International Tables for Crystallography (2018). Vol. H, Section 3.8.9, pp. 340–342.

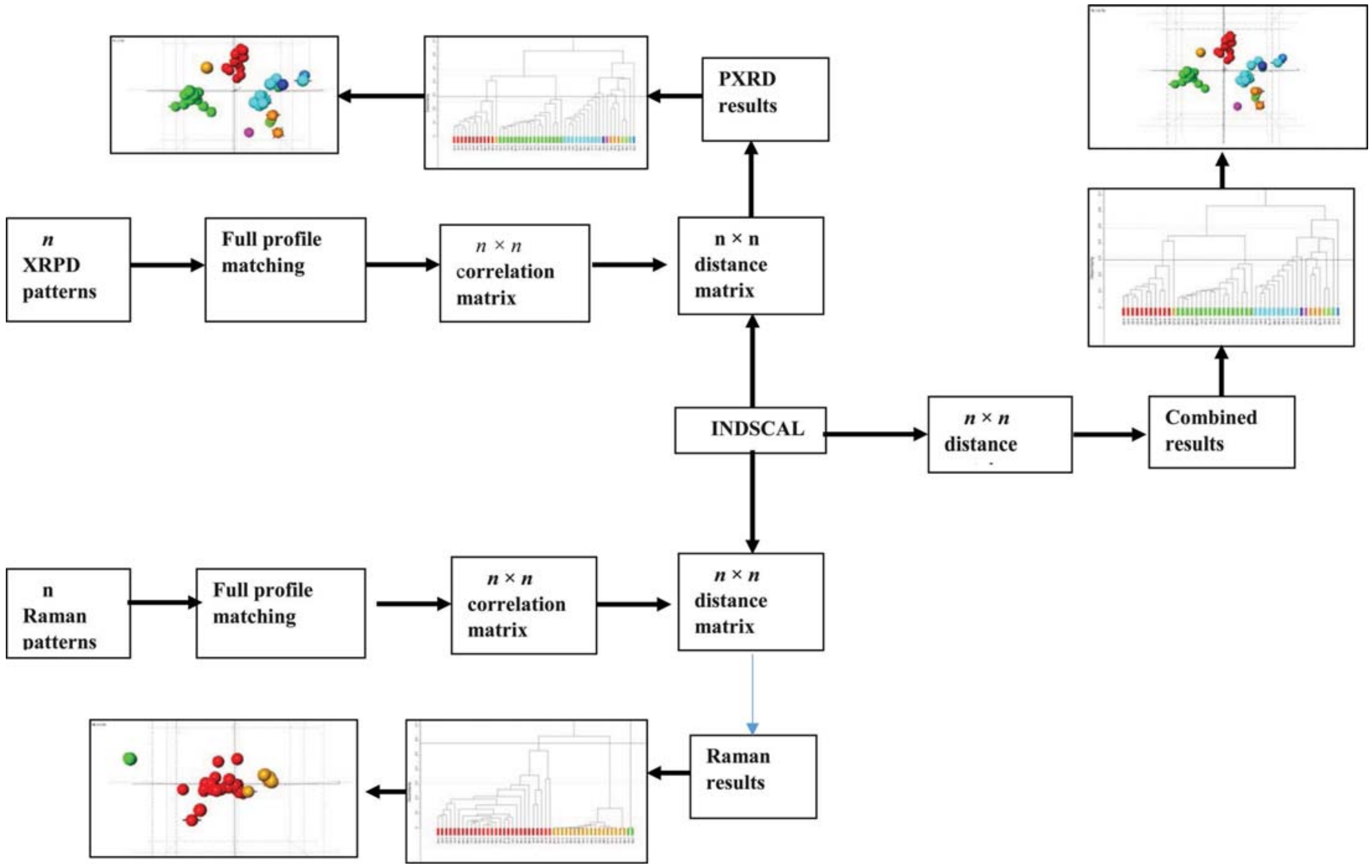

Figure 3.8.15
A flowchart for the INDSCAL method using Raman and PXRD data. Note that any combination of any 1D data can be used here.

morphs of mefenamic acid; the dark blue contains phenilbutazone; and finally the purple cluster contains sulfamerazine. The MMDS plot gives a complementary visualization of the data that supports the clustering.

It is also possible to use derivative data in place of the original spectra for clustering. The results of this for the 74 Raman spectra without initial background subtraction followed by the generation of first-derivative data are shown in Fig. 3.8.14. The clusters are well defined but now the carbamazepine data have split into two clusters. These correspond to forms I and III of carbamazepine, although the differences in the Raman spectra for these three species are small (O'Brien *et al.*, 2004). At the same time, both furosemide and mefenamic acid are each split into two groups. This is probably the best description of the data in terms of clustering and cluster membership corresponding to the chemical differences in the samples. The dendrogram also has the feature that the tie bars between samples are higher, *i.e.* the similarities are lower, reflecting the fact that the use of first derivatives accentuates small differences in the data.

It is interesting to note that, in general, PXRD works less well with derivative data. The reason for this is not clear, but possibly the presence of partial overlapping peaks and the associated issues of peak shape are partly responsible.

3.8.9. Combining data types: the INDSCAL method

It is now common to collect more than one data type, and some instruments now exist for collecting spectroscopic and PXRD data on the same samples, for example the Bruker D8 Screenlab, which combines PXRD and Raman measurement for high-throughput screening (Boccaleri *et al.*, 2007).

A technique for combining the results of more than one data type is needed. One method would be to take individual distance matrices from each data type and generate an average distance matrix using equation (3.8.3), but this leaves open the question of how best to define the associated weights in an optimal, objective way. Should, for example, PXRD be given a higher weight than Raman data? The individual differences scaling method (INDSCAL) of Carroll & Chang (1970) provides an unbiased solution to this problem by, as the name suggests, scaling the differences between individual distance matrices.

In this method, let $\mathbf{D}_k$ be the squared distance matrix of dimension $(n \times n)$ for data type k with a total of K data types. For example, if we have PXRD, Raman and differential scanning calorimetry (DSC) data for each of n samples, then $K = 3$. A group-average matrix $\mathbf{G}$ (which we will specify in two dimensions) is required that best represents the combination of the K data types. To do this, the $\mathbf{D}$ matrices are first put into inner-product form by the double-centring operation to give

$$\mathbf{B}_k = -\tfrac{1}{2}(\mathbf{I} - \mathbf{N})\mathbf{D}_k(\mathbf{I} - \mathbf{N}), \qquad (3.8.33)$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{N}$ is the centring matrix $I - \mathbf{11}'/N$; $\mathbf{1}$ is a column vector of ones. The inner-product matrices thus generated are matched to the weighted form of the group average, $\mathbf{G}$, which is unknown. To do this the function

$$S = \sum_{1}^{K} \left\| \mathbf{B}_k - \mathbf{G}\mathbf{W}_k^2\mathbf{G}' \right\| \qquad (3.8.34)$$

is minimized. The weight matrices, $\mathbf{W}_k$, are scaled such that

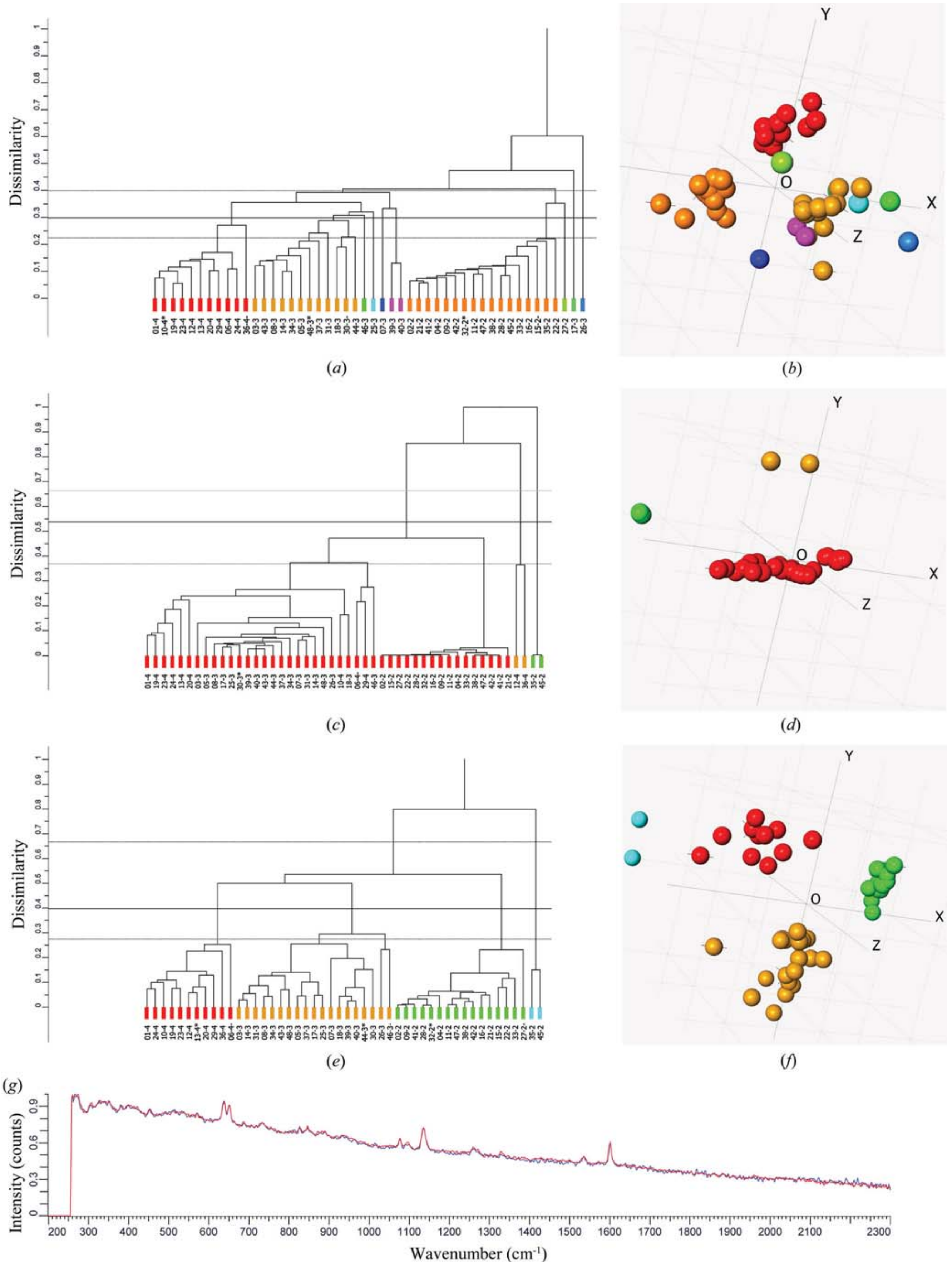

Figure 3.8.16

Clustering 48 PXRD spectra with background corrections applied for three polymorphs of sulfathiazole. (*a*) The dendrogram. Each sample is identified by a four-digit code. The first two digits are the well number, and the last digit defines whether the sample is form 2, 3 or 4 of sulfathiazole. (*b*) The MMDS plot: the red cluster is well defined but the rest of the spheres are diffuse and intermingled. (*c*) The dendrogram derived from clustering 48 Raman spectra of sulfathiazole with background corrections applied. (*d*) The corresponding MMDS plot. The clusters are poorly defined. (*e*) The results of the INDSCAL method. The dendrogram is shown with the default cut level. The clustering is correct; all the samples are placed in the correct group except for patterns 35-2 and 45-2. (*f*) The MMDS plot validates the dendrogram. (*g*) The Raman patterns for 35-2 and 45-2 superimposed. They are primarily background noise.

$$\sum_{k=1}^{K} \mathbf{W}_k^2 = K\mathbf{I}. \qquad (3.8.35)$$

The INDSCAL method employs an iterative technique to solve equation (3.8.7) in which one parameter is kept fixed whilst the other is determined by least-squares refinement. An initial estimate for **G** is taken either from the average of the **D** matrices for each sample or as a random matrix. This is then used to estimate the weight matrices, and the whole process repeated until a minimum value of S is obtained. The algorithm derived by Carroll and Chang was used in the example below. When random matrices are used to generate the initial **G** matrix, the INDSCAL procedure is repeated 100 times and the solution with the minimum value of S is kept. In practice, there is very little difference in the results of these two procedures. The resulting **G** matrix is used as a standard-distance matrix, and used in the standard way to generate dendrograms, MMDS plots *etc*. The method has the property that where data types show samples to be very similar this is reinforced, whereas where there are considerable variations the differences are accentuated in the final **G** matrix. For a fuller description of the INDSCAL method with examples see Gower & Dijksterhuis (2004), Section 13.2, and for a useful geometric interpretation see Husson & Pagès (2006).

3.8.9.1. *An example combining PXRD and Raman data*

We now present an example of the INDSCAL method applied to data collected on sulfathiazole using PXRD and Raman spectroscopy (Barr, Cunningham *et al.*, 2009). A flowchart is shown in Fig. 3.8.15. Three polymorphs of sulfathiazole were prepared and PXRD data were collected on a Bruker C2 GADDS system. Each sample was run for 2 min over a 3–30° range in 2θ using Cu $K\alpha$ radiation. Raman data were collected on a Bruker SENTINEL. The Raman probe was integrated into the PXRD instrument.

The only data pre-processing performed was background removal. Fig. 3.8.16(*a*) shows the resulting dendrogram (with the default cut level) and Fig. 3.8.16(*b*) shows the corresponding MMDS plot. To identify each sample they are numbered *via* a four-digit code: the first two digits are the well number, and the last digit defines whether the sample is form 2, 3 or 4 of sulfathiazole. It can be seen that the clustering is only partly successful: form 4 (red) is correctly clustered; form 3 (orange) gives five clusters and form 2 gives three clusters.

Fig. 3.8.16(*c*) shows the clustering from the Raman spectra. The results are poor: most of form 2 is correctly clustered, but forms 4 and 3 are intermixed, and the MMDS plot in Fig. 3.8.16(*d*) is diffuse with little structure.

The INDSCAL method is now applied starting from random **G** matrices and the results are shown in Fig. 3.8.16(*e*) and (*f*) with the dendrogram cut level at its default value. The clustering is almost correct; all the samples are placed in the correct groups except that there are two outliers coloured in blue. Fig. 3.8.16(*g*) shows the Raman patterns for these samples: they are primarily background with very little usable signal.

3.8.10. Quality control

Quality control (Gilmore, Barr & Paisley, 2009) is designed for situations where the stability of a material is being monitored over time, for example as part of a production-line system, or for periodic equipment alignment. A set of reference patterns is collected that represents acceptable measurements – any measurement sufficiently close to these references represents a good measurement. Various sample patterns are then imported and compared with those reference patterns, and any that vary significantly from the ideal are noted and highlighted.

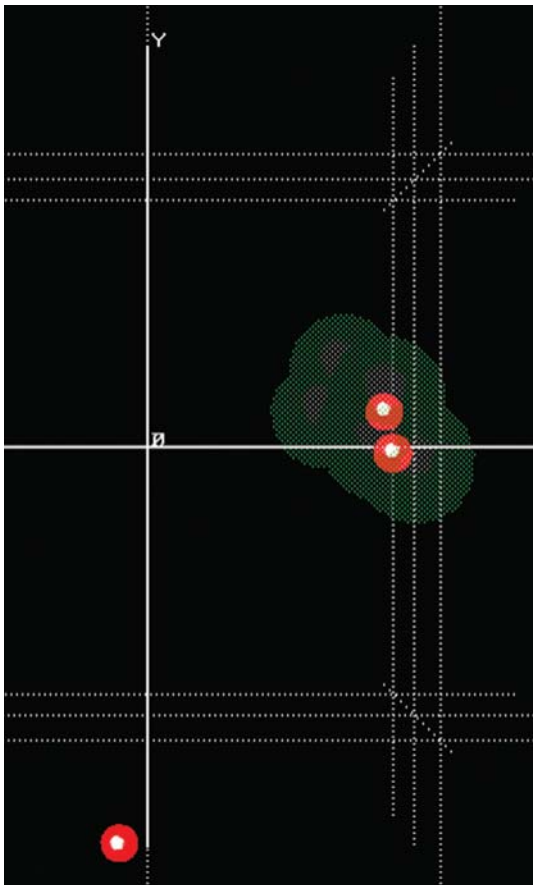

Figure 3.8.17
Visualization tools for quality-control procedures using a modified MMDS plot. The red outlier is a sample unacceptably far from the cluster of reference measurements.

The results are best displayed graphically using a variant of the MMDS method, of which an example is shown in Fig. 3.8.17. The reference patterns define a green shaded surface with acceptable sample patterns, coloured red, shown within it, and potentially problematic sample patterns appearing outside it. The volume of the green shape is defined by intersecting spheres around each reference sample and these can be altered to allow more- or less-stringent quality control.

3.8.11. Computer software

These calculations can be carried out using MATLAB (http://www.mathworks.co.uk/products/matlab/) or the open-source R software (http://www.r-project.org/; Crawley, 2007) with graphics using the *GGobi* software (Cook & Swayne, 2007). There are four commercial packages for handling powder data: *DIFFRAC.EVA* and *PolySNAP 3*, both from Bruker (http://www.bruker.com/; Gilmore, Barr & Paisley, 2004; Barr, Dong & Gilmore, 2009), *Jade* from Materials Data Inc. (http://www.materialsdata.com) and *HighScore* from Malvern PANalytical (http://www.panalytical.com/).

Data were kindly provided by Gordon Cunningham at Glasgow University, Arnt Kern of Bruker AXS in Karlsruhe and Michael Herrman at the Fraunhofer Institute, Germany.

references