International Tables for Crystallography (2018). Vol. H, Section 3.8.5, pp. 331–333.

viewed in this way and that it is maintained when the data are examined in six dimensions.

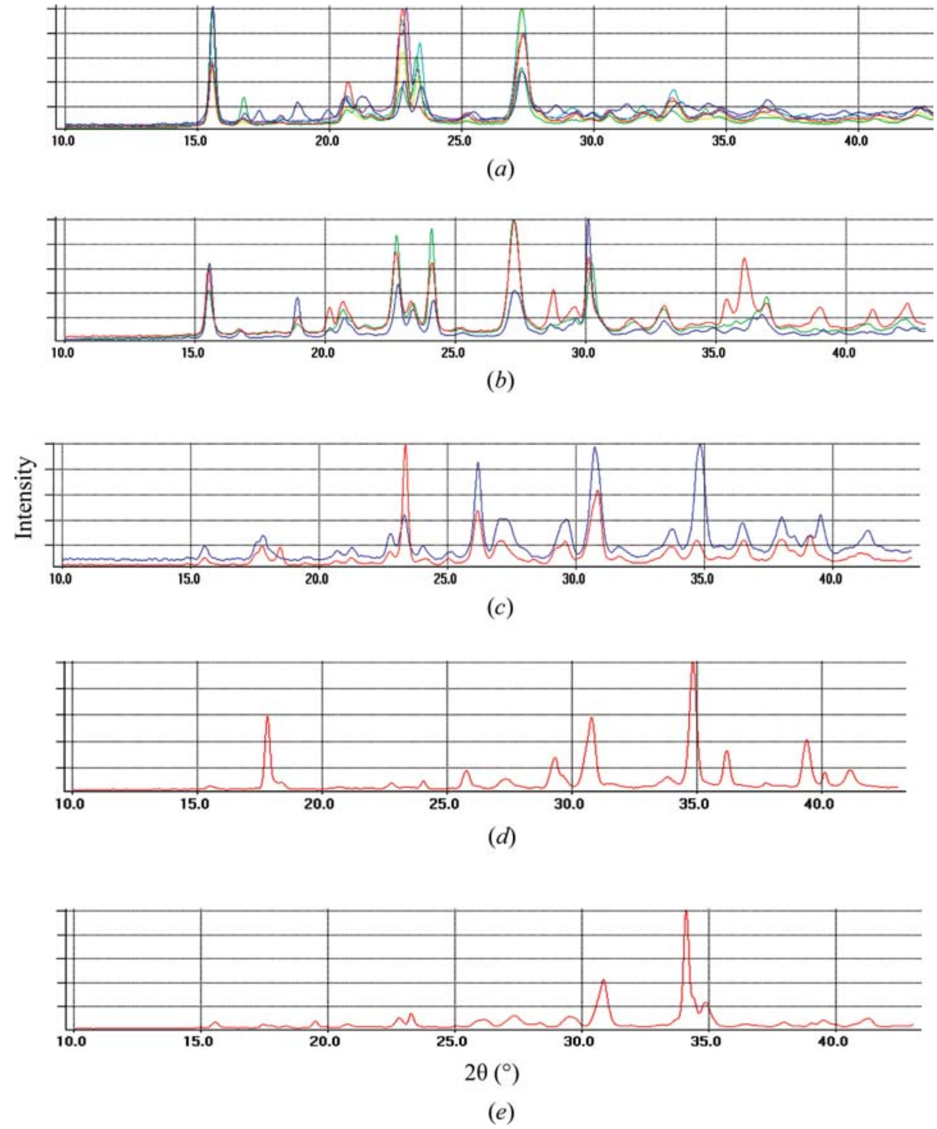

Figure 3.8.5
Powder patterns for 13 commercial aspirin samples partitioned into five sets. The patterns are in highly correlated sets: (*a*) comprises patterns 1, 3, 5, 6, 9 and 12; (*b*) comprises patterns 10, 11 and 13; (*c*) contains patterns 2 and 4; (*d*) contains pattern 7 and (*e*) contains pattern 8.

3.8.4.2.2. *The grand tour*

The grand tour is a method of animating the parallel-coordinates plot to examine it from all possible viewpoints. Consider a 3D data plot using orthogonal axes: a grand tour takes 2D sections through these data and displays them in parallel-coordinates plots in a way that explores the entire space in a continuous way. The former is important, because the data can be seen from all points of view, and the latter allows the user the follow the data without abrupt discontinuities. This concept was devised by Asimov (1985) and further developed by Wegman (1990). In more than three dimensions it becomes a generalized rotation of all the coordinate axes. A *d*-dimensional tour is a continuous geometric transformation of a *d*-dimensional coordinate system such that all possible orientations of the coordinate axes are eventually achieved. The algorithm for generating a smooth and complete view of the data is described by Asimov (1985).

To do this, the restriction of $p = 3$ in the MMDS calculation is relaxed to 6, so that there is now a 6D data set with six orthogonal axes. The choice of six is somewhat arbitrary – more can be used, but six is sufficient to see whether the clustering is maintained without generating unduly complex plots and requiring extensive computing resources. The data are plotted as a parallel-coordinates plot. The grand-tour method is then applied by a continuous geometric transformation of the 6D coordinate system such that all possible orientations of the axes are achieved. Each orientation is reproduced as a parallel-coordinates plot using six axes.

Figs. 3.8.9(*j*) and (*k*) show an example from the clustering of the 13 aspirin samples using PXRD data. Fig. 3.8.9(*j*) shows the default parallel-coordinates plot. Fig. 3.8.9(*k*) shows alternative views of the data taken from the grand tour. In Fig. 3.8.9(*j*) there appears to be considerable overlap between clusters in the 4th, 5th and 6th dimensions (X4, X5 and X6), but the alternative view given in Fig. 3.8.9(*k*) show that the clustering is actually well defined in all six dimensions (Barr, Dong & Gilmore, 2009).

3.8.4.2.3. *Powder data as a tree: the minimum spanning trees*

The minimum spanning tree (MST) displays the MMDS plot as a tree whose points are the data from the MMDS calculation (in three dimensions) and whose weights are the distances between these points. The minimum-spanning-tree problem is that of joining the points with a minimum total edge weight. (As an example, airlines use minimum spanning trees to work out their basic route systems: the best set of routes taking into account airport hubs, passenger numbers, fuel costs *etc.* is the minimum spanning tree.) Because a tree is used, each point is only allowed a maximum of three connections to other points.

To do this Kruskal's (1956) algorithm can be used, in which the lowest weight edge is always added to see if it builds a spanning tree; if so, it is added or otherwise discarded. This process continues until the tree is constructed. An example is shown in Figs. 3.8.7 for the 13-sample aspirin data. A complete tree for this data set using three dimensions and the MMDS-derived coordinates is shown in Fig. 3.8.7(*a*). This has 12 links between the 13 data points. Reducing the number of links to 10 gives Fig. 3.8.7(*b*).

3.8.5. Further validating and visualizing clusters: silhouettes and fuzzy clustering

Other techniques exist to validate the clusters, and these are discussed here.

3.8.5.1. *Silhouettes*

Silhouettes (Rousseeuw, 1987; Kaufman & Rousseeuw, 1990) are a property of every member of a cluster and define a coef-

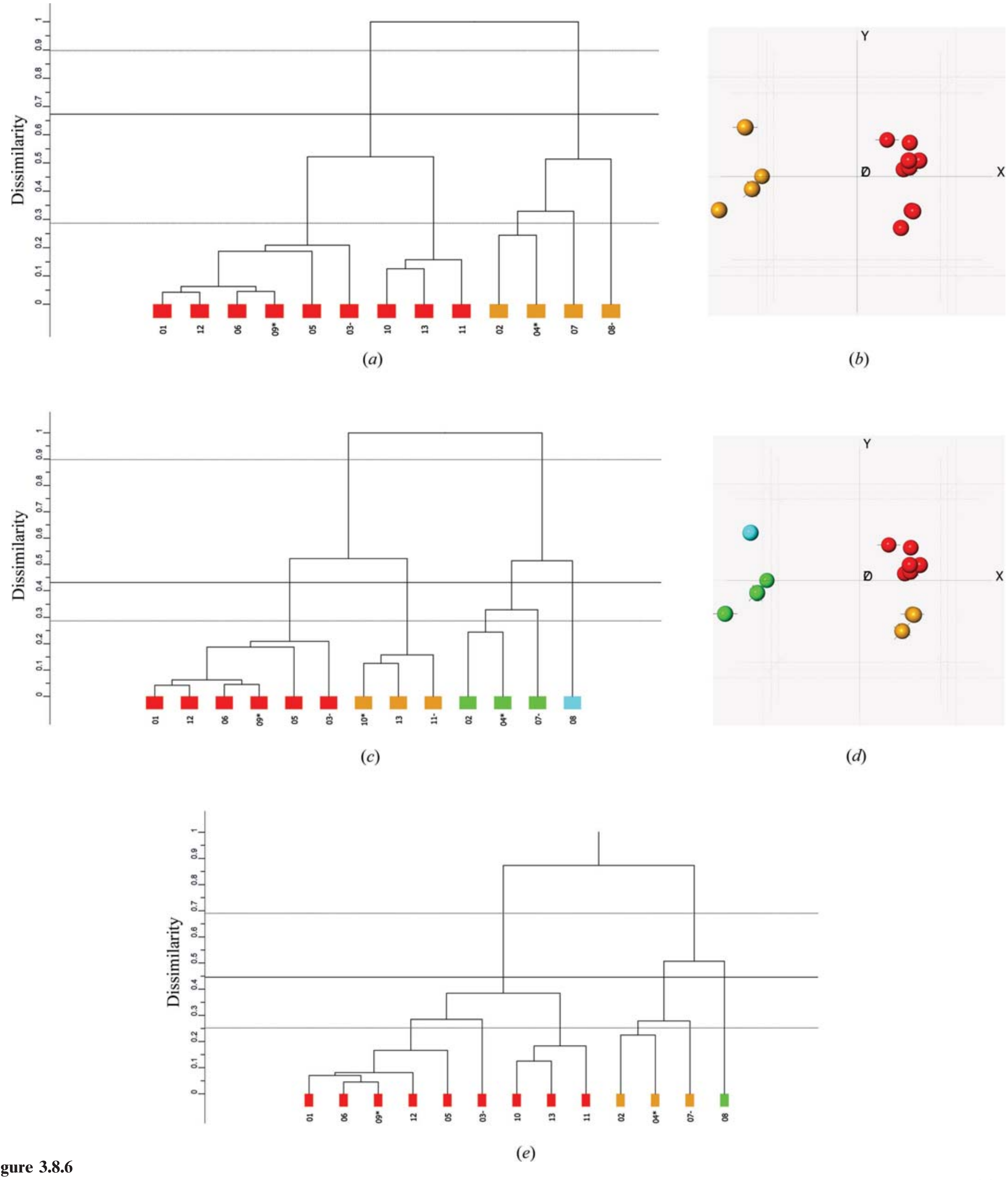

Figure 3.8.6
(*a*) The initial default dendrogram using the centroid clustering method on 13 PXRD patterns from 13 commercial aspirin samples. (*b*) The corresponding MMDS plot. It can be seen that both clusters have a natural break in them and should be partitioned into two clusters. (*c*) The dendrogram cut line is reduced. (*d*) The corresponding MMDS plot. The red cluster is now partitioned into two; the remaining patterns are a light-blue singleton and a green triplet cluster. (*e*) The default dendrogram using the single-link method.

ficient of cluster membership. To compute them, the dissimilarity matrix, $\boldsymbol{\delta}$, is used. If the pattern i belongs to cluster C_r which contains n_r patterns, we define

$$a_i = \sum_{\substack{j \in C_r \\ j \neq i}} \delta_{ij}/(n_r - 1). \qquad (3.8.19)$$

This defines the average dissimilarity of pattern i to all the other patterns in cluster C_r. Further define

$$b_i = \min_{s \neq r}\left\{\sum_{j \in C_s} \delta_{ij}/n_s\right\}. \qquad (3.8.20)$$

The silhouette for pattern i is then

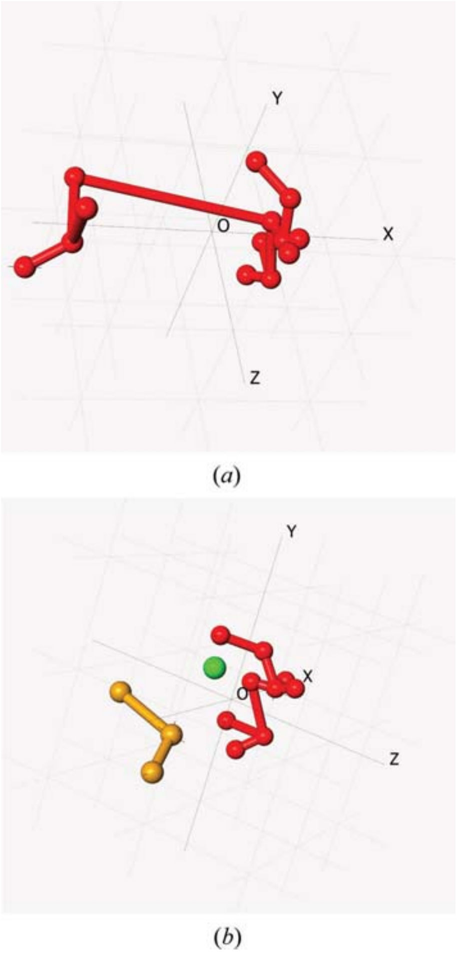

Figure 3.8.7
The use of minimum spanning trees (MSTs). (*a*) The MST with 12 links. (*b*) The MST with 10 links; three clusters are now present.

$$h_i = \frac{b_i - a_i}{\max(a_i,\ b_i)}. \qquad (3.8.21)$$

Clearly $-1 \le h_i \le 1.0$. It is not possible to define silhouettes for clusters with only one member (singleton clusters). Silhouettes are displayed such that each cluster is represented as a histogram of frequency plotted against silhouette values so that one can look for outliers or poorly connected plots.

From our experience with powder data collected in reflection mode on both organic and inorganic samples (Barr *et al.*, 2004*b*), we conclude that for any given pattern

(1) $h_i > 0.5$ implies that pattern *i* is probably correctly classified;
(2) $0.2 < h_i < 0.5$ implies that pattern *i* should be inspected, since it may belong to a different or new cluster;
(3) $h_i < 0.2$ implies that pattern *i* belongs to a different or new cluster.

The use of silhouettes in defining the details of the clustering is shown for the aspirin data in Fig. 3.8.8. The silhouettes for the red cluster corresponding to the dendrogram in Fig. 3.8.6(*a*) are shown in Fig. 3.8.8(*a*) and those for the corresponding orange cluster are shown in Fig. 3.8.8(*b*). Both sets of silhouettes have values < 0.5, which indicates that the clustering is not optimally defined. When the cut line is moved to give the dendrogram in Fig. 3.8.6(*c*), the silhouettes for the red cluster are shown in Fig. 3.8.8(*c*). The entry centred on a silhouette value of 0.15 is pattern 3. This implies that pattern 3 is only loosely connected to the cluster and this is demonstrated in Fig. 3.8.8(*d*) where pattern 3 and the most representative pattern for the cluster (No. 9) are superimposed. Although there is a general sense of similarity there are significant differences and the combined correlation coefficient is only 0.62. In Fig. 3.8.8(*e*), the silhouettes for the orange cluster are shown. They imply that this is a single cluster without outliers. The silhouettes for the green cluster corresponding to the dendrogram in Fig. 3.8.6(*c*) are shown in Fig. 3.8.8(*f*). The clustering is poorly defined here.

3.8.5.2. *Fuzzy clustering*

In standard clustering methods a set of *n* diffraction patterns are partitioned into *c* disjoint clusters. Cluster membership is defined *via* a membership matrix $\mathbf{U}(n \times c)$, where individual coefficients, u_{ik}, represent the membership of pattern *i* of cluster *k*. The coefficients are equal to unity if *i* belongs to *c* and zero otherwise, *i.e.*

$$u_{ik} \in [0, 1] \quad (i = 1, \ldots, n;\, k = 1, \ldots, c). \qquad (3.8.22)$$

If these constraints are relaxed, such that

$$0 \le u_{ik} \le 1 \quad (i = 1, \ldots, n;\, k = 1, \ldots, c), \qquad (3.8.23)$$

$$0 < \sum_{i=1}^{n} u_{ik} < n \quad (k = 1, \ldots, c) \qquad (3.8.24)$$

and

$$\sum_{k=1}^{c} u_{ik} = 1, \qquad (3.8.25)$$

then fuzzy clusters are generated, in which there is the possibility that a pattern can belong to more than one cluster (see, for example, Everitt *et al.*, 2001; Sato *et al.*, 1966). Such a situation is quite feasible in the case of powder diffraction, for example, when mixtures can be involved. It is described in detail by Barr *et al.* (2004*b*).

3.8.5.3. *The PolySNAP program and DIFFRAC.EVA*

All these techniques have been incorporated into the *PolySNAP* computer program (Barr *et al.*, 2004*a*,*b*,*c*; Barr, Dong, Gilmore & Faber, 2004; Barr, Dong & Gilmore, 2009), which was developed from the *SNAP-D* software (Barr, Gilmore & Paisley, 2004). *PolySNAP* has subsequently been incorporated into the Bruker *DIFFRAC.EVA* program (Bruker, 2018), and the following sections are based on its use.

3.8.6. Examples

All the elements for clustering and visualization are now in place. Fig. 3.8.4 shows this as a flowchart. Hitherto we have looked at elements of the aspirin data to demonstrate how methods work; we now examine the aspirin data in detail as a single analysis.

3.8.6.1. *Aspirin data*

In this example we use 13 powder patterns from commercial aspirin samples collected in reflection mode on a Bruker D8 diffractometer. Since these samples include fillers, the active pharmaceutical ingredient (API) and other formulations, it is not surprising that peak widths are high: ~0.5° full width at half maximum (FWHM). The data-collection range was 10–43° in 2θ using Cu $K\alpha$ radiation. The 13 powder data sets are shown in Fig. 3.8.5 arranged into groups based on similarity. We have already described the methods of analysis and have shown typical results in Figs. 3.8.6 to 3.8.8, and now present detailed examples. The correlation matrix derived from equation (3.8.3) is shown in Fig.