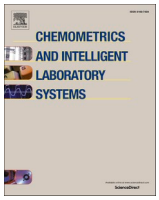
Chemometrics and Intelligent Laboratory Systems 252 (2024) 105174

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Exploratory analysis of hyperspectral imaging data

Alessandra Olarini [a,b,*], Marina Cocchi [b], Vincent Motto-Ros [c], Ludovic Duponchel [a], Cyril Ruckebusch [a]

[a] Université de Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000, Lille, France
[b] University of Modena and Reggio Emilia, Department of Chemical and Geological Sciences, Via Campi 103, 41125, Modena, Italy
[c] Université Claude Bernard Lyon 1, Institut Lumière Matière, CNRS, UMR 5306, Villeurbanne, 69622, France

ARTICLE INFO

Keywords:
Spectral imaging
Essential information
Clustering
Spectral unmixing
Raman
LIBS

ABSTRACT

Characterizing sample composition and visualizing the distribution of its chemical compounds is a prominent topic in various research and applied fields. Integrating spatial and spectral information, hyperspectral imaging (HSI) plays a pivotal role in this pursuit. While self-modelling curve resolution techniques, like multivariate curve resolution - alternating least squares (MCR-ALS), and clustering methods, such as K-means, are widely used for HSI data analysis, their effectiveness in complex scenarios, where the structure of the data deviates from the models' assumptions, deserves further investigation. The choice of a data analysis method is most often driven by research question at hand and prior knowledge of the sample. However, overlooking the structure of the investigated data, i.e. linearity, geometry, homogeneity, might lead to erroneous or biased results. Here, we propose an exploratory data analysis approach, based on the geometry of the data points cloud, to investigate the structure of HSI datasets and extract their main characteristics, providing insight into the results obtained by the above-mentioned methods. We employ the principle of essential information to extract archetype (most linearly dissimilar) spectra and archetype single-wavelength images. These spectra and images are then discussed and contrasted with MCR-ALS and K-means clustering results. Two datasets with varying characteristics and complexities were investigated: a powder mixture analyzed with Raman spectroscopy and a mineral sample analyzed with Laser Induced Breakdown Spectroscopy (LIBS). We show that the proposed approach enables to summarize the main characteristics of hyperspectral imaging data and provides a more accurate understanding of the results obtained by traditional data modelling methods, driving the choice of the most suitable one.

1. Introduction

Understanding the composition and distribution of the chemical compounds within a sample stands as a priority in many research and applicative fields [1,2]. In this respect, hyperspectral imaging (HSI) is a key analytical tool as it combines spatial information about the distribution of the chemicals across the image pixels with the corresponding spectral signatures. HSI finds applications throughout a wide range of scientific disciplines, spanning from remote sensing to macro- and micro-imaging [3–5]. The information provided by HSI is usually organized in a third-order tensor with two spatial dimensions and a spectral one. To identify individual sources of spectral variation and determine their respective contribution to the mixed signal in each pixel, self-modelling curve resolution techniques are among the most popular approaches [6–8]. One of the principal algorithms in this category is multivariate curve resolution - alternating least squares (MCR-ALS) [9]. Based on the matrix formulation of Beer-Lambert's law, the results of the data decomposition provided by MCR-ALS can be interpreted as concentration distribution maps (spatial distributions) and spectral signatures of the individual components of the spectral mixture. The MCR-ALS algorithm minimizes the difference between the reconstructed data and the original data, by iteratively optimizing the concentration profiles and the spectra profile in each least square iteration, until convergence is achieved. Constraints can be imposed on the components profiles to enforce physically or chemically meaningful solutions. Imposing constraint would also contribute to reduce rotational ambiguity which, except in very specific conditions, remains inevitable [10,11]. MCR-ALS initially found extensive application in spectroscopic data analysis, particularly in fields such as analytical chemistry and process analysis. Later, its utility expanded to include image analysis

* Corresponding author. Université de Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000, Lille, France.
E-mail address: alessandra.olarini@unimore.it (A. Olarini).

https://doi.org/10.1016/j.chemolab.2024.105174
Received 17 May 2024; Received in revised form 4 July 2024; Accepted 4 July 2024
Available online 9 July 2024

and a broad range of other applications [12–14]. However, in complex scenarios, the MCR bilinear decomposition may not fully capture the complexity of the physics/chemistry underlying the analyzed data, due to e.g. interactions between individual species. This and other effects might result in a deviation from the ideal linear mixture model [15].

Another approach for the analysis of HSI datasets can be found in the framework of clustering techniques, aiming at grouping pixels based on their spectral similarity (hierarchical and partitional clustering) [16] or on density criteria [17], and highlighting spatial patterns in the image. Unlike spectral unmixing which aim at identifying the contributions of the individual mixture components for each pixel, clustering methods, such as K-means [18] assigns each pixel to one cluster only, characterized by a centroid, serving as a prototype of the cluster, and results should be interpreted as spectral pixel classification or image segmentation approaches. The determination of the number of clusters, as well the algorithm initialization step's requiring a random selection of the mean spectrum of each cluster, represent the major challenges for this method. Even though solutions trying to overcome these issues, such as the use of indices and replicates, have been proposed over the years, these questions remain open in the field [19,20].

In practice, although the assumptions and goals of MCR and K-means approaches are different, both can provide complementary results when applied to HSI data, the first can be used for data decomposition and the second for clustering, to obtain interpretable factors or clusters albeit in different ways [21]. Clustering techniques have been used in conjunction with other multivariate analysis methods [22,23], for instance, as a powerful tool for examining data homogeneity, in terms of chemical composition or properties, together with Principal Component Analysis (PCA) [24,25]. Other works have explored the use of clustering as a constraint in unmixing methods such as MCR-ALS or vertex component analysis (VCA) for complex samples [26,27].

While MCR and clustering methods are powerful tools, the first step of any multivariate data analysis should be exploratory [28,29], summarizing the main characteristics of the investigated data set, and this is even more the case for HSI data. Users should be aware of the specific characteristics of the structure of the data and carefully consider the methods' assumptions and limitations in order to ensure the reliability of their interpretation. To this aim, the extraction of the essential information (EI) can reveal very useful as it is not based on data variance but on the geometry of the data points cloud [15,30–34]. Essential information consists of archetype points that outline the convex hull of the data points cloud in a normalized abstract data space. Recent studies have highlighted the potential and usefulness of identifying essential rows and columns of a data matrix [32,35–37]. A key aspect is that the corresponding samples (spectral pixels) and variables (single--wavelength images) contain all the information needed to reproduce the measured data [38].

This paper introduces an exploratory approach for analysing HSI data of complex samples, especially for scenarios where the results obtained from MCR-ALS and K-means are difficult to obtain and interpret. Through identification of the archetype points of the data cloud, we aim to extract some of the most linearly dissimilar spectra and single-wavelength images measured. A powder mixture analyzed with Raman spectroscopy and a mineral sample, characterized mostly by pyrite, analyzed by Laser Induced Breakdown Spectroscopy (LIBS) were investigated. The first dataset is a perfect example for using unmixing approach, the second is a dataset where both the unmixing and clustering approaches could be used considering the analysis task [39]. However, this dataset has an intrinsic complication due to the weathering products of pyrite in LIBS technique and the lack of selectivity, making the hyperspectral imaging data at hand deviating from the ideal model underlying both MCR and clustering techniques. Comparison to the result obtained by applying MCR-ALS [40,41] and K-means [18] is also provided and discussed. We argue that this approach is very useful to extract the main characteristics of a hyperspectral imaging dataset and provide accurate information to be used for spectral unmixing and clustering. Moreover, it has been observed that spatial distributions and spectral signatures extracted by this approach are not always retrievable using conventional methods like MCR-ALS and K-means.

2. Materials and methods

2.1. Datasets

2.1.1. Raman powder dataset

Powders of three salts i.e. calcium carbonate ($CaCO_3$), sodium nitrate ($NaNO_3$) and sodium sulfate (Na_2SO_4) were mixed and pressed in a tablet, obtaining a three-component system. Sample preparation and Raman imaging acquisition features were described by Coic et al. in Ref. [31]. The sample was investigated in the range 901.2 cm^{-1} to 1280.5 cm^{-1} with a spectral resolution of 1.11 cm^{-1}. A 101 × 101 pixels image was mapped using point-by-point raster-scanning mode with a 1 μm step between successive acquisitions. The dataset corresponds to a third-order tensor of dimensions 101 × 101 × 341, which was subsequently analyzed without any spectral pretreatment.

2.1.2. LIBS mineral dataset

A thin section of a mineral sample from the Nishapur turquoise deposit (Iran) was prepared and polished for LIBS imaging. Sample preparation, equipment and LIBS acquisition are detailed in Moncayo et al. in Ref. [42]. The sample is constituted by three main mineral phases: pyrite FeS_2, silica (mainly quartz) SiO_2 and turquoise $CuAl_6(PO_4)_4(OH)_8{\cdot}4H_2O$. The LIBS image was recorded considering a 15 μm step between successive acquisitions over 2048 spectral channels in the spectral range from 250 to 330 nm. From the full acquired dataset, only a region of interest has been selected, resulting in a third-order tensor of dimensions 300 × 300 × 1930, which was then analyzed without spectral preprocessing.

2.2. Data analysis

The data analysis methodologies employed in subsequent sections of the paper are here introduced. Section 2.2.1 provides a detailed description of the proposed data analysis approach, which investigates the geometry of the data point cloud resulting from a singular value decomposition (SVD). Sections 2.2.2 and 2.2.3 describe well-known chemometric methods: Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS) and K-means clustering, respectively. These methods are employed for data analysis to compare with the proposed approach.

2.2.1. Selection of the most relevant archetype points for exploratory analysis

The HSI tensor is first unfolded into a matrix **D** of dimensions (n, p) with rows corresponding to pixels and columns corresponding to spectral channels (unfolded single-wavelength images). The matrix **D** is then decomposed by SVD [43] according to Eq. (1):

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where **U** of dimensions (n, k) is the matrix containing the left singular vectors, **S** of dimensions (k, k) is the diagonal matrix of singular values and $\mathbf{V}^{\mathrm{T}}$ of dimensions (k, p) is the matrix of the right singular vectors transposed, k is the number of factors of the decomposition and **E** of dimensions (n, p) the error matrix. The matrices **X** and **Y** of dimensions (n, k) and (p, k) are calculated as in Eqs. (2) and (3), respectively, and contain the coordinates of the data points in the column- and row-vector space, respectively:

$$\mathbf{X} = \mathbf{U} \times \mathbf{S} \tag{2}$$

$$\mathbf{Y} = \mathbf{V} \times \mathbf{S} \tag{3}$$

All column vectors of **X** (resp. **Y**) are then normalized to constant projection on the first column vector of **X** (resp. **Y**) to enforce convexity of the data points cloud [44,45]. The archetypes of the data points cloud of **X** and **Y** can be identified by computing the corresponding convex hulls, as in Eqs. (4) and (5) [35]. They correspond to the most linearly dissimilar spectral pixels and single-wavelength images, respectively.

Convex hulls of matrices **X** and **Y** are computed:

$$conv(\mathbf{X}) = \left\{ \mathbf{x} \in \mathbf{X} \,|\, \begin{matrix} \sum \alpha \mathbf{x}; \alpha \geq 0 \\ and \sum \alpha = 1 \end{matrix} \right\} \tag{4}$$

$$conv(\mathbf{Y}) = \left\{ \mathbf{y} \in \mathbf{Y} \,\middle|\, \begin{matrix} \sum \beta \mathbf{y}; \beta \geq 0 \\ and \sum \beta = 1 \end{matrix} \right\} \tag{5}$$

where α and β are coefficients of the convex linear combinations. The number of components to consider into convex hull calculation is left to the user [44,45]. Analogous to exploratory PCA, inspection of the information carried by the most dissimilar spectra/images can guide the selection.

The most relevant archetype points are then selected by visual inspection and the corresponding (essential) spectra and (essential) single-wavelength images extracted, as illustrated in Fig. 1.

2.2.2. Multivariate curve resolution - alternating least squares (MCR-ALS)

The MCR-ALS algorithm provides pure spectral signature of the components and their corresponding component distribution maps, as in Eq. (6):

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{T} + \mathbf{E} \tag{6}$$

where **D** of dimensions (n, p) is the unfolded tensor, **C** (n, c) is the pure concentration matrix, with component distribution maps of the c components as columns, and $\mathbf{S}^{T}$ of dimension (c, p) is the matrix of pure spectra, with spectra profiles of the c components as rows. **E** (n, p) contains the variation unexplained by the MCR model. To solve Eq. (6), alternating least-squares (ALS) optimization [46] is used as well-established approach, and both concentration and spectra profiles are constrained with non-negativity. The first step of the optimization requires spectra or distribution profiles that will be implemented and optimized during the iteration process. Simple to use interactive self-modelling mixture analysis (SIMPLISMA) [47] was used throughout this work to calculate initial spectral estimates. The optimization procedure stops when the convergence criterium is reached, expressed as a threshold (0.1 %) based on the relative difference of the lack of fit (LOF) during consecutive iterations. The LOF and the explained variance, defined in Eqs. (7) and (8), are used as parameters to evaluate the quality of the MCR model:

$$\mathrm{LOF} = 100 \times \sqrt{\frac{\sum e^2}{\sum d^2}} \tag{7}$$

$$r^2 = 100 \times \left(1 - \frac{\sum e^2}{\sum d^2}\right) \tag{8}$$

where e and d are elements of **D** and **E** respectively.

2.2.3. K-means clustering

As one of the most used partitioning clustering techniques in image analysis, the K-means algorithm can be applied to **D**. In K-means, once the number of clusters is defined (c), the first iteration selects c clusters randomly, then at each iteration samples are reassigned to minimize the sum of point-to-centroid distances, summed over all c clusters (*sumd*). The algorithm stops when clusters assignments do not change, or the maximum number of iterations is reached. As distance measure, the Pearson correlation distance, defined as one minus the correlation coefficient calculated between the point and centroid spectra, has been used [48]. In order to stabilize the results, 50 replicate runs of K-means clustering are performed for each analysis and the run with lowest *sumd* has been selected; the number of iterations was set to 200. Silhouette [49] and Pakhira-Bandyopadhyay-Maulik (PBM) [50] indices were used to evaluate the optimal number of clusters. These were compared with the number of most informative pixels/spectral wavelengths suggested by the proposed exploratory approach. Here, the explicit use of c to denote both the number of components (MCR-ALS) and clusters (K-means), is adopted because the results, for sake of comparison, are presented considering the same number of components and clusters.

2.3. Software

All computations were performed using MATLAB© 2022a (MathWorks Massachusetts, USA). For the cluster analyses the K-means function of the Statistical and Machine Learning Toolbox was used, with the addition of the MATLAB© Parallel Computing Toolbox to improve

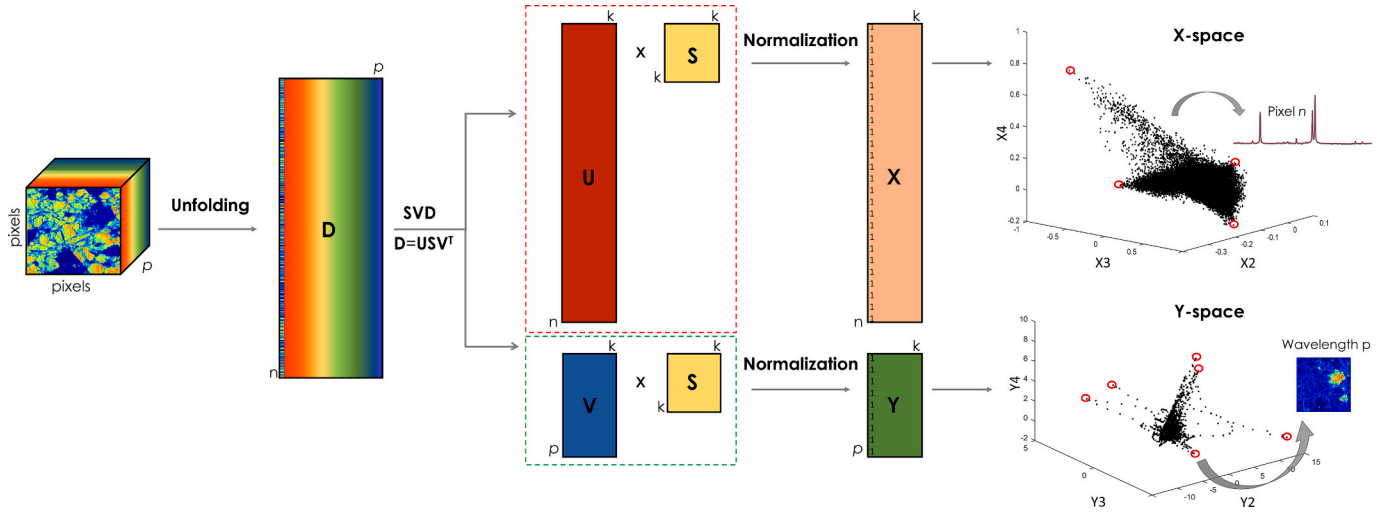


Fig. 1. Graphical representation of the data exploratory approach: third-order tensor is unfolded into a matrix (**D**) and decomposed through SVD algorithm (**D** = **USV**^{T}). Matrices **X** and **Y** are calculated and normalized [45], resulting in a unit first column vector **X1** (resp. **Y1**) to which all other column vectors of **X** (resp. **Y**) are orthogonal. Convex hulls of essential spectra and essential variables are computed for **X** and **Y**, and the most relevant archetype points are identified by visual inspection (red circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the speed of the algorithm. *Pure* and *als* routines from Tauler and De Juan (2003) were used for the MCR-ALS analysis and *convhulln* is the built in MATLAB© function used to compute convex hulls.

3. Results and discussion

For each dataset, the information extracted from exploratory analysis is put in perspective of the results obtained by applying both MCR-ALS and K-means.

3.1. Raman powder dataset

The three-component Raman hyperspectral imaging dataset, described in 2.1.1, exhibits well-defined characteristics: (i) a clear distribution of the salts building up the chemical composition of the sample (Fig. 2A), (ii) good signal-to-noise ratio data (Fig. 2B) and (iii) selective spectral regions (Fig. 2C).

Fig. 3A provides a representation of the (X2, X3) data points clouds, where the number 2 and 3 refer to the second and third column of the normalized matrix **X**. As expected, the observed data structure corresponds to a triangular geometry (as would be obtained for simplex data), the 3 vertices being expected to correspond to the pure compounds [45]. Similarly, the data points representation of the second and third column of the normalized matrix **Y**, (Y2, Y3), (Fig. 4A), enable to identify vertices pointing at clearly distinct directions.

Convex-hull computation provided 14 archetypes points in the (X2, X3) space corresponding to essential spectra and 3 archetype points in the (Y2, Y3) space corresponding to essential single-wavelength images (black circles in Fig. 3A and 4A). Considering that the number of components is known, 3 archetype points were selected in both sub-spaces (filled green circles in Fig. 3A and 4A, respectively), which are expected to correspond to the purest spectral pixels and most selective wavelengths measured (see Fig. 3B and 4B, respectively). The provided spectral and image information can be readily interpreted for this simple data set (1070 cm^{-1} maximum selective peak for $NaNO_3$, 1090 cm^{-1} maximum selective peak for $CaCO_3$, 996 cm^{-1} maximum selective peak for Na_2SO_4). For the sake of comparison, the results obtained by SIMPLISMA are provided (Fig. S1 in Supplementary Material).

Fig. 5 shows the results obtained for a three-component MCR-ALS model (LOF = 10 %, r^2 = 99 %) and for the application of K-means considering 3 clusters. The selection of the number of clusters was set as 3 according to the mixture composition. For each cluster the class assignation vector has been refolded in the original image dimensions and shown with the pixels recognized as cluster member coloured in brown (Fig. 5B third column). For the sake of comparison, the results obtained from the previous archetype identification are also reported (Fig. 5). The similarity between the essential spectra and essential single-wavelength images obtained from our approach and the spectra and component distribution maps obtained applying MCR-ALS is striking.

Focusing on the spectra provided in Fig. 5A, the ones shown for K-means correspond to "centroid" spectra and are, as expected, not the pure ones, though in quite good agreement. It is worth noting that the centroid spectrum corresponding to the $NaNO_3$ salt is more similar to the pure one than for the 2 other salts. This can be explained by considering the density of points for each of the 3 clusters modelled by K-means (see Fig. S2 in Supplementary Material). As for K-means, the maps (Fig. 5B) obtained for each cluster are also very comparable (considering that the information is segmented).

This dataset was introduced to clearly show that in cases in which we have prior information on the number of components, high spectral and spatial selectivity, as well as a high number of pure pixels, MCR-ALS and K-means solutions are very comparable, with selection of the method depending on specific analysis goals. Also, the information retrieved with the 2 approaches can be readily extracted from the analysis of the geometry of the data.

3.2. LIBS mineral dataset

The mean image and the LIBS spectra obtained for the mineral sample are shown in Fig. 6A and B, respectively. In Fig. 6A it is important to note that the pixel size is 15 μm. Considering the scale of mineral phases, the presence of many pure spectral pixels is, therefore, not expected. Fig. 6B highlights data characterized by low spectral selectivity. An additional complexity of this sample arises from its composition, which includes iron. Iron has numerous emission lines across the entire spectral range. Additionally, pyrite typically exists in various oxidative forms [39,42,51], and the iron ions within pyrite can easily exchange with copper or aluminium ions present in turquoise [52]. In fact, this kind of rocks are often referred to as "solid mixtures" [53]. Furthermore, within quartz, the predominant silica phase in this sample, aluminium impurities are quite common, while iron inclusions are also possible, albeit less frequent [54]. All these peculiarities translate into a very challenging LIBS HSI dataset to analyse and investigate with classical chemometric tool. Indeed, this scenario is not ideal for approaches such as MCR-ALS as pure pixels may not be present, spectral selectivity is low and different phases with very similar spatial distribution are present. Similarly, K-means clustering is not ideal as it may have difficulty assigning different minority phases to distinct clusters, as pixels may belong to multiple clusters due to low spectral selectivity.

The geometry of the data in the (X2, X3) and (Y2, Y3) spaces is illustrated in Fig. 7A and B, respectively. While more complex than the geometry observed in the previous example, the observed data points clouds exhibit some degree of structure. However, determining the appropriate number of components to consider is not straightforward given the absence of clear a priori information with this dataset. Convex-

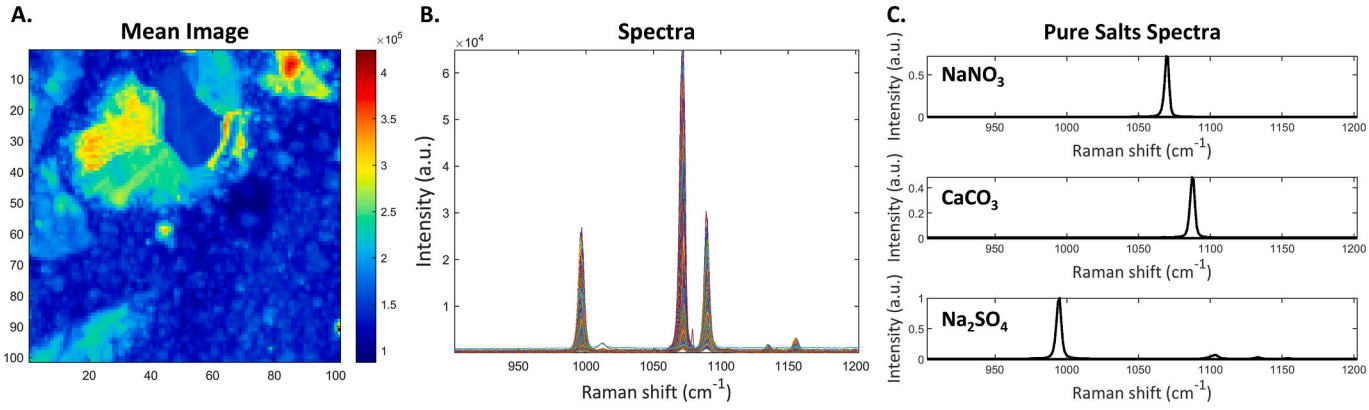


Fig. 2. Raman powder dataset: mean image (A), spectra (B) and spectra of pure salts (C).

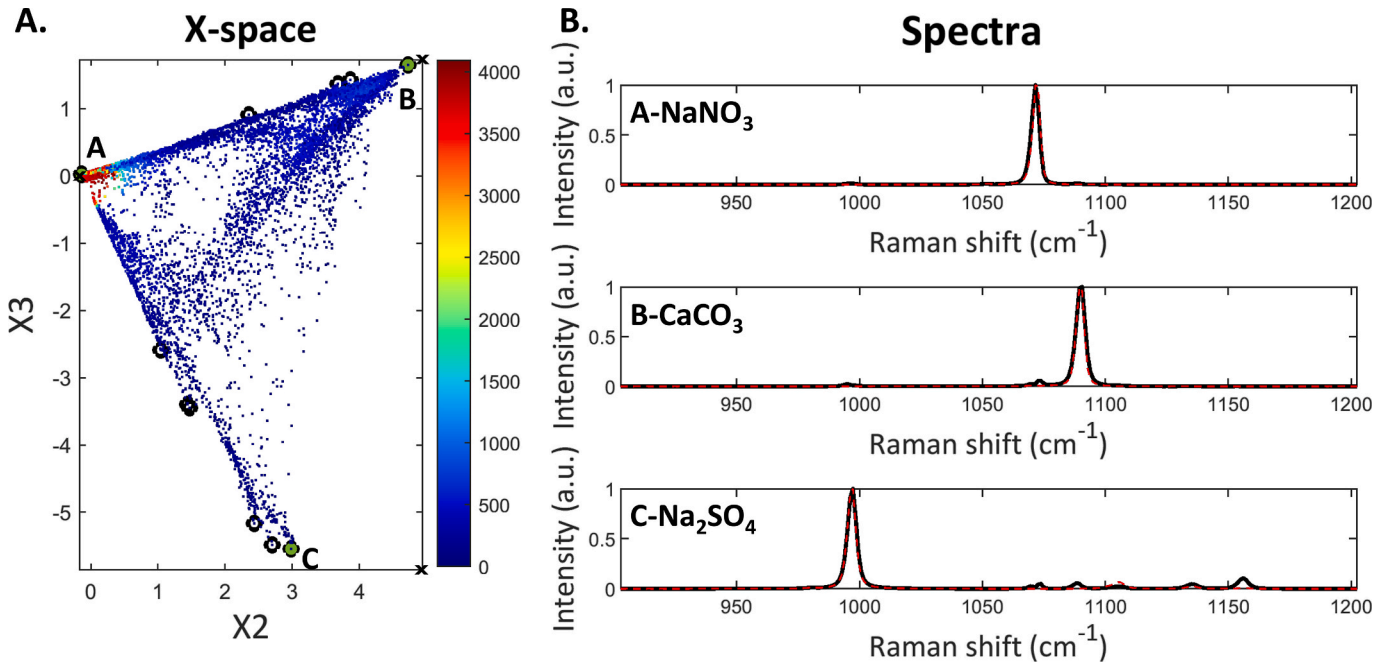


Fig. 3. 2-D representation of the X-space (A) colour-coded by point density. Black circles mark the archetypes points (some are close and result overlapped in the plot) at the vertices of the convex hull computed in the (X2, X3) normalized space. Filled green circles are the selected points and black crosses are the projection of the pure reference spectra in the (X2, X3) normalized space. In panel B, the spectra corresponding to the green points (black line) with overlapped the pure spectrum (red line) of the corresponding component. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

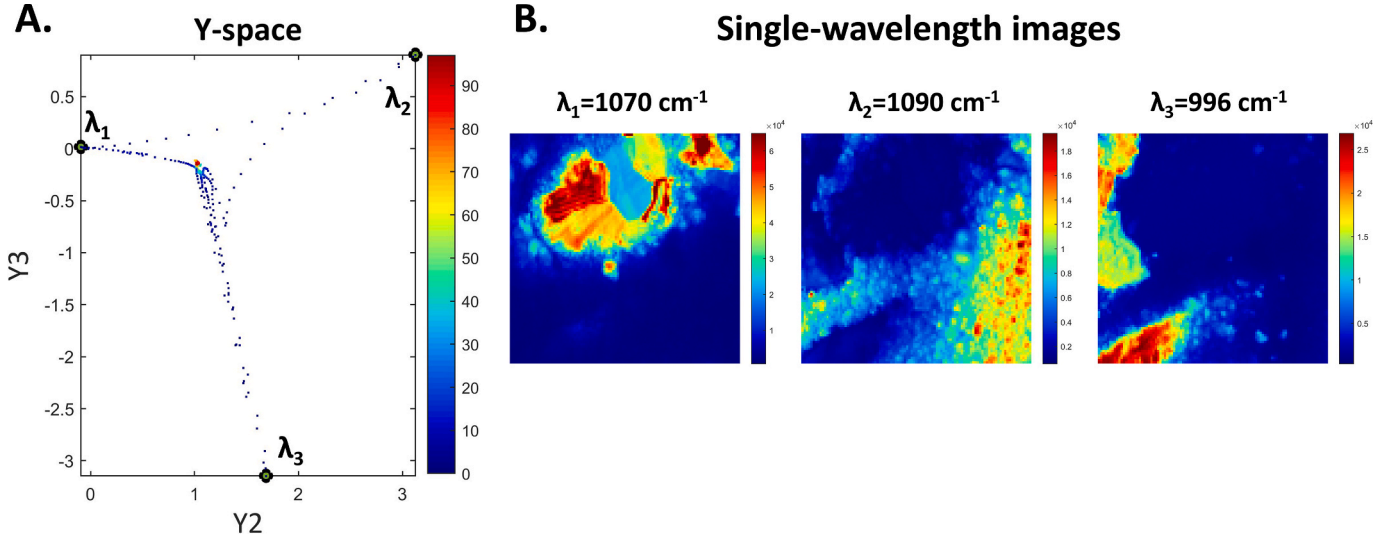


Fig. 4. 2-D representation of the Y-space colour-coded by point density (A). Black circles mark the archetypes points of the convex hull computed in the (Y2, Y3) normalized space. Filled green circles are the points identified looking at the structure of the data. In panel B, the essential single-wavelength images corresponding to the 3 identified selective wavelengths. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

hull computation provided 19 archetype points in (X2, X3), and 13 archetype points in (Y2, Y3), (black circles in Fig. 7A and B, respectively). Considering the geometry of the data observed in Figs. 7A and 6 archetype points were selected (filled green circles) and the corresponding essential spectra are shown. However, considering Fig. 7B–is clear that some relevant points, corresponding to clear directions, were not identified as archetypes, as they are not found at vertices of the data points cloud in the two-dimensional Y-space. It should be noted that by applying convex hull calculation to a six-dimensional **Y** matrix (see Supplementary Material Fig. S3), these points could be selected, but the total number of archetypes would be very large. This is not really needed, though, since they can be manually pointed out in the (Y2, Y3) plot, resulting in the extraction of 7 essential single-variable images.

The spectra corresponding to points labelled A, C and D in Fig. 7A correspond to the main mineral phases of pyrite, silica, and turquoise respectively. Spectrum B shows spectral features corresponding to a phase where silica has iron inclusions, somehow in between the pyrite and the main phase of silica. Spectrum F features another pyrite phase, different from the one observed in A. Lastly, the spectrum corresponding to pixel E characterizes an intermediate phase between turquoise and pyrite, where iron and mainly aluminium exchanges occur. The spectral regions used for the identification are highlighted in blue referring to

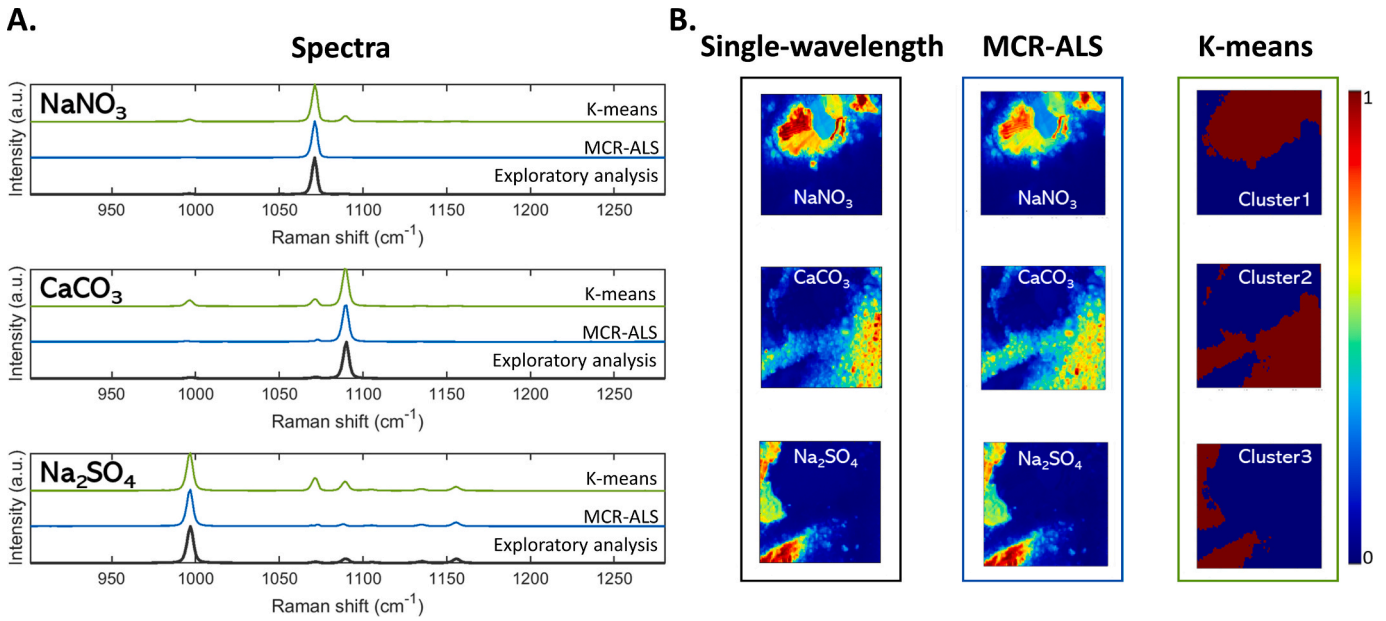


Fig. 5. Panel A shows the spectra for the purest pixels obtained by the exploratory analysis (black line), the purest components resolved spectra by MCR-ALS (blue) and the K-means centroids spectra (green). Centroids spectra are calculated as the average of the spectra of all the pixels belonging to a given cluster. For sake of clarity an arbitrary vertical offset was added to the MCR-ALS and K-means results. Panel B shows the single-wavelength images extracted with the exploratory approach, the concentration distribution maps retrieved by MCR-ALS and the clustering maps obtained by K-means clustering. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

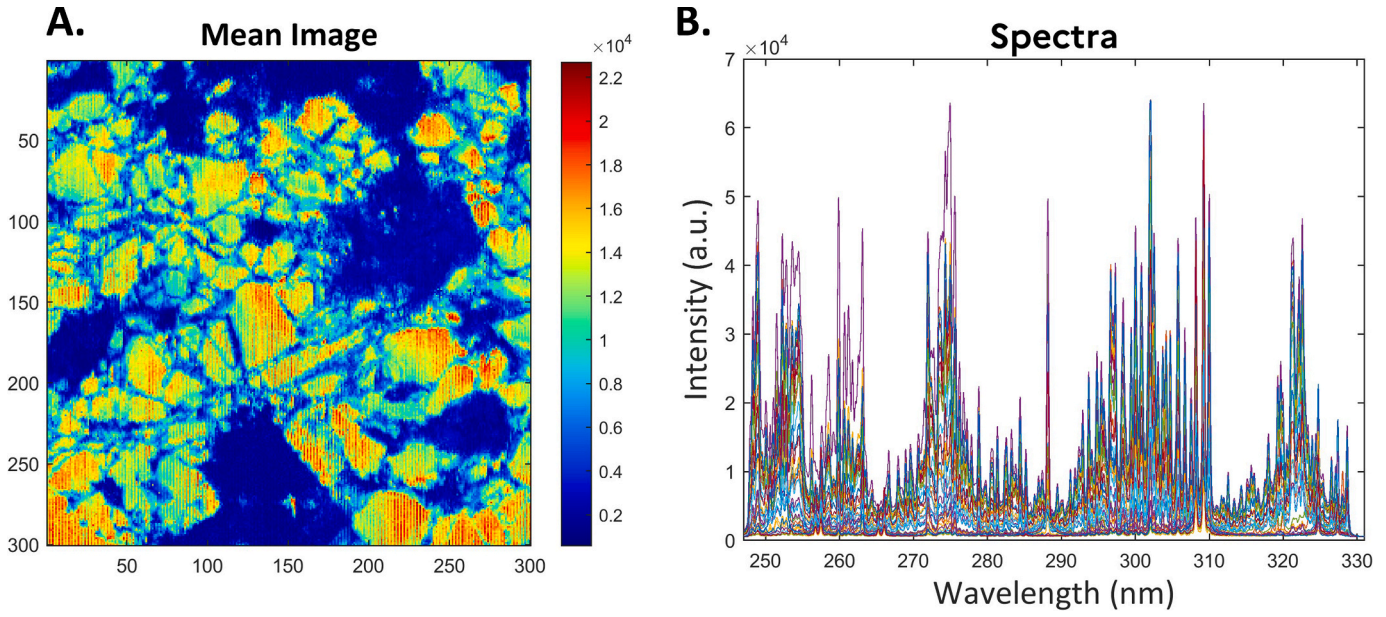


Fig. 6. Mean image (A) and overlapped spectra (B) of the mineral sample dataset.

Kurucz LIBS database [55], following the assessed procedure of Moncayo et al. [42]. These spectra are the purest spectra identified and can be interpreted as such without further analysis of the data.

In the same way, the essential single-wavelength images extracted correspond to the information obtained at the most selective wavelengths. Images λ_1, λ_3 and λ_7 which are linked to point A, C and D in Fig. 7B, respectively describe pyrite, silica and turquoise. Image λ_2, corresponding to point B, describes a situation where both pyrite and silica are present and image λ_4 shows the distribution of pyrite, turquoise and silica. It is worth noting that when looking at image λ_5, which does not show any correspondence in the (X2, X3) plot, it could be hypothesised that it represents a mineral phase where both silica and turquoise are present. In fact, it lies between image λ_3 and λ_7 in the (Y2, Y3) plot. Image λ_6, linked to point F, is identified as another form of pyrite. In addition, it can be noticed that in the right area of both the (X2, X3) and (Y2, Y3) plots, there is a higher density of points (either pixels or spectral wavelengths). Since points A and F correspond to spectra that are associated to pyrite phases, it can be concluded that pyrite is identified as the major phase in this mineral sample. For comparison purposes, the results obtained by SIMPLISMA are also provided (Fig. S4 in Supplementary Material). A six-component MCR-ALS model could then be fitted (LOF = 3 %, r^2 = 99 %) and the results are shown in Fig. 8.

The spectra of the first 2 components of the MCR-ALS model are identified as silica and turquoise phases, respectively. The

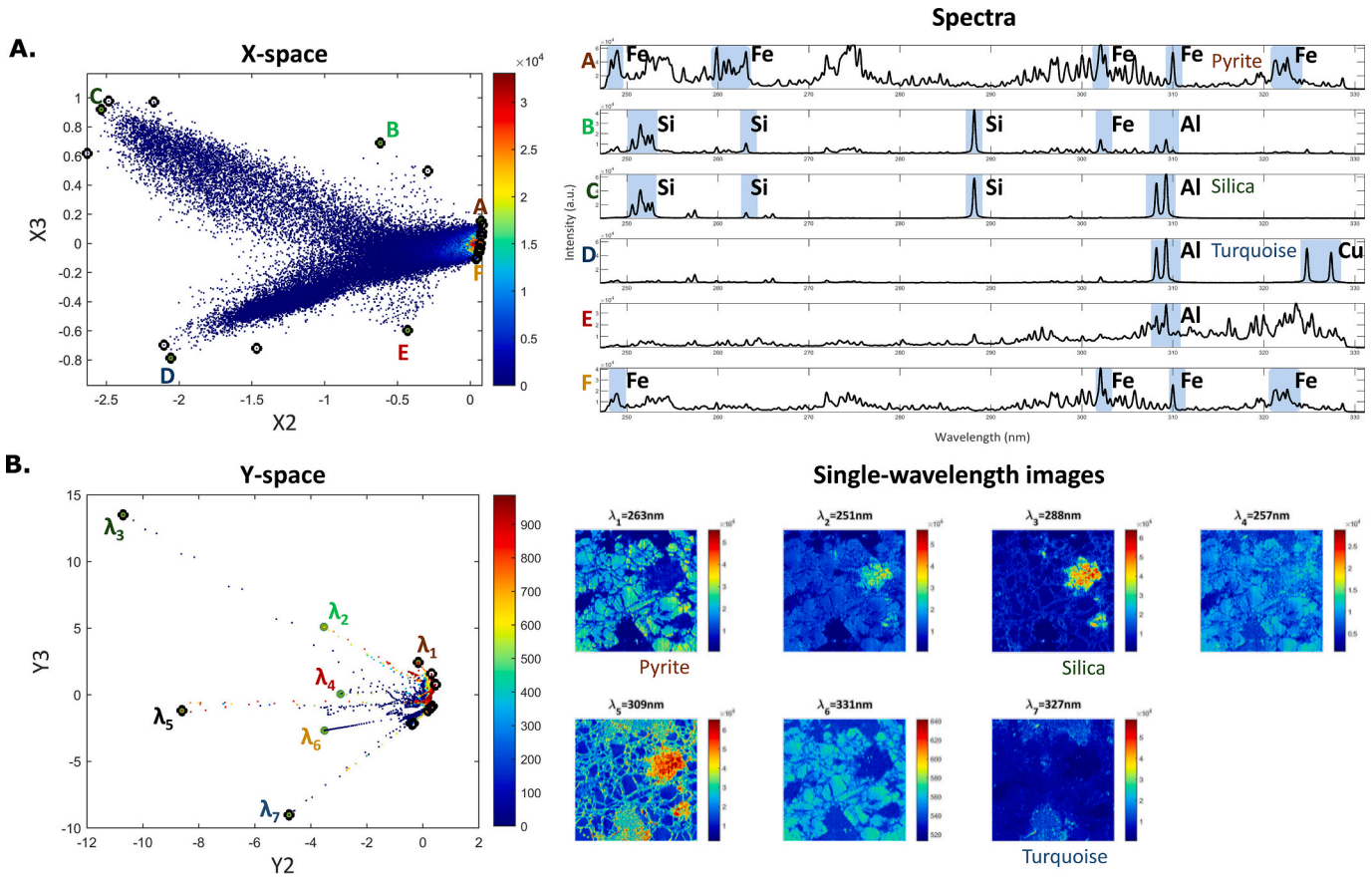


Fig. 7. (A) 2-D representation of the X-space of the mineral sample dataset, colour-coded by point density. Black circles mark the archetypes points at the vertices of the convex hull computed in the (X2, X3) normalized space. Letters and filled green circles represent the selected points, while the corresponding spectra are shown in the right panel. (B) 2-D representation of the Y-space, colour-coded by point density. Black circles mark the archetypes points of the convex hull computed in the (Y2, Y3) normalized space. Filled green circles represent the selected wavelengths, the corresponding refolded images are shown in the right panel. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

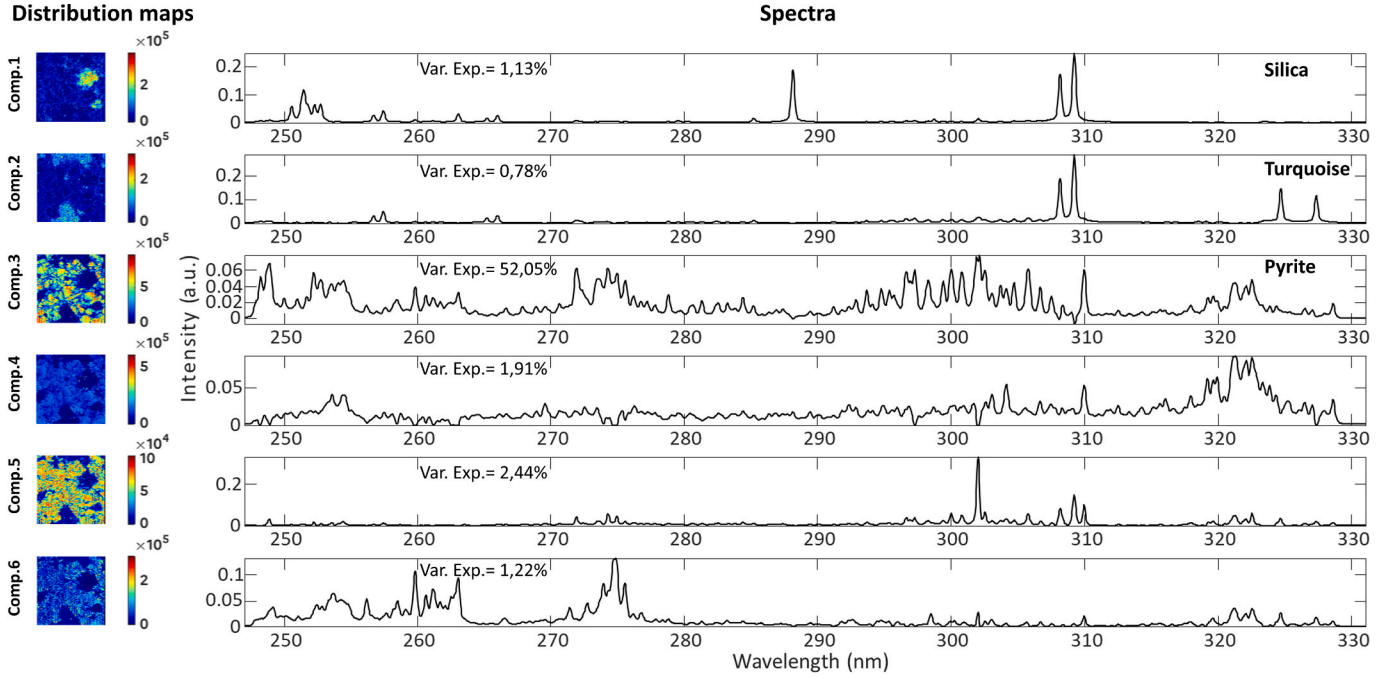


Fig. 8. MCR-ALS solutions for the mineral sample dataset. Refolded concentration profiles (left) and resolved spectra (right) are shown for each of the 6 components with the corresponding data variance.

corresponding concentration distribution are in agreement with the images retrieved by the exploratory analysis. Pyrite is identified primarily in the spectral profile of the third component. However, the spectral profiles observed in the remaining 3 components suggest the potential occurrence of different pyrite phases characterized by ion exchanges, which present challenges for interpretation. This is further complicated by the fact that the corresponding concentration distribution maps show very similar distributions. The third MCR component is the one explaining most of the data variance (52 %) confirming that pyrite is the major phase. By contrast the variance explained by other components, is very low, less than 3 % for silica, turquoise and the other phases of pyrite.

For the sake of comparison, a K-means model was computed setting the number of clusters to 6. The results are shown in Fig. 9. Cluster 1, 2 and 3 can be associated to silica, turquoise and pyrite phases, respectively. The clustering maps for clusters 4 to 6 reveal distributions spanning the boundaries between pyrite and the phases described by the first 2 clusters. The centroid spectra of these clusters are challenging to interpret, suggesting possible exchanges between iron and aluminium.

MCR-ALS and K-means clustering provide complementary information that leads to a more complete understanding of the sample. The proposed methodology allows for observing the potential complexity of data exploration prior to implementing MCR and/or K-means. It is important to note that the exploratory approach not only provides the same information as the one obtained from data modelling, but also enables to extract the spectral and spatial features related to the presence of minority components resulting from ion exchanges between the main mineral phases. The results obtained for silica, turquoise, and pyrite are comparable. The centroid spectra obtained by K-means and identified as pyrite is very comparable with the one extracted exploring the X-space and MCR-ALS, again because of the high number of pure pixels in that cluster (being pyrite the major phase, high number of pixels correspond only to pyrite). The concentration maps of 3 of the MCR-ALS components and the clustering maps of 3 of the clusters show the same distribution observed in the purest images extracted from the Y-space, while the other differ and as discussed above, are not easily interpretable.

Overall, we may remark that in this challenging scenario, that deviates from the ideal model underlying both MCR and clustering techniques, exploratory analysis driven by archetypes identification can provide insight into the number of components (when going for an unmixing approach) or clusters (when using clustering) to select. In fact, traditional methods such as eigenvalues, scree plots, and cluster indices may not provide unambiguous answers, as illustrated in Fig. S5 in the Supplementary Material. The exploratory approach employed in this study offers notable advantages, particularly in the extraction of spectra and images without the need for complex modelling. Also, convex hulls need to be calculated for more than 2 components, in order to retrieve the archetype points for each direction in the Y-space. These findings emphasize the feasibility and efficiency of our methodology in obtaining informative data without excessive computational load.

4. Conclusion

Understanding the structure of the data is a key step in the data analysis workflow of any application. In particular, exploring HSI datasets, because of their nature and dimensionality, is nontrivial. An exploratory approach, like the one proposed in this work, demonstrate to be able to guide extracting the useful information encrypted in the spectral image of complex samples and furnishing a comprehensive understanding of the investigated system.

Two different datasets were analyzed in this work by the exploratory approach and compared with the conventional methods of two widely used approaches in spectral image analysis: spectral unmixing (MCR-ALS) and clustering (K-means), with the aim of envisioning their applicability domain. The shared information, among all methods, in terms of distribution and spectral signature of retrieved common components, concerned major phases and/or the one with selective spectral profile. In cases, where the application of very well-known methodologies revealed its limits, looking at the geometry of the data resulted in an extremely easy and fast way to have better and more complete insights, with respect to the MCR-ALS and/or K-means ones. The analysis of the structure of the data could be considered, as any exploratory tool, as preliminary to allow a more rational choice of the next steps of data analysis and also to help solve all the cases of limitations for the two methods, such as the choice of the number of components and clusters,

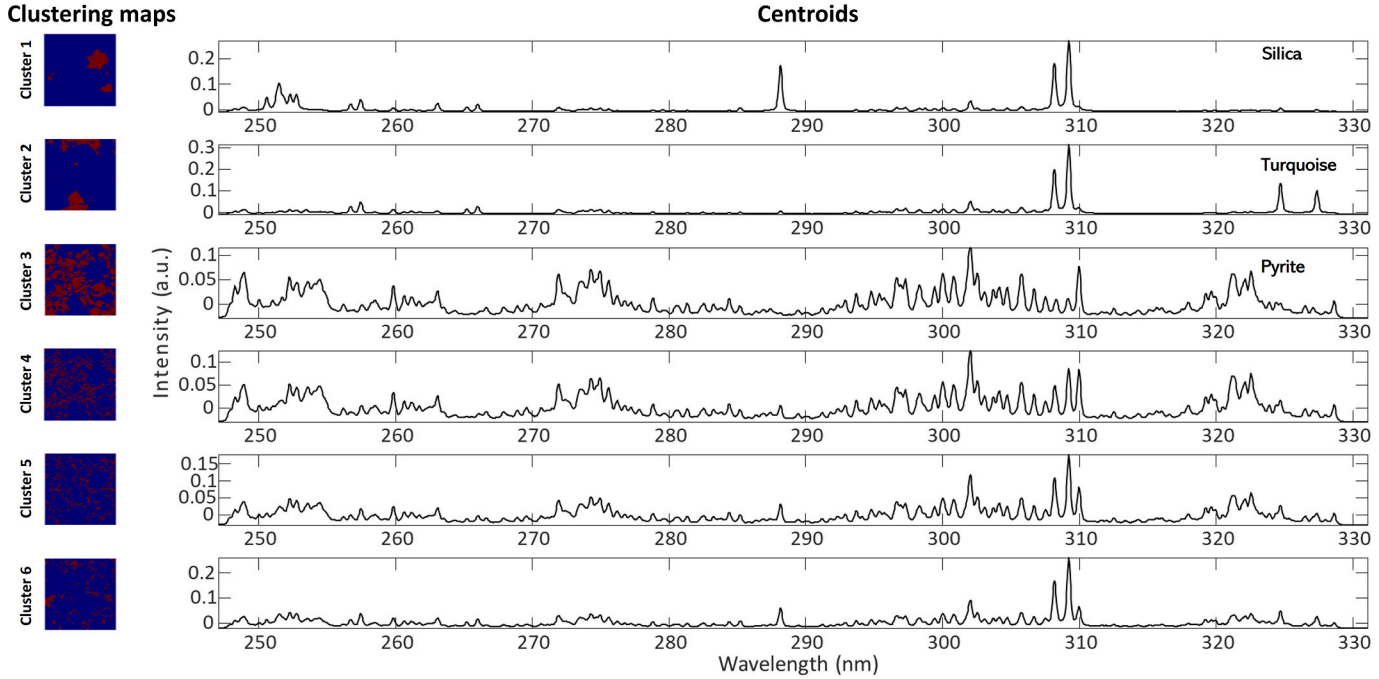


Fig. 9. K-means results for the mineral sample dataset. Cluster membership maps and the mean spectra are shown for each of the 6 clusters.

and in the retrieval and identification of the purest species as well as minor components.

This exploratory approach may have limitations when the data present a quite uniform distribution with no clear structures, thus rendering difficult finding the archetype points. However, to the best of our knowledge, applying appropriate spectral pre-processing could remove those effects, such as baseline, scatter, etc., that go into making the data less geometrically structured. In this way, a change in the "data shape" can be obtained making this approach therefore applicable. Furthermore, while automation of this process could be considered, it bears the risk of yielding inaccurate results, as extreme points may also include noise points requiring visual inspection before selection. Moreover, any automated implementation must carefully consider relevant parameters and considering convex hull algorithm proves significantly more reliable in this regard.

In conclusion, this paper highlights also the potential synergy between the exploratory analysis and the unsupervised methods of clustering and unmixing. Further exploration of their combined application, which remains relatively unexplored in the scientific community, is warranted, thus paving the way for a new research direction.

CRediT authorship contribution statement

Alessandra Olarini: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marina Cocchi:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Investigation, Funding acquisition, Conceptualization. **Vincent Motto-Ros:** Resources, Data curation. **Ludovic Duponchel:** Writing – review & editing, Visualization, Validation, Supervision, Investigation, Funding acquisition, Conceptualization. **Cyril Ruckebusch:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge Laureen Coïc for making data available. A. O. grant received support from Fondo Dipartimentale per la Ricerca (FDR2020) Università degli Studi di Modena e Reggio Emilia and Erasmus+ program University of Modena and Reggio Emilia. A.O. acknowledges LASIRE-DyNaChem research team for fruitful discussion.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2024.105174.

References

[1] J.M. Amigo, Hyperspectral and multispectral imaging: setting the scene, Data Handling Sci. Technol. 32 (2020) 3–16, https://doi.org/10.1016/B978-0-444-63977-6.00001-8.
[2] B. Gaci, F. Abdelghafour, M. Ryckewaert, S. Mas-Garcia, M. Louargant, F. Verpont, Y. Laloum, R. Bendoula, G. Chaix, J.M. Roger, A novel approach to combine spatial and spectral information from hyperspectral images, Chemometr. Intell. Lab. Syst. 240 (2023) 104897, https://doi.org/10.1016/j.chemolab.2023.104897.
[3] L. Coic, P.Y. Sacré, A. Dispas, C. De Bleye, M. Fillet, C. Ruckebusch, P. Hubert, E. Ziemons, Pixel-based Raman hyperspectral identification of complex pharmaceutical formulations, Anal. Chim. Acta 1155 (2021), https://doi.org/10.1016/j.aca.2021.338361.
[4] G. Lu, B. Fei, Medical hyperspectral imaging: a review, J. Biomed. Opt. 19 (2014) 010901, https://doi.org/10.1117/1.jbo.19.1.010901.
[5] B. Lu, P.D. Dao, J. Liu, Y. He, J. Shang, Recent advances of hyperspectral imaging technology and applications in agriculture, Rem. Sens. 12 (2020) 1–44, https://doi.org/10.3390/RS12162659.
[6] N. Keshava, J.F. Mustard, Spectral unmixing, IEEE Signal Process. Mag. 19 (2002) 44–57, https://doi.org/10.1109/79.974727.
[7] V. Olmos, L. Benítez, M. Marro, P. Loza-Alvarez, B. Piña, R. Tauler, A. de Juan, Relevant aspects of unmixing/resolution analysis for the interpretation of biological vibrational hyperspectral images, TrAC, Trends Anal. Chem. 94 (2017) 130–140, https://doi.org/10.1016/j.trac.2017.07.004.
[8] C. Ruckebusch, Resolving spectral mixtures: with applications from ultrafast time-resolved spectroscopy to super-resolution imaging. Data Handling in Science and Technology 30, Elsevier, 2016.
[9] R. Tauler, Multivariate curve resolution applied to second order data, Chemometr. Intell. Lab. Syst. 30 (1995) 133–146, https://doi.org/10.1016/0169-7439(95)00047-X.
[10] H. Abdollahi, M. Maeder, R. Tauler, Calculation and meaning of feasible band boundaries in multivariate curve resolution of a two-component system, Anal. Chem. 81 (2009) 2115–2122, https://doi.org/10.1021/ac8022197.
[11] J. Jaumot, R. Tauler, MCR-BANDS: a user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemometr. Intell. Lab. Syst. 103 (2010) 96–107, https://doi.org/10.1016/j.chemolab.2010.05.020.
[12] B. Celik, QLSU (QGIS Linear Spectral Unmixing) Plugin: an open source linear spectral unmixing tool for hyperspectral & multispectral remote sensing imagery, Environ. Model. Software 168 (2023) 105782, https://doi.org/10.1016/j.envsoft.2023.105782.
[13] J. Chaumel, M. Marsal, A. Gómez-Sánchez, M. Blumer, E.J. Gualda, A. de Juan, P. Loza-Alvarez, M.N. Dean, Autofluorescence of stingray skeletal cartilage: hyperspectral imaging as a tool for histological characterization, Discov Mater 1 (2021), https://doi.org/10.1007/s43939-021-00015-x.
[14] N. Cavallini, L. Strani, P.P. Becchi, V. Pizzamiglio, S. Michelini, F. Savorani, M. Cocchi, C. Durante, Tracing the identity of Parmigiano Reggiano "Prodotto di Montagna - Progetto Territorio" cheese using NMR spectroscopy and multivariate data analysis, Anal. Chim. Acta 1278 (2023), https://doi.org/10.1016/j.aca.2023.341761.
[15] C. Ruckebusch, R. Vitale, M. Ghaffari, S. Hugelier, N. Omidikia, Perspective on essential information in multivariate curve resolution, TrAC, Trends Anal. Chem. 132 (2020) 116044, https://doi.org/10.1016/j.trac.2020.116044.
[16] R.S. Michalski, Knowledge acquisition through conceptual clustering: a theoretical framework and an algorithm for partitioning data into conjunctive concepts, Int. J. Pol. Anal. Inf. Syst. 4 (1980) 219–244.
[17] T.S. Madhulatha, An overview on clustering methods, IOSR J. Eng. 2 (2012) 719–725, https://doi.org/10.9790/3021-0204719725.
[18] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability vol. 1, 1967, pp. 281–297. Statistics.
[19] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 1650–1654, https://doi.org/10.1109/TPAMI.2002.1114856.
[20] B. Desgraupes, Package clusterCrit: clustering indices, CRAN Package (2017) 1–34, cran.r-project.org/web/packages/clusterCrit.
[21] A. Kaarna, P. Zemcik, H. Kälviäinen, J. Parkkinen, Compression of multispectral remote sensing images using clustering and spectral reduction, IEEE Trans. Geosci. Rem. Sens. 38 (2000) 1073–1082, https://doi.org/10.1109/36.841986.
[22] S. Piqueras, C. Krafft, C. Beleites, K. Egodage, F. von Eggeling, O. Guntinas-Lichius, J. Popp, R. Tauler, A. de Juan, Combining multiset resolution and segmentation for hyperspectral image analysis of biological tissues, Anal. Chim. Acta 881 (2015) 24–36, https://doi.org/10.1016/J.ACA.2015.04.053.
[23] L. Massart, D. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, 1983.
[24] T. Celik, Unsupervised change detection in satellite images using principal component analysis and κ-means clustering, Geosci. Rem. Sens. Lett. IEEE 6 (2009) 772–776, https://doi.org/10.1109/LGRS.2009.2025059.
[25] I.E. Kaya, A.Ç. Pehlivanlı, E.G. Sekizkardeş, T. Ibrikci, PCA based clustering for brain tumor segmentation of T1w MRI images, Comput. Methods Progr. Biomed. 140 (2017) 19–28, https://doi.org/10.1016/j.cmpb.2016.11.011.
[26] P. Firmani, S. Hugelier, F. Marini, C. Ruckebusch, MCR-ALS of hyperspectral images with spatio-spectral fuzzy clustering constraint, Chemometr. Intell. Lab. Syst. 179 (2018) 85–91, https://doi.org/10.1016/j.chemolab.2018.06.007.
[27] D. ChengX, Z. Cai, J. Li, M. Wen, Y. Wang, A. Zeng, spatial-spectral clustering-based algorithm for endmember extraction and hyperspectral unmixing, Int. J. Rem. Sens. 42 (2021) 1948–1972.
[28] J.W. Tukey, Exploratory Data Analysis, vol. 2, Addison-Wesley Publishing Company, 1977.
[29] M. Li Vigni, C. Durante, M. Cocchi, Exploratory Data Analysis, first ed., Elsevier, 2013 https://doi.org/10.1016/B978-0-444-59528-7.00003-X.
[30] M. Ghaffari, N. Omidikia, C. Ruckebusch, Essential spectral pixels for multivariate curve resolution of chemical images, Anal. Chem. 91 (2019) 10943–10948, https://doi.org/10.1021/acs.analchem.9b02890.
[31] L. Coic, R. Vitale, M. Moreau, D. Rousseau, J.H. de Morais Goulart, N. Dobigeon, C. Ruckebusch, Assessment of essential information in the fourier domain to

accelerate Raman hyperspectral microimaging, Anal. Chem. 95 (2023) 15497–15504, https://doi.org/10.1021/acs.analchem.3c01383.
[32] S.V. Zade, K. Neymeyr, M. Sawall, C. Fischer, H. Abdollahi, Data point importance: information ranking in multivariate data, J. Chemom. 37 (2023) 1–15, https://doi.org/10.1002/cem.3453.
[33] V.H.C. Ferreira, V. Gardette, B. Busser, L. Sancey, S. Ronsmans, V. Bonneterre, V. Motto-Ros, L. Duponchel, Enhancing diagnostic capabilities for occupational lung diseases using LIBS imaging on biopsy tissue, Anal. Chem. (2024), https://doi.org/10.1021/acs.analchem.4c00237.
[34] Q. Wu, C. Marina-Montes, J.O. Cáceres, J. Anzano, V. Motto-Ros, L. Duponchel, Interesting features finder (IFF): another way to explore spectroscopic imaging data sets giving minor compounds and traces a chance to express themselves, Spectrochim. Acta Part B At. Spectrosc. 195 (2022), https://doi.org/10.1016/j.sab.2022.106508.
[35] M. Ghaffari, N. Omidikia, C. Ruckebusch, Joint selection of essential pixels and essential variables across hyperspectral images, Anal. Chim. Acta 1141 (2021) 36–46, https://doi.org/10.1016/j.aca.2020.10.040.
[36] S. Khodadadi Karimvand, J. Mohammad Jafari, S. Vali Zade, H. Abdollahi, Practical and comparative application of efficient data reduction - multivariate curve resolution, Anal. Chim. Acta 1243 (2023) 340824, https://doi.org/10.1016/j.aca.2023.340824.
[37] M. Sawall, C. Ruckebusch, M. Beese, R. Francke, A. Prudlik, K. Neymeyr, An active constraint approach to identify essential spectral information in noisy data, Anal. Chim. Acta 1233 (2022) 340448, https://doi.org/10.1016/j.aca.2022.340448.
[38] R. Vitale, C. Ruckebusch, On a black hole effect in bilinear curve resolution based on least squares, J. Chemom. 37 (2023) 1–7, https://doi.org/10.1002/cem.3442.
[39] E.C. Muñoz, F. Gosetti, D. Ballabio, S. Andò, O. Gómez-Laserna, J.M. Amigo, E. Garzanti, Characterization of pyrite weathering products by Raman hyperspectral imaging and chemometrics techniques, Microchem. J. 190 (2023), https://doi.org/10.1016/j.microc.2023.108655.
[40] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – a review, Anal. Chim. Acta 1145 (2021) 59–78, https://doi.org/10.1016/j.aca.2020.10.051.
[41] A. De Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964–4976.
[42] S. Moncayo, L. Duponchel, N. Mousavipak, G. Panczer, F. Trichard, B. Bousquet, F. Pelascini, V. Motto-Ros, Exploration of megapixel hyperspectral LIBS images using principal component analysis, J Anal At Spectrom 33 (2018) 210–220, https://doi.org/10.1039/c7ja00398f.
[43] C. Golub, G. H Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (1970) 403–420, https://doi.org/10.1007/BF02163027.
[44] R. Rajkó, Studies on the adaptability of different Borgen norms applied in self-modeling curve resolution (SMCR) method, J. Chemom. 23 (2009) 265–274, https://doi.org/10.1002/cem.1221.
[45] B.V. Grande, R. Manne, Use of convexity for finding pure variables in two-way data from mixtures, Chemometr. Intell. Lab. Syst. 50 (2000) 19–33, https://doi.org/10.1016/S0169-7439(99)00041-6.
[46] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS gui 2.0: new features and applications, Chemometr. Intell. Lab. Syst. 140 (2015) 1–12, https://doi.org/10.1016/j.chemolab.2014.10.003.
[47] W. Windig, J. Guilment, Interactive Self-Modeling Mixture Analysis, 1991, pp. 1425–1432.
[48] V. Kumar, J.K. Chhabra, K. Dinesh, Performance evaluation of distance metrics in the clustering algorithms, INFOCOMP J. Comput. Sci. 13 (2014) 38–51.
[49] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.
[50] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recogn. 37 (2004) 487–501, https://doi.org/10.1016/j.patcog.2003.06.005.
[51] G. Hu, K. Dam-Johansen, S. Wedel, J.P. Hansen, Decomposition and oxidation of pyrite, Prog. Energy Combust. Sci. 32 (2006) 295–314, https://doi.org/10.1016/j.pecs.2005.11.004.
[52] X. Wang, Y. Guo, The impact of trace metal cations and absorbed water on colour transition of turquoise, R. Soc. Open Sci. 8 (2021), https://doi.org/10.1098/rsos.201110.
[53] J.J. Rushchitsky, Interaction of waves in solid mixtures, Appl. Mech. Rev. 52 (1999) 35–74.
[54] https://www.mindat.org/, last access 18/October/2023.
[55] https://www.atomtrace.com/elements-database/, last access 18/October/2023.