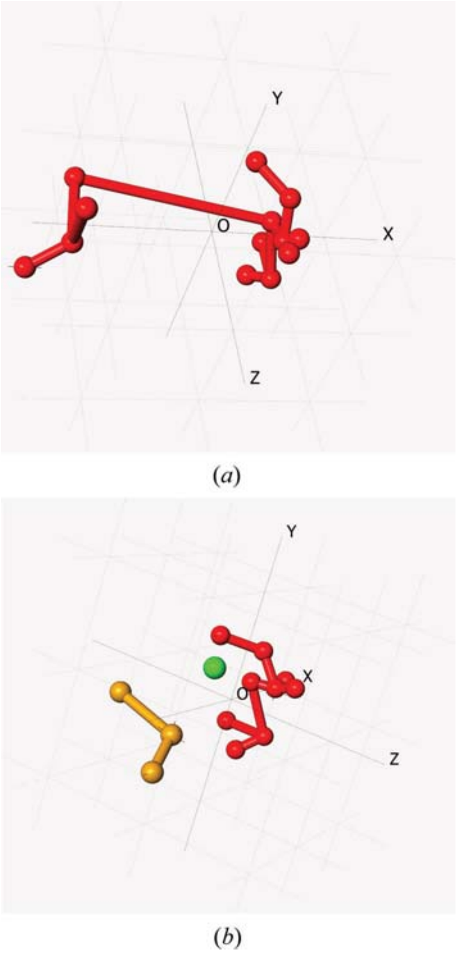

Figure 3.8.7
The use of minimum spanning trees (MSTs). (*a*) The MST with 12 links. (*b*) The MST with 10 links; three clusters are now present.

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

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.*

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

If these constraints are relaxed, such that

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

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

and

$$\sum_{k=1}^{c} u_{ik} = 1, \tag{3.8.25}$$

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.