

High-throughput powder diffraction. IV. Cluster validation using silhouettes and fuzzy clustering

Gordon Barr, Wei Dong and Christopher J. Gilmore*

Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland, UK.
Correspondence e-mail: chris@chem.gla.ac.uk

In two previous papers [Gilmore, Barr & Paisley (2004). *J. Appl. Cryst.* **37**, 231–242; Barr, Dong & Gilmore (2004). *J. Appl. Cryst.* **37**, 243–252], it was demonstrated how to generate a correlation matrix by comparing full powder diffraction patterns, and then partition the diffractograms into groups using multivariate statistics and associated classification procedures. For clustering the patterns into related sets, dendrograms, metric multidimensional scaling and three-dimensional principal-components analysis score plots are employed. However, sometimes cluster membership for certain patterns is not always very clear or other ambiguities may arise; this paper describes cluster validation techniques using silhouettes and fuzzy clustering. The two methods operate in a complementary way: in some cases silhouettes are the most useful, and in others fuzzy clustering is more applicable. These procedures are available as options in the commercial computer program *PolySNAP*.

1. Introduction

In previous papers (Gilmore *et al.*, 2004; Barr *et al.*, 2004a; Barr, Dong, Gilmore & Faber, 2004, referred to as I, II and III, respectively; see also Storey *et al.*, 2004) we have shown how to use the full powder diffraction pattern to partition collections of diffractograms into sets by generating a correlation matrix derived from matching the full profiles of all the powder patterns with one another, and then applying the relevant techniques of multivariate statistics and classification. For clustering the patterns into related sets, we use dendrograms coupled with metric multidimensional scaling (MMDS) and three-dimensional principal-components analysis (PCA) score plots. Sometimes cluster membership for certain patterns is not always very clear or other ambiguities arise; this paper describes some additional calculations and algorithms that can be used to validate cluster membership, in particular the use of silhouettes and fuzzy clustering. They operate in a complementary way: in some cases silhouettes are the most useful, and in others fuzzy clustering is more applicable. In §§2 and 3 we describe these techniques in detail, and follow this in §4 with a set of examples. These procedures are available as options in the *PolySNAP* computer program, licensed to Bruker-AXS (Barr *et al.*, 2004b).

2. Silhouettes

We start the high-throughput diffraction analysis by generating a correlation matrix, ρ . To do this, powder patterns are treated as bivariate samples with n measured points $[(x_1, y_1), \dots, (x_n, y_n)]$ and are compared with one another using a weighted mean of parametric and non-parametric correlation

coefficients (the Pearson and Spearman coefficients, respectively) using every measured intensity data point (Gilmore *et al.*, 2004) From this we generate a distance matrix, \mathbf{d} , where

$$d_{ij} = 0.5(1.0 - \rho_{ij}); \quad 0.0 \leq d_{ij} \leq 1.0, \quad (1)$$

or a similarity matrix \mathbf{s} where

$$s_{ij} = 1.0 - d_{ij}/d_{ij}^{\max}; \quad 0.0 \leq s_{ij} \leq 1.0, \quad (2)$$

where d_{ij}^{\max} is the maximum element in the distance matrix. These matrices are used as input for the generation of dendrograms, the MMDS and PCA computations, which give the primary partition of data into clusters.

Silhouettes (Rousseeuw, 1987; Kaufman & Rousseeuw, 1990) are a property of every member of a cluster and define a coefficient of membership. To compute them, we use a dissimilarity matrix, δ , in place of the distance matrix. The relationship between the two is defined *via*

$$\delta_{ij} = d_{ij}/d_{ij}^{\max}. \quad (3)$$

If the pattern i belongs to cluster C_r , which contains n_r patterns, define

$$a_i = \sum_{\substack{j \in C_r \\ j \neq i}} \delta_{ij} / (n_r - 1). \quad (4)$$

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

$$b_i = \min_{s \neq r} \left(\sum_{j \in C_s} \delta_{ij} / n_s \right). \quad (5)$$

The silhouette for pattern i is then

$$h_i = (b_i - a_i) / \max(a_i, b_i). \quad (6)$$

Clearly $-1 \leq h_i \leq 1.0$. Furthermore, it is not possible to define silhouettes for clusters with only one member (singleton clusters).

From our experience with powder data collected in reflection mode on both organic and inorganic samples with peak widths varying from 0.1 to 0.05° FWHM, we conclude that for any given pattern:

(i) $h_i > 0.5$ implies that pattern i is probably correctly classified;

(ii) $0.2 < h_i < 0.5$ implies that pattern i should be inspected since it may belong to a different or new cluster;

(iii) $h_i < 0.2$ implies that pattern i belongs to a different or new cluster.

We display each cluster as a histogram, frequency plotted against silhouette values, and look for outliers or poorly connected plots.

3. Fuzzy clustering

We have already described the theory of fuzzy clustering as applied to high-throughput diffraction pattern analysis in paper III, but we present a brief overview of the principles again here for clarity. In standard clustering methods we partition a set of n diffraction patterns into c disjoint clusters. We can express cluster membership *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.*

$$u_{ik} \in [0, 1] \quad (i = 1, \dots, n; k = 1, \dots, c). \quad (7)$$

If we relax these constraints and insist only that

$$0 \leq u_{ik} \leq 1 \quad (i = 1, \dots, n; k = 1, \dots, c), \quad (8)$$

$$0 < \sum_{i=1}^n u_{ik} < n \quad (k = 1, \dots, c) \quad (9)$$

and

$$\sum_{k=1}^c u_{ik} = 1, \quad (10)$$

then we have the concept of fuzzy clusters or fuzzy sets 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 (see §4.4).

In this paper we will relax the constraint imposed by equation (10) by allowing the membership coefficients to be un-normalized; such coefficients are then sometimes called ‘possibilities’.

The generation of the \mathbf{U} matrix is not simple and, as described in paper III, we have explored two methods as discussed in detail by Sato *et al.* (1966).

(a) Additive clustering in which \mathbf{U} is determined by minimizing the difference between the observed and calculated

similarity matrices coupled with steepest descents for optimization. The function minimized is

$$\eta_1^2 = \sum_{i \neq j=1}^n \left(s_{ij} - \alpha \sum_{k=1}^c u_{ik} u_{jk} \right)^2 / \sum_{i \neq j=1}^n (s_{ij} - \bar{s})^2, \quad (11)$$

where

$$\bar{s} = [1/n(n-1)] \sum_{i \neq j=1}^n (s_{ij}) \quad (12)$$

and α is a constant that scales \mathbf{s} and \mathbf{U} .

(b) The use of a more general algorithm using aggregation operators and also coupled with steepest descents. In this case we minimize

$$J = \sum_{i \neq j=1}^n \left[s_{ij} - \sum_{k=1}^c \min(u_{ik}, u_{jk}) \right]. \quad (13)$$

These will be referred to as methods 1 and 2, respectively. Both techniques need starting values of \mathbf{U} . We use the initial cluster assignments from the dendrogram such that if powder pattern i is deemed to belong to cluster j , the initial value of $u_{ij} = 0.8$; otherwise it is given a random value scaled in accordance with equation (10).

The two methods minimize different functions and thus give different results, although they do not usually differ significantly. Method 2 tends to give values of u_{ij} with a wider dynamic range. Where relevant, we present the results of both calculations in §4.

Finally, membership coefficients $u_{ij} < 0.3$ can usually be treated as zero.

4. Using silhouettes and fuzzy clusters

All the results presented here are derived using the silhouette and fuzzy clustering options in *PolySNAP* (Barr *et al.*, 2004*b,c*) employing real experimental data (except the simulated mixtures in §4.4 which are sums of experimental patterns) collected on a variety of diffractometers. We start with a situation in which the initial clustering is well behaved, and show that the silhouettes and fuzzy clusters add additional evidence that this is so, then move on to a series of situations where there is ambiguity in some cluster assignments that these validation methods can help to resolve. All the data sets are relatively small in order to preserve the clarity and presentation of our argument, but the techniques are equally (if not more so) valid when used with larger data sets. From our experience with data sets of up to 2000 patterns, there are no limits on the validity of the silhouette formalism with pattern numbers, but fuzzy clustering techniques become less useful with more than 100–200 data sets.

4.1. Well defined clusters

We begin with an example where the clusters are well defined. The data come from a proprietary pharmaceutical compound and were collected on a Bruker D8-GADDS system in reflection mode with a 2θ range of 5–43°. Peak

Table 1
Silhouettes and membership coefficients.

(a) Silhouettes corresponding to Fig. 1. Each cluster is well defined.

Pattern	Cluster	Silhouette
7	1	0.603
11	1	0.589
3	2	0.760
6	2	0.623
12	2	0.586
16	3	0.817
1	3	0.799
2	3	0.790
15	3	0.697
10	4	0.906
14	4	0.905
4	5	0.841
5	5	0.829
9	5	0.812
13	5	0.751

(b) The membership coefficients, u_{ij} . The entries in bold face correspond to the cluster to which the pattern belongs. The membership functions are all >0.8 , and there are no other entries >0.23 , i.e. there is no evidence of patterns belonging to more than one cluster. This behaviour is indicative of well defined clusters correctly assigned.

Pattern No.	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	0.09	0.08	0.88	0.11	0.17
2	0.10	0.10	0.88	0.12	0.16
3	0.19	0.88	0.09	0.15	0.14
4	0.12	0.14	0.16	0.13	0.89
5	0.11	0.12	0.14	0.11	0.89
6	0.17	0.84	0.10	0.18	0.21
7	0.84	0.21	0.06	0.08	0.09
8	0.10	0.13	0.17	0.12	0.81
9	0.10	0.11	0.13	0.11	0.88
10	0.10	0.17	0.15	0.87	0.16
11	0.84	0.23	0.16	0.13	0.19
12	0.17	0.83	0.10	0.10	0.09
13	0.09	0.10	0.12	0.11	0.86
14	0.11	0.18	0.17	0.87	0.18
15	0.08	0.08	0.84	0.10	0.12
16	0.10	0.10	0.88	0.12	0.17

widths are *ca* 0.5° FWHM. There are 16 samples. Fig. 1(a) shows the dendrogram calculated using the complete-link method (Barr *et al.*, 2004a). It can be seen that the data are partitioned into five clusters connected with tie bars that represent high similarity between the members of each cluster. This is reinforced by the corresponding metric multi-dimensional scaling (MMDS) plot in Fig. 1(b). Here each sphere represents a single diffraction pattern, and each cluster is also well defined. In Figs. 1(c)–1(f) typical silhouette histograms for four of the clusters are shown: they are compact with no outliers and have no entries less than 0.5 for any silhouette. Table 1(a) shows the corresponding results in numerical form.

The fuzzy cluster coefficients are equally well behaved and shown in Table 1(b) using method 2 (method 1 gives very similar results). The membership functions are all >0.8 and there are no anomalous entries, i.e. patterns with either low membership coefficients in the class to which they are assigned

Table 2

The silhouettes for 13 powder diffraction patterns collected on a Bruker D8 diffractometer from commercial aspirin tablets.

Fig. 4 shows the initial dendrogram in which patterns 7 and 8 are separated into two singleton clusters. In (a) the silhouettes for the three clusters containing more than one sample are presented. The clusters are well defined with no silhouette <0.58 . When the cut level on the dendrogram is adjusted to merge patterns 7 and 8 into a single cluster (see Fig. 4c), the resulting silhouettes are displayed in (b). It can be seen that the cluster formed by patterns 7 and 8 is poorly defined and that cluster 4 now also contains possible outliers. This confirms the singleton status of patterns 7 and 8.

(a) Silhouettes for the three clusters containing more than one sample.

Pattern	Cluster No.	Silhouette
2	4	0.712
4	4	0.691
10	5	0.749
11	5	0.700
13	5	0.661
9	6	0.807
6	6	0.794
1	6	0.792
5	6	0.696
12	6	0.584

(b) Silhouettes after merging patterns 7 and 8 into a single cluster.

Pattern	Cluster No.	Silhouette
8	1	0.412
7	1	0.348
2	2	0.669
4	2	0.663
1	4	0.654
6	4	0.646
9	4	0.636
13	4	0.572
5	4	0.534
10	4	0.433
12	4	0.385
11	4	0.337

or high memberships in alternative clusters. We can therefore be confident in the cluster assignments made by *PolySNAP*.

4.2. Ambiguous cluster definition

The second case is not so simple. The data comprise 106 pharmaceutical samples, also collected on a Bruker D8-GADDS system in reflection mode. Peak widths are *ca* 0.5° FWHM. The dendrogram shown in Fig. 2(a) is ambiguous: there are three clusters, but the large red and yellow coloured groups are connected by a relatively low tie bar with a third, more isolated, small group in green. Furthermore, the MMDS plot in Fig. 2(b) shows that the two large clusters are in close proximity. The green cluster is still well isolated from the others. The silhouettes for the green and yellow clusters are well defined with no entries <0.5 , but the red cluster is more diffuse and has several entries <0.5 . These silhouettes are displayed in Figs. 2(c)–2(e).

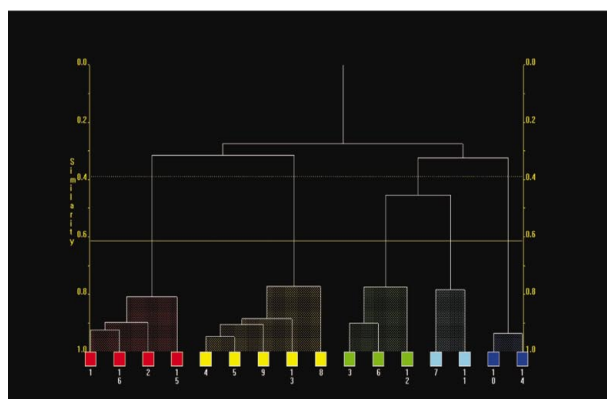
In Fig. 3(a) the tie bar in the dendrogram is raised so that the two large clusters amalgamate into one. The associated MMDS plot in Fig. 3(b) looks convincing, although there are

several potential outliers. The silhouettes, shown in Fig. 3(c), however, are very well defined with no entry <0.6 . In this way we can be sure that the data comprise one large cluster and a small unrelated one without investigating any individual powder diffraction patterns, although one should still inspect any potential outliers in the final stages of analysis.

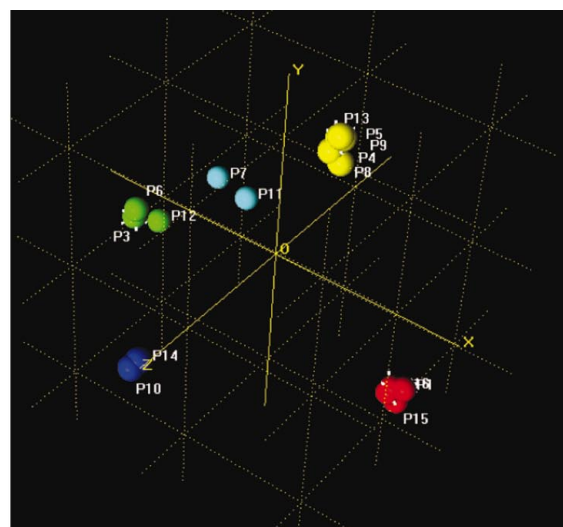
Fuzzy clusters are of limited value here, and do not indicate the need for amalgamation of the two large groups.

4.3. Are two patterns to be clustered together?

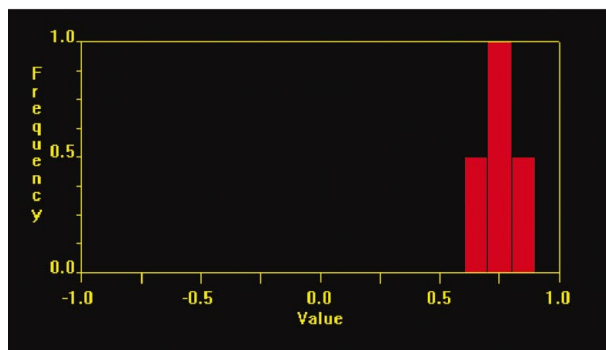
In this example we use 13 powder patterns from commercial aspirin samples collected in reflection mode on a Bruker D8



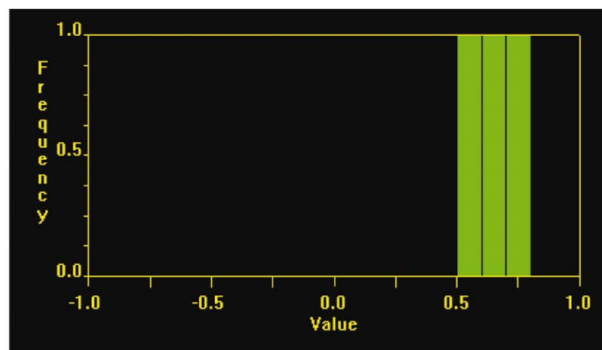
(a)



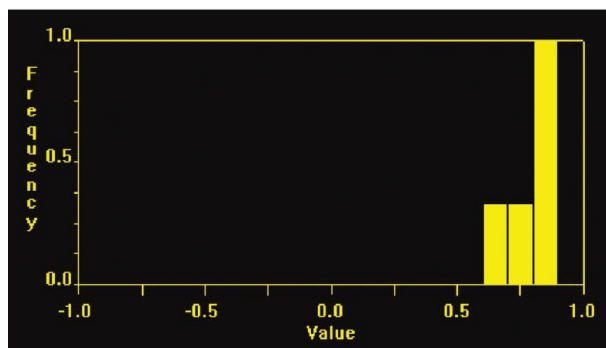
(b)



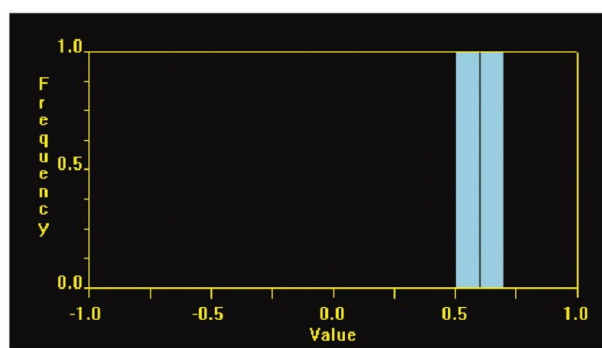
(c)



(d)



(e)



(f)

Figure 1

(a) The dendrogram for 16 powder diffraction patterns calculated using the complete link method. The data are partitioned into five clusters each with a unique colour. (b) The corresponding metric multidimensional scaling (MDS) plot in which each sphere represents a single diffraction pattern. The sphere colours are taken from the dendrogram. Each cluster is well defined. (c)–(f) show typical silhouettes for four of the clusters. They are compact and have no entries less than 0.5 in value of silhouette. (The fifth cluster is equally well defined and omitted for brevity.) Table 1 shows these results numerically.

system. Since these samples include fillers, aspirin itself and other formulations, it is not surprising that peak widths are *ca* 0.5° FWHM. The data collection range was $10\text{--}43^\circ 2\theta$. A default run of *PolySNAP* gives the dendrogram shown in Fig. 4(a); the data are partitioned into five sets with patterns 7 and 8 forming singleton clusters. The silhouettes for all the clusters containing more than one pattern are tabulated in Table 2(a); they are all well defined with no entries <0.58 . However, Fig. 4(b) presents the corresponding MMDS plot, and it can be seen that patterns 7 and 8 are relatively close. The question is therefore posed as to whether they should form a 2-pattern cluster.

In Fig. 4(c) the dendrogram cut level is raised so that this amalgamation takes place. Table 2(b) shows the resulting silhouettes. Both clusters 1 (formed by patterns 7 and 8) and 4 are now poorly defined with low silhouettes and possible outliers, indicating that there are significant differences between these patterns.

We now inspect the patterns themselves, shown superimposed in Fig. 4(d). There are considerable similarities and there is evidence of possible preferred orientation, but the peaks at *ca* 18 and $34^\circ 2\theta$ make it clear that these are different samples.

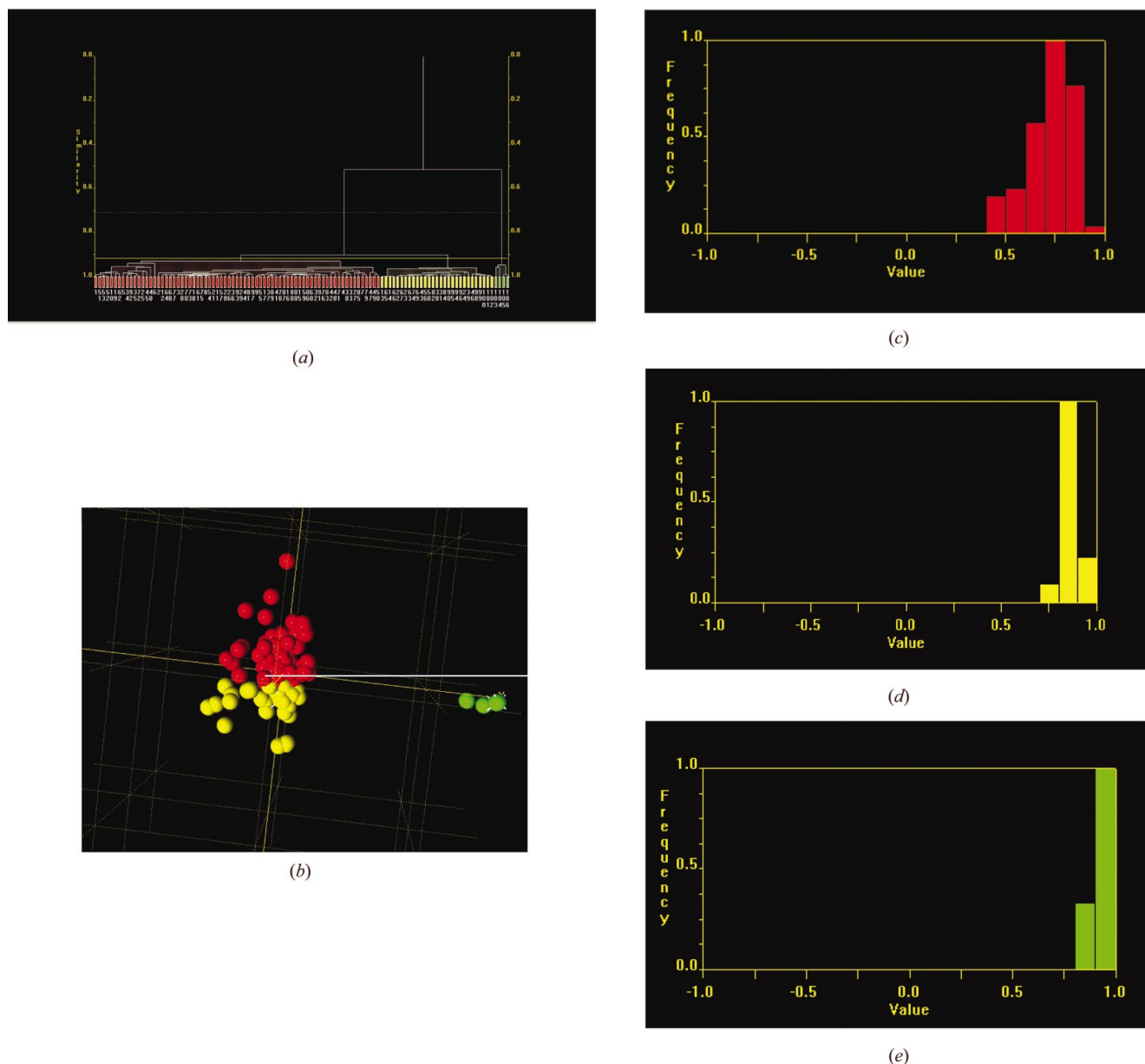


Figure 2 (a) The dendrogram for 106 powder diffraction patterns collected on a Bruker GADDS system. There are three clusters. The yellow and red groups are connected by a relatively low tie line and thus have a high similarity. The green cluster is quite distinct. (b) The corresponding MMDS plot. This reinforces the evidence from the dendrogram: the red and yellow clusters are in close proximity and almost overlapping, but the green coloured group remains separate. (c)–(e) show the silhouettes for the three clusters. The red cluster is somewhat diffuse and has six entries <0.5 (0.44–0.49), which indicates that the clustering pattern needs inspection. The green cluster has values between 0.93–0.89 and is well defined.

Although this is a simple case to resolve, cases where there are more than 1000 patterns are much more complex, and silhouettes can provide a powerful tool for resolving membership ambiguities of this type. It is interesting to note that fuzzy clustering was again of minimal value in this situation.

4.4. Mixtures

Mixtures are a common occurrence in high-throughput experiments and *PolySNAP* has numerous tools to process them in both qualitative and quantitative mode. However,

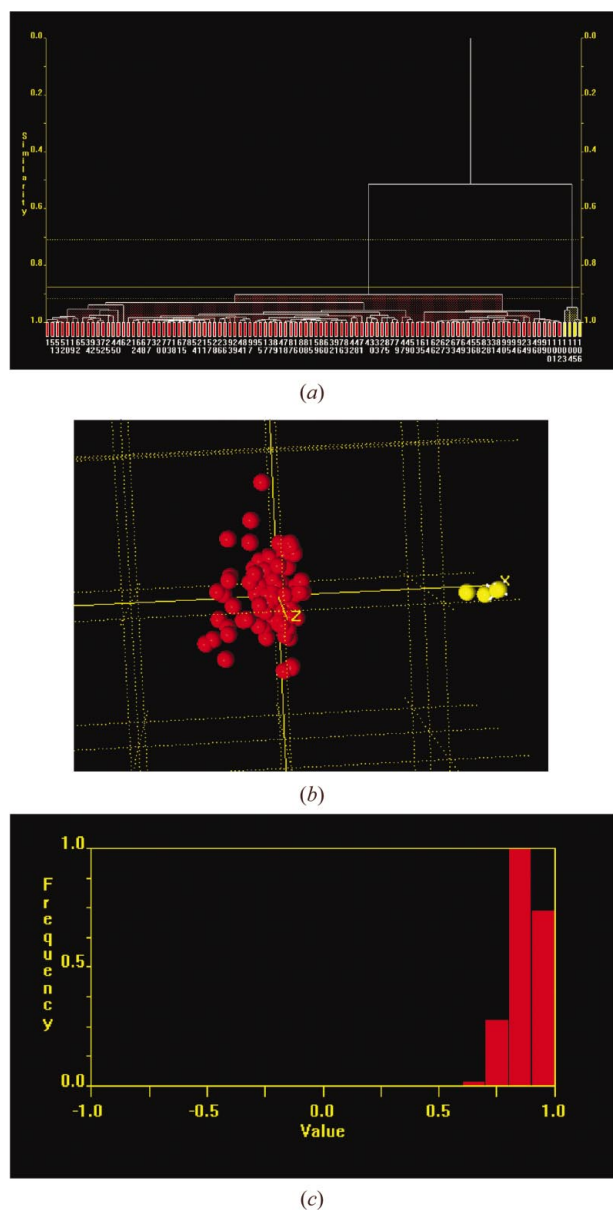


Figure 3
 (a) The same data as Fig. 2, but the dendrogram cut line has now been raised so that the red and yellow clusters are joined into one large group. (b) The corresponding MMSD plot. The red cluster is well defined. (c) The corresponding silhouette plot. The silhouettes are now tightly clustered with a minimum value of 0.64, providing strong evidence for cluster amalgamation.

fuzzy clustering is also useful. As an example, we present data from a proprietary pharmaceutical compound collected in reflection mode on a Bruker D8-GADDS system. The data

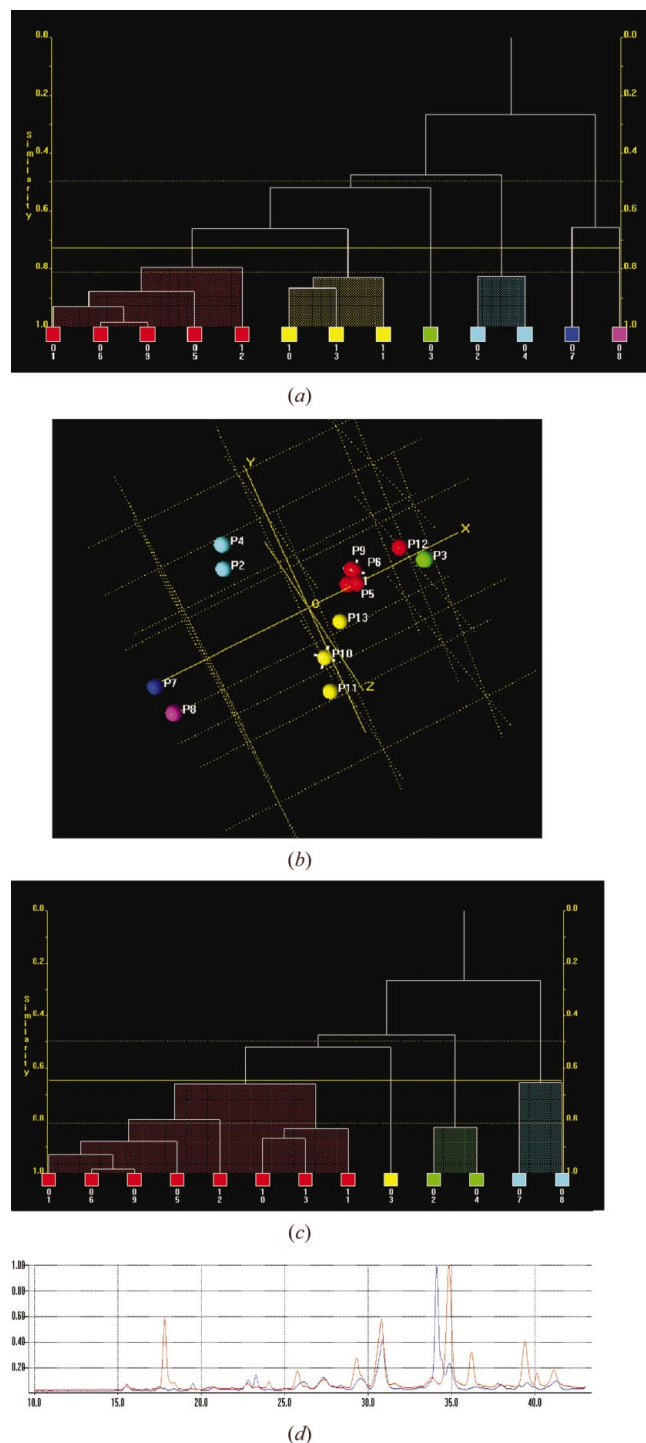


Figure 4
 (a) The powder patterns for the 13 commercial aspirin samples partitioned into five sets; patterns 7 and 8 form singleton clusters. (b) The corresponding MMSD plot; patterns 7 and 8 are relatively close, and the question is posed as to whether they form a cluster together? (c) The dendrogram cut level is adjusted so that this takes place. (d) The superposition of the powder diffraction patterns of samples 7 (blue) and 8 (red).

Table 3

The membership coefficients, u_{ij} , corresponding to Fig. 4.

The results of both fuzzy clustering methods are displayed; the output from method 1 is in columns 3 and 4, whilst that of method 2 is in columns 5 and 6. They are very similar. Samples 1–4 are pure form A and all have values of u_{ij} corresponding to membership of a single cluster (number 2). Patterns 8–11 are all pure form C, and they too have membership coefficients indicating that they belong to cluster 1 and no other. Pattern 5 comprises 40% B and 60% C; pattern 6 is 50% B and 50% C, while set 7 contains 60% B and 40% C. (The patterns for the mixtures have been generated artificially by adding the most representative sample of the pure forms with the required proportions.) These four patterns have significant membership coefficients of both clusters, and thus the mixtures are clearly identified.

Pattern	% composition	Method 1		Method 2	
		Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	Pure A	0.11	0.76	0.07	0.72
2	Pure A	0.10	0.76	0.04	0.73
3	Pure A	0.19	0.71	0.18	0.68
4	Pure A	0.10	0.76	0.05	0.73
5	A 40% B 60%	0.46	0.63	0.42	0.61
6	A 50% B 50%	0.40	0.72	0.36	0.66
7	A 60% B 40%	0.32	0.76	0.29	0.70
8	Pure B	0.73	0.21	0.71	0.18
9	Pure B	0.75	0.24	0.71	0.19
10	Pure B	0.70	0.26	0.70	0.21
11	Pure B	0.68	0.23	0.71	0.20

collection range was 12–45° 2θ. Peak widths were *ca* 0.5° FWHM. There are two polymorphic forms present: A (patterns 1–4) and B (patterns 8–11). Patterns 5–7 are mixtures generated by adding the patterns of the pure forms in the following proportions: pattern 5 is A 40%, B 60%; pattern 6 is A 50%, B 50%, and pattern 7 comprises A 60%, B 40%. The default dendrogram from *PolySNAP* on this data set is shown in Fig. 5. The data are partitioned into two clusters with three of the mixtures in the red coloured cluster and one in the yellow. There is little indication of mixtures from this display. The silhouettes also show nothing unusual: cluster 1 has silhouette values between 0.76 and 0.81 and cluster 2 between 0.68 and 0.85.

The fuzzy cluster memberships tell a different story; this is shown in Table 3. Both fuzzy clustering methods are used and the results are very similar. Samples 1–4 all have values of u_{ij} corresponding to membership of a single cluster (number 2). Patterns 8–11 are all pure form B, and they too have membership coefficients indicating that they belong to cluster 1 and no other. Patterns 5–7, however, have significant membership coefficients of both clusters, and thus the possibility of mixtures is clearly identified. *PolySNAP* could now be re-run in quantitative mode with a database of pure forms used as additional input.

4.5. Optimum shifts

One of the commonest sources of systematic error in matching powder patterns, especially in high-throughput situations linked to crystallization robotics, is the occurrence of 2θ shifts arising from variability of the instrumental zero point, sample height, transparency, *etc.* (see Klug & Alex-

ander, 1974; Wilson, 1963). The *PolySNAP* software provides three possible corrections:

$$\Delta(2\theta) = a_0 + a_1 \cos \theta, \quad (14)$$

which corrects for the zero-point error *via* the a_0 term and, *via* the $a_1 \cos \theta$ term, for varying sample heights in reflection mode, or

$$\Delta(2\theta) = a_0 + a_1 \sin \theta, \quad (15)$$

which corrects for transparency errors, or

$$\Delta(2\theta) = a_0 + a_1 \sin 2\theta, \quad (16)$$

which provides transparency coupled with thick-specimen error corrections. The parameters a_0 and a_1 are refinable constants determined by maximizing pattern–pattern correlations, although this greatly increases the run time of the program (see paper II). A problem can arise as to which of the equations (14), (15) or (16) is most suitable in a given experiment; we show here the applicability of fuzzy clusters to this problem.

The test data for this example comprise 15 patterns from the ICDD database of clay minerals where the full diffraction profiles are available (ICDD, 2003). The data were collected on a wide variety of instruments in reflection mode; typical peak widths were *ca* 0.05–0.1° FWHM (for further details see Barr *et al.*, 2004). The *PolySNAP* program partitions the data into five distinct clusters. Table 4(a) shows the membership coefficients using clustering method 2 before the application of any shifts, and then after the shift function $a_0 + a_1 \sin \theta$ has been applied. The maximum shift for both coefficients was 0.1. The entries in bold face correspond to the cluster to which the pattern has been assigned by the dendrogram. The average membership coefficient, u_{ij} , is 0.74, with a minimum value of 0.65, whereas after the application of optimal shift they take the corresponding value of 0.80 with a minimum value of 0.76. All the membership coefficients increase. Attempts to use the two other shifts [equations (14) and (16)] resulted in no significant change in the fuzzy cluster values.

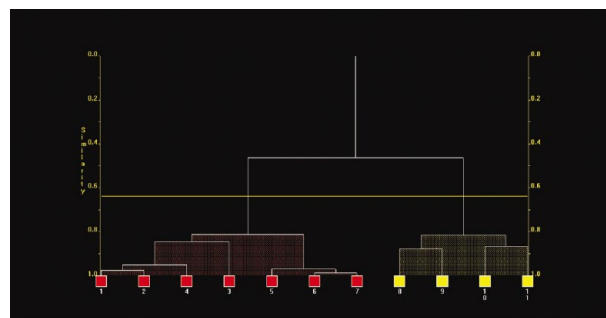


Figure 5

The dendrogram for a set of 11 organic powder patterns with two polymorphs. Samples 1–4 are form A; 8–11 are all form B; pattern 5 comprises 40% A and 60% B; pattern 6 is 50% A and 50% B, while set 7 contains 60% A and 40% B. The patterns for the mixtures have been simulated by adding the most representative sample of the pure forms with the required proportions. The data are partitioned into two groups, with the mixtures belonging to one or other of the sets.

Table 4

Using membership coefficients to determine the optimum formula for shifting powder diffraction patterns relative to each other.

The data comprise 15 patterns from the ICDD database which form five distinct clusters. (a) shows the membership coefficients using clustering method 2 before the application of any shifts and then after the shift function $a_0 + a_1 \sin \theta$ has been applied. The entries in bold face correspond to the cluster to which the pattern belongs. The average membership coefficient is 0.74 with a minimum value of 0.65, whereas after the application of optimal shifts it is 0.80 with a minimum value of 0.76. (b) shows the corresponding values of the silhouettes. These are much less sensitive to the shift function: the mean value before the shift is 0.709, whereas after its application it is 0.755; some patterns show a decrease in silhouette, whereas all the patterns show an increase in membership coefficient.

(a) Membership coefficients.

Pattern No.	u before $a_0 + a_1 \sin \theta$ shift					u after $a_0 + a_1 \sin \theta$ shift				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	0.07	0.75	0.06	0.08	0.05	0.14	0.81	0.03	0.15	0.04
2	0.00	0.05	0.00	0.76	0.06	0.05	0.06	0.03	0.80	0.09
3	0.04	0.07	0.00	0.76	0.06	0.09	0.12	0.05	0.83	0.07
4	0.04	0.12	0.00	0.68	0.09	0.05	0.17	0.10	0.78	0.09
5	0.00	0.00	0.78	0.00	0.00	0.08	0.03	0.79	0.09	0.05
6	0.72	0.12	0.17	0.06	0.06	0.79	0.21	0.19	0.14	0.07
7	0.72	0.02	0.30	0.02	0.03	0.79	0.11	0.27	0.01	0.06
8	0.05	0.14	0.00	0.73	0.05	0.08	0.21	0.01	0.83	0.06
9	0.00	0.00	0.00	0.04	0.77	0.02	0.01	0.05	0.07	0.84
10	0.05	0.72	0.04	0.12	0.05	0.11	0.79	0.03	0.18	0.03
11	0.15	0.02	0.75	0.00	0.00	0.22	0.06	0.76	0.00	0.05
12	0.09	0.65	0.07	0.14	0.06	0.13	0.78	0.04	0.17	0.06
13	0.00	0.00	0.78	0.00	0.00	0.10	0.01	0.83	0.09	0.06
14	0.03	0.00	0.78	0.00	0.00	0.16	0.05	0.80	0.02	0.06
15	0.00	0.02	0.00	0.05	0.77	0.06	0.02	0.02	0.09	0.84

(b) Corresponding silhouettes.

Pattern No.	Cluster	Silhouette before $a_0 + a_1 \sin \theta$ shift	Silhouette after $a_0 + a_1 \sin \theta$ shift
7	1	0.680	0.641
6	1	0.649	0.639
1	2	0.792	0.782
10	2	0.751	0.737
12	2	0.705	0.685
13	3	0.792	0.779
14	3	0.712	0.703
5	3	0.683	0.672
11	3	0.635	0.611
8	4	0.823	0.848
3	4	0.819	0.824
2	4	0.757	0.766
4	4	0.732	0.720
15	5	0.905	0.966
9	5	0.903	0.965

Table 4(b) shows the corresponding values of the silhouettes. These are much less sensitive to the shift function: the mean value before the shift is 0.709, whereas after its application it is 0.755 with some patterns showing a decrease in silhouette values while others increase.

5. Conclusions

We have shown how silhouettes and fuzzy clusters can be used as a secondary technique to validate cluster assignments when using powder diffraction data. They are not primary sources of the generation of clusters [although Rousseeuw (1987) has used them in that way], but serve in this instance as a tool for checking the final assignments, especially highlighting potential problem data sets in the presence of a large number of patterns.

The two methods are complementary: often one technique is insensitive to clustering ambiguities, whilst the other will

highlight possible problems, and for this reason *PolySNAP* allows the use of both automatically. Both are robust with respect to data defects, *e.g.* preferred orientation, large peak widths and high backgrounds.

Cluster analysis and related methods have a large literature, and we have not yet exhausted the possibilities in the area of high-throughput powder diffraction. We are now studying the use of neural networks, especially Kohonen self-organizing maps (Kohonen, 1997) and minimum spanning trees (see, for example, Graham & Hell, 1985). The methods described here should also be applicable to any one-dimensional data set such as Raman and IR spectroscopy or DSC, and we are currently investigating such applications.

We wish to thank Bob Docherty, Chris Dallman, Neil Feeder and Paul Higginson of Pharmaceutical Sciences, Pfizer Global R and D, UK, for data, many useful discussions and

suggestions, and for pioneering and inspiring this project, Bruker-AXS for the aspirin data, and the International Center for Diffraction Data for the data used in §4.5.

References

- Barr, G., Dong, W. & Gilmore, C. J. (2004a). *J. Appl. Cryst.* **37**, 243–252.
- Barr, G., Dong, W. & Gilmore, C. J. (2004b). *PolySNAP: a Computer Program for the Analysis of High-Throughput Powder Diffraction Data*. University of Glasgow and Bruker-AXS. (See also <http://www.chem.gla.ac.uk/staff/chris/snap.html>.)
- Barr, G., Dong, W. & Gilmore, C. J. (2004c). *J. Appl. Cryst.* **37**, 658–664.
- Barr, G., Dong, W., Gilmore, C. J. & Faber, J. (2004). *J. Appl. Cryst.* **37**, 635–642.
- Everitt, B. S., Landau, S. & Leese, M. (2001). *Cluster Analysis*, 4th ed. London: Arnold.
- Gilmore, C. J., Barr, G. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 231–242.
- Graham, R. L. & Hell, P. (1985). *Ann. Hist. Comput.* **7**, 43–57.
- ICDD (2003). *The Powder Diffraction File*. International Center for Diffraction Data, 12 Campus Boulevard, Newton Square, Pennsylvania 19073-3273, USA.
- Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data*. New York: Wiley.
- Klug, H. P. & Alexander, L. E. (1974). *X-ray Diffraction Procedures*, 2nd ed. New York: Wiley.
- Kohonen, G. (1997). *Self-Organizing Maps*, 2nd extended ed. Berlin: Springer-Verlag.
- Rousseeuw, P. J. (1987). *J. Comput. Appl. Math.* **20**, 53–65.
- Sato, M., Sato, Y. & Jain, L. C. (1966). *Fuzzy Clustering Models and Applications*. New York: Physica-Verlag.
- Storey, R., Docherty, R., Higginson, P., Dallman, C., Gilmore, C., Barr, G. & Dong, W. (2004). *Crystallogr. Rev.* **10**, 45–56.
- Wilson, A. J. C. (1963). *Mathematical Theory of X-ray Powder Diffractometry*. New York: Gordon and Breach.