Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 21 March 2012
Accepted 11 June 2012

Application of clustering techniques to electron-diffraction data: determination of unit-cell parameters

Sebastian Schlitt,[a] Tatiana E. Gorelik,[b] Andrew A. Stewart,[b] Elmar Schömer,[c] Thorsten Raasch[a] and Ute Kolb[b]*

[a]Institute for Mathematics, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128, Mainz, Germany, [b]Institute for Physical Chemistry, Johannes Gutenberg University Mainz, Jakob Welder Weg 11, 55128, Mainz, Germany, and [c]Institute of Computer Science, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128, Mainz, Germany. Correspondence e-mail: kolb@uni-mainz.de

A new approach to determining the unit-cell vectors from single-crystal diffraction data based on clustering analysis is proposed. The method uses the density-based clustering algorithm DBSCAN. Unit-cell determination through the clustering procedure is particularly useful for limited tilt sequences and noisy data, and therefore is optimal for single-crystal electron-diffraction automated diffraction tomography (ADT) data. The unit-cell determination of various materials from ADT data as well as single-crystal X-ray data is demonstrated.

1. Introduction

The description of a periodic object is based on the determination of its internal regularity – the periodicity law. In crystallography the determination of the unit-cell vectors is based on the analysis of the periodicity of three-dimensional reflection positions and is essential for the analysis of a crystalline structure. Automated determination of unit-cell vectors for single-crystal X-ray data relies on the projection of the three-dimensional reflection positions onto a certain direction, and the periodicity along this direction is assessed by one-dimensional Fourier transformation (Steller *et al.*, 1997; Sauter *et al.*, 2004; Rossmann, 2001). This works by scanning the projection vector over the complete solid angle of the data set. Significant Fourier terms are observed in the one-dimensional fast Fourier transform (FFT) when directions orthogonal to widely separated planes of reflections are encountered. Given these directions and the periodicity parameters, the orientation matrix can be calculated.

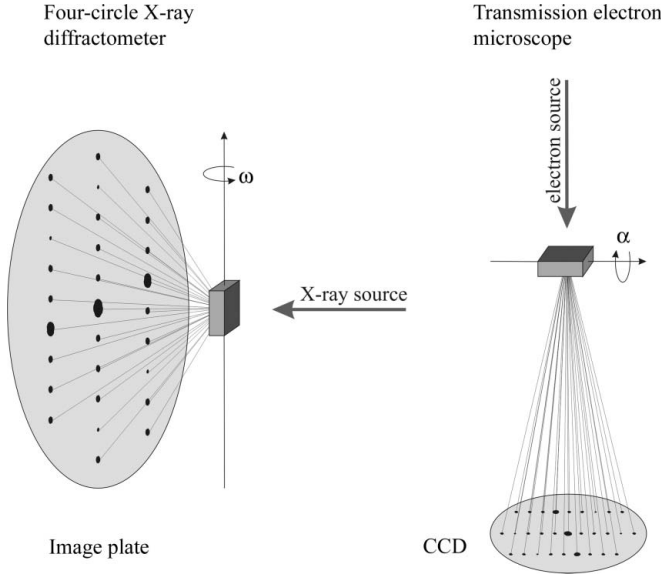


Figure 1
Schematic view of (*a*) the experimental geometry of an ω scan in a four-circle X-ray diffractometer and (*b*) the ray path in a transmission electron microscope in diffraction mode.

For a long time the use of the single-crystal *electron* diffraction method was restricted to analysis of low-index zonal patterns (Dorset, 1995) with unit-cell basis vectors determined either manually using a Vainshtein plot (Vainshtein, 1964) or dedicated automated routines (Zou *et al.*, 2004; Jiang, Georgieva, Nederlof *et al.*, 2011; Jiang, Georgieva & Abrahams, 2011). The situation changed drastically with the introduction of the automated diffraction tomography (ADT) technique, employing electron-diffraction data collection through a fine-step tilt around an arbitrary crystallographic axis (Kolb *et al.*, 2007, 2008). Conceptually ADT is comparable to ω-scan X-ray data collection using an area detector (Fig. 1). The data are collected in transmission geometry; the crystal is tilted around the primary goniometer axis α. Diffraction patterns are usually recorded on a charge-coupled device (CCD) camera, although image plates or film can also be used. Frames are collected using a tilt step of about 1° within a certain tilt range. The range over which the data are collected depends on the purpose of the data acquisition (determination of unit-cell parameters can be performed with relatively short

Table 1
Samples used for testing the algorithm.

Sample	Chemical composition	Space group, unit-cell parameters	Reference
A	Barite ($BaSO_4$)	*Pnma*: a = 8.8842, b = 5.4559, c = 7.1569 Å	Jacobsen *et al.* (1998)
B	Pseudo-spinel ($LiTi_{1.5}Ni_{0.5}O_4$)	$P\bar{3}c1$: $a = b$ = 5.059, c = 32.537 Å	Kolb *et al.* (2011)
C	Paracetamol (orthorhombic)	*Pcab*: a = 11.805, b = 17.164, c = 7.393 Å	Haisa *et al.* (1974)
D	Propellan (X-ray data)	$P2_1/c$: a = 8.0258, b = 36.389, c = 7.8184 Å, β = 99.493°	Mirion *et al.* (2012)

tilt sequences, *ab initio* structure solution requires data with high completeness) and on the experimental setup (some transmission electron microscopes can have a very short pole-piece gap restricting the tilt range of the sample holder).

Crystal structures of diverse materials have been solved *ab initio* using direct methods (Kolb *et al.*, 2011) from ADT data, just as is done in single-crystal X-ray analysis. Single-crystal ADT data suitable for *ab initio* structure analysis have been collected from crystals with dimensions as small as 50 nm (Mugnaioli *et al.*, 2012; Birkel *et al.*, 2010). In principle, ADT mimics single-crystal X-ray analysis at the nanoscale; however, some significant differences occur due to the differing radiation types and their associated interactions with the sample. Electrons interact with matter much more powerfully than X-rays and, therefore, a much smaller sample volume can deliver substantial diffraction intensity data. The electron wavelength is smaller, which in turn makes the Ewald sphere very large, to the extent that it almost represents planar cuts through reciprocal space. Other important features of ADT data are:

(*a*) The size of the electron beam used for diffraction is freely adjustable by the microscope lenses. Different situations can be realised: the complete crystal can be illuminated ('bathed' in the beam) or only a part of the crystal can be illuminated. The latter setup is especially useful for beam-sensitive materials, as consecutive frames can be collected from a fresh unexposed crystal area. In this case, *crystal bending* may add distortion to the data geometry.

(*b*) The *excitation error*, which manifests itself as significant intensity appearing on diffraction patterns far away from the reflection centre; this can be relatively large for electron-diffraction data.

(*c*) The major consideration in a transmission electron microscopy (TEM) study is the thickness of the sample in the direction of transmission. The elongation of the reflection due to the *size effect* causes additional ambiguity in the reflection's true position.

(*d*) In many nanocrystalline materials the crystals are tightly agglomerated, making it difficult to collect a tilt series from a clearly separated single crystal. The number of reflections appearing from *additional crystals* is usually low, but nonetheless they present a significant source of noise in the data set and can cause problems during the data-reduction process.

(*e*) At present, most data sets are collected using CCD cameras. This recording medium often suffers from the *hot-spot* problem: fast X-rays stochastically produce high-contrast spots on the camera, which may be falsely interpreted as reflections.

(*f*) In X-ray diffraction, the *background* has a well defined and characterized pattern. This does not hold for electron diffraction. The background profile can be considerably affected by dynamical effects, including formation of Kikuchi lines. On the other hand, the strong interaction between the incident electrons and the CCD phosphor can also contribute to the diffraction pattern in a very nonlinear fashion.

The combination of all the aforementioned problems gives rise to the fact that the Bragg reflections' positions and therefore the reciprocal-lattice nodes are not as clearly defined for electron ADT data as they are for single-crystal X-ray data. This additional noise in the data often causes unit-cell search methods based on Fourier analysis to fail. Fig. 2 shows a one-dimensional Fourier transformation of main projections of reflection positions. The top row (Fig. 2*a*,*b*) are Fourier transforms of reflection-position projections of sample D (see Table 1) onto the main crystallographic directions **c** and **b** calculated from single-crystal X-ray data. The periodicity along both directions is clearly resolved. The bottom row (Fig. 2*c*,*d*) shows Fourier transforms of reflection-position projections onto **a** and **c** of sample B (Table 1) from its ADT data. The periodicity of the lattice along the **a** direction (corresponding to the unit-cell parameter of 5.059 Å) is well resolved, while the periodicity along the **c** direction (32.537 Å, comparable to the lattice parameter shown in Fig. 2*b* of 36.389 Å) is not evident at all.

Here we propose an alternative method for determining the orientation matrix from electron-diffraction single-crystal data collected using the ADT technique. The method is based on the cluster analysis of difference vectors, calculated from three-dimensional Bragg peak positions. We review a variety of clustering algorithms, highlighting their strengths and weaknesses with regards to determining unit-cell parameters from ADT data, and give some examples of how the selected algorithm performs.

2. Difference-vector space built from ADT data: a closer look

Owing to the small wavelength, electron-diffraction patterns are almost planar cuts through reciprocal space. This means that if a short tilt sequence is collected within a certain tilt range, the three-dimensional information within the reciprocal volume will only be present within this wedge, while for single-crystal X-ray diffraction, due to the smaller radius of the Ewald sphere, the information within the reciprocal volume is distributed differently. Basic unit-cell vectors do not necessarily lie within the measured fraction of the reciprocal space,

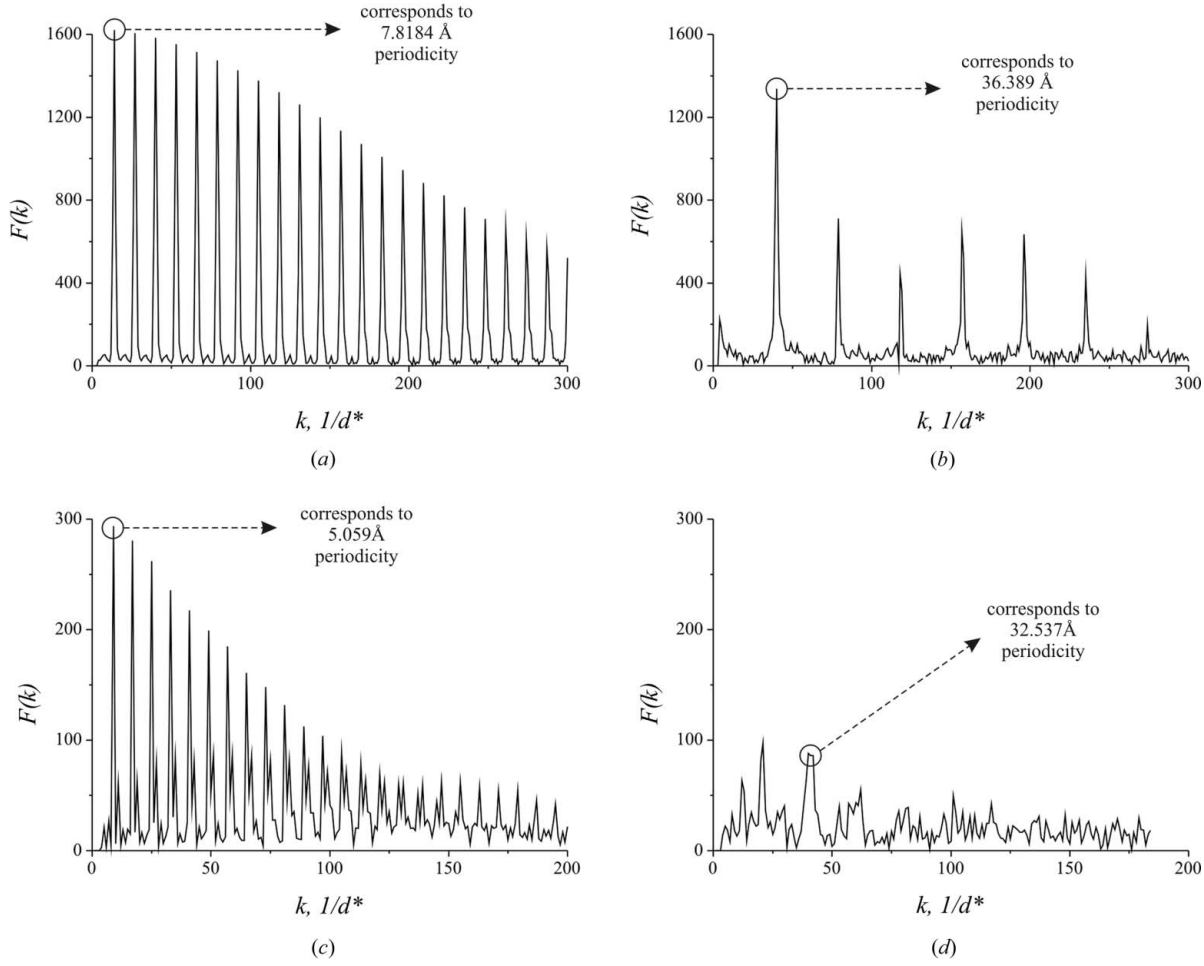


Figure 2
One-dimensional Fourier transform of reflection positions projected onto the main crystallographic directions. Projections of single-crystal X-ray data for sample D (see Table 1) onto (*a*) direction **c** and (*b*) direction **b**; and ADT tilt-series data for sample B (Table 1) onto (*c*) direction **a** and (*d*) direction **c**.

therefore in a first step difference vectors for all measured reflection positions are calculated. In a way, difference vectors represent the autocorrelation of the lattice and therefore enhance the short vectors around the origin. These short vectors are likely to include the basis vectors of the lattice. Unit-cell basis-vectors determination within the difference-vectors space (DVS) is routinely done in single-crystal X-ray crystallography (Kabsch, 1993, 2010); nevertheless, due to the nuances of electron-diffraction data as listed above, the DVS from ADT data also has specific characteristics.

2.1. Determination of reflection positions

An ADT tilt series represents a stack of electron-diffraction patterns with an angular relationship between them – usually an equidistant tilt step of 1°. Preprocessing of the data includes an optional *background correction*, *shifting* of the patterns to a common origin (the exact position of the central beam is usually unknown and has to be determined; furthermore, the shift can be different for different patterns in the same stack) and *rotation* of the patterns to the tilt axis [the tilt-axis azimuthal rotation within a stack is *a priori* not known and can be calculated for each stack using the internal consistency of the lattice as a criterion, see Kolb *et al.* (2009)]. The preprocessing steps are detailed in Kolb *et al.* (2008).

Once the preprocessing has finished, the positions of the reflections can be extracted for unit-cell determination. The reflections can be found within the two-dimensional diffraction patterns – frames (two-dimensional peak search), or within the reconstructed three-dimensional diffraction volume (three-dimensional peak search). Two-dimensional peaks are more sensitive to position uncertainty, due to the problems listed above, while peaks found within the three-dimensional diffraction volume are better defined due to averaging: neighbouring voxels are fused to a single reflection. Therefore hereafter we concentrate on the use of three-dimensional peaks.

A three-dimensional peak-search procedure locates three-dimensional objects with an overall intensity above a given background threshold. Besides, a limit for the minimum volume of a reflection (in voxels) is set to filter out hot spots, and for a maximum volume to exclude fused reflections. For each reflection found, the arithmetical mean of all voxels assigned to the *i*th reflection $r_i^{\rm mean}$, the centre of gravity $r_i^{\rm w.mean}$

and the position corresponding to the maximal intensity voxel $r_i^{\max}$ are calculated. If a hot spot falls onto a reflection, the position of $r_i^{\max}$ will be significantly influenced, while $r_i^{\text{w.mean}}$ is less sensitive to it. Nevertheless, this is a rare situation, and in practice all three positions are very close to each other, so any of them can be used.

2.2. Difference vectors

After a set of three-dimensional reflection positions is extracted, difference vectors are calculated between all pairs of positions except for the primary beam. Fig. 3 schematically shows the idealized two-dimensional lattice, difference vectors calculated for the lattice and the difference-vector space (DVS). Difference vectors point exactly to a node of the lattice, therefore the lattice of the DVS coincides exactly with the original lattice.

For experimental data the lattice will never be perfect. The distortion of the lattice leads to a large variety of difference vectors with varying but similar lengths. Owing to the variation in difference-vector lengths, it is no longer possible to assume the difference vectors to be the unit-cell lengths or a multiple of these. Thus, the difference-vector space consists of groups of spots around each lattice node. The extent of points spread within the groups is correlated with the amount of distortion in the original lattice.

The analysis of these kinds of data is done through data clustering (see Kaufman & Rousseeuw, 1990). The aim of clustering is to group or classify the data based on a specific property, fitting the problem. In our case the property is the mutual proximity of points in the difference-vector space. Thus, the DVS should include *clusters*, the arrangement of which reflects the periodicity of the lattice. The number of clusters is related to the lattice parameters and cannot be known *a priori*. Apart from clusters, *noise* can be present in the data originating from reflections not coherent with the major lattice: additional crystals or hot spots.

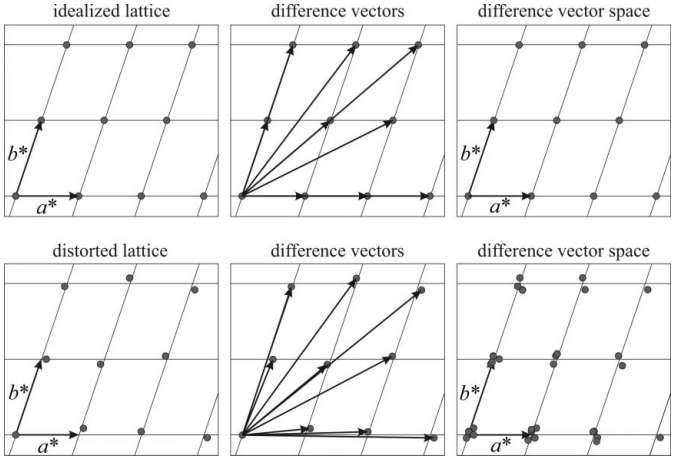


Figure 3
Top row: idealized two-dimensional lattice, difference vectors calculated for this lattice and difference-vector space where the positions of vectors overlap; bottom row: distorted lattice, difference vectors calculated for the distorted lattice and the difference-vector space. For the distorted lattice the difference-vector space consists of groups of points around each lattice node.

The outcome of the clustering procedure is a list of clusters with their coordinates. These clusters should represent the reciprocal lattice and they are typically arranged around the origin of the lattice. Once a sufficient number of clusters are found, the unit-cell basis vectors can usually be determined within the clusters as three shortest non-coplanar vectors. Subsequent Niggli cell reduction (Niggli, 1928) or a basis transformation in order to detect possible lattice centring finally delivers the unit-cell metric and the orientation matrix related to the ADT data set.

The number of difference vectors for n reflection positions is $n(n-1)$. The difference vectors carry inherent inversion symmetry, thus there are only $n(n-1)/2$ independent vectors. For a set with 1000 reflection positions there will be around half a million difference vectors which need to be analysed. Even though the lattice basis vectors are likely to be represented by the shortest difference vectors, and therefore there is no need to analyse longer difference vectors, a fast and efficient algorithm is sought. In the next section we will explore the available options for clustering the data and finding unit-cell vectors.

3. Overview of clustering approaches

Clustering analysis of data is used in many fields to group data into similar (homogeneity) and dissimilar (heterogeneity) categories, and is used in fields as diverse as data mining and image analysis. There are three basic types of clustering methods: partitioning methods, hierarchical methods and density-based methods, each of which is described in brief below. In our case we wish to find a clustering algorithm that will group the difference vectors together to produce a difference-vector lattice which corresponds to the lattice of the unit cell under investigation, given no prior knowledge of how many clusters will be produced or what shape the clusters will form.

3.1. Partitioning methods

Partitioning methods divide the data D into k clusters, where k is a user-defined input parameter. In order to determine a good choice of k a significant amount of knowledge about D is required, which in many cases is not available. Principally, it is possible to run the clustering for various values of k, and then use a suitable figure of merit to evaluate the results. The additional evaluation of the results can significantly increase the run time depending on the range of k. Typically, partitioning methods start with a random or given partition of the data D into a set C of k clusters. Each cluster C_i ($i = 1, \ldots, k$) is represented by one object x_i. This can, for instance, be the centre of gravity or the mean of the cluster (this is known as *k-means* method, see MacQueen, 1967), or a data point closest to the physical centre of the cluster (*k-medoid* method, see Kaufman & Rousseeuw, 1990). The rest of the data are then assigned to a cluster in an iterative

process, using a given criterion, most commonly the square-error criterion (*e.g.* *k*-means):

$$\sum_{i=1}^{k} \sum_{x \in C_i} ||x - x_i||_2^2.$$

Minimizing of the square-error criterion leads to high homogeneity within clusters. In each iteration a two-step procedure is used: first, *k* representatives optimizing the criterion function (*e.g.* the centre of gravity for each of the *k* clusters) are determined; subsequently, the clusters are rearranged by assigning each object to the 'closest' representative. The iteration stops when no changes occur after the rearrangement. All data points are assigned to a particular cluster. Therefore, noise in the data makes the assignment of points to clusters ambiguous, making the algorithm not reliable.

A well known method, related to *k*-means, is Expectation Maximization (EM; Dempster *et al.*, 1977), which differs from *k*-means by determining a *fuzzy* clustering partition. In a fuzzy partition certain probabilities for belonging to a cluster are assigned to each data point, then the clusters are formed based on a probability criterion. Other common methods related to *k*-medoid are Partitioning Around Medoids (PAM; Kaufman & Rousseeuw, 1990), Clustering LARge Applications (CLARA; Kaufman & Rousseeuw, 1990), and Clustering Large Applications based on RANdomized Search (CLARANS; Ng & Han, 1994). All these methods are computationally expensive. To reduce the computation time CLARA divides the data into subsets and uses PAM for every subset. This leads to faster computation time but a deterioration of the clustering quality, because the global minimum of the criterion function is not guaranteed to be reached.

3.2. Hierarchical methods

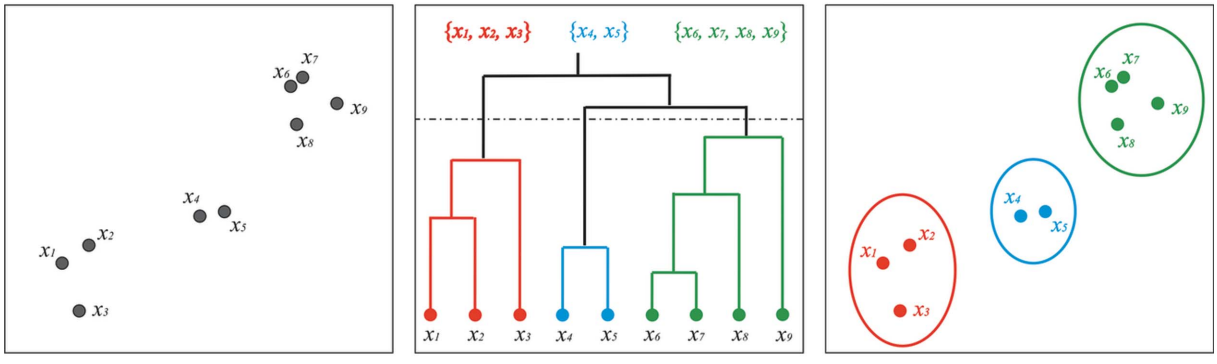
Hierarchical methods are often the most recognizable of the clustering methods because of the tree or *dendrogram* used to represent the data. They have already been used in crystallography (for instance for the analysis of X-ray powder-diffraction profile matching; see Barr *et al.*, 2004). A *dendrogram* has a branch at every node representing a cluster and a horizontal cut at any height is a possible partition of the data *D*. In the root node *D* consists of only one cluster. Below the data are sequentially divided into smaller, finer partitions until each cluster consists of only one object (leaves of the tree). The *dendrogram* can be built from root to leaves (top-down divisive methods) or from leaves to root (bottom-up agglomerative methods). In contrast to partitioning methods, the number *k* of clusters is not needed as an input parameter. The clusters are merged or divided until a termination criterion is reached.

The *single-linkage* (Sibson, 1973) algorithm is an agglomerative method which in each step merges the two nearest clusters (Fig. 4). In this case (and generally for agglomerative methods) a termination criterion is the minimal distance $d_{\min}$ between clusters. For the single-linkage algorithm, for example, the iteration stops if the distance between the two nearest clusters is smaller than $d_{\min}$. The selection of the termination criterion is the key point for all hierarchical methods, as it assures that the data are partitioned into clusters of appropriate size to reveal the desired information in the data. If the termination criterion is not known initially, the full dendrogram should be calculated. The appropriate data partitioning can then be selected.

Hierarchical methods can tolerate a low amount of noise in the data. Hierarchical methods are typically slow [$O(n^2)$ or higher, where *n* is the number of points in the data set]. Typical hierarchical methods are Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH; Zhang *et al.*, 1996) and CURE (Guha *et al.*, 1998). BIRCH first organizes the data into a height-balanced cluster-feature tree (CF tree) in order to divide the clustering into smaller problems. Then an arbitrary clustering method (*e.g.* CLARANS) is applied to the leaves of the tree to determine the partitioning of the data.

3.3. Density-based methods

The principal idea of density-based methods is to find high-density regions separated by regions of low density. The main criterion is that the density of the noise has to be lower than the density of any cluster. These methods are particularly stable in the presence of noise in the data. There are two major ideas behind density-based methods. One defines dense



(*a*) (*b*) (*c*)

Figure 4
Scheme of data clustering in the single-linkage method: (*a*) a data set consisting of nine points; (*b*) a dendrogram of the single-linkage procedure; (*c*) clusters determined with single linkage and an optimal termination criterion.

regions through neighbourhoods of points (regions with a minimum number of points inside). This is the main idea of Density Based Spatial Clustering of Applications with Noise (DBSCAN; Ester *et al.*, 1996), which will be discussed in more detail in §4. The other group of algorithms uses kernel density estimation (KDE) through kernel functions (*e.g.* Gaussian kernel), for instance DENsity-based CLUstEring (DENCLUE; Hinneburg & Keim, 1998).

Density-based methods are able to deal with noisy data and have no limitations on the cluster shape. These algorithms can be implemented very efficiently with index-based data structures [$O(n \log(n))$, *e.g.* DBSCAN and DENCLUE]. Among other algorithms extending the idea of DBSCAN, Ordering Points To Identify the Clustering Structure (OPTICS; Ankerst *et al.*, 1999), Density Differentiated Spatial Clustering (DDSC; Borah & Bhattacharyya, 2008), Density Clustering Based on Outlier Removal (DCBOR; Fahim *et al.*, 2008) and the Shared Nearest Neighbor clustering algorithm (SNN; Ertoz *et al.*, 2003) should be mentioned.

3.4. Applicability of clustering algorithms to ADT data

In order to cluster the DVS produced from ADT data efficiently, an appropriate clustering technique is required. Partitioning methods are not appropriate in this case because they cannot handle noisy data. Additionally, the number of clusters k can vary over a large range, thus making the computations very time consuming.

The hierarchical methods are not best suited to the task either, because they are not robust enough against particularly noisy data such as can be encountered in ADT. Furthermore, hierarchical methods are not computationally efficient enough to deal with large data sets (as mentioned above, half a million difference vectors can easily be produced from ADT data).

The density-based methods are robust to very noisy data and can be efficiently implemented. In addition, there is no restriction on the cluster shape. Therefore they are best suited for DVS clustering. DBSCAN is preferred over DENCLUE due to its easier implementation. Other extensions of DBSCAN require more input parameters, which may give additional flexibility during the clustering, but makes the underlying philosophy less intuitive and understandable, and thus the choice of the optimal parameters more problematic. Other algorithms which do not require additional inputs are not as efficient in implementation and have a runtime of $O(n^2)$ or higher.

On the basis of the robustness to noise and computational efficiency, as well as ease of implementation, we have opted for the DBSCAN algorithm as our preferred method for cluster analysis of ADT data in order to obtain unit-cell parameters.

4. DBSCAN

4.1. The algorithm

The DBSCAN algorithm (Ester *et al.*, 1996) uses two input criteria to define the minimum density of a cluster: ε, the size of the neighbourhood which will be reviewed for each data point, and the minimum number of points *minPts* which have to be inside the neighbourhood to define a dense region as a part of a cluster. In terms of DVS built from ADT data, these two criteria define how many difference vectors should be present within a specified region to form a cluster.

The DBSCAN algorithm clusters data in an iterative procedure. First of all *core points* are identified. A core point is a data point which has at least *minPts* different data points in its ε-neighbourhood (Fig. 5). The ε-neighbourhood of a data point $p \in D$ is defined as

$$U_\varepsilon(p) := \{q \in D | dist(p, q) \leq \varepsilon\}$$

for $\varepsilon > 0$ and an arbitrary distance function dist (*e.g.* Euclidean distance). The choice of the distance function affects the clustering and should be adapted to the problem. In the case of three-dimensional DVS, the Euclidean distance is a proper choice. A data point p is called a core point iff $|U_\varepsilon(p)| \geq$ *minPts*. Core points are parts of very dense regions within a data set. After the core points are found, the algorithm identifies sets of *directly density-reachable*, *density-reachable* and *density-connected* points of data.

A data point $p \in D$ is called *directly density-reachable* from $q \in D$ with respect to ε and *minPts* if

$$\text{(i) } p \in U_\varepsilon(q) \quad \text{and} \quad \text{(ii) } |U_\varepsilon(q)| \geq minPts.$$

These say that p is *directly density-reachable* from q if p lies within the ε-neighbourhood of a core point q (Fig. 5).

A data point $p \in D$ is called *density-reachable* from $q \in D$ with respect to ε and *minPts* if there is a chain of points $p_1, \ldots, p_m \in D$, $p_1 = q, p_m = p$, such that p_{i+1} is *directly density-reachable* from p_i. This essentially determines whether a series of points in a chain belong to the same dense region or not (Fig. 5).

As Fig. 5 shows, *density-reachable* is a *non-symmetric* relation: r is density-reachable from q but q is not density-reachable from r. Both objects r and q should belong to the same cluster.

A data point $p \in D$ is called *density-connected* to $q \in D$ with respect to ε and *minPts* if there exists a data point $o \in D$ such that both p and q are *density-reachable* from o with respect to ε and *minPts* (Fig. 5). *Density-connected* is the symmetric extension of *density-reachable*.

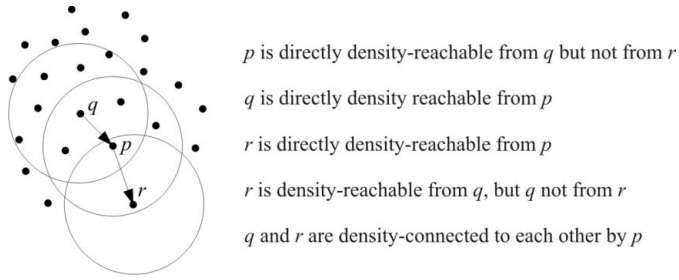


Figure 5
A schematic representation of DBSCAN terminology (*minPts* = 5; *p*: core point $|U_\varepsilon(p)| = 6$; *q*: core point $|U_\varepsilon(q)| = 9$; *r*: border point $|U_\varepsilon(r)| = 2$).

A *density-based cluster* is defined as a maximal set of *density-connected* points with respect to *density reachability*. Therefore we are trying to maximize the number of *density-connected* clusters in the data set, given the parameters for the ε-neighbourhood and the minimum number of points *minPts*. All the points in the data set that do not satisfy the density-connected definition are considered to be noise.

Clusters are formed by iteratively looking at each point in the data set. If a given point satisfies the neighbourhood and minimum number of points criteria, then it is considered to be part of a cluster. This is achieved by first identifying a core point. From each core point all *density-reachable* points are collected to build a cluster. Then the iteration starts with the identification of a new core point in the data set that does not already belong to an existing cluster. For the proof of this algorithm and further details see Ester *et al.* (1996).

During the calculation it is necessary to compute the ε-neighbourhood of each point only once. The naive implementation of this calculation computes the distance to each data point and needs $O(n)$ time for each calculation. Therefore the whole algorithm has a run time of $O(n^2)$. By utilizing efficient data structures such as R*-tree (Beckmann *et al.*, 1990) to determine the neighbourhood it is possible to complete the calculation in $O(\log(n))$ run time. If this is realised within the DBSCAN algorithm, the complete run time is $O(n\ \log(n))$, which is suitable for large data sets.

4.2. Clustering control parameters – ε and *minPts*

DBSCAN is implemented in the *ADT3D* package (Nanomegas, Belgium) as a core routine of unit-cell-parameter determination. The procedure can run automatically; however, in some cases it requires additional adjustment of the control parameters. The parameters can be changed interactively by the user. Here, once more, we summarize the action of ε and *minPts*.

Each data set has a different reciprocal lattice given by the unit-cell metric. Therefore, it can be necessary to adjust the parameters ε and *minPts*. Fig. 6 shows different situations which can be realised when different control parameters are used. For a given value of *minPts*, too high a value of the ε-neighbourhood can lead to a situation where a point of a cluster falls into the ε-neighbourhood of another cluster (Fig. 6*a*), causing a fusion of separate clusters (Fig. 6*b*). This can either be resolved by reducing the ε-neighbourhood size, or

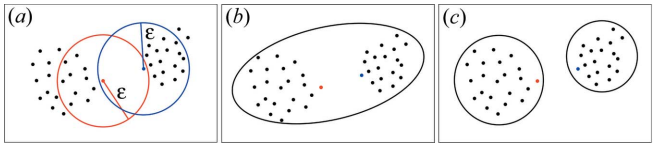


Figure 6
Different situations of DBSCAN clustering for different control parameters: (*a*) a critical situation with ε smaller than the distance between border points belonging to different clusters; (*b*) cluster fusion with the same value of ε as in (*a*) and *minPts* = 12; (*c*) cluster separation with *minPts* = 13.

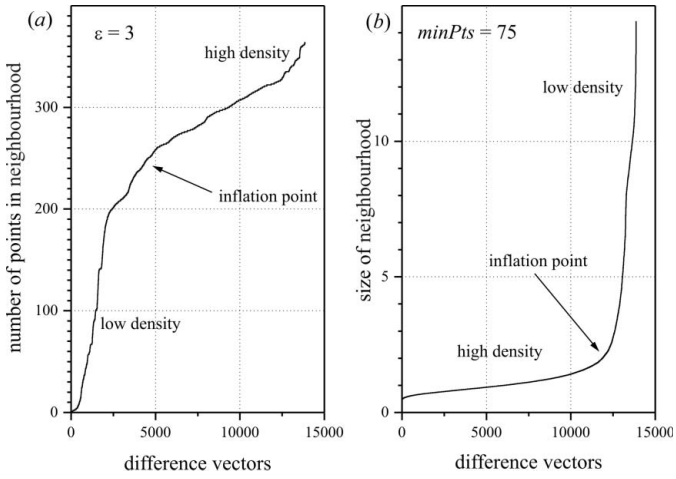


Figure 7
Mutual influence of the two DBSCAN parameters *minPts* and ε: (*a*) plot of *minPts* for all difference vectors of a data set of material A keeping ε fixed (ε = 3); (*b*) plot of ε for all difference vectors keeping *minPts* fixed to 75.

alternatively, *minPts* can be increased to result in the ideal cluster separation (Fig. 6*c*). In contrast, too small a value of ε can lead to a splitting of a cluster. If the chosen minimum density for clustering (defined through the combination of the two control parameters) is too high, it may happen that no clusters will be found at all.

For high-quality data, the default clustering parameters are often good enough to provide a satisfactory result. For very limited data or materials with long lattice constants, additional adjustment of the parameters is required. As ε and *minPts* substantially influence each other, it is recommended to change them separately. One parameter can be fixed and the result of the clustering can be analysed for various values for the second parameter. The result of such an evaluation is presented for data for material A in Fig. 7. All difference vectors are sorted according to their ordinate value in order to produce a smooth graph. On the left side of the figure (Fig. 7*a*) ε is fixed to the value of 3, and for the complete set of the difference vectors the number of points lying inside the 3-neighbourhood is calculated. Now we should search for regions with a high density. The start of the plot with a high gradient (difference vectors 0 to approximately 2000) corresponds to regions with a low density. The dense regions start at the inflation point. To avoid including too many points from low-density regions as border points into a cluster, the optimal *minPts* should be slightly larger than that corresponding to the inflation point. In Fig. 7(*b*) in contrast, *minPts* is fixed to the value of 75, and for the complete set of the difference vectors the minimum size of the neighbourhood is calculated so that 75 points are inside this neighbourhood. The optimal ε value should be chosen close to the inflation point, slightly shifted to the region of higher density.

These plots do not give strict values for the parameters, but show in which regions they should be. Since the parameters are not completely independent, there is no single optimal set of the parameters, but many pairs which give a similar result,

Table 2
Data for the clustering procedure for the test samples.

	Sample			
	A	B	C	D
Tilt range (°)	121	111	61	180
No. of reflections	978	424	246	1626
No. of difference vectors	17748	5816	6112	79932
No. of clusters	42	44	52	108
No. of difference vectors assigned to clusters	16296	3478	1958	79462
No. of difference vectors assigned to noise	1452	2338	4154	470
Fraction of noise in the difference vectors (%)	8	40	68	<1
a' (Å)†	8.90	32.50	16.97	36.65
b' (Å)†	7.23	5.11	11.15	8.04
c' (Å)†	5.53	5.08	7.37	7.85
α (°)	90.14	119.57	90.12	99.49
β (°)	90.27	90.31	90.05	90.08
γ (°)	89.52	89.89	88.14	90.04

† The unit-cell parameters a', b', c' are automatically found after the clustering has been done, and therefore appear in descending order. The order does not correspond to the standard crystallographic settings used in Table 1.

for instance, as seen in Fig. 7, $\varepsilon = 3$, *minPts* = 250 and $\varepsilon = 1.8$, *minPts* = 75.

5. Examples

Hereafter some examples of unit-cell basis-vectors determination using DBSCAN from ADT data are presented. The clustering routine has worked successfully when the clusters that are found form an equidistant three-dimensional lattice reasonably describing all difference vectors. From all the clusters that are found, three clusters are selected automatically, representing three shortest non-coplanar vectors. The three vectors are sorted by their length and assigned to **a***, **b***, **c*** (therefore, the lattice parameters may appear in nonstandard crystallographic settings). For a primitive lattice these vectors directly describe the unit cell, for a centred lattice they have to be transformed into the correct settings. The test materials are summarized in Table 1. The lattice-parameters determination of two inorganic (A and B) and two organic (C and D) samples are described. The DBSCAN clustering procedure was used within the *ADT3D* program (Nanomegas, Belgium).

For the materials listed, the positions of the reflections were found within the reconstructed reciprocal volume. Then for these positions difference vectors were calculated. The difference vectors were then subjected to a DBSCAN clustering routine. The unit-cell parameters were determined based on the clusters closest to the origin. A summary of the data clustering is presented in Table 2.

5.1. A: barite

The barite tilt series is the easiest and most straightforward example. The data were collected within a large part of the reciprocal volume (2/3 of the complete reciprocal space). A little less than one thousand reflections were found for the unit-cell determination procedure and from these 17 748 difference vectors were produced (Table 2). These are not all the difference vectors that can be calculated between the reflection positions; only the difference vectors closest to the origin within a given resolution shell are calculated, as these are the most likely to include the basis vectors of the lattice. The clustering procedure with parameters $\varepsilon = 2$, *minPts* = 100 resulted in 42 clusters forming a clear lattice (see Fig. 8). Only a few difference vectors (8%) were rejected from clustering and assigned to noise in the data. The lattice parameters are found automatically as the three shortest non-coplanar vectors and assigned to **a***, **b***, **c***, and therefore always appear in the descending sequence: 8.90, 7.23, 5.53 Å. The lengths of the vectors match well with the expected values (Table 1). The angles are all close to 90° with an accuracy of better than 0.5°.

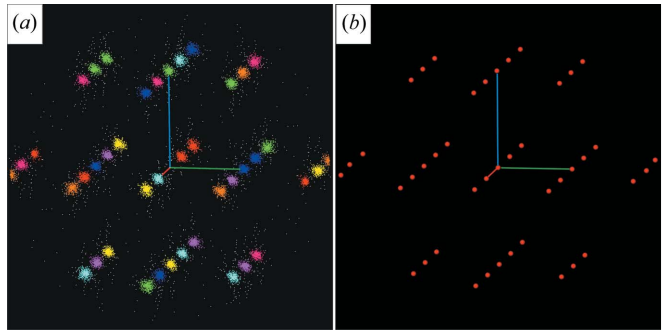


Figure 8
Difference-vector space of barite. (*a*) The points assigned to clusters are coloured, noise points are grey; (*b*) the centres of clusters building a periodic lattice are shown in red. The unit-cell vectors **a***, **b***, **c*** are colour-coded as red, green and blue.

5.2. B: pseudo-spinel

Long lattice parameters especially in combination with short lattice vectors usually require additional tuning of the clustering parameters. The pseudo-spinel example demonstrates unit-cell determination for a material with very different lengths of the unit cell in different directions. The data were collected within a large tilt range (Table 2) and 5816 difference vectors were selected for clustering. Selecting the value of the neighbourhood parameter ε close to or higher than the distance between the lattice nodes causes fusion of clusters. Fig. 9(*a*) shows a result of a clustering procedure using too high a value for the neighbourhood ε. The situation is resolved by decreasing ε to 2 with the subsequent adjustment of *minPts* (here *minPts* = 15). With these clustering parameters compact equidistant clusters are produced along the problematic direction.

Remarkably, the same data were not able to deliver the unit-cell vectors using the one-dimensional Fourier transform routine (Fig. 2*c*,*d*). This demonstrates the advantage of the clustering approach for electron-diffraction data with long lattice constants.

This example demonstrates the intrinsic limitation of the clustering regarding long crystallographic axes. As long as the

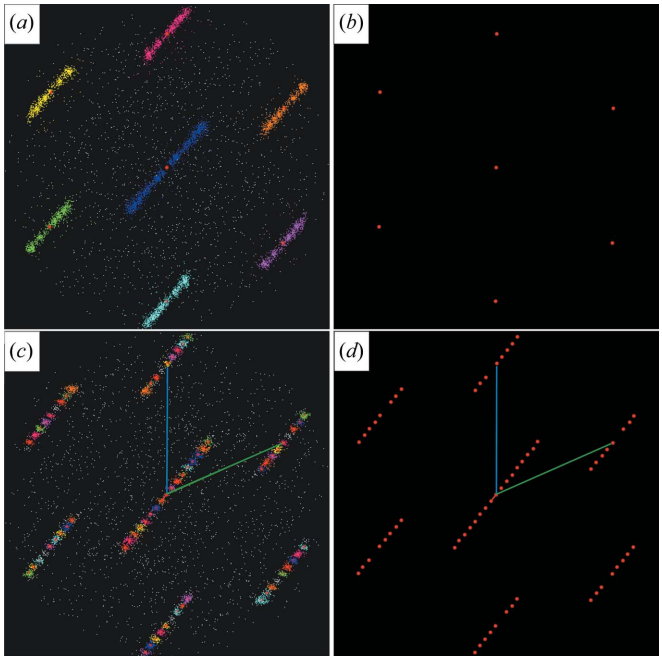


Figure 9
Pseudo-spinel, a structure with one large lattice parameter. (*a*) Clustered difference vectors using too high a value of ε; (*b*) centres of the clusters found using a high value of the neighbourhood; (*c*) clustering procedure using a reduced ε with adjusted *minPts*; (*d*) cluster centres using a low value of ε forming a lattice.

distance between the nodes of the lattice is significantly larger than the neighbourhood ε and the amount of noise in the data is not too high, there is no problem in determining a proper value for *minPts* able to separate the clusters and thereafter determine the lattice basis vectors. When ε is close to or higher than the distance between the lattice nodes, difficulties may occur in separation of clusters. The effective size of the cluster is determined by the data (crystal) quality, so it is not possible to put a rigid limit on the crystallographic axis length. Principally, if there is only one long axis in the structure, and the cluster separation along this direction failed, the data can be extracted along these lines and treated specially in order to elucidate the periodicity. Further discussion on this point is beyond the scope of this paper.

5.3. C: orthorhombic paracetamol

ADT data from organic crystals have two major problems: (i) since the crystalline lattice degrades fast under the beam, the tilt sequences are usually very short; and (ii) the diffracted intensities are very weak, so noise peaks will be found together with the reflection positions. In short, the ADT data of organic crystals are particularly noisy and provide a very limited part of reciprocal space.

Fig. 10 shows the difference-vector space of orthorhombic paracetamol. The amount of noise in the difference-vector data is particularly high – 68% (Table 2). Nevertheless, the basis vectors were found automatically. The high amount of noise in the data results in more spread out (less compact and less dense) clusters. This can significantly affect the resulting lattice parameters. The angle between the two long axes (two clusters nearest to the origin) is predominantly sensitive to noisy data: thus the highest deviation of the angles here was observed for γ (88.14 *versus* 90°).

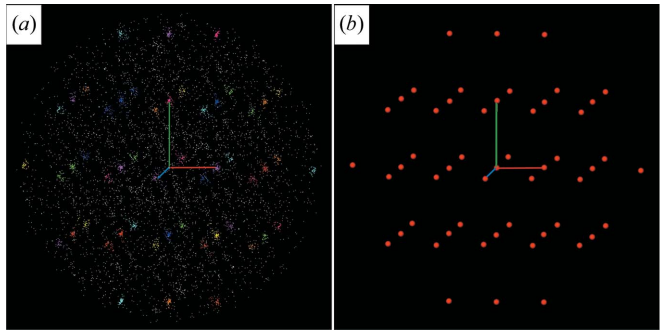


Figure 10
Difference-vector space of orthorhombic paracetamol. (*a*) The points assigned to clusters are coloured, noise points are grey; (*b*) the centres of clusters building a periodic lattice are shown in red.

The two examples B and C demonstrate two different cases of problems in data clustering: a long crystallographic axis and low quantity of data/noisy data. For these two cases the clustering parameters were adjusted manually to achieve the final result. Nevertheless, for data with insufficient sampling along an axis or when there is too much noise in the data, the method may fail to find the unit-cell vectors that describe the reciprocal lattice.

5.4. D: propellan – X-ray data

The data for propellan were collected on a Stoe diffractometer (Stoe & Cie GmbH, Darmstadt, Germany) with Mo $K\alpha$ radiation. The reciprocal volume was created within the *X-Area* software (Stoe) and then converted to an MRC file with an x3D-to-MRC converter. The three-dimensional diffraction volume was created from a single ω run of 180 frames with an ω increment of 1°. For the three-dimensional interpolation, the data were taken up to 0.6 Å diffraction resolution and binned in steps of 0.003 Å. The resulting volume had dimensions (in voxels) of 401 × 401 × 401. The MRC volume was then opened in *ADT3D* and processed (peak search and determination of unit-cell parameters) as usual.

The unit-cell parameters found *via* DBSCAN clustering match well with those from *X-Area* (Tables 1 and 2). The striking difference to electron-diffraction data is the low amount of noise in the data – less than 1% of the difference vectors were rejected by the clustering routine. Fig. 11 shows the difference-vector space of propellan. No noise is seen in the data.

The high quality of X-ray data compared to electron ADT data explains why the projection Fourier analysis was able to find the periodicity in X-ray data, and often fails for electrons. The clustering approach applied here provides a reliable unit-cell metric, and can in principle be used for single-crystal

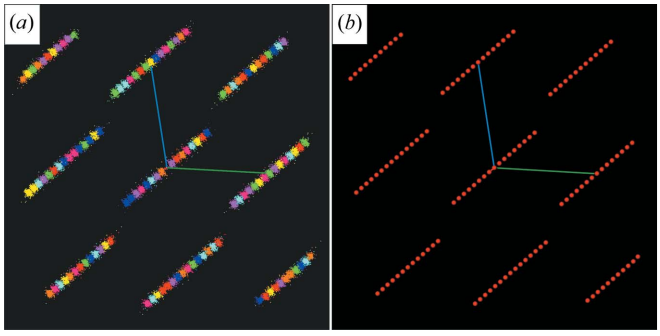


Figure 11
Difference-vector space of propellan built using X-ray data. (*a*) The points assigned to clusters are coloured, noise points are grey; (*b*) the centres of clusters building a periodic lattice are shown in red.

X-ray data when the data quality is not satisfactory for the Fourier analysis method.

6. Conclusion

The clustering is an essential part of automated diffraction tomography (ADT) data processing, delivering the lattice basis vectors for single-crystal electron-diffraction data. In the present paper the concepts of cluster analysis and various methods were reviewed. From the list of clustering approaches a density-based algorithm DBSCAN was selected and applied to electron ADT data.

The determination of the unit-cell metric using DBSCAN was demonstrated by several examples. In all cases, the clustering approach was able to find the unit-cell vectors. The robustness of the method was demonstrated by an example of the determination of the unit cell of orthorhombic paracetamol represented by particularly noisy data. An example of unit-cell parameters determination from single-crystal X-ray data was also given.

The lattice-basis-vector determination approach through difference-vectors clustering as presented in this paper has a general character and can be applied to any kind of single-crystal diffraction data. It is robust for low amounts of data and high levels of noise, and can therefore be applied when other methods have difficulties. Thus, the clustering approach can be applied where limited single-crystal data are available, such as in the case of diamond anvil or environmental cells, or if a crystal has degraded prematurely in the beam, leaving a limited data set to work with.

The authors are grateful to Dr Dieter Schollmeyer (University of Mainz, Germany) for proving the single-crystal X-ray data for propellan and fruitful discussions. The authors highly appreciate the help of Dr Friedemann Hahn (Stoe, Darmstadt, Germany) with the x3D-to-MRC data conversion. The sample of orthorhombic paracetamol was kindly provided by Professor Elena Boldyreva (Novosibirsk State University, Russia). The work was done under the financial support of the Center of Computational Science of the Johannes Gutenberg University Mainz and and the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 625, Schwerpunktprogram SPP1415).

References

Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. (1999). *SIGMOD 1999. Proc. ACM SIGMOD Intl Conf. Manag. Data*, edited by A. Delis, C. Faloutsos & S. Ghandeharizadeh, pp. 49–60. Philadelphia: ACM Press.
Barr, G., Dong, W. & Gilmore, C. J. (2004). *J. Appl. Cryst.* **37**, 243–252.
Beckmann, N., Kriegel, H.-P., Schneider, R. & Seeger, B. (1990). *Proc. ACM SIGMOD Intl Conf. Manag. Data*, pp. 322–331. Atlantic City: ACM Press.
Birkel, C., Mugnaioli, E., Gorelik, T., Panthöfer, M., Kolb, U. & Tremel, W. (2010). *J. Am. Chem. Soc.* **132**, 9881–9889.
Borah, B. & Bhattacharyya, D. K. (2008). *J. Comput.* **3**, 72–79.
Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). *J. R. Stat. Soc. Ser. B*, **39**, 1–38.
Dorset, D. L. (1995). *Structural Electron Crystallography*. New York: Plenum Publishing Corporation.
Ertoz, L., Steinbach, M. & Kumar, V. (2003). *Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. Proc. Third SIAM Intl Conf. Data Min.* p. 47. Philadelphia: SIAM.
Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996). *Proc. 2nd Intl Conf. Knowl. Discov. Data Min.* pp. 226–231. Portland: AAAI Press.
Fahim, A. M., Saake, G., Salem, A. M., Torkey, F. A. & Ramadan, M. A. (2008). *World Acad. Sci. Eng. Technol.* **45**, 171–176.
Guha, S., Rastogi, R. & Shim, K. (1998). *SIGMOD '98. Proc. 1998 ACM SIGMOD Intl Conf. Manag. Data*, pp. 73–84. New York: ACM.
Haisa, M., Kashino, S. & Maeda, H. (1974). *Acta Cryst.* B**30**, 2510–2512.
Hinneburg, A. & Keim, D. (1998). *Proc. Fourth Intl Conf. Knowl. Discov. Data Min.* pp. 58–65. New York: AAAI Press.
Jacobsen, S. D., Smyth, J. R., Swope, R. J. & Downs, R. T. (1998). *Can. Mineral.* **36**, 1053–1060.
Jiang, L., Georgieva, D. & Abrahams, J. P. (2011). *J. Appl. Cryst.* **44**, 1132–1136.
Jiang, L., Georgieva, D., Nederlof, I., Liu, Z. & Abrahams, J. P. (2011). *Microsc. Microanal.* **17**, 879–885.
Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: John Wiley and Sons.
Kolb, U., Gorelik, T., Kübel, C., Otten, M. T. & Hubert, D. (2007). *Ultramicroscopy*, **107**, 507–513.
Kolb, U., Gorelik, T. & Mugnaioli, E. (2009). *Mater. Res. Soc. Symp. Proc.* 1184-GG01–1184-GG05.
Kolb, U., Gorelik, T. & Otten, M. T. (2008). *Ultramicroscopy*, **108**, 763–772.
Kolb, U., Mugnaioli, E. & Gorelik, T. E. (2011). *Cryst. Res. Technol.* **46**, 542–554.
MacQueen, J. (1967). *5th Berkeley Symp. Math. Stat. Prob.* **1**, 281–297.
Mirion, M., Waldvogel, S. R. & Schollmeyer, D. (2012). Private communication (deposition number 871116). CCDC, Cambridge, England.
Mugnaioli, E., Andrusenko, I., Schüler, T., Loges, N., Dinnebier, R., Panthöfer, M., Tremel, W. & Kolb, U. (2012). *Angew. Chem. Int. Ed.* **51**, 7041–7045.
Ng, R. T. & Han, J. (1994). *Proc. 20th Intl Conf. Very Large Databases, Santiago, Chile*, pp. 144–155. San Francisco: Morgan Kaufmann Publishers.
Niggli, P. (1928). *Handbuch der Experimentalphysik*, Vol. 7, Part 1, pp. 108–176. Leipzig: Akademische Verlagsgesellschaft.

Rossmann, M. G. (2001). *International Tables for Crystallography*, Vol. F, edited by M. G. Rossmann & E. Arnold, pp. 209–211. Dordrecht: Kluwer Academic Publishers.
Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. (2004). *J. Appl. Cryst.* **37**, 399–409.
Sibson, R. (1973). *Comput. J.* **16**, 30–34.
Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *J. Appl. Cryst.* **30**, 1036–1040.
Vainshtein, B. K. (1964). *Structure Analysis by Electron Diffraction.* Oxford: Pergamon Press.
Zhang, T., Ramakrishnan, R. & Miron Livny, M. (1996). *Proc. 1996 ACM SIGMOD Intl Conf. Manag. Data Montreal, Quebec, Canada*, edited by H. V. Jagadish & Inderpal Singh Mumick, pp. 103–114. ACM Press.
Zou, X. D., Hovmöller, A. & Hovmöller, S. (2004). *Ultramicroscopy*, **98**, 187–193.