WARNING

CONCERNING COPYRIGHT RESTRICTIONS

The Copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or reproduction.

One of three specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research.

If electronic transmission of reserve material is used for purposes in excess of what constitutes “fair use”, that user may be liable for copyright infringement.

This policy is in effect for the following document:

TITLE: Some final comments and guidelines (Chapter 9) / from Cluster analysis
AUTHOR: Everitt
SOURCE: Chichester: Wiley, 2011 pp.257-287

NO FURTHER TRANSMISSION OR DISTRIBUTION OF THIS MATERIAL IS PERMITTED

Cluster Analysis

5th Edition

Brian S. Everitt • Sabine Landau
Morven Leese • Daniel Stahl
King's College London, UK

WILEY
A John Wiley and Sons, Ltd., Publication

This edition first published 2011

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Everitt, Brian.
Cluster Analysis / Brian S. Everitt. – 5th ed.
p. cm. – (Wiley series in probability and statistics ; 848)
Summary: "This edition provides a thorough revision of the fourth edition which focuses on the practical aspects of cluster analysis and covers new methodology in terms of longitudinal data and provides examples from bioinformatics. Real life examples are used throughout to demonstrate the application of the theory, and figures are used extensively to illustrate graphical techniques. This book includes an appendix of getting started on cluster analysis using R, as well as a comprehensive and up-to-date bibliography."– Provided by publisher.
Summary: "This edition provides a thorough revision of the fourth edition which focuses on the practical aspects of cluster analysis and covers new methodology in terms of longitudinal data and provides examples from bioinformatics"– Provided by publisher.
Includes bibliographical references and index.
ISBN 978-0-470-74991-3 (hardback)
1. Cluster analysis. I. Title.
QA278.E9 2011
519.5'3–dc22

2010037932

A catalogue record for this book is available from the British Library.

Print ISBN: 978-0-470-74991-3
ePDF ISBN: 978-0-470-97780-4
oBook ISBN: 978-0-470-97781-1
ePub ISBN: 978-0-470-97844-3

Set in 10.25/12pt Times Roman by Thomson Digital, Noida, India
Printed and bound in Great Britain by TJ International Ltd, Padstow, Cornwall

9

Some final comments and guidelines

9.1 Introduction

It should by now be obvious to most readers that the use of cluster analysis in practice does not involve simply the application of one particular technique to the data under investigation, but rather necessitates a series of steps, each of which may be dependent on the results of the preceding one. It is generally impossible *a priori* to anticipate what combination of variables, similarity measures and clustering techniques is likely to lead to interesting and informative classifications. Consequently, the analysis proceeds through several stages, with the researcher intervening if necessary to alter variables, choose a different similarity measure, concentrate on a particular subset of individuals, and so on. The final, extremely important stage concerns the evaluation of the clustering solutions obtained. Are the clusters real or merely artefacts of the algorithms? Do other solutions exist which are better? Can the clusters be given a convincing interpretation? A long list of such questions (which are full of traps for the unwary; see Dubes and Jain, 1979) might be posed.

It should also be clear from the preceding chapters that no one clustering method can be judged to be 'best' in all circumstances. The various studies that have compared a variety of clustering procedures on artificially generated data all point to the existence of a 'method × data type' interaction. In other words, particular methods will be best for particular types of data. Table 9.1 shows some possible method–data combinations.

In many applications it might be reasonable to apply a number of clustering methods. If all produce very similar solutions, the investigator might, justifiably

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

Table 9.1 Overview of data types and applicable clustering methods.

Data type	Method	Notes
Continuous (Ordered data can be treated as continuous, possibly with standardization based on range)	Hierarchical agglomerative or partitioning methods	Ward's method, average linkage, and k-means or methods based on det($\mathbf{W}$) are popular choices; contiguity-constrained versions are possible
	Density search, mode analysis	
	Mixtures models	Especially for multivariate normal data; may need to restrict number of parameters fitted and/or take account of structure (such as repeated measures)
	Fuzzy k-means	Where clusters other than the 'best' are of interest
	Direct data clustering	Useful for data representing positive associations; both objects and variables clustered; need to consider scaling of data
	Kohonen self-organizing map (SOM)	Useful for large data sets; produces low-dimensional plot of clusters
Binary (Categorical data can be converted to binary)	Hierarchical or partitioning methods using appropriate proximity measure	Need to consider negative match issue; some special proximity measures available for categorical data
	Monothetic divisive method	Useful for developing diagnostic keys
	Latent classes or grade of membership (GOM)	These are 'fuzzy' methods; GOM often used in health applications
	Hierarchical classes (HICLAS)	Objects and variables clustered; overlapping clusters, useful for psychological data
Mixed mode	Hierarchical or partitioning methods using appropriate proximity measure or using ranks	Gower's similarity measure can be used
	Two step (SPSS)	Note possible problems with commensurability
	Model-based models for mixed data types	

Proximity matrix (Either computed from data or measured directly)	Hierarchical agglomerative or partitioning methods that do not require raw data	Examples are single, average and complete linkage, or partitioning around medoids (PAM)
	Additive clustering (ADCLUS and variants)	Overlapping clusters
	Kaufman and Rousseeuw fuzzy method (FANNY)	Similar to fuzzy k-means but raw data not required
	Tree-fitting methods	Dendrograms, additive trees; pyramids (limited overlap)
Large data sets	Data reduction methods such as principal components, spectral analysis	Reduction methods do not generally respect the cluster structure (so may be misleading)
	Clustering a subsample; using the clusters as seeds for further processing	Sampling process may need to be adapted to likely cluster structure

perhaps, have more confidence that the results are worthy of further investigation. Widely different solutions might be taken as evidence against any clear-cut cluster structure.

Most all-purpose statistical packages contain the 'standard methods' described in Chapters 4 and 5, and a few of the model-based methods described in Chapter 6. In `Stata`, k-means and k-medians, and hierarchical clustering methods are available using the `cluster` command or via a user-supplied dissimilarity matrix, using the `clustermat` command. Linkage methods are single, average, complete, weighted average, median, centroid and Ward's method. Dissimilarities for continuous variables include Euclidean and squared Euclidean, city block and a generalization of Euclidean and city block (Minkowski), Canberra, and similarity coefficients include correlation and angular separation. A large number of binary coefficients are available including simple matching and Jaccard similarities, and Gower's mixed data dissimilarity coefficient. `Stata` also includes the partitioning methods k-means and k-medians, and commands for the Calinski–Harabasz and the Duda–Hart indices are available. Clustering in `SPSS` can be obtained via the menu system or using `cluster` in syntax. Hierarchical methods are: between-groups linkage, within-groups linkage, nearest neighbour, furthest neighbour, centroid clustering, median clustering, and Ward's method, and the partitioning method k-means. The range of similarity and dissimilarity measures is similar to that in `Stata`, and slightly more extensive for binary variables, although Gower's method is not available. Both values and dissimilarity measures can be transformed in a number of ways (e.g. z-scores, or to a range of 0–1, and by case or by variable) within the command. Dendrograms and icicle plots can be obtained, but no stopping rules. The `twostep` method for mixed variable types is also available. Most of the other main statistical packages such as `SAS`, `Genstat`, `BMDP` and `Statistica` contain a similar range of methods. `S-PLUS` provides traditional hierarchical methods in `hclust` (single, complete, average, McQuitty, median, centroid, Ward's), but also some more unusual hierarchical methods such as `agnes`, and two divisive methods, DIANA and the monothetic method MONA, the partitioning method PAM as well as the model-based `mclust`. The specialized packages `Clustan` and `Clustan Graphics` are also available. There is an increasing number of specialized routines available in `R`.

There is no optimal strategy for either applying clustering or evaluating results, but some suggestions, which might be helpful in many situations, are discussed in this chapter, starting with an overview of the steps in a typical analysis. Those steps concerned with cluster validation and interpretation will be discussed in more detail in Sections 9.3–9.6, and two applications that illustrate many of the issues involved will be discussed in Section 9.7.

9.2 Using clustering techniques in practice

Milligan (1996) identifies seven steps that generally constitute a cluster analysis, based on findings from a number of studies, and gives a series of very useful tables

summarizing the results of these studies. The framework suggested by Milligan is outlined below, with some further comments. The first six steps have largely been covered in previous chapters. The topics mentioned in the final steps, interpretation, testing and replication, are discussed in more detail in this chapter, in sections on graphical methods for interpretation, testing for absence of cluster structure, comparing clusterings (both partitions and trees) and checks for robustness and stability.

The steps in a typical cluster analysis suggested by Milligan (1996) are as follows (the comments are a combination of Milligan's original suggestions and some additional points which we consider important):

(i) *Objects to cluster*: These should be representative of the cluster structure believed to be present, and they should be randomly sampled if generalization to a larger population is required. However, since cluster analysis is not an inferential technique, over-sampling of small populations and the placement in the sample of 'ideal types' (representatives of clusters suspected of being present) may be acceptable so long as generalization is not required.

(ii) *Variables to be used*: Variables should only be included if there is good reason to think they will define the clusters. Irrelevant or *masking* variables should be excluded if possible. A possible solution to the problem of masking variables is to employ the data matrix to suggest variable weights as described in Section 3.7, and an alternative solution is to use model-based variable selection as described in Section 6.6.

(iii) *Missing values*: Where the proportion of missing values is low, imputation of the raw data matrix may be acceptable, for example based on the clusters obtained in an initial pass, followed by re-clustering. Alternatively, the elements in a similarity or dissimilarity matrix can be imputed using only variables that are present, sometimes known as 'pairwise deletion'. For methods that use (raw) categorical data, for example latent class or grade of membership analysis, it is possible to include an additional response level to denote 'missing'. Model-based methods may be able to accommodate missing values as part of the expression of the likelihood.

(iv) *Variable standardization*: Standardization is not necessarily always indicated and can sometimes be misleading, as shown in Section 3.8. Standardization using the range showed good recovery of clusters in the simulations of Milligan and Cooper (1988) and should be considered as an alternative to the more usual standardization using standard deviations. Another solution to the problem of choosing an appropriate unit of measurement is to employ a cluster method that is invariant under scaling – see Chapter 6.

(v) *Proximity measure*: There are few general guidelines for this (see Section 3.9), but knowledge of the context and type of data may suggest suitable choices from those given in Chapter 3.

(vi) *Clustering method*: Methods should be: designed to recover the types of clusters suspected; effective at recovering them; insensitive to error; and

available in software. It is also advisable to consider data-generating processes, and this might suggest the application of a model-based method, as described in Chapters 6 and 7.

(vii) *Number of clusters*: One of the most difficult decisions to make is the number of clusters. Where different stopping rules suggest different numbers, the highest should also be considered for safety (unless external information from the subject matter suggests a suitable choice). An alternative would be to consider the possibility that there are no clusters present (see Section 9.3).

(viii) *Replication and testing*: Replication can involve cross-validation techniques, to investigate how far clusters identified in a subsample are still identifiable among the subsample of objects *not* used in the clustering. Another useful technique is the perturbation of the sample by omitting or slightly changing particular data points. Section 9.5 describes some of the techniques available for assessing internal validity. Goodness-of-fit statistics can be calculated to compare the clusters to the data used to derive them. Quality assessment may also involve comparing results between subsets, between the sample and a second sample or an external classification, using, for example, the Rand index for comparing partitions, or the cophenetic correlation for comparing dendrograms (see Section 9.4).

(ix) *Interpretation*: This may require graphical representation and descriptive statistics. Section 9.6 describes some graphical techniques helpful for cluster interpretation. It is important to note that standard statistical tests such as analysis of variance are inappropriate for comparing the *clustering* variables between clusters, since the clustering technique will have maximized between-cluster differences on these variables in some way.

The logical starting point for a cluster analysis would be a test for the absence of cluster structure. However, such tests are not usually employed in practical applications of clustering. This may be because the available tests are of limited usefulness. Their power is generally unknown and depends on the cluster structure, and so a test might simply not detect any departures from the null model due to lack of power. Nevertheless, the subject has some theoretical interest and therefore a short overview of this topic, based on the excellent review by Gordon (1998), is given in the next section. Subsequent sections deal with comparing partitions and dendrograms, measures of internal validity, and graphical methods for interpretation.

9.3 Testing for absence of structure

A test for the absence of cluster structure may not be necessary if the reason for clustering is practical (e.g. for organizational purposes). However, if it is aimed at detecting an unknown underlying structure then testing becomes more relevant. What is required is a model that describes the data-generating process in the

absence of clustering, and a test statistic which will reflect departures from the model.

The *Poisson model* assumes that, in the absence of cluster structure, the n individual p-variate observations arise from a uniform distribution over some region A of p-dimensional space. Equivalently, the underlying frequency distribution is assumed to have no mode. The (random) number of individuals observed within any subregion A_s follows a Poisson distribution with mean $\lambda/|A_s|$ where λ is the constant intensity and $|A_s|$ denotes the p-dimensional volume of A_s. In the absence of theoretical results for finite samples, the null distribution of a test statistic can be generated by Monte Carlo simulation, by repeated sampling of p-variate observations from a uniform distribution over A. This hypothesis has been referred to as the *uniformity hypothesis* (Bock, 1985), as the *random position hypothesis* (Jain and Dubes, 1988) or, in the framework of spatial statistics, as the *complete spatial randomness hypothesis* (Diggle, 1983).

Departures from random positioning can be due to regularity or clustering. The number of interpoint distances below a specified threshold (Strauss, 1975; Kelly and Ripley, 1976; Saunders and Funk, 1977) and the largest nearest-neighbour distance within the set of individuals (Bock, 1985) can be used to assess departures from the Poisson model. The distances from randomly selected positions to the nearest points can be compared with distances between those points and their nearest neighbour (Cross and Jain, 1982; Panayirci and Dubes, 1983; Zeng and Dubes, 1985). Ripley (1981) and Diggle (1983) have generalized tests for more than two dimensions.

The *unimodal null hypothesis* assumes that the p-dimensional observations are generated from a frequency distribution with one mode. Most tests for this are limited to univariate data, but Bock (1985) assessed the distributions of the mean of all pairwise similarities and the minimum total within-group sum of squared distances when the data are partitioned into a fixed number of groups. Hartigan (1988) suggested a generalization of a one-dimensional test based on the difference between the empirical distribution function and the theoretical distribution function of the unimodal distribution nearest to it. Hartigan and Mohanty (1992) introduced a test for detecting bimodality based on single linkage clustering, and assessed the distribution of a relevant test statistic under both the Poisson and unimodal null models.

The *random dissimilarity matrix model* assumes that, under the null hypothesis of absence of cluster structure, all permutations of the ranks of the pairwise dissimilarities are equally likely. Hence the approach makes use only of the ranks of the dissimilarities and might be relevant when the dissimilarities are considered to be on an ordinal scale or when analysis by a clustering algorithm that only uses ranks is contemplated. In graph theory this model is also referred to as the *random graph hypothesis* (Jain and Dubes, 1988).

This model has some serious drawbacks. It generates an unrealistic distribution of any test statistic under the null hypothesis of absent cluster structure because it ignores existing relationships in the data. For example, if the dissimilarity between the ith and jth individuals, δ_{ij}, is small then the dissimilarities δ_{ik} and δ_{jk} would be

expected to have similar ranks (for further criticism see Jain and Dubes, 1988). However, Ling (1973) pointed out that the model could be used to obtain a lower limit for the p-value, and claimed that if the random dissimilarity model was not able to detect clustering then no other null model would.

9.4 Methods for comparing cluster solutions

Comparing partitions or trees, either with each other or with data, is a common requirement in cluster validation. For example, one might hope that different subsamples of the same data set, or different methods applied to the data, would give similar results. This might be considered as an aspect of robustness. It is also possible that an external classification is available and it is wished to investigate the similarity between this and the clustering, as an aspect of external validity. Sometimes it is more appropriate to compare proximity matrices, without clustering. A number of techniques are available to compare two partitions, two dendrograms or two proximity matrices. These are now discussed.

9.4.1 Comparing partitions

Two classifications may be represented as a $c_1 \times c_2$ matrix $\mathbf{N} = n_{ij}$, where n_{ij} is the number of objects in group i of partition 1 ($i = 1, \ldots, c_1$) and group j of partition 2 ($j = 1, \ldots, c_2$). The two classifications might be two different partitions of the same data set based on different clustering methods, or one might be a clustering and the other might be some externally defined classification. The labellings of the two partitions are arbitrary. If the number of clusters in the two partitions is the same and agreement is good, the correspondence of labels is usually obvious from inspection, and one partition can be relabelled to match the other. Partitions with equal numbers of clusters can, after relabelling, be compared using a simple percentage agreement, or the kappa coefficient (see Cohen, 1960).

However, if the number of clusters differs between the two partitions, the *Rand index* (Rand, 1971) can be used, since it is based on the agreement or otherwise of every pair of n objects rather than the simple cross-tabulation of frequencies. The index computes the proportion, of the total of $\binom{n}{2}$ object pairs, that agree; that is, are either (i) in the same cluster according to partition 1 *and* the same cluster according to partition 2, or (ii) in different clusters according to 1 *and* in different clusters according to 2. The index is defined as

$$I_R = 2A/n(n-1), \tag{9.1}$$

where

$$A = \binom{n}{2} + 2\sum_{i=1}^{C_1}\sum_{j=1}^{C_2} n_{ij}^2 - \left(\sum_{i=1}^{C_1} n_{i.}^2 + \sum_{j=1}^{C_2} n_{.j}^2\right),$$

in which

$$n_{i.} = \sum_{i=1}^{C_1} n_{ij}, n_{.j} = \sum_{j=1}^{C_2} n_{ij}. \tag{9.2}$$

As Fowlkes and Mallows (1983) point out, this index tends to increase as the number of clusters increases, and the possible range of values is quite narrow. To counter these problems, Hubert and Arabie (1985) developed the *adjusted* (or *corrected*) *Rand index*; this has been recommended for general use by Milligan and Cooper (1986). This adjusted coefficient, I_{HA}, is analogous to the κ coefficient, since it measures the agreement over and above that expected by chance.

An alternative formulation for the (unadjusted) Rand index when $c_1 = c_2 = c$ is as follows:

$$I_{\mathrm{R}} = \left[T_c - \tfrac{1}{2}P_c - \tfrac{1}{2}Q_c + \binom{n}{2}\right] \Big/ \binom{n}{2}, \tag{9.3}$$

where

$$T_c = \sum_{i=1}^{c}\sum_{j=1}^{c} n_{ij}^2 - n,$$

$$P_c = \sum_{i=1}^{c} n_{i.}^2 - n,$$

and

$$Q_c = \sum_{i=1}^{c} n_{j.}^2 - n.$$

To illustrate these concepts, the values of the Rand and adjusted Rand indices are calculated for the matrix of n_{ij} values obtained from two normal mixtures classifications of the famous Fisher (1936) data on sepal and petal widths and lengths of irises. One solution assumes equal covariances, and the other allows for different covariance matrices. The two classifications are shown in Table 9.2. The agreement between the two classifications is clearly very good (Rand index 0.87), and well above that expected by chance (adjusted Rand index 0.72).

9.4.2 Comparing dendrograms

Hierarchical classifications can also be compared with each other using measures such as the cophenetic correlation or Goodman and Kruskal's γ (see Section 4.4.2).

Table 9.2 Two alternative classifications of irises into three groups, assuming equal and unequal covariance matrices.

		Clusters assuming equal covariance matrices			
		1	2	3	Total
Clusters assuming unequal covariance matrices	1	0	0	50	50
	2	2	47	0	49
	3	36	15	0	51
	Total	38	62	50	150

Cell entries are the numbers of objects classified into each of three clusters by the two methods.

Random tree models can be used to generate appropriate null distributions, in the context of both comparing dendrograms with proximity matrices and, as discussed here, comparing two dendrograms. Such models assume that all possible trees based on the n objects are equally likely, the universe of possible trees depending on whether the tree is binary, labelled or ranked (see Section 4.4.1). Where n is large, the number of possible trees generated in a Monte Carlo simulation will be enormous, and a random sample of trees may be used. A problem with this method is that each type of clustering produces a distinct type of tree, incorporating, for example, the chaining effect in single linkage. Ideally, then, the null distribution should be based on random sets of, say, single linkage trees. Hubert (1974) considers the degree of distortion which single linkage and complete linkage impose on data generated under the three null models discussed in Section 9.3 (Poisson, unimodal and random dissimilarity). Further information on tree generation is given by Gordon (1998) and Furnas (1984).

Lapointe and Legendre (1995) compared three methods of randomization – (i) labels; (ii) labels and topology; and (iii) labels, topology and heights – as used to assess the statistical significance of the cophenetic correlation. The first of these is the well-known test of Mantel (1967). The authors show that the Mantel test is too conservative and conclude that the test based on all three, the *double permutation test*, is optimal in the sense that the universe of dendrograms sampled in this way is the most comprehensive. Published tables (Lapointe and Legendre, 1992) can be used to assess the significance of correlations without actually performing the permutations. Section 8.6.1 describes an application of these methods.

The B_c coefficient of Fowlkes and Mallows (1983) is an alternative to the Rand index, also for the case $c_1 = c_2 = c$, and is defined as follows (see Equation (9.3) for definitions):

$$B_c = T_c / \sqrt{P_c Q_c}. \tag{9.4}$$

The coefficient can be used in conjunction with dendrograms by plotting its value against the number of clusters; that is, plotting the pairs (c, B_c), $c = 2, \ldots,$

$n-1$, for each pair of partitions containing c clusters obtained from two dendrograms. A series of Monte Carlo studies reported by Fowlkes and Mallows reveal the potential of this procedure for comparing classifications. Additionally, the plots appear to have the potential for selecting the appropriate number of clusters. However, if a hierarchical method has been used but only certain partitions are of interest, it may be more natural to compare these particular partitions, using the adjusted Rand index, than to compare the complete hierarchies.

9.4.3 Comparing proximity matrices

Arabie and Hubert (1996) point out that analysts sometimes inappropriately compare the clustering output (for example, dendrograms) when they are in fact interested in the correlation between the input proximity matrices. This would typically be the case if the aim of the analysis is to establish whether clustering according to one proximity measure corresponds to clustering according to another. In this case the cophenetic matrices can be thought of as containing simplified information about the underlying cluster structures, the full information about which is contained in the proximity matrices. Schneider and Borlund 2007a, 2007b) review methods for comparing proximity matrices.

Applications which investigate the association between two proximity matrices, derived from the same cases, can be found in epidemiology. An example might be when a disease is suspected to be aetiologically associated with an infectious agent, so that temporal dissimilarities may be associated with spatial dissimilarities (*time-space clustering*). Tests for association are typically derived according to the techniques suggested by Mantel (1967), Ederer *et al.* (1966) and Knox (1964); for an application to Hodgkin's disease, see Smith and Pike (1974). Chen *et al.* (1984) assessed the power of tests for time–space clustering under three alternative models for the distribution, transmission and development of Hodgkin's disease.

9.5 Internal cluster quality, influence and robustness

Internal cluster quality can be taken to refer to the extent to which clusters meet the requirements for good clusters, as defined by Cormack (1971), namely isolation and cohesion. *Robustness* refers to the effects of errors in data or missing observations, and changes in the data or methods. Similar solutions should be obtained from different data, methods or subsets of variables when the data are clearly structured. Milligan (1980) gives an example of how this can be achieved through error perturbation. For k-means and other hill-climbing techniques, different seeds for the initial clusters should not affect the cluster solutions. *Influence* refers to the deletion of a particular point and consequent changes to the cluster assignments of the other elements. In an extension of this idea, deletion of a small number of variables from the analysis should not, in most cases, greatly alter the clusters found, if these clusters are 'real' and not mere artefacts of the particular technique being used. These three topics are discussed in the following sections.

9.5.1 Internal cluster quality

Numerical measures of isolation and cohesiveness (or compactness) are generally based on indices reflecting the relative magnitudes of intra- and inter-cluster similarity. Some of these have been discussed earlier, in connection with the problem of the number of clusters (Section 5.5). The *silhouette index* was introduced earlier as a measure of cluster quality, in connection with the silhouette plot. Bailey and Dubes (1982) define indices for a particular cluster on the basis of (i) the numbers of edges in a graph between cluster members, and (ii) the number of edges between cluster members and non-members. The probabilities of these indices, given a null random graph model (see Section 9.3), are then used to produce a *cluster validity profile* for a given cluster. The uncertainty in each of the individual clusters in hierarchical cluster analysis can also be examined using the bootstrapping approach of Suzuki and Shimodaira (2006). This is available in R as the function `Pvclust`, which calculates probability values for each cluster using bootstrap resampling techniques. Two types of probabilities are available: approximately unbiased (AU) probability and bootstrap probability (BP) value. Multiscale bootstrap resampling is used for the AU probability, which has lower bias than the BP value calculated by the ordinary bootstrap resampling. The probability (*'p-value'*) is the proportion of bootstrapped samples that contain the cluster so that larger p-values indicate more support for the cluster. Kapp and Tibshirani (2007) develop a measure of cluster quality, the IGP (in-group proportion), which quantifies how often points near each other are predicted to belong to the same group (when classified to their nearest cluster). They compare the method with a number of other measures of cluster quality. Software in R, `clusterRepro`, is available through http://cran.r-project.org.

Cohen *et al.* (1977) describe a number of other useful graphical techniques for evaluating cluster analysis solutions. The first of these involves consideration of the relative tightness of a k-point group (a potential cluster) compared to other k-point neighbourhoods in the data. The $k-1$ closest neighbours of each data point are found, and then the average interpoint distance, d_i, among these k individuals over all $k(k-1)/2$ pairs is determined. Data points contained in 'real' k-point clusters should give d_i values substantially smaller than data points not in such groupings. Cohen *et al.* suggest comparing the d_is for a given k by means of a normal probability plot. A tight cluster of size j should produce j points with nearly equal d_is which are well separated from the others at the bottom of the plot for $k=j$. Figure 9.1 shows such plots for a data set, using $k=2$, 5, 9 and 14. The behaviour at the lower ends of these plots suggests the existence of a group of size of approximately $k=9$ in these data.

A further plot described by Cohen *et al.* (1977) is that of squared distances from certain cluster centroids to individuals that are near the centroid. This is useful for examining the internal cohesiveness of a cluster. Figure 9.2 shows an example of such a plot. The symbol plotted corresponds to the cluster in which the individual resides. From this plot it is clear that there is a large distance between the members of cluster A and the closest individuals that are not assigned to A. Gnanadesikan

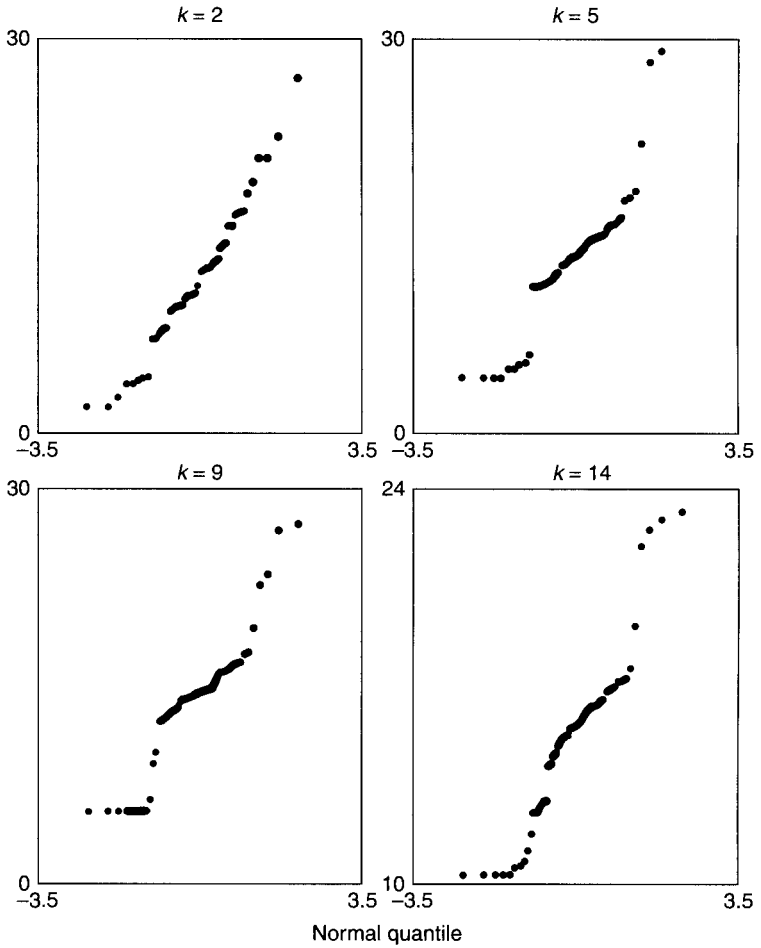

Figure 9.1 *Quantile–quantile plots to display clusters.* (Source: Cohen et al., 1977.)

et al. (1977) suggest a similar method in which the horizontal plotting positions are slightly perturbed, so that objects which would otherwise be coincident are distinguishable.

9.5.2 Robustness – split-sample validation and consensus trees

One approach to assessing the effects of perturbations of the data is the 'split-sample' method, randomly dividing the data into two subsets and performing an analysis on each subset separately. A scheme for performing split-sample cross-validation was proposed by McIntyre and Blashfield (1980), and was shown to give good results using Monte Carlo simulation. Their method involves the following steps:

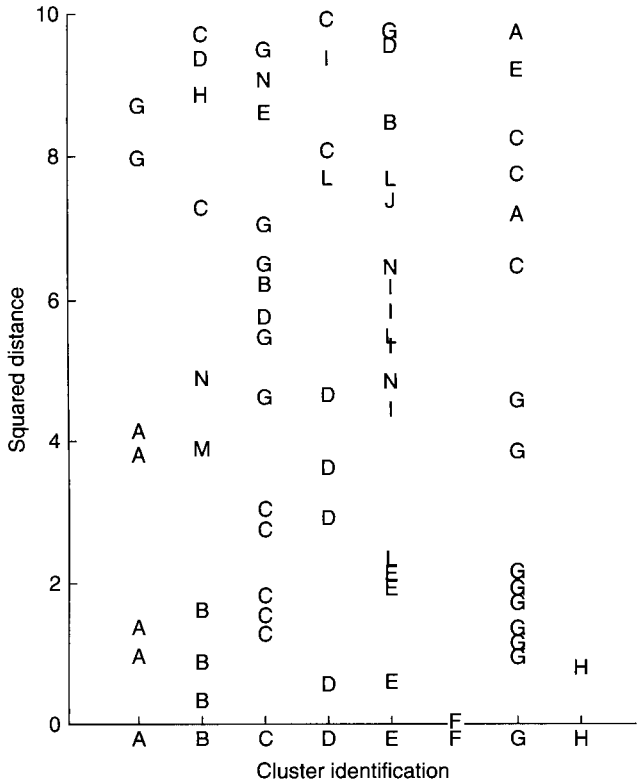

Figure 9.2 *Plot of squared distances of selected individuals from cluster centroids.* (Source: Cohen et al., 1977.)

- Divide the sample in two, and perform a cluster analysis on one of the samples, having a fixed rule for the number of clusters.
- Determine the centroids of the clusters, and compute proximities between the objects in the second sample and the centroids, classifying the objects into their nearest cluster.
- Cluster the second sample using the same methods as before, and compare these two alternative clusterings for the second sample, using, for example, the adjusted Rand index.

Breckenridge (1989) proposed a variation on this procedure in which a nearest-neighbour criterion was used to classify the second sample.

Such techniques have been criticized by Krieger and Green (1999), particularly when used for suggesting the number of clusters present, as proposed by Overall and Magee (1992). Krieger and Green showed through simulations of univariate data that validation improved for larger sample sizes, independently of the number

of clusters fitted, and also that the parity of the number of clusters could influence the level of validity inferred. For multivariate data, they concluded that the performance in determining the number of clusters degraded for unequal sized clusters and/or highly correlated data.

Where robustness is assessed by fitting different clusterings to the same data (rather than, as above, by using the same method on different subsets), it may be reasonable to synthesize the results as a *consensus clustering*. An alternative (and simpler) approach for clusterings based on different variable subsets is to reanalyse using the complete set of variables (De Querioz, 1993). Dendrograms may be combined into either a *strict consensus tree*, where each subset in the consensus tree is in every constituent tree (Sokal and Rohlf, 1981), or a *majority consensus tree* where each subset in the consensus tree is in a majority of the constituent trees (Margush and McMorris, 1981). Further types of consensus tree allow partial agreement between the constituent trees and the consensus tree. An example of a *consensus graph*, in which different clusterings are indicated by vertices in a graph, with edges joining those with a minimum level of agreement, is given in Section 9.7.1. Such graphs can indicate which trees might reasonably be combined. The formation of consensus trees is discussed in more detail by Gordon (1999).

9.5.3 Influence of individual points

Gnanadesikan *et al.* (1977) and Jolliffe *et al.* (1988) studied the deletion of selected individual points. The latter also considered the influence of the individual points on complete single and complete linkage dendrograms. Cheng and Milligan (1996) extended this approach to define a number of measures of the influence of individual points. The first part of this work compared the cluster solution of simulated data with external criterion clusters. The value of the adjusted Rand index for a clustering compared to an external grouping was used to assess what they call *cluster recovery*, with the clustering and assessment based on all n objects. Where the recovery is measured for $n-1$ observations (but based on clustering all n objects) it is termed the *adjusted cluster recovery* (relating to the depleted point).

The differences between the adjusted recovery index and that based on clustering the $n-1$ points can be used to measure the impact of the point. Where a positive difference is found, the point is regarded as a *facilitator*, whereas if it is negative it is considered as an *inhibitor*. Once an influential point is found, if it is an inhibitor, then the suggestion is to omit it. As Edelbrock (1979) has argued, in many applications there is no need to assign every point. This would usually be the case where the sample was purposive (see step (ii), Section 9.2), since the composition of the sample to be clustered is arbitrary.

Table 9.3 illustrates the calculation of these indices for a small example with one candidate influential point. In this example the influence of the point is $0.9183 - 0.7652 = 0.1531$ (a facilitator to the clustering method).

In applications where the true clustering is unknown, the comparisons must be internal using HA_{ii}. This contains no information about whether a point is a facilitator or inhibitor. For such cases, Goodman and Kruskal's γ (Section 4.4.2)

Table 9.3 Four types of influence index for a small example.

		Frequency tables			Adjusted Rand index	Cheng and Milligan name and interpretation
		Clustering method (n)				
		1	2	Total		
1 External criterion clusters	1	25	0	25	0.9200	HA_{cr} Cluster recovery
	2	1	24	25		
	Total	26	24			
		Clustering method (n)				
		1	2	Total		
2 External criterion clusters	1	25	0	25	0.9183	HA_{acr} Adjusted cluster recovery
	2	1	23	$24 = 25 - 1$		
	Total	26	$23 = 24 - 1$			
		Clustering method ($n-1$)				
		1	2	Total		
3 External criterion clusters	1	23	2	25	0.7652	HA_{ei} External influence
	2	1	23	$24 = 25 - 1$		
	Total	24	25			
		Clustering method ($n-1$)				
		1	2	Total		
4 Clustering method (n)	1	22	4	26	0.5611	HA_{ii} Internal influence
	2	2	21	$23 = 24 - 1$		
	Total	24	25			

1. Complete sample used for both clustering and index.
2. Complete sample used for clustering; candidate point omitted from index.
3. Candidate point omitted from clustering and index.
4. Two clusterings: (i) based on complete sample (method n) and (ii) omitted candidate point (method $n-1$); index computed from reduced sample. (Taken with permission of Springer-Verlag, from Cheng and Milligan, 1996.)

and the *point biserial correlation index* were suggested by Cheng and Milligan as indices for determining if an influential point is a facilitator or an inhibitor. The point biserial index is based on the correlation between the elements of the proximity matrix and a dummy variable indicating, for each pair of points, whether they were placed in the same or different groups by the clustering method. A positive (negative) value of either type of index indicates a facilitator (inhibitor). A further index, *internal influence*, compares the clustering obtained with and without a candidate point.

9.6 Displaying cluster solutions graphically

Before applying any clustering method, some graphical representation of the data should be obtained, and a number of possibilities were discussed in Chapter 2. Classical principal components analysis is commonly employed to obtain a low-dimensional mapping of the data, although this is not guaranteed to reflect any clustering present. Other, potentially more useful, ordinations may be obtained from other methods, for example those described by Sammon (1969), and the *projection pursuit* techniques discussed by Jones and Sibson (1987). In addition, multidimensional scaling techniques may be used to extract similar visual displays from a calculated proximity matrix. Some authors suggest that if these displays produce no evidence of clustering in the data, then more formal clustering procedures are not required; see, for example, Chatfield and Collins (1980).

Specialized techniques specifically designed for cluster analysis can be used. Examples of the latter have been given in earlier parts of the book: for example, banner plots (Figure 4.11) and silhouette plots (Figures 5.5 and 8.13). To illustrate the process of hierarchical clustering, as distinct from particular partitions, dendrograms are used (Figures 4.10, 4.13 and 4.14). The Kohonen self-organizing map (e.g. Figure 8.18) can be regarded as a clustering method which provides at the same time a complete visual representation of the clusters. Techniques that are designed to cluster both variables and objects with binary data may have their own specialized graphs (Figure 8.8).

Once the cluster analysis has been performed, the partition(s) found can be added to low-dimensional plots (Figure 5.4), and the minimum spanning tree (see Section 4.4.5) can indicate possible distortions in this. *Correspondence analysis* (Greenacre, 1984) can be used to produce a low-dimensional plot that shows how the clustering obtained corresponds to another classification (either a second clustering or an externally defined classification). It is generally important for interpretation to be able to associate the values of particular variables with the clusters, a simple approach being to describe the clusters by profiles, bar charts or scatterplots of pairs of variables. These three graphical approaches are illustrated by Stopford *et al.* (1991) in an application concerned with the chemical composition of decorated medieval tiles. This paper gives an example of (i) a principal components plot illustrating the compositional affinities of tile clusters; (ii) a correspondence plot showing how the compositional clusters relate to externally defined production groups; and

(iii) profile plots of the chemical compositions of selected tile clusters and related clay sources.

Leisch (2009) describes several graphical displays that can be used to visualize solutions from k-means-type clustering methods. The basis of a number of these graphics is the *shadow value*, $s(\mathbf{x})$ of each multivariate observation, $\mathbf{x}$, defined as follows:

$$s(\mathbf{x}) = \frac{2d[\mathbf{x}, c(\mathbf{x})]}{d[\mathbf{x}, c(\mathbf{x})] + d(\mathbf{x}, \tilde{c}(\mathbf{x})]}, \tag{9.5}$$

where $d[\mathbf{x}, c(\mathbf{x})]$is the distance of the observation $\mathbf{x}$ from the centroid of its own cluster, and $d[\mathbf{x}, \tilde{c}(\mathbf{x})]$ is the distance of $\mathbf{x}$ from the second-closest cluster centroid. If $s(\mathbf{x})$is close to zero then the observation is close to its cluster centroid; if $s(\mathbf{x})$is close to one then the observation is almost equidistant from the two centroids (a similar approach is used in defining silhouette plots – see Chapter 5.) The average shadow value of all observations where cluster i is closest and cluster j is second closest can be used as a simple measure of cluster similarity:

$$s_{ij} = n_i^{-1} \sum_{\mathbf{x} \in A_{ij}} s(\mathbf{x}), \tag{9.6}$$

where n_i is the number of observations which are closest to the centroid of cluster i and A_{ij} is the set of observations for which the centroid of cluster i is closest and the centroid of cluster j the second closest. The denominator of s_{ij} is taken to be n_i rather than n_{ij}, the number of observations in the set A_{ij}, to prevent inducing large cluster similarity when n_{ij} is small and the set of observations consists of poorly clustered points with large shadow values.

For a cluster solution derived from bivariate data, a *neighbourhood graph* can be constructed using the scatterplot of the two variables and where two cluster centroids are joined if there exists at least one observation for which these two are closest and second closest, with the thickness of the joining lines being made proportional to the average value of the corresponding s_{ij}. When there are more than two variables in the data set, the neighbourhood graph can be constructed on some suitable projection of the data into two dimensions; for example, the first two principal components of the data could be used. Such plots may help to establish which clusters are 'real' and which are not, as we will try to illustrate with two examples.

The first example uses some two-dimensional data generated to contain three clusters. The neighbourhood graph for the k-means five-cluster solution from the application of k-means clustering is shown in Figure 9.3. The thicker lines joining the centroids of clusters 1 and 5 and clusters 3 and 4 strongly suggest that both pairs of clusters overlap to a considerable extent and are probably each divisions of a single cluster.

For the second example we return to the pottery data previously met in Chapter 2. From previous analysis it is clear that these data contain three clusters; Figure 9.4 shows the neighbourhood plot for the k-means *four*-cluster solution in the space of the first two principal components of the data. The very thick line joining the centroids of clusters 2 and 4 suggests that the pottery in these two clusters really belongs in a single cluster.

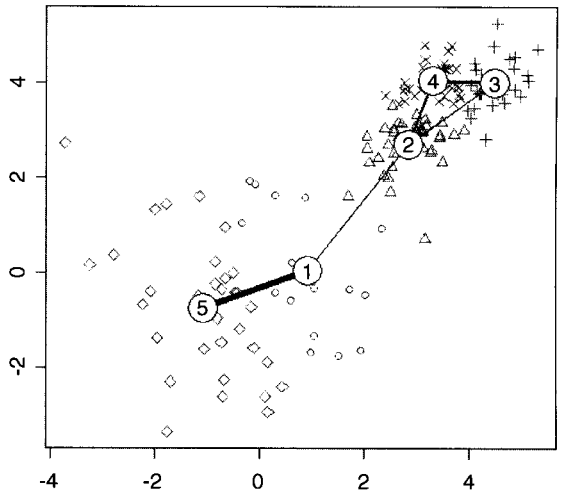

Figure 9.3 *Neighbourhood plot of k-means five-cluster solution for bivariate data containing three clusters.*

A simple graphical aid for evaluating clustering solutions, again suggested by Cohen *et al.* (1977), can be used for examining the clusters in terms of either variables used to form the clusters or other variables of interest. Here the clusters are again identified along the x-axis, and above each label the values of a particular variable are plotted, for each individual in the cluster. The median of the cluster is also plotted. The plot can then be used to compare individuals in the same cluster

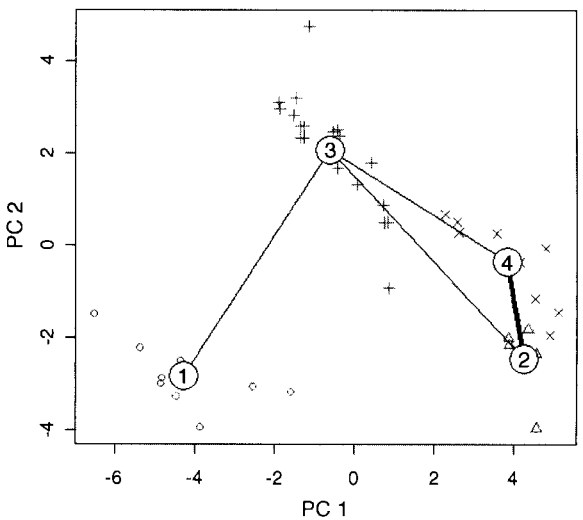

Figure 9.4 *Neighbourhood plot for the k-means four-cluster solution on the pottery data (see Chapter 2); the plot is shown in the space of the first two principal components of the data.*

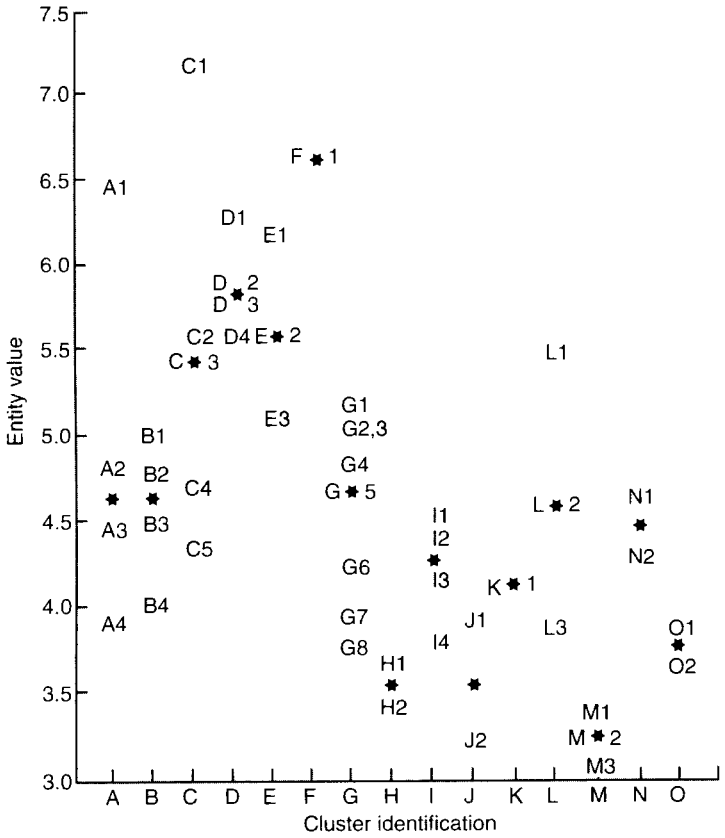

Figure 9.5 *Plot of values of a single variable for selected individuals in various clusters; *denotes the median of a cluster.* (Source: Cohen et al., 1977.)

on the variable in question, and to make multiple comparisons across clusters. An example of such a plot appears in Figure 9.5. This shows that clusters A and B are quite similar on the variable apart from one individual, Al in cluster A. Cluster E tends to contain individuals with large values on this variable and these have moderate spread, whilst cluster M has much smaller values with small spread.

A graphic for displaying clustering solutions, similar to the one in Figure 9.5 in certain respects, is suggested by Leisch (2009). Known as a *stripes plot*, this graphic is a simple but often effective way of visualizing the distance of each point from its closest and second-closest cluster centroids. For each cluster, $k = 1, \ldots, K$, a stripes plot has a rectangular area which is vertically divided into K smaller rectangles, with each smaller rectangle, i, containing information about distances of the observations in cluster i from the centroid of that cluster, along with the corresponding information about observations that have cluster i as their second-closest cluster. The explanation of how the plot is constructed becomes more transparent if we look at an actual example, and Figure 9.6(a) shows a stripes plot for a five-cluster solution on a set of data generated to contain five relatively distinct clusters. Looking first at the

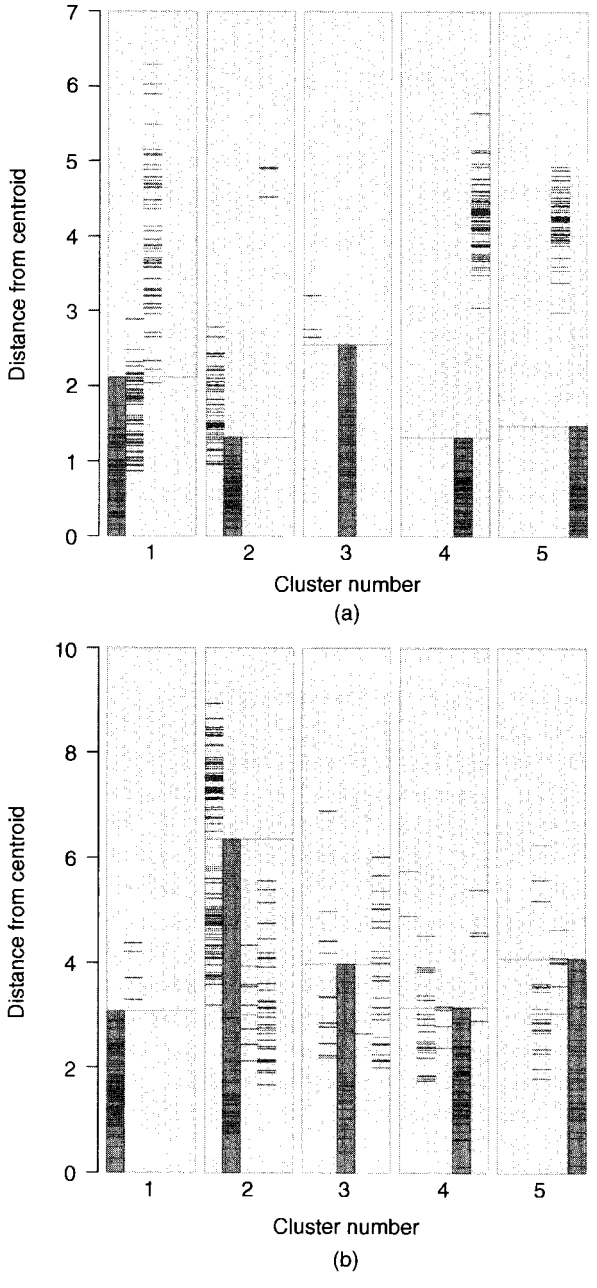

Figure 9.6 *Stripes plot for k-means five-group solution on (a) a simulated data set with five relatively distinct clusters; and (b) a second data set, where the plot suggests that the five-group solution is not appropriate in this case.*

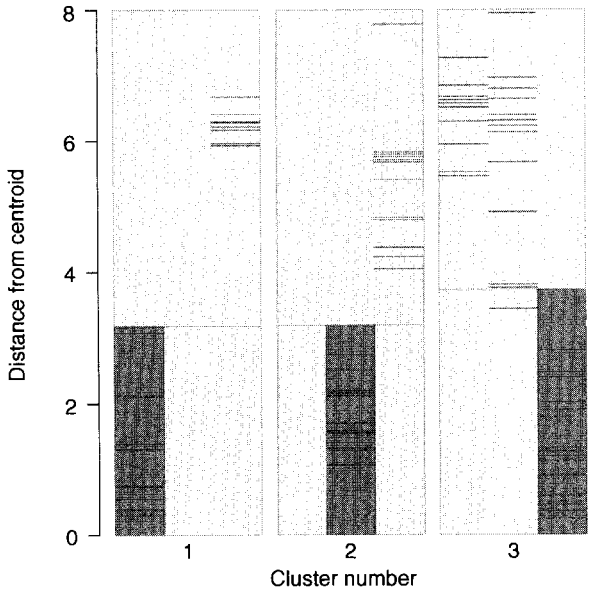

Figure 9.7 *Stripes plot for k-means three-group solution on pottery data.*

rectangle for cluster 1, we see that observations in clusters 2 and 3 have the cluster 1 centroid as their second closest. These observations form the other two stripes within the rectangle. Observations in cluster 3 are further away from cluster 1, but a number of observations in cluster 3 are at a similar distance from the centroid of cluster 1 as those observations that belong to cluster 1. Overall though, the stripes plot in Figure 9.6(a) suggests that the five-cluster solution matches quite well the actual structure in the data. The situation is quite different in Figure 9.6(b), where the stripes plot for the *k*-means five-group solution suggests that the clusters in this solution are not well separated, implying perhaps that the five-group solution is not appropriate for the data in this case. Lastly, the stripes plot for the *k*-means three-group solution on the pottery data is shown in Figure 9.7. The graphic confirms the three-group structure of the data.

All the information in a stripes plot is also available from a neighbourhood plot, but the former is dimension independent and may work well even for high-dimensional data where projections to two-dimensions lose a lot of information about the structure in the data.

Neither neighbourhood graphs nor stripes plots are infallible, but both offer some help in the often-difficult task of evaluating and validating the solutions from a cluster analysis of a set of data.

9.7 Illustrative examples

In this section, some further illustrative examples are given. The first two show how data matrices, proximity matrices and alternative hierarchical clusterings

can be compared. The third is an application using a model-based technique, which illustrates how a number of steps can be applied to reduce the complexity of data. Finally, an example shows how a combination of internal and external evidence can be used to choose between methods in genetics applications.

9.7.1 Indo-European languages – a consensus tree in linguistics

Atkinson *et al.* (2005) examined divergence time estimates for various Indo-European languages, with a focus on the date of the original parent language. Swadesh lists of words (see Chapter 3) were employed, but characters (grammatical and phonological features) were also considered in the analyses as well as words.

Two types of model were developed using ideas from evolutionary biology; this approach contrasts with previous research in glottochronology, which was based on standard hierarchical clustering methods. One type was based on the so-called 'finite sites' model (where changes in a fixed number of characters are considered); the other was based on a stochastic model of language creation, loss and splitting. MCMC sampling and Bayesian analysis were used to compare the performance of the models with different parameters and on various types of data set, including synthetic data. The overall conclusions from these two new types of model were surprisingly consistent. One of the benefits of this new analysis is the ability to assign uncertainties to estimates, since the Bayesian analysis allowed the quantification of phylogenetic uncertainty in date estimates. A simplified consensus tree is shown in Figure 9.8, in which the posterior probabilities for each branch (the degree of support for the branch) is given as the percentage of time that it appeared in the Bayesian MCMC sample.

A summary of recent developments, a discussion of the differences between this statistical analysis and other analyses from linguistic palaeontology, and an example of another type of figure, the consensus network, can be found in Nicholls (2008).

9.7.2 Scotch whisky tasting – cophenetic matrices for comparing clusterings

Lapointe and Legendre (1994) applied a hierarchical clustering method to binary characteristics of 109 Scotch whiskies, the 68 variables describing feature types such as colour, nose, body, and so on, derived from an expert's description (Jackson, 1989). Jaccard's similarity measure was used in conjunction with Ward's method, and results were compared with those from two other methods. A spatially constrained k-means approach was also used (see Section 8.5). Here we concentrate on the use of cluster comparison, rather than the clustering methods as such.

Comparisons were made between raw data matrices, distance matrices, both derived from the whisky features and geographic distances, and the cophenetic matrix representing the complete clustering. From the clustering with 2, 3, 6 and 12 group partitions, binary matrices were computed, with entries equal to 1 when two whiskies were in the same partition, and 0 otherwise. Since Scotch whiskies

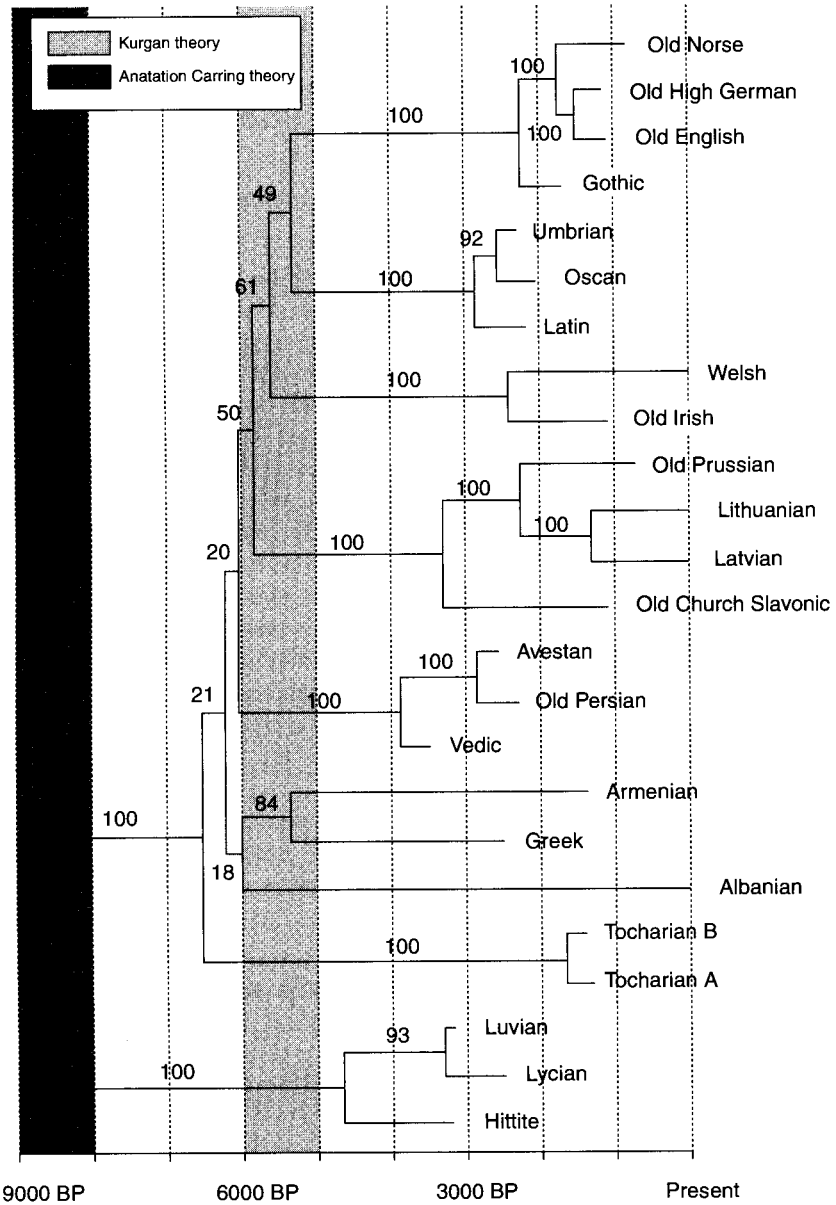

Figure 9.8 *A consensus tree of Indo-European languages (from Atkinson, et al., 2005). The shaded bars represent two hypotheses about the time of common origin of the languages. Figures above branches indicate the degree of support for that branch.*

are known to derive some of their characteristics from water, soil, air quality, and so on, which vary by region, two geographic matrices were also computed. The first of these was a matrix of distances between the distilleries based on map coordinates. The second was a proximity matrix containing binary entries: 1 if two distilleries were in the same region (Highland, Lowland or Islay) and 0 otherwise.

The authors computed cophenetic correlations to compare the clustering with the geographical classification. The significance of this was tested using both random permutation of labels (Mantel test) and the more realistic double permutation of labels, topology and fusion levels (see Section 9.4.2). A further detailed analysis of the feature types (colour, nose, body, palate and finish) was undertaken to assess the congruence of the five features. The analysis involved comparing raw data tables (for which canonical correlations and associated permutational test statistics were derived), distance matrices and dendrograms. Table 9.4 shows the results of these tests of significance. Entries in the tables are correlations (standardized Mantel statistics) or coefficients based on canonical correlations, indicated as significant at $p = 0.01$ or 0.001 according to the appropriate test. In Table 9.4(a), the authors interpreted the significant results obtained for the 6- or 12-group partition as indicating that geographic information seems more relevant in defining smaller clusters of closely related whiskies.

The remaining subtables relate to feature types. Some discrepancies between the comparison tests among feature types are explained in terms of the loss of information when moving from data to distance matrix, and from distance matrix to dendrogram. Figure 9.9 shows the significant and nonsignificant results in Table 9.4 (b)–(d) as a consensus graph, and from this it is evident that only three comparisons were congruent at all levels: palate–nose; nose–colour and colour–body; finish was always left unconnected with other features.

9.7.3 Chemical compounds in the pharmaceutical industry

Guttiérrez *et al.* (1999) described the application of a model-based hierarchical method to a large data set of binary 'fingerprints' identifying the molecular structure of chemical compounds used in the pharmaceutical industry. The fingerprints are an abstract representation of molecular patterns in the form of a sequence of bits of length 1024. Their account of the clustering process illustrates some of the steps described in Section 9.2

- The choice of objects (molecules), variables (binary fingerprints) and an appropriate proximity measure, followed by multidimensional scaling to reduce the dimensionality of the data.
- The choice of statistical model, and hence method of analysis – in this case normal mixtures.
- Consideration of the number of clusters.
- The assessment of the stability of the clusters by comparing the results for different MDS solutions; the use of another similar clustering method to check

Table 9.4 Correlations and comparison tests between clustering of Scotch whiskies and geography, and between different whisky feature groups.

(a) Clustering versus geographic distance matrix based on map coordinates

		Partitions into:			
Standardized Mantel statistics	Complete dendrogram	2 groups	3 groups	6 groups	12 groups
Mantel test	0.031**	−0.027	0.007	0.065**	0.064**
Double permutation test	0.031*				

(b) Feature types (comparisons between raw data matrices)

Redundancy coefficient	Nose	Body	Palate	Finish
Colour	0.147*	0.121*	0.151	0.152
Nose		0.121*	0.164**	0.153
Body			0.105	0.116
Palate				0.129

(c) Feature types (comparisons between distance matrices)

Standardized Mantel statistics	Nose	Body	Palate	Finish
Colour	0.032*	0.067**	0.027	−0.011
Nose		0.042*	0.074*	0.009
Body			0.054*	0.018
Palate				−0.012

(d) Feature types (comparisons between cophenetic matrices)

Standardized Mantel statistics	Nose	Body	Palate	Finish
Colour	0.048*	0.064**	0.046*	−0.025
Nose		0.021	0.071**	−0.001
Body			0.0510**	0.008
Palate				0.002

* $0.001 < p < 0.01$.
** $p < 0.001$.
(Taken with permission of the publisher, Blackwell, from Lapointe and Legendre, 1994.)

for robustness; comparison against previous findings and assessment of the usefulness of the solution.

The Jaccard coefficient (the proportion of bits present in either compound which were common to both) was used as a measure of similarity between two compounds (see Section 3.2), and this was subsequently converted to a distance measure. Because of the size of the distance matrix (460 320 elements), metric

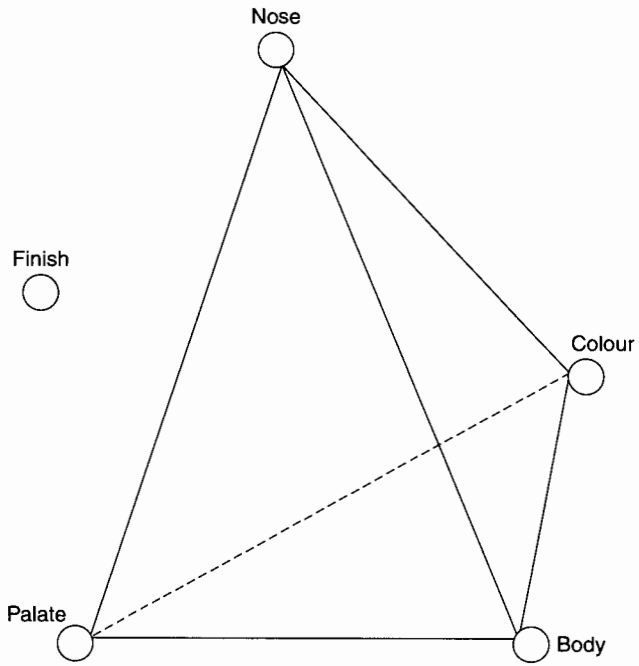

Figure 9.9 *Consensus graph of whisky feature types. Thick edges depict congruent relationships at all three levels (raw data tables, distance matrices and dendrograms); unbroken thin edges are congruent at two levels, whereas the broken edge indicates a relationship which is significant at one level only; see also Table 9.4.* (Taken with permission of the publisher, Blackwell, from Lapointe and Legendre, 1994.)

scaling was performed on a number of subsamples and then applied to the rest of the data to produce coordinates in a low dimension (five in this case). Before clustering, the data transformation procedure was as follows: binary data → similarity → distance → continuous data. Each scaling (from a different subsample) produced slightly different sets of continuous data. However, they all, when displayed on the first two (of the five) axes, showed an obvious clump of points in the middle, with ellipses emanating from the centre (see Figure 9.10); plots on other axes were similar.

The elliptical nature of the apparent clusters indicated the advisability of using a method that can identify such clusters. The large sample sizes and low dimensionality meant that normal mixtures models with different orientations and sizes (but the same shape; criterion S^* in Table 6.1) could be considered without computational problems. They were fitted using the method of Banfield and Raftery (1993) which suggested eight groups, whichever set of scaled data was used. Further examination of the data suggested the subdivision of the largest group into two, to give a final nine-cluster solution.

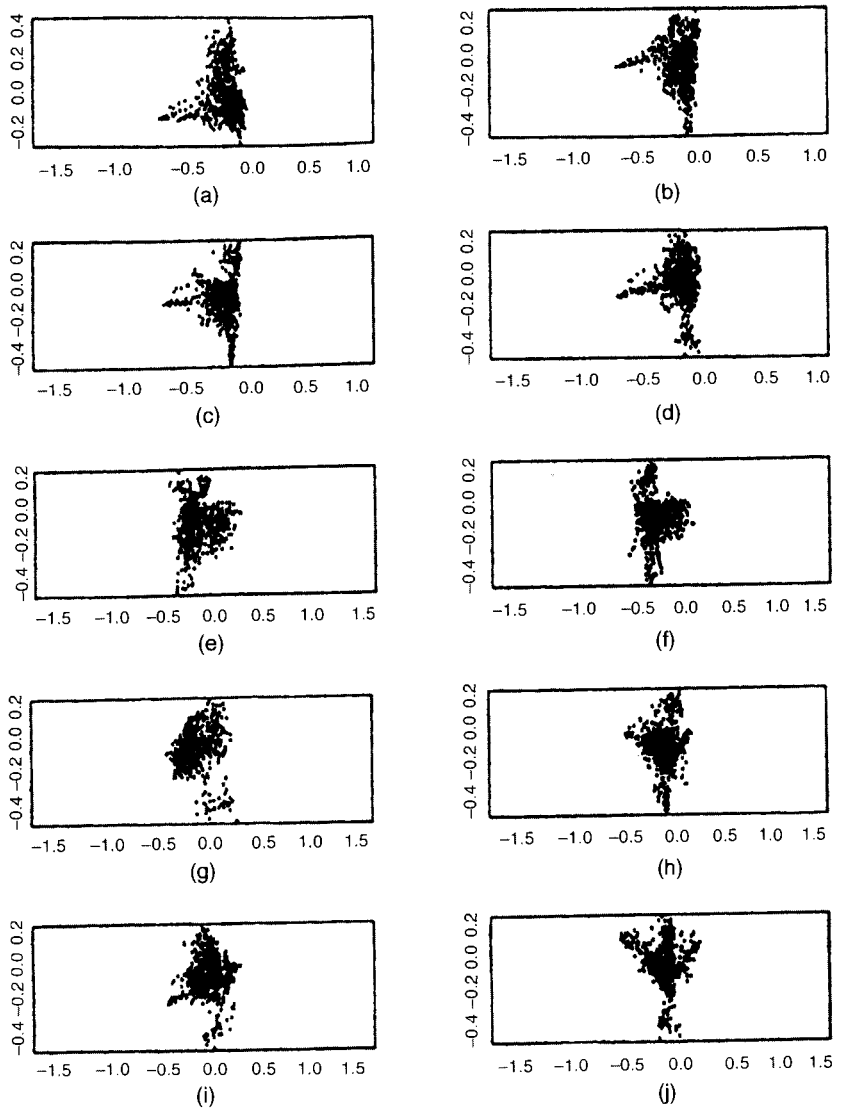

Figure 9.10 *Pairwise scatterplots of pharmaceutical data, showing elliptical clusters; axes 1 vs. 2–5; 2 vs. 3–5; 3 vs. 4–5; 4 vs. 5 in order (a) to (j) in a five-dimensional metric scaling based on a subsample of the data.* (Taken with permission of the publisher, Blackwell, from Guttiérrez et al., 1999.)

Table 9.5 shows the eight-cluster solutions after MDS using two different subsamples. There appears to be a good correspondence between these results, showing robustness to small changes in the input data. The agreement was 75% (after making the obvious correspondence of clusters), the Rand index was 0.79 and the adjusted Rand index was 0.51.

Table 9.5 Model-based clustering using MDS based on two subsamples of pharmaceutical data[a]).

Clusters when scaling based on subsample 2	Clusters when scaling based on subsample 1							
	1	2	3	4	5	6	7	8
1	5	0	0	5	0	0	0	**15**
2	9	6	**15**	0	0	2	2	0
3	0	0	6	**38**	0	2	0	0
4	0	**74**	0	0	0	0	0	0
5	59	0	0	0	**0**	0	0	0
6	1	0	1	0	0	0	**158**	3
7	40	0	0	0	0	**40**	0	0
8	**380**	8	5	0	40	12	11	23

[a]Corresponding clusters shown in bold.

The data were also analysed using a mixtures method, developed by Cheeseman and Stutz (1995), that can handle large data sets; in five dimensions the method worked well and gave reasonable agreement with the previous analysis, although the number of clusters suggested was higher for this second analysis.

Many of the clusters were confirmed by subject matter specialists to correspond to already known compounds, and the results proved to be useful in identifying groups of similar compounds for further pharmaceutical testing.

9.7.4 Evaluating clustering algorithms for gene expression data

In a study by Datta and Datta (2006), six clustering algorithms, UPGMA (average linkage hierarchical), *k*-means (partitioning), DIANA (divisive hierarchical method), FANNY (a fuzzy method), model-based (mixtures of Gaussians) and SOM (self-organizing map: a type of neural network – see Chapter 8) were evaluated on two publicly available genetic data sets, one on breast cancer patients and one from microarray data on yeast. The methods are those that might be considered reasonable choices for this type of data and have been mentioned earlier in this book. Two types of validation measures are proposed (each with two variants), the first 'statistical' (VS1 and VS2) which are based on statistical stability when a unit is deleted from the gene expression profile, and the second 'biological' (VB1 and VB2), where it is assumed that gene pairs with similar biological functions can be identified, and should appear in the same cluster. The two variants 1 and 2 use proportion of overlap of different clusters containing similar gene pairs, and average distance between clusters with and without the deleted gene, respectively. The aim was to assess the performance on statistical and biological validity, both separately and together, for different numbers of clusters. An illustration is given in Figure 9.11.

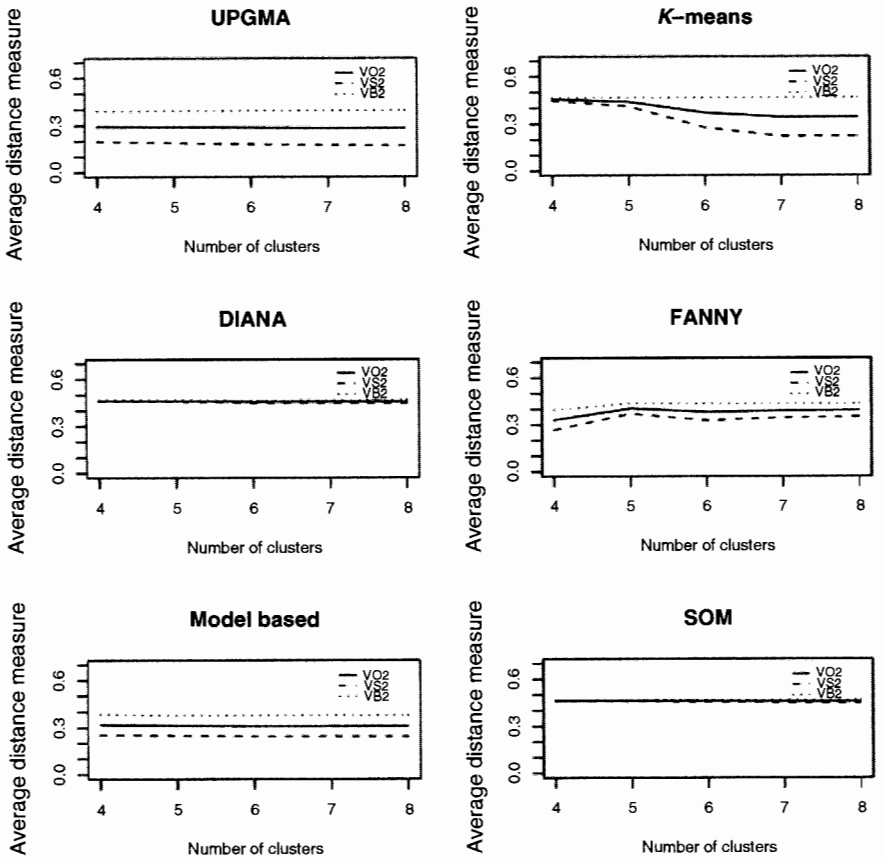

Figure 9.11 *The performance for various clustering methods and numbers of clusters on the cancer data, based on the average distance between the clusters containing similar gene pairs (dotted – VB2), and with and without each unit (dashed – VS2), and VO2 (solid – the average of VB2 and VS2). From Datta and Datta (2006).*

The model-based mixtures method and average linkage did well for this data set and measure, but taking into account both data sets and types of measures, no method was an overall winner. Although the results were inconclusive, the authors show that a fairly simple and systematic approach is possible in assessing clusters. It should be noted, however, that the 'biological' validation measure depends on additional information on which genes share functionality, which is not always available, and also that the relative weight of the statistical and biological measures is subjective (here they were equally weighted).

A much larger study (Souto *et al.*, 2008) did, however, come to firmer conclusions. The authors compared 'classic' methods and those designed to take advantage of the specific nature of gene expression data on 35 publicly available data sets (see http://algorithmics.molgen.mpg.de/Supplements/CompCancer/).

One characteristic of gene expression data is, as the authors point out, the high dimension (compared to say clustering genes themselves, which typically are described by a lower number of states). The criterion here was the recovery of known cancer types – thus using a 'biological' rather than 'statistical' criterion (in the terms used by Datta and Datta). Methods considered were hierarchical clustering with single, complete and average linkage, k-means and mixture of multivariate Gaussians, and more recent methods of spectral clustering (Ng *et al.*, 2002) and a nearest-neighbour-based method (Ertoz *et al.*, 2002). Proximity measures considered were Pearson's and Spearman's correlations, cosine and Euclidean distance, the latter with three types of pre-processing: standardization, scaling and ranking. The adjusted Rand coefficient was used to assess agreement with the best partition and the partition obtained by assuming that the correct number of clusters was known. The conclusion was that a mixture model, followed by k-means, was optimal in recovering the known clusters.

9.8 Summary

The methods of cluster analysis can be valuable tools in the exploration of multivariate data. By organizing such data into subgroups or clusters, clustering may help the investigator discover the characteristics of any structure or pattern present. Applying the methods in practice, however, requires considerable care if over-interpretation of the solutions obtained is to be avoided. Much attention needs to be given to questions of cluster validity, although such questions are rarely straightforward and are full of traps for the unwary. Simply applying a particular method of cluster analysis to a data set and accepting the solution at face value is in general not adequate.