Pre-treatment of soil X-ray powder diffraction data for cluster analysis

Benjamin M. Butlera,⁎, Andrew M. Silab, Keith D. Shepherdb, Mercy Nyamburab,c, Chris J. Gilmorec, Nikolaos Kourkoumelisd, Stephen Hilliera,e

a The James Hutton Institute, Craigiebuckler, Aberdeen AB15 8QH, UK
b World Agroforestry Centre (ICRAF), P.O. Box 30677-00100 GPO, Nairobi, Kenya
c School of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK
d Department of Medical Physics, University of Ioannina, 45110 Ioannina, Greece
e Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), SE-75007 Uppsala, Sweden

1. Introduction

X-ray powder diffraction (XRPD) is widely applied for the qualitative and quantitative analysis of soil mineralogy. In recent years, high-throughput XRPD has resulted in soil XRPD datasets containing thousands of samples. The efforts required for conventional approaches of soil XRPD data analysis are currently restrictive for such large data sets, resulting in a need for computational methods that can aid in defining soil property – soil mineralogy relationships. Cluster analysis of soil XRPD data represents a rapid method for grouping data into discrete classes based on mineralogical similarities, and thus allows for sets of mineralogically distinct soils to be defined and investigated in greater detail. Effective cluster analysis requires minimisation of sample-independent variation and maximisation of sample-dependent variation, which entails pre-treatment of XRPD data in order to correct for common aberrations associated with data collection.

A 2^4 factorial design was used to investigate the most effective data pre-treatment protocol for the cluster analysis of XRPD data from 12 African soils, each analysed once by five different personnel. Sample-independent effects of displacement error, noise and signal intensity variation were pre-treated using peak alignment, binning and scaling, respectively. The sample-dependent effect of strongly diffracting minerals overwhelming the signal of weakly diffracting minerals was pre-treated using a square-root transformation. Without pre-treatment, the 60 XRPD measurements failed to provide informative clusters. Pre-treatment via peak alignment, square-root transformation, and scaling each resulted in significantly improved partitioning of the groups (p < 0.05). Data pre-treatment via binning reduced the computational demands of cluster analysis, but did not significantly affect the partitioning (p > 0.1). Applying all four pre-treatments proved to be the most suitable protocol for both non-hierarchical and hierarchical cluster analysis. Deducing such a protocol is considered a prerequisite to the wider application of cluster analysis in exploring soil property – soil mineralogy relationships in larger datasets.
in the intensity of either individual mineral peaks or of the full patterns (i.e. all peaks) of each mineral to its concentration in the sample, usually expressed in weight %. Such quantitative analyses are most often made on multiple samples, so that variation in mineral abundance can be compared from one soil sample to another and ultimately interpreted in relation to soil properties and functions that are dependent on soil mineralogy. Again the procedures that may be used to perform quantitative mineralogical analyses of soils rely heavily on the experience of the analyst to ensure that the results obtained are reliable and fit for purpose (Raven and Self, 2017).

The data of modern XRPD methods which all these procedures require is a precisely measured digital diffraction pattern typically comprised of discrete ‘Bragg’ diffraction peaks varying in intensity (y), expressed for example in counts, along an experimental axis (x) usually expressed in degrees 2θ. The ‘Bragg’ peaks from crystalline mineral phases rise above a background, which may also include diffuse scattering from X-ray amorphous phases. For example organic matter and volcanic glass can be common amorphous phases in many soils (Dixon and Schulze, 2002).

In recent years the availability of digital XRPD soil data has increased, and attempts are now being made to generate datasets containing thousands of spatially referenced XRPD measurements (e.g. those collected for the National Soil Inventory of Scotland (NSIS) and the Africa Soil Information Service (AfSIS)). Since many soil properties are closely related to soil mineralogy (Butler et al., 2018; Newman, 1984), such datasets in combination with computational data analysis represent unique opportunities to advance the understanding of the role of soil minerals in governing or influencing many soil properties, processes and functions.

In data-rich cases like NSIS (Butler et al., 2018) and AfSIS (Towett et al., 2015), computational methods of XRPD data analysis become particularly attractive because they are time-efficient and do not necessarily require any classical expert interpretation of the XRPD patterns, at least not until the final stages of such an analysis. This may seem like the analysis is initially disconnected from the crystallographic origins of the data, but this is the foundation of data-driven, or digital, approaches to soil mineralogy (Butler et al., 2018; Hillier and Butler, 2018). Cluster analysis is one such ‘digital’ approach that can aid the
interpretation of large XRPD datasets by classifying the data into discrete clusters that can be more manageably interpreted and described in their groups (Barr et al., 2004a, 2004b, 2004c, 2004d; Gilmore et al., 2004). Many clustering methods exist, including the fuzzy-c-means algorithm (Bezdek et al., 1984) that has been considered particularly appropriate for the heterogeneous nature of soils (Viscarra Rossel et al., 2016), and hierarchical clustering that is currently implemented in the DIFFRAC.EVA software (Bruker) for XRPD data analysis. For soil XRPD data, a rational cluster should be comprised of soils with similar mineralogy (i.e. minimised intra-group variation), whilst separate clusters should reflect distinct soil mineralogies (i.e. maximised inter-group variation).

In effect, the application of approaches such as cluster analysis to digital XRPD data collected from soil samples allows the diffraction patterns to be treated simply as digital signatures which encode information on soil mineral (peaks) and amorphous components (background). Subsequent to cluster analysis, identification and quantification may then be more efficiently introduced to understand how differences in soil mineralogy have influenced the clusters observed and how they relate to soil property data.

1.1. Pre-treatment of XRPD data for cluster analysis

Measurement reproducibility, i.e. precision, plays a critical role in facilitating effective cluster analysis. It is therefore crucial to appreciate that the precision of XRPD measurements may be commonly affected by several experimental and specimen related factors that bear no relation to compositional differences between samples (Jenkins, 1989). Such factors may alter peak positions, relative and absolute intensities, and signal to noise ratios. Some of these variations may be further emphasised and potentially become more problematical when samples are prepared by multiple personnel, as is often the case in the collection of larger datasets. The aim of data pre-treatment prior to cluster analysis is to minimise these experimental and specimen related factors and promote computation of suitable clusters. Here we evaluate how peak alignment, noise, sample composition, and count intensities can affect XRPD data (Fig. 1).

Peak positions in XRPD data commonly shift in response to small variations in specimen height in the instrument, the so called 'specimen displacement error'. An example of this error is illustrated in Fig. 1a. Even seemingly small misalignments between peaks can hinder the comparison of XRPD data by multivariate methods such as principal component analysis (Andersson and Hämäläinen, 1994; Arneberg et al., 2007), and hence reduce the effectiveness of cluster analysis. One approach to deal with such peak shifts is to use a mineral with essentially invariant peak positions as an internal standard (e.g. the common mineral quartz), resulting in well aligned data (Fig. 1b).

Noise in XRPD data relates to small fluctuations in response and detection, and to some extent is linked to the signal intensity and counting statistics (i.e. signal to noise ratios). Many approaches are available to reduce noise in XRPD data including moving averages and Savitsky-Golay filtering (Savitsky and Golay, 1964). Alternatively, binning data (combining a fixed number of adjacent data points into one combined variable) can be used to reduce noise whilst also acting as a method of data reduction that can be advantageous for computationally intensive exercises (Wehrens, 2011). A suitable bin width should be chosen that reduces noise but retains sufficient information within the peak shapes and intensities (Fig. 1c).

When based on principal component scores (Viscarra Rossel et al., 2016), cluster analysis of soil XRPD data has potential to be influenced by the signal of 'strong' diffraction factors, which is dependent upon sample mineralogy. Examples of strong diffraction factors include the omnipresent soil mineral quartz, which can overwhelm the less intense signal of weaker diffraction factors (e.g. feldspars and clay minerals) or minor but potentially important soil minerals (e.g. iron oxides). Because the information value related to what may appear as minor peaks can exceed that associated with intense peaks, by applying logarithmic or nth root transformations, skewness of the data can be reduced as the intensity of the strongest peaks relative to weaker peaks is rescaled. As outlined in Arneberg et al. (2007), logarithmic transforms modify linear correlations within data, whereas nth root transforms preserve any perfect linear correlations but reduce correlations in regions of partial correlation. Pre-treatment by nth root transform is therefore considered here to be most suitable for XRPD data (Fig. 1d).

Lastly, variations in signal intensities can also affect the outcome of cluster analysis. Such variations can be caused by several instrumental factors, including X-ray tube degradation, power settings, and choice of scan-time. Differences in signal intensities due to such factors may be scaled to comparable values by scaling methods such as mean centering (Wehrens, 2011; Viscarra Rossel et al., 2006), which subtracts the mean from all values in a sample vector of XRPD count intensities (Fig. 1e).

Further to the four factors outlined above, preferred orientation, or specimen texture, can be particularly troublesome in XRPD data since it alters the relative intensity of a mineral's Bragg peaks. Preferred orientation effects can be especially pronounced for minerals with prominent cleavage planes, which would include most clay minerals, feldspars, and carbonates for example. It can also be a very operator dependent issue, since it occurs during the loading of the specimen into its holder. Whilst the effects of preferred orientation cannot be accounted for when using data pre-treatment, it can be eliminated during sample preparation using methods such as spray drying (Hillier, 1999) but in cases where this method is not available the specimen preparation protocol must be demonstrably reproducible (Zhang et al., 2003).

The aim of this present study was to assess the requirement for and effectiveness of data pre-treatment methods for minimising the influence of the experimental and specimen aberrations on soil XRPD cluster analysis. Here, a 2^4 factorial design was used to assess the effects of four data pre-treatment protocols (alignment, binning, square-root transformation, and scaling) on the clustering of soil XRPD data primarily using the fuzzy-c-means algorithm. The factorial design employed all possible combinations of the four pre-treatments, allowing examination of individual effects as well as potential interactions between them. Ultimately the investigation is used to outline the most appropriate data pre-treatment protocol for soil XRPD cluster analysis using a combination of objective measures for group variance, number of clusters and cluster validity. The study is viewed as a necessary prelude to the application of cluster analysis to large datasets of soil XRPD patterns, such as NSIS and ASIS, and to exploiting their value in the increasingly data-driven field of digital soil analysis.

2. Materials and methods

2.1. Soil samples

Soils were sourced from 10 of the 60 ASIS ‘sentinel’ sites spread across Sub-Saharan Africa (Vågen et al., 2010). Each sentinel site covers 10 x 10 km, divided into 16 grid cells of equal size. Each grid cell contains 10 randomly selected plots. One composite topsoil (0–20 cm) sample from each of the 16 grid cells at each site was selected for analysis, yielding a dataset of 160 samples measured by XRPD (see Section 2.2). For this investigation, 12 soils were then selected from this dataset (n = 160) using the Kennard Stone algorithm (applied to untreated XRPD measurements) to provide a range of minerallogically distinct diffractograms (Kennard and Stone, 1969).

Although the mineralogical composition of the samples is not considered explicitly in the results of this study, it is useful to mention that the samples cover a wide range of mineralogical compositions that are commonly observed in the complete ASIS data set (Towett et al., 2015). Specifically, phase identification using the Powder Diffraction File database (ICDD, 2016) and subsequent quantification by a full pattern summation method (Omotoso et al., 2006) indicates that the minerals/mineral groups encountered across the set of twelve samples...
and estimates of their range of relative abundance in weight percent include: quartz (13–93), feldspars (0–55), amphibole (0–3), epidote (0–12), ilmenite (0–15), goethite (0–15), hematite (0–8), gibbsite (0–13), mica/illite (0–13), kaolin (3–64), and expandable/mixed-layer (0–40) clay minerals.

2.2. X-ray powder diffraction

Subsampling for XRPD analysis was conducted by coning and quartering. Subsamples were then prepared for XRPD by McCrone milling 3 g of sieved (< 2 mm) and air-dried soil for 12 min in ethanol. Excess ethanol was removed by centrifugation and each sample was suspended in 1.5 ml of hexane. The samples were then oven dried at 80°C before being disaggregated and ground by hand in a mortar and pestle before passing through a 250 μm sieve. Each sample (Section 2.1) was loaded onto the instrument by five different personnel, and analysed, one of these runs being the original AFIS data collection. Loading was carried out by loosely filling the sample holders with the finely ground powders, before flattening the surface with the sharp edge of a razor blade, with personnel instructed to take care to apply minimum pressure and to avoid shearing motion. The combination of methodical milling and loading was designed to produce samples with appropriate particle statistics and to minimise preferred orientation (Zhang et al., 2003) since the most effective procedure of spray drying the samples (Hillier, 1999) was not available in the AFIS lab.

After loading, XRPD data were collected on a Bruker desktop D2 PHASER diffractometer, with Ni-filter, Cu-Kα radiation with the X-ray tube operated at 30 kV and 10 mA. The beam was collimated using a 0.6 mm divergence slit, a 1 mm anti-scatter slit and a 2.5 mm Soller slit. Samples were rotated continuously at 15 rpm during data collection over the angular range of 3 to 75°2θ, counting for 96 s per 0.02° step with a Lynxeye position sensitive detector.

2.3. Pre-treatments and factorial design

A 2^4 factorial design (based on that used in Arneberg et al., 2007 for pre-treatment of mass spectrometry data) was performed on the dataset of 60 diffractograms (12 soils, each measured 5 times). The 4 pre-treatments were selected to correct for 2θ shifts (alignment), noise/data reduction (binning), overwhelming contributions from strong diffractors [using the square-root of the intensities, Kourkoumelis, 2013] and count intensity variation (scaling). The design allowed examination of individual pre-treatment effects on subsequent analysis (Section 2.5) as well as potential interactions between them. The 16 runs of the factorial design encompassed peak alignment (+1) versus no peak alignment (−1), binning (+1) versus no binning (−1), square-root transform (+1) versus no square-root transform (−1), and scaling (+1) versus no scaling (−1). The 16 factorial combinations are provided in Table 1. The same 16 combinations were also performed with data that had been subjected to de-noising using a discrete wavelet transform (Nason, 2016), but no noticeable effect on group variance or clumping was observed, and the results are therefore not presented.

Like almost all soils, all samples contained some quartz, therefore the quartz peaks in each diffractogram were used as an internal standard for an empirical peak alignment. The 2θ corrections were approximated by maximising the correlation between each sample and a ‘standard quartz’ pattern [PDF 00-046-1045, ICDD, 2016] by adjusting the 2θ scale of the sample, before converting all data to the same 2θ scale using linear interpolation. This procedure empirically corrects for sample displacement and zero error simultaneously, though its primary purpose here is to ensure that peak positions between samples are internally consistent. For binning, a constant width of 5 points was used. Square-root transformations were applied to the vector of count intensities. Mean centering (Viscarra Rossel et al., 2006) was used as the method of scaling, which subtracts the mean from all values in a sample vector of XRPD count intensities. The order with which these pre-treatments were applied is reflected in the column order of Table 1.

2.4. Spearman and Pearson correlation coefficients

As illustrated by Gilmore et al. (2004), the Spearman correlation coefficient (rs) and Pearson correlation coefficient (rp) are sensitive to different features of XRPD data. The Pearson coefficient can be sensitive to strong Bragg peaks; with intense, aligned peaks resulting in high correlation. In contrast, the Spearman coefficient is less sensitive to major peaks, and high correlation coefficients can still arise from misaligned data. Combining these coefficients with equal weighting is therefore considered to provide a more robust measure of correlation. To combine two or more correlation coefficients, their non-linear attributes need to be accounted for (Alexander, 1990). A combined correlation matrix was therefore derived by applying a Fisher transformation to rs and rp (i.e. their inverse hyperbolic tangents), taking the mean of the two Fisher transformed coefficients, and then back-transforming the averaged value (i.e. the hyperbolic tangent). All correlation coefficients described hereafter relate to data subjected to this averaging procedure. The average correlation coefficients, r, were computed only for untreated data, and are thus denoted as rer untreated hereafter. Whilst not specifically used in the fuzzy-c-means clustering presented here, their subsequent use is intended as a simple parameter that can summarise the similarity of samples prior to pre-treatment, and aid in interpretation of the results.

2.5. Multivariate analysis

All methods of data analysis described hereafter were performed using the R software environment (R Core Team, 2017) unless otherwise stated. Principal component analysis (PCA) was the first multivariate method to be applied to the treated data to derive scores for the first three principal components, which together described 95.1% of total variance on average. The three PCA scores were then used in a multivariate analysis of variance (MANOVA) in order to derive an F-statistic from Wilks’ lambda. The F-statistic is computed by division of the inter-group variance by the intra-group variance, therefore the objective of XRPD pre-treatment presented here is to maximise this parameter.

Non-hierarchical cluster analysis was employed using the fuzzy-c-means algorithm (Bezdek et al., 1984; Meyer et al., 2017), which was recently applied to soil spectral data since the fuzzy approach provides information on class overlaps that are inevitable in the case of soil data (Viscarra Rossel et al., 2016). Further, the partition coefficient of the fuzzy-c-means clustering statistics can be used as an objective measure to define the most suitable number of clusters (Bezdek et al., 1984; Viscarra Rossel et al., 2016). The first three PCA scores were used as inputs for fuzzy-c-means clustering, and the degree of fuzzification set to the default of 2 throughout (Meyer et al., 2017). The optimum number of clusters for each factorial treatment was derived by applying the algorithm to 11 model iterations with cluster nodes (nchauf) ranging from 2 to 12, and selecting the instance with the highest partition coefficient from the clustering statistics (Bezdek et al., 1984; Viscarra Rossel et al., 2016). As a measure of cluster validity, the adjusted Rand index (ARI, Fraley et al., 2017) was calculated for the vector containing the classification derived from fuzzy-c-means clustering relative to the groups of the original data (i.e. 12 clusters, each comprised of 5 replicates). An adjusted Rand index of 1 would represent perfect agreement between the two vectors. As will be described below, the objective measures of F, ARI and nchauf were used to provide recommendations for the most appropriate data pre-treatments for cluster analysis. To ensure that these recommendations were also relevant to hierarchical clustering methods, the untreated and pre-treated (i.e. all pre-treatments) data were subjected to cluster analysis using DIFFRAC.EVA software (Bruker), and the resulting dendrograms visually inspected. The implementation of
Hierarchical cluster analysis within DIFFRACT.EVA is based on that of PolySNAP (Barr et al., 2004a, 2004b, 2004c, 2004d; Gilmore et al., 2004), and uses a correlation matrix of all XRPD samples (full patterns, as outlined in Section 2.4) to compute dendrograms based on dissimilarity, d, where \( d = 1 - r \).

### 2.6. Re-analysis of Bondigui samples

As will be discussed below (Section 3), two of the Bondigui measurements from the original AFIS archive were identified as having different count intensities and peak to background ratios to the remaining measurements in their respective sets, which affected their clustering within the dataset. These samples, belonging to the Bondigui-A and Bondigui-B sample sets, were therefore retrieved during the course of this investigation from the AFIS soil archive and re-analysed for a sixth time using procedures described above (Section 2.2). The samples identified as having different count intensities and peak to background ratios were then substituted with the re-analysed data, and subjected to cluster analysis using the F16 pre-treatment combination (i.e. all pre-treatments; 'F16-rerun', Table 1).

### 3. Results and discussion

Providing all samples in the set of 12 presented here are mineralogically distinct from one-another (and hence produce different diffractograms), the XRPD data would be ideally expected to cluster into 12 groups, because the data are simply re-run samples of the same 12 soils. Within the selected set of 12 soils there are, however, notable similarities between some samples that can be gleaned from interpretation of the mean diffractograms (Fig. 2). To further aid in understanding this mineralogical similarity, a correlation matrix of the untreated data is also provided (Fig. 3). Together, Figs. 2 and 3 highlight notable correlation for several of the selected soils. More specifically, the \( r_{\text{untreated}} \) between Bimbe-B and Bondigui-A is 1.00; Didy-A, Didy-B and Didy-D exhibit an average \( r_{\text{untreated}} \) of 0.96; and Didy-D and Bimbe-A display an \( r_{\text{untreated}} \) of 0.97.

The F-statistic displays substantial changes across the 16 pre-treatment combinations tested here (Table 1). To assess the overall effects that pre-treatments had on F, the treatment assignments (~1 or +1) were regressed against F using multiple linear regression (\( R^2 = 0.69 \)), and resulting coefficients for each pre-treatment extracted. All pre-treatments were found to increase F relative to the untreated data (Fig. 4). Binning displayed the lowest effect on F, but nevertheless its contribution remained positive (i.e. increased group variance). Although the effect of binning was not significant (p > 0.1), the data reduction (in this case an 80% reduction in numeric length) associated with this pre-treatment would be computationally advantageous whilst retaining the outcome of analysis in this case. Of the three remaining pre-treatments (alignment, square-root and scaling), all displayed a significant positive influence on F (p < 0.05; Fig. 4). The combination of peak alignment and square-root transformation proved to be particularly effective at increasing F, indicating a synergistic interaction between these pre-treatments. This highlights how small variations in peak alignment can affect inter- and intra-group variance, and that a square-root transformation is a powerful method of rescaling information relating to weakly diffracting minerals (relative to quartz in this case). When both applied, the peak alignment and square-root transformation resulted in a near 4- to 5-fold increase in F compared to when individually applied (with and without scaling). Lastly, scaling was also found to have a strong positive increase on F, resulting in a near 3-fold increase in this statistic across the pre-treatment factors compared to when it was not applied.

The number of clusters (\( n_{\text{clust}} \), selected from an objective measure (partition coefficient, Section 2.5), was found to be particularly dependent upon pre-treatment. No single pre-treatment was sufficient to result in \( n_{\text{clust}} \) exceeding 2. Notably, as expected from the synergistic effects on F, the combination of alignment and square-root transformation proved to be key for forming a more appropriate number of clusters in this case. It was only in cases where alignment and square-root transformation were applied together that \( n_{\text{clust}} \) exceeded 2 (increasing to 11 or 12; see F6, F8, F14 and F16, Table 1). Factors 6 and 14 (\( n_{\text{clust}} = 14 \)) represent the same pre-treatments except for scaling, as do factors 8 and 16 (\( n_{\text{clust}} = 12 \)). Although these pairs produce the same number of clusters, the ARI further highlights how scaling improves the clustering of this dataset, increasing by 0.16 in both cases.

The clustering of this data by fuzzy-c-means can be visually inspected by plotting the 3 PCA scores against one-another. For comparison, the PCA scores of F1 (no pre-treatment), F8 (alignment, binning and square-root transformation) and F16 (all pre-treatments) are provided in Fig. 5. It is evident that untreated data (i.e. F1) are difficult to classify and only two clusters (Fig. 5, top row) are observed, separated largely by the PCI score. From this it is clearly evident that sample-independent variation of untreated data is resulting in a spread of the measurements in principal component space. In contrast to F1, the pre-treatments associated with F8 result in a marked improvement in clustering (Fig. 5, middle row). Clusters begin...
to become distinguishable from one-another through a reduction in intra-group variation (largely due to peak alignment) combined with enhanced inter-group variation (resulting from the square root transformation). Nonetheless, despite improvements, the 12 clusters described by F8 display several inconsistencies. These first relate to clusters 6 and 12 (Fig. 5, middle row), which can be described as singletons (clusters of one measurement) and is often an indication that manual inspection is required. Cluster 6, a Moiben soil measurement (Fig. 2), should be classified within group 1 (associated with the remaining Moiben measurements), whilst cluster 12, a Bondigui-A measurement (Fig. 2), should be classified within cluster 11 (comprised of the Bondigui-A measurements as well as all Bimbe-B measurements). In addition to the two singular outliers there are further misclassifications with the F8 clustering output. First, cluster 1 contains a Bondigui-B soil sample (Fig. 2) that would ideally be clustered within cluster 2 (comprised primarily of Bondigui-B samples). As it stands, cluster 2 also contains two misclassifications - the first relating to a Lambussie sample that should belong to cluster 7, and the second relating to a Didy-C sample that should belong to cluster 3 (Fig. 2). Lastly in relation to F8 clustering, it is worth noting how Didy-A, and -B samples (Fig. 2) are grouped together within cluster 5 (Fig. 5, middle row), which reflects the relatively high correlation of their untreated measurements (runtreated = 0.95; Fig. 3).

In terms of F and ARI, F16 displays the best performance of all pre-treatment combinations. This performance can also be observed from inspection of the PCA score plots and corresponding clustering (Fig. 5, bottom row): The only pre-treatment difference between F16 and F8 (described above) is the application of scaling. The result of this scaling is a notably increased F (Table 1) due to variances in signal intensity (predominantly resulting from the instrumental operation and collimation parameters) being minimised, which then facilitates improved fuzzy-c-means clustering (Fig. 5, bottom row). As suggested by the ARI value, F16’s clustering is more appropriate than that provided by F8. Nonetheless two misclassification’s still remain. The first relates to cluster 11, which is a sole Bondigui-A outlier that should be classified within cluster 8, where the remaining Bondigui-A measurements are clustered along with all Bimbe-B measurements. The classification of Bondigui-A and Bimbe-B into the same cluster is unsurprising given their similarity (runtreated = 1.00, Figs. 2 and 3). The only other misclassification from F16’s clustering relates to a Bondigui-B measurement classified with cluster 2, which should belong to cluster 12 which consists of the remaining Bondigui-B measurements. It is worth noting here that despite the Didy-A and -B samples existing within the same cluster from F8’s clustering (described above), the added pre-treatment of scaling is sufficient to separate the measurements from these two soils into two clusters (clusters 7 and 5, respectively; Fig. 5, bottom row). In summary, when excluding the two misclassification’s from F16’s clustering, the resulting output of 11 clusters, which includes the grouping of two soils with very similar mineralogies, can be considered an effective application of soil XRPD cluster analysis.

As mentioned above (Section 1.1), the effects of preferred orientation from minerals with good cleavage can be a primary factor limiting the reproducibility of XRPD measurements. None of the 4 pre-treatments tested here were designed to correct for preferred orientation.
and instead care taken in sample preparation and loading by a procedure similar to that described by Zhang et al. (2003) was used to minimise the tendency for measurements to be affected by such variation. The fact that only 2 samples were misclassified when all 4 pretreatments were applied, combined with the conclusion that this misclassification was caused by count intensities, indicates that measurement variation caused by preferred orientation, and its potential variability due to different operators performing the sample loading into the XRPD instrument, was sufficiently constrained.

It is also useful to further scrutinise the two misclassified samples that remained in F16’s clustering. These relate to the Bondigui-A and Bondigui-B sets, the diffractograms of which are plotted in Fig. 6 with all pre-treatments applied. To the trained expert eye, the diffractograms within each group are clearly representative of the same sample, however the multivariate analysis applied here is more sensitive to subtle variations distributed throughout the diffractograms. In both

![Correlation matrix](image)

**Fig. 3.** Correlation matrix ($r_{untreated}$) of the 12 selected soil samples calculated on the mean diffractograms of untreated data. Correlations represent Fisher-transformed means of Spearman and Pearson correlation coefficients (see Section 2.4). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

![Multiple regression coefficients](image)

**Fig. 4.** Multiple regression coefficients used as a measure of pre-treatment effects on $F$ (Section 2.5).
cases it was found that the misclassified measurements are identifiable by small but significant variations in peaks heights and background intensities relative to the remaining 4 measurements. The source of this variation can be traced back to the untreated data, and their corresponding peak and background intensities. For the Bondigui-A sample, the 4 correctly classified measurements displayed similar count intensities (maximum intensity = 44,265 ± 3942) whilst that of the misclassified measurement was approximately half (maximum = 19,584). The same goes for the Bondigui-B sample, with the correctly classified measurements having maximum count intensities of 4789 ± 217, and the misclassified measurement a maximum of 2321. This misclassification in both of these cases therefore relates to count intensities and peak to background ratios, which cannot be completely accounted for using the pre-treatment steps applied here. Though the exact cause of these variations for the two Bondigui samples could not be identified, it is likely due to changes in instrument operation or collimation parameters. The finding is able to highlight the importance of ensuring comparable count intensities and peak to background ratios in soil XRPD datasets if effective cluster analysis is to be applied, particularly in the case of large datasets where assembly can span months-years.

3.1. Clustering of re-analysed data

After substituting the Bondigui-A and Bondigui-B samples, identified as having low count intensities and different peak to background ratios, with re-analysed data (Section 2.6), clustering of the dataset was notably improved. The improvement is reflected in the F-statistic, which showed a near 3-fold increase compared to the F16 results containing the original data, and also in the ARI, which increased to 0.90 (Table 1). Fuzzy clustering of the updated dataset yielded 11 clusters (Fig. 7). Of these clusters, 10 represented groups containing 5 correctly classified measurements, whilst the remaining cluster (cluster 1) contained 10 measurements. Cluster 1 combined all measurements from the Bondigui-A and Bimbe-B soils, which again is appropriate given the similarity of the untreated data reflected in both Fig. 2 and their high correlation (untreated = 1.00; Fig. 3). As a whole, the cluster analysis applied to re-analysed data produced groups that agree with what may be reasonably expected for the given dataset. This re-analysis again highlights how care taken to facilitate long-term instrument reproducibility is also a key step in ensuring effective cluster analysis of soil XRPD data.

3.2. Application of pre-treatments to hierarchical cluster analysis

The most effective pre-treatments have so far been discussed and refined in relation to cluster analysis via the fuzzy-c-means algorithm. To test whether these pre-treatments are applicable to other cluster analysis methods, the hierarchical cluster analysis implemented in DIFFRAC.EVA software (Section 2.5) was applied to the untreated and pre-treated (all pre-treatments, i.e. F16) datasets using the group average link method, which in both cases included the re-analysed Bondigui-A and Bondigui-B samples.

Selecting the appropriate number of clusters from the hierarchical cluster analysis dendrogram of untreated data (Fig. 8a) proved difficult. Setting the cut-line to a dissimilarity (d) of 0.045 yielded 31 clusters, 22 of which were singletons, and only two of which displayed appropriate grouping of the 5 replicates. Various cut-lines were tested on the untreated data, but none provided a suitable classification of the data into the clusters that would be expected of this dataset. In contrast, the dendrogram of the pre-treated data (Fig. 8b) displays much improved clustering. Setting the cut-line again to a dissimilarity of 0.045 yielded
Fig. 6. The 5 XRPD measurements of Bondigui-A (top) and Bondigui B (bottom) with all pre-treatment steps applied (i.e. F16). The misclassified measurement from each group (according to F16 clustering) is plotted in black, with the remaining 4 measurements in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 7. The first 3 PCA scores plotted against one-another with the identified Bondigui-A and -B samples removed and replaced with re-analysed data. Data were subject to all pre-treatments (i.e. F16). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
13 clusters, of which 3 were singletons and the remaining were classified appropriately. Consistent with their low dissimilarity (Fig. 8b), the Didy-B and Didy-D measurements are grouped together ($d = 0.024$), as are the Bimbe-B and Bondigui-A measurements ($d = 0.033$). Upon closer inspection of the 3 singletons, each of them are close to being correctly classified with their correct clusters (Bondigui-B, Lambussie and Didy-C), and it again seems that very subtle variations in count intensities and peak to background ratios may be causing their dissimilarities to be slightly higher than the remaining 4 measurements in each set. Nevertheless the dendrograms of untreated and pre-treated data illustrate that pre-treatment by alignment, binning, square-root transformation and scaling is highly beneficial in hierarchical cluster analysis of soil XRPD data.

3.3. Limitations of soil XRPD cluster analysis

The sensitivity of cluster analysis to several aspects of XRPD data presented here illustrates that reproducible measurements are a key component in deriving informative clusters. A clear limitation of the approach, even with pre-treatment of the data, is that substantial
variation in count intensities and peak to background ratios can detrimentally affect the clustering. Given these sensitivities it seems likely that cluster analysis of XRPD data from multiple soil datasets, where different preparation and measurement protocols are implemented on various instruments, would currently prove challenging. This therefore limits application of soil XRPD cluster analysis at present to datasets measured in a single laboratory on a single instrument. If the approach is to be applied to larger international collaborations of soil XRPD data [e.g. Viscarra Rossel et al., 2016], methods of data harmonisation will need to be carefully considered. Nevertheless, there is promising scope for the application of cluster analysis to individual soil XRPD datasets such as NSIS (Butler et al., 2018) and AFSIS (Towett et al., 2015).

4. Conclusions

A 2⁴ factorial design to assess the effects of four diffractogram pre-treatments has shown that peak alignment, square-root transformation, and scaling each significantly enhance the effectiveness of fuzzy-c-means clustering of soil XRPD data. Further, binning the data can reduce the computational demands substantially whilst not significantly effecting the clustering, and would therefore be a useful addition for analysis of larger XRPD datasets. The combination of peak alignment and square-root transformation pre-treatments was found to have synergistic effects on the F-statistic. This can be interpreted as the peak alignment acting to reduce sample-independent variation, whilst the square-root transformation acts to rescale the signal of strong diffractors (in this case quartz) relative to that of minerals with weaker diffraction intensities, thus emphasising more subtle variations in mineralogical compositions of the soils. When combined, the four factors tested here resulted in accurate classification of the soil dataset into 12 clusters (using fuzzy-c-means), but two misclassifications remained that resulted from variations in count intensities and peak to background ratios relative to the remaining measurements in each set, and were presumably related to measurement problems not identified at the time of original AFIS data collection. Upon re-analysis of these samples, the dataset clustered into 11 groups with completely appropriate classifications given that two of the soils investigated here display near perfect correlation even in the untreated data. Together this highlights how appropriate data pre-treatment combined with care taken to facilitate reproducible XRPD measurements can provide an effective basis for cluster analysis of soil XRPD data. A thorough understanding of factors that include sample-independent variation of soil XRPD data, the masking effects of strong diffractors, and data quality issues - all of which may influence the results of methods such as cluster analysis - is a prerequisite to the wider and hopefully profitable application of these techniques for exploration of soil property – soil mineralogy relationships. Ultimately the pre-treatments of alignment, binning, square-root transformation and scaling together result in significantly improved cluster analysis of XRPD data via both the fuzzy-c-means clustering algorithm and hierarchical cluster analysis.

Acknowledgements

This work was supported by a Macaulay Development Trust Fellowship, United Kingdom, grant number MDT-50. The XRPD analysis was financially supported by the World Agroforestry Centre (ICRAF) through the Africa Soil Information Service (AFSIS) project, funded by the Bill and Melinda Gates Foundation, United States, grant number 51353 and Alliance for a Green Revolution in Africa (AGRA). The authors acknowledge the regional AFSIS field crews who provided the soil samples, Beatrice Owaria and Bella Kauma for assistance with XRPD analysis, as well as other technical staff of the ICRAF Soil-Plant Spectral Diagnostics Laboratory for assistance with sample preparation. We also thank Altea Lorenzo-Arribas for reviewing the manuscript prior to submission, and the two anonymous reviewers whose constructive comments helped improve this paper. The support of the Scottish Government’s Rural and Environment Science and Analytical Services Division (RESAS) is also gratefully acknowledged.

References

Alexander, R.A., 1990. A note on averaging correlations. Bull. Psychon. Soc. 28 (4), 335–336.
Andersson, R., Hämäläinen, M.D., 1994. Simplex focusing of retention times and latent variable projections of chromatographic profiles. Chemom. Intell. Lab. Syst. 22 (1), 49–61.
Arneberg, R., Rajalabati, T., Filika, K., Berven, P.S., Kroksveen, A.C., Berle, M., Myhr, K.M., Vedeler, C.A., Utlvik, R.J., Kvalheim, O.M., 2007. Pretreatment of mass spectral profiles: application to genomics data. Anal. Chem. 79 (18), 7014–7026.
Barr, G., Dong, W., Gilmore, C., Faber, J., 2004a. High-throughput powder diffraction. III. The application of full-profile pattern matching and multivariate statistical analysis to round-robin type data sets. J. Appl. Crystallogr. 37 (4), 635–642.
Barr, G., Dong, W., Gilmore, C.J., 2004b. High-throughput powder diffraction. II. Applications of clustering methods and multivariate data analysis. J. Appl. Crystallogr. 37 (2), 243–252.
Barr, G., Dong, W., Gilmore, C.J., 2004c. High-throughput powder diffraction. IV. Cluster validation using silhouettes and fuzzy clustering. J. Appl. Crystallogr. 37 (6), 874–882.
Barr, G., Dong, W., Gilmore, C.J., 2004d. PolySnap: a computer program for analysing high-throughput powder diffraction data. J. Appl. Crystallogr. 37 (4), 658–664.
Beck, T.G., Ehlich, R., Full, W., 1984. FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. 10 (2–3), 191–201.
Bish, D.L., 1994. Quantitative X-ray diffraction analysis of soils. In: Amone, J.E., Zelazny, L.W. (Eds.), Quantitative Methods in Soil Mineralogy. Soil Science Society of America, Madison, WI, pp. 1–19.
Brindle, G., Brown, G., 1980. Crystal structures of clay minerals and their X-ray identification. In: Mineralogical Society of Great Britain and Ireland. vol. 5.
Butler, B.M., O’Rourke, S.M., Hillier, S., 2018. Use of rule-based regression models to predict and interpret soil properties from X-ray powder diffraction data. Geoderma 329, 43–53.
Dixon, J.B., Schulze, D.G., 2002. Soil Mineralogy With Environmental Applications. SSSA Book Series 7 Soil Science Society of America, Madison, WI.
Fraleys, C., Bayerley, A., Scrucca, L., Murphy, T.B., Pop, M., Scrucca, M.L., 2017. Package mlclust.
Gilmore, C.J., Barr, G., Paisley, J., 2004. High-throughput powder diffraction. I. A new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles. J. Appl. Crystallogr. 37 (2), 231–242.
Harriss, W., White, G.N., 2008. X-ray diffraction techniques for soil mineral identification. In: Methods of Soil Analysis, Part 5, pp. 81–115.
Hillier, S., 1999. Use of an air brush to spray dry samples for X-ray powder diffraction. Clay Miner. 34 (1), 127–135.
Hillier, S., Butler, B.M., 2018. New XRD data-based approaches to soil mineralogy. Spectroscopy 33 (7), 34–36.
ICDD, 2016. PDF-4+ [2016 (Database). International Center for Diffraction Data, Newtown Square, PA, USA.
Jenkins, R., 1989. Experimental procedures. In: Bish, D., Post, J. (Eds.), Modern Powder Diffraction. Mineralogical Society of America, pp. 47–72.
Kendall, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148.
Kourkoumelis, N., 2013. PowDLL: a reusable .net component for interconverting powder diffraction data. Recent developments. In: O’Neill, Lisa (Ed.), ICDD Annual Spring Meeting.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2017. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071). TU Wien URL: https://cran.r-project.org/package=e1071.
Nason, G., 2016. wavethresh: wavelets transforms and statistics. R package version 4.6.8.
URL: https://CRAN.R-project.org/package=wavethresh.
Newman, A.C.D., 1984. The significance of clays in agriculture and soils. Philos. Trans. R. Soc. Lond. A 311, 375–389.
Omontos, O., McCarty, D.K., Hillier, S., Kleeberg, R., 2006. Some successful approaches to predict and interpret soil properties from X-ray powder diffraction data. Geoderma 131 (1–2), 61–75.
Viscarra Rossel, R.A., Wolvaart, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1–2), 61–75.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Dematté, J.A.M., Shepherd, C.J., 2004. High-throughput powder diffraction. II. Applications of clustering methods and multivariate data analysis. J. Appl. Crystallogr. 37 (2), 243–252.
K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aichi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodzky, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.I.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudworth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth Sci. Rev. 155, 198–230.

Wehrens, R., 2011. Chemometrics With R. Multivariate Data Analysis in the Natural Sciences and Life Sciences. Springer Science and Business Media.

Zhang, G., Germaine, J.T., Martin, R.T., Whittle, A.J., 2003. A simple sample-mounting method for random powder X-ray diffraction. Clay Clay Miner. 51 (2), 218–225.