Journal of
Applied
Crystallography

ISSN 0021-8898

Received 9 January 2004
Accepted 7 June 2004

High-throughput powder diffraction. III. The application of full-profile pattern matching and multivariate statistical analysis to round-robin-type data sets

Gordon Barr,[a] Wei Dong,[a] Christopher Gilmore[a]* and John Faber[b]

[a]Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland, and [b]International Center for Diffraction Data, 12 Campus Boulevard, Newton Square, Pennsylvania 19073-3273, USA. Correspondence e-mail: chris@chem.gla.ac.uk

Powder pattern matching techniques, using all the experimentally measured data points, coupled with cluster analysis, fuzzy clustering and multivariate statistical methods are used, with appropriate visualization tools, to analyse a set of 27 powder diffraction patterns of alumina collected at seven different laboratories on different instruments as part of an International Center for Diffraction Data Grant-in-Aid program. In their original form, the data factor into six distinct clusters. However, when a non-linear shift of the form $\Delta(2\theta) = a_0 + a_1 \sin\theta$ (where a_0 and a_1 are refinable constants) is applied to optimize the correlations between patterns, clustering produces a large 25-pattern set with two outliers. The first outlier is a synchrotron data set at a different wavelength from the other data, and the second is distinguished by the absence of $K\alpha_2$ lines, *i.e.* it uses Ge-monochromated incident X-rays. Fuzzy clustering, in which samples may belong to more than one cluster, is introduced as a complementary method of pinpointing problematic diffraction patterns. In contrast to the usual methodology associated with the analysis of round-robin data, this process is carried out in a routine way, with minimal user interaction or supervision, using the *PolySNAP* software.

1. Introduction

In three previous papers (Gilmore *et al.*, 2004, subsequently referred to as I; Barr *et al.*, 2004*a*, subsequently referred to as II, and Storey *et al.*, 2004), we have presented a series of techniques for processing and matching powder diffraction data generated from high-throughput experiments using the full pattern profiles. We have shown that the data may be partitioned into sets by generating a correlation matrix derived from matching all the powder patterns with each other, and then applying the relevant techniques of multivariate statistics and classification. In this way unusual or unexpected patterns can be readily identified, even if there are more than 1000 patterns present for a wide rage of polymorphs and solvates. However, the methods are equally applicable to other data, and we present here an analysis of a small set of 27 patterns collected on alumina in a Grant-in-Aid program organized by the International Center for Diffraction Data (ICDD) using the computer program *PolySNAP* (Barr *et al.*, 2004*b*,*c*). Although such a data set could be processed manually, this process points the way to handling large data sets.

There is a continuing effort by the ICDD to ensure that new patterns being added to the powder diffraction file (PDF) contain a significant proportion of phases that represent current needs and trends in industry and research. The effort is implemented, in part, by sponsoring a Grant-in-Aid (GiA) program, which is a competitive financial assistance package designed to encourage scientists working on new phases to submit high-quality diffraction data for inclusion in the PDF, and also for the production of new patterns of phases of current interest or the preparation of the phases themselves. In 2002, GiAs were awarded to ~60 universities and research laboratories from 23 different countries for the collection of new and improved data on compounds currently under study. This has created a continual flux of new and potentially technologically relevant entries (approximately 800–1000 patterns) in the PDF.

As an additional support feature in the GiA program, NIST 1976 corundum plate samples are distributed to all GiA recipients. The ICDD requests that a protocol be established to submit NIST 1976 reference material results along with submission data. Digitized diffraction data are received by the ICDD and reviewed on a periodic basis. The historical records are used to track instrument alignment, instrumental resolution *etc*. The 27-sample set used here was chosen arbitrarily from these reference records.

DOI: 10.1107/S0021889804013743

2. The method

A brief summary of the method may be useful at this point. Papers I and II contain full details.

Data can be imported in a variety of formats. Each pattern is interpolated, if necessary, to give increments of 0.02° in 2θ using local fifth-order polynomials and Neville's algorithm (Press *et al.*, 1992). Background removal is optional and, where used, employs local nth-order polynomial functions (where n is selected by the algorithm), which are fitted to the data and then subtracted to produce a pattern with a flat baseline. Smoothing of the data is carried out using wavelets *via* the SURE (Stein's unbiased risk estimate) thresholding procedure (Donoho & Johnstone, 1995). Peak positions are found using Savitsky–Golay filtering (Savitzky & Golay, 1964).

Powder patterns are treated as bivariate samples with n measured points $[(x_1, y_1), \ldots, (x_n, y_n)]$. Patterns are compared with each other using a weighted mean of parametric and non-parametric correlation coefficients (the Pearson and Spearman coefficients, respectively) using every measured intensity data point. The Pearson coefficient is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2\right]^{1/2}}, \qquad (1)$$

where x_i and y_i are the measured data points for the two patterns. In contrast, the Spearman coefficient is defined as

$$\rho_{xy} = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n\left(\frac{n+1}{2}\right)^2}{\left[\sum_{i=1}^{n} R(x_i)^2 - n\left(\frac{n+1}{2}\right)^2\right]^{1/2} \left[\sum_{i=1}^{n} R(y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right]^{1/2}}, \qquad (2)$$

where $R(x_i)$ and $R(y_i)$ are the ranks of the data points rather than their values. From these two coefficients a weighted mean, r_w, is calculated, and from this a correlation matrix, $\boldsymbol{\rho}$, can be derived in which every pattern is correlated with every other. This matrix can be converted to a distance matrix, $\mathbf{d}$, using the relationship

$$d_{ij} = 0.5\left(1.0 - \rho_{ij}\right), \quad 0.0 \leq d_{ij} \leq 1.0, \qquad (3)$$

or a similarity matrix, $\mathbf{s}$, where

$$s_{ij} = 1.0 - d_{ij}/\left(d_{ij}\right)_{\max}, \quad 0.0 \leq s_{ij} \leq 1.0. \qquad (4)$$

In this way, highly correlated patterns with large correlation coefficients give small corresponding distances or high similarity coefficients and *vice versa.* We then use the matrices $\mathbf{d}$, $\mathbf{r}$ and $\mathbf{s}$ as input to a set of clustering and multivariate data analysis methods with associated visualization tools:

(i) Cluster analysis, which partitions the patterns into individual clusters or sets defined by their similarity.

(ii) Estimation of the number of clusters present.

(iii) Three-dimensional data plots derived from either metric multidimensional scaling (MMDS) or three-dimensional score plots from principal-components analysis (PCA). Each sphere in this plot represents a powder diffraction sample; the further the distance apart of the spheres the greater the corresponding distance as measured by (3) and the lower the corresponding correlation.

Table 1
Summary of the 27 data sets used in the analysis.

The initial cluster number is the cluster to which the pattern is assigned before non-linear shifts are applied to the data. The cluster number after shifting is in column 5. Relevant details concerning individual data sets are in column 6.

Sample number	Start 2θ (°)	Finish 2θ (°)	Initial cluster No.	Cluster No. after shifts	Comments
1	4	100	1	3	Investigator *A*
2	4	100	2	3	Investigator *A*
3	4	100	2	3	Investigator *A*
4	4	100	2	3	Investigator *A*
5	4	100	2	3	Investigator *A*
6	4	100	2	3	Investigator A
7	4	100	2	3	Investigator *A*
8	4	100	1	3	Investigator *A*
9	4	100	1	3	Investigator *A*
10	9	71	3	2	Investigator *B*; synchrotron data set, λ = 0.7907 Å
11	10	159	4	3	Investigator *C*
12	2	158	4	3	Investigator *C*
13	2	158	4	3	Investigator *C*
14	5	75	5	3	Investigator *D*
15	5	75	5	3	Investigator *D*
16	2	150	4	3	Investigator *E*
17	2	150	4	3	Investigator *E*
18	2	150	4	3	Investigator *E*
19	2	150	4	3	Investigator *E*
20	2	150	4	3	Investigator *E*
21	2	150	4	3	Investigator *E*
22	2	150	4	3	Investigator *E*
23	2	150	4	3	Investigator *E*
24	2	150	4	3	Investigator *E*
25	2	150	4	3	Investigator *E*
26	5	95	1	3	Investigator *C*, 3 years later than 11–13
27	2	100	6	1	Ge monochromator; only $K\alpha_1$ radiation

3. Data

The data comprised 27 patterns for corundum. In terms of background and peak noise levels, they were of relatively high quality and consequently no wavelet smoothing or background subtraction was carried out. The data were collected in Bragg–Brentano geometry. Every sample used 0.02° increments in 2θ, and so no data interpolation was used. This minimal data pre-processing is ideal; trials involving the removal of backgrounds and/or smoothing resulted in no significant difference in the results. Table 1 summarizes the 2θ measurement ranges of the data, including details of the investigator and other relevant experimental options. Six investigators were involved over a period of three years; all the data come from laboratory sources, except data set 10, which comes from a synchrotron using a wavelength of 0.7907 Å. All non-synchrotron data were collected without a

Table 2
Estimating the number of clusters (*a*) before applying non-linear pattern shifts and (*b*) after the non-linear shifts have been applied.

(*a*) The maximum estimate of the number of clusters is 7; the minimum estimate is 3; the combined weighted estimate of the number of clusters is 6, and the median value is 7. Only the tests that were able to find suitable estimates are quoted; missing test values result from the lack of optimum points using the tests. CH is the Calinški–Harabasz statistic (Calinški & Harabasz, 1974).

Principal-components analysis (non-transformed matrix)	4
Principal-components analysis (transformed matrix)	3
Metric multidimensional scaling	5
CH statistic using single linkage	7
CH statistic using group averages	7
CH statistic using Ward method	7
CH statistic using complete linkage	7

(*b*) The maximum estimate of the number of clusters is 6; the minimum estimate is 2; the combined weighted estimate of the number of clusters is 3, and the median value is 3.

Principal-components analysis (non-transformed matrix)	3
Principal-components analysis (transformed matrix)	2
Metric multidimensional scaling	6
CH statistic using single linkage	3
CH statistic using complete linkage	3

monochromator, except for set 27, which was collected with a Ge monochromator with only $K\alpha_1$ radiation present. Cu radiation was used throughout (except for sample 10).

4. Results

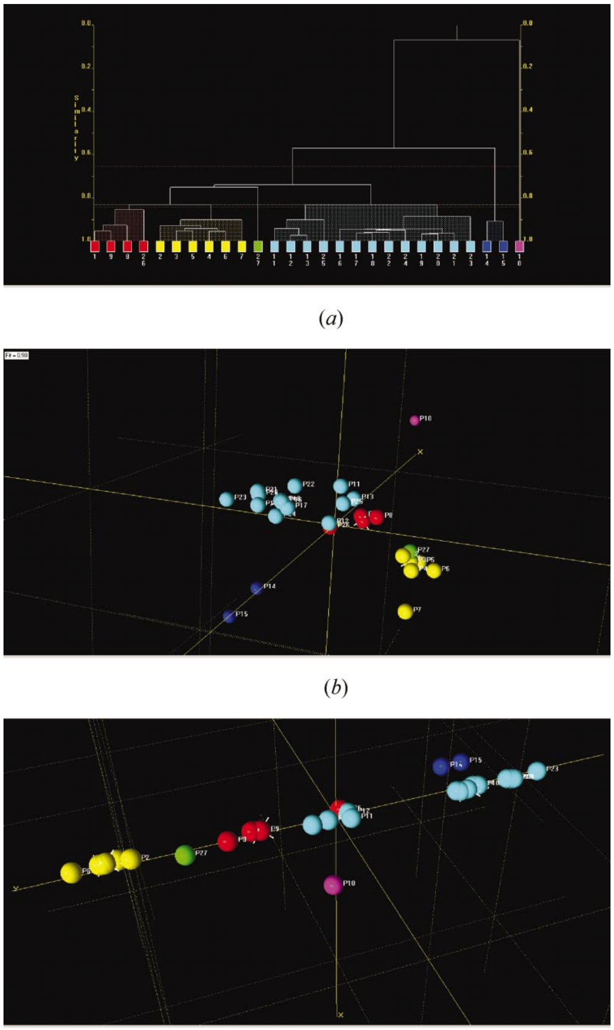
We present two sets of results. The first uses the data without any optimal 2θ shift, and in the second analysis each pattern is optimally shifted with respect to every other.

4.1. Unshifted data

The 27 patterns were used to generate a correlation matrix $\boldsymbol{\rho}(27\times27)$ using the unweighted mean of the Spearman and Pearson correlation coefficients computed using every measured data point in the profile, not just the peaks. Where the measurement ranges of the two patterns being correlated were not the same, only the overlapping 2θ range was used. The results are summarized in Tables 1 and 2(*a*) and in Fig. 1.

To examine the results, we commence with the dendrogram shown in Fig. 1(*a*). Each of the 27 diffraction patterns begins at the bottom of this plot in a separate class, and these amalgamate in stepwise fashion and become linked by horizontal tie bars. The height of the tie bar represents the similarity between the samples as measured by the relevant distance statistic. Sample 10 in this case is the least tightly linked, whereas, in complete contrast, patterns such as 12 and 13 are very tightly coupled and thus very similar. The dendrogram technique used was the group average link method (paper II) which was chosen automatically by the *PolySNAP* program using the maximal consistency algorithm also described in the same paper.

It is useful to be able estimate the number of clusters present and thus 'cut' the dendrogram in an optimal way, so that all the tie lines above the cut line are ignored and only the connections below this line are retained. This process results in the partitioning of the data into clusters. Accurately determining cluster numbers is difficult (see, for example, Meloun *et al.*, 2000). We used estimates based on the eigenanalysis of the correlation and related matrices integrated with techniques based on cluster analysis, developed by Goodman & Kruskal (1954), Calinški & Harabasz (1974) and Milligan & Cooper (1985) (see paper II). The individual estimates of cluster numbers are shown in Table 2(*a*). Only those methods that yielded an optimum value are listed; several of the tests gave no usable indication. The proposed cut line is shown on the dendrogram in Fig. 1(*a*) as a horizontal yellow line and results in the definition of six clusters. The tie bars in the



(*a*)

(*b*)

(*c*)

Figure 1
(*a*) Dendrogram for unshifted powder diffraction data. The optimum cut level partitions the data into six clusters. Pattern 10 is the least well joined pattern. (*b*) The MMDS plot; each sphere in this plot represents a powder diffraction sample; the greater the separation of the spheres, the smaller the corresponding correlation. (*c*) The three-dimensional PCA plot. The distance properties mirror those of the MMDS plot, although the shape is different. The colour of each sphere in (*b*) and (*c*) is taken from the dendrogram to allow comparison of the methods.

dendrogram lie at low levels, indicating a high degree of similarity between the samples, even when they are in different clusters. The one exception involves pattern 10, which is minimally connected to the rest of the data.

The data can also be summarized using three-dimensional plots derived from either metric multidimensional scaling (MMDS) or three-dimensional score plots from principal-components analysis (paper II). These act independently of the dendrogram. The MMDS plot is shown in Fig. 1(*b*). Each sphere in this plot represents a powder diffraction sample; the greater the separation of the spheres the greater the corresponding distance as measured by (1) and the lower the corresponding correlation. The colour of each sphere is taken from the dendrogram, but there is no other interdependence.

It can be seen that the patterns form natural clusters, some of which correspond to the investigator. Thus patterns 14 and 15 form a natural set and this can also be seen in the MMDS plot. The MMDS calculation correlates well with the observed distance matrix from pattern matching, with a correlation coefficient of 0.98. The PCA plot (Fig. 1*c*) is less clear, however; it tends to one-dimensionality. This is not always the case with the method, however, and it still clusters the data in an appropriate way. Patterns 2–7 also form a set and these all come from investigator *A*. The remaining three patterns from this researcher are numbers 1, 8 and 9; in the MMDS and PCA plots these lie close to the 2–7 set. Patterns 16–25 belong to investigator *E*, and these also form a set with the addition of patterns 11–13 from investigator *C*. The rationale of this clustering can be seen by visual inspection of the diffraction data: the patterns are almost identical.

The dendrogram also presents strong evidence that pattern 10 is less well linked than the others, and in both three-dimensional plots the sphere corresponding to this pattern is wholly isolated from other patterns. This result is discussed further in §6.

In general, the partitioning of the data using these different methods is remarkable, given the very close similarities between the pattern profiles.

4.2. Pattern shifts

One of the commonest sources of systematic error in matching powder patterns is a consequence of θ shifts arising from variability of the zero point, instrumental setup, sample height, transparency *etc.* There is a full discussion of this topic by Wilson (1963), Zevin & Kimmel (1995, ch. 3) and Jenkins & Snyder (1996). *PolySNAP* provides three possible corrections, although this by no means encompasses all the possible correction geometries that can arise. These take the form

$$\Delta(2\theta) = a_0 + a_1 \cos\theta, \tag{5}$$

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

$$\Delta(2\theta) = a_0 + a_1 \sin\theta, \tag{6}$$

which corrects for transparency errors or, for example, transmission geometry with constant specimen–detector distance, and

$$\Delta(2\theta) = a_0 + a_1 \sin 2\theta, \tag{7}$$

which provides transparency and thick-specimen error corrections. The parameters a_0 and a_1 are constants that can be determined by maximizing the correlation between patterns (paper I, §4; Barr *et al.*, 2003, 2004). It is of course possible to combine equations (5)–(7) into a single expression involving four constants and the trigonometric functions $\cos\theta$, $\sin\theta$ and $\sin 2\theta$. This poses problems of computer times and potential high correlations between coefficients, which will be explored in later versions of the *PolySNAP* computer program. Given the complexities of this problem, an argument can be made for selecting a suitable function on the criterion of improving pattern–pattern correlations. It is difficult to obtain suitable expressions for the derivatives $\partial a_0/\partial r_w$ and $\partial a_1/\partial r_w$ for use in the optimization, so we use the downhill simplex method (Nelder & Mead, 1965) to obtain values of a_0 and a_1 in all the cases (5)–(7).

This process does not require the calculation of derivatives. Both a_0 and a_1 were constrained to lie between ±0.4. There can be problems with the high correlations between a_0 and a_1. The use of the downhill simplex method with full-pattern correlation coefficients seems to be robust in this respect.

For the 27 patterns under study, there was an increase in peak separation with increasing 2θ, which indicates a correction using either (6) or (7). Before the application of (6), the mean correlation coefficient (excluding pattern 10) between the 26 patterns was 0.75. Again excluding pattern 10, refinement of a_0 and a_1 *via* the downhill simplex method gave $-0.21 \le a_0 \le 0.07$ and $-0.21 \le a_1 \le 0.28$, and the mean correlation coefficient increased to 0.83. For (7) there was very little change in the correlation or the associated clustering, and visual inspection of the pattern matching confirmed that the use of (6) was optimal.

Before any shifts were applied, the mean correlation coefficient for pattern 10 with the other 26 patterns was 0.097, with a maximum value of 0.13 and a minimum value of −0.067. After optimal shifts were calculated, the mean correlation coefficient for pattern 10 was 0.13, with a maximum value of 0.19 and a minimum value of 0.032. This result indicates the unique status of this pattern, its lack of linkage to the other 26 patterns in the data and the fact that the problem does not arise from the experimental setup.

The resulting dendrogram is shown in Fig. 2(*a*); all the patterns now belong to a single cluster, except numbers 10 and 21. Table 2(*b*) shows the available estimates of the number of clusters present; the median value is now three, with a variation from 2 to 6.

The MMDS plot is shown in Fig. 2(*b*) and the three-dimensional PCA score plot is shown in Fig. 2(*c*). The two plots have a very similar form, and the data are clustered much closer than in the unshifted case. 25 patterns are deemed to belong to a single cluster, with only patterns 10 and 27 in clusters of their own. Pattern 19 is the most representative

Table 3
The results of fuzzy clustering calculations after non-linear shifts to maximize the correlations between patterns.

The entries are the membership or possibility values, u_{ij}, where i is the pattern number and j the cluster number; $i = 1, \ldots, 27$; $j = 1, \ldots, 3$. Two fuzzy clustering methods are employed: additive and using aggregation operators. All values of $u_{ij} > 0.7$ are highlighted.

	Using additive clustering			Using aggregation operators		
Pattern number, i	Cluster k = 1 (u_{i1})	Cluster k = 2 (u_{i2})	Cluster k = 3 (u_{i3})	Cluster k = 1 (u_{i1})	Cluster k = 2 (u_{i2})	Cluster k = 3 (u_{i3})
1	0.02	0.05	**0.84**	0.05	0.04	**0.99**
2	0.02	0.05	**0.83**	0.06	0.02	**0.97**
3	0.02	0.05	**0.84**	0.06	0.02	**0.99**
4	0.02	0.05	**0.84**	0.06	0.02	**0.98**
5	0.02	0.05	**0.85**	0.06	0.03	**1.00**
6	0.02	0.05	**0.84**	0.06	0.03	**0.98**
7	0.02	0.03	**0.85**	0.06	0.00	**0.99**
8	0.02	0.05	**0.82**	0.06	0.03	**0.96**
9	0.02	0.05	**0.84**	0.06	0.04	**0.98**
10	0.00	**0.97**	0.10	0.00	**1.00**	0.11
11	0.00	0.00	**0.90**	0.04	0.03	**0.98**
12	0.00	0.00	**0.91**	0.05	0.02	**1.00**
13	0.00	0.00	**0.91**	0.04	0.04	**0.99**
14	0.00	0.01	**0.81**	0.04	0.05	**0.93**
15	0.00	0.00	**0.80**	0.04	0.03	**0.92**
16	0.00	0.00	**0.94**	0.04	0.00	**1.00**
17	0.00	0.00	**0.95**	0.06	0.00	**1.00**
18	0.03	0.00	**0.96**	0.06	0.01	**1.00**
19	0.02	0.00	**0.94**	0.06	0.00	**1.00**
20	0.01	0.00	**0.94**	0.06	0.01	**1.00**
21	0.01	0.00	**0.94**	0.06	0.01	**1.00**
22	0.00	0.01	**0.95**	0.05	0.02	**1.00**
23	0.00	0.00	**0.93**	0.04	0.00	**0.99**
24	0.02	0.00	**0.95**	0.06	0.00	**1.00**
25	0.00	0.00	**0.91**	0.03	0.02	**0.99**
26	0.01	0.01	**0.86**	0.05	0.03	**0.98**
27	**0.97**	0.03	**0.73**	**0.89**	0.00	**0.80**

sample of the large cluster, as defined as that pattern having the minimum mean distance from all other patterns in the same cluster (paper II).

It is possible to modify the dendrogram cut line manually from its initially calculated position. It should be remembered that this is a perfectly valid procedure since the estimation of the number of clusters present is at best an imperfect procedure, and the program estimates vary from 2 to 6 for this calculation. When the cut line is lowered to a similarity level of *ca* 0.88 the data are partitioned into 6 sets, as shown in Fig. 3(*a*). The first set includes patterns 1–9 from investigator *A*; the second contains 13 patterns comprising all those from investigators *C* and *E*; the patterns from *D* are also clustered, and patterns 10, 26 and 27 are isolated. Pattern 26 comes from investigator *C* three years after the measurements composing 11–13. The partitioning is now almost perfect. The MMDS and PCA plots in Figs. 3(*b*) and 3(*c*) also show this to be a natural partition of the data.

It now remains to investigate the relationship of patterns 10 and 27 to the remainder of the data. Whereas for simple cases like this visual inspection of the patterns will suffice, for large data sets visual inspection could be difficult. In addition, this case provides an excellent opportunity to use another classification technique: fuzzy sets and clusters.

5. Fuzzy sets

In standard clustering methods we partition a set of n objects (or patterns) into c disjoint sets or clusters. We can express this partitioning *via* a cluster matrix, $\mathbf{U}(n \times c)$, where u_{ik} represents the membership of pattern i of cluster k and is equal to unity if i belongs to c and zero otherwise, *i.e.*

$$u_{ik} \in [0, 1] \quad (i = 1, \ldots, n;\ \ k = 1, \ldots, c). \tag{8}$$

If we relax these constraints and insist only that

$$0 \le u_{ik} \le 1 \quad (i = 1, \ldots, n;\ \ k = 1, \ldots, c), \tag{9}$$

$$0 < \sum_{i=1}^{n} u_{ik} < n \quad (k = 1, \ldots, c) \tag{10}$$

and

$$\sum_{k=1}^{c} u_{ik} = 1, \tag{11}$$

then we have the concept of fuzzy clustering or fuzzy sets, in which we have the possibility that a pattern can belong to more than one cluster (see, for example, Gordon, 1999; Sato *et al.*, 1966). Such a situation is quite possible in the case of powder diffraction when mixtures can be involved.

If the restraint represented by (11) is omitted then the u_{ik} values are sometimes referred to as 'possibilities'. We will use this option in this paper.

The calculation of the **U** matrix is not simple, and we have explored two methods as discussed in detail by Sato *et al.* (1966) for ordinal data, as follow.

(*a*) Additive clustering in which the **U** matrix is determined by minimizing

$$\eta_1^2 = \frac{\sum\limits_{i \neq j=1}^{n} \left(s_{ij} - \alpha \sum\limits_{k=1}^{c} u_{ik} u_{jk} \right)^2}{\sum\limits_{i \neq j=1}^{n} \left(s_{ij} - \bar{s} \right)^2}, \tag{12}$$

where

$$\bar{s} = [1/n(n-1)]^{-1} \sum_{i \neq j=1}^{n} \left(s_{ij} \right) \tag{13}$$

and α is a scaling constant that scales **s** and **U**; **s** is the similarity matrix defined *via* (4). Sato *et al.* (1966) recommend random values of u_{ij} as a starting point, followed by a form of steepest descents to obtain optimal values. With powder diffraction data, we have found that it is much faster in terms of computing time to use the initial cluster assignments from the dendrogram: if powder pattern i is deemed to belong to cluster j the initial value of u_{ij} is 0.8; otherwise it is given a random value scaled such that, for cluster j,

$$\sum_{i=1}^{n} u_{ij} = 1.0, \tag{14}$$

although this normalization condition is not imposed in the subsequent calculations. A steepest-descents method is used for minimizing (12).

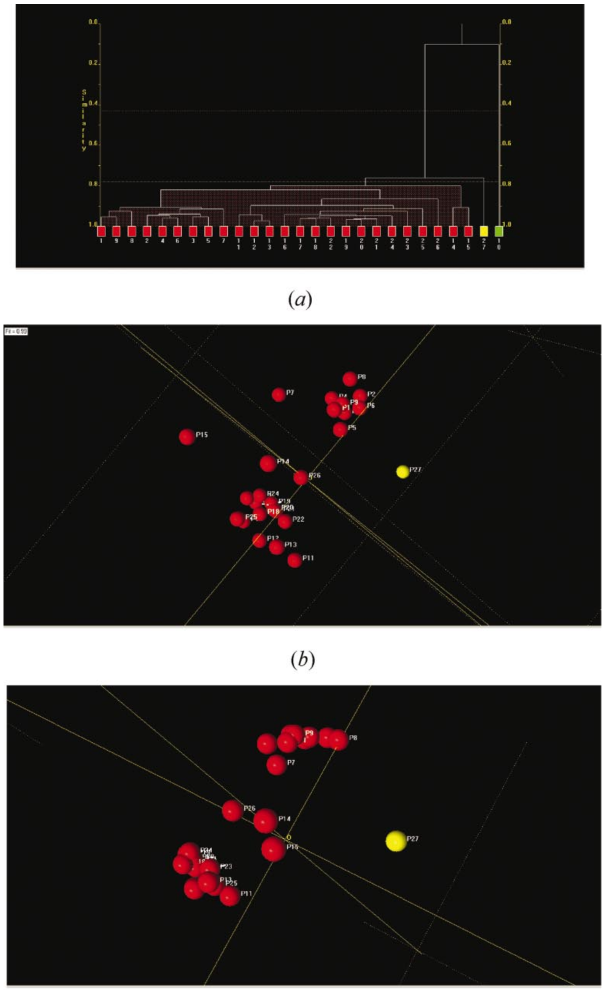
(*b*) The use of a more general algorithm using aggregation operators. In this case we minimize

$$J=\sum_{i\neq j=1}^{c}\left[s_{ij}-\sum_{k=1}^{c}\min\left(u_{ik},u_{jk}\right)\right] \qquad (15)$$

to obtain the membership or possibility values u_{ik}. We use the same starting point as (*a*), and again use a steepest descents method in the optimization of J in (15).

The results of both calculations are shown in Table 3, using a similarity matrix derived from the optimally shifted data. These results are highly informative. Pattern 10 has cluster membership {0.00, 0.97, 0.1} from additive clustering and {0.00, 1.00, 0.11} from aggregation methods, *i.e.* it belongs almost exclusively to cluster number 2. (Small values of u_{ik}, less than 0.25, have only marginal significance.) No other pattern has any significant membership, u_{j2}, for cluster number 2, reinforcing the point that this sample constitutes an isolated data set. All the other patterns, except 10, have values of $u_{j3} > 0.70$, indicating that they all belong to a single cluster (in this case cluster 3 and no other). The exception to this rule is pattern 27; the membership values are {0.97, 0.03, 0.73} and {0.89, 0.00, 0.80} from the two clustering methods. This result indicates that 27 belongs both to the large cluster and to a cluster of its own, no other pattern having a significant value of u for cluster 1. In other words, while having similarities to the large cluster, it also has some unique features not displayed by other diffraction patterns, and these are discussed in §6.

Fuzzy clustering is an easily calculated semi-independent way of assessing patterns that do not conform to the dominant

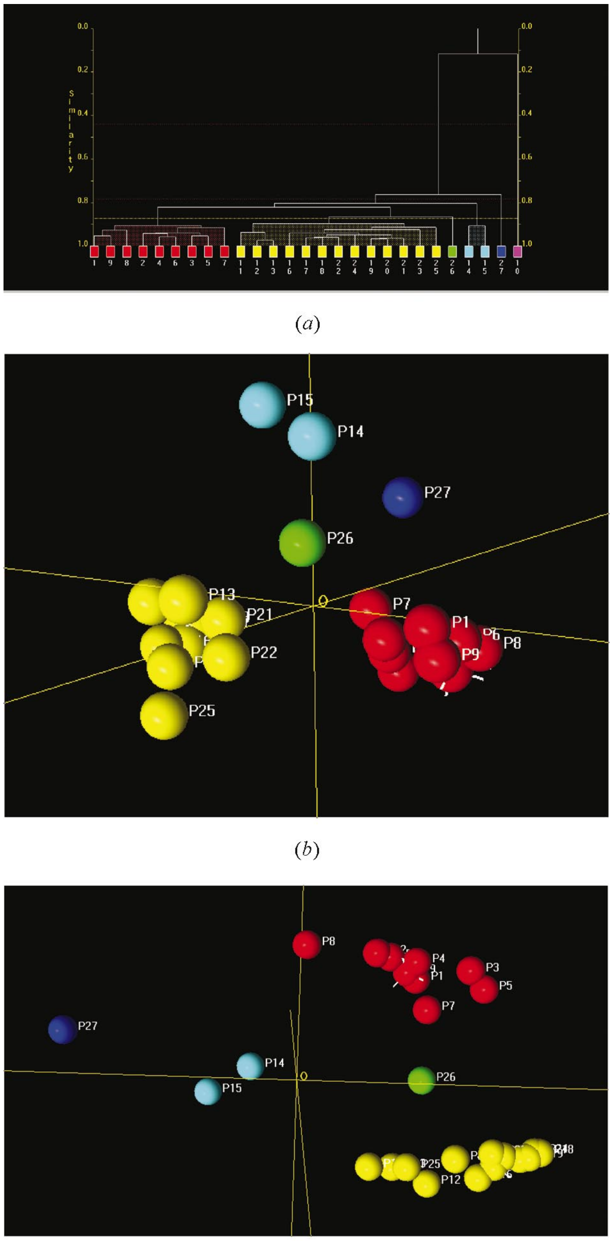


(*a*)

(*b*)

(*c*)

Figure 2
(*a*) Dendrogram where an optimum shift has been calculated. The optimum cut level partitions the data into three clusters. The red cluster contains all but two of the diffraction patterns. (*b*) The MMDS plot. (*c*) The three-dimensional PCA plot. The colour of each sphere in (*b*) and (*c*) is taken from the dendrogram.



(*a*)

(*b*)

(*c*)

Figure 3
(*a*) Dendrogram where an optimum shift has been calculated and the cut line has been reduced to a similarity level of 0.87, thus partitioning the data into six clusters. (*b*) The MMDS plot. (*c*) The three-dimensional PCA plot.

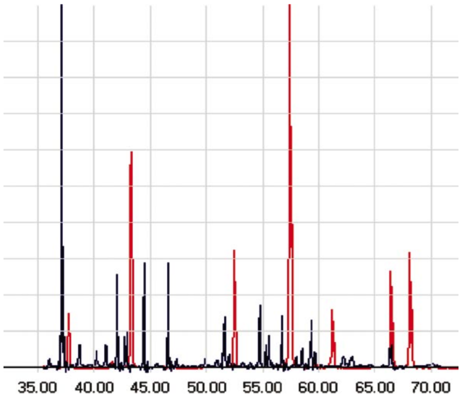


Figure 4
Pattern 10 (in blue) compared with the most representative pattern of the set (number 19, in red). The lack of correspondence is obvious. The Pearson correlation coefficient between the two patterns is −0.013 and the Spearman correlation coefficient is −0.121.

trend of the data. The method will be discussed thoroughly with other, more complex, examples in a future paper in this series.

6. Patterns 10 and 27

We now visually inspect the individual patterns that appear to behave anomalously. Pattern 10 is easily dealt with; it is measured on a synchrotron with an incident X-ray wavelength of 0.7907 Å. The difference is shown in Fig. 4, where pattern 10 is superimposed on the most representative pattern, number 19, and there is no point of correspondence whatsoever. There are two reasons for including pattern 10: (*a*) it demonstrates that the mathematical methods retain their sensitivity to pattern differences even in the presence of outliers, and (*b*) the use of fuzzy clustering.

The situation regarding pattern 27 is more subtle. The *PolySNAP* computer program permits the comparison of two patterns on a peak-by-peak basis; this process was carried out for patterns 19 and 27, and typical results are shown in Fig. 5. The source of the discrepancy is now clear and lies in the monochromation of the incident X-rays. Sample 27 was collected with the use of an incident Ge monochromator and the missing Cu $K\alpha_1$–$K\alpha_2$ splitting is clear. It is significant that this methodology can automatically identify such differences even though they are relatively small.

7. Conclusions

We have presented a method for routine analysis of a typical set of data produced by a round-robin or related set of experiments in which different laboratories collect data on the same sample. We have shown that the techniques hitherto used for analysis in high-throughput crystallography are equally applicable here when examining data sets that are very similar and where differences are expected to be small. The calculation of non-linear 2θ shifts to optimize pattern correlation plays a key role in identifying anomalous data sets, and when these shifts are applied the method becomes sensitive to quite small differences in the patterns; the absence or presence of Cu $K\alpha_1$–$K\alpha_2$ splitting, for example, is clearly indicated. This is a small data set, and the computations can be carried out manually, but it is easy to see that the technique will easily scale to hundreds and even thousands of data sets where manual inspection is not possible. The visual tools associated with dendrograms, MMDS and three-dimensional PCA, are essential components of the analysis and also make it easy to find anomalous data sets, as well as to visualize the clusters they form. The method also has obvious applications for quality control. Fuzzy clustering has great potential, especially as a potential tool for identifying mixtures.

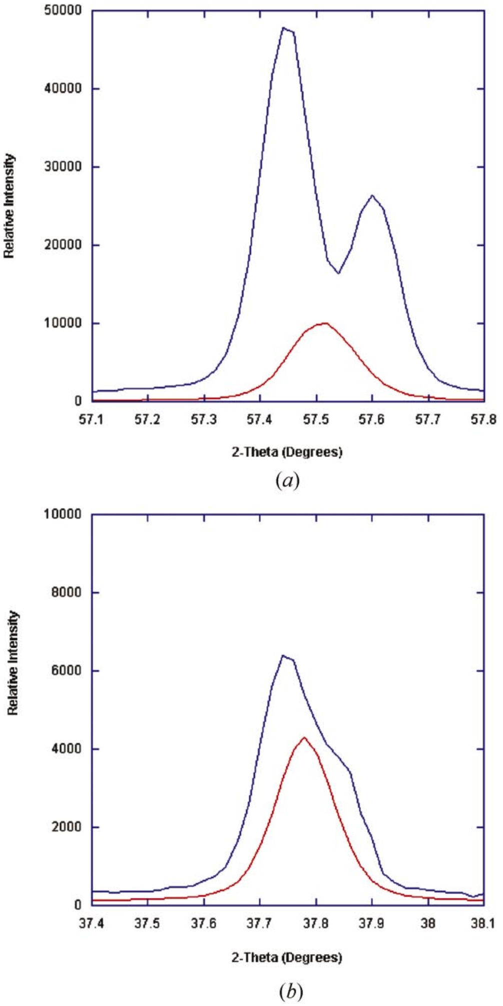


Figure 5
Two individual peak comparisons between pattern 27 (in red) and pattern 19 (in blue). The latter is the most representative sample of the large 25-sample set of patterns. The Cu $K\alpha_1$–$K\alpha_2$ splitting is present in pattern 19 but not in 27. The patterns have not been shifted relative to each other.

We thank the contributors to the ICDD Grant-in-Aid program for their data and Dr Lowe-Ma for very helpful discussions.

References

Barr, G., Dong, W. & Gilmore, C. J. (2004*a*). *J. Appl. Cryst.* **37**, 243–252.

Barr, G., Dong, W. & Gilmore, C. J. (2004*b*). *J. Appl. Cryst.* **37**, 658–664.

Barr, G., Dong, W. & Gilmore, C. J. (2004*c*) *PolySNAP*: *a Computer Program for the Analysis of High Throughput Powder Diffraction Data.* University of Glasgow, Scotland. (See also http://www.chem.gla.ac.uk/staff/chris/snap.html.)

Barr, G., Gilmore, C. J. & Paisley, J. (2003). *SNAP-1D*: *Systematic Non-parametric Analysis of Patterns – a Computer Program to Perform Full-Profile Qualitative and Quantitative Analysis of Powder Diffraction Patterns.* University of Glasgow, Scotland, and Bruker AXS Inc., Madison, Wisconsin, USA. (See also http://www.chem.gla.ac.uk/staff/chris/snap.html.)

Barr, G., Gilmore, C. J. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 665–668.

Calinški, T. & Harabasz, J. (1974). *Commun. Stat.* **3**, 1–27.

Donoho, D. L. & Johnstone, I. M. (1995). *J. Am. Stat. Assoc.* **90**, 1200–1224.

Gilmore, C. J., Barr, G. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 231–242.

Goodman, L. A. & Kruskal, W. H. (1954). *J. Am. Stat. Assoc.* **49**, 732–764.

Gordon, A. D. (1999). *Classification*, 2nd ed. Boca Raton: Chapman and Hall/CRC.

Jenkins, R. & Snyder, R. L. (1996). *Introduction to X-ray Powder Diffractometry*, pp. 194–195. New York: John Wiley.

Milligan, G. W. & Cooper, M. C. (1985). *Psychometrika*, **50**, 159–179.

Meloun, M., Čapek, J., Mikšík, P. & Brereton, G. (2000). *Anal. Chim. Acta*, **20736**, 1–18.

Nelder, J. A. & Mead, R. (1965). *Comput. J.* **7**, 308–313.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C.* Cambridge University Press.

Sato, M., Sato, Y. & Jain, L. C. (1966). *Fuzzy Clustering Models and Applications.* New York: Physica-Verlag.

Savitzky, A. & Golay, M. J. E (1964). *Anal. Chem.* **36**, 1627–1639.

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.

Zevin, L. S. & Kimmel, G. (1995). *Quantitative X-ray Diffractometry.* New York: Springer-Verlag.