# Chapter 10

# Cluster analysis

## 10.1 Introduction

### 10.1.1 Main ideas

Cluster analysis is a generic term for a range of methods aimed at identifying groups in a set of data. It is probably the most widely used multivariate method in archaeology. To give only a few examples, cluster analysis has been used to group artifacts on the basis of their dimensions or chemical compositions; assemblages on the basis of the similarity of their profiles; and to spatial clustering on the basis of the location of artifacts in space.

Many methods of cluster analysis result in the identification of $G$ groups, with the hope that cases in a group are similar to each other and dissimilar from cases in other groups. This introduces the idea of (dis)similarity, which is crucial to an understanding of how many methods of cluster analysis work. Many measures of (dis)similarity can be defined, contributing to the many methods of cluster analysis available. Another reason for this proliferation is that, given a measure of (dis)similarity, a large number of clustering algorithms have been proposed for the subsequent grouping exercise. This chapter discusses the most common methods used in archaeological practice.

In some ways these have not changed much since the earlier days of exploration in the 1960s and 1970s (Doran and Hodson, 1975; Hodson, 1969, 1970; Bieber *et al.*, 1976), which are often more interesting in the way cluster analysis was exploited that it commonly is now. A lot of research has subsequently been undertaken on more complex methods but most have, as yet, found limited archaeological application. Some of these are discussed in Section 10.3. Before discussing methods in more detail a small example is provided to illustrate ideas and issues.

### 10.1.2 Example – Blue medieval window glass

The data given in Table B.16 in Appendix B were originally published by Cox and Gillies (1986) and have subsequently been reanalyzed by Baxter (1989), Bell and Croson (1998) and others, usually for the purpose of methodological illustration. The measurements, the chemical composition in percentages for 11 oxides, are for 27 specimens of blue medieval glass from the windows of York Minster and elsewhere. It was of interest to see if the Minster glass was distinct from other sources. A basic cluster analysis is shown in Figure 10.1 and can be obtained in one line of code using

```
plot(hclust(dist(scale(york)), method = "a"))
```

though it makes for easier reading if broken down into its component parts. To herald discussion in Section 10.2 some of these components are discussed now.

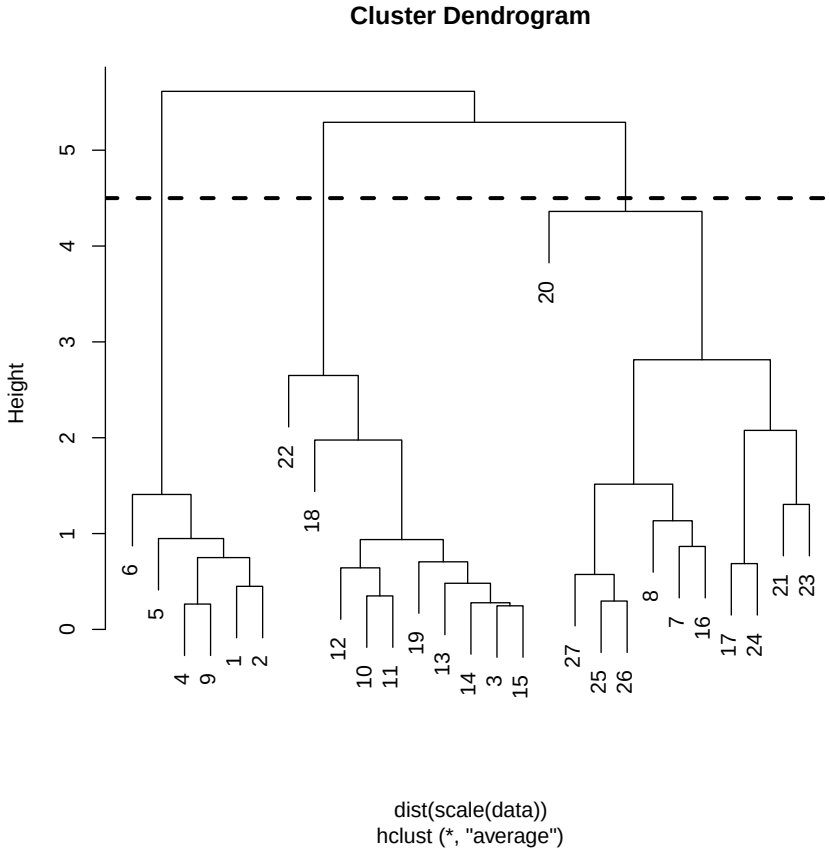


Figure 10.1: *An average-link cluster analysis for the standardized medieval glass compositional data from Table B.16.*

The `method = "a"` argument specifies that the average-link method of clustering is to be used. This is a *hierarchical* clustering method of which several alternatives are available in `R`. The same considerations concerning data transformation occur with cluster analysis as with principal component analysis (Section 7.2); thus `scale(york)` standardizes the $27 \times 11$ data matrix (named `york`). The `dist` argument computes Euclidean distances (Section 7.3) between the rows of scaled data; `hclust` executes the analysis; and `plot` displays the result in the form of a *dendrogram*, as in Figure 10.1.

This can be thought of as a tree consisting of branches and leaves which are the individual cases. The idea is to identify branches that contain leaves that are similar to each other, in terms of the distance between them, and at some distance from leaves associated with other branches. This is not always easy and not always possible. The most common practice is to cut the tree at some chosen height to identify distinct branches. A cut that gives three clusters is shown at a height of 4.5. The choice is subjective; if a cut is made at a slightly lower height four clusters are obtained, once containing a single case, 20, that may be an outlier. Splitting the coding above as

```
clus <- hclust(dist(scale(data)), method = "a"); plot(clus)
```

allows the object `clus` to be interrogated. Thus, `cutree(clus, h = 4.5)` identifies cluster membership for the cut at a height of 4.5, as follows,

```
 1 1 2 1 1 1 3 3 1 2 2 2 2 2 2 3 3 2 2 3 3 2 3 3 3 3 3
```

which is useful for labeling purposes in further analysis. It is possible to specify the number of clusters required, using `k = 3` rather than `h = 4.5`.

Checking the validity of a proposed cluster is not always easy. Clustering algorithms are designed to identify clusters even when they do not exist. Given the clusters, a simple method of assessing their integrity is to use labeled principal component plots. This is shown for a scatterplot matrix of the first three components in Figure 10.2, where the three clusters are mostly clearly distinct. Clusters 1 and 2 are very tightly defined, the former in particular, apart from one case that is outlying relative to the rest of the cluster. Cluster 3 is rather more dispersed but plots coherently on the first two components.

The default output from the `plot` command is invaluable for a quick look at the data, but some sort of enhancement is desirable for presentational and interpretive purposes. Figure 10.3 illustrates some possibilities. Default titles have been removed or replaced so that the figure is more informative. Readers familar with applications of cluster analysis may not be familiar with the default style of presentation used in `R` in Figure 10.1. Figure 10.3 may be a more familiar representation, where all the leaves 'descend' to a base of zero. This is obtained by including the argument `hang = -1` in the `plot` command.

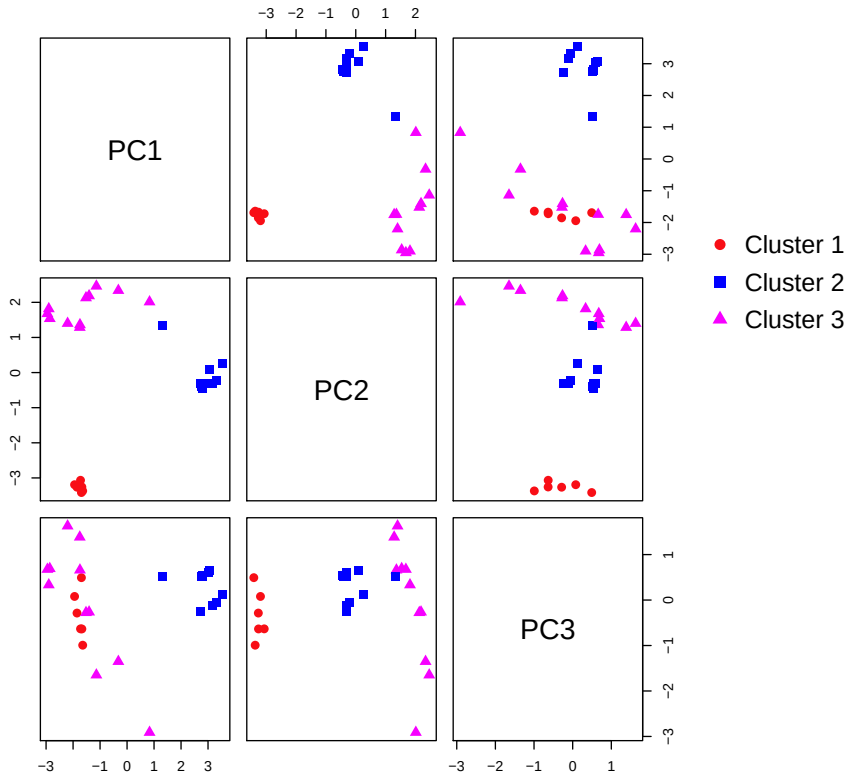


Figure 10.2: *A scatterplot matrix for the first three components of a PCA of the medieval glass compositional data from Table B.16, showing cluster labeling from an average-link cluster analysis of the data.*

The coloring requires more explanation. The default labeling in R is by case number. This can be replaced by other text, such as cluster identifications (1, 2, 3). With many observations the labeling can be unreadable unless corrective action is taken, by splitting the dendrogram into component parts, for example. The use of different colors and symbols as labels can make it easier to read the dendrogram. Color labeling is illustrated in Figure 10.3. It is uninformative in this example because the dendrogram itself defines the labeling. It is useful, however, for looking at the extent to which other methods of clustering reproduce the results. This is illustrated in Section 10.2.2 .

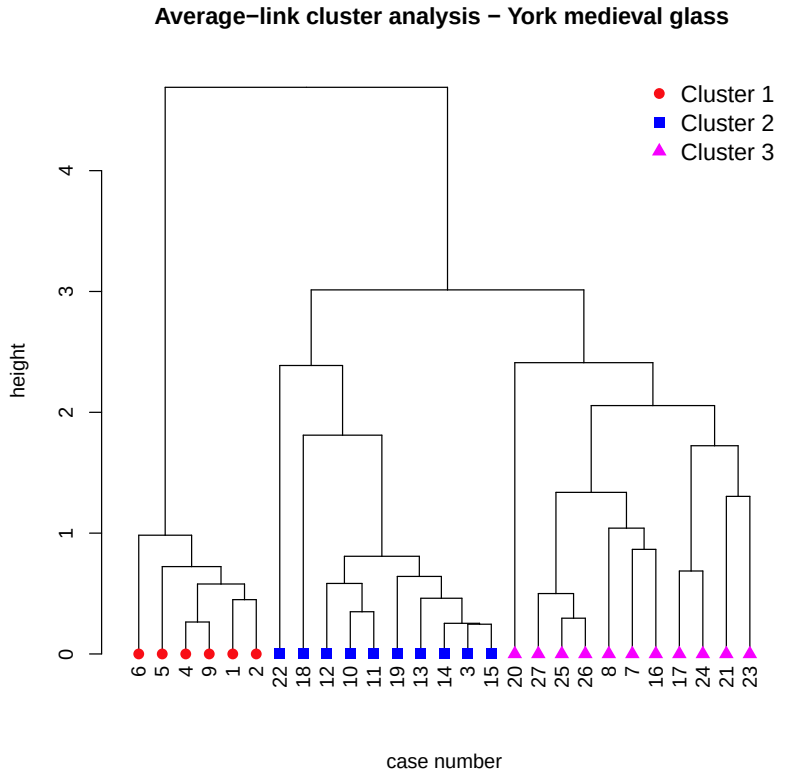


Figure 10.3: *An average-link cluster analysis for the standardized medieval glass compositional data from Table B.16 - an enhanced version of Figure 10.1.*

## 10.2 Hierarchical Clustering

### 10.2.1 The most commonly used methods

Hierarchical agglomerative methods of cluster analysis are those most commonly used in practice. Each case is initially treated as a single cluster so there are $n$ in all. The two most similar cases are merged to form a cluster of two cases, giving $(n-1)$ clusters. Thereafter, clusters are successively merged (treating single cases as clusters) on the basis of which pair is most similar at any stage. Eventually all cases are merged into a single cluster. It is possible to start by assuming that all cases belong to a single cluster and then successively split clusters up, one case at a time, until all cases are distinct. This method, hierarchical divisive clustering, has had comparatively limited use, and will not be considered further.

To merge clusters a measure to determine how similar clusters are is needed. Similarity can be defined in different ways. In *single-link* cluster analysis the

similarity of two clusters is measured by the smallest distance between two cases, one from each cluster. The two clusters merged are those for which this smallest distance is smallest. In *complete-link* cluster analysis, similarity is defined by the largest distance between two cases, one from each cluster, the clusters being merged for which this largest distance is smallest.

Single-link cluster analysis is rarely used because it tends to produce uninterpretable results unless the structure is obvious. It is sometimes useful for detecting outliers. A criticism of both single- and complete-link clustering is that the measure of similarity between clusters depends only on two cases, and fails to take account of group structure. *Average-link* cluster analysis attempts to overcome this problem by defining similarity between clusters as the average distance between all possible pairs of cases, one from each cluster. It has probably been the most widely used method of cluster analysis in archaeology. Ward's method (Section 10.3) also takes group structure into account.

The results from a hierarchical cluster analysis need to be validated and interpreted. This is usually done using a dendrogram, useful in conjunction with PCA. Cases that merge at a low level (e.g., 4 and 9 in Figure 10.1) show a high level of chemical similarity. The appearance of a dendrogram depends on the style of presentation, choice of method, and the distance measure used.

### 10.2.2 Example – Levantine glass compositions

The York data used so far is not especially suitable for exploring the issues raised above. The structure is obvious and recovered by all the methods mentioned. For further illustration a $67 \times 5$ data matrix showing the compositions of Levantine glass found at primary glass-production sites in Israel, from the first centuries AD, is used (Table B.17).

A Ward's method analysis using standardized data is shown in Figure 10.4. The appearance is 'cleaner' compared to the average-link analyses previously presented, making it easier to select a level of clustering to work with. A cut at a height of 7 produces seven reasonably convincing looking clusters. A six-cluster 'solution' is also defensible; a three-cluster solution ignores some of the structure in the lower part of the plot. The default dendrogram configuration from `R` is used but this is not very evident from the plot. This is a function of the way Ward's method can tend to work, where the depth to which the leaves hang look mostly the same.

The exception to this comment is one case from Cluster 4 which, as will be seen later, is an extreme outlier. Ward's method can be poor for outlier detection. This is a consequence of the 'model' that implicitly underpins the method and is discussed further in Section 10.3.

In a sense single-link analysis is at the opposite pole to Ward's method. The cluster analysis for single-link for the Levantine data is shown in Figure 10.5.

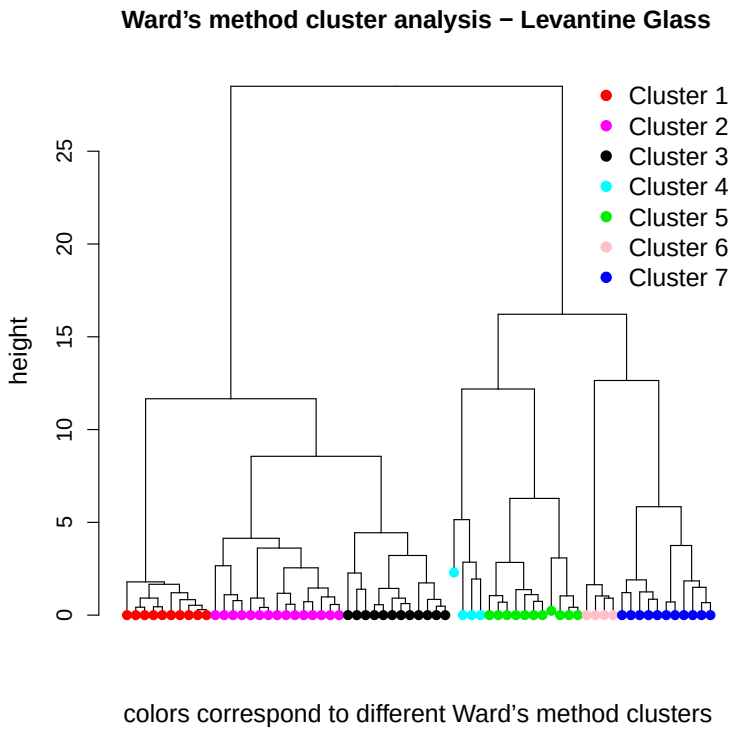


Figure 10.4: *A Ward's method cluster analysis for the standardized Levantine glass compositional data.*

It should be emphasized that the cluster identifications are those derived from the Ward's method analysis. A reason for the lack of use of single link in the archaeological literature is evident from the plot, and this is the phenomenon called 'chaining'. This arises because otherwise distinct clusters can be linked because of the effect of a small number of cases that are intermediate between the otherwise disparate clusters. In its purest manifestation the dendrogram will have a 'staircase' like appearance that makes cluster identification impossible without the use of externally derived cues. The dendrogram in the figure is not quite that bad but apart from some outliers to the left and two small groups there is no obvious structure. If the 'cues' provided by the clustering from the Ward's method analysis are used it can be seen that apart from Cluster 6, and the partial exception of Cluster 1, there is no real match, with cases from different clusters scattered throughout the dendrogram.

The contrast with the average-link cluster analysis in Figure 10.6 is more interesting. Average-link is possibly the most widely used, in archaeology, of the available methods. There are several reasons for this; is is the default in several

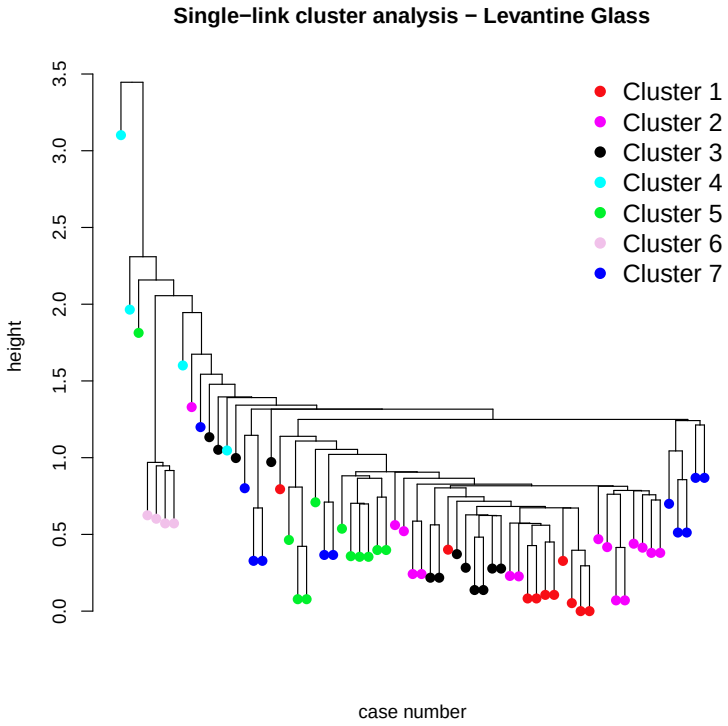


Figure 10.5: *A single-link cluster analysis for the standardized Levantine glass compositional data.*

popular software packages; it was 'promoted' from an early stage once archaeology engaged with quantitative ideas (e.g., see Doran and Hodson, 1975: 177); and (one hopes) practitioners have found it useful.

Coming to the dendrogram 'cold', without additional cues, interpretation is not straightforward. The eleven cases to the left can be treated as 'outlying' and include the two small Clusters 4 and 6. Visually cutting the dendrogram at between 2 and 3 (allowing the cut-height to vary) suggests three clusters. Stray cases apart, that to the right can be identified with cluster 5 from the Ward's method analysis, and that to the left with Cluster 7. The larger central cluster could be cut at just below a height of 2 to give a subdivision of three clusters, one of which can be identified with Cluster 1 from the Ward's method analysis. There is a small cluster of six cases, all from Cluster 2, with the remaining cluster mixing cases from the Ward's method Clusters 2 and 3.

Thus the correspondence between the Ward's method and average-link results is reasonable, with 5/7 of the clusters from the former method blocking together on the dendrogram for the latter method. The average-link analysis might thus

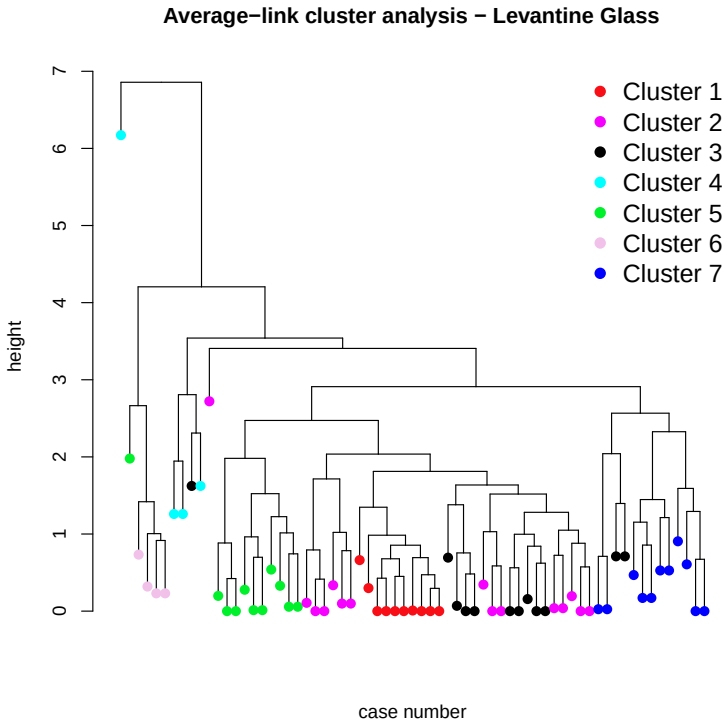


Figure 10.6: *An average-link cluster analysis for the standardized Levantine glass compositional data.*

be regarded as producing a more nuanced analysis than Ward' method.

Just because the different methods produce moderately similar results doesn't mean they are 'right'. It is advisable to check on this and PCA is one way of doing so[1]. Figure 10.7 shows plots based on the first three PCs.

In the plot of the first two components two outlying points below the bottom of the plot, which were from Cluster 4, are not shown, for easier reading of the rest of the plot. Essentially Cluster 4 consists of outliers, so that it plots separately but not coherently. Cluster 6 is a small group that plots coherently on both plots and is 'extreme' relative to other clusters. A stray case apart, Cluster 1 plots separately on the plot for the first and third components, and much the same can be said for Cluster 7 on the first two components. Cluster 5 is rather less compact than either cluster analysis might suggest. It can be largely separated from other clusters, though not perfectly. It does, however, separate out on a plot of the third

[1]In practice exploratory analyses would be carried out before a cluster analysis; PCA can be used for this. It may reveal outliers that one could consider omitting from the cluster analysis, or obvious clusters that can be separated out before undertaking further analysis.

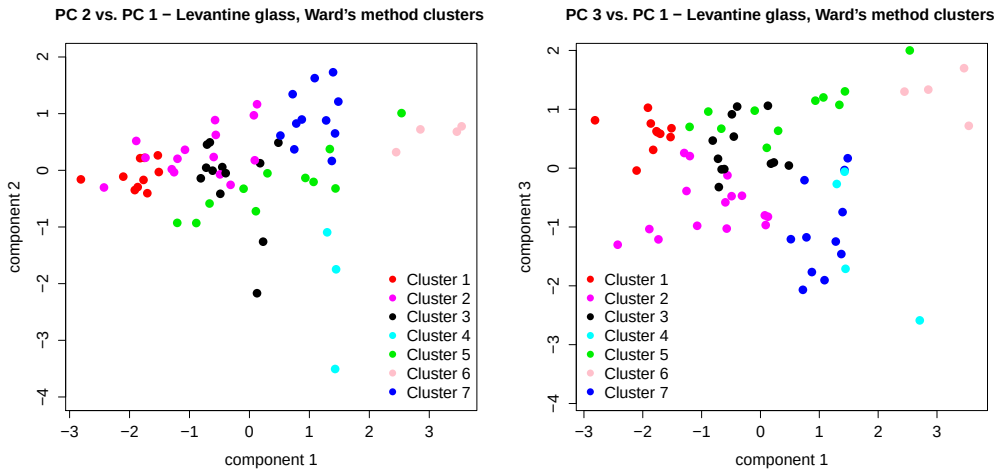


Figure 10.7: *Principal components plots for the standardized Levantine glass compositional data, labeled after the Ward's method clustering. Two outlying values from Cluster 4 are not shown in the left-hand plot.*

and fourth components (not shown).

This leaves Clusters 3 and 6 which were not especially well-separated on the average-link dendrogram, the same being true for the plot on the first two components. One case from each cluster apart they separate out on a plot of the first and third components, though they are contiguous. The plot on the first two components suggests that two cases from Cluster 3 are fairly clear outliers.

Overall the PCA suggests that the clustering produced by Ward's method is acceptable, provided one examines more than the first two components. The average-link cluster analysis, while confirming much of what can be inferred from the Ward's method analysis, fails to distinguish between two Ward's method clusters that the PCA shows can be distinguished. Both methods of cluster analysis provide little information on the coherence or otherwise of clusters. Ward's method is of little use for detecting outliers; average-link is much more satisfactory for this. The overall message is that a combination of both methods of cluster analysis, allied to checking using PCA, is a much better way of interrogating the data than relying on a single method of cluster analysis (which many publications give the impression of having done so).

It remains to ask if the sites can be distinguished. There is an imbalance between the sample sizes of 53 and 14. They are not readily separated; analyses are not shown, but the best that can be done is with the second and third PCs which separate out 9/14 of the smaller sample. This can be examined using PCA without recourse to cluster analysis which can, however, be useful for identifying

sub-groups within site assemblages if sites plot separately. The small Cluster 6 comes from the site with the smaller sample size.

## 10.3 Ward's method and model-based methods

### 10.3.1 Ward's method

Ward's method is an exception to the generalization that most commonly used methods of cluster analysis in archaeology are just grouping algorithms with no firm basis in statistical theory. These methods have been widely used because they have seemed sensible to the people who devised them, and have found favor with practitioners. Statisticians have been less impressed (see Cormack, 1971, for an early and damning review from a statisticians perspective) and this has led more recently to the development of *model-based* clustering methods. Ward's method is discussed in further detail here, partly to introduce some of the ideas used in model-based and other methods.

In contrast to the other linkage methods, Ward's method attempts to optimize an explicit objective function, $S_G$. All the agglomerative methods discussed suffer from the drawback that once a merge is made it cannot be undone. Ward's method, as usually applied, is no exception and the word 'attempts' was used above because often $S_G$ will not be optimized. That is, given any specific partition into $G$ clusters produced by Ward's method, it may be possible to improve $S_G$ by relocating cases between clusters. This is the basis of *k-means* methodology (Section 10.3.3).

Ward's method was popular in the 1970s and 80s, partly because it was the default in `CLUSTAN`, one of the earliest software packages designed specifically for cluster analysis. Applications of the method often use squared Euclidean distance as a dissimilarity measure. This distance forms the basis of measuring the variability *within* a cluster. At any given stage of clustering let $G$ be the number of clusters. For a single case, $i$, and single cluster the overall closeness to the cluster centroid, $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p)$, can be measured by

$$\sum_j (y_{ij} - \bar{y}_j)^2$$

where summation is over the $p$ variables. Summing this over the $n_g$ cases in cluster $g$ gives $S_g$, a measure of the 'compactness' of the cluster, and summing over $g$ gives $S_G$ a measure of the compactness of the clustering[2]. Any merge in the clustering process will increase $S_G$ and the merge is chosen for which this increase is least.

[2]For a singleton cluster (i.e. consisting of a single case) which predominate in the early stages of clustering, $n_g = 1$ and $S_g = 0$. It is easy to see that merging two non-identical cases will increase $S_G$ and this is generally true regardless of cluster size.

### 10.3.2 Model-based clustering

It has already been noted that Wards method can suggest clusters quite clearly, even when none exist. This behaviour can be understood by viewing Ward's method as a special case of a *model-based* method. Specifically, Ward's method will tend to produce (hyper-)spherical clusters of the same size. It can be shown to be an 'optimal' method if the assumption that clusters have a spherical normal distribution of equal size is correct. The method will tend to impose this kind of structure on clusters even if the assumption is not true.

Model-based methods were developed partly in response to the lack of theoretical justification for more heuristic methods. Such models depend on assumptions; it is unfortunate that archaeological data rarely satisfy the assumptions for Ward's method to be optimal, but the theory also explains the typical appearance of a Ward's method dendrogram.

Only a very brief account of model-based methods is attempted here – the mathematics is beyond the level of these notes. Banfield and Raftery (1993) provide a technical account that led on to the development of a package in `R`, `mclust`, to implement some of the methods; Papageorgiou *et al.* (2001) provide a detailed archaeological application, but the methodology has not been much used. Some possible reasons for this are discussed after describing the ideas involved.

What follows is less general than would be possible but covers the most common uses. Assume that clusters are (hyper)-ellipsoidal in $p$-dimensional space; a special case is when the clusters are (hyper)-spherical. Assume the sample is from $K$ sub-populations and that the data for a single cluster, $k$, are sampled from a multivariate normal distribution. In principle the size and orientation of the clusters can vary. Given these assumptions it is possible to define a probabilty density function for the data, and parameters can be estimated using the method of maximum-likelihood; that is, an objective function is optimized but the optimization depends on the assumptions just listed. The parameters are those associated with the underlying probability densities and cluster labels, the estimation of which is the object of the exercise.

In its most general form the model is rarely used, as the number of parameters allied to the size of the data make estimation impractical. Banfield and Raftery (1993) simplify matters by developing models that impose constraints on the size, orientation and shape of the clusters. Thus, clusters may be constrained to have the same orientation, but allowed to vary in size, or *vice-versa*. Note that all this involves moving away from the theoretically preferred model. The most extreme simplification assumes that clusters are spherical and of the same size (orientation is irrelevant) and this leads to what is essentially Ward's method.

Thus Ward's method approximates the solution to a model that *assumes* that clusters have identical multivariate normal distributions other than their variation

in location. This results in a method that 'looks for', and tends to find, spherical clusters of the same size, resulting in a typical dendrogram having the appearance illustrated in Figures 10.4. The issue of choosing $K$ remains. Banfield and Raftery (1993) develop an approximate Bayesian method, rather complicated mathematically, that involves investigation of a range of values for $K$. In passing it can be noted that fully-fledged Bayesian approaches to clustering have been developed that are model-based and even more complex (Buck *et al.*, 1996). They have been little used in archaeology and such papers I have seen mostly use data for illustrative purposes where the cluster structure is obvious.

That model-based methods, other than in their simplest form, are rarely used is (apart from mathematical complexity) possibly because data sets are often too small or of too high a dimension to exploit the power of the methodology and/or because the structure of the data invalidates the assumptions of the methods. One area of application where model-based methodology is more practical is in spatial clustering, where the (usually) two-dimensional nature of the data (i.e. its low dimensionality) allows the use of more complex models.

Having said all this, Ward's method and other, *ad-hoc*, methods have been widely used and this is likely to continue. Ward's method is interesting in that it only approximates an ideal solution, and it may be possible to improve on this. This leads into the idea of k-means clustering.

### 10.3.3 K-means clustering

At its simplest the idea behind k-means clustering is straightforward. Newer and more complex methods that extend the ideas have been developed (e.g., Hastie *et al.* 2009). Attention here is confined to Ward's method. The basic idea is to take an initial starting position for the group centroids, either randomly chosen or from an initial Ward's method analysis (involving a choice of $G$), and reallocate cases between clusters until the optimum is, hopefully, attained.

For the seven-cluster Ward's method solution we follow Venables and Ripley (2002) and take the centroids of each cluster as a starting point. The distance from each case to each centroid can be calculated along with the criterion to be optimized, and cases can be reallocated if this improves the optimization. Table 10.1 compares the clustering from the original analysis with that resulting after reallocation.

Summing down the diagonal. 59/67 (88%) of cases are allocated to the same cluster as produced by Ward's method with three clusters remaining unaltered. Of the reallocated cases 5/8 were originally in Cluster 2; this result perhaps is not surprising given the earlier evidence from the average-link analysis and PCA. This kind of analysis is less common in the archaeological literature than one might expect; a possible reason is that such applications as exist often show very little

| Ward' method | K-means clusters | | | | | | |
|---|---|---|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 10 | 0 | 0 | 0 | 0 | 3 |
| 3 | 0 | 0 | 11 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 3 | 1 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 10 | 1 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |

Table 10.1: *A comparison of the clusterings obtained by Ward's method and subsequent reallocation using the k-means algorithm from the* `kmeans` *function in* `R`.

difference, if any, in the clusterings obtained. The results are not readily presented in a simple form such as the dendrogram.

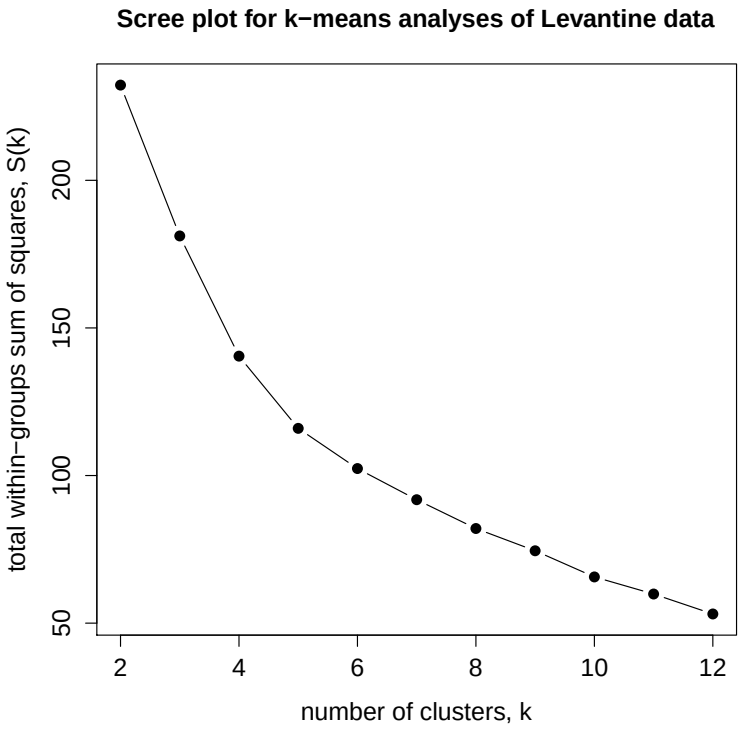


Figure 10.8: *A scree plot of $S(k)$ against $k$ for the Levantine data k-means analysis.*

The question 'what is the correct number of clusters' remains; it is not necessarily an easy question to answer. One approach that has been suggested is to look at a scree plot of the total within sums of squares for the clusters against the number of clusters, or $S(k)$ against $k$, where $S(k)$ is the criterion that k-means

optimizes for $k$ clusters (Section 10.3). The idea then is to look for a clear 'elbow' that suggests the appropriate number of clusters. A scree plot for the Levantine data is shown in Figure 10.8.

The idea is simple but, as with other uses of the scree plot (for example, determining the appropriate number of components in a PCA in Section 7.4), it is frequently of little use in practice. This is because a clear elbow is often not apparent. This is the case for Figure 10.8. It is inevitable that $S(k)$ will decrease as $k$ increases. In the figure the decay is greater up to $k = 5$ compared to larger values of $k$ where the decrease is almost linear. There is not a clear elbow; it could be argued that a value of $k = 6$ seems appropriate, but the graphical evidence is not especially compelling.

### 10.3.4 Fuzzy clustering

#### Levantine glass compositions

The methods discussed so far produce what are called *hard clusters* where each case is assigned to a single cluster. In *fuzzy clustering* cluster membership is distributed across all clusters. Baxter (2009) discusses the ideas involved with some technical detail that is not repeated here. Implementation in `R` is straightforward and the function used here, `cmeans` from the `e1071` package, is almost identical in structure to the `kmeans` function. An alternative implementation, not explored here, is the function `fanny` from the `cluster` package.

The one difference between the command line for `cmeans` and `kmeans` is the inclusion of a *fuzzification factor* for which the default in `cmeans` is the argument `m = 2`. This controls the 'crispness' of the clustering; as `m` approaches 1 a hard clustering is obtained; as it becomes large a totally fuzzy clustering results. The use of `m = 2` is arbitrary but claimed to work in a generally satisfactory way. Baxter (2009) provides examples where values of `m` less than 2 were judged to produce better results. Output from the two functions differs in various ways, the most important for present purposes being a table of membership values.

An illustration is provided in Table 10.2 for a subset of the Levantine data. Cluster 2 from the original Ward's method analysis contained 15 cases; 5/15 cases were reallocated using k-means analysis. Fuzzy clustering provides further insight into this and Table 10.2 shows how c-means distributes the membership of this cluster across the seven suggested by the Ward's method analysis.

The c-means analysis has the highest membership in Cluster 2 for 11/15 cases, though not always dramatically so. This is fairly similar to what is suggested by the k-means analysis but with values of the membership for 10 of these 11 cases below 50 it suggests, as might be inferred from earlier analyses, that the clustering is not a very crisp one. Of the four cases where Cluster 2 does not have the highest

| | Cluster | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Case | Membership | | | | | | |
| 2 | 23 | 48 | 9 | 8 | 1 | 10 | 2 |
| 3 | 21 | 33 | 11 | 22 | 3 | 8 | 3 |
| 5 | 17 | 41 | 22 | 8 | 2 | 7 | 2 |
| 6 | 22 | 42 | 10 | 8 | 2 | 14 | 2 |
| 15 | 26 | 49 | 7 | 8 | 1 | 8 | 2 |
| 25 | 32 | 22 | 4 | 32 | 1 | 8 | 1 |
| 29 | 38 | 17 | 3 | 36 | 1 | 4 | 1 |
| 34 | 22 | 32 | 11 | 23 | 3 | 7 | 2 |
| 35 | 21 | 48 | 10 | 12 | 2 | 6 | 2 |
| 43 | 26 | 43 | 5 | 17 | 1 | 6 | 1 |
| 45 | 3 | 94 | 1 | 1 | 0 | 1 | 0 |
| 46 | 18 | 38 | 27 | 6 | 2 | 8 | 2 |
| 63 | 19 | 26 | 12 | 21 | 6 | 12 | 4 |
| 66 | 12 | 26 | 47 | 5 | 1 | 7 | 2 |
| 67 | 12 | 23 | 48 | 5 | 2 | 8 | 3 |

Table 10.2: *Results from a c-means clustering of the Levantine data showing membership values for cluster 2 from the Ward's method analysis with seven clusters.*

membership, two could plausibly be associated with Clusters 2 or 4 and the other two with Cluster 3.

### Medieval glass compositions

For a different and simpler illustration the York medieval glass data are subjected to a similar analysis in Table 10.3. The analysis is simpler in the sense that the clustering is quite a crisp one with 25/27 cases clearly associated with a single cluster (taking, a little arbitrarily, 'clearly associated' to imply that the smallest membership value is 65 or greater).

The two exceptions to this observation are cases 20 and 22 which were the most outlying in Figure 10.1, particularly case 20. Case 22 can be seen from Table 10.3 to have a much higher membership value for Cluster 2 than 3 (60 compared to 29). The impression given by the plot for the first two PCs does not suggest this very clearly; further investigation reveals that the case is much closer to Cluster 2 if plots using the third component are examined[3]. It is clear, though, from the dendrogram and the PCA of the first two components, that this case is best regarded as an outlier. Case 20 is allocated to Cluster 3 in the original analysis; it has the highest membership value for this group but the difference compared to Cluster 2 is marginal (42 compared to 40). Inspection of plots using the third

[3]The plots of Figure 10.2 need to be labeled by case number to see this.

component show it is at some distance from the rest of Cluster 3, so the conclusion is that this is also a clear outlier.

| Id. | Cluster | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| | Membership | | |
| 1 | 100 | 0 | 0 |
| 2 | 98 | 1 | 1 |
| 3 | 1 | 98 | 1 |
| 4 | 98 | 1 | 1 |
| 5 | 95 | 2 | 2 |
| 6 | 92 | 4 | 5 |
| 7 | 8 | 7 | 85 |
| 8 | 4 | 3 | 93 |
| 9 | 99 | 1 | 1 |
| 10 | 1 | 97 | 2 |
| 11 | 1 | 97 | 1 |
| 12 | 1 | 97 | 2 |
| 13 | 1 | 98 | 1 |
| 14 | 1 | 98 | 1 |
| 15 | 1 | 99 | 1 |
| 16 | 7 | 7 | 86 |
| 17 | 5 | 6 | 89 |
| 18 | 6 | 86 | 8 |
| 19 | 1 | 98 | 1 |
| 20 | 19 | 40 | 42 |
| 21 | 12 | 21 | 67 |
| 22 | 11 | 60 | 29 |
| 23 | 11 | 14 | 76 |
| 24 | 6 | 7 | 86 |
| 25 | 5 | 4 | 91 |
| 26 | 5 | 4 | 91 |
| 27 | 5 | 3 | 92 |

Table 10.3: *Results from a c-means clustering of the York medieval glass data showing membership values.*

.

## 10.4 Summary

Given the difficulties of applying more sophisticated model-based methods to typical archaeological data (with the possible exception of spatial clustering), it is not surprising that practitioners continue to rely largely on the older and more *ad hoc* methods. Presumably they are commonly found to give archaeologically interpretable results, with the caveat that results judged to be unsatisfactory usually never get published.

Some space has been devoted to Ward's method, but not because it is a preferred choice. It is good practice to compare more than one clustering method, and Ward's method is a useful starting point because it usually suggests clear, if

possibly illusory, clusters. Given an initial and provisional identification the results can then be compared with those from other methods, as illustrated above.

Ward's method is also a useful peg on which to hang a discussion of model-based clustering methods. Although, for the reasons outlined, they have had limited archaeological use the underlying theory explains why Ward's method can be a potentially misleading method. Other methods lack a theoretical basis but can be subject to analogous or other problems. The method also provides an entrée into k-means and c-means clustering. Both have probably been underused by archaeologists (fuzzy clustering in particular). Both are very easily implemented in `R`. Fuzzy clustering is capable of producing a more nuanced view of the data than hard clusterings afford, and deserves more attention than it has received.

Books written specifically for archaeologists, that discuss cluster analysis, include, in ascending order of difficulty, Shennan (1997), Baxter (1994) and Baxter (2003). Not all the methods discussed in this chapter are covered in the last two books. General statistical texts, with a wider coverage, include Everitt *et al.* (2011), which is devoted to cluster analysis, accessible, and includes some archaeological examples. Good statistical texts on multivariate analysis, with treatments of CA, abound. They include Everitt and Dunn (2001), Krzanowski and Marriott (1995), and Seber (1984). This is in rough order of difficulty.

For 'newer' approaches to cluster analysis Everitt *et al.* (2011) is probably the most accessible statistical text for a non-statistical readership and includes material on mixture models and fuzzy cluster analysis. Hastie *et al.* (2009) covers several newer methods at a more advanced level. They are primarily concerned with methods of supervised pattern recognition (e.g., discriminant analysis, classification trees, neural networks), but have chapters on unsupervised pattern recognition that cover many more recent methods. Banfield and Raftery (1993) is a useful starting point for a statistical treatment of model-based clustering. Buck *et al.* (1996) is the best starting point for an exposition of the uses of Bayesian methods in archaeology, with references to, and applications of, cluster analysis to archaeological data; their pioneering work has not been emulated much.

## 10.5 `R` notes

*Figure 10.2*

```
pca <- prcomp(scale(york))
pairs(pca$x[,1:3], oma=c(4,4,6,12))
par(xpd=TRUE)
legend(0.85, 0.7)
```

This shows how, if you wish, placement of the pairs plot and a legend can

be controlled. The `oma` argument in the call to the `pairs` function controls the placement of the plots by adjusting the outer margins; `par(xpd = TRUE)` clips the plot to the figure region with the effect of emsuring the legend is visible. See the help on `par` for details. The first two arguments in the call to `legend` specify its location. The arguments `col`, `pch` and `cex` are available for the `pairs` function but have been omitted, as they and other arguments have in `legend`.

### *Figures 10.3 to 10.6*

The code for Figure 10.4 is given first since it is a prerequisite for Figures 10.5 and 10.6; Figure 10.3 is produced in a similar way. Normally this would be best written as a function, but it is convenient for expository purposes, to split the code that would be in the function into blocks. It is assumed that the standardized data for the Levantine glass compositions are held in `Levantine`. Version 3.1.2 of `R` was used for this; earlier work on these notes used `R` 2.13.1. For the most part it doesn't matter much, but in the later version, in `hclust`, it is necessary to select the version of Ward's method (there are now two possibilities); `method = "ward.D"`, based on Euclidean distance, was used here.

A Ward's method analysis is undertaken first since the clusters identified determine the labeling used in later analyses. The first block of code produces and plots the dendrogram, from which it was decided that a seven-cluster solution woud be used as a starting point.

```
data <- Levantine
clus <- hclust(dist(scale(data)), method = "ward.D")
plot(clus)
```

Next the `cutree` function is used to associate each case with a cluster label, 1 to 7, and these, in turn, are used to define a character variable, `Colour`, that associates each cluster with a color label.

```
clus.id <- cutree(clus, h = 7)
Colour <- c(rep("black",dim(data)[1]))
Colour <- ifelse(clus.id == 2, "magenta", Colour)
Colour <- ifelse(clus.id == 3, "blue", Colour)
Colour <- ifelse(clus.id == 4, "red", Colour)
Colour <- ifelse(clus.id == 5, "cyan", Colour)
Colour <- ifelse(clus.id == 6, "green2", Colour)
Colour <- ifelse(clus.id == 7, "pink", Colour)
```

Folllowing this the `as.dendrogram` function converts the previously defined object `clus` to a dendrogram structure that can then be manipulated to achieve the

desired effect using other functions. These include `order.dendrogram` which extracts the ordering (left to right) of case labels as they appear on the dendrogram.. The small function that follows then associates each of the re-ordered labels with its appropriate color as defined from the initial clustering. A plotting character (`pch = 16`) is also defined for the colored labels that are to be used subsequently in plotting the dendrogram. The main reason for doing this was so that dendrograms using clustering methods other than Ward's have the same colors associated with case labels, so the similarity to Ward's method results is more apparent[4].

```
hcd <- as.dendrogram(clus, hang = 0.1)
newindex <- order.dendrogram(hcd)
i <- 0
colLab <- function(n) {
    if (is.leaf(n)) {
     i <<- i + 1
     a <- attributes(n)
     attr(n, "label") <- NULL
     attr(n, "nodePar") <- c(a$nodePar, list(lab.col = "black",
     pch = 16, col = Colour[newindex[i]], cex = 1.2))
     attr(n, "frame.plot") <- TRUE
    }
    n
}
```

Having done all this the desired dendrogram can be plotted. The function `dendrapply` uses the function `colLab` defined above to color the symbols used for each label according to their cluster. The presentational arguments for `plot` have been omitted, as has the legend.

```
clusDendro = dendrapply(hcd, colLab)
plot(clusDendro)
```

To obtain single- and average-link clusterings replace `"ward.D"` wih `"s"` or `"a"` at the start of the above code, omitting the second block where the coloring is defined.

[4]I'd assumed that this would be a fairly simple thing to do, but based on what I found on the web apparently not. I eventually hit on what I used on a couple of sites, but did not make a note of the source. At least two packages, `dendroextras` and `dendextend`, have been written to facilitate dendrogram manipulation. I suspect the latter can do what I wanted, but how to do it is not especially obvious.

*Figure 10.8*

The `kmeans` function is used to obtain the total within-groups sums-of-squares for $k = 2, 3, \ldots, 12$ cluster solutions. The second argument `centers = initial` specifies centers of the initial clusters to use. These are the centroids of $k$ clusters obtained from an initial Ward's method analysis using code based on Venables and Ripley (2002: 318). It is simpler to use `centers = k` which will randomly sample $k$ (distinct) rows from the data matrix as starting centroids. If this is done the appearance of the scree plot is at the mercy of the random selection, and the scree plot may vary if an analysis is repeated.

```
WithinSS <- NULL
k = 2
hh <- hclust(dist(Levantine), method = "ward.D")

while(k < 13) {
initial <- tapply(Levantine,
           list(rep(cutree(hh, k), ncol(Levantine)), col(Levantine)),
           mean)

km <- kmeans(Levantine, centers = initial)
WithinSS <- c(WithinSS, km$tot.withinss)
k <- k + 1
}
 plot(2:12, WithinSS, type = "b")
```

*Table 10.2*

This proceeds much as the code for the k-means analysis above. In the call to `cmeans` the argument `m = 2` is the (default) 'fuzzification' factor, and can be varied. The object `cmmembership` can be printed and edited to get Table 10.2.

```
hh <- hclust(dist(Levantine), method = "ward.D")
cluslev.id <- cutree(hh, 7)

initial <- tapply(Levantine,
           list(rep(cutree(hh,7), ncol(Levantine)),
           col(Levantine)), mean)

cm <- cmeans(Levantine, initial, m = 2)
cmmembership <- 100 * round(cm$membership[cluslev.id == 2,], 2)
```