biometrics*, *61*, 10-16.

TUKEY, P.A., and TUKEY, J.W. (1981), "Data-driven View Selection; Agglomeration and Sharpening," in *Interpreting Multivariate Data,* Ed. V. Barnett, New York: Wiley, pp. 215-243.

VAN REES, K., and VAN EIJCK, K. (2003), "Media Repertoires of Selective Audiences: the Impact of Status, Gender, and Age on Media Use," *Poetics*, *31*, 465-490.

WEGMAN, E.J. (2003), "Hyperdimensional Data Analysis Using parallel Coordinates," *Journal of the American Statistical Association*, *85*, 664-675.

XU, R. and WUNSCH, D. (2005), "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, *16,* 645-678.

YEUNG, K.Y., FRALEY, C., MURUA, A., RAFTERY, A.E., and RUZZO, W.L. (2001), "Model-based Clustering and Data Transformations for Gene Expression Data," *Bioinformatics*, *17*, 977-987.

YEUNG, K.Y., and RUZZO, W.L. (2001), "Principal Component Analysis for Clustering Gene Expression Data," *Bioinformatics*, *17*, 763-774.

ZHU, X. (2005). "Semi-Supervised Learning Literature Survey," Computer Sciences Technical Report 1530, University of Wisconsin, available at http://www.cs.wisc.edu /~jerryzhu/pub/ssl_survey.pdf.