High-throughput powder diffraction. I. A new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles

Christopher J. Gilmore,* Gordon Barr and Jonathan Paisley‡

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

A new integrated approach to full powder diffraction pattern analysis is described. This new approach incorporates wavelet-based data pre-processing, non-parametric statistical tests for full-pattern matching, and singular value decomposition to extract quantitative phase information from mixtures. Every measured data point is used in both qualitative and quantitative analyses. The success of this new integrated approach is demonstrated through examples using several test data sets. The methods are incorporated within the commercial software program *SNAP-1D*, and can be extended to high-throughput powder diffraction experiments.

1. Introduction

The identification of unknown materials *via* X-ray powder diffraction patterns has until recently relied on simplified patterns in which the full diffraction profile is reduced to a set of point functions selected from the strongest normalized peaks. Each of these functions uses *d*-spacings (or 2θ values) and intensities (the *d–I* system) to represent the diffraction peaks. This simplified approach to the analysis of powder diffraction patterns has advantages primarily in computer storage requirements, and with respect to the speed of search algorithms especially in very large databases (ICDD, 2003). However, problems arise from the use of such data.

(i) Accurate determinations of the peak positions may be difficult to obtain, especially in cases where peak overlap occurs or there is significant peak asymmetry.

(ii) The hardware and sample preparation used can also affect the *d*-spacing (or 2θ value) that is recorded for the peak. Shoulders to main peaks and broad peaks can also be problematic.

(iii) There is an objective element in choosing the number of peaks to select. Different software packages produce a range of different numbers of peaks from an identical pattern. For example, an ICDD round robin using a standard corundum pattern returned values varying from 23 to 81 for the number of peaks, when the correct number was 42 (Jenkins, 1998).

(iv) Many weak peaks are discarded. This can affect quantitative analysis of mixtures if one component diffracts weakly or is present in small amounts.

(v) Sample preparation and instrumentation can induce significant differences in near-identical samples. Preferred orientation is a very difficult problem.

(vi) The reduction of the pattern to point functions can also make it difficult to design effective algorithms.

In order to use the extra information contained within the full profile, search–match algorithms are required that utilize each measured data point in the analysis. Recent drastic reductions in the price of computer storage, and corresponding increases in speed and processing power, means that storing and handling large numbers of full-profile data sets is much more practical than it would have been just a few years ago, and a new approach would be timely. However, databases of full profiles are not widely available.

2. Existing search–match software overview

Most existing search–match programs do not use the full profile data. Peak search and indexing programs are used firstly to extract a *d*-spacing and corresponding intensity for each identified peak, although indexing is not a pre-requisite. The pattern is thus reduced to a stick pattern. As an example of such pre-processing, see *N-TREOR* (Altomare *et al.*, 2000). The most popular search algorithm used with such 'stick' patterns is the Hanawalt search index (Hanawalt *et al.*, 1938). Based upon a method developed for manual search–match, this utilizes the eight strongest peak lines to identify the pattern. Likely matches are ranked using various figures of merit (FoM) or goodness of match (GoM) indicators (for example, see Johnson & Vand, 1967).

An intermediate approach between reduced-pattern matching programs and true full-profile programs, are programs that take a full-profile unknown pattern and

‡ Current address: Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland.

DOI: 10.1107/S002188980400038X **231**

compare it to a database of reduced patterns. An example of a computer program that includes such features is *DIFFRACT-AT* (Nusinovici & Winter, 1994). Patterns are assigned scores based upon a calculated figure of merit, and the best matches are displayed graphically, with their stick profiles superimposed over the unknown full profile for visual comparison and verification. The approach used allows small database peaks, which could potentially be obscured in the unknown profile by part of the full profile of a peak, not to be penalized as they would be in an approach based solely on a *d–I* system.

In contrast, true full-profile search–match programs compare full-profile unknowns to databases consisting of full profiles. As such databases are not yet commercially available, they must be either built up gradually from existing, often locally collected, experimental patterns, or generated from stick patterns by pattern simulation software (see for example Steele & Biederman, 1994).

The latter approach is that taken by *MATCHDB* (Smith *et al.*, 1991) where each unknown pattern data point is compared in turn with the corresponding database-pattern data point. Overall figures of merit for each database pattern are then calculated, and the top 15 matches are listed. The figures of merit used evaluate the patterns point-by-point in regions where the intensity is greater than a previously selected cut-off level. Several different proprietary full-profile search–match systems also exist, but since they are commercial products they are not discussed in any detail in the literature.

An excellent web site containing downloadable pattern-matching software is available (CCP14, 2003).

3. Qualitative pattern matching using the full diffraction pattern

Although much less dependent on the quality of data than reduced-pattern methods, the reliability of full-profile pattern matching can be improved by accurate pre-processing that involves smoothing and background removal. A flow chart of the process is shown in outline in Fig. 1.

3.1. Data pre-processing

Data are imported either as ASCII *xy* data (2θ, intensity), CIF format (Hall *et al.*, 1991) or a Bruker raw format. We have also developed a platform-independent binary format for this data that is used internally in the associated software. The data are normalized such that the maximum peak intensity is unity.

The pattern is interpolated if necessary to give increments of $0.02°$ in 2θ. High-order polynomials are used, employing Neville's algorithm (Press *et al.*, 1992).

To remove the background, local *n*th-order polynomial functions are fitted to the data, and then subtracted to produce a pattern with a flat baseline. The value of *n* is selected by the algorithm. Three domains are usually defined, but this can be modified for difficult cases.

Smoothing of the data is then carried out using wavelets (Gilmore, 1998; Smrčok *et al.*, 1999) *via* the *SURE* (Stein's unbiased risk estimate) thresholding procedure (Donoho & Johnstone, 1995; Ogden, 1997).

Peak positions are found using Savitsky–Golay filtering (Savitzky & Golay, 1964). Smoothing *via* a digital filter replaces each data point x_i with a linear combination of itself and a number of nearest neighbours. (This smoothing is distinct from the wavelet–*SURE* procedure and is only used to determine peak positions in the formalism that we use.) We can write any point g_i as a linear combination of the immediate neighbours:

$$g_i = \sum_{n=-n_l}^{n_r} c_n x_{i+n}. \tag{1}$$

Figure 1
A flowchart of data pre-processing before pattern matching. Items marked with an asterisk (*) are optional.
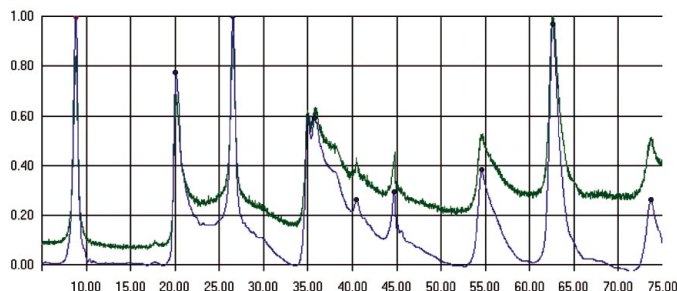

Figure 2
Pre-processing the powder data. The green line is the raw data. The blue line is the result of (*a*) removal of background using local *n*th-order polynomials, (*b*) smoothing *via* wavelets and the *SURE* procedure, and (*c*) peak searching using Golay–Savitsky filtering; peaks are marked with a bullet (●).
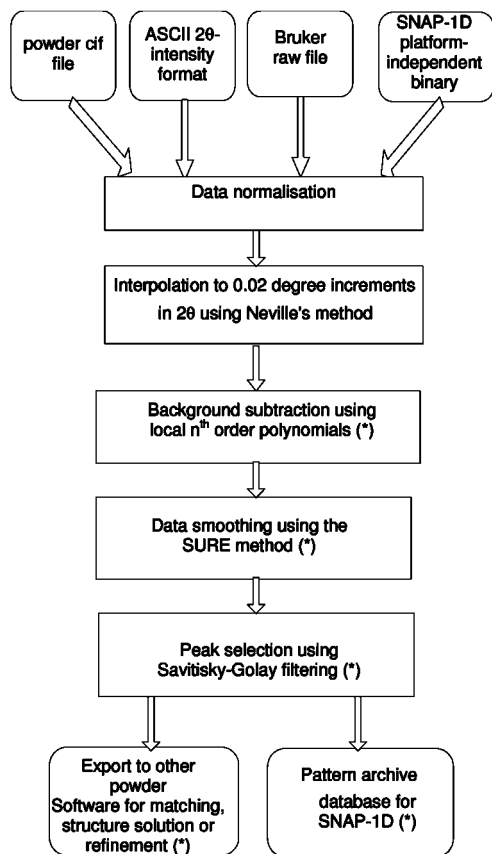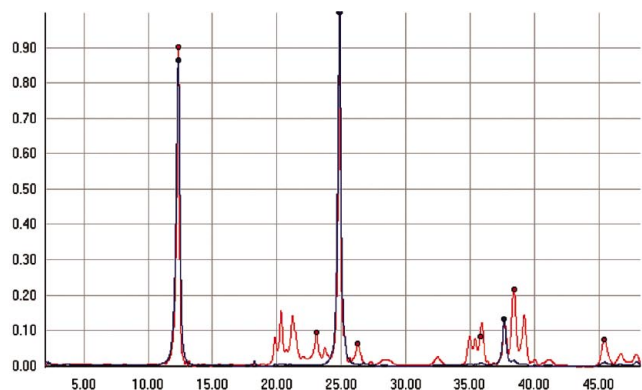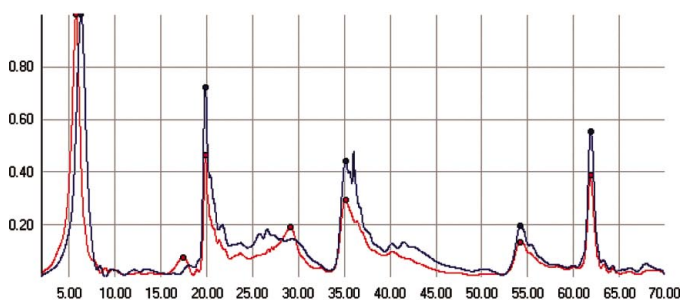
Savitsky–Golay filtering provides an efficient way to determine the coefficients c_n by the least-squares fit of a polynomial of degree M in i,
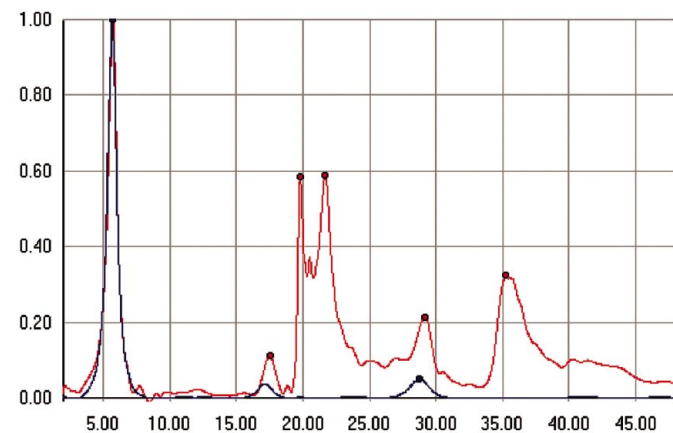
$$a_0 + a_1 i + a_2 i^2 + \ldots a_M i^M, \qquad (2)$$

(a)

(b)

(c)

Figure 3
The use of the Pearson (r) and Spearman (ρ) correlation coefficients. (a) r = 0.93; ρ = 0.68. The high value of r arises from the perfect match of the two biggest peaks around 12 and 25° in 2θ, but the much lower Spearman coefficient acts as a warning that there are unmatched regions in the two patterns. (b) r = 0.79; ρ = 0.90. The lower value of r is due to peak offsets around 6° and the peak at 29°. Visual inspection of the two patterns indicates a high degree of similarity however, which is reflected in the Spearman coefficient of 0.9. (c) r = 0.66; ρ = 0.22. The value of r reflects the peak at 6°; the low value of ρ indicates a poor match in other regions.

to the values $x_{-n_l}, \ldots x_{n_r}$. For finding peaks we need the first-order derivative and thus require a_1. To distinguish maxima and minima the gradient change is inspected. This procedure is robust with respect to noise, peak shape and peak width.

As an example, Fig. 2 shows the pre-processing of powder data for a clay mineral including normalization, the removal of background using local nth-order polynomials, followed by smoothing *via* wavelets, then peak searching.

3.2. Non-parametric statistics

The full-pattern-matching tools described here utilize, in part, non-parametric statistics. In general, non-parametric statistics are little used in crystallography where the statistical distributions are well defined or, at least, well approximated. In contrast, the use of non-parametric statistics involves no assumptions about the underlying distributions of data; instead it works using ranks. A set of n data points $x_1, x_1, \ldots x_n$ is represented by the data ranks in which the data are sorted into descending order and this order is used rather than the data value itself. Identical ranks are designated 'ties'. Correlation, for example, becomes a processing of correlating ranks. This has special advantages for comparing powder patterns on a point-by-point basis, since the distribution of the data is unknown. Furthermore, such statistics are robust and resistant to unplanned defects, outliers, *etc.* (see, for example, Conover, 1971). In the case of powders, this robustness will encompass peak asymmetry and preferred orientation.

The first step when dealing with non-parametric statistical tests is to convert the diffraction pattern from actual data values to the ranks of those values. If there are n data points in the pattern, the smallest intensity value is assigned a rank of 1
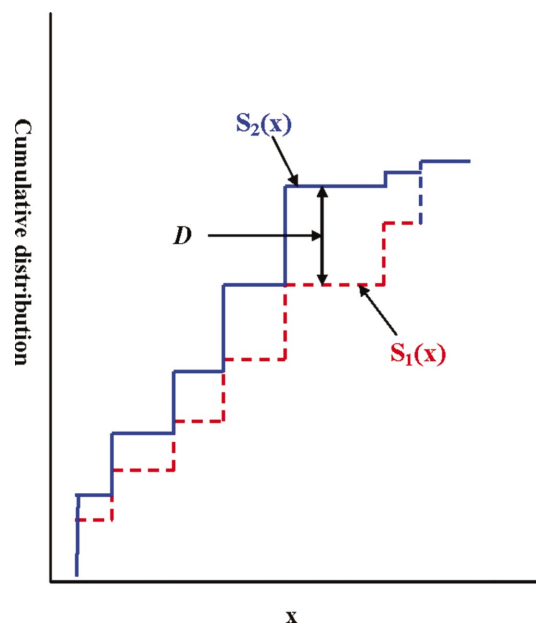

Figure 4
The Kolmogorov–Smirnov two-sample test. The two data sets are converted to ranks then further transformed to cumulative distributions, $S_1(x)$ and $S_2(x)$, and D is calculated as the maximum distance between $S_1(x)$ and $S_2(x)$. The associated probability is computed *via* equation (6).

$[R(x) = 1]$, the largest a rank of n $[R(x) = n]$ and the ith largest intensity a rank of I [denoted $R(x_i) = I$]. If any tied ranks exist (*i.e.* from data points of equal value) they are assigned a rank corresponding to the average value of the ranks they would have taken if they were not the same. Having transformed the data into such a form, non-parametric tests may then be applied.

3.3. Matching powder patterns

We employ up to four statistics for matching powder patterns with each other.

(i) The non-parametric Spearman rank over the full collected intersecting 2θ range employed on a point-by-point basis.

(ii) The Pearson correlation coefficient also taken over the same range.

(iii) The Kolmogorov–Smirnov test, also on a point-by-point basis, but only involving regions of the patterns where there are marked peaks.

(iv) The Pearson correlation coefficient that is the parametric equivalent of (iii).

Each statistic will now be discussed in turn.

3.4. Spearman's rank order coefficient

Consider two diffraction patterns, each with n measured points $n[(x_1, y_1), \ldots (x_n, y_n)]$. These are transformed to ranks $R(x_i)$ and $R(y_i)$. The Spearman test (Spearman, 1904) then gives a correlation coefficient ρ_{xy}, in the form (Conover, 1971; Press *et al.*, 1992)

$$\rho_{xy} = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n\left(\frac{n+1}{2}\right)^2}{\left[\sum_{i=1}^{n} R(x_i)^2 - n\left(\frac{n+1}{2}\right)^2\right]^{1/2}\left[\sum_{i=1}^{n} R(y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right]^{1/2}}. \quad (3)$$

This produces a coefficient in the range $-1 \leq \rho_{xy} \leq 1$. As with the conventional correlation coefficient, a score of zero would indicate no correlation between the two data sets. A negative score indicates anti-correlation, *i.e.* that large values of x are paired with small values of y, and *vice versa*. A positive score means large x values are paired with large y values, and *vice versa*. Usually the whole pattern is used, but some regions, *e.g.* areas where standards are present, can be excluded.

3.5. Pearson's r

Pearson's r is a parametric linear correlation coefficient widely used in crystallography. It has a similar form to Spearman's test, except that the data values themselves, and not their ranks, are used:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2\right]^{1/2}} \quad (4)$$

(where $\bar{x}$, $\bar{y}$ are the means of intensities taken over the full diffraction pattern). Again, r can range from -1.0 to 1.0.

Fig. 3 shows the use of the Pearson and Spearman correlation coefficients. In Fig. 3(*a*), $r = 0.93$ and $\rho = 0.68$. The high parametric coefficient arises from the perfect match of the two biggest peaks, but the much lower Spearman coefficient acts as a warning that there are unmatched regions in the two patterns. In Fig. 3(*b*), the situation is reversed: the Pearson $r = 0.79$, whereas $\rho = 0.90$, and it can be seen that there is a strong measure of association with the two patterns, although there are some discrepancies in the region 15–35°. In Fig. 3(*c*), $r = 0.66$ and $\rho = 0.22$; in this case the Spearman test is again warning of missing match regions. Thus, the use of the two coefficients acts as a valuable balance of their respective properties when processing complete patterns.

3.6. Kolmogorov–Smirnov two-sample test

The third test we use is the Kolmogorov–Smirnov (KS) two-sample test (also known as the Smirnov test) which we apply to individual peaks rather than the complete diffraction pattern, *i.e.* only peaks that occur at the same 2θ values (within a user-specified tolerance) in both patterns are compared, and this is done on a point-by-point basis. For further details of the KS test, see work by Smirnov (1939) with a fuller discussion by Steck & Smirnov (1969). The original Kolmogorov test was designed to compare an empirical distribution function to a hypothetical distribution function. The Smirnov variation compares two empirical distribution functions. As the correct function is generally not known, the Smirnov variation is more widely useful. Unlike tests such as the chi-squared, the KS test gives exact results for small data sets and does not require a large number of observations.

The two peak profiles each have n_p points, which are transformed to ranks then converted to cumulative distributions $S_1(x)$ and $S_2(x)$, respectively. The test then looks for the maximum value of the absolute difference between the two over the full range of n_p:

$$D = \sup_{x} |S_1(x) - S_2(x)|. \quad (5)$$

The process is shown graphically in Fig. 4. To establish the validity of the null hypothesis, H_0, that the peaks are drawn from the same sample, the associated probability can be calculated *via* the approximation

$$p(H_0|D) = Q_{KS}\left[\left(n_p^{1/2} + 0.12 + 0.11/n_p^{1/2}\right)D\right], \quad (6)$$

where

$$Q_{KS}(t) = 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2t^2), \quad (7)$$

with the limits $Q_{KS}(0) = 1$ and $Q_{KS}(\infty) = 0$. The larger the value of D, the less likely it represents the same data and the two peaks are different. Just as with the Spearman coefficient, the KS test is a robust non-parametric statistic.

An example of the KS test applied to real data is shown in Fig. 5. In Fig. 5(*a*) the peaks have similar, although not identical shapes with identical peak positions; $D = 0.22$, with an associated probability for the null hypothesis of $p(H_0|D) =$

0.98, *i.e.* there is a 98% chance that the null hypothesis is valid. In Fig. 5(*b*), where peak shapes are very different and there is a small offset of the peak maxima, the corresponding statistics are D = 0.51, with $p(H_0|D)$ = 0.25. In this case the null hypothesis is not accepted at the usual limits of 95 or 99%.

3.7. Peak matching using Pearson's r

In the same way as the KS test, peaks can also be matched using their full profile by employing the Pearson r on a point-by-point basis but confining the match to the region of peak overlap(s) in the two samples. In general, this test is the least useful of the four, and is highly correlated with the r coefficient computed over the whole diffraction pattern.

3.8. Combining the coefficients

It is usually advantageous to combine individual correlation coefficients to give an overall measure of similarity. The Pearson r and the Spearman ρ are usually used together in a weighted mean to give an overall rank coefficient r_w:

$$r_w = (w_1\rho_{xy} + w_2r_{xy})/(w_1 + w_2). \qquad (8)$$

Usually $w_1 = w_2 = 0.5$. This argument is, of course, heuristic: there is no particularly rigorous statistical validity in doing this, but in practice the combination has considerable discriminating power.

The KS test gives $p(H_0|D)$. In principle, this allows us to mix the KS test with r and ρ, but, in reality, we have here two classes of test: one is based on the entire pattern and the other uses only specified peaks, and it is not easy to combine the two classes, since the second is a function of the number of peaks and there remains the problem of processing problems where a peak is present in the reference sample but not in another, and *vice versa*. In consequence, we tend to keep the two classes separate.

4. Full-profile qualitative pattern matching in action

The method proceeds as follows.

(i) A database of known samples is created. Each sample is optionally pre-processed as described in §3.1. Note that peak identification is only necessary if the KS or the related parametric test are to be used: it is not required for the Spearman or full-pattern Pearson tests.

(ii) The sample pattern to be matched against this database is selected, and pre-processed as necessary.

(iii) The intersecting 2θ range of the two data sets is calculated, and each of the pattern-matching tests is performed using only that region. The user may also define excluded regions.

(iv) A minimum intensity is set, below which profile data are set to zero. This eliminates noise and does not reduce the discriminating power of the method. This is set to $0.1I_{max}$ as a default, where I_{max} is the maximum measured intensity, but the parameter may be varied.

(v) An optimal shift in 2θ between patterns is often required, arising from equipment settings and data collection protocols. We use the form

$$\Delta(2\theta) = a_0 + a_1 \sin\theta, \qquad (9)$$

where a_0 and a_1 are constants that can be determined by maximizing equation (5). 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) which does not require them. The efficiency of this procedure is discussed in §4.5.

(vi) A parametric Pearson's test is performed using all the measured data points.

(vii) The Spearman ρ is computed in the same way.

(viii) Peak lists for the sample and database patterns are compared. Where a peak is located within a user-controllable tolerance at the same 2θ in both patterns, a KS test is performed utilizing the full profiles of each coinciding peak. A final KS probability is calculated as the average of the individual KS peak test scores.

(ix) Procedure (viii) is repeated using the parametric Pearson test in exactly the same way as the KS test.

(x) Results from each of the four tests are stored and displayed by the program for each pattern in the database.

(xi) An overall rank value is calculated for each database sample after completion of the various calculations. It comprises the sum of weighted values of the available statistics. The weights applied are user-definable.

(xii) The matching results are then sorted in rank order, r_w, or *via* any of the individual tests described above as required.

4.1. Test data

To provide suitable examples of *SNAP-1D* full-profile pattern matching, a database of 98 patterns in CIF format was imported into the program. These comprise a subset of the ICDD database for the analysis of clay minerals (Smith *et al.*, 1996; Smith, 1999; ICDD, 2003). Clay minerals are layer silicates, in which layer stacking-sequence errors give rise to broad peaks which are often highly asymmetric, and are thus poorly represented by the standard *d–I* formalism, and so represent a suitable challenge for full-profile matching procedures. There is a good monograph on the use of powder diffraction and clay minerals by Moore & Reynolds (1997).

4.2. Pattern matching on montmorillonite using the ICDD database of clay minerals

There are three samples of montmorillonite in the database. One of these was selected as the reference pattern and matched against the remaining 97 patterns. The results are shown in Fig. 6 and tabulated in Table 1, sorted on the r_w value. The three montmorillonite samples are clearly identified with the top r_w values; the next pattern in the list is nonite and there is a clear and significant drop in r_w for this sample. There are substantial differences in the three montmorillonite patterns, especially in the region 18–35° 2θ, but the combined use of the Pearson and Spearman coefficients allows the patterns to be

Table 1
Pattern matching on a sample of montmorillonite using an ICDD database of clay minerals.

The results are sorted on r_w. This table needs to be read in conjunction with Fig. 6. There are three montmorillonite samples in the database and these have been successfully identified as the top three matches. The values of the Pearson, Spearman and the Pearson coefficient applied only to matching peaks are quite similar, but the KS test indicates significant detailed difference in the patterns.

Mineral	Rank	Pearson	Spearman	KS	Pearson peaks	Line colour
Montmorillonite	1.00	1.00	1.00	1.00	1.00	Red
Montmorillonite	0.87	0.87	0.88	0.47	0.92	Dark blue
Montmorillonite	0.79	0.71	0.89	0.18	0.71	Green
Nonite	0.54	0.48	0.60	0.19	0.56	Light blue

Table 2
Pattern matching on opal using an ICDD database of clay minerals.

This should be read in conjunction with Fig. 7. The first entry was input as the reference. There are two other opal samples in the database and these are identified as the top entries in the sorted r_w list, even though there are considerable differences between them, especially for Opal-A. This is highlighted by the low values of the KS test.

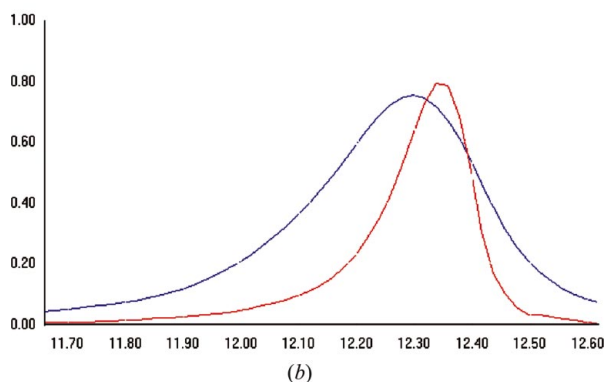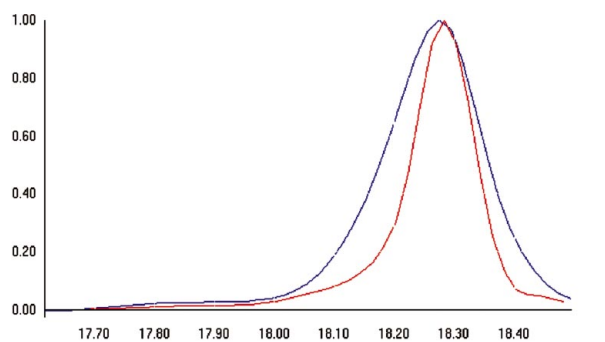| Sample	Rank	Pearson	Spearman	KS	Pearson peaks	Line colour
Opal-CT	1.0000	1.0000	1.0000	1.0000	1.0000	Red
Opal-CT	0.7942	0.9308	0.6577	0.6218	0.7905	Blue
Opal-A	0.6313	0.7286	0.5341	0.0102	0.5271	Green

Figure 5
The Kolmogorov–Smirnov two-sample test applied to single peaks from two patterns which occur at the same value of 2θ. (a) $D = 0.22$; the associated probability $p(H_0|D) = 0.98$, i.e. the null hypothesis that the two peaks are drawn from identical samples, is accepted with an associated probability of 0.98. (b) $D = 0.51$, with $p(H_0|D) = 0.25$. In this case the peaks are drawn from different samples, which can be seen via the offset in 2θ and the very different peak shapes.

successfully matched. The KS test highlights the fact that significant peak profile differences are present. As expected, the Pearson peak correlation coefficient is less sensitive, and less useful, and is closely correlated to the full Pearson r coefficient.

4.3. Opal

Opal is a quartz mineral. Opaline silicates form a diagenetic series which begins with amorphous opal (opal A) and progresses through opal-CT to opal C, ending with low-quartz (Moore & Reynolds, 1997). An opal-CT sample was matched against the database. The results are shown in Fig. 7 and tabulated in Table 2, sorted on the r_w value. There are only three opal samples in the database as used. They have all been identified despite considerable difference in peak shapes, widths and offsets, especially those involving opal-A. As before, the KS test highlights the differences in peak shape. Sample matching using d–I values would be very difficult with these data.

4.4. Using the Kolmogorov–Smirnov test

As an example of the use of the KS test to monitor small peak shape differences, the KS test was applied to quartz in
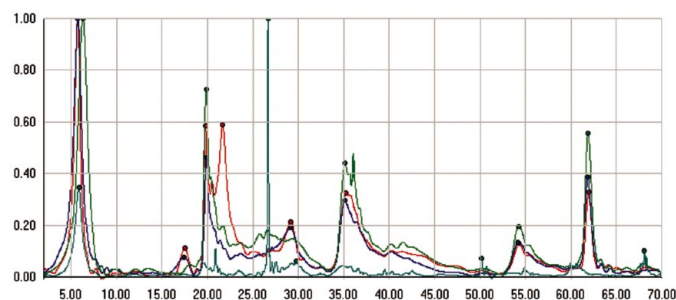

Figure 6
Pattern matching for montmorillonite using the ICDD clay minerals database. This needs to be read in conjunction with Table 1, which includes the key for line colours. There are three montmorillonite samples in the database and these have been successfully identified as the top three matches despite considerable profile differences. The next pattern in the sorted list of r_w values is the unrelated nonite mineral, which is quite different, having a sharp peak around 27°. This is reflected in the low value of 0.55 for r_w.
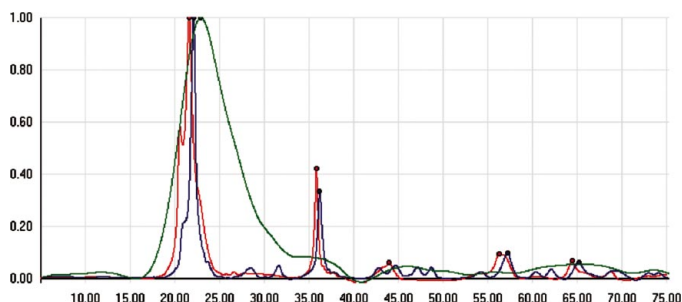

Figure 7
Pattern matching using an Opal-CT sample in the ICDD clays database as a reference. This figure needs to be examined in conjunction with Table 2. The top two matches (excluding the reference opal) are opal CT and opal-A; the latter has a very different peak profile compared with the remaining samples. There are also problems with peak offsets.

Table 3
The use of the downhill simplex method to determine the shift parameters a_0 and a_1 from equation (9).

Columns 1 and 4 contain the values of a_0 and a_1 that were used to generate the offset data. The calculated values from the downhill simplex method optimizing r_w are given in columns 2 and 5, with the absolute differences in columns 3 and 6. The mean deviation of a_0 from the true values is 0.02, while that of a_1 is 0.005. This compares well with the resolution of the data, which is $0.02°$ 2θ.

a_0	a_0^{calc} from simplex method	$\|\Delta_0\| = \|a_0 - a_0^{\text{calc}}\|$	a_1	a_1^{calc} from simplex method	$\|\Delta_1\| = \|a_1 - a_1^{\text{calc}}\|$
0.00	0.00	0.00	0.00	0.00	0.00
0.20	0.18	0.02	0.40	0.41	0.01
1.00	0.99	0.01	0.00	0.00	0.00
0.00	−0.04	0.04	0.40	0.41	0.01
0.40	0.37	0.03	0.40	0.41	0.01
0.40	0.37	0.03	0.20	0.22	0.02
1.00	1.00	0.00	1.00	1.00	0.00
−1.00	−0.97	0.03	1.00	1.00	0.00
1.00	1.00	0.00	−1.00	−1.00	0.00

the 2θ range 79.0–84.5°. The Pearson and Spearman correlation coefficients are 0.88 and 0.87, respectively; the Pearson coefficient applied to the peaks only is 0.82, but the KS test gives a coefficient of 0.19, highlighting the fact that there are significant differences. Fig. 8 shows the two patterns superimposed; it can be seen that there are differences in peak widths and data resolution, although overall the peaks are very similar, especially as characterized by d–I values.

4.5. Pattern shifts

To test the efficacy of the downhill simplex method for determining the parameters a_0 and a_1 in equation (9), a series of eight shifted patterns were generated for an organic powder sample in the range $0 \leq 2\theta \leq 35°$ using values of a_0 and a_1 in the range −1.0 to 1.0. The simplex method was then used to compare the calculated values of the shift parameters with those used to generate the offset patterns. The method uses multiple starting points: if the maximum search values for a_0 and a_1 are defined as $(a_{0\text{max}}, a_{1\text{max}})$, we use the starting points $(0.0, 0.0)$, $(a_{0\text{max}} + 0.1, 0.0)$, $(0, a_{1\text{max}} + 0.1, 0.0)$. Once an optimum point has been found, it is usually recommended that the calculation is restarted from the optimum point, but we found this to be unnecessary.

Table 3 summarizes the results; the average deviation between true and calculated values of the a_0 coefficients is
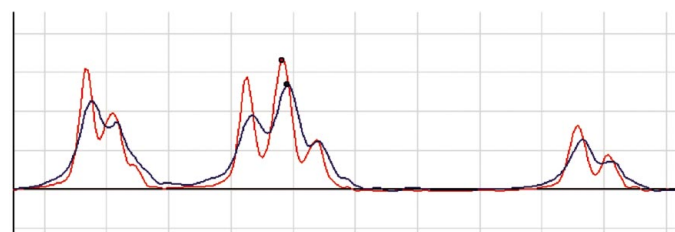
Figure 8
The KS test applied to quartz in the 2θ range 79.0–84.5°. The Pearson and Spearman correlation coefficients are 0.88 and 0.87, respectively; the Pearson coefficient applied to the peaks only is 0.82, but the KS test gives a coefficient of 0.19, highlighting the difference in detail between the two.

$0.02°$, and for a_1 is 0.005. This is within the resolution of the data, which is $0.02°$.

5. Quantitative analysis without Rietveld refinement

Quantitative analysis seeks to identify the components of a mixture given the powder diffraction patterns of the pure components and that of the mixture itself. It is obvious that the full profile data will, in general, be invaluable in these cases, and should give more accurate answers than d–I-based calculations, but will be less tractable mathematically. In this section we first review existing techniques and then demonstrate the use of least-squares combined with singular value decomposition to use full-profile diffraction data to obtain quantitative analyses of mixtures, without the use of Rietveld refinement and thus without knowledge of the crystal structures of the components.

5.1. Overview of existing quantitative techniques

There is an excellent text by Zevin & Kimmel (1995) covering all aspects of quantitative X-ray diffractometry. Quantitative analyses of powder diffraction patterns may be roughly divided into two categories: those involving the use of either an internal or an external standard, or those utilizing a full diffraction profile. The latter approach may be subdivided into the Rietveld method, profile stripping and least-squares best-fit summation.

The Rietveld approach requires crystal structures to be known for all individual phases in the mixture. A calculated full profile is produced based upon that knowledge, and crystallographic parameters refined to produce the best fit to the experimental data. See, for example, works by Bish & Howard (1988) and Hill (1993).

In the profile-stripping method (also known as pattern subtraction), figures of merit are used to identify a phase that best fits the overall mixture pattern. This pure-phase profile is then subtracted from the mixture profile, after scaling has been performed. The process is then repeated until no residual pattern remains, showing that all phases have been accounted for. Our approach is related to this but works in the opposite direction, taking all candidate patterns simultaneously, then reducing the possible candidates.

The best-fit summation approach, described by Smith *et al.* (1988), is suited to situations where the user has prior knowledge of likely candidate phases, and can therefore select them individually for inclusion. Using least-squares techniques, the best-fit of the weighted sum of combined phase patterns to the mixture pattern is obtained. Weight fractions are then calculated using the reference intensity ratio method (RIR) (Hill & Howard, 1987). A modification of this by Chipera & Bish (2002) obtains weight fractions using the pre-scaled patterns and the internal standard approach, and is implemented as an *Excel* worksheet.

5.2. Quantitative analysis using full profiles and singular value decomposition

Assume we have a sample pattern, S, which is considered to be a mixture of up to N components. S comprises m data points, S_1, S_2, ... S_m. The N patterns can be considered to make up fractions p_1, p_2, p_3, ... p_N of the sample pattern. We want the best possible combination of database patterns to make up the sample pattern. A system of linear equations can be constructed in which x_{11} is measurement point 1 of pattern 1, *etc.*:

$$x_{11}p_1 + x_{12}p_2 + x_{13}p_3 + \ldots + x_{1N}p_N = S_1,$$
$$x_{21}p_1 + x_{22}p_2 + x_{23}p_3 + \ldots + x_{2N}p_N = S_2,$$
$$\vdots$$
$$x_{m1}p_1 + x_{m2}p_2 + x_{m3}p_3 + \ldots + x_{mN}p_N = S_m. \quad (10)$$

Writing these in matrix form:

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mN} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{pmatrix} = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{pmatrix} \quad (11)$$

or

$$\mathbf{x} \cdot \mathbf{p} = \mathbf{S}. \quad (12)$$

We seek a solution for $\mathbf{S}$ that minimizes

$$\chi^2 = \left| \mathbf{x} \cdot \mathbf{p} - \mathbf{S} \right|^2. \quad (13)$$

Since $N \ll m$, the system is heavily overdetermined, and we can use least squares.

The condition number of a matrix is the ratio of the largest to the smallest values of its corresponding diagonal matrix $\mathbf{W}$. It is called singular if its condition number is or approaches infinity, and ill-conditioned if the value of the reciprocal of the condition number begins to approach the precision limit of the machine being used to calculate it (see, for example, Searle, 1999). Normal least-squares procedures can have difficulties attempting to invert very poorly conditioned matrices, such as will arise with powder data where every data point is included. Singular value decomposition (SVD) is ideal in such cases as it allows singular and ill-conditioned matrices to be dealt with. In particular, not every $m \times N$ matrix has an inverse. However, every such matrix does have a corresponding singular value decomposition.

SVD decomposes the $\mathbf{x}$ matrix to several constituent matrices to give the solution (Press *et al.*, 1992)

$$\mathbf{p} = \mathbf{V} \cdot \text{diag}(1/w_j) \cdot \mathbf{U}^T \cdot \mathbf{S}. \quad (14)$$

$\mathbf{W}$ is a diagonal matrix with positive or zero elements. If most of its components are unusually small, then it is possible to approximate the matrix $\mathbf{p}$ with only a few terms of $\mathbf{S}$ (*i.e.* we can make up the sample pattern using only a combination of just a few database patterns) so that combinations of equations that do not contribute to the best possible final solution are effectively ignored. This system of least squares is highly stable, and the use of $\mathbf{W}$ gives us a flexible and powerful way of producing a solution for the composition of an unknown number of pure phases contributing to a measured pattern.

Although computationally the method is, relatively speaking, quite a slow and memory-hungry one, as it involves calculations dealing with several large matrices, it is exceptionally stable, and, when dealt with properly, rarely causes computational problems. The method has found use in powder indexing (Coelho, 2003).

The variance–covariance matrix can also be obtained from the $\mathbf{V}$ matrix and the diagonal of $\mathbf{W}$:

$$\text{cov}(p_j, p_k) = \sum_{i=1}^{N} \left(\frac{V_{ji}V_{jk}}{w_i^2} \right). \quad (15)$$

From this an estimate of the variances of the component percentages can be found.

Powder diffraction yields the fractional percentages arising from the scattering power of the component mixtures, $p_i - p_N$. The values of p can be used to calculate a weight fraction for that particular phase provided that the atomic absorption coefficients are known, and this in turn requires the unit-cell dimensions and cell contents, but not the atomic coordinates (Smith *et al.*, 1993; Cressey & Schofield, 1996). The general formula for the weight fraction of component n in a mixture comprising N components is (Leroux *et al.*, 1953)

$$c_n = p_n \mu^* / \mu_n^*, \quad (16)$$

where

$$\mu^* = \sum_{j=1}^{N} c_j \mu_j^* \quad (17)$$

and

$$\mu_j^* = \mu_j / \rho_j, \quad (18)$$

where μ_j is the atomic X-ray absorption coefficient and ρ_j is the density of component j. The variance of c_n can be computed *via*

$$\sigma^2(c_n) = \left[\frac{1}{(1-p_n)\mu_n^*} \right]^2 \left[\frac{1}{(1-p_n)^2} \left(\sum_{\substack{j=1 \\ j \neq n}}^{N} \mu_j^* c_j \right)^2 \sigma^2(p_n) \right.$$
$$\left. + p_n^2 \sum_{\substack{j=1 \\ j \neq n}}^{N} (\mu_j^*)^2 \sigma^2(c_j) \right]. \quad (19)$$

(see Appendix *A* for details). Clearly the variance of any component depends on the variances of the other phases which are themselves unknown at the start of the calculation. Equation (19) is solved by assigning equal variances of 1.0 to each $\sigma^2(c_j)$ and iterating until there is no significant change in variance.

5.3. Applications of the SVD method

This method requires a database of full-profile patterns, and assumes that the patterns of the individual pure phases are

Table 4
Quantitative analysis test using a subset of an ICDD clay minerals database.

Only the scale percentages are calculated. (*a*) A synthetic mixture of equal proportions of fluorite, anatase and gibbsite tested against the whole database, which has two gibbsite entries. The scale fraction for gibbsite sums to 0.330. (*b*) A mixture of equal proportions of fluorite, anatase and gibbsite tested against the whole database with one of the two gibbsite entries removed. (*c*) A synthetic mixture of unequal proportions of fluorite, anatase and gibbsite tested against the whole database with one of the two gibbsite entries removed.

Name	Actual scale fraction	Calculated scale fraction
(*a*)		
Fluorite	0.333	0.337 (8)
Anatase	0.333	0.332 (7)
Gibbsite	0.333	0.293 (20)
Gibbsite		0.037 (13)
(*b*)		
Fluorite	0.333	0.329 (8)
Anatase	0.333	0.335 (9)
Gibbsite	0.333	0.336 (13)
(*c*)		
Fluorite	0.750	0.750 (45)
Anatase	0.150	0.149 (80)
Gibbsite	0.100	0.101 (51)

included within that database. Obviously, the quality of the overall results is strongly dependent on the quality of the measured data and care is needed to use suitable protocols. As in qualitative analysis, data interpolation followed by optional background subtraction and wavelet smoothing procedures are performed upon all the patterns.

Depending upon user preferences, either the entire database, or just a subset of it can be used as possible phase input. The subset is selected using a user-controlled correlation cut-off level. In this case only those patterns that have a weighted mean correlation, r_w, greater than a given cut-off value are subsequently used in the SVD-based least squares. The full angle range of the unknown sample is used by default in the calculations, but a smaller sub-range may be employed if required. The method selects the top 15 results as measured by the **p** matrix from this solution vector as long as the associated weights from the **W** matrix are significantly greater than zero,

and builds another matrix **p** with them, carrying out the entire procedure again.

Finally, the top *j* patterns (where *j* is a user-controllable integer between 1 and 15) are put through the matrix decomposition process once more. The results returned are the fractions of each pattern included in the test pattern. These are scaled to a percentage, and the number of possible phases is limited to *j*. The displayed results are effectively the scale fraction for each phase; weight percentages may be calculated from these if required. Any patterns that are considered to be incorrect can be marked as such by the user, and may then be ignored and the analysis repeated.

6. Examples of quantitative analysis

6.1. Simulated mixtures

To provide a test for the method, the powder diffraction patterns of mixtures were simulated by combining various experimental patterns from the ICDD clay database, and then adding 5% Gaussian noise to the resulting pattern.

The first example of this involved three individual minerals: gibbsite, anastase and fluorite. A powder pattern was generated by combining the individual patterns in equal proportions. A qualitative search was first carried out of the entire database, and all patterns with an r_w value of <0.01 were excluded from the quantitative analysis which then followed. The results are shown in Table 4(*a*). Note that these are the only suggested constituent phases returned by the program; all other phases in the database were discarded by the analysis process. That two different gibbsite phases are suggested is a consequence of the database, which contains multiple patterns for some minerals, and two for gibbsite. Marking one of the two gibbsites to be ignored and re-running the analysis gave the results shown in Table 4(*b*).

A mixture containing the same phases, but in different proportions, was then constructed. The same data handling and options were used as previously. The results from this run are in Table 4(*c*).

Table 5
Quantitative analysis of mixtures of corundum, fluorite, zincite and brucite (or alternatively silica) from the IUCr round robin (Scarlett *et al.*, 2002; Masden *et al.*, 2001).

Sample 3 contains silica rather than brucite (indicated by asterisks *). For sample 3, the weight percentage corrundum was provided by the organizers and the results have been scaled to this value. Standard errors are in brackets. The average deviation for corundum is 20%, that for fluorite is 18%, and that of zincite is 3%, with an overall average deviation of 23%. The results for 1a–1h are also shown graphically in Fig. 8.

Sample	Published value				SNAP-1D result			
	Corundum	Flourite	Zincite	Brucite or silica	Corundum	Fluorite	Zincite	Brucite or silica
1a	1.2	94.8	4.0		1.3 (6)	93.8 (2)	4.8 (2)	
1b	94.3	4.3	1.4		94.9 (6)	3.7 (2)	1.4 (2)	
1c	5.0	1.4	93.6		3.8 (10)	1.1 (4)	95.1 (4)	
1d	13.5	53.6	32.9		12.2 (9)	49.6 (3)	38.2 (3)	
1e	55.1	29.6	15.3		53.4 (10)	28.6 (3)	17.9 (3)	
1f	27.1	17.7	55.2		22.8 (11)	15.9 (4)	61.3 (4)	
1g	31.4	34.4	34.2		28.3 (10)	32.0 (3)	39.7 (3)	
1h	35.1	34.7	30.2		30.1 (15)	33.4 (5)	36.5 (5)	
2	21.3	22.5	19.9	36.3	18.3 (20)	17.4 (3)	19.6 (1)	44.7 (2)
3	30.8	19.7	20.1	29.5*	30.8 (3)	21.9 (4)	18.1 (4)	29.0 (1)*

These calculations give values of the scale factors p_1, p_2, *etc.* in equation (23), rather than weight percentages. The average deviation of the calculated value from the true value is 0.2 and is always less than 2% in error, indicating that the method is capable, with good sample preparation techniques and with well characterized samples, of a viable level of accuracy.

6.2. IUCr round robin

The International Union of Crystallography Commission on Powder Diffraction (CPD) sponsored a round robin on the determination of quantitative phase abundance from diffraction data. The results were published in two papers (Madsen *et al.*, 2001; Scarlett *et al.*, 2002). We have used the data supplied for samples 1a–1h, 2 and 3 to test the viability of the full-pattern SVD methodology. Sample 1 is a three-phase system prepared with eight different and widely varying compositions. It was possible for participants to collect their own data or use that supplied by the CPD; we chose the latter approach. The papers identified a large variation in reported results arising from incorrect data processing and program usage. The results from the full-pattern SVD method are tabulated in Table 5 and are shown graphically in Fig. 9 (which is partially taken from Fig. 2 of Madsen *et al.*, 2001). The average deviation between true and calculated weight percentages is 2.0% for corundum; that for fluorite is 1.8%, and that of zincite is 3%, with an overall average deviation for all components of 2.3%. Given the simplicity and speed of our calculations, this is quite satisfactory. It should be noted that we are not proposing that this formalism is a substitute for Rietveld methods when high
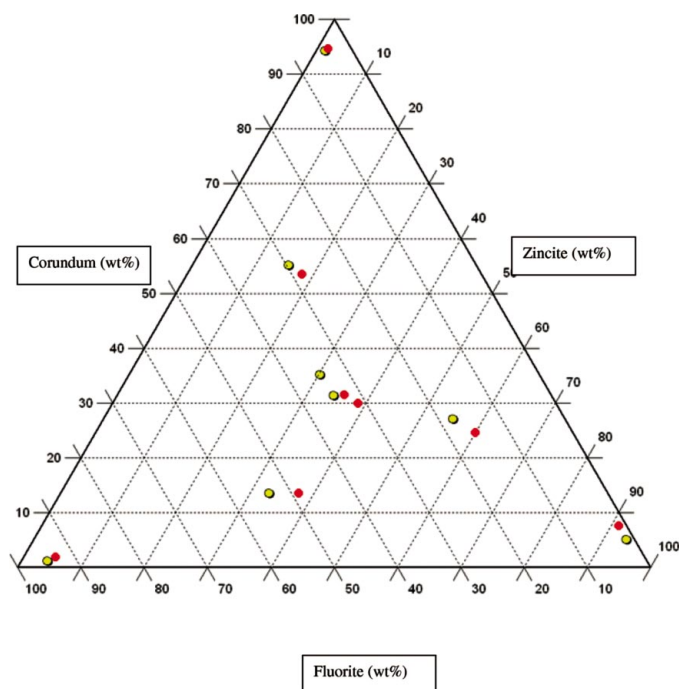

Figure 9
The results of the full-pattern quantitative analysis on mixtures of corundum, zincite and fluorite used in the IUCr round robin (Madsen *et al.*, 2001). The values are in weight percentage. The yellow circles are the correct results; the red are those calculated by *SNAP-1D*. This figure has been adapted from the paper by Madsen *et al.* (2001).

accuracy is required. However, it should also be emphasized that the total time for all these calculations is less than 1 min on a modest PC once the data are in a suitable format.

The errors seem to be underestimated, however. The source of this is probably due to systematic errors associated with peak shapes and background that do not find their way into our current model.

6.3. BCA round robin

The BCA 2003 Industrial Group Quantitative round robin (Cockcroft & Frampton, 2003) used a two-phase sample comprising paracetamol and lactose. Samples of mixture and pure phases were provided. Data collection was carried out on a Bruker D5000 diffractometer in reflection mode and analysed using the quantitative mode of *SNAP-1D*. There were noticeable preferred orientation effects in the lactose sample. The correct results were paracetamol 84.92% and lactose 15.08%. The values obtained by *SNAP-1D* were 86.2 and 14.8%, respectively. This represents an average deviation of 0.8%, which is very satisfactory.

7. Conclusions

We have shown that a mixture of parametric and non-parametric statistical tests using full-profile powder diffraction patterns is useful in both qualitative and quantitative powder diffractometry. The method is relatively simple and overcomes the problems that arise when only the peaks or representations of the peaks in the pattern are used. In quantitative mode, the use of singular value decomposition gives a stable mathematical formalism capable of being used with full diffraction data where every measured point is included. This can act as a simple alternative to Rietveld refinement and does not require atomic coordinates, although it does need X-ray absorption coefficients, and thus the unit-cell dimensions and contents, unless one is dealing with polymorphic mixtures. It is not as accurate as the Rietveld method, but can give percentage weight compositions with an estimated uncertainty of 1–5% depending on data quality. The limit of detectability for a given component is well below 5%.

The methodology is incorporated into the commercial computer program *SNAP-1D* (Barr *et al.*, 2003) that runs on PCs using the Windows 2000/XP operating systems and is marketed by Bruker-AXS.

The title of the paper concerns high-throughput crystallography, and the link with this technique now needs to be made: it is possible to generate a correlation matrix in which every pattern in a database of n patterns is matched with every other to give an $n \times n$ correlation matrix $\boldsymbol{\rho}$ using a weighted mean of the Spearman and Pearson coefficients with the optional inclusion of the KS and PP coefficients. The matrix $\boldsymbol{\rho}$ can be converted to a distance matrix, **d**, of the same dimensions *via*

$$\mathbf{d} = 0.5(1.0 - \boldsymbol{\rho}). \tag{20}$$

At this point, the tools of cluster analysis and multivariate data analysis are now available to classify patterns, identify clusters, estimate the number of pure components and to visualize them. This topic is addressed in the following paper. It can be used with more than 1000 patterns, and so is relevant to high-throughput techniques.

The methods described here can also be applied to other one-dimensional spectroscopies, such as Raman and IR, and results will also be presented elsewhere.

APPENDIX A
Error propagation in quantitative analysis

The general formula for the weight fraction of component n in a mixture comprising N components is:

$$c_n = p_n \mu^* / \mu_n^*, \qquad (21)$$

where

$$\mu^* = \sum_{j=1}^{N} c_j \mu_j^* \qquad (22)$$

and

$$\mu_j^* = \mu_j / \rho_j, \qquad (23)$$

where μ_j is the absorption coefficient and ρ_j the density of component j. Rearranging (21),

$$c_n = \frac{p_n}{1 - p_n} \left(\sum_{\substack{j=1 \\ j \neq n}}^{N} \mu_j^* c_j \right) \frac{1}{\mu_n^*} \qquad (24)$$

for $0.0 \leq s_n \leq 1.0$. We require the standard deviation $\sigma(c_n)$:

$$\left(\frac{\partial c_n}{\partial s_n} \right)_{c_n} = \frac{1}{(1 - p_n)^2} \left(\sum_{\substack{j=1 \\ j \neq n}}^{N} \mu_j^* c_j \right) \frac{1}{\mu_n^*}, \qquad (25)$$

$$\left(\frac{\partial c_n}{\partial c_j} \right)_{j \neq n} = \frac{p_n}{1 - p_n} \frac{\mu_j^*}{\mu_n^*}. \qquad (26)$$

Error propagation theory gives

$$\sigma^2(c_n) = \left(\frac{\partial c_n}{\partial p_n} \right)^2 \sigma^2(p_n) + \sum_{\substack{j=1 \\ j \neq n}}^{n} \left(\frac{\partial c_n}{\partial c_j} \right)^2 \sigma^2(c_j), \qquad (27)$$

so

$$\sigma^2(c_n) = \left[\frac{1}{(1 - p_n)\mu_n^*} \right]^2 \left[\frac{1}{(1 - p_n)^2} \left(\sum_{\substack{j=1 \\ j \neq n}}^{N} \mu_j^* c_j \right)^2 \sigma^2(p_n) \right.$$

$$\left. + p_n^2 \sum_{\substack{j=1 \\ j \neq n}}^{N} (\mu_j^*)^2 \sigma^2(c_j) \right]. \qquad (28)$$

References

Altomare, A., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Rizzi, R. & Werner, P. (2000). *J. Appl. Cryst.* **33**, 1180–1186.

Barr, G., Dong, W. & Gilmore, C. J. (2004). *J. Appl. Cryst.* **37**, 243–253.

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. (See also http://www.chem.gla.ac.uk/staff/chris/snap.html.)

Bish, D. L. & Howard, S. A. (1988). *J. Appl. Cryst.* **21**, 86–91.

Chipera, S. J. & Bish, D. L. (2002). *J. Appl. Cryst.* **35**, 744–749.

Coelho, A. A. (2003). *J. Appl. Cryst.* **36**, 86–95.

CCP14 (2003). http://www.ccp14.ac.uk/.

Cockcroft, J. & Frampton, C. (2003) *British Crystallographic Association Spring Meeting*, York, UK. Session P/L002. (No abstract.)

Conover, W. J. (1971). *Practical Nonparametric Statistics*. New York: John Wiley.

Cressey, G. & Schofield, P. F. (1996). *Powder Diffr.* **11**, 35–39.

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

Gilmore, C. J. (1998). Presented at the ICDD Spring Meeting, Newton Square, Pennsylvania, USA.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Hanawalt, J. D., Rinn, H. W. & Frevel, L. K. (1938). *Ind. Eng. Chem. Anal.* **10**, 457–512.

Hill, R. J. (1993). *The Rietveld Method*, edited by R. A. Young, pp. 95–101. Oxford University Press.

Hill, R. J. & Howard, C. J. (1987). *J. Appl. Cryst.* **20**, 467–474.

ICDD (2003). *The Powder Diffraction File.* International Center for Diffraction Data, 12 Campus Boulevard, Newton Square, Pennsylvania 19073–3273, USA.

Jenkins, R. (1998). 1988 Denver X-ray Conference, Workshop W9. http://www.dxcicdd.com/98/wkshopt.htm.

Johnson, G. G. & Vand, V. (1967). *Ind. Eng. Chem.* **59**, 19–31.

Leroux, J., Lennox, D. H. & Kay, K. (1953). *Anal. Chem.* **25**, 740–743.

Madsen, I. C., Scarlett, N. V. Y., Cranswick, L. M. D. & Lwin, T. (2001). *J. Appl. Cryst.* **34**, 409–426.

Moore, D. M. & Reynolds, R. C. Jr (1997). *X-ray Diffraction and the Identification and Analysis of Clay Minerals.* Oxford University Press.

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

Nusinovici, J. & Winter, M. J. (1994). *Adv. X-ray Anal.* **37**, 59–66.

Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*, pp. 144–148. Boston: Birkhäuser.

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

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

Scarlett, N. V. Y., Madsen, I. C., Cranswick, L. M. D., Lwin, T., Groleau, E., Stephenson, G., Aylmore, M. & Agron-Olshina, N. (2002). *J. Appl. Cryst.* **35**, 383–400.

Searle, S. R. (1999). *Matrix Algebra Useful for Statistics*, pp. 316–317. New York: John Wiley.

Smirnov, N. V. (1939). *Bull. Moscow Univ.* **2**, 3–16.

Smith, D. K. (1999). *Defect and Microstrucure Analysis by Diffraction*, edited by R. L. Snyder, J. Fiala & H. J. Bunge, pp. 597–610. Oxford University Press.

Smith, D. K., Hoyle, S. Q. & Johnson, G. G. (1993). *Adv. X-ray Anal.* **36**, 287–299.

Smith, D. K., Johnson, G. G. & Hoyle, S. Q. (1991). *Adv. X-ray Anal.* **34**, 377–385.

Smith, D. K., Johnson, G. G. & Jenkins, R. (1996). *Adv. X-ray Anal.* **38**, 117–125.

Smith, D. K., Johnson, G. G. & Wims, A. M. (1988). *Aust. J. Phys.* **41**, 311–321.

Smrčok, Ĺ., Ďurík, M. & Jorík, V. (1999). *Powder Diffr.* **14**, 300–304.

Spearman, C. (1904). *Am. J. of Psychol.* **15**, 72–101.

Steck, G. P. & Smirnov, G. N. (1969). *Ann. Math. Stat.* **40**, 1449–1466.

Steele, J. K. & Biederman, R. R. (1994). *Adv. X-ray Anal.* **37**, 101–107.

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