Université de Lille – Sciences et Technologies

École Doctorale des Sciences de la Matière, du Rayonnement et de l'Environnement

Thèse de Doctorat

En vue de l'obtention du grade de

Docteur de l'Université de Lille

Discipline: Optique et Lasers, Physico-Chimie et Atmosphère

Alessandro Nardecchia

Chemometric exploration in hyperspectral imaging in the framework of big data and multimodality

Soutenue le 16 mai 2022:

Président du jury:	
M. Bruno BOUSQUET	Professeur, Université de Bordeaux
Rapporteurs:	
Mme Aoife GOWEN	Professeur, University College Dublin
M. Federico MARINI	Professeur, Università di Roma "La Sapienza"
M. Jorge O. CÁCERES GIANNI	Professeur, Universidad Complutense de Madrid
Examinateurs:	
Mme Delphine NEFF	Chercheuse HDR, CEA Saclay
Invités:	
Mme Anna DE JUAN	Professeur, Universitat de Barcelona
M. Vincent MOTTO-ROS	Maître de Conférences HDR, Université de Lyon 1
Directeur de Thèse:	
M. Ludovic DUPONCHEL	Professeur, Université de Lille

Université de Lille – Sciences et Technologies

École Doctorale des Sciences de la Matière, du Rayonnement et de l'Environnement

Thèse de Doctorat

En vue de l'obtention du grade de

Docteur de l'Université de Lille

Discipline: Optique et Lasers, Physico-Chimie et Atmosphère

Alessandro Nardecchia

Exploration chimiométrique en imagerie hyperspectrale dans le cadre du big data et de la multimodalité

Soutenue le 16 mai 2022:

Président du jury:

M. Bruno BOUSQUET	Professeur, Université de Bordeaux

Rapporteurs:

Mme Aoife GOWEN	Professeur, University College Dublin
M. Federico MARINI	Professeur, Università di Roma "La Sapienza"
M. Jorge O. CÁCERES GIANNI	Professeur, Universidad Complutense de Madrid

Examinateurs:

Mme Delphine NEFF	Chercheuse HDR, CEA Saclay

Invités:

Mme Anna DE JUAN	Professeur, Universitat de Barcelona
M. Vincent MOTTO-ROS	Maître de Conférences HDR, Université de Lyon 1

Directeur de Thèse:

M. Ludovic DUPONCHEL	Professeur, Université de Lille

“Ad maiora…Semper”

ACKNOWLEDGMENTS

Three years have been passed so fast. It seems like yesterday my first day starting this PhD, but if I compare myself from back then and now, I realize that I have deeply changed and I have gained a lot of experience, from every point of view. I have spent many good moments, as well as some hard ones. However, I have never given up, showing to myself that I have the capacity to stand always up again. Clearly, this is not only my merit. I have been surrounded by amazing people that made it possible to enjoy this incredible experience. I know it is not possible, but I wish I could thank properly everyone.

First of all, I have to thank my supervisor, Ludovic. Working with me has not been always that easy, am I right? However, no matter what, you have seen in me something good and you have taught me a lot. You have been very understanding with me when needed, as well as strict when it was the moment to be. I owe you a lot, I cannot deny it. I will always cherish the time spent together and I will never forget your teachings. Thanks a lot, Ludo!

A big thank to all the members of the committee of my PhD defense. Thanks to have accepted to read and evaluate my work.

Following, all the people from LASIRE. They have been always kind and patient with me and have the possibility of communicating in French with them has been a good way to improve it. Particularly, thanks to Hervé Vezin, director of LASIRE, to have welcomed me in this group. Cyril, we have never worked together, but many times we have discussed about chemometrics and other stuff, you have given me different suggestions and you have been always well-disposed to help me... And I have repaid the favor satisfying your insatiable hunger with cookies and pain au chocolat. Therefore, I guess we are even, is not it? Aude, spending some moments with you during these three years has been terrific! Thanks to you... Nacer, you have been a good friend to me, we have had a good time together and you have always been ready to help me, when needed. Thanks for your support! Of course, my thanks go to Ákos and Bálint too, the best Hungarian friends ever! Every time I see Unicum, my thoughts turn to you, for obvious reasons. Afterward thanks to the "crew" of PhD students, who has finished and who is still there, people with which I had the possibility of sharing good and bad moments. However, I have to thanks three people from the University in particular. Ali! You have been, you are and you will always be a very good friend of mine. When I needed, you have been every time there for me, and I hope I have done the same with you. Thanks for your support, buddy. It has been fundamental for me. Alessandra, it is true, we didn't have a lot of time to be together, due to the fact that you arrived

when I was almost finishing the PhD and that I was facing a particular moment of my life. And, also if the first approach has not been the best one, we created a fantastic friendship. You have been so nice with me, we have spent so many good moments together... I have faith in you, I know you will rock! So go on like this, don't change ever! And then, finally you Raffa... You have been super-patient (maybe too much, sometimes) with me, both from a professional point of view (you have explained me plenty of things), but especially you have shown me your friendship in three years here in Lille. I have been very lucky to have you as a friend, thanks for all the good memories. Thanks to be my friend.

A mis compañeros de Barcelona! David, Rodrigo, Adri, Iker and of course you, Dario... Four months there with you have been amazing, I have found a second family there and I have loved every single moment spent together. Las comidas, las cervecitas y todo el resto, claramente muy gustoso! Gracias por todo, mis amigos! And naturally, thanks to you Anna... You have welcomed me in your research group and you have taught me many things. The experience I had there have been just amazing... I will never be able to thanks you enough.

To my friends, the ones that I have met in Lille and the ones of a whole life (the ones that have hated me when I have told them that I was leaving Rome to start a PhD of three years in France). Without your support, these three years would have been harsher, for sure. Your words, your presence, your pep talks have been fundamental for me, always. Thanks to you all (and you are many!)... I owe you more than I will ever be able to give you in my whole life. Thank you all from deep in my heart. A special thanks to Massimo and its bar... The best part of the Cité Scientifique for sure! My friend, thanks for all the good experiences you have made me enjoy in Lille! And you Marilù, we had ups and downs, but when there was the moment, you were always there for me, as I was for you... I will AUUUUUUUUlways bring you with me in my hearth! Giulia, you stress me a lot (really, A LOT!), that for sure, but that's you, so we have to love you in the way you are, I suppose... Thanks for all the good memories (and drinks)!

Lastly, but especially, to my family. Probably the ones affected the most by my departure have been them. Despite this, they have been always proud of me and they have always supported my decisions because they just want the best for me. Without them, I would not be the person that I am today, and I am grateful for all their efforts and sustain. Thanks for everything you have done and do for me. I love you.

Sincerely, thanks to you all.

Alessandro

CONTENTS

ABBREVIATIONS AND ACRONYMS

ANN: Artificial Neural Network

AsLS: Asymmetric Least Squares

CCD: Charge-Coupled Device

CV: Cross-Validation

CWT: Continuous Wavelet Transform

DWT: Discrete Wavelet Transform

EDX: Energy-Dispersive X-ray spectroscopy

EKM: Embedded K-Means clustering

FCS: Fully Conditional Specification

FT: Fourier Transform

FTIR: Fourier Transform Infrared spectroscopy

HELP: Heuristic Evolving Latent Projections

IR: Infrared

ITTFA: Iterative Target Transformation Factor Analysis

JM: Joint Modeling

KM: K-Means clustering

LDA: Linear Discriminant Analysis

LIBS: Laser-Induced Breakdown Spectroscopy

LOF: Lack of Fit

LV(s): Latent Variable(s)

MC: Mean Centering

MCR: Multivariate Curve Resolution

MCR-ALS: Multivariate Curve Resolution-Alternating Least Squares

MIA: Multivariate Image Analysis

MICE: Multivariate Imputation by Chained Equations

MLR: Multiple Linear Regression

MSPC: Multivariate Statistical Process Control

Nd:YAG: Neodymium-doped Yttrium Aluminum Garnet

PBM: Pakhira-Bandyopadhuay-Maulik index

PC(s): Principal Component(s)

PCA: Principal Component Analysis

PIL: Plasma Induced Luminescence

PLS-DA: Partial Least Squares-Discriminant Analysis

PLSR: Partial Least Squares Regression

SEM: Scanning Electron Microscope

SFA: Subwindow Factor Analysis

SG: Savintzky-Golay derivative

SIMCA: Soft Independent Modelling of Class Analogies

SIMPLISMA: SIMPLe-to-use Interactive Self-modelling Mixture Analysis

SNV: Standard Normal Variate

SSE: Sum of Squared Errors

SVM: Support Vector Machine

SWT: Stationary Wavelet Transform

SWT 2-D: 2-D Stationary Wavelet Transform

WFA: Window Factor Analysis

ABSTRACT

Nowadays, it is widely known that hyperspectral imaging is a very good tool used in many chemical-related research areas. Indeed, it can be exploited for the study of samples of different nature, whatever the spectroscopic technique used. Despite the very interesting characteristics related to this kind of acquired data, various limitations are potentially faced. First of all, modern instruments can generate a huge amount of data (big datasets). Furthermore, the fusion of different spectroscopic responses on the same sample (multimodality) can be potentially applied, leading to even more data to be analyzed. This aspect can be a problem, considering the fact that if the right approach is not used, it could be complicated to obtain satisfying results or even lead to a biased vision of the analytical reality of the sample. Obviously, some spectral artifacts can be present in a dataset, and so the correction of these imperfections has to be taken into account to obtain better results. Another important challenge related to the use of hyperspectral image analysis is that normally, the simultaneous observation of spectral and spatial information is almost impossible. Clearly, this leads to an incomplete investigation of the sample of interest. Chemometrics is a modern branch of chemistry that can perfectly match the current limitations related to hyperspectral imaging. The purpose of this PhD work is to give to the reader a series of different topics in which many challenges related to hyperspectral images can be overcome using different chemometric facets. Particularly, as it will be described, problems such as the generation of big amount of data can be faced using algorithms based on the selection of the purest information (i.e., SIMPLISMA), or related to the creation of clusters in which similar components will be grouped (i.e., KM clustering). The problem related to the correction of instrumental artifacts (i.e., saturated signals) will be faced using a methodology based on the statistical imputation, in order to recreate in a very elegant way the missing information and thus, obtain signals that otherwise would be irremediably lost. A significant part of this thesis has been related to the investigation of data acquired using LIBS imaging, a spectroscopic technique that is currently obtaining an increasing interest in many research areas, but that, still, has not really been exploited to its full potential by the use of chemometric approaches. In this manuscript, it will be shown a general pipeline focusing on the selection of the most important information related to this kind of data cube (due to the huge amount of spectral data that can be easily generated) in order to overcome some limitations faced during the analysis of this instrumental response. Furthermore, the same approach will be exploited for the data fusion analysis related to LIBS and other spectroscopic data. Lastly, it will be shown an interesting way to use wavelet

transform, in order to not limit the analysis only to spectral data, but also to spatial ones, to obtain a more complete chemical investigation.

RÉSUMÉ

Nous sommes aujourd'hui tous conscients que l'imagerie hyperspectral est un outil très utile dans de nombreux domaines de recherche liés à la chimie et qu'elle peut être exploitée pour l'étude d'échantillons de nature différente, quelle que soit la technique spectroscopique utilisée. Malgré les caractéristiques très intéressantes liées à ce type de données, diverses limitations sont potentiellement rencontrées. Les instruments modernes peuvent tout d'abord générer une énorme quantité de données (big datasets). De plus, la fusion de différentes réponses spectroscopiques acquises sur le même échantillon (multimodalité) peut être potentiellement appliqué, conduisant à encore plus de données à analyser. Cet aspect peut être problématique, compte tenu du fait que si la bonne approche n'est pas utilisée, il peut être compliqué d'obtenir des résultats satisfaisants. Bien évidemment, certains artefacts spectraux peuvent être présents dans les jeux de données acquis, et donc la correction de ces imperfections doit être prise en compte pour obtenir de bons résultats. Un autre défi important lié à l'utilisation de l'analyse d'images hyperspectrales est que normalement, l'observation simultanée d'informations spectrales et spatiales est presque impossible avec la plupart des méthodes actuelles. De toute évidence, cela conduit à une exploration incomplète des données à disposition acquises sur l'échantillon d'intérêt. La chimiométrie est une branche moderne de la chimie qui peut parfaitement répondre aux limitations actuelles liées à la structure des données en imagerie hyperspectrale. Le but de ce travail de thèse est de présenter au lecteur une série de sujets différents dans lesquels de nombreux défis liés aux images hyperspectrales peuvent être surmontés en utilisant différentes facettes de la chimiométrie. En particulier, les problèmes liés à la génération d'une grande quantité de données peuvent être surmontés à l'aide d'algorithmes basés sur la sélection de l'information la plus pure (i.e., SIMPLISMA), ou liés à la création de clusters dans lesquels des composants similaires seront regroupés (i.e., KM clustering). Afin de corriger les artefacts instrumentaux tels que les signaux saturés, une méthodologie originale qui exploite l'imputation statistique sera utilisée, afin de recréer de manière très élégante les informations manquantes et ainsi obtenir des signaux qui autrement seraient irrémédiablement perdus. Une partie importante de cette thèse est liée à l'investigation des données acquises à l'aide de l'imagerie LIBS, une technique qui suscite actuellement un intérêt croissant dans de nombreux domaines de recherche, mais qui n'a pas encore vraiment été exploitée à son plein potentiel par l'utilisation des approches chimiométriques. Dans ce manuscrit, nous introduirons un pipeline général axé sur la sélection des informations les plus importantes liées à ce type de structure de données cubique (en raison

de l'énorme quantité de données spectrales qui peuvent être facilement générées) afin de surmonter certaines limitations rencontrées lors de l'analyse de cette réponse instrumentale. De plus, la même approche sera exploitée pour les problématiques de fusion de données spectrales, liée à la LIBS et à d'autres données spectroscopiques. Enfin, nous introduiront une manière intéressante d'utiliser la transformée en ondelettes (wavelet transform), afin de ne pas limiter l'analyse uniquement aux données spectrales, mais aussi spatiales, pour obtenir une exploration chimique plus complète des échantillons complexes.

INTRODUCTION

Analytical chemistry is nowadays a very important research area exploited in many different investigation fields, for many scientific purposes. Spectroscopic techniques represent very powerful tools to deal with the complexity and heterogeneity shown by real samples of different nature. Indeed, due to the instrumental developments in the last decades, the quality and quantity of the acquired data is constantly increasing. Factors such as faster acquisitions and more sensitive detection chains are just few examples of the reason of this phenomenon. Furthermore, nowadays it is common to refer to multimodality, i.e., the analysis of the same sample merging the information obtained using different spectroscopic instruments. For sure, this rapid development in analytical chemistry is leading to a real challenge in finding an adequate way to interpret the information of a given specimen and, thus, obtain satisfactory outcomes. Due to the heterogeneity of the investigated sample, it is fundamental to observe it in its entirety. Nowadays, a bulk analysis can be limitative, and at this moment it is fundamental to find a way to overcome the constraints shown by routine analyses, answering always more questions about the observed samples. Hyperspectral imaging is one of the possible solutions that currently can be used in any chemical investigation area. In fact, modern instrumentations can be easily coupled with an imaging setup, leading to new analytical exploration horizons. First, using a hyperspectral imaging system, it is clearly possible to observe a sample from a global perspective. One of the most important aspects of this kind of technique is that the sample is observed also from a spatial and not only the spectral point of view. This means that more information could be carried out by the acquisition of a sample, considering different facets. For example, the limitations of a bulk analysis are in this way overcome, and the heterogeneity of the specimen can be finally really studied, observing the spatial distribution of the various constituents of a complex matrix. Nevertheless, some important obstacles have to be faced. Due to the complexity and the quantity of acquired data (hundreds of thousands to millions of spectra obtained in very reasonable times), it is complicated to directly analyze the raw information. Deal with aspects such as multimodality, big data, and considering the fact that a good exploration is done when a good preprocessing is applied, is fundamental. The correction of artifacts, the data reduction with the purpose of using only the most relevant part of the information contained in the considered sample, the use of not exclusively the spectral, but also the spatial information, taking into account this particular aspect when a hyperspectral image is analyzed, are all factors that are nowadays essential. Chemometrics can be a good solution to all these problems. This discipline, in fact, is applied with the intent of learning the underlying relationships and structures of complex samples in order to obtain more particular information. It is known that in the last decades chemometrics has been vastly exploited also in the hyperspectral imaging context.

Nevertheless, more facets regarding the link between these two domains are possible and clearly required. Hyperspectral imaging potentialities are still not really exploited, and it is a duty of research to broaden its horizons. Too many limitations are nowadays correlated to this important kind of data. A hyperspectral image is a data cube made of pixels (the spectra of the given specimen) in which not only the spectral, but also the spatial details are available, leading to the possibility of observing the sample from a different point of view. The challenges are multiple. For example, normally, in order to carry out a routine analysis pipeline, this three-dimensional data cube is unfolded in its corresponding two-dimensional form, leading to the loss of all the information related to the spatial details (all the time of the chemometric analysis), and so, to an incomplete investigation. It is also true that, as already stated, the quantity of produced data is normally very big. This is for sure an important aspect, due to the fact that in this way it is possible to obtain more spectroscopic information related to the investigated sample. At the same time, the analysis of big datasets can be very complicated for different reasons. First, it is a hard task to deal with a huge number of spectra from a computational perspective (proper devices are required to work with). Then, and probably more importantly, the possibility of missing some specific and fundamental information that are related to very small areas of a given sample is a very common scenario. If the right approach is not taken into consideration, an inaccurate analysis would certainly lead to outcomes far from meeting the expectations.

Here finally the main purpose of this PhD project: exploiting various classical and emerging chemometric methodologies and algorithms (e.g., SIMPLISMA, K-Means, wavelet transform), work on big datasets acquired with different spectroscopic techniques (the most commonly used today, and also some recent ones that are nowadays obtaining always more importance and interest) and on the multimodality (operation that is possible coupling different device responses, due to the modern instrumental developments) with the perspective of providing new ways to deal with the limitations that are currently related to hyperspectral image analysis. The various concepts will be described in the present manuscript, facing different problems and giving some possible solutions based on already existent and new chemometric methods. In order to evaluate the quality of the presented research line, various data matrices acquired with different spectroscopies will be investigated and, depending on the main task taken into consideration, particular approaches will be proposed and described in detail.

CHAPTER 1

1. HYPERSPECTRAL IMAGE ANALYSIS: AN OVERVIEW

1.1. Introduction to hyperspectral imaging

In an important context such as the data analysis, the study of samples by the use of the only spectral information can be restrictive. In fact, despite the impressive improvement from a technical and instrumental point of view for the different acquisition methodologies [1–4], it is important to highlight a constraint: a bulk spectroscopic analysis can produce only a mean spectrum-based average measurement of the observed sample. This clearly leads to a non-representative local analysis of heterogeneous samples and objects, and so, to the complete loss of the information related to the spatial distribution of the different constituents [5–11]. Indeed, besides the chemical information contained in a specimen, nowadays the interest in the spatial structure and distribution of the composition of the sample of interest has a crucial role. Implementing the scientific investigation and acquisition of a sample by the use of not only the bulk spectroscopic techniques, but also, and most importantly, by the use of the spatial information coming from the specimen is currently fundamental [12–16]. Clearly, the easier way to obtain information from the spatial point of view is the use, and so the investigation, of an image (a picture) of the analyzed sample. By definition, an image is a two-dimensional graphical depiction of a subject, normally a physical object. From a technical point of view, an image is composed by pixels, where a pixel is the smallest element in a raster image. By literature, a first attempt to exploit this kind of analysis was obtained by the use of grayscale images [17–21]. For this specific kind of matrix, each pixel contains only one channel, based on a precise amount of light coming from that part of the image, and so, it can carry only the intensity information related to that pixel. Afterwards, the interest in using pictures as specimen led to the use of color-based objects, capturing images using three filters centered on red, green and blue (RGB) spectral domains [22–25]. Nevertheless, the use of these few channels cannot be comparable to the vastness of information captured and described by the use of a wider interval such as the one represented by the whole spectral domain, containing hundreds or even thousands of spectral variables, which can range from frequencies of a few Hz up to very large values. In other words, different spectroscopic techniques, each of them focusing on a specific spectral domain (ultraviolet, visible, infrared, etc.), will excite the sample with a specific amount of energy, so that different effects (from molecular, ionic, atomic, etc. point of view) will be generated and therefore, observed [26]. In simple terms, although the possibility to observe external attributes such as surface texture, defects, color, shape of the sample, the chemical composition cannot be captured with these kinds of images, due to the lack in spectral information. In the last decades, from an engineering point of view, the interest in fusing spectral and spatial information has enormously increased. Finally, after the first attempts, an evolution in the use of image analysis was achieved. Nowadays in fact, it is possible to acquire images that contain not only few channels, but instead the complete spectral domain of a selected range. Clearly, we are talking about the well-known hyperspectral imaging technology [14,27,28]. The main difference referring to this kind of acquisition methodology is that a hyperspectral image is a spatial picture of a sample in which each pixel contains a spectrum of a series of contiguous wavelengths, and not a single value. Undoubtedly, the amount and the importance of information that can be obtained using this kind of acquisition system (from both the spectral and the spatial point of view) is enormously larger compared with the one observable with other spectroscopic techniques, leading to the necessity of new ways of interpret the acquisition outcomes. Particularly, dealing with very large amount of data can be usually counterproductive, limiting the possibility of maximize the total extractable useful and meaningful information coming from the raw data. Nowadays, one of the main ways to face this issue is based on the use of chemometrics, which has shown to be a very useful discipline, able to help in this complex task [29]. In fact, using different tools related to this field it is possible, for instance, to reduce the dimensionality of the data, to extract only the most important information, and optimize the obtainable results. Despite this, the applications of chemometrics in the hyperspectral imaging area are nowadays still limited. This is the reason why a rising interest in overcoming the common restrictions related to this technique, finding new ways to couple the various chemometric methods with this particular kind of matrix, is constantly in the spotlight of many research lines. Here finally the main purpose of the present doctoral thesis: exploiting different algorithms and tools of chemometrics, trying to exceed the routine applications and particularly, improve the way to investigate the results coming from big datasets, also when various spectroscopic techniques are used simultaneously.

1.2. Hyperspectral image characteristics

From a general point of view, an image is a two-dimensional representation of an object. Normally, referring to a picture, the two spatial dimensions related to it, and so the number of pixels in the two directions, are represented by the letters *x* and *y*, respectively associated to the horizontal (the rows) and the vertical (the columns) directions of the image. The main and more evident difference between a simple image and a hyperspectral image is the extent of the third dimension. In fact, while in the first case only one single value, based on an intensity, is associated to each pixel, a hyperspectral image is a three-dimensional cube, in which this new side of the matrix, normally characterized by the letter *λ*, is related to the spectral information. More specifically, *λ* will represent the spectral range (e.g., wavelengths) related to the used spectroscopic technique in the acquisition. By way of example, a representation of a hyperspectral image is shown in Fig. 1:

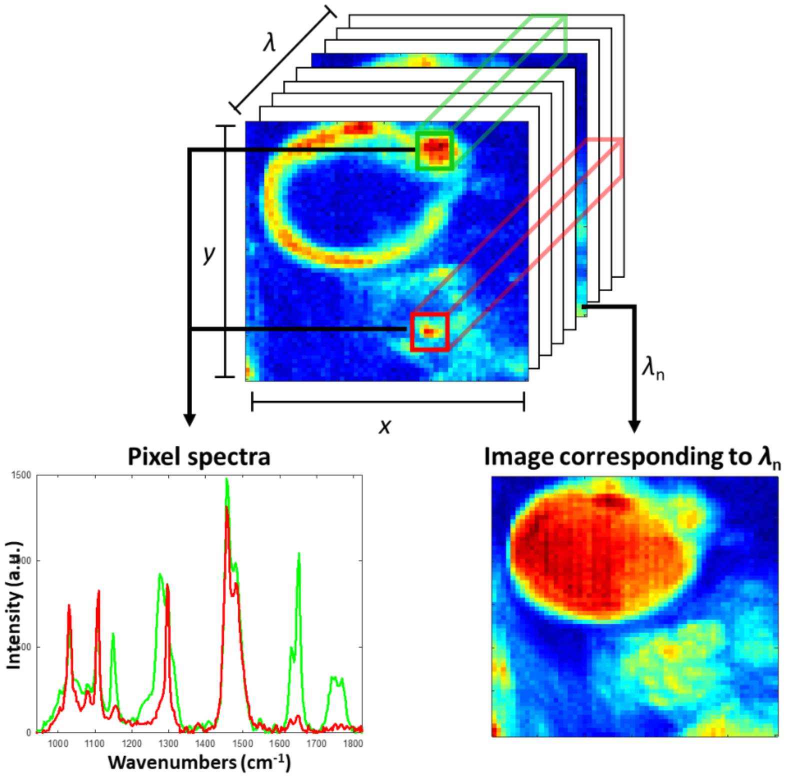


Fig. 1 – Schematic representation of a hyperspectral image. Here, the spatial dimensions of the image are represented by the directions labelled as *x* and *y*, while the spectral information is related to *λ*.

In general, the hyperspectral image can be observed both as an image at each single wavelength *λ* or as a spectrum, at each individual pixel (*x* and *y*). It is clear how this technique can be useful in many areas. Starting from remote sensing, which has been the first investigation field in which hyperspectral imaging was applied [11,27,30–33], mineralogy [34–37], food [14,28,38–40], forensic [41–44], medical [13,45–49], pharmaceutical [50–54], and biological [55–58] analyses are some of the most important research example areas related to the use of this discipline. This thanks to the peculiarity of the hyperspectral image analysis that leads to the simultaneous use and investigation of both the spectral and the spatial data related to a particular sample, providing decisive and precise information. In fact, while a simple image can provide only physical characteristics of the represented object, a hyperspectral image provides at the same moment the spatial and the spectral information. When a bulk analysis is performed by spectroscopy, it is possible to characterize homogenous materials, providing an average spectrum of the sample. Contrariwise, if the specimen is inhomogeneous, this bulk analysis could lead to non-representative information. In order to avoid this situation, the measurement should be repeated many times in a systematic way, acquiring the spectra from several positions of the sample. Clearly, this procedure is not practical in a real context. Hyperspectral imaging spectroscopy easily overcomes these limitations, identifying and quantifying the chemicals of the sample, as well as the precise location and spatial distribution.

1.3. Instrumental perspective

From a general point of view, three different hyperspectral image-acquisition approaches can be used, all of them presenting pro and cons: the point scanning (or mapping), the line scanning, and the area scanning methods, reported in the Fig. 2. The first one, also called whiskbroom approach, measures the complete spectrum of a single position (the pixel) at the time (Fig. 2a). Then, the sample is moved and another spectrum is collected for this new position. This procedure is iteratively repeated until the whole surface is captured. Clearly, a grid is defined a priori, in order to create a map based on the different acquisition points that compose the surface of the sample. The main advantage of this technique is that all the points pass for the same path of the optical system. In addition, this approach is very convenient for analyses in which is necessary to find out minor compounds. The con is that this kind of acquisition turns out to be very slow, particularly if a large area of the sample has to be explored. The second method, also called pushbroom configuration, is an extension of the previous one. The main difference is that in this case, not a point but a whole line of the image is acquired each time (Fig. 2b). This is possible using a two-dimensional dispersing element and a two-dimensional detector array. This kind of technique results to be very practical, faster than the previous one, and versatile. Usually, it can be used for food and industry applications, in which the samples are scanned using a control chain. At last, the third method, also called staring imaging, show an evident difference, compared with the previous methods. In this case, the whole image is acquired, but one spectral band per time (Fig. 2c). It means that it is not necessary to move the sample, because the whole scene is scanned in one shot, but on the other side, this procedure is not advisable if the number of needed wavelengths is too large.

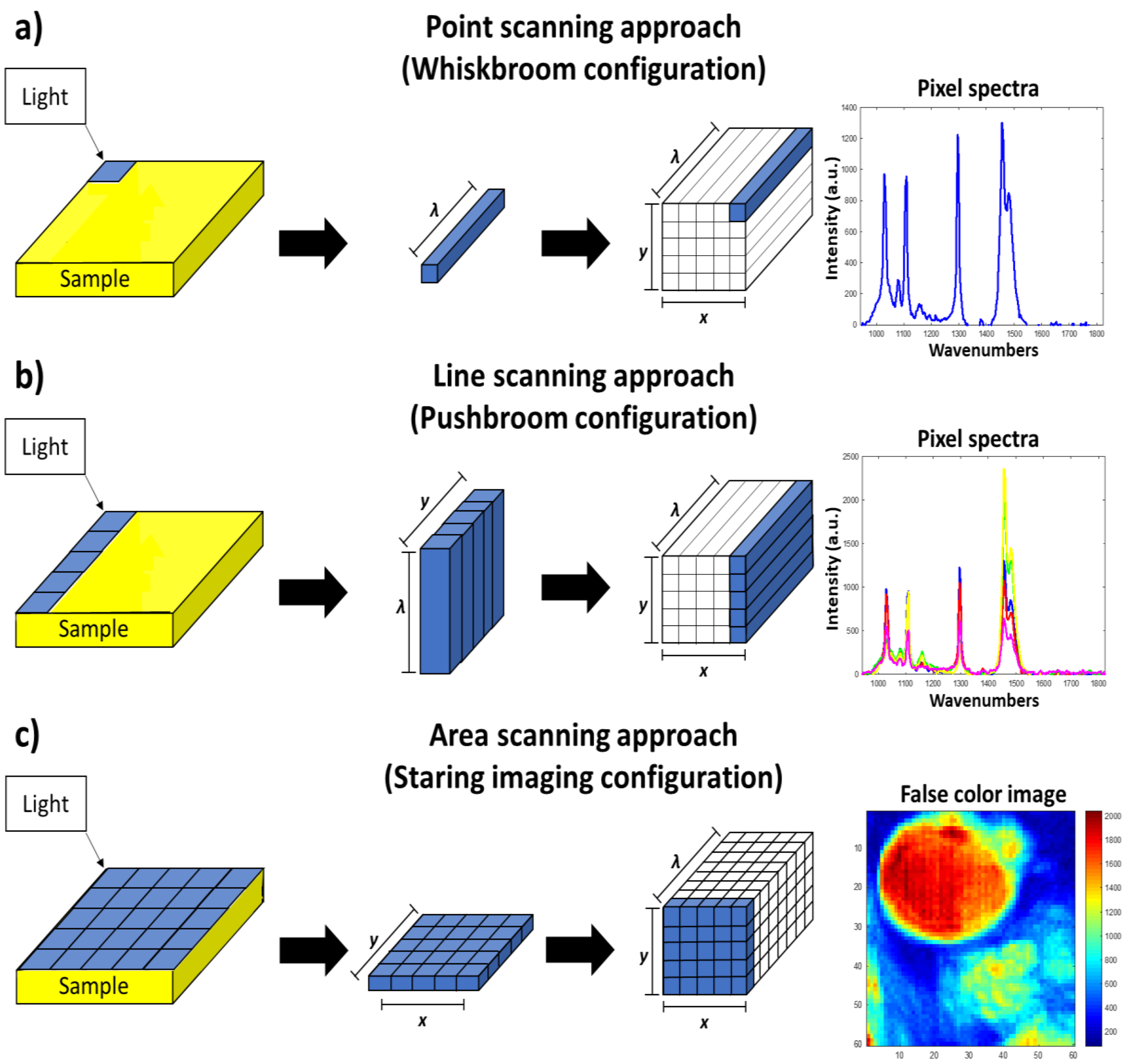


Fig. 2 – Schematic representation of the different hyperspectral imaging acquisition approaches. a) Point scanning approach. b) Line scanning approach. c) Area scanning approach.

Beyond these considerations, it should be noted that these three acquisition modes are not always feasible for each spectroscopy, mainly due to instrumental reasons. By way of example and taking into account the previous methods of acquisition in the infrared region, it is also possible to distinguish three different modalities, based on the disposition of the light source and the optical unit in a given spectroscopic equipment: the reflectance, the transmittance, and the interactance, as shown in Fig. 3. The first acquisition mode, that is without any doubt the most used nowadays, is based on a reflectance phenomenon (Fig. 3a). The second one, in which light source and detector are on the opposite sides of the sample, can show limited applications, because the light needs to penetrate and go through the specimen (Fig. 3b). Finally, the third method is a combination of the previous ones (Fig. 3c), in which both the light source and the detector are on the same side of the acquisition system. Nevertheless, a light seal is needed to avoid any interference coming from exterior light. In any case, from a general point of view is always advisable to avoid phenomena that could invalidate the quality of the acquisition, like refraction, specular reflectance and scattering.

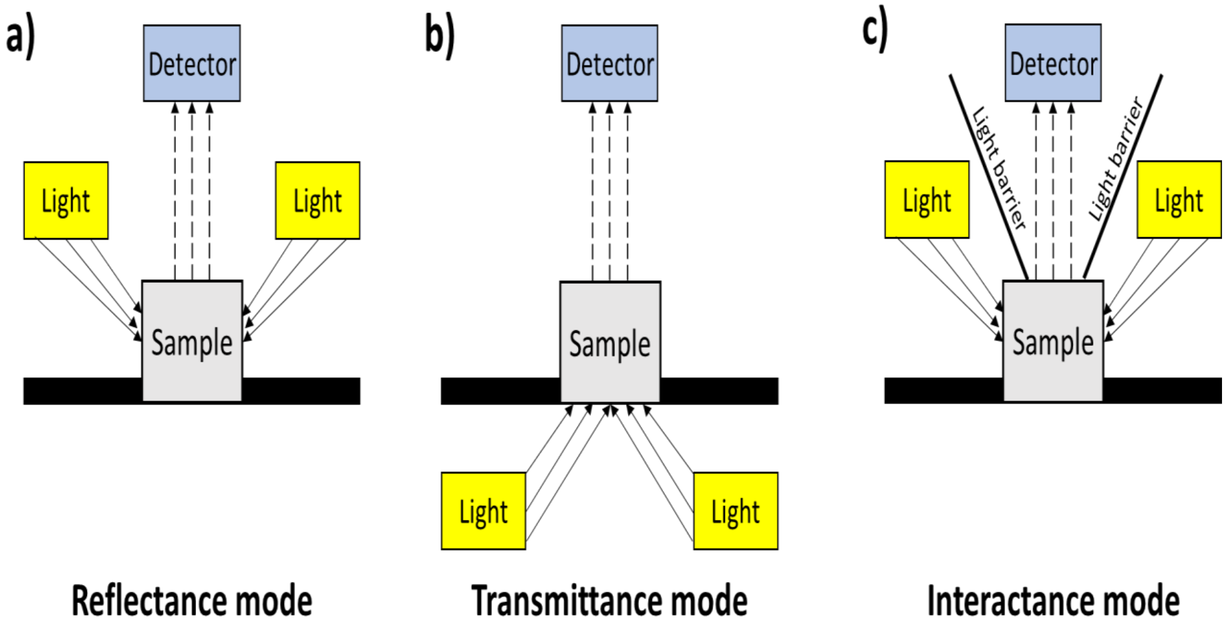


Fig. 3 – Schematic representation of the different modes to generate a hyperspectral image. a) Reflectance mode. b) Transmittance mode. c) Interactance mode.

Due to the wide interest in hyperspectral imaging, a rapid evolution in the various spectroscopic instrumentations related to this kind of technique is obvious. Nowadays, it is possible to acquire a hyperspectral image with different spectroscopic systems. Despite this, the description of all these techniques is not one of the main purposes of this doctoral thesis. This is the reason why only a brief illustration of the most interesting spectroscopies used during this period is here reported. In addition, as information, it is important to highlight the fact that the most of the datasets used during this PhD for the various works described in this manuscript were not directly acquired from our group. In fact, the different data were collected in the framework of collaborations with other research groups, while for this PhD work, only the chemometric approaches were studied and carried on. The only exception is represented by a specific dataset discussed in Chapter 2, related to the exploration of biological samples using a synchrotron beamline facility in Paris (France), namely the SOLEIL. Specifically, this has been a collaboration between our group, the team of the synchrotron DISCO beamline, and the INRAE group of research in Nantes (France). More details will be given in the corresponding aforementioned section of this manuscript.

1.3.1. Infrared (IR) spectroscopy

IR spectroscopy is without any doubt one of the most common and useful instrumentations applied to many analytical investigation areas that shows the perfect combination with the chemometric methodologies [59–61]. This kind of technique measures the interaction of infrared radiation with matter by different ways, i.e., absorption, scattering, or reflection for the study of chemical substances and functional groups. The electromagnetic spectrum of the IR region is very vast, which is why normally one can distinguish among three different spectral subregions: the near-, mid-, and far- infrared (respectively: NIR, MIR and FIR), acquired by different instruments. From a general point of view, the first one is related to overtones, or combinations of molecular vibrational modes; the second spectral domain is dedicated to fundamental vibrational modes; finally, the last region is associated to low vibrational frequencies mainly observed in minerals and crystals. The most interesting aspect of this spectroscopy is that the different molecules can absorb the frequencies generated in the IR region characteristics of their structures. This phenomenon corresponds to the possibility of observing spectra that show bands able to distinguish various structures that can be compare and recognized by the use of libraries containing the specific fingerprints of different chemical functional groups. Here, an important aspect has to be stressed. Compared with the other IR regions, NIR shows broadened bands, very informative, but at the same time hardly interpretable. This is the main reason why this kind of spectroscopy has been underestimated for years, before the introduction of chemometrics as a routine tool to study this spectral response [62,63]. Naturally, in the last decades, abreast of the evolution of the modern instruments, new devices able to make full use of this important spectroscopy coupled to hyperspectral image analysis were developed and used in different areas [51,64–66]. An important aspect to be discussed regarding IR spectroscopy, which is a dispersive spectrometer, is that it can normally measure the intensity over a narrow range of wavelengths at a time. This is why modern instruments are based on the use of the Fourier Transform (FT). In fact, Fourier Transform Infrared spectroscopy (FTIR), allows the simultaneous acquisition of high-resolution spectral data over a wide spectral range. A general scheme of the FTIR spectrometer is reported in Fig. 4:

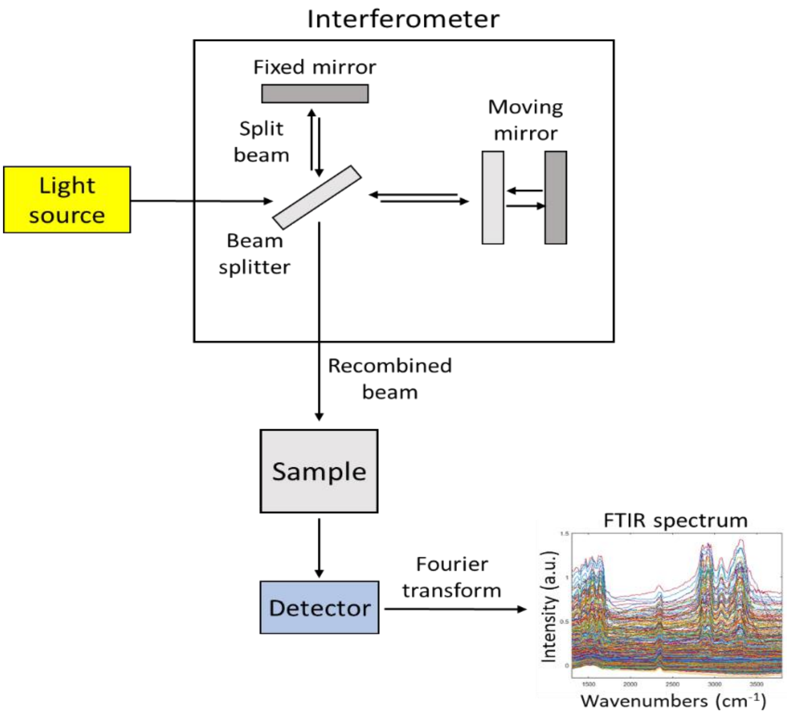


Fig. 4 – Scheme of a FTIR instrument.

If normally absorption spectroscopy measures the quantity of light absorbed by a sample at each wavelength using a monochromatic light beam, FT spectroscopy works in a less intuitive way. Instead of using a monochromatic light, which uses a single wavelength at a time, this technique generates a beam containing many frequencies in one shot, measuring the amount of light absorbed by the sample. The possibility of interpreting this complex signal is given by the use of an interferometer. This kind of instrument contains a beam splitter and two mirrors, one fixed and the other one moving. The incoming light is at first split into two equal quantities, directed to the different mirrors. The moving mirror, shifting, introduces an optical path difference, which will generate coming back to the splitter, a constructive or destructive interference with the part of the ray reflected by the fixed mirror. In this way, it is possible to obtain an interferogram that shows the representation of the intensity in the time domain for a specific signal. Then, using the FT, it is possible to pass from this domain to the corresponding frequency one, generating the corresponding IR spectra that can be interpreted in the investigation analysis.

1.3.2. Raman spectroscopy

Raman phenomenon was detected the first time in 1928 by the Indian physicist C. V. Raman and K. S. Krishnan [67,68]. Compared with other spectroscopic techniques, as infrared, it is not based on the absorption of photons but on the light scattering effect correlated to the vibrational energy state of the molecules. For many decades, this kind of response could not be used due to the weakness of the corresponding signal. In fact, when a light source is used to excite a sample, as shown in Fig. 5, it is possible to distinguish different responses:

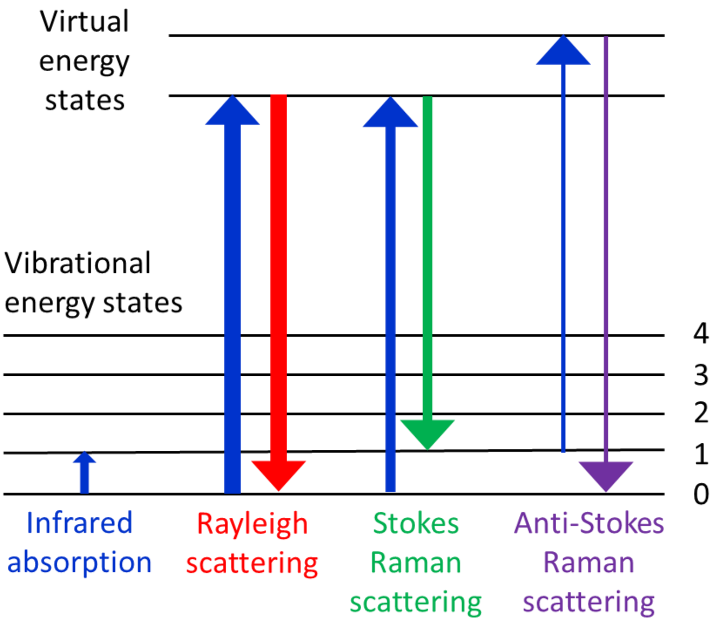


Fig. 5 – Different vibrational energy responses.

Excepting the absorption phenomenon, an overwhelming majority of the scattered photons show the Rayleigh scattering effect, in which the energy intensity of the incident light is equal to the one of the scattered light. More precisely, only one of a thousand or ten thousand of the scattered light (that is anyway a thousandth of the initial incident light) will correspond to the Raman effect that in other words, represents only a millionth of the incident light. Due to this intensity weakness, this kind of spectroscopy originally did not obtain the right interest. Only in the early 1960s, when laser was introduced as excitation source (high radiation intensity), this kind of instrumentation was recognized as one of the most important tools for many different research areas. This is true particularly because nowadays many libraries containing the fingerprint of various compounds exist, as well as for other spectroscopic techniques (e.g., FTIR), driving to an easier chemical interpretation using this kind of spectroscopy. Considering hyperspectral image analysis, nowadays many different studies based on this technique are available [69–78]. In addition, the use of visible photons, linked to lower diffraction limits, can lead to a better spatial resolution and so, to an increase of the exploratory potential of samples. Regarding the technical characteristics of this spectroscopy, a scheme is represented in Fig. 6. Formerly, a mercury vapor lamp was commonly used. The limitation of this kind of source is that it has many strong bands that could lead to partial overlapped spectra, if a filter is not applied to select a particular emission frequency.

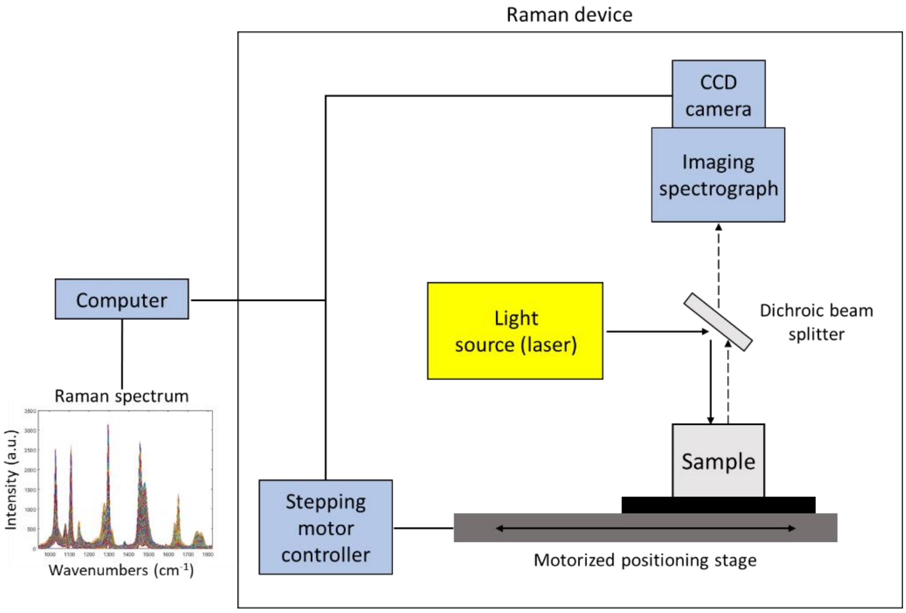


Fig. 6 – Scheme of a Raman instrument.

In modern Raman instruments, as already described, a laser is usually used as emission light [79]. Particularly, a crystal of Nd:YAG (neodymium-doped yttrium aluminum garnet) is the most common solid-state laser source. Moreover, also laser diodes are gaining importance, particularly for their emission power and the different obtainable emission wavelengths. A dichroic filter is used in order to make selectively pass only one fraction of the light that will hit the sample, generating the scattering response. Then, this light is collected passing through a spectrometer, which disperses the light into a spectrum. Normally, a CCD (charge-coupled device) detector is used to record the final spectra.

1.3.3. Energy-Dispersive X-ray (EDX) spectroscopy

This kind of spectroscopy is an analytical technique that can be used to probe the elemental composition of solid materials [80]. Considering quantum mechanics, depending on the observed element, an atom consists of different energy levels, each of them containing a certain number of electrons spinning around the orbit of the core. In detail, when a surface is properly excited, an electron from the first level (the closest to the core) can be expelled, leading to a drop of more distant electrons to fill the resulting ‘holes’ around the center of the atom. The principle behind this methodology relies on the transition of electrons from higher energy levels to the ones close to the core of the atom. Transitions between energy levels follow the law of conservation of energy. Excitation of an electron to a higher energy state requires an input of energy from the surroundings, and relaxation to a lower energy state releases energy to the surroundings. Specifically, this change into the structure of the atom generates a set of X-ray emissions at different frequencies, specific for each element, allowing the possibility of a qualitative analysis. Generally, two different methods are the most commonly used to excite the core electrons. The first one uses a high-energy electron beam, which is produced by an electron gun. Another possibility is the one of using X-rays instead of electrons, to excite the core electrons to the point of ionization. No matter the excitation source, the purpose using this energy is to excite core electrons to high-energy states, creating a low-energy vacancy in the electronic structure of the atom. This phenomenon leads to a cascade of electrons from higher energy levels, in order to recreate the minimum-energy state of the atom. Due to the conservation of energy, the electrons emit X-rays in the moment that they transit to lower energy levels. The interesting aspect of this spectroscopy is that since each element has a different nuclear charge, the energies of the core shells and the spacing between them vary from one element to the next. Giving sufficient resolving power it is possible, using the EDX, to determine the composition of the sample based on the observation of the characteristic peaks. Nevertheless, some limitations are evident. First, not every peak in the spectrum of an element is exclusive to that element, and this is the main reason why all the peaks need to be matched with preexistent libraries and using standards. Also important is the fact that a combination of elements can act differently than a single element alone, leading to the necessity of knowing the general composition of the investigated sample. Another limitation is related to the impossibility in observing elements lighter than boron, which represents a problem due to the natural abundance of hydrogen in materials. It is also important to consider the fact that EDX needs to be coupled with a microscopy such as the Scanning Electron Microscope (SEM) to provide both the spectral and spatial information of a given sample. Secondary electrons may cause additional excitation and emission of spectral lines, generating the possibility of overlap with the lines related to real elements. Lastly, sample needs some preparations. EDX is a near-surface technique, so the specimen has to be exempt from any trace of grime, to avoid false results. Furthermore, the sample must be stable under vacuum because the instrument works in an environment preventing the presence of any atmosphere, which could interfere with the electron beam. Nevertheless, this instrument is nowadays used for many different purposes, as shown in literature [43,81–84]. A general representation of the EDX instrument is here reported, in Fig. 7:

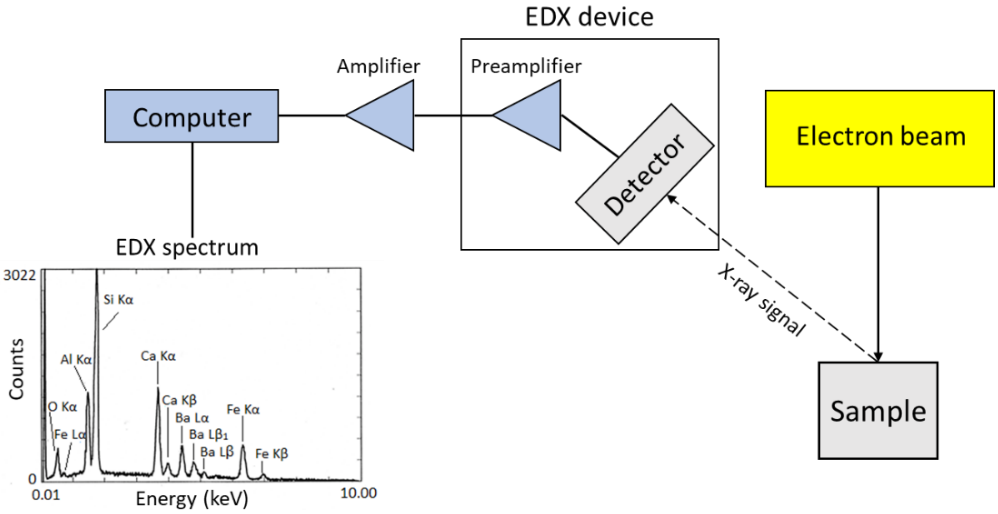


Fig. 7 – Scheme of an EDX instrument.

1.3.4. Laser-Induced Breakdown Spectroscopy (LIBS) and Plasma Induced Luminescence (PIL)

In the framework of the elemental analysis, it is necessary to focus especially on one instrumentation. LIBS is a very suitable spectroscopy that shows many advantages, if compared with other techniques such as atomic absorption, inductively coupled atomic emission, X-ray fluorescence, etc. In fact, despite the interesting detection limits and accuracy, these methods require a complex sample preparation and a long detection time. Furthermore, these spectroscopic techniques are destructive compared with LIBS, in which only a small portion of the sample is ablated. Despite the initial interest in the early 1960s, this spectroscopic technique started to be really in the spotlight after the 1980s, due to an increasing development of the used laser and detector technologies. Considering the main merits of LIBS, it is important to stress some aspects. Besides the absence of a sample preparation and pretreatment, this analytical technique is very fast, performing acquisitions within a fraction of a second and allowing a multi-elemental analysis. It is also sensitive to light elements, which are not observable with other techniques, and can be used for analysis for all states of matter. Last but not least, LIBS can be coupled with other analytical techniques, e.g., Raman spectroscopy, to obtain simultaneously multi-elemental and molecular surface analysis. Of course, some weak points are observable. The limit of detection, which is in the range of the part per million, can be limitative compared with other techniques that reach the part per billion. Furthermore, a self-absorption phenomenon can occur. Specifically, emissions from hotter regions can be absorbed by the colder atoms surrounding the plasma, affecting the spectral intensity of the signals and so the quantitative analysis. Another limitation is represented by the matrix effect. Depending on the nature of the sample (from both the physical and chemical perspectives), it can affect the ablation phenomenon and so the quality of the final spectral signals. Nowadays, LIBS is applied for the analysis of a wide range of materials [85–89]. A scheme of this instrumentation is shown in Fig. 8. A pulsed laser, normally generated using a crystal of Nd:YAG, is used to ablate a minute amount of material from the sample surface. The ablated mass will produce a vaporous plume on the surface of the sample. The interaction between the pulse laser beam and the plume will generate a plasma, which will prevent the beam from entering into the sample in a process named ‘plasma shielding’. In addition to stopping the ablation from the surface of the sample, this phenomenon will generate an increase of temperature that will ionize the plasma. In this way, a luminous plasma is generated. The phenomena of excitation, de-excitation, expansion and condensation for each species in the plasma plume will produce electromagnetic radiations that contain information about the different species present in the sample. Eventually, this light is collected and directed to a spectrometer by the use of an optical fiber, in order to generate a spectrum that will facilitate the element detection. In fact, a good aspect in LIBS spectra is that each element shows specific emission wavelengths, leading to an easier way to recognize and so, identify a specific atom, making this kind of spectrometry a very suitable technique in many scientific areas.

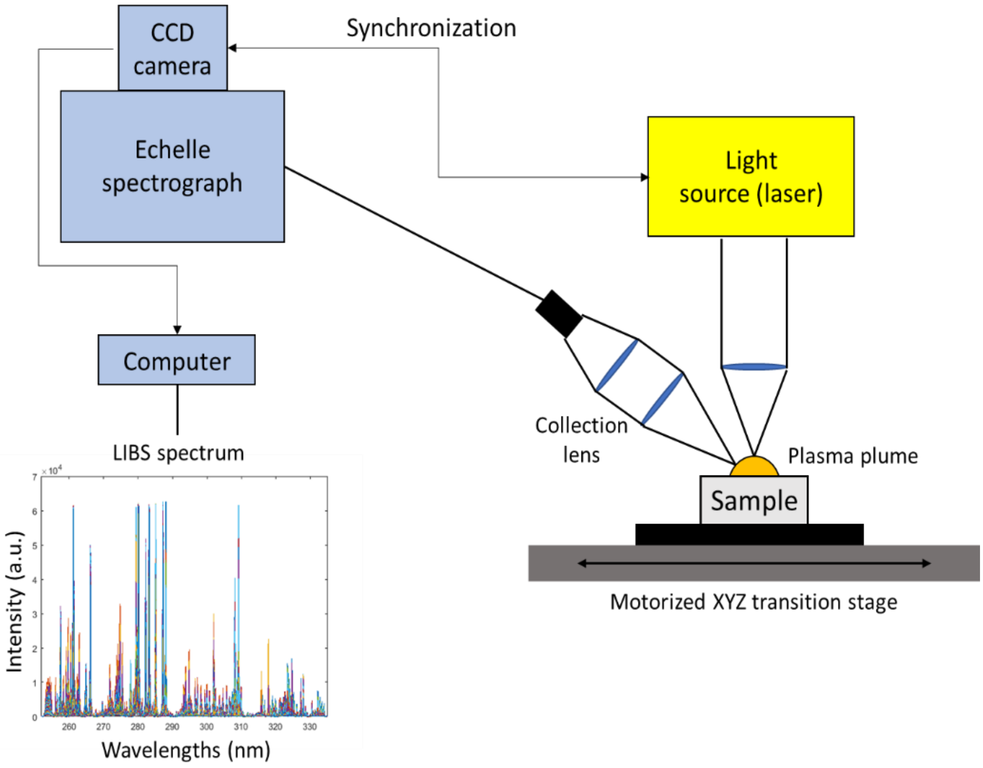


Fig. 8 – Scheme of a LIBS instrument.

More recently, another interesting phenomenon that can occur when LIBS is used as excitation source has been observed. In fact, using the same instrument, the plasma generated by the LIBS laser shot can act as an excitation source and produce the emission of a luminescence response for specific elements present on the sample surface, with a delay of some milliseconds. This kind of phenomenon is called Plasma Induced Luminescence (PIL) [87,90–92]. Clearly, the possibility of obtaining distinct chemical answers using the same equipment is a very interesting aspect, from a research point of view. Nevertheless, nowadays the ability to interpret this kind of phenomenon is still limited, leading to the interest in deepening this chemical response. Particularly, as it will be shown in the Chapter 4, one way to extract more information related to PIL, taking the advantage of using the same instrument for the acquisition, is the data fusion between this kind of response and the LIBS spectra. In fact, using this approach, it seems possible to obtain details and correlations that otherwise would not be observable, when only PIL spectra are investigated. Lastly, LIBS can also be used in order to acquire simultaneously different spectral ranges. Particularly, it is possible to use the same instrumentation, without any necessity of changing the platform, for the acquisition of, for example, also Raman spectra. Clearly, this is another very interesting aspect because in this way it would be possible to obtain at the same time elemental and molecular information, respectively from LIBS and Raman spectroscopies, leading to more accurate and interesting results from the analysis. By way of example, it is plausible that some elements or compounds can be detectable using exclusively one spectroscopy or the other one. Therefore, using a data fusion approach, it would be possible to generate more details that otherwise will be missed. For informational purposes, also this argument will be discussed more in detail into the present manuscript, in the aforementioned dedicated chapter.

1.3.5. Synchrotron beamlines and associated spectroscopies

Synchrotron radiation is the term used to describe an electromagnetic radiation emitted by a charged particle beam in a circular accelerator. It represents nowadays a very interesting excitation source for chemical and biological investigation purposes, and recently it has been also applied to hyperspectral image analysis [93–96]. The principle of this phenomenon can be explained by the equation of Maxwell, based on the notion that changing the charge density, it is possible to radiate electromagnetic waves. Throughout the decades, more and more facilities in which this kind of technology is used have been built all around the world. A synchrotron, whose scheme is reported in Fig. 9, is an accelerator of electrons where they are trapped and forced to travel at the light speed along a circular path with a constant radius in a bending magnet. The acceleration of the electrons gives rise to the emission of radiations, emitted in discrete quanta or photons, each of them with an energy depending on the frequency of the radiation itself. Despite the interest in using this beamline, many cons can be highlighted. First, building a synchrotron facility is related to very high-cost machines. Also, the storage ring in which the particle beam is kept circulating in to create the photons must show some characteristics, in order to generate a relevant electron flow. In fact, a loss of energy due to the emission of synchrotron radiation is normal, and the radiofrequency must provide a sufficient power to accelerate the electrons. Another important factor to be highlighted is the possible collisions with the walls of the ring, if the radiofrequency is too intense. This is a very limiting factor in the construction of a circular electron accelerator. An option would be to reduce the bending magnet field strength, but this means to build larger instruments, and so an increase in the costs.

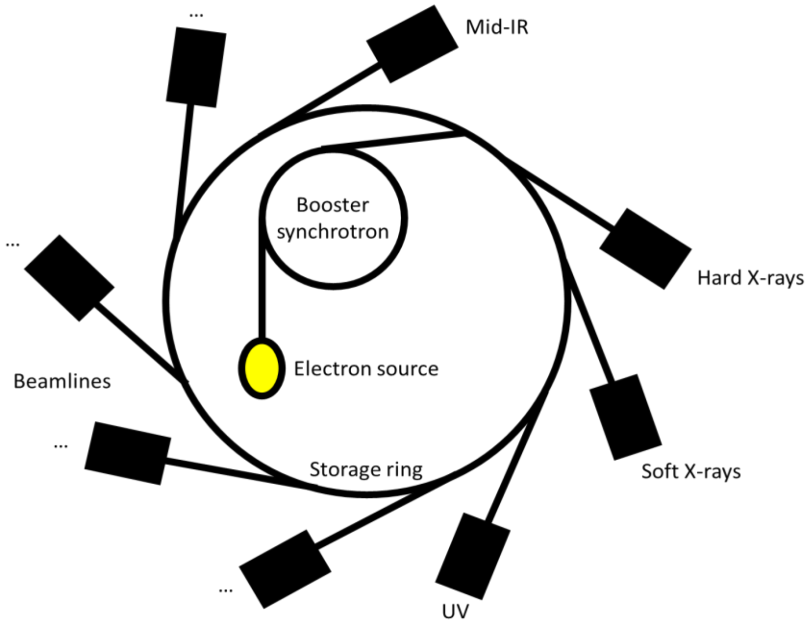


Fig. 9 – Scheme of a synchrotron beamline facility.

Another essential aspect to be considered is the vacuum system, used to obtain an optimal beam lifetime. The power of the synchrotron radiation can be very high, and so water-cooled absorbers must be provided. The back bombardment of ions generated from residual gases (created by the photons hitting the vacuum chamber) can desorb gas molecules from the surface, and so decrease the lifetime of the beam. Related to the energy of the photons is also the possibility of generating high-energy particles that can escape from the circular trajectory and thus, damage the ring components. For this reason, some precautions have to be taken into account, i.e., the use of special alloys covering the internal walls of the ring, or the use of extra magnets to shield.

Nevertheless, the synchrotron beamline is nowadays a very suitable technique for the acquisition of hyperspectral images with high resolution for the investigation of the whole range of basic and applied sciences, for samples of different nature [97–99]. Some of the properties making the synchrotron radiation so attractive are the high intensity or photon flux, and the fact that a continuous spectrum covering a broad range from the far infrared to hard X-rays is obtainable. Normally, this is possible because many different beamlines are built around the ring, each of them used for the acquisition of a specific spectral response (e.g., diffraction, X-ray absorption, crystallography, autofluorescence etc.). Regarding this kind of radiation source, it is important to highlight the fact that during this PhD thesis it has been possible to collaborate with the national synchrotron facility in Paris, namely the SOLEIL, using the DISCO beamline acquiring biological samples to observe the phenomenon of autofluorescence coming from excitation using UV and visible spectral ranges. More details will be given in the Chapter 2 of the present manuscript.

1.4. Methodological perspective

So far, in this manuscript, hyperspectral image analysis has been described from a general point of view, especially highlighting its main strong features compared with the use of the classical bulk spectroscopic acquisition methods. To give some examples, one can refer to the possibility of acquiring a quantity of spectra that will result greater than the one obtainable with a bulk analysis (in some cases up to millions of spectra) [100,101]. Furthermore, a hyperspectral image has the great advantage of showing not only the spectral information, but also, and particularly, the spatial distribution of the components in the acquired sample [15,102–104]. This is a very interesting aspect, considering the fact that if the specimen is heterogeneous, a routine analysis could lead to non-representative results, and so, to wrong analysis conclusions. Nevertheless, hyperspectral imaging is still a very recent methodology and, despite its very promising and rising characteristics, some limitations and constraints are clearly unavoidable. In fact, dealing with a too massive amount of information can be challenging, if not impossible. In addition, despite the introduction of the spatial information, currently the use of this kind of details is still very limited, when compared with the spectral domain. Without any doubt, chemometrics is nowadays one of the most interesting approaches that perfectly matches and overcomes the limitations concerning the hyperspectral image analysis [29,100,105–108]. This discipline can in fact extract the most meaningful information from the massive quantity of data by the use of mathematical and statistical methods. In this way, it becomes possible, decomposing multivariate complex data, to obtain more interpretable information and so, leading to the interpretation of the chemical, physical and biological aspects of the sample. Nevertheless, it is also important to highlight the fact that normally the analysis of the raw spectral data is not recommended and counterproductive. Data need to be pretreated in order to extract the most important information in the right way, giving the possibility to all the details to be observed and correctly used. This is the reason why before any analysis, usually the matrix is treated with some pretreatments with the aim of leading to more interpretable results [109], as it will also be briefly discussed in this manuscript. Multivariate data analysis, that is the core of the chemometric approach, is able to show the hidden chemical information of the investigated specimen, showing important details that otherwise would be missed using more simple and traditional approaches [110–112]. Nowadays, a vast quantity of methods can be used to dig into complex matrices and obtain interesting results. From a general point of view, one can discern between two big different kinds of analyses. On one side, chemometrics and so multivariate data analysis can be used for the qualitative analysis. From the other side, this discipline is also used for the quantitative analysis. A general scheme representing the main chemometric methods that are currently applied is shown in Fig. 10:

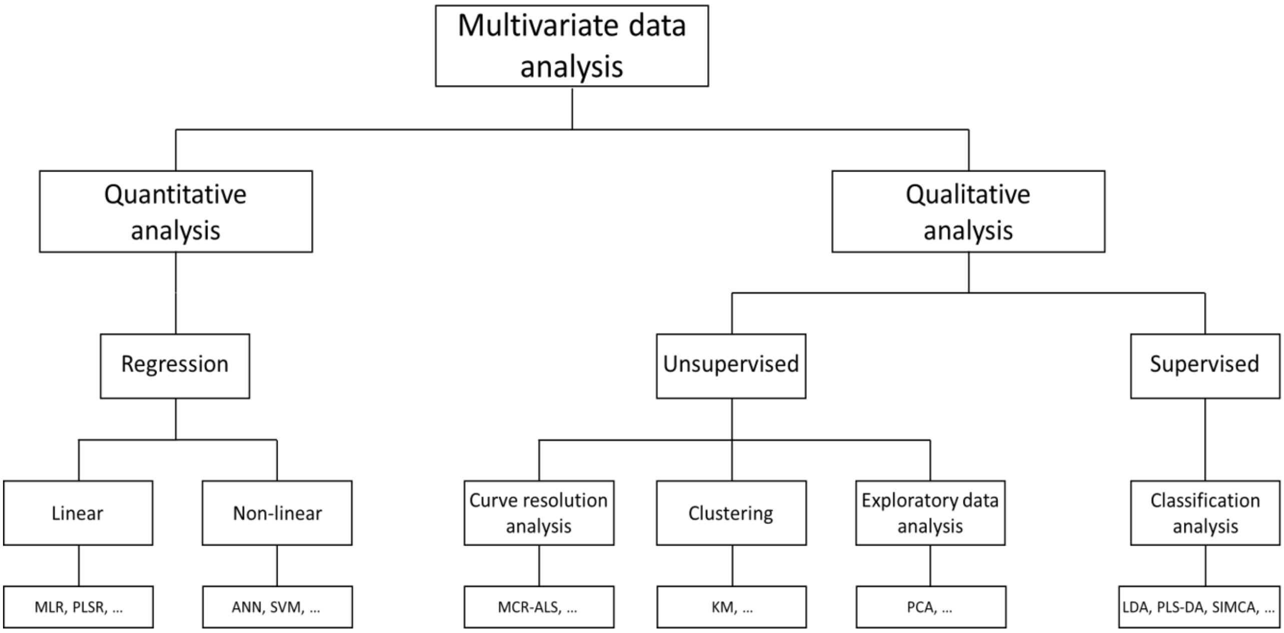


Fig. 10 – General scheme of multivariate data analysis methods. MLR: Multiple Linear Regression, PLSR: Partial Least Squares Regression, ANN: Artificial Neural Network, SVM: Support Vector Machine, MCR-ALS: Multivariate Curve Resolution-Alternating Least Squares, KM: K-Means clustering, PCA: Principal Component Analysis, LDA: Linear Discriminant Analysis, PLS-DA: Partial Least Squares-Discriminant Analysis, SIMCA: Soft Independent Modelling of Class Analogies.

Hereafter, a brief description of the scheme and so, of the main chemometric approaches is reported. Nevertheless, only the most important algorithm methods used during this work will be described in detail in the present manuscript.

Regarding the quantitative analysis, regression methods are the most commonly used [113]. The main purpose of this methodology is to find a relationship between a desired information (chemical, biological or physical) and the spectrum responses, in order to predict the numeric values coming from new data related to the ones used to build the model [114–117]. Two different kind of regression analyses can be used. The most common one is the linear method. In this approach, several explanatory variables are used to predict the outcome of a response variable using the linear relations between the spectra data and the target attributes. Different algorithms are nowadays available to face this particular task, each of them showing some differences. Multiple Linear Regression (MLR) is the simplest method in which it is normally observed the correlation between the measured variables and the response of interest. Anyway, this kind of analysis shows a limitation with regard to the robustness of the model. In fact, spectra show often a high co-linearity, leading to overfitting problems. Partial Least Squares Regression (PLSR) is a more recent linear method, which shows a better robustness compared with the previous one. Being a bilinear modelling method, PLSR creates models using a large number of independent variables (mainly predictors or wavelengths), in order to predict a set of dependent variables (concentrations or chemical information). The reliability of predictions is normally achieved by the extraction and the observation of a certain amount of Latent Variables (LVs), orthogonal factors related to the information contained in the variables used to create the model [118]. From the other side, when spectral data and target attributes are not linearly related, non-linear approaches can be applied [119]. Interesting methods related to this discipline are Artificial Neural Network (ANN) and Support Vector Machine (SVM). These methods are very suitable to deal with complex and nonlinear correlations, hard to be interpreted differently, using networks that can extract hidden information.

Referring to qualitative analysis, they are generally used with the main purpose of classify and distinguish between different categories of elements present in the same matrix. Normally, it is possible to discern among a massive number of different methods. A first differentiation can be made between two main groups: supervised and unsupervised methods. In the first case, the main purpose of the analysis is to create classification models able to make it possible to classify new unknown samples based on the previous classified and known measurements. This kind of methodology is also known with the name of classification analysis [120]. Linear Discriminant Analysis (LDA) is the first supervised technique used in data analysis [121,122]. Nowadays, it is still used due to its robustness, simplicity, and reliability. The concept behind this classification method is based on the assumption that the conditional probabilities for each class are normally distributed and that the variance/covariance matrices for all the classes are identical. In other words, a partition of the space among the different classes is applied by maximizing the ratio of between-class variance and minimizing the ratio of within-class variance. Nevertheless, this method is a very simple one and so, it is affected by some classification problems, particularly when collinearity between data is present, such as in complex chemical matrices. This is the reason why another method is normally applied, to overcome the limitation of LDA. Partial Least Squares-Discriminant Analysis (PLS-DA) is a chemometric approach based on the use of partial least squares regression method [123]. Traditional regression methods can be used for classification analysis, in which case the relationship between a multivariate independent vector and a qualitative vector of responses is searched. From a general point of view, the solution given by this technique is statistically equivalent to the one obtainable using LDA, being the resulting model a linear one. The difference is given by the fact that this method is related to the use of the LVs, and not the original spectra, taming the constraints described above. In fact, once the model based on the data obtained by the known samples is created, a new unknown sample can be analyzed, computing a predicted vector of responses. Then, it will be compared with the different classes present in the model and assigned to the category that show the highest similarity with the investigated sample. The previous described methods are known as discriminant techniques [124]. Another type of supervised method is represented by the modelling approaches, in which the main task is to capture the similarities among samples belonging to the same category [125]. One of the most famous and used methods in this group of chemometric approaches is for sure the Soft Independent Modelling of Class Analogy (SIMCA) [126]. Differently from discriminant methods, when using this algorithm, each class is separately considered and so, an individual model is constructed for every one of the categories present in the dataset.

As previously introduced, the other group of qualitative analysis is represented by the unsupervised methods. In this case, a previous knowledge about the different kind of samples is not necessary, and they will be classified only according to their natural groups and the similarities among the samples of the different classes. Various methods can be applied, depending on the main investigation purpose. Some of the most important ones are here reported. Exploratory data analysis is for sure one of the cornerstones of chemometrics, vastly used in many research areas. Principal Component Analysis (PCA) [127], the most representative exploratory data technique, is a method in which the spectral data are decomposed into several orthogonal factors, the principal components, which are a linear combination of the original variables of the analyzed matrix. Generally, the use of principal components will create a new set of uncorrelated data, ordered in terms of decreasing variance (the scores) that can also be used as input for other techniques, with the purpose of maximize the usable information. Based on the scaling coefficient given by the scores, each of them is related to a particular set of loadings, which contains the maximum variations common to all the spectra in the dataset. PCA can be used to extract the most important part of the details contained into the array, filtering the redundant data, and so reducing the total amount of used information. Another technique used for data exploration is the one represented by the clustering analysis. K-Means (KM) clustering is for sure the most representative method in which the clustering approach is applied [128]. The main point behind this technique is, after selecting the right *K* number of classes in a matrix, to classify each element into one of the different clusters. This procedure is applied by trying to minimize the sum of squares of distance between each spectrum and the corresponding cluster centroid. Despite its simplicity and efficiency, this method has the con of being influenced by the operator choice, because if the selected number of clusters is not right, the results will be biased. In addition, the presence of unbalanced classes and/or sub-populations can lead to not very precise results. Finally, qualitative investigation can be conducted by the use of the curve resolution analysis, also known as signal unmixing. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is for sure the most important algorithms used in chemometrics [129]. The main purpose of this kind of analysis is to resolve the mixture analysis problem by expressing the original data using a bilinear model of pure component meaningful contributions. In other words, during normal experiment conditions the spectroscopic data can be approximated using a bilinear model whose elements are directly interpretable in chemical terms. It is possible, using MCR-ALS, to perform a data unmixing of the studied matrix, and so, estimate the number of constituents in the mixture, obtaining their pure concentration profiles and pure spectral ones from the information encoded in the recorded data. Furthermore, being a self-modeling method, in principle it does not require any specific preliminary information about the data. The only limitation using this technique is given by the fact that the investigated sample must satisfy the bilinearity and that some generic characteristics of the pure concentration or spectral profiles are known.

An important aspect that has to be taken into consideration is that these methods generally cannot be applied directly on the hyperspectral image. In fact a data cube, which is represented by three dimensions (namely, x and y regarding the spatial direction, and λ for the spectral information), needs at first to be unfolded into its corresponding two-dimensional matrix. This represents an important limitation that to date afflicts the use of the full potential of a hyperspectral image. Unfolding an image, the spatial information is naturally lost, leading to the fact that these chemometric strategies will take into consideration only the spectral part of the data. Nevertheless, as it will be described into the present manuscript (more precisely, Chapter 5), some interesting methods that overcome this challenge have been recently investigated and applied to hyperspectral image analysis. Particularly, wavelet transform [130] represents a new approach, in which some particular filters are used with the aim of decompose the original signal of the image into the contribution of particular coefficients that will save the most important information related to the spatial features of the data cube.

So far, the main point of the present paragraph of this manuscript has been the one of giving a general idea of the different chemometric pretreatments, methods and algorithms that are nowadays commonly used in the investigation of the complex nature of real samples. Hereafter, a more detailed description regarding the most relevant approaches used during this PhD work will be reported.

1.4.1. Data preprocessing in chemometrics

As already described, it is of fundamental importance to use the correct pretreatment before any further analysis, in order to extract the most important information from the data, avoiding to obtain unclear results [109,131]. Here following is reported a brief description of only some of the most used and common preprocessing approaches in spectroscopy, to give a general idea of their necessity and applications.

1.4.1.1. Mean Centering (MC) and autoscaling

Despite its simplicity, MC [132] is for sure one of the most important preprocessing steps that has to be applied to a data matrix when needed, particularly in hyperspectral imaging, and more specifically in analyses such as PCA. It is an additive transformation of a continuous variable *m*. The mean of the resulting variable is zero. In other words, mathematically, MC calculates the mean of each column of the matrix and subtracts this from the column. Using this technique, the distribution of the variables will be shifted and centered to the zero, changing the scaling of a variable, but not its units. In this way, the standard deviation of an observed information will not be affected, and so, to each variable will be given the same distribution, but the relative importance will be conserved. It is also important to remember that this kind of approach is applied normally as last method, in a series of preprocessing steps. Another method related to MC is the autoscaling. In this case, the matrix is at first mean centered and then, each column is divided by the standard deviation of that column. In this way, each column of the corresponding matrix will have a mean of zero and a standard deviation of one. This method is very useful when it is necessary to correct different variable scaling and units if the main reason of the variance of the variables is related to signal rather than noise. At the end, each variable will be scaled such that its meaningful signal has an equal footing with the signal of other variables. Also in this case, it is important to consider that autoscaling preprocessing has to be used for the right type of data (e.g., wavelet transform), in order to avoid the generation of wrong magnitudes and information.

1.4.1.2. Standard Normal Variate (SNV)

SNV [133] is particularly useful to correct imprecision carried out by IR instruments, due to the scattering light phenomenon. Due to the interaction between the light and sample particles, a baseline shift can be generated, resulting in a more complicated spectral interpretation. Normally, this scattering can produce a background signal that varies with the wavelength, leading to a baseline shift and curvature, which can vary among samples. Using this algorithm, it is possible to reduce multiplicative effects of scattering and particle size, also reducing the differences in the global intensity of the signals. From a mathematical point of view, each spectrum is centered and then scaled dividing it by its standard deviation.

1.4.1.3. Savitzky-Golay (SG) derivative

SG derivative [134] is commonly used as a signal pretreatment for spectral data. Despite the fact that it is possible to use higher-order derivatives, first and second ones are the most frequently used in the analysis, because they result to be generally adequate to obtain optimal results. Normally, this kind of preprocessing is used to resolve peak overlaps, enhance the resolution and eliminate constant and linear baseline drift among the samples. Nevertheless, it is important to remember that, using a derivative approach as pretreatment, noise level of the spectra can be increased, as well as the fact that spectral interpretation becomes more complicated. From a mathematical point of view, derivatives are defined as the slope of the line (the acquired spectrum) at any given point. SG first derivative method fits a curve through a small section of the spectrum, and then finds the slope of the tangent to this curve at the central point. Second derivative can be computed directly from the first one. It corresponds to the slope of the first derivative, generating new peaks in correspondence of the less interpretable zones of the first derivative results, leading to the possible observation of signals hided during the first part of the calculation. Also important is that if the spectra are preprocessed in a too extreme way, artifacts could be generated, leading to a general misinterpretation of the spectral information.

1.4.1.4. Baseline correction using Asymmetric Least Squares (AsLS)

Lastly, another important kind of preprocessing method that can be applied to many different spectroscopic data is the correction of the baseline, due to a possible offset that can be generated during the acquisition linked to instrumental problems or specific photon-matter interactions. Despite the existence of various methods (that show different pros and cons), here is reported only one approach, which provides an automatic baseline correction that overcomes many of the limitations related to other procedures. This method is based on the Asymmetric Least Squares (AsLS) algorithm [135], using the well-known Whittaker smoother, in which the baseline offset is automatically removed by the use of a piecewise method, to get a slowly varying estimate of the baseline. This method results to be very interesting due to many reasons. Particularly, compared with other approaches, it is relatively fast, and only two parameters are required to obtain a suitable baseline, with completely reproducible computations: one is needed to tune the flexibility of the baseline, and the other to adjust its position. Using this approach, once given a signal, it will be combined with a series that has to follow two properties: be smooth and be faithful to the used given signal. These two goals can be combined by minimizing a penalized least squares function in which the fit to the data and a penalty on non-smooth behavior of the series are measured. From a general point of view, while a light smoothing will remove noise, a strong one gives the slowly varying trend of a signal. Nevertheless, when using this smoother, it is also important to use a parameter to compute the obtained weights to the residuals (based on the principle of asymmetric least squares) that otherwise will be both positives and negatives. The resulting equation is complex, based on the mutual interaction of weights and smooth curve. Despite this, it can be transformed into iterative application of two easy computations until the moment a convergence is obtained. The two used parameters are respectively p for asymmetry, and λ for smoothness, chosen by the operator.

1.4.2. Principal Component Analysis (PCA)

Without any doubt, PCA [127,136] is one of the most commonly used approaches applied in every research area, when chemometrics is needed. As already described, it can be used as an exploratory analysis to obtain a general idea of the information contained in the investigated specimen. In addition, this kind of method allows the compression of the original matrix, making possible the description and interpretation of large sets of multidimensional data. From a general point of view, it is possible to describe a matrix **D** as sized *n* x *m*, where *n* denotes the different objects in the matrix and *m* the number of variables (normally the wavelengths, in spectroscopy) registered during the analysis. As already introduced above, modern instruments can acquire huge amount of data, using very large spectral intervals, and it is particularly true when hyperspectral imaging is the used technique, where thousands, hundreds of thousands, or even millions of spectra can be obtained. Very easily, the meaningful information in **D** can be related to correlations among different variables over the whole set. PCA can be a very useful solution in this kind of situation. In fact, it is possible, using this approach, to reduce the original *m*-dimensional space of the variables into a new subspace with a lower dimension of size *a* (based on the correlation of the initial acquired information) in which all the *n* samples can be projected and represented as new points. From a mathematic point of view, PCA can be described as a bilinear model, as shown in the Equation (1):

$$\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{1}$$

where **T** (*n* x *a*) represents the scores of the PCA, i.e., the projection coordinates of the original *n* rows of **D** in the new low-dimensional space created using **P** (*m* x *a*) that is the array of the loadings, which determines the basis vectors, namely the factors of the PCA subspace expressing the highest variance of the data. Lastly, **E** (*n* x *m*) is the matrix of the residuals, the not modelled information that is not explained, at a chosen rank, from the model. If PCA is correctly conduced, the obtained **E** matrix should contain the information related not to the real details of the studied sample, but only a variation linked to factors such as the instrumental noise of the acquisition. An important aspect in PCA is that each Principal Component (PC) is orthogonal to the others, so the related information will be completely uncorrelated to the previous ones. A general representation of PCA in the framework of hyperspectral imaging is reported in Fig. 11. It is important to highlight again (as shown in the figure below) that the most of the multivariate analyses, and so PCA, cannot be applied directly on a hyperspectral image. This means that the data cube must be unfolded first in the corresponding two-dimensional data matrix and then, after the computational calculations, be again refolded, in order to obtain as many score images as the number of the selected PCs.

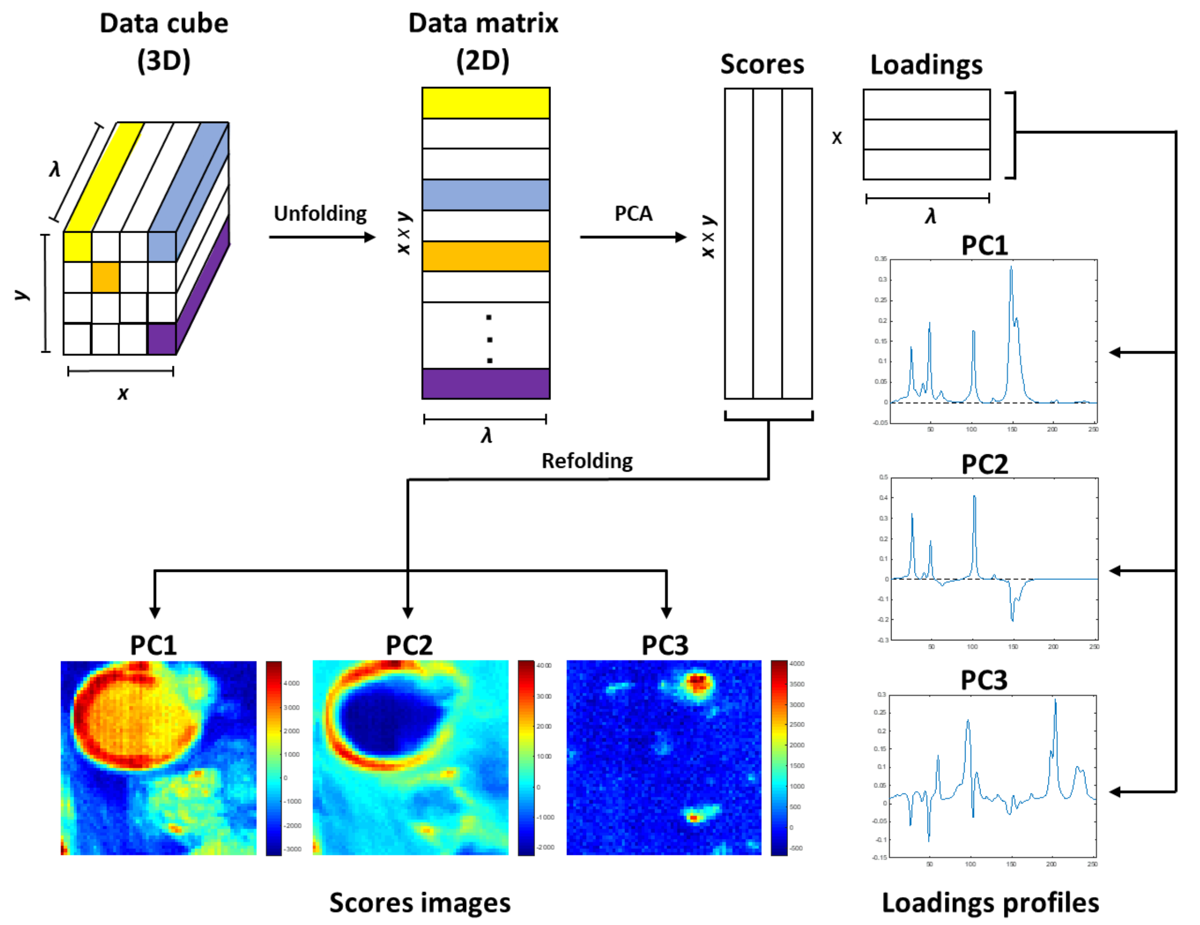


Fig. 11 – General scheme of PCA on a hyperspectral image. The data cube is first unfolded in the corresponding two-dimensional dataset and after the PCA, refolded to obtain the score images and loading profiles related to the contribution of the different selected PCs.

A relevant aspect to be stressed is related to the use of this technique and the information contained in the **E** matrix. If the quantity of selected PCs is not correct, this could lead to not precise results. In other words, if a number of PCs lower than the optimal value is selected, some information would be missing, being still contained in the **E** matrix. On the contrary, a too big number of selected PCs would lead to the use of residuals not related to the chemical information, but factors like the instrumental noise. This is the reason why it is mandatory to choose the appropriate number of PCs. Nowadays, various approaches to drive the operator in this choice are available, and they can be classified in three different categories [137]. The first and the most common method is based on the observation of the scree plot of the eigenvalues, firstly introduced by Cattell [138]. The general criterion is the examination of the eigenvalues, normally using the logarithmic unit, to find a threshold able to describe and distinguish the useful information contained in the matrix from the residuals. The second selection method uses statistical tests [139], based on the rate of decrease of the remaining residual sum of squares. Lastly, the third method is based on computational criteria and permutation testing, like Cross-Validation (CV) [140], and bootstrapping [141]. In the first case, part of the data is kept out of the model development, in order to predict then their values based on the use of the generated model to observe its robustness. In the second one, the residuals are used to simulate a large number of data similar to the original ones to observe the distribution of the model parameters over these data.

1.4.3. K-Means (KM) clustering

This kind of clustering methodology is designed with the purpose of partitioning a data matrix **D**, represented by *n* objects and *m* variables into *K* classes ($\mathbf{C_1}$, $\mathbf{C_2}$, ... $\mathbf{C}_{\boldsymbol{K}}$), where $\mathbf{C}_{\boldsymbol{K}}$ is the set of *n* objects in the cluster *k*, for a given total number of *K*. From a general point of view, once chosen a precise *K* number of classes, *K* centroids will be randomly generated. Each centroid is a point in the *m*-dimensional space found by averaging the values on each variable over the objects within the cluster. From a mathematical point of view, the centroid of the *j*th variable in cluster $\mathbf{C}_{\boldsymbol{K}}$ is, as reported in the Equation (2):

$$\bar{d}_j^{(k)} = \frac{1}{n_k} \sum_{i \in \mathbf{C}_{\boldsymbol{K}}} d_{ij} \tag{2}$$

and the complete centroid vector for cluster $\mathbf{C}_{\boldsymbol{K}}$ is given by the Equation (3)

$$\bar{d}^{(k)} = (\bar{d}_1^{(k)}, \bar{d}_2^{(k)}, \dots, \bar{d}_m^{(k)}) \tag{3}$$

The partition into the different clusters is based on the concept that the distance between the row vector for a particular object *i* belonging to *n* and the centroid of its corresponding cluster is at least as small as the distance to the centroids of the other clusters. Nevertheless, the task of optimizing KM outcomes can be very challenging for different reasons. From a general point of view, this algorithm operates following an iterative procedure, here explained:

1) *K* initial points are defined by *m*-dimensional vectors ($s_1^{(k)}$, …, $s_m^{(k)}$). For each object *i* belonging to *n*, the distance l^2 (*i*, *k*) between it and the *k*th seed vector is calculated as following, assigning each object to the cluster where the value of l^2 is minimum:

$$l^2(i,k) = \sum_{j=1}^{m} (d_{ij} - s_j^{(k)})^2 \tag{4}$$

2) After this first step, new centroids based on the Equation (3) are calculated. Each object is then newly examined, and reallocated to the cluster for which the distance with the new centroid is lower.
3) Again, new centroids are calculated with the updated version of the cluster membership.
4) Steps 2 and 3 are repeated in an iterative way, until the moment in which no objects can be moved between the clusters.

Another important factor that is taken into consideration while trying to partition the objects of a matrix into the different clusters, is to minimize a particular loss criterion, the Sum of Squared Errors (SSE) [142], as reported in the Equation (5):

$$SSE = \sum_{j=1}^{m} \sum_{k=1}^{K} \sum_{i \in \boldsymbol{C_K}} (d_{ij} - \bar{d}_j^{(k)})^2 \tag{5}$$

Using also this value, it is possible to obtain a better optimization of the clustering results, moving the objects from one cluster to another, trying to minimize the final value of the SSE. Nevertheless, despite the robustness shown by the use of KM clustering, it remains a method influenced by some limitations. Particularly, one can refer to two main challenges. The first one is related to the obtained final global optimum. In fact, depending on the starting values used as centroids, the algorithm will provide results that can show some local optima, but hardly a verifiably global one. Clearly, some solutions have been proposed [143–145], in order to overcome this problem. For example, one method is to perform the KM clustering several times, with different starting values, choosing at the end the best SSE solution. Another way is represented by the choice of *K* data point as the initial cluster seeds, or subdivide randomly the data units into *K* mutually exclusive partitions and calculate the mean for each of them, in order to use these values as centroids. Nevertheless, these methods might suffer of the influence of the initial selected data. By way of example, a further proposed approach is based on a deterministic method. Once defined a distance l_1, it is computed the number of data points within l_1, defined to as the density. The group represented by the highest density is chosen as the first cluster. The remaining clusters are selected by decreasing density, as long as they are at another defined distance l_2 from the already defined seed. The second problem, when using KM clustering, is given by the estimation of the right number of clusters *K* [146]. From a general point of view, it is possible to distinguish between three different kinds of methods. The first one is the algorithmic method. Normally, it is the operator that decides the number of clusters to be used. Nevertheless, the algorithm has the opportunity to modify the user-provided value, if some clusters result to be closer than a certain calculated value, optimizing the partition into *K*-new clusters. The second approach is the graphical interpretation, and probably it is the most used, despite the weakness of being a highly subjective method. Various K values are attempted, and the resulting curve is observed. Normally, a ‘flattening’ of the curve indicates the right number of clusters to be selected. Lastly, one can refers to formulaic methods, in which an equation is computed across a range of K and the results are investigated with the aim of minimizing or maximizing the selected criterion, in order to select the right quantity of clusters.

1.4.4. Partial Least Squares Regression (PLSR) and Partial Least Squares-Discriminant Analysis (PLS-DA)

In spectroscopy it is common the use of the acquired data with the purpose of obtaining quantitative information. Nevertheless, in order to accomplish this task, it is necessary to have a potential relation between the measured signal and the response to be estimated and, particularly from a chemometric point of view, the postulation of a mathematical formulation that can express, or at least approximate, this relation. In other words, it is mandatory to find a functional relation *f,* which can allow quantifying the value of a property *y*, based on the experimental measurements of a spectroscopic signal $\mathbf{d}$, as shown in the Equation (6):

$$y = f(\mathbf{d}) \tag{6}$$

The limitation is that, from an experimental point of view, this function is unknown and so, it has to be found in an empirical way, by the use of the so-called calibration [147]. In this way, it becomes possible by the use of an approximation to calculate the experimental value *y*, based on the function presented in the Equation (6). Nevertheless, this step can be challenging. In fact, in order to obtain good approximations, a sufficient number of samples (namely the training set) showing the properties of interest must be used to train the model and obtain a good robustness. It is also important to highlight the fact that the function *f*(**d**) does not depend only on **d**, but also, and more importantly, on the values of some parameters, the coefficients, that are the principal key related to the quality of the calibration. So, assuming that the function *f*(**d**) is linear, as generally it is in spectroscopy, for an acquisition of m spectral variables, Equation (6) can be rewritten as:

$$y = \hat{y} + e = b_1 d_1 + b_2 d_2 + \cdots + b_m d_m + e \tag{7}$$

where $\hat{y}$ is the approximation of y related to the linear function defined by the coefficients b_1, b_2,…, b_m and e, the residuals, which explain the difference between y and $\hat{y}$. The coefficients are then calculated by the use of the regression analysis in calibration, using the available pairs of (**d**, *y*) constituting the training set of the data. In order to obtain the best prediction of *y* through the measured signal **d**, normally the most common criterion is based on the use of the minimization of the residuals, using the least squares [148]. From a general point of view, in a dataset **D** containing *n* samples, for which both **d** and *y* are known, the coefficients $b_1, b_2, \ldots, b_m$ are identical for all the elements, due to the fact that the same functional relation is valid for all the samples. Therefore, it is possible to express the general formula of **y** as reported below, in the Equation (8):

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e} = \mathbf{Db} + \mathbf{e} \tag{8}$$

where $\hat{\mathbf{y}} = \mathbf{Db}$ represents the approximation of the response of **y**, the column vector **b** contains the regression coefficients of $b_1, b_2, \ldots, b_m$ and **e** constitutes the residuals. Lastly, assuming that there is not difference from a mathematical point of view between building a model for each of all the single properties of interest and constructing a single one calibration set for the whole system, Equation (8) can be summarized as:

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{E} = \mathbf{DB} + \mathbf{E} \tag{9}$$

As previously described in the manuscript, nowadays different regression approaches can be used, such as MLR [149]. Nevertheless, without any doubt PLS is the most promising method, overcoming the limitations of the other methods [116,150]. In fact, MLR faces difficulties in the situation in which the number of training samples is lower than the number of recorded variables (an easy scenario, due to the modern instrumentations), and/or when the variables are highly correlated. Contrarily, PLS uses the information in **Y** coming from the already compressed data, so that the scores extracted from **D** are relevant for describing simultaneously both the variance in the descriptors and in the properties of interest. In other words, PLS is based on the extraction of LVs from the **Y**-block. This is particularly interesting because it means that if multiple responses are observed at the same time, individual models can lead to different responses compared with the outcomes obtainable with a single global model, contrarily to MLR approach. The calculation of a single model for calibrating all the responses at the same time means that it can contain a part of the information that is related to the systematic variability, i.e., a certain degree of intercorrelation between the variables. Extracting two sets of scores, one from the independent and one from the dependent data block, which have maximum covariance, using the selected components, it is possible at the end to summarize the relevant information in **D** and **Y** as following shown:

$$\mathbf{T} = \mathbf{DR} \tag{10}$$

$$\mathbf{U} = \mathbf{YQ} \tag{11}$$

$$\hat{\mathbf{Y}} = \mathbf{UQ}^{\mathbf{T}} \tag{12}$$

$$\hat{\mathbf{U}} = \mathbf{TC} \tag{13}$$

where the columns of **T** and **U**, **R** and **Q** are the scores and the coefficient matrices of, respectively, the **X**-weights and **Y**-loadings associated with the individual retrieved components of the original data matrix **D**. Lastly, **C** is a diagonal array in which the non-zero elements represent the inner regression coefficients. Finally, the regression model can be rewritten as in the Equation (14):

$$\hat{\mathbf{Y}} = \mathbf{UQ}^{\mathbf{T}} = \mathbf{TCQ}^{\mathbf{T}} = \mathbf{DRCQ}^{\mathbf{T}} = \mathbf{DB} \tag{14}$$

where **B** is the matrix of PLS coefficients. Due to the linearity of the projection, it is possible to express the regression model that is calculated at the level of the scores, also in terms of the original variables. In addition, due to the fact that only a part of the information of the original matrix is used for the regression, the PLS coefficients can be rewritten as:

$$\mathbf{B} = \mathbf{RCQ}^{\mathbf{T}} \tag{15}$$

Despite the robustness of PLS, it is important to stress the fact that it is necessary to select the right number of components to create a model and so, to avoid biased results. Normally, this procedure is based on the selection of the values leading to the minimum prediction error that is found using the CV approach [151]. A very interesting aspect of PLS is that a regression problem can be considered as a classification method in which the class belonging of a sample (the dependent variable) is to be estimated from the set of variables (such as the spectra) obtained using a particular instrument. This kind of approach is known with the name of PLS-DA [123,152,153]. Discriminant analysis is a particular kind of classification approach in which the main task is to highlight the differences between samples of distinct classes. The multivariate space is divided into a number Z of subregions equal to the number of the selected categories. Then, each object is assigned to a particular class, i.e., the one for which the point corresponding to its measurement vector falls into the region of a particular category. In order to work, discriminant analysis requires some characteristics. First of all, a training set composed by samples belonging to all the classes is used, in order to calibrate a balanced model. Furthermore, it is important that each single sample be assigned to one and only one of the different categories.

Lastly, it is fundamental to consider that if a sample is not coming from any of the classes (a new kind of specimen not considered in the initial study), it will be anyway always assigned to one of the categories, despite the fact that it is misclassified. The principle of this technique is that, given a data matrix **D**, it is regressed on a dummy binary-coded response array (namely **Y**), made of a set of *Z*-dimensional row vectors. Taken an object, if it is a member of the *z*th class, the corresponding vector will correspond to a 1-value in the *z*th entry, and 0-value in the other ones. By way of example, considering a simple case in which only two classes are available, samples belonging to the first one will be described by a vector [1 0]. In the same way, samples from the second class will be represented by a vector [0 1]. Once the model is built, new objects will be assigned to a particular class according with their similarities with the samples already available.

1.4.5. Soft Independent Modelling of Class Analogies (SIMCA)

As previously described, SIMCA [126,154] is a modelling approach, meaning that instead of highlighting differences between samples belonging to different classes, as in discriminant analysis, the main task is to capture the similarities among the samples of the same category. This is the main reason why each class is modeled individually, handling the samples coming from each category separately and independently from the ones belonging to the other classes. From a general point of view, a multivariate boundary will be defined for each class, which delimits a specific region, which will describe a particular category. This means that, if the projection of a particular sample falls into this region, it will be assigned to that particular class, otherwise it will be considered as an outlier, and so rejected [155]. In addition, one of the most interesting aspects compared with the discriminant analysis is that in modelling techniques it is not mandatory to divide the total original space into the considered classes, because only a multivariate boundary space for each category is defined. It means that the various class spaces do not necessarily have to cover completely the totality of the original variable space. In other words, if using approaches such as PLS-DA one sample will be always assigned to a specific class, no matter if it is really part of that category or is a completely new element, using algorithms such as SIMCA, the same specimen can be assigned to one, none or multiple classes. Also important is that, due to its characteristics, modelling analysis can be used in studies in which a unique class of interest has to be identified. The concept behind SIMCA is very simple. Each class is separately defined on the basis of a principal component model of opportune dimensions. Considering Equation (1) previously introduced to explain PCA, it is possible to obtain for a specific class *c*, described by a principal component model:

$$\mathbf{D_c} = \mathbf{TP^T} + \mathbf{E} \tag{16}$$

where $\mathbf{D}_c$ is the sub-matrix of the original data obtained by the use of only the samples being part of the class *c*, **T** and **P** are respectively the matrices of the selected scores and loadings, and **E** the residuals not used to create the model. At this point, it is possible to create a multivariate boundary delimiting the specific class *c* and, using a distance-to-the-model criterion that is based on Multivariate Statistical Process Control (MSPC) [156], it is possible to detect if a new observed sample is part of the considered class or not, depending if it falls within or not the limits of this space. More specifically, these borders in SIMCA are calculated using two values, computed observing the score matrix **T** and the residual matrix **E**. They are the probability distributions for the distances within the model spaces (T^2 statistics) and the orthogonal distance to the model space (Q statistics). A threshold value corresponding to a precise confidence level (that statistically is normally equal to 95%) is chosen, and the class space will be calculated by the Equation (17):

$$\sqrt{\left(\frac{\mathrm{T}^2}{\mathrm{T}^2_{0.95}}\right)^2 + \left(\frac{\mathrm{Q}}{\mathrm{Q}_{0.95}}\right)^2} \leq \sqrt{2} \tag{17}$$

where T and Q are the statistic values found for a particular sample, while $\mathrm{T}^2_{0.95}$ and $\mathrm{Q}_{0.95}$ are their corresponding 95% confidence level threshold values. In other words, the Equation (17) is used to generate the boundaries that will be used to determine if a specific sample is part or not of the considered class *c*, if its projection falls within or outside the limits of the statistic results.

1.4.6. Multivariate Curve Resolution (MCR) or signal unmixing

The purpose of MCR methods [157] is to extract the relevant information in a mixture system to obtain the pure components through a bilinear model decomposition. It means that the experimental data matrix **D** of dimensions *n* x *m* is decomposed into the product of the concentration profiles matrix **C** (*n* x *k*) containing the concentration of the pure components present in the system and their corresponding pure spectral profiles matrix $\mathbf{S^T}$ (*k* x *m*). In this notation, *n* represents the mixture spectra in rows measured at *m* wavelengths that follow the bilinear model, while *k* is the number of pure components supposed to underlie **D**. The algorithm can be resumed as an extension of the Lambert-Beer's law, which can be described by the use of a vector notation as in the Equation (18):

$$\mathbf{d} = \mathbf{cS^T} \tag{18}$$

in which **d** represents the measurement vector (1 x *m*), **c** (1 x *k*) is the vector of the concentrations and **S** (*k* x *m*) is the matrix of the absorption for each species at each wavelength. An important aspect in order to use a MCR algorithm is that the system has to describe at least a second-order data, so a set of *n ≥ 2* spectral mixtures at *m* wavelengths. The MCR model, used when spectral or calibration information are not available in order to obtain the contribution from the different pure components [158], will be then described as reported in the Equation (19):

$$\mathbf{D} = \sum_k \mathbf{c}_k \mathbf{S}_k^{\mathbf{T}} + \mathbf{E} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \tag{19}$$

in which **E** (*n* x *m*) is the residual matrix, containing the variability of **D**, which is not explained by the model and should be close to the experimental error. Normally, MCR is also described graphically as shown below, in Fig. 12:

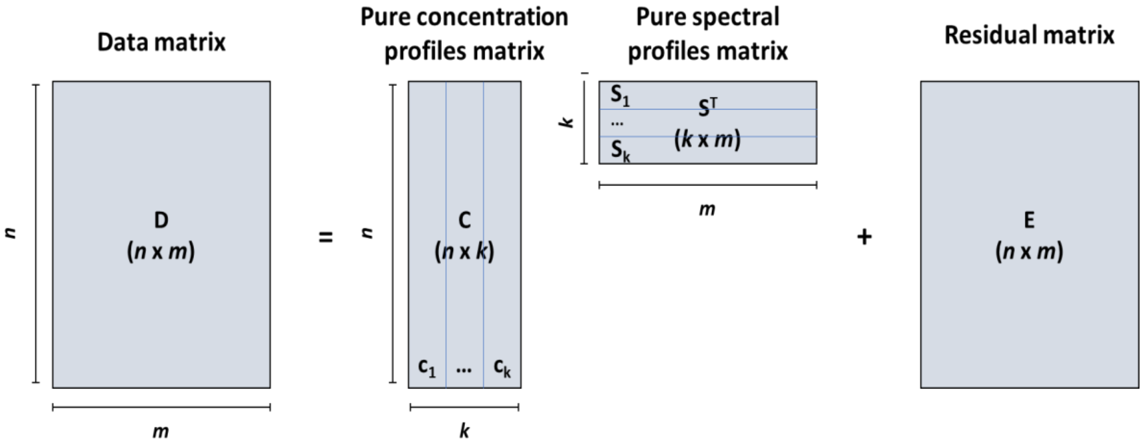


Fig. 12 – Graphical representation of a MCR model for a data matrix **D** containing *n* mixture spectra at *m* wavelengths for *k* pure components.

There are different approaches that can be used in order to decompose correctly the data matrix **D**, mainly grouped into non-iterative and iterative approaches. The first ones are based on combining information of small sections of the data obtained from global and local rank information that can contain particular properties, as the presence and/or absence of a particular component. Just to mention a few, some of the most used methods are Window Factor Analysis (WFA) [159], Subwindow Factor Analysis (SFA) [160] and Heuristic Evolving Latent Projections (HELP) [161]. On the other hand, iterative methods start from initial estimates of **C** or $\mathbf{S}^{\mathbf{T}}$ that will evolve to yield profiles with chemically meaningful shapes. Examples are MCR-ALS [6,162,163] and Iterative Target Transformation Factor Analysis (ITTFA) [164,165]. Iterative methods are probably the most popular and used in chemometrics because they allow the introduction of external information with the purpose of calculating better results. In fact, non-unique solutions are ordinarily obtained through the presence of ambiguities [166], due to the fact that for a given rank, sets of paired $\mathbf{C}$ and $\mathbf{S}^{\mathbf{T}}$ matrices can bring to the same quality of fit during the MCR decomposition. By way of example, two of these ambiguities are the intensity ambiguity, related to the fact that different profiles with the same shape but different relative scales will fit the results equally well, for which a normalization can be applied to avoid this behavior, and the permutation ambiguity. However, the most critical kind of ambiguity is represented by the so-called rotational ambiguity. The basic equation associated with **D** can be rewritten as:

$$\mathbf{D} = \mathbf{C}(\mathbf{T}\mathbf{T}^{-\mathbf{1}})\boldsymbol{S}^{\mathbf{T}} \tag{20}$$

$$\mathbf{D} = (\mathbf{C}\mathbf{T})(\mathbf{T}^{-\mathbf{1}}\mathbf{S}^{\mathbf{T}}) \tag{21}$$

$$\mathbf{D} = \mathbf{C}'\boldsymbol{S}'^{\mathbf{T}} \tag{22}$$

where $\mathbf{C}'$ = $\mathbf{CT}$, $\mathbf{S}'^{\mathbf{T}}$ = $\mathbf{T}^{-1}\mathbf{S}'^{\mathbf{T}}$ and in which **T** represents a rotation matrix. Mathematically $\mathbf{C}'$ and $\mathbf{S}'^{\mathbf{T}}$ lead to solutions that will fit the experimental data **D** equally correctly as the true **C** and $\mathbf{S}^{\mathbf{T}}$ matrices, though $\mathbf{C}'$ and $\mathbf{S}'^{\mathbf{T}}$ are not the sought solutions from a physical point of view.

For this reason, constraints are applied during the ALS process in order to refine initial estimates, but also and more importantly, to reduce the possible ambiguities [167]. Constraints are chemical or physical properties implemented as mathematical conditions with the aim of driving the MCR optimization to the final solutions, taking care of not introducing wrong information that could lead to artifacts. They can be grouped into hard and soft constraints, depending on the strictness to force the optimization process to obtain the MCR decomposition [168], though nowadays the implementation of physicochemical models make possible to take together the advantages of both the methods [169,170]. Some of the most interesting and used constraints are: non-negativity, maybe the most common and used constraint, applied for many datasets to correct the fact that many signals are naturally positive or zero [164,166,171]; unimodality, in the cases in which only one maximum per profile can exist, as in chromatography elution time peaks; closure constraints, applied on the rows of the matrix **C** and normally used in the reaction systems in order to equal all the elements of each row of the matrix **C** to a known constant, summing them [172,173]; selectivity constraints, associated with the concept of local rank (how the number and the distribution of the components vary locally along a particular dataset, referring to the fact that in a particular spectral range it can be assumed that a specific species can exist while others are known to be absent) and related to mathematical features, they can be applied to all datasets, regardless of their chemical nature [166,174]; equality, using chemical information associated with the knowledge of pure spectra or concentration profiles, when some elements are known, in order to set them to be invariant along the iterative process [175,176]. Once a constraint has been used and implemented, it will act as the driving force of the iterative process for the optimization. However, care must be taken using constraints because also a potentially applicable one could play a negative role if some factors such as experimental noise or instrumental problems distort the related profile. In order to obtain acceptable results, the MCR-ALS algorithm is based on the following steps:

1) Determine the rank of the dataset, by the use of PCA [177,178] and the corresponding calculated eigenvalues [138], as better discussed in the following paragraph. Despite this, one of the works of the present thesis focused on the description of a different method that is able to perform this task in a different way based on a well-known algorithm, SIMPLe-to-use Interactive Self-modelling Mixture Analysis (SIMPLSIMA) [179], named Randomised SIMPLISMA [180], and better described in the Chapter 2 of this manuscript.
2) Generation of initial estimates (**C** or $\mathbf{S^T}$ matrix).
3) Calculate respectively $\mathbf{S^T}$ or **C** depending on the previous step using the iterative method (MCR-ALS) under the right constraints, to avoid any artifact.
4) Starting from the previous results, calculate the other matrix using least squares under constraints.
5) Using the product of the obtained results, reproduce the dataset **D** and evaluate its reproduction.
6) Repeat the procedure from step (3) until convergence.

Normally, convergence is achieved when in two consecutive iterative cycles, relative differences in standard deviations of the residuals between experimental and ALS calculated data values are less than a selected value, usually 0.1%. The final quality of the model depends on two important figures of merit: Lack of Fit (LOF), representing the difference among the input data **D** and the data reproduced from the $\mathbf{CS^T}$ product obtained by MCR-ALS, and the percentage of variance explained (r^2), shown respectively in Equation (23) and Equation (24):

$$LOF\ (\%) = \ 100 \text{ x} \sqrt{\frac{\sum_{ij} e_{ij}^2}{\sum_{ij} d_{ij}^2}} \tag{23}$$

$$r^2 = \frac{\sum_{ij} d_{ij}^2 - \sum_{ij} e_{ij}^2}{\sum_{ij} d_{ij}^2} \tag{24}$$

where d_{ij} is the spectral value of the mixture i at the wavelength j and e_{ij} is the associated error.

MCR-ALS has been largely applied to solve many complex matrices in different fields [181–184] as well as for the resolution of spectroscopic image analysis [185–189]. Again, it is worth stopping on the same important aspect highlighted in the previous paragraphs, regarding the investigation of a hyperspectral image. In this case, the three-dimensional data cube of dimensions *x* x *y* x *λ* where *x* and *y* represent the number of pixels in the two spatial directions and *λ* the direction of the spectral variables, will have to be unfolded in the corresponding two-dimensional dataset prior the MCR calculations in order to allow the decomposition of **D** into the contribution of the **C** and $\mathbf{S}^T$ matrices. In the last step, the **C** matrix will be refolded in order to retrieve concentration maps of each pure compound extracted by MCR-ALS, as showed in Fig. 13:

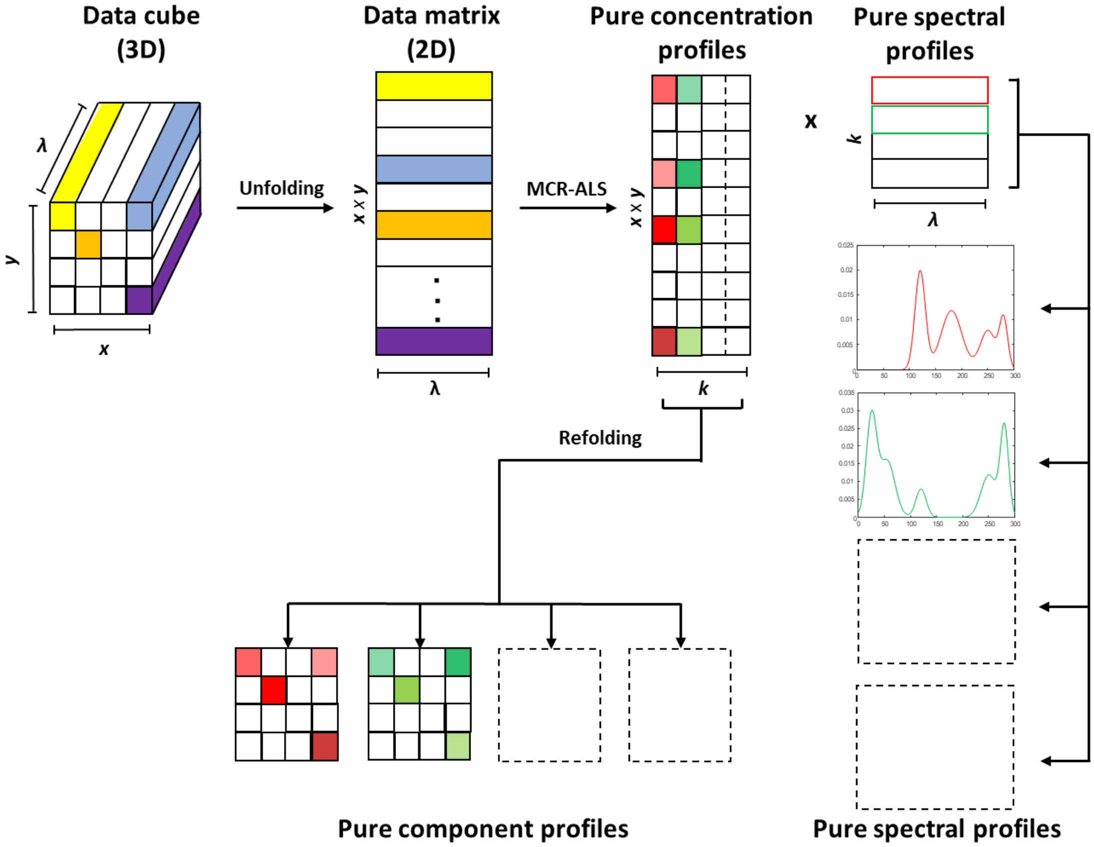


Fig. 13 – Graphical representation of an MCR-ALS procedure when applied to hyperspectral images. The cube needs as a prerequisite step to be unfolded into the product of *x* and *y* towards *λ* and after the optimization process, to be folded again into the contribution of the found pure components.

1.4.6.1. Rank evaluation using PCA

The most used method in order to evaluate the rank of a data matrix prior MCR-ALS analysis, as introduced in the previous paragraph, is based on the observation of the scree plot of the eigenvalues associated with principal components obtained by PCA. This method was introduced the first time by Cattell [138] and it is based on the principle of the meaningful information expressed by a particular factor (or component). It is displayed as a downward curve in which eigenvalues are compared on the Y-axis, the most important first, while on the X-axis are reported the various components. Because the useful information decreases gradually taking into account the subsequent eigenvalues (as also the distance in the graphic between them), the strategy is to consider only a limited number of components. Using a scree plot, the choice of this value is carried out by the interpretation of the curve, in which the right value should correspond to the 'elbow' in the graph, where the eigenvalues level off. In this way, it will be considered a threshold above which the eigenvalues carry a meaningful chemical information, leaving out the ones that contain noise and redundant information. Despite the method can seem easy and immediate, a subjective and arbitrative interpretation is often observed, especially when noise is significant. Moreover, it also remains very challenging using this method to detect minor compounds, which result to be very close to the noise level, as in the analysis of complex data matrices, due to the small value of explained variance related to their information.

1.4.6.2. SIMPLe-to-use Interactive Self-modelling Mixture Analysis (SIMPLISMA)

SIMPLISMA [179] is a pure variable method and particularly, it has been one of the very first multivariate curve resolution approaches used in spectroscopy [190–192]. Normally, a mixture consists of hundreds of variables, each of them represented by the contribution of one or more components. A pure variable is a variable that depends on the contribution of only one component. The central task of this approach is the selection of the so-called pure variables from the data matrix **D**. It is important to stress the fact that using this approach, the presence of pure components in the matrix is not required as long as pure variables are present. By the use of a spatial representation, the variables can be presented as vectors, which positions give a direct measure of the contributions of the components. This means that the purer a variable is, the more it will coincide with a particular component axis. Furthermore, because the purity of a component is related to the length on the variable vectors, a variable with a high intensity will be relatively pure. In this way, the first pure variable will be found by determining the vector with the largest length in the plot. For a data matrix **D** with dimensions n x m, the length l_i of a variable i is, as shown in the Equation (25):

$$l_i = \sqrt{\frac{\sum_{v=1}^{m}(d_{i,v})^2}{m}} \tag{25}$$

l_i is strongly related to the mean of variable i, μ_i (26), and the standard deviation of variable i, σ_i (27), as shown in the Equation (28):

$$\mu_i = \frac{\sum_{v=1}^{m} d_{i,v}}{m} \tag{26}$$

$$\sigma_i = \sqrt{\frac{\sum_{v=1}^{m}\left(d_{i,v} - \mu_i\right)^2}{m}} \tag{27}$$

$$l_i^2 = \mu_i^2 + \sigma_i^2 \tag{28}$$

The first purity value related to a variable *i*, based on these two statistical tools, is then estimated with the index $p_i^{(1)}$:

$$p_i^{(1)} = \frac{\sigma_i}{\mu_i + \alpha} \qquad \text{for } i = 1, \ldots, n \tag{29}$$

The user-defined parameter α avoids giving a high purity value to a variable with a low mean. This factor will be negligible if the noise is low (high values of μ_i) and vice versa, in which situation α will correct the noise influence. Once the first purest variable, the one showing the highest $p_i^{(1)}$value, is calculated, the second one will be the most independent from the previous one. It is necessary, in order to calculate it, to subtract the contribution of the first pure variable from the matrix **D** before continuing the calculation. A weighting parameter $w_i^{(2)}$ is thus considered in order to reduce the influence of other variables that would be correlated to the first pure variable. More details about this parameter are given in other works [193]. The second purity value $p_i^{(2)}$ related to a variable *i* is then defined by:

$$p_i^{(2)} = w_i^{(2)} \frac{\sigma_i}{\mu_i + \alpha} \qquad \text{for } i = 1, \ldots, n \tag{30}$$

Again, the next purest variable has the highest $p_i^{(2)}$ value. The following purest variables are of course obtained by iterating this calculation until the number of variables corresponding to a given rank is obtained. It is often forgotten that this extraction of pure variables can be done in both the dimensions of the matrix **D**. In this way, the selection of variables along the columns of **D** allows to obtain the estimations related to the concentration profiles, while using the rows are obtained the estimations of the purest spectra.

1.4.7. Multivariate Image Analysis (MIA)

The main limitation using chemometric approaches is that a given data cube has to be unfolded in the corresponding two-dimensional dataset before any analysis, as previously stated. In this way, the problem is that the spatial information is completely lost, leading to an incomplete exploitation of the real potentials of this kind of matrix. In chemometrics, the analysis and interpretation of a chemical image has always mainly been based on the spectroscopic part of the data, and not the spatial information related to the image. Each pixel is considered as an independent sample, and the whole image is represented as a set of vectors of intensity values. Multivariate Image Analysis (MIA) is a particular kind of field of chemometrics in which the main task is to represent the results in a graphical way, trying to give a new interpretation to the original data cube [194–196]. Nevertheless, also this kind of technique focuses only on the spectral information. In fact, the first applied operation using MIA is the unfolding step of the three-dimensional matrix. Then, different chemometric approaches can be applied (e.g., PCA, MCR-ALS, etc.), considering each pixel as a single and independent sample in the dataset. Finally, each pixel model component obtained from the multivariate analysis can be refolded to the original spatial structure and represented as a false-color image with the same dimensions as the original image. This means that this kind of procedure does not find a solution in using the spatial information obtainable from the image, in which normally one could assume that neighbor pixels can easily show correlations and anti-correlations from the chemical and physical point of view. Nevertheless, using MIA is possible to observe the results coming from different chemometric approaches, such as PCA, on the folded image, in order to lead to a more practical interpretation of the distribution of the various components present in the sample of interest.

1.4.8. Wavelet transform

Despite the variety and vastness of chemometric techniques nowadays used in hyperspectral image analysis, it seems that it is impossible to use equally both the spectral and spatial information of an investigated sample. As explained, the main constraint when observing this kind of matrix within the use of chemometrics is the mandatory unfolding procedure of the data cube into its corresponding two-dimensional dataset. Despite this, by the use of this procedure the totality of the spatial information is lost, leading to a limited use of the information related to the image. Nevertheless, the interest in this problem has been recently in the spotlight of many research studies. Different ways to deal with this limitation have been investigated, but one of the most interesting ones is for sure related to the use of the wavelet transform algorithm.

Generally, wavelet transform [197,198] shows many similarities with Fourier Transform (FT) [199]. Both of them can be used with the goal of obtaining a signal or image clearance and simplification. Fourier analysis has been first used in the framework of the signal analysis varying with time and it results to be useful because the content of the frequency of the signal is of great importance. Despite this, the drawback of the FT technique is that transforming a signal in the corresponding frequency domain, time information is lost and so, it is impossible to tell when a particular event took place. Differently, wavelet analysis overcomes this aspect by the use of a windowing technique with variable-sized regions [200]. Particularly in spectroscopy, this peculiarity confers a great importance to the wavelet transform, which is related not to the time domain, but to the wavelengths one. One major advantage of the wavelets is the ability to perform local analysis, namely to analyze a localized area of a larger signal. Mathematically, the FT equation $F(\omega)$, where ω is the frequency, is the sum over all time of the signal $f(t)$ multiplied by a complex exponential, as shown in Equation (31). Contrarily, the first kind of studied wavelet, called Continuous Wavelet Transform (CWT), is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function ψ, as shown in Equation (32):

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt \tag{31}$$

$$C(scale, position) = \int_{-\infty}^{\infty} f(t)\psi(scale, position, t)dt \tag{32}$$

The CWT can operate at every scale, but the con is that in this way an awful amount of data will be generated. For this reason, the Discrete Wavelet Transform (DWT) was introduced, in order to save the low-frequency contents, which contain the signal identity, removing the rest of the unnecessary information [201], using low- (g) and high-pass (h) filters, as reported in Equation (33):

$$y[n] = (x \times g)[n] = \sum_{k=-\infty}^{\infty} x[k]g[n-k] \tag{33}$$

where x is a particular signal and g is the impulse response of the applied low-pass filter. This procedure is performed by the decomposition of the signal into a mutually orthogonal set of wavelets, leading to the elimination of the redundancy in coefficients, though subsampling is operated by this step, leading to the removal of half the frequencies of the signal, and so half the samples, according to Nyquist's rule. Finally, Stationary Wavelet Transform (SWT) was designed to overcome the lack of translation-invariance of DWT, removing its downsampling, and upsampling the filter coefficients by a factor of $2^{(j-1)}$ in the jth level of the algorithm [202].

By the use of this algorithm, the output of each level of SWT contains the same number of points as the input, contrarily to DWT. The comparison of the two mechanisms is shown in Fig. 14:

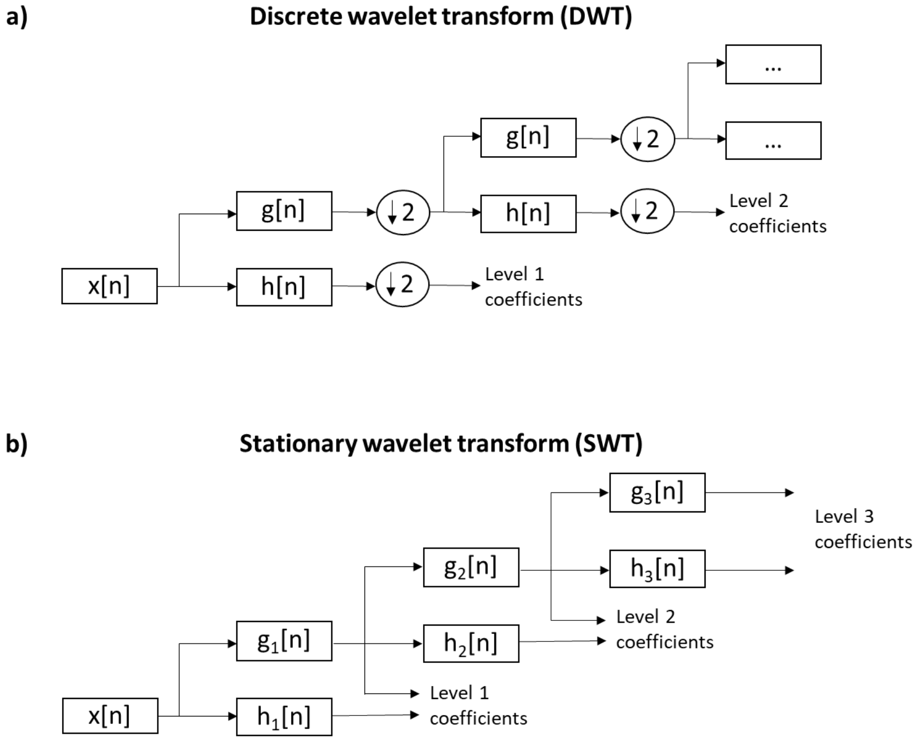


Fig. 14 – General schemes for a) Discrete Wavelet Transform (DWT) and b) Stationary Wavelet Transform (SWT).

Moreover, an improvement of DWT (and so SWT) compared with CWT is that various wavelet families presenting different wavelet functions ψ have been introduced, with the aim to better fit the kind of signal to be interpreted [203,204]. One example is the first and simplest one, the Haar wavelet, which is represented by a discontinuous and step-size function. Despite this, the Daubechies family remains the most used nowadays, which is a set of compactly supported orthonormal wavelets. Another interesting family is represented by the biorthogonal function, which exhibits the property of linear phase. Particularly, this kind of approach uses two wavelets (one for the decomposition and another one for the reconstruction), which results to be useful in the context of the signal and image reconstruction, showing interesting properties. Finally, during this PhD has been explored a way to use wavelet transform in the framework of hyperspectral image analysis. In fact, in the Chapter 5 of the present manuscript, it will be discussed the exploitation of this same principle applied to images extracted from hyperspectral data cubes for a better consideration of the spatial dimension of the cube merged together with the corresponding spectral part of the data.

CHAPTER 2

2. ON THE SELECTION OF THE MOST IMPORTANT INFORMATION IN HYPERSPECTRAL IMAGE ANALYSIS

2.1. General overview of the challenge in selecting the most important variables/spectra in hyperspectral imaging framework

As discussed in Chapter 1, hyperspectral image analysis has surely broadened the horizons of the investigation in many different research areas [38,56,73,76,78,205]. Nevertheless, the use of this technique leads also to the possibility of generating a huge amount of data, which need to be filtered and observed with the right techniques, in order to keep all the important information, preventing the loss of very specific details [29,106,206]. Over the years, chemometrics has been used for this challenge, trying to maximize the quality of the obtainable results in this context, using various algorithms and strategies. One part of this PhD has been focusing in this aspect, trying to understand how to help in this delicate and fundamental task. From a general point of view, by way of example, PCA is nowadays vastly used in many exploration analyses, with the aim of obtaining a main idea about the chemical structure of complex matrices [6,137,207]. In addition, this chemometric tool can be used to reduce the dimensionality of the dataset, selecting only the most important and meaningful variables, as well as find and remove outliers present in the matrix. Nevertheless, PCA is based on the interpretation of the operator, reason why it can lead to inaccurate results. In addition, the division of the information into the different PCs is based on the total explained variance. This means that if very specific, but few spectra are showing some information that is pure and different from the rest of the data, they might be lost, due to the fact that the total variance that they express very low compared with the rest of the data in the matrix. This is a very common scenario in hyperspectral imaging, where hundreds to thousands of spectra can be easily acquired. Another method that is vastly used for the distinction of the various chemical groups into a matrix is KM clustering [143,208]. As previously described, this approach separates the classes of components using as criterion the distance of each spectrum, considered as a point in the multidimensional space, from the different centroids, that are in a first step randomly selected. Then, in an iterative way, the centroids are recalculated using the identified clusters obtained in the first step, and so each point is assigned again to the new, closer class. This procedure is repeated until the moment that any spectrum cannot be anymore moved from one cluster to another. Naturally, some issues can be faced. First of all, the operator needs to select the right number of classes to be considered, and so the number of initial centroids to be used. It means that, if the starting selected value is not right, the clustering will lead to a certain degree of inaccuracy, and samples of different nature will be considered to contain the same information, or vice versa, samples of the same class will be split into different groups. In addition, also the starting selected clusters could lead to wrong results, due to some computational mistakes given by the initial used values. Nevertheless, as already discussed, nowadays some methods to help the operator in these tasks are available. Anyway, it is important to stress the fact that is fundamental to carefully use this method, because a lack of attention could lead to unsuitable outcomes, as for PCA. Lastly, KM clustering can be affected by some ulterior problems. Considering hyperspectral image analysis, as already explained, it is possible to obtain an enormous quantity of produced data. In this kind of situation, it is an obvious statement that some classes can be represented by a small number of spectra or that pixels being part of two or more different families can show very similar spectral information and so be erroneously grouped together, as described in Fig. 15:

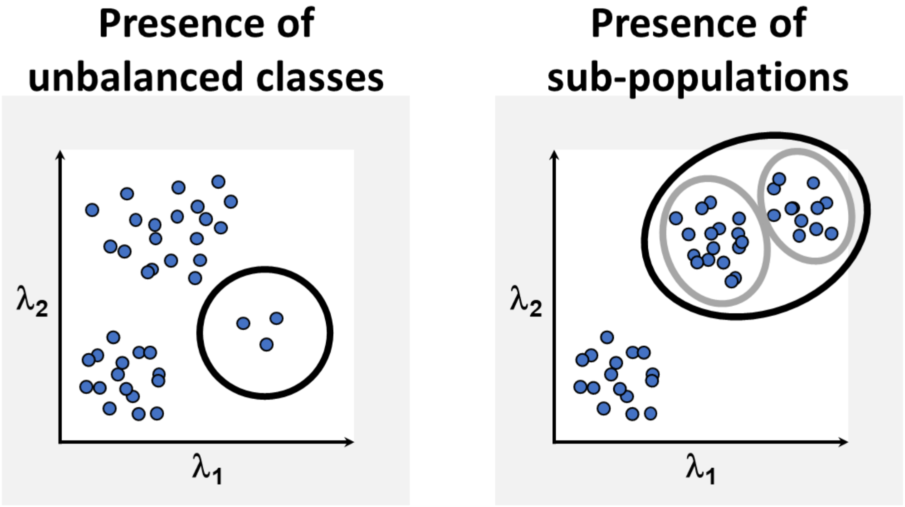


Fig. 15 – Graphical representation of two common problems when KM clustering is used as classification method.

In this scenario, the possibility of missing particular clusters is very common, and it can result very challenging to find the right experimental values. Lastly, MCR-ALS is for sure one of the most interesting approaches currently used in hyperspectral imaging [33,209]. In fact, one of the main requested tasks in many research areas is the spectral unmixing of the matrix of interest into its corresponding pure components. Finding a way to separate and observe the signal contributions of the different elements composing a matrix is a very important mission nowadays, but it can be more challenging than expected. As introduced, MCR-ALS is a procedure that in an iterative way refines the obtained results to eventually yield profiles with chemically meaningful shapes. In order to do this, the algorithm is based on some important steps. First of all, similarly to the previous explained PCA and KM clustering, it is fundamental to determine the right rank of the dataset, i.e., the right number of pure components present in the sample. This procedure is normally conducted by the use of PCA and the observation of the corresponding eigenvalues of the extracted PCs. Again, this procedure is made by the operator and so, it can be affected by an experimental error. In fact, due to different reasons (e.g., the noise level of the acquired data, the presence of minor compounds, which explain a low quantity of information, etc.) it is not always easy to determine the right number of components, leading to inexact outcomes. Also important is that MCR-ALS is based on the use of initial estimates in order to drive the computation of the results, and so calculate the pure matrix concentration profiles **C** and their corresponding spectral profiles $\mathbf{S^T}$, respectively. Normally, initial estimates can be calculated by the use of some algorithms. Currently, one of the most used approaches in the routine analyses is SIMPLISMA [192], as previously described. Nevertheless, if some inputs are incorrect, as the rank of the matrix, the initial estimates could not perfectly fit the resolution of the unmixing procedure, leading to problems in the decomposition of the signal into the pure contributions of the original data.

As introduced, the purpose of this chapter is the description of the work that has been conducted during this PhD to face this kind of problem, i.e., the selection of the most important information in a complex matrix. In brief, two different lines will be investigated. First, this manuscript will focus on the use of SIMPLISMA in a new and more intuitive way, in order to facilitate the task of the optimal rank selection and extract the purest contributions to be used as initial estimates for the MCR-ALS calculation. This will be the opportunity to introduce the first publication resulting from this thesis work. Then, a second part of the chapter will be dedicated to LIBS imaging. This kind of instrumentation is related to very interesting characteristics that make it very suitable for different chemical areas. For example, LIBS shows a high acquisition rate (up to 1000 spectra/s), and a high sensitivity (major elements to traces can observed). Nevertheless, these aspects can result to be a problem. First of all, it is not easy to deal with a huge amount of data as in LIBS imaging, where millions of spectra can be acquired in a short time (this aspect will also be better described in the Chapter 4 of the present manuscript). In addition, despite the fact that minor compounds can be observed, this task can be very complicated because these pure spectra are represented by a very small quantity of pixels compared with their totality. For this reason, KM clustering was applied in a specific way, trying to overcome the problems faced by a typical investigation and extract more details, i.e., classify major, minor compounds and even trace elements.

2.2. Randomised SIMPLISMA: Using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging

2.2.1. Introduction

SIMPLISMA is a very suitable algorithm in MCR-ALS framework used with the aim of obtaining the initial estimates for the signal unmixing procedure. Nevertheless, it is mandatory to give as an input an optimal rank in order to extract the purest variables and use them in an iterative approach to refine the final results. Normally, the rank estimation is carried out by the investigation of the eigenvalues obtained by a first exploratory analysis using PCA. The limitation shown by this procedure is that if some components are present as few pixels, or with a signal close to the signal-to-noise ratio, it can lead to results that are underestimating the complexity of the original observed matrix. In the same way, considering a rank higher than the real one, this could lead to the extraction of wrong profiles. The first work here discussed, and published in Talanta, Volume 217 (2020) [180], shows an alternative way to use SIMPLISMA. The main purpose using this approach is the one of selecting first the right rank using a graphical interpretation in the PCA space, and then extract the information obtained from the different groups, in order to use the pure spectral signals as initial estimates to obtain at the end the signal unmixing using MCR-ALS. Randomised SIMPLISMA (this is the given name to the presented approach) has shown interesting results, a good rapidity of calculation, and particularly, it can be used in cases in which SIMPLISMA can experience difficulties, such as the investigation of big datasets. Nevertheless, it is important to understand that also randomised SIMPLISMA is influenced by the operator decisions, so it is not an error-free method. On the other hand, offering a graphical interpretation based on the distribution of the purest pixels into the PCA space (as explained in the corresponding paper), this method can clearly be considered a good alternative to deal with, in particular, complicated situations. By way of example, randomised SIMPLISMA can be used when a complex matrix is investigated, in which doubts regarding the real rank of the dataset may arise (e.g., if some minor components related to a small number of pixels are present). Another situation in which this approach has been applied is given by the case in which the dataset is made by thousands (or millions) of spectra. In this case, SIMPLISMA, as other algorithms, can face some issues due to calculation problems.

Talanta 217 (2020) 121024

Contents lists available at ScienceDirect

Talanta

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

Randomised SIMPLISMA: Using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging

Alessandro Nardecchia, Ludovic Duponchel*

Univ. Lille, CNRS, UMR 8516 - LASIRe – LAboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, F-59000, Lille, France

ARTICLE INFO

Keywords:
MCR-ALS
SIMPLISMA
Rank estimation
Initial estimates
Hyperspectral imaging
Big dataset

ABSTRACT

Hyperspectral imaging opens the opportunity in analytical chemistry to investigate always more complex samples by the use of Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) and other signal unmixing techniques, but not without difficulties. Nowadays, one of the principal challenges regarding this kind of analysis is the awkward estimation of the correct chemical rank of the dataset, which represents the total number of pure compounds present in the chemical system. Despite the existence of various algorithms able to focus on this rank evaluation, the method very often used for this task is finally quite simple since it is based on the observation of the eigenvalues generated by the Principal Component Analysis (PCA). Although this method has shown some potential for rank evaluation, it is still difficult to use it on complex and big datasets or when the signal to noise ratio is relatively weak. In this paper, we introduce a new method, based on the SIMPLE-to-use Self-modeling Mixture Analysis (SIMPLISMA) algorithm that we call Randomised SIMPLISMA. The main idea is thus to use random selections of spectra from the initial dataset and to apply the SIMPLISMA approach to each of them. At the end of this step, all selected spectra are observed using PCA where observed clusters can potentially be highlighted and exploited for the tasks we are interested in. With the present paper, we want to highlight in particular the possibility of an easier rank estimation and initial estimates generation when this approach is considered. Datasets of different complexity acquired with various spectroscopic techniques will be explored in order to evaluate the potential of this approach.

1. Introduction

Nowadays, hyperspectral imaging is a useful technique employed in analytical chemistry with the aim to deeply investigate complex matrices of various types [1–6]. In this perspective, one of the most important tasks is to decompose the spectra of mixtures into purest contributions of the components present in the matrix [7–11]. Among all the available techniques used with the purpose of spectral unmixing (also called source separation method in the signal processing community), Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) [12–15] is probably the most suitable method applied in the chemometrics community, as shown in many works, to datasets acquired with various techniques, for instance, separation methods [16–18], and different spectroscopies such as Raman [19,20], UV–Vis [21,22], Mid-Infrared (MIR), Near Infrared (NIR) [23–26], fluorescence [27] and even on very specific techniques such as ion mobility spectrometry (IMS) [28]. In a natural way, MCR-ALS is also widely applied to the resolution of hyperspectral images [29–32]. The basic assumption of MCR-ALS is that the considered data matrix or the multicomponent system has to follow a bilinear model in order to propose its decomposition into the pure individual contributions of concentration profiles and corresponding pure spectra. Despite this great potential, it is well-known that non-unique solutions can be potentially extracted through the presence of rotational ambiguities. This lack of trueness is due to the fact that different sets of pure individual contributions can reproduce the original dataset with the same fit quality. To avoid as much as possible this uncertainty, different constraints can be applied to MCR-ALS in order to force the concentration and spectral profiles to obey certain conditions, as described in many works [33–38]. For instance, non-negativity is the most natural and classical constraint applied in the field of signal unmixing. Another important aspect regarding the use of the MCR-ALS method, but also all source separation methods, is the evaluation of the rank of the data matrix, i.e. finding the appropriate number of pure components present in the system. This task is particularly crucial with MCR-ALS, because it is not nested, contrarily to Principal Component Analysis (PCA) [39]. It means that if the selected rank is incorrect, the algorithm could lead to the extraction of wrong profiles even in the case of a rank

* Corresponding author.
E-mail address: ludovic.duponchel@univ-lille.fr (L. Duponchel).

https://doi.org/10.1016/j.talanta.2020.121024
Received 22 February 2020; Received in revised form 4 April 2020; Accepted 8 April 2020
Available online 18 April 2020

overestimation. The common method used to achieve this task is the observation of the eigenvalues generated by PCA [40]. In a second step, we must also generate initial estimates of pure spectral profiles or concentration profiles which will be refined afterword by the MCR-ALS algorithm. It is generally managed by the use of the simple-to-use interactive self-modeling mixture analysis (SIMPLISMA) [41], a technique based on the concept of variable purity. In general, this tool can lead to good initial estimates though there is not any guarantee about the positivity of the solution. Nevertheless, these guesses can be used as a starting point and refined during the MCR-ALS process. Despite the effectiveness of SIMPLISMA, it can nevertheless show weaknesses, for instance with complex chemical systems, but also when the signal to noise ratio is limited. Moreover, SIMPLISMA cannot sometimes be simply applied when the number of spectra is too large.

The aim of this work is to present an alternative approach based on the SIMPLISMA algorithm, called randomised SIMPLISMA leading to an easier estimation of the rank by the simultaneous use of a pixel-pulling technique and the SIMPLISMA algorithm. Then a graphical exploration of the selected pixels in the PCA-space will allow us to estimate the rank but also observe groups of pixels from which initial estimates will be generated. In order to show the potential of this approach, three different datasets obtained from different instruments (Raman, Autofluorescence and EDX) were selected and processed by the use of randomised SIMPLISMA with the aim of investigating the number of pure components in the data matrix and the generation of initial estimates necessary for MCR-ALS calculations. The last step will be the extraction of all pure contributions on the basis of the information extracted by our method.

2. Material and methods

2.1. Multivariate curve resolution – Alternating Least Squares (MCR-ALS)

The purpose of Multivariate Curve Resolution methods [42] is to extract the relevant information in a mixture system to obtain the pure components through a bilinear model decomposition of the experimental data matrix **D** of dimensions $n \times m$ into the product of the concentration profiles matrix **C** ($n \times k$) containing the concentration of the components present in the system and the corresponding spectral profiles matrix $\mathbf{S}^T$ ($k \times m$). In this notation, n represents the mixture spectra in rows measured at m wavelength that follows the bilinear model, while k is the number of pure components supposed to underlie **D**. The algorithm can be resumed in equation (1), that represents the multiwavelength extension of Lambert-Beer's law in a matrix form:

$$\boldsymbol{D} = \boldsymbol{C}\boldsymbol{S}^T + \boldsymbol{E} \tag{1}$$

with **E** the residual matrix, containing the variability of **D** which is not explained by the model and should be close to the experimental error.

As discussed in the *Introduction*, in order to obtain acceptable results, the MCR-ALS algorithm needs first the rank evaluation and second the generation of initial guesses of the pure components without requiring prior information about the composition of the sample. These tasks will be discussed in the next sections. Furthermore, constraints are applied during the ALS process in order to refine initial estimates, but also and more importantly, to reduce rotational ambiguity due to the non-uniqueness of the pure component MCR decomposition. The final quality of the model depends on two important figures of merit: Lack of Fit (*LOF*), representing the difference among the input data **D** and the data reproduced from the $\mathbf{CS}^T$ product obtained by MCR-ALS and the percentage of variance explained (r^2), shown respectively in equations (2) and (3):

$$LOF\,(\%) = 100\sqrt{\frac{\sum_{ij} e_{ij}^2}{\sum_{ij} d_{ij}^2}} \tag{2}$$

$$r^2 = \frac{\sum_{ij} d_{ij}^2 - \sum_{ij} e_{ij}^2}{\sum_{ij} d_{ij}^2} \tag{3}$$

where d_{ij} is the spectral value of the mixture i at the wavelength j and e_{ij} is the associated error. It is worth stopping on another important aspect regarding the investigation of a hyperspectral image in the MCR-ALS framework. In this case, the three-dimensional cube of dimensions $x \times y \times \lambda$ where x and y represent the number of pixels in the two spatial directions and λ the direction of the spectral variable will have to be unfolded prior MCR calculations in order to allow the decomposition of **D** into the contribution of **C** and $\mathbf{S}^T$ matrices. In the last step, the **C** matrix will be refolded in order to retrieve concentration maps of each pure compound extracted by MCR-ALS.

2.2. Rank evaluation using PCA

As already discussed, the most used method in order to evaluate the rank of a data matrix prior MCR-ALS analysis is based on the observation of the scree plot of the eigenvalues associated with principal components obtained by PCA. This method was introduced the first time by Cattell [43] and it is based on the principle of the meaningful information expressed by a particular factor (or component). It is displayed as a downward curve in which eigenvalues compare on the Y axes, the most important first, while on the X are reported the various components. Because the useful information decreases gradually taking into account the subsequent eigenvalues (as also the distance in the graphic between them), the strategy is to consider only a limited number of components. Using a scree plot, the choice of this value is done by the interpretation of the curve, in which the right value should correspond to the 'elbow' in the graph, where the eigenvalues level off. In this way, we try to set a threshold above which we will consider significant eigenvalues that carry chemical information. Despite the method can seem easy and immediate, a subjective and arbitrative interpretation is often observed, especially when the noise level is significant. Moreover, it also remains very delicate with this method to detect minor compounds that are then very close to the noise level.

2.3. Simple-to-use interactive self-modeling mixture analysis (SIMPLISMA)

SIMPLISMA [41] was one of the very first multivariate curve resolution methods used in spectroscopy [44–46]. It is based on two basic statistical tools which are the mean and the standard deviation. In fact, the central task of this approach is the selection of so-called pure variables from the data matrix **D**. A pure variable is a variable that depends on the contribution of only one component. The first purity of the variable i is then estimated with the purity index $p_i^{(1)}$:

$$p_i^{(1)} = \frac{\boldsymbol{\sigma}_i}{\mu_i + \boldsymbol{\alpha}} \quad \textit{for } i = 1, \&, n \tag{4}$$

with $\mu_i = \frac{\sum_{v=1}^{m} d_{i,v}}{m}$ and $\boldsymbol{\sigma}_i = \sqrt{\frac{\sum_{v=1}^{m} d_{i,v}-\mu_i}{m}}$ for $i = 1, \ldots, n$ (5).

The user-defined parameter α avoids giving a high purity value to a variable with a low mean and therefore could only be noise. Then the first purest variable will have the highest $p_i^{(1)}$ value. In a second step, it is necessary to subtract the contribution of this first pure variable from matrix **D** before continuing the search for a second pure variable. A weighting parameter $w_i^{(2)}$ is thus considered in order to reduce the influence of other variables that would be correlated to the first pure variable. More details about this parameter are given in other works [47]. The second purity $p_i^{(2)}$ of a variable i is then defined by:

$$p_i^{(2)} = w_i^{(2)}\frac{\boldsymbol{\sigma}_i}{\mu_i + \boldsymbol{\alpha}} \quad \textit{for } i = 1, \&, n \tag{6}$$

Again, the next purest variable has the highest $p_i^{(2)}$ value. The following purest variables are of course obtained by iterating these calculations until the number of variables corresponding to a given rank is

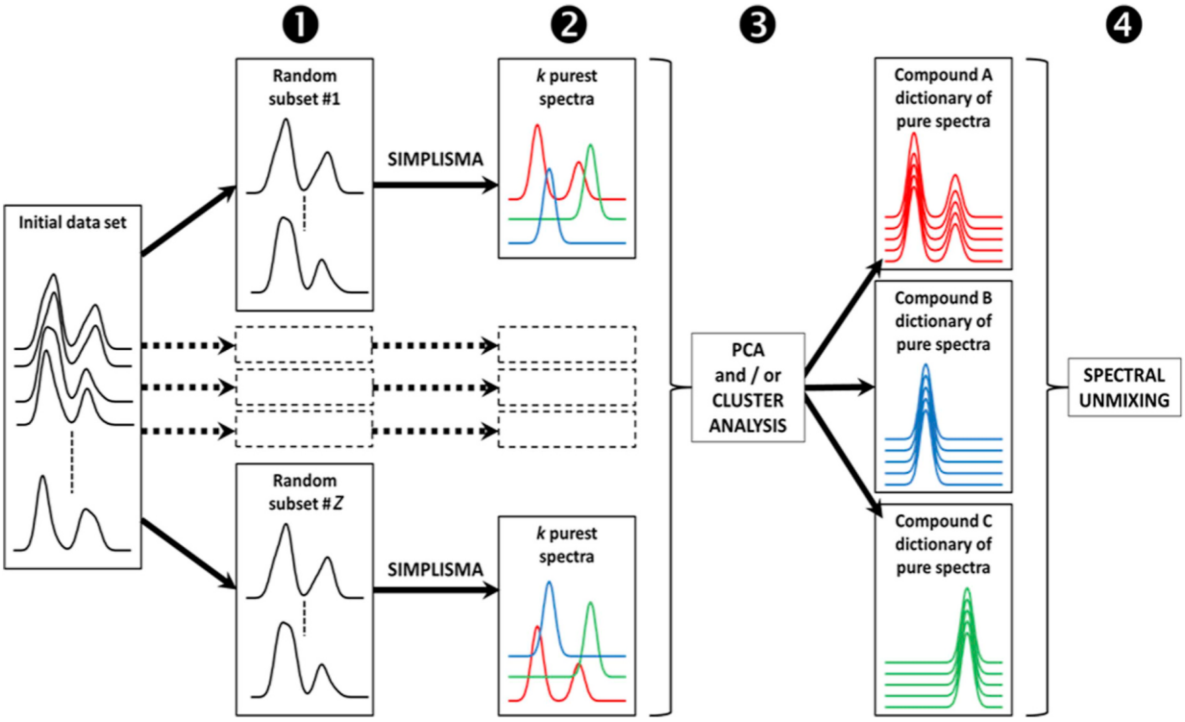

❶
❷
❸
❹
Initial data set
Random subset #1
Random subset #Z
SIMPLISMA
SIMPLISMA
k purest spectra
k purest spectra
PCA and / or CLUSTER ANALYSIS
Compound A dictionary of pure spectra
Compound B dictionary of pure spectra
Compound C dictionary of pure spectra
SPECTRAL UNMIXING

Fig. 1. Principle of the randomised SIMPLISMA approach.

obtained. It is often forgotten that this extraction of pure variables can be done in both dimensions of the matrix **D**. In this way, the selection of variables along the columns of **D** allows us to obtain estimates of concentration profiles while following the rows we obtain estimates of purest spectra. We will consider the latter case in this work because we potentially have a higher probability to select purest pixels for the considered spectroscopies.

2.4. Randomised SIMPLISMA

The main concept of the proposed randomised SIMPLISMA approach is very simple because it is based on random pixel selections on which SIMPLIMA will be applied. Thus it can be resumed in four steps, as reported in the scheme in Fig. 1:

1) The first and fundamental step is to generate *z* random subsets of the whole dataset, with the idea of taking a small percentage of all pixels in the matrix **D.** In this way, we give all pixels a chance to be explored whether they belong to a major or minor class of compound.
2) Then the SIMPLISMA algorithm is applied to each generated subset. However, because we don't want to fix the rank in advance, SIMPLISMA is systematically applied to each subset considering a varying number of pure contributions *k* from 2 to $k_{\max}$. At this level, *k* pixels are selected per subset, which represent a total number of selected spectra at most equal to *zk*. However, the total number of selected pixels is often much weaker since identical spectra can be selected from different subsets.
3) In the next step, the new dataset of selected pixels is explored with PCA. Natural groupings of pixels corresponding to pure compounds are then observed in scores plots. The idea is then to count these clusters in order to estimate the rank. Given the rank, spectra belonging to a specific class of compound are selected by hand in the scores plot. In this way, we can say that we generate a dictionary of spectra for each pure compound in the investigated chemical system. We are fully aware that our approach based on visual inspection may appear subjective. So naturally we could say that an automatic cluster analysis should be used. Nevertheless, we know that the literature is full of so-called ultimate metrics to automatically count the number of clusters in an optimal way. The only problem is that there is an optimal metric for each considered data set, which adapts to the variations in point densities in the clusters but also to their insect structures, which are not always Gaussian. It is for all these reasons that we have preferred a visual approach which is finally no more debatable than an arbitrarily chosen metric.
4) In order to exploit the results of the previous step, the mean spectrum of each dictionary can be used as an initial guess in MCR-ALS.

2.5. Dataset #1

The first dataset corresponds to a Raman analysis of an oil-in-water emulsion sample. It has been acquired by Andrew et al. [48]. The data cube consists of 60 pixels by 60 pixels corresponding to a 1 μm^2 area each on the sample surface. The spectral range is between 950 cm^{-1} and 1800^{-1} corresponding to 253 wavenumbers. Further details about the instrumental and acquisition setup may be obtained through the original work [48].

2.6. Dataset #2

The second dataset has been acquired using an auto-fluorescence imaging microscope. It is focused on the growing process of wheat plants, a precise stage of the wheat grain development being investigated. The freshly harvested grain samples were frozen and cut in the equatorial region using a cryotome (HM 500 OM, Microm) into 20 μm cross-sections. The sample was analyzed using a confocal laser-scanning system (A1, Nikon) equipped with an ×40 objective for confocal imaging in order to obtain an auto-fluorescence response. Three excitation wavelengths have been considered: 375 nm (UV), 488 nm (blue) and 561 nm (green). As a consequence, three hyper-spectral images have been acquired by collecting emitted light from 404 to 714 nm for the UV excitation, 504–744 nm for the blue excitation and 574–744 nm for the green one with a 10 nm step between spectral variables. The size of the image is 512 pixels by 512 pixels (0.62 μm per pixel) corresponding to a total of 262,144 emission spectra for 75 variables, obtained by a data augmentation strategy apply on the wavelength dimension from each excitation wavelength range. Further details of this specific dataset are described in the work of Ghaffari et al. [49].

2.7. Dataset #3

The last dataset is a hyperspectral image of two cancer cells treated with a bromine-containing prodrug. More specifically, P31 cells were grown a gold-coated silicon wafer. Spectra have been acquired with a Fei Quanta 200 electron microscope with an EDX detector. Scanning electron microscope images were obtained in secondary and backscattered electron mode using an acceleration voltage of 5 kV. The size of the image is 101 pixels by 176 pixels, with a size of 0.4 μm per pixel, for a total of 17,776 emission spectra for 13 spectral variables, corresponding to 13 different elements (Au, Br, C, Ca, Cl, K, Mg, N, Na, O, P, Pd, S). Further details about the dataset are described in the work of Ofner et al. [50].

3. Results and discussion

3.1. Dataset #1

Before entering into a real chemometric exploration of a hyperspectral dataset, it is always interesting to generate a global integration image. This procedure is very simple since it consists of the summation of all intensities for each pixel over the whole spectral domain. Of course, we lose the chemical information but it is nevertheless possible to observe structures within the explored sample. The global integration image of the oil-in-water emulsion is presented in Fig. 2. Intuitively, it is possible to recognize at least three structures: a big drop (upper left area with highest intensity values) in contrast with a surrounding area (lowest values of intensity) and a less well-defined area in the lower right part of the image. Logically and without taking too much risk, we can imagine having drops of oil and the aqueous phase. Furthermore, the border of the drop shows different levels of intensity compared with its internal part, which could suggest the potential presence of a more complex chemical structure. To get a better idea of the complexity of the dataset, an investigation by using PCA is performed (Fig. 3). More specifically, Fig. 3a presents the first 8 score maps in decreasing order of explained variance. The first 7 components seem to have structures even if the last ones are rather noisy. Beside PC1 and PC2 that mainly describe the big drop, some specific aspects are highlighted: as an example, PC3 focuses on two small drops while PC5 and PC6 seem to describe the oil-water interface. The remaining PCs contain supplementary information, but the amount of noise hinders a meaningful interpretation at this level. Of course, another aspect of PCA is the observation of eigenvalues in order to potentially estimate the rank of the considered data cube. However, we quickly observe in Fig. 3b that the rank evaluation is difficult due to a smooth decrease of the eigenvalues in the scree plot. Indeed, it is quite impossible to select a threshold above which we could consider significant variances from chemical compounds alone. It is also interesting to represent all spectra in a three-dimensional representation of scores along PC1, PC2, and PC3, which we often do in chemometrics but finally not very much in the specific framework of imaging. From Fig. 3c, we quickly understand that it is indeed very difficult to extract information despite a dataset with not so many spectra. Thus, even if PCA usually allows us to estimate the rank, it remains difficult here to propose unambiguously a number of pure compounds present in this chemical system from all previous representations. However, it is interesting to know that despite these conflicting elements, a rank of 4 is often used for this particular dataset.

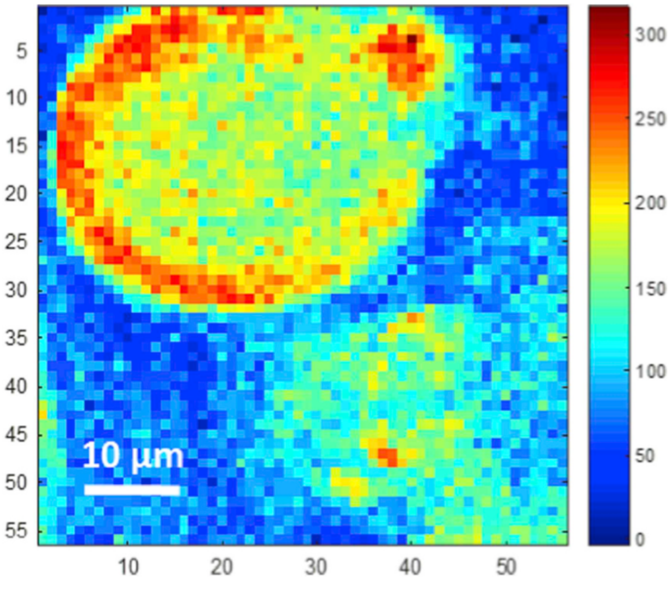


Fig. 2. Global integration image of Oil-in-water emulsion dataset.

After this first conventional chemometric investigation, we now want to show what our strategy can bring to the exploration of this same dataset. In a first pixel resampling step, 1000 datasets have been generated by randomly selecting 10% of pixels from the whole dataset (i.e. 360 pixels on a total of 3600 in this case).

We can then ask ourselves the question of this specific choice for these two parameters, which will be approximately the same for the other two data sets. First of all, we observed that at least 500 subsets were needed to observe reproducible results if the whole procedure was replicated several times. Second, regarding the percentage of selected pixels, a value below 10% did not allow the observation of minor compounds, while a higher value densified the clusters in such a way that they tended to overlap or even merge into a single one. As a conquence, more than 500 subsets and 10% of selected pixels was a good compromise for all the explored data sets in this work.

The second step of our strategy was to apply randomised SIMPLISMA to each of the 1000-pixel subsets for different values of k, i.e. the number of purest pixels extracted with SIMPLISMA. The idea is not, of course, to set the value of k in advance, since one of the objectives of this method is specifically to determine its optimal value. We will, therefore, observe the evolution of pixel selection as the value of k increases. Fig. 4a shows a PCA of all the purest pixels selected from the 1000 subsets for $k = 3$. Only 150 spectra are finally present in the score plot because many of them have been selected several times in different subsets. On theory, we could effectively extract 3 × 1000 spectra from the initial dataset. Then, we notice that pixels are organized in 4 clusters in this PCA space. It is precisely the principle of this approach to consider each cluster as a representation of a pure compound allowing some variations around a mean point. In this way, we have a kind of spectral dictionary for each of them. It is also interesting to see in Fig. 4a more differences between eigenvalues in the corresponding scree plot compared with Fig. 3b. Thus at this stage, we detect at least 4 chemical species. The idea is now to look at the evolution of the pixel selection when the k value increases. Thus for two successive values of k, any appearance of a new cluster would correspond to the detection of the new family of a compound and therefore mechanically to an increase of the rank. Fig. 4b shows PCA results when considering simultaneously purest pixels for $k = 3$ (in blue) and $k = 4$ (in red). By the way, 220 over 3600 spectra are now selected when $k = 4$. It should be noted that most of the red dots ($k = 4$) are projected into clusters already described by blue dots ($k = 3$). Nevertheless, a number of spectra (in red) are located in a new area represented by a solid line ellipse. Thus a new compound is detected. In Fig. 4c, the comparison of the pixel selections between for $k = 4$ (in blue) and $k = 5$ (in red) highlights the presence of a new cluster (also represented by a solid line ellipse). If we continue this process, pixel selections between for $k = 5$ (in blue) and $k = 6$ (in red) are compared in Fig. 4d. Again, a number of red dots are located in a new area of the PCA space highlighting a new

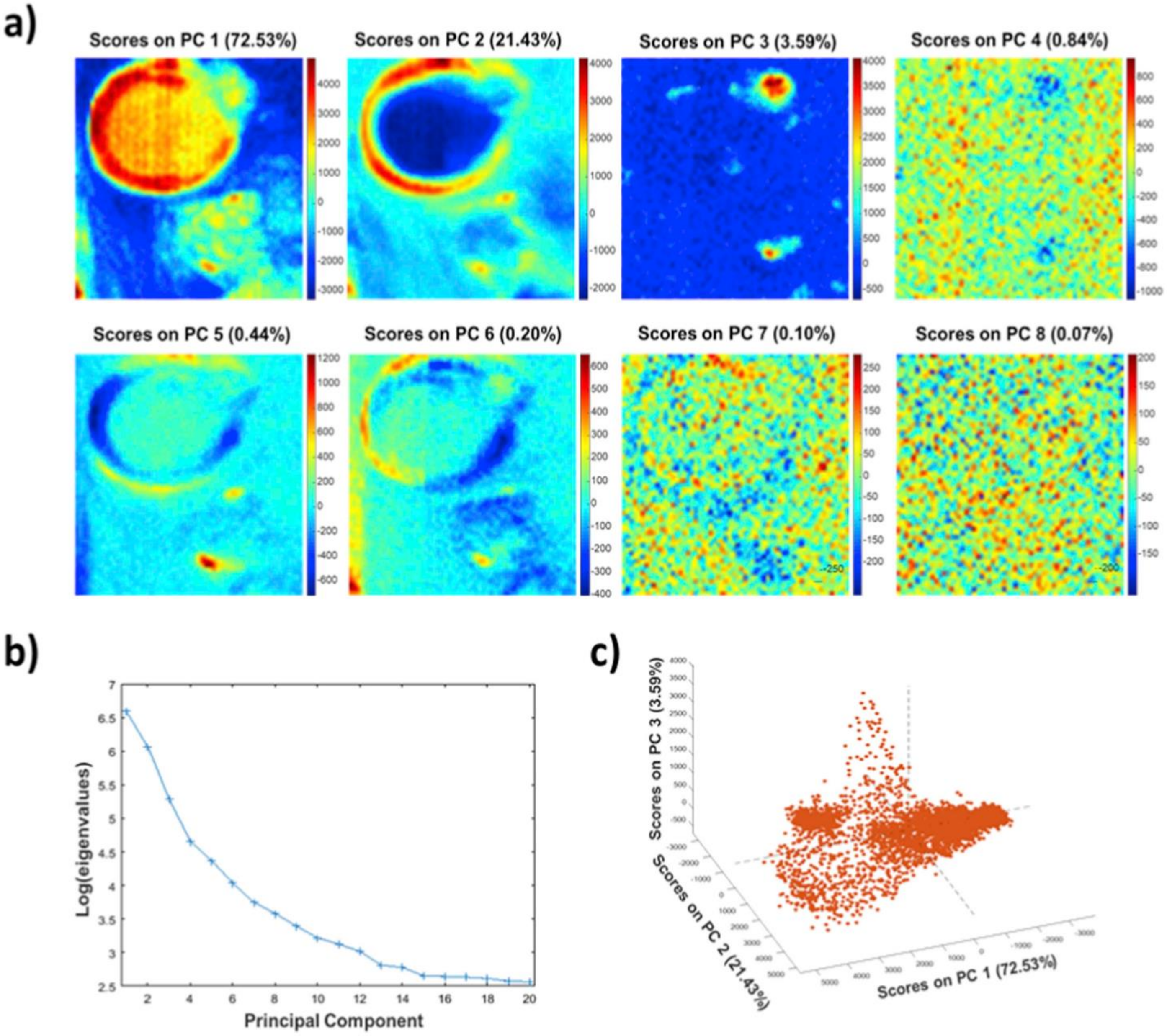


Fig. 3. Principal Component Analysis of the Oil-in-water emulsion dataset. a) The first eight score maps. b) The scree plot of eigenvalues. c) A three-dimensional representation of scores along PC1, PC2, and PC3.

class of compounds. At this step, a rank of 7 is considered. At the same time, we see that the corresponding scree plot shows clearer differences between the eigenvalues on which we could begin to consider a threshold. From these first results, we could easily believe that simply increasing the value of k is enough to increase the number of clusters and thus the chemical rank. This is not the case as we can see in Fig. 4e where pixel selections between $k = 6$ (in blue) and $k = 7$ (in red) are compared. Indeed, all red dots are projected in all areas already defined by the previous calculations. In this way, we observe a certain stabilization of the cluster structure. Thus, we can say that the pixel selection obtained for $k = 6$ with its 7 clusters is a good representation of the complexity of the dataset. We now propose to look in detail at each of these 7 clusters. Fig. 5 shows the spectra contained in each cluster in overlay mode and their localization (yellow pixels) on the sample surface. Thus for each of the clusters, we first notice a good consistency between the spectra. With regard to spatial distributions, it is already observed that some classes are located on specific parts of the emulsion. For example, class 1 is specifically located on two small drops, class 3 seems to describe the inside of the largest drop, and classes 2 and 7 are located on the edge of the large drop. It is more difficult to define a location for the other three classes, but we can still say that the aqueous phase must be part of it. We must not lose sight of the fact that our approach aims to select the purest pixels. In other words, the spectra shown in Fig. 5 can be weakly mixed, which makes the task of spectral interpretation all the more difficult. It is in this sense that the MCR-ALS method is then used to refine these solutions.

As part of the MCR-ALS method, we now need initial guesses of spectra for each of the 7 classes. For simplicity, we use here the average spectrum of each class. It is interesting to compare these 7 spectra with the 7 estimates that could be obtained directly with SIMPLISMA on the whole dataset of 3600 spectra (figure S1 in supplementary material). It is then not difficult to see that a better signal-to-noise ratio is obtained on the generated spectra with randomised SIMPLISMA.

Finally, Fig. 6 shows the MCR-ALS results considering different configurations. In the top and the middle panel, a rank of 4 and 7 have been respectively considered with an estimation of initial guesses obtained from the standard SIMPLISMA algorithm on the whole dataset. The bottom panel shows the MCR-ALS decomposition when initial estimates are generated from randomised SIMPLISMA. If we look at the rank of 4, we see that the first pure chemical map describes the small

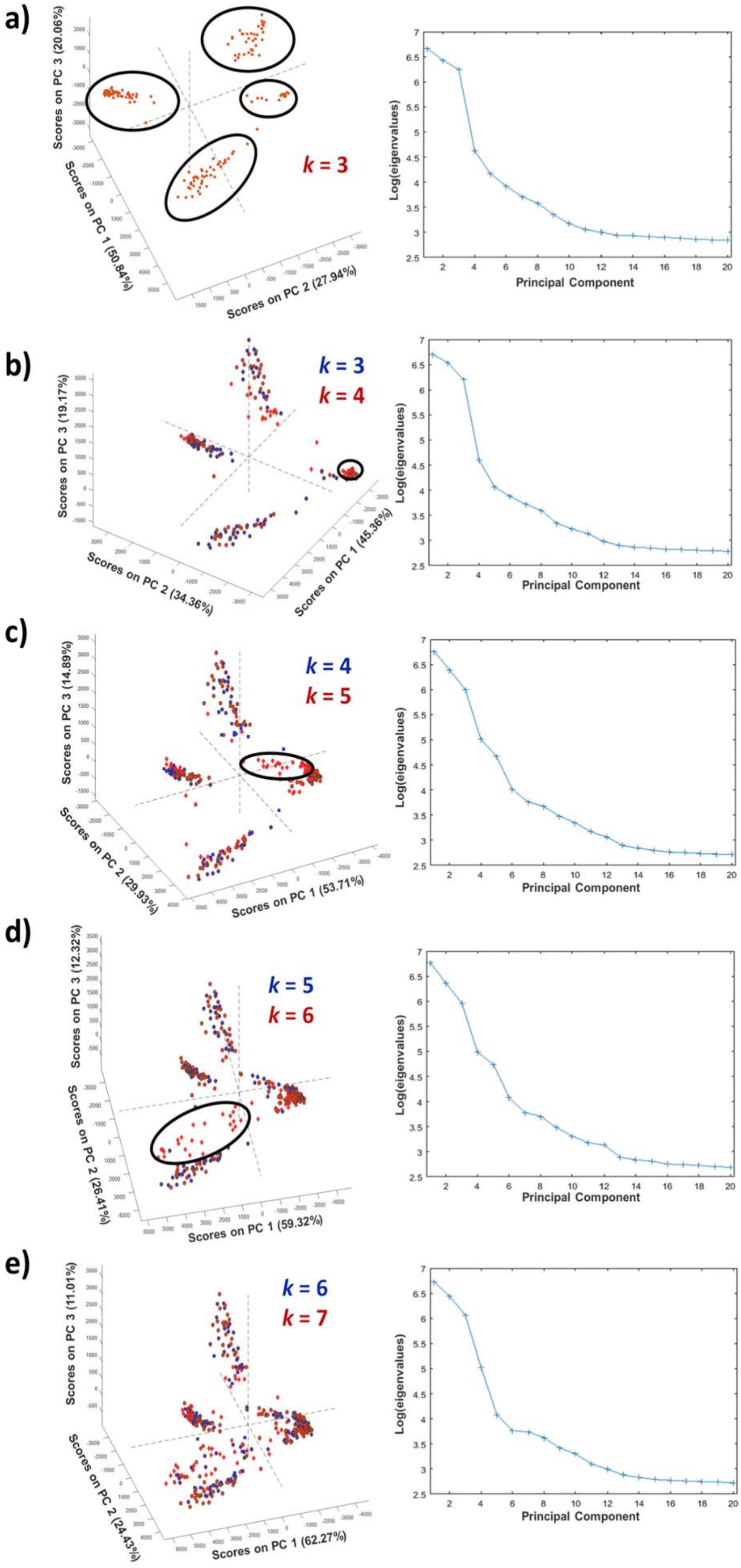


Fig. 4. PCA exploration of pixels selected by randomised SIMPLISMA as a function of *k*.

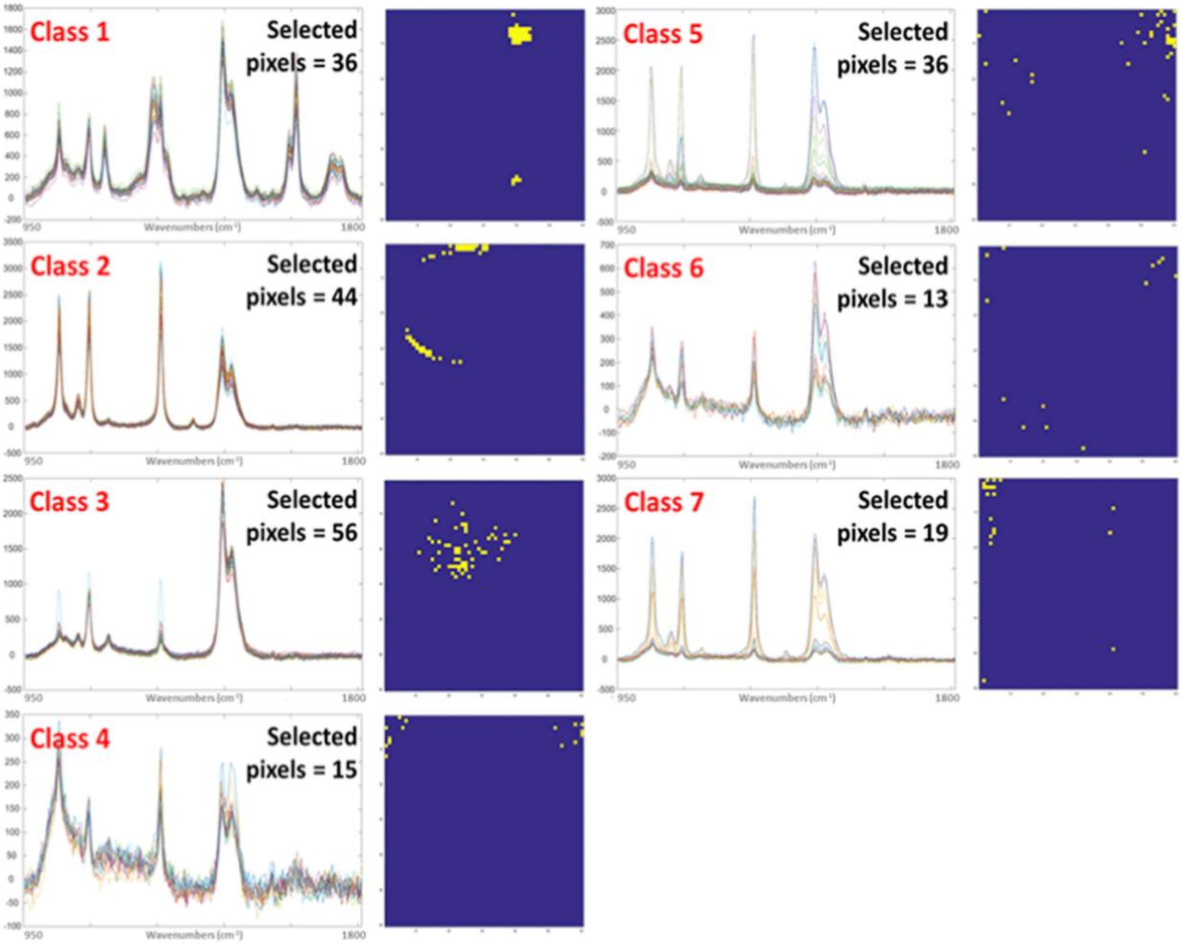


Fig. 5. Spectra of the purest pixels for each of the 7 classes et their location.

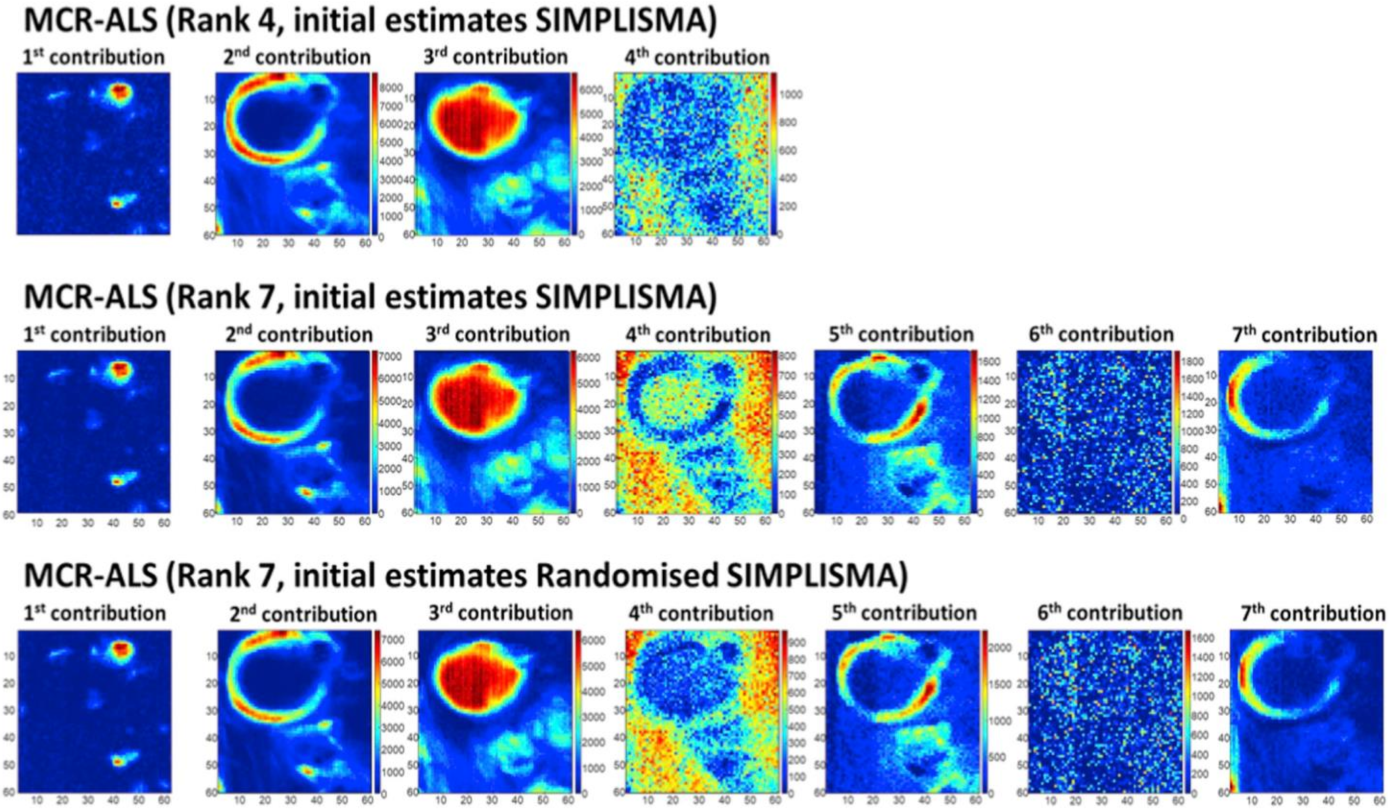


Fig. 6. Extracted concentration maps of MCR-ALS considering different rank and initial estimates.

drops, the second and third ones the big drop and its edge, and the last one the aqueous phase. A lack of fit and R^2 values of 7.76% and 99.39% are respectively obtained in this condition which is not so bad for the considered signal-to-noise ratio. If we now look at the results obtained for a rank of 7 for the two strategies (i.e. the classical and randomised SIMPLISMA), almost the same figures of merit are observed (LOF ≈ 3.23%, R^2 ≈ 99.65%). Contributions 1 and 3 are very similar to those previously obtained with a rank of 4. The greatest change is observed for the description of the border of the drops since three distinct contributions are now extracted (contributions 2, 4 and 7) against only one in the previous resolution. As for contribution 6, it seems a priori questionable because it is not particularly structured but the corresponding pure spectra (bold curves, figure S2 in supplementary material) mainly explain a variance related to a baseline deformation badly corrected by the spectral-preprocessing. This contribution is, therefore, an important part of this resolution. The greatest difference is observed for the contribution 5 corresponding to the aqueous phase. A more contrasted and less ambiguous image is extracted with randomised SIMPLISMA. Indeed, we observe a significant contribution inside the drop when the classical SIMPLISMA method is used which is rather incoherent compared to the knowledge about the behaviour of the molecules considered in this chemical system. From a general point of view, some people might say that for a rank of 7 the extraction results are not so different between the classical method and our strategy but we must not forget that we would never have used such a rank with the classical approach.

3.2. Dataset #2

The second data set is particularly interesting because of its size and the type of spectroscopy envisaged. Indeed, we have a much larger number of spectra, namely 262,144 in the data cube. We also selected very original spectroscopy, namely the autofluorescence one with only 75 emission wavelengths, which is much lower than in the previous case. Beyond this small number of spectral variables, we can expect a significant spectral overlap between chemical species related to an intrinsically large bandwidth in this spectroscopy.

As in the previous case, it is natural to make first a PCA of the complete dataset. Thus, from the score images (Figure S3), we observe that the two first principal components explain 93.81% of the total variance. At this stage, even if it seems possible to observe some details on the score maps from the third principal component, it remains difficult to certify at this level of the investigation that they correspond to relevant chemical information. It is now interesting to look at the evolution of the eigenvalues in the scree plot given in Fig. 7a. As we can see, a sharp decrease in values is observed after the second principal component, which would potentially indicate a rank of two with the traditional threshold-based method. It is also interesting to look at the three-dimensional representation of scores along PC1, PC2, and PC3 in this same figure. With these 262,144 spectra, the point density is so high that it is impossible to see details on the intrinsic structure of the dataset except for a global V-shape.

Then randomised SIMPLISMA has been applied considering the generation of 500 subsets with 10% of pixels randomly selected from the whole dataset. Fig. 7b shows a PCA of the 131 pixels selected from these 500 subsets for $k = 3$. Then it is very easy to detect the presence of 3 clusters. Moreover, the third eigenvalue of the corresponding scree plot is now detached from the noise level. As with the previous dataset, we will now analyze the evolution of pixel selection as the k value increases. PCA results considering simultaneously purest pixels for $k = 3$ (in blue) and $k = 4$ (in red) are given in Fig. 7c. By the way, 227 spectra are now selected when $k = 4$. It is then obvious that a new cluster is detected indicating a rank of 4. By continuing this exploration for the values of k equals 4 and 5, new clusters are not highlighted. As a conclusion, pixels selected with $k = 4$ represent the intrinsic structure of this dataset, which finally has a rank of 4. Spectra contained in each cluster are represented in figure S4 in the supplementary material. Again a good consistency between the spectra of a cluster is observed. With regard to pixel location (yellow pixels in figure S4), it is particularly difficult here to link this information to the biological structure because of the very small number of spectra selected compared to the 262,144 pixels of the sample surface. Finally, Figure S5 proposes the 4 pure concentration maps and corresponding pure spectra extracted by MCR-ALS using the mean spectrum of each cluster as an initial estimate. It can, therefore, be concluded that without this approach, we would certainly not have extracted contributions 3 and 4. These two contributions have very small variations in concentration at a very local level but specific spectral contributions. We must also insist on the fact that it is not possible to apply the classical SIMPLISMA approach to the 262,144 spectra of the dataset for RAM problems, even on very large computers.

3.3. Dataset #3

The originality of this last dataset does not lie in its size but in the chemical information it contains. Thus, the variables describing each pixel in this data cube are elemental concentrations obtained from Energy Dispersive X-Ray Analysis (EDX). Two tumor cells treated with a bromine-containing prodrug are explored in this case. As usual, we will start with a PCA of the complete dataset. Score images are given in figure S6. It is then possible to observe 4 or even 5 chemical contributions defining both the cells and the support. In parallel with that, the scree plot in Fig. 8a seems to indicate a rank of 3. This observation is very interesting because it is quite symptomatic of the use of PCA in imaging when the number of spectra is very large. Indeed, we can see for example that there is a potential contribution expressed on the fourth score maps but the number of pixels it concerns is so small compared to the total number of pixels, that they only induce a very small variance of 0.67% almost undetectable in the scree plot. On the basis of this information, many of us would certainly have selected a chemical rank of 3. Fig. 8a also proposes the three-dimensional representation of scores along PC1, PC2, and PC3 for the 17,776 pseudo-spectra. Once again, the density of points is so high that it is impossible to see a particular data structure.

Randomised SIMPLISMA has been applied to this dataset considering the generation of 500 subsets with 10% of pixels randomly selected from the whole dataset. Fig. 8b shows a PCA of the 89 pixels selected when $k = 2$. Thus 2 clusters are detected at this step. Then purest pixels obtained from $k = 2$ (in blue) and $k = 3$ (in red) have been explored with PCA (Fig. 8c). We, therefore, observe in this representation two new clusters. Spectra of these two clusters are also shown in Fig. 8c in order to highlight the chemical differences. The chemical rank is now 4. For the next PCA on the purest pixels for $k = 3$ (in blue) and $k = 4$ (in red), the last cluster is detected (Fig. 8d). No additional clusters are observed in Fig. 8e which makes it possible to set a rank of 5. Figure S7 in the supplementary material shows spectra selected in each of the 5 clusters. We observe a good consistency between the spectra of a cluster. Except for class 3 located mainly on a cell, it is rather delicate to strictly associate the others to a sample structure. Finally, Figure S8 (in the supplementary material) shows the MCR-ALS extractions obtained from cluster averages. Thus we quickly notice that each pure contribution is mainly influenced by one particular element. The first contribution contains bromine so we can localize the prodrug in the cell volume. Palladium particles are also detected inside the cells from contribution 2. Obviously, contribution 4 corresponds to the gold on the surface of the wafer. Contribution 5 reports on the presence of other elements such as phosphorus also present on the surface mainly related to cell preparation. Finally, contribution 3 is very interesting since it expresses the presence of the minor compound present only on a few isolated pixels.

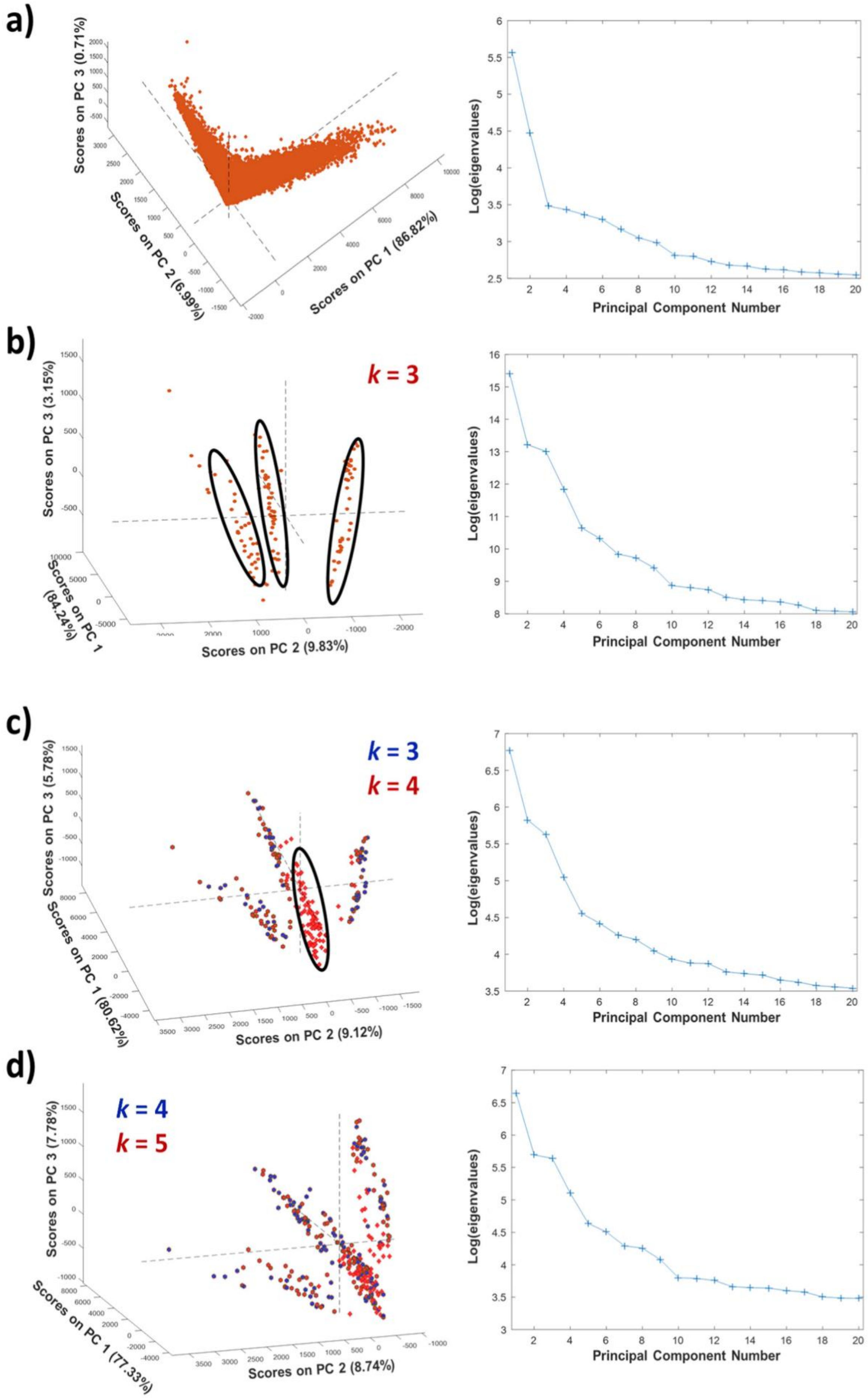


Fig. 7. a) PCA of the whole dataset. b-d) PCA exploration of pixels selected by randomised SIMPLISMA as a function of *k*.

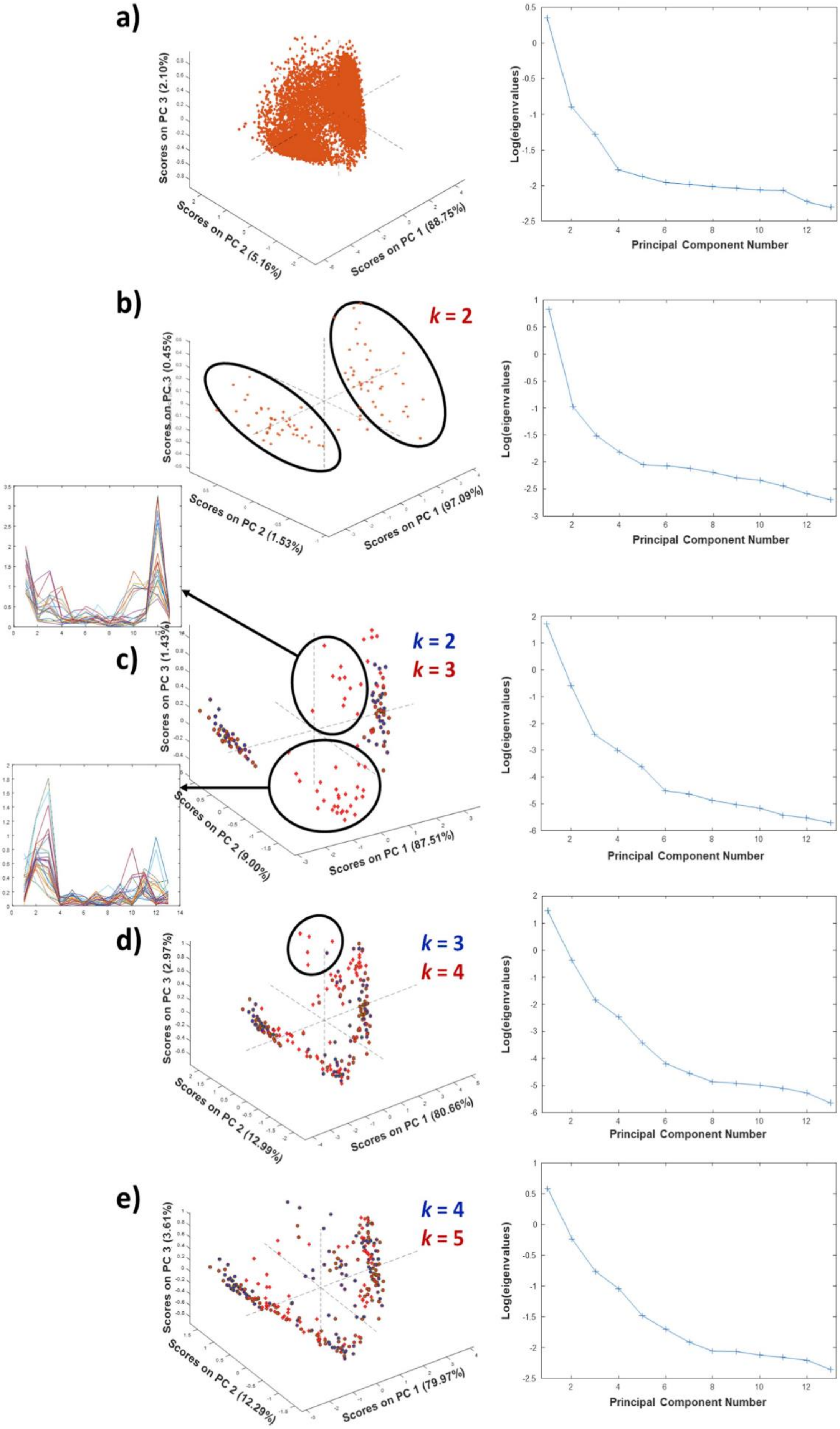


Fig. 8. a) PCA of the whole dataset. b-e) PCA exploration of pixels selected by randomised SIMPLISMA as a function of *k*.

4. Conclusion

Numerous publications demonstrate on a daily basis the strong potential of the MCR-ALS method for a priori-free extraction of the contributions of all pure compounds present in complex chemical systems. This approach is particularly useful in the spectroscopic imaging framework, for which unsupervised exploration is often the only alternative in view of the complexity of the samples and the absence of reference methods. Nevertheless, the main constraint of a signal unmixing approach such as MCR-ALS remains the chemical rank evaluation which thus conditions all the relevance of the extractions. Of course, principal component analysis can help us in this task, but this study has shown on different datasets that it is not always suitable for the simultaneous detection of major and minor compounds. It is also very sensitive to the signal-to-noise ratio and not well-suited to big datasets. The aim of this work was then to present a new concept called randomised SIMPLISMA based on pixel resampling and the original SIMPLISMA algorithm. Through the different datasets, we were able to show that our approach not only facilitates the estimation of rank but also provides initial estimates of pure compounds. Moreover, it has been possible to manage datasets containing several hundred thousand spectra where SIMPLISMA simply cannot be directly applied. Another peculiarity of our approach also lies in the generation of a real dictionary of spectra for each pure compound. This makes it possible to better locate a given contribution even before curve resolution.

The perspectives of this work are twofold. First, randomised SIMPLISMA will be evaluated on even bigger datasets containing several million spectra. Second, as can be seen, the variability present in a dictionary of a given contribution is rather little exploited since finally, we use its mean as an initial estimate prior MCR-ALS. Thus, as can be done today in the remote sensing community, we could, for example, consider that the spectra of a dictionary would be different pure representations of a given compound. We would then consider a non-linear model, which may make sense in some cases of matter-radiation interaction.

CRediT authorship contribution statement

Alessandro Nardecchia: Data curation, Formal analysis, Investigation, Software, Supervision, Validation, Writing - original draft, Writing - review & editing. **Ludovic Duponchel:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing - original draft, Writing - review & editing.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2020.121024.

References

[1] G. Lu, B. Fei, Medical hyperspectral imaging: a review, J. Biomed. Optic. 19 (2014) 010901, , https://doi.org/10.1117/1.JBO.19.1.010901.

[2] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, Trends Food Sci. Technol. 18 (2007) 590–598, https://doi.org/10.1016/j.tifs.2007.06.001.

[3] V. Studer, J. Bobin, M. Chahid, H.S. Mousavi, E. Candes, M. Dahan, Compressive fluorescence microscopy for biological and hyperspectral imaging, Proc. Natl. Acad. Sci. Unit. States Am. 109 (2012) E1679–E1687, https://doi.org/10.1073/pnas.1119511109.

[4] H. Liang, Advances in multispectral and hyperspectral imaging for archaeology and art conservation, Appl. Phys. A 106 (2012) 309–323, https://doi.org/10.1007/s00339-011-6689-1.

[5] G.J. Edelman, E. Gaston, T.G. van Leeuwen, P.J. Cullen, M.C.G. Aalders, Hyperspectral imaging for non-contact analysis of forensic traces, Forensic Sci. Int. 223 (2012) 28–39, https://doi.org/10.1016/j.forsciint.2012.09.012.

[6] G. ElMasry, N. Wang, A. ElSayed, M. Ngadi, Hyperspectral imaging for non-destructive determination of some quality attributes for strawberry, J. Food Eng. 81 (2007) 98–107, https://doi.org/10.1016/j.jfoodeng.2006.10.016.

[7] K. Awa, T. Okumura, H. Shinzawa, M. Otsuka, Y. Ozaki, Self-modeling curve resolution (SMCR) analysis of near-infrared (NIR) imaging data of pharmaceutical tablets, Anal. Chim. Acta 619 (2008) 81–86, https://doi.org/10.1016/j.aca.2008.02.033.

[8] B. Vajna, G. Patyi, Z. Nagy, A. Bódis, A. Farkas, G. Marosi, Comparison of chemometric methods in the analysis of pharmaceuticals with hyperspectral Raman imaging, J. Raman Spectrosc. 42 (2011), https://doi.org/10.1002/jrs.2943 1977–1986.

[9] X. Zhang, A. de Juan, R. Tauler, Multivariate curve resolution applied to hyperspectral imaging analysis of chocolate samples, Appl. Spectrosc. 69 (2015) 993–1003, https://doi.org/10.1366/14-07819.

[10] A.M. Siddiqi, H. Li, F. Faruque, W. Williams, K. Lai, M. Hughson, S. Bigler, J. Beach, W. Johnson, Use of hyperspectral imaging to distinguish normal, precancerous, and cancerous cells, Cancer 114 (2008) 13–21, https://doi.org/10.1002/cncr.23286.

[11] J.M. Amigo, I. Martí, A. Gowen, Hyperspectral imaging and chemometrics, Data Handling in Science and Technology, Elsevier, 2013, pp. 343–370, https://doi.org/10.1016/B978-0-444-59528-7.00009-0.

[12] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemometr. 9 (1995) 31–58, https://doi.org/10.1002/cem.1180090105.

[13] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemometr. Intell. Lab. Syst. 76 (2005) 101–110, https://doi.org/10.1016/j.chemolab.2004.12.007.

[14] A. de Juan, R. Tauler, Multivariate curve resolution (MCR) from 2000: progress in concepts and applications, Crit. Rev. Anal. Chem. 36 (2006) 163–176, https://doi.org/10.1080/10408340600970005.

[15] 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.

[16] F. Berbel, E. Kapoya, J.M. Díaz-Cruz, C. Ariño, M. Esteban, R. Tauler, Multivariate resolution of coeluted peaks in hyphenated liquid chromatography - linear sweep voltammetry, Electroanalysis 15 (2003) 499–508, https://doi.org/10.1002/elan.200390060.

[17] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods. 6 (2014) 4964–4976, https://doi.org/10.1039/C4AY00571F.

[18] K. Johnson, A.D. Juan, S.C. Rutan, Three-way Data Analysis of Pollutant Degradation Profiles Monitored Using Liquid Chromatography-Diode Array Detection, (1999), p. 11.

[19] Y. Xie, W. Cao, S. Krishnan, H. Lin, N. Cauchon, Characterization of mannitol polymorphic forms in lyophilized protein formulations using a multivariate curve resolution (MCR)-Based Raman spectroscopic method, Pharm. Res. (N. Y.) 25 (2008) 2292–2301, https://doi.org/10.1007/s11095-008-9624-1.

[20] K.R. Fega, A.S. Wilcox, D. Ben-Amotz, Application of Raman multivariate curve resolution to solvation-shell spectroscopy, Appl. Spectrosc. 66 (2012) 282–288, https://doi.org/10.1366/11-06442.

[21] S. Mas, A. de Juan, S. Lacorte, R. Tauler, Photodegradation study of decabromodiphenyl ether by UV spectrophotometry and a hybrid hard- and soft-modelling approach, Anal. Chim. Acta 618 (2008) 18–28, https://doi.org/10.1016/j.aca.2008.04.044.

[22] A. Jayaraman, S. Mas, R. Tauler, A. de Juan, Study of the photodegradation of 2-bromophenol under UV and sunlight by spectroscopic, chromatographic and chemometric techniques, J. Chromatogr. B 910 (2012) 138–148, https://doi.org/10.1016/j.jchromb.2012.03.038.

[23] S. Navea, A. de Juan, R. Tauler, Modeling temperature-dependent protein structural transitions by combined near-IR and mid-IR spectroscopies and multivariate curve resolution, Anal. Chem. 75 (2003) 5592–5601, https://doi.org/10.1021/ac0343883.

[24] C. Ruckebusch, L. Duponchel, B. Sombret, J.P. Huvenne, J. Saurina, Time-resolved step-scan FT-IR spectroscopy: focus on multivariate curve resolution, J. Chem. Inf. Comput. Sci. 43 (2003) 1966–1973, https://doi.org/10.1021/ci034094i.

[25] B. Czarnik-Matusewicz, S. Pilorz, J.P. Hawranek, Temperature-dependent water structural transitions examined by near-IR and mid-IR spectra analyzed by multivariate curve resolution and two-dimensional correlation spectroscopy, Anal. Chim. Acta 544 (2005) 15–25, https://doi.org/10.1016/j.aca.2005.04.040.

[26] T.R.M. De Beer, P. Vercruysse, A. Burggraeve, T. Quinten, J. Ouyang, X. Zhang, C. Vervaet, J.P. Remon, W.R.G. Baeyens, In-line and real-time process monitoring of a freeze drying process using Raman and NIR spectroscopy as complementary process analytical technology (PAT) tools, J. Pharmaceut. Sci. 98 (2009) 3430–3446, https://doi.org/10.1002/jps.21633.

[27] M. Bosco, M. Callao, M. Larrechi, Resolution of phenol, and its di-hydroxyderivative mixtures by excitation–emission fluorescence using MCR-ALSApplication to the quantitative monitoring of phenol photodegradation, Talanta 72 (2007) 800–807, https://doi.org/10.1016/j.talanta.2006.12.004.

[28] L. Cao, P. de B. Harrington, J. Liu, SIMPLISMA and ALS applied to two-way nonlinear wavelet compressed ion mobility spectra of chemical warfare agent

simulants, Anal. Chem. 77 (2005) 800–807, https://doi.org/10.1021/ac0486286.

[29] L. Duponchel, W. Elmi-Rayaleh, C. Ruckebusch, J.P. Huvenne, Multivariate curve resolution methods in imaging spectroscopy: influence of extraction methods and instrumental perturbations, J. Chem. Inf. Comput. Sci. 43 (2003) 2057–2067, https://doi.org/10.1021/ci034097v.

[30] S. Piqueras, L. Duponchel, R. Tauler, A. de Juan, Monitoring polymorphic transformations by using in situ Raman hyperspectral imaging and image multiset analysis, Anal. Chim. Acta 819 (2014) 15–25, https://doi.org/10.1016/j.aca.2014.02.027.

[31] S. Piqueras, L. Duponchel, R. Tauler, A. de Juan, Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares, Anal. Chim. Acta 705 (2011) 182–192, https://doi.org/10.1016/j.aca.2011.05.020.

[32] D. Zhang, P. Wang, M.N. Slipchenko, D. Ben-Amotz, A.M. Weiner, J.-X. Cheng, Quantitative vibrational imaging by hyperspectral stimulated Raman scattering microscopy and multivariate curve resolution analysis, Anal. Chem. 85 (2013) 98–106, https://doi.org/10.1021/ac3019119.

[33] P.J. Gemperline, E. Cash, Advantages of soft versus hard constraints in self-modeling curve resolution problems. Alternating least squares with penalty functions, Anal. Chem. 75 (2003) 4236–4243, https://doi.org/10.1021/ac034301d.

[34] M.H. Van Benthem, M.R. Keenan, D.M. Haaland, Application of equality constraints on variables during alternating least squares procedures, J. Chemometr. 16 (2002) 613–622, https://doi.org/10.1002/cem.761.

[35] S. Hugelier, S. Piqueras, C. Bedia, A. de Juan, C. Ruckebusch, Application of a sparseness constraint in multivariate curve resolution – alternating least squares, Anal. Chim. Acta 1000 (2018) 100–108, https://doi.org/10.1016/j.aca.2017.08.021.

[36] A. Golshan, H. Abdollahi, S. Beyramysoltan, M. Maeder, K. Neymeyr, R. Rajkó, M. Sawall, R. Tauler, A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data, Anal. Chim. Acta 911 (2016) 1–13, https://doi.org/10.1016/j.aca.2016.01.011.

[37] 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.

[38] T. Azzouz, R. Tauler, Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples, Talanta 74 (2008) 1201–1210, https://doi.org/10.1016/j.talanta.2007.08.024.

[39] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, (n.d.) 16.

[40] R. Tauler, A. Izquierdo-Ridorsa, R. Gargallo, E. Casassas, Application of a new multivariate curve resolution procedure to the simultaneous analysis of several spectroscopic titrations of the copper(II)-polyinosinic acid system, Chemometr. Intell. Lab. Syst. (1995) 12.

[41] Willem Windig, Jean Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432, https://doi.org/10.1021/ac00014a016.

[42] W.H. Lawton, E.A. Sylvestre, Self modeling curve resolution, Technometrics 13 (1971) 617–633, https://doi.org/10.1080/00401706.1971.10488823.

[43] R.B. Cattell, The scree test for the number of factors, Multivariate Behav. Res. 1 (1966) 245–276, https://doi.org/10.1207/s15327906mbr0102_10.

[44] W. Windig, J.L. Lippert, M.J. Robbins, K.R. Kresinske, J.P. Twist, A.P. Snyder, Interactive self-modeling multivariate analysis, Chemometr. Intell. Lab. Syst. 9 (1990) 7–30, https://doi.org/10.1016/0169-7439(90)80050-G.

[45] A.P. Snyder, W. Windig, J.P. Toth, Interactive self-modeling multivariate analysis of thermolysis mass spectra, Chemometr. Intell. Lab. Syst. 11 (1991) 149–160, https://doi.org/10.1016/0169-7439(91)80062-U.

[46] W. Windig, C.E. Heckler, F.A. Agblevor, R.J. Evans, Self-modeling mixture analysis of categorized pyrolysis mass spectral data with the SIMPLISMA approach, Chemometr. Intell. Lab. Syst. 14 (1992) 195–207, https://doi.org/10.1016/0169-7439(92)80104-C.

[47] S. Gourvénec, D.L. Massart, D.N. Rutledge, Determination of the number of components during mixture analysis using the durbin–watson criterion in the orthogonal projection approach and in the SIMPLe-to-use interactive self-modelling mixture analysis approach, Chemometr. Intell. Lab. Syst. 61 (2002) 51–61, https://doi.org/10.1016/S0169-7439(01)00172-1.

[48] J.J. Andrew, M.A. Browne, I.E. Clark, T.M. Hancewicz, A.J. Millichope, Raman imaging of emulsion systems, Appl. Spectrosc. 52 (1998) 790–796, https://doi.org/10.1366/0003702981944472.

[49] M. Ghaffari, A.-L. Chateigner-Boutin, F. Guillon, M.-F. Devaux, H. Abdollahi, L. Duponchel, Multi-excitation hyperspectral autofluorescence imaging for the exploration of biological samples, Anal. Chim. Acta 1062 (2019) 47–59, https://doi.org/10.1016/j.aca.2019.03.003.

[50] J. Ofner, F. Brenner, K. Wieland, E. Eitenberger, J. Kirschner, C. Eisenmenger-Sittner, S. Török, B. Döme, T. Konegger, A. Kasper-Giebl, H. Hutter, G. Friedbacher, B. Lendl, H. Lohninger, Image-based chemical structure determination, Sci. Rep. 7 (2017) 6832, https://doi.org/10.1038/s41598-017-07041-x.

2.2.2. Additional work and use of randomised SIMPLISMA

Another important aspect to be deepened here is the use of this kind of approach on even bigger datasets, containing several millions of spectra. Concerning this topic, some ideas were proposed during this PhD. In detail, a big dataset (a section of a wheat grain) has been acquired in the French national synchrotron facility, namely the SOLEIL, thanks to a collaboration with Dr. Frédéric Jamme, beamline scientist in SOLEIL, and working also with the INRAE group of research in Nantes (France) headed by Dr. Marie-Françoise Devaux. Regarding the acquisition information, a microscope (zoom 40x) has been used to generate the image of the dataset in which the phenomenon of autofluorescence coming from the excitation using UV and visible spectral ranges (excitation wavelength of 275 nm) has been observed. The size of this data cube, whose global integration image and corresponding spectral profiles are reported in Fig. 16, is 1024 pixels by 1024 pixels, with a resolution of 0.3 µm per pixel, for a total of 1048575 emission spectra for 9 spectral variables. More precisely, these spectral variables correspond to different filters, each of them corresponding to a specific domain of wavelengths, in order to distinguish the emission coming from different biological molecules of the sample.

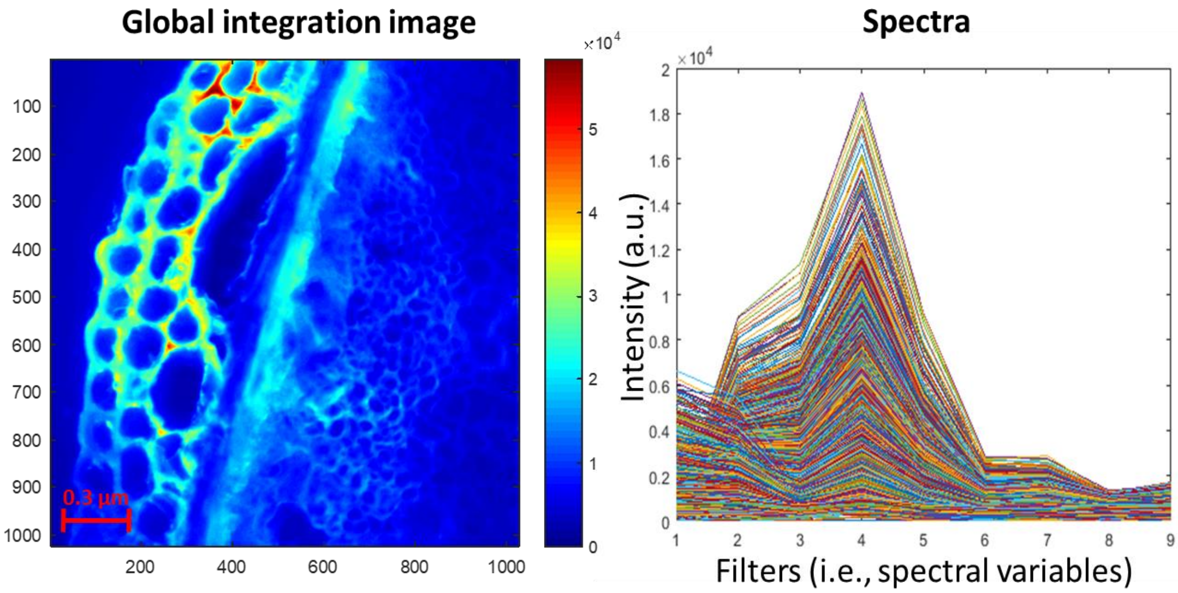


Fig. 16 – Global integration image and spectra of the wheat dataset.

In detail, here are reported the emission wavelength ranges for the different variables, and a brief description of which kind of molecules they correspond to. The first filter covers the range between 327 and 353 nm, while the second one the interval between 370 and 410 nm. They are potentially used to probe proteins and small molecules. The third, fourth and fifth filters are respectively related to the following spectral ranges: 412-438 nm, 420-480 nm, and 435-455 nm. This interval leads to the observation of small molecules and particularly to hydroxycinnamic acids (e.g., ferulic acid, para-coumaric acid, etc.) that are a class of aromatic acids (or phenylpropanoids) with a C6-C3 skeleton, derivatives of cinnamic acid. Finally, the last four filters cover respectively the intervals between 484 and 504 nm, 499 and 529 nm, 530 and 570 nm, 535 and 607 nm. This range is particularly important to observe lignin compounds, a class of complex organic polymers that form key structural materials in the support tissues of vascular plants. Evidently, the size of this image is enormous, reason why it can be very complicated to find the right rank, considering for example the fact that some information can be extremely pure, but represented in very specific and small areas of the data cube. In fact, observing Fig. 17, it is obvious that it is impossible to obtain any clear information from PCA, considering all the pixels:

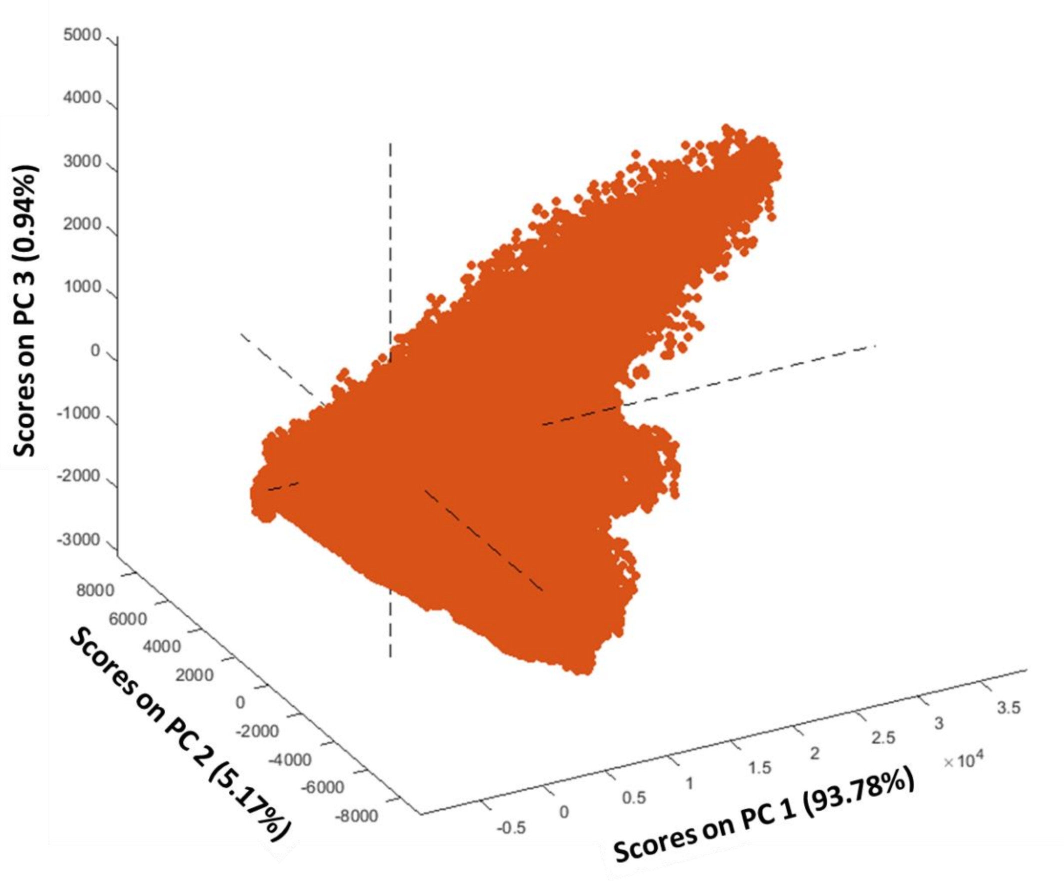


Fig. 17 – PCA on the totality of the pixels of the wheat dataset.

Evidently, randomised SIMPLISMA can be a good alternative, in order to avoid the loss of details and so, obtain better outcomes. Nevertheless, due to the fact that the matrix is represented by more of one million of spectra, it can be challenging to select the right inputs to use (e.g., the percentage of random pixels per subset). In fact, imprecise values would lead to difficulties in visualizing the separated clusters, i.e., situations such as the overlapping of different classes or the possibility of losing some very small details. For this reason, as a first step, the whole data cube has been divided into four sections, each of them now equal to a sub-image of 512 pixels by 512 pixels, corresponding to 262144 spectra. Then, randomised SIMPLISMA has been separately applied on each of the new reduced images, and the right rank (that can obviously vary through the different sub-images, depending on the information carried by each of them) selected. Finally, the chosen clusters from the single sub-images have been merged together in the same dataset. Then, randomised SIMPLISMA has been used a second time, in order to reduce again the total amount of spectra. At the end, the selected purest spectra have been observed to select the global rank of the initial data matrix. A graphical representation of the final clustering, in which were used only the selected spectra from each of the sub-images, is shown below in Fig. 18:

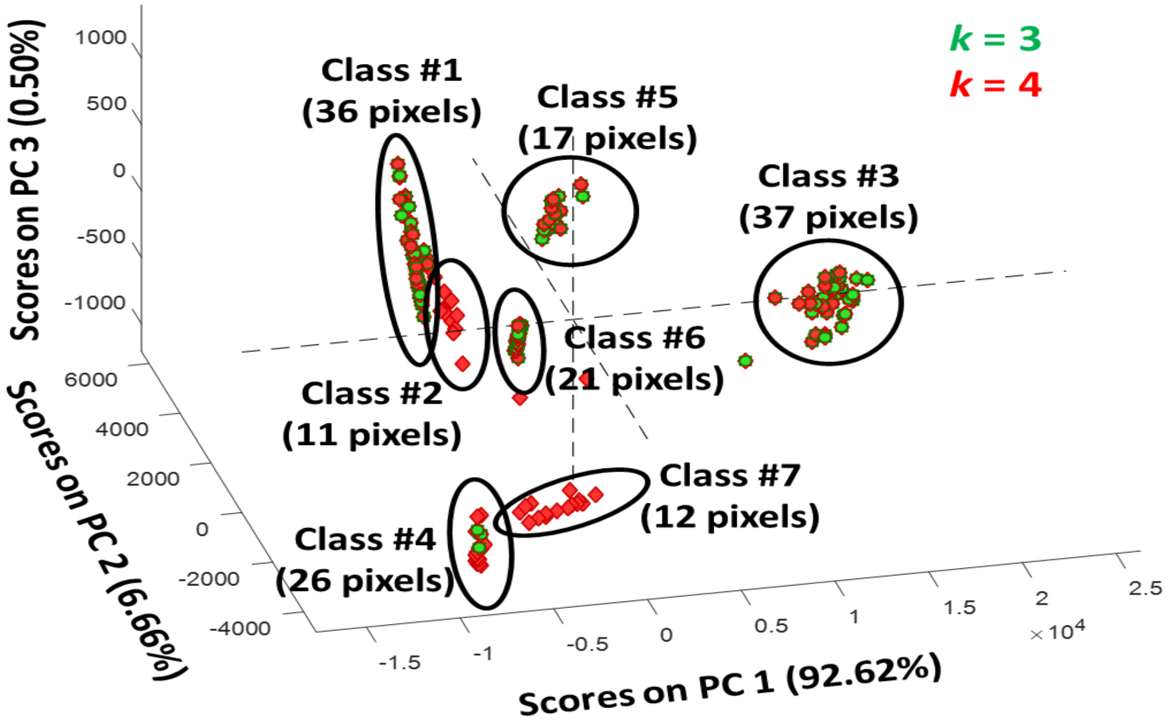


Fig. 18 – PCA exploration of pixels selected by the ‘double’ randomised SIMPLISMA approach as a function of *k*.

As observable, due to this ‘double’ randomised SIMPLISMA approach, it is possible to select the purest spectra from the whole dataset, and particularly, to use only a very small quantity of the initial pixels for the investigation in the PCA space. In fact, in this way it is easier to observe very clear clusters, each of them related to particular spectral information. As a final step, using the same approach previously explained, the mean spectrum from each cluster has been used to generate the initial estimates for the MCR-ALS calculation, whose first outcomes (both pure concentration and corresponding spectral profile matrices) are here reported in Fig. 19:

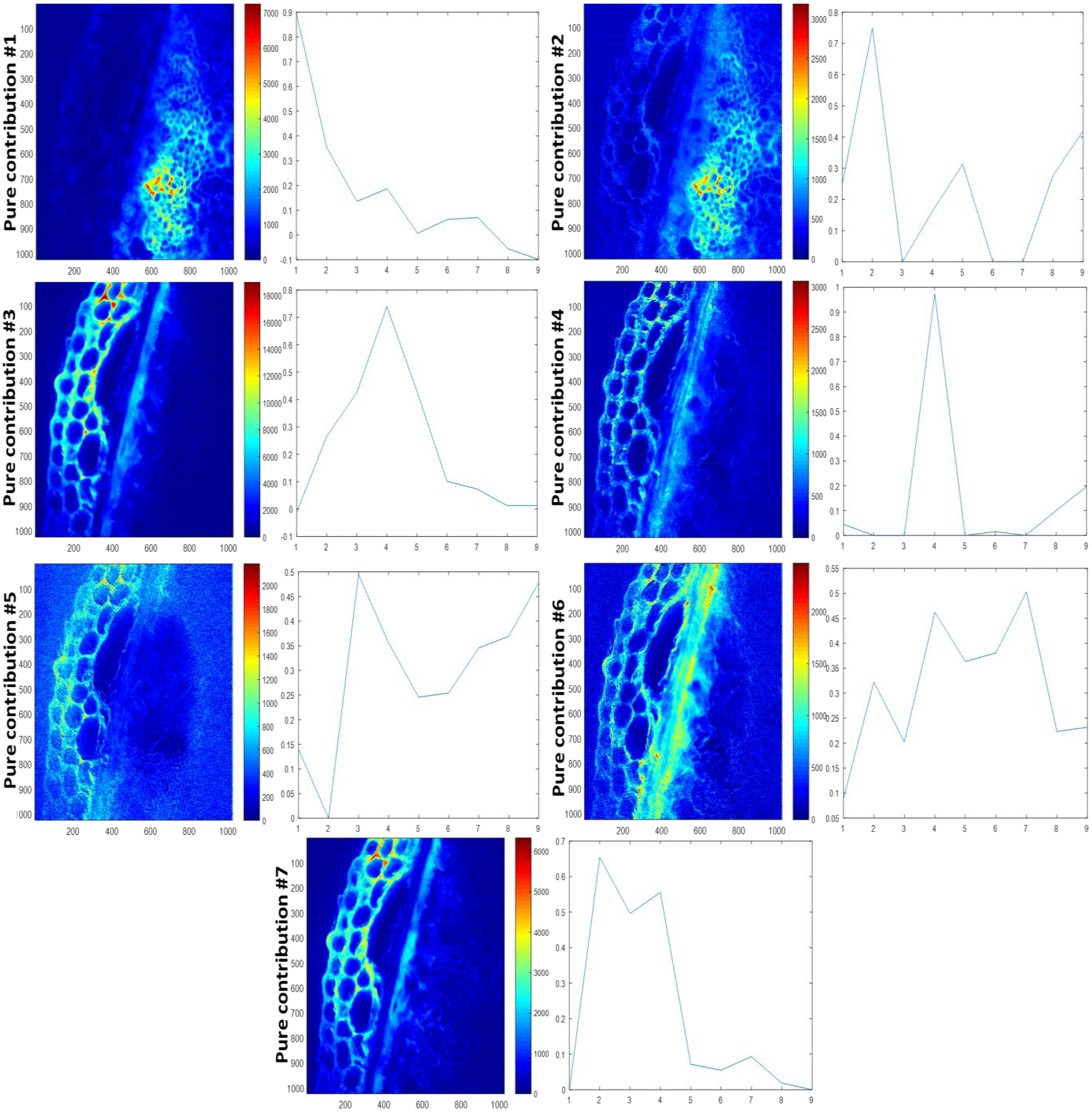


Fig. 19 – Final results using the ‘double’ randomised SIMPLISMA approach. Here are reported the seven extracted concentration maps corresponding to the MCR-ALS calculation and their corresponding spectral profiles.

Naturally, further investigations are required in order to confirm these results, understanding their biological nature, and clearly, refine the use of this new approach based on randomised

SIMPLISMA. Nevertheless, first outcomes seem to be very promising, enough to lay claim new possibilities in order to obtain always better results in the hyperspectral image analysis and particularly, the spectral unmixing framework.

2.2.3. Conclusions and future perspectives

Modern instruments show the possibility of acquiring a big quantity of data, normally related to very high resolution. MCR-ALS is a chemometric tool that has the potential of being used in many different areas, leading to the decomposition of complex matrices into the contribution of the pure components, in order to dig the chemical composition of samples of various natures. Nevertheless, some limitations are still faced. In an interesting domain such as the hyperspectral image analysis, choosing the right rank can be challenging, selecting all the information contained in a data cube, particularly when contributions related to a small quantity of pixels is available. Finding a way that can automatize the selection of the real number of components is a very important task. Randomised SIMPLISMA is an algorithm based on SIMPLISMA (a useful tool used in the MCR-ALS framework) that has shown the capability to help the operator in this important purpose. Anyway, some limitations using this approach are still present. For example, it is necessary to highlight the fact that randomised SIMPLISMA requires some inputs to work correctly. First of all, the percentage of data to select for each subset. Then, the number of subsets to be generated. If the wrong number is chosen, it could lead to incorrect results. In fact, too many selected points could create an issue in the possibility of observing the right number of groups in the PCA space, due to an overlap of the classes. Contrariwise, too few spectra could lead to the loss of some specific pixels, very pure, but present as a small quantity compared with the rest of the information. This is the reason why, during the experimental part of this work, many attempts were carried out in order to find the right values to be applied in order to obtain reliable results. In addition, as previously shown, an interesting alternative would be the one of at first divide the whole image into a certain quantity of sub-images (the number of them depending on the original size and complexity of the data cube), in order to use randomised SIMPLISMA in a first step separately on each of them. In this way, the possibility of observing more interesting details would be easier. Clearly, more studies to optimize this procedure are required. Another important aspect of randomised SIMPLISMA is that the observation of the different clusters is carried out by the use of the PCA space, and so, it is related to a subjective interpretation. The main problem here is that, despite the existence of various methods that can automatically count the number of clusters, they depend on the nature of the investigated sample.

This means that based on the data matrix, different information has to be given, to avoid clustering errors. Therefore, it would be interesting in future to find a way to make this selection automatic, for example by the use of KM clustering, or other clustering approaches. In addition, as previously described, at first the spectra of the different clusters are collected, and then the mean spectrum for each class is calculated, in order to use them as initial estimates. However, it is important to consider that the variability presents in a dictionary of a given contribution is rather little exploited for this reason. An important improvement would be to have the capability of considering the spectra of a dictionary as different pure representations of a given compound, such as it can be done nowadays in the remote sensing community.

2.3. Detection of minor compounds in complex mineral samples from millions of spectra: A new data analysis strategy in LIBS imaging

2.3.1. Introduction

Among the different spectroscopies that are currently used in different research domains, one in particular is obtaining a constant arising interest in many communities, i.e., LIBS imaging. The development of this spectroscopic technique has led to the necessity of finding new ways to extract the information, and chemometrics is for sure a very interesting area that can be involved in this task, due to its characteristics and strong points. One of the reasons why LIBS is nowadays a very promising technique for the analytical investigation of complex matrices is that thousands to millions of spectra can be obtained in very short times, thanks to the high acquisition rate (up to 1000 spectra/s) of this instrumentation. This means that it is easily possible to acquire images made of millions of pixels, also helped by the use of powerful microscopes. Naturally, this characteristic involves the fact that the generated hyperspectral images show a high spatial resolution that, linked to the multi-elemental capabilities of LIBS, arises the interest in using this instrument in different research areas [85,210,211]. Nevertheless, one of the most important challenges currently faced in handling this kind of data cubes is that still, it is not easy to find a strategy able to deal with this huge quantity of data in a suitable way. In fact, many issues can be experienced. First of all, the amount of data is big, reason why it can result complicated, if not even impossible, to work with the whole dataset. Therefore, a strategy able to collect the most important information and reduce the quantity of spectra to be observed is a fundamental task. In addition, it is reasonable to think that, no matter the quality of this spectroscopy, the acquired data can be affected by some problems. For example, it is a common scenario the possibility of generating some saturated signals that naturally can lead to a poor interpretation of the spectra. Lastly, as already explained in many parts of this manuscript, it is reasonable to think that while some elements are present as main components in a sample, others (probably the majority) are available as minor compounds or even traces. Therefore, if the right approach is not selected, there is a real risk of losing the information related to these details, leading to only partial and insufficient outcomes. Chemometrics can be a good alternative to the routine analyses, capable of overcoming these problems, and showing new interesting approaches for LIBS image analysis. Many of these aspects have been faced during this PhD, and so discussed in the manuscript. In detail, while the other aspects will be shown in Chapter 3 and Chapter 4 of the present work, in this section it will be described a method used to investigate in an easy way a big dataset made of more than two million spectra, finding interesting results coming from not only the main components, but also the minor ones and the traces of the sample of interest. Specifically, the investigated sample is a complex mineral containing various elements, such as W, Au, Pb, Zn, Ag, and others. This work, published in Analytica Chimica Acta, Volume 1114 (2020) [212], shows the use of KM clustering [146,208] in a particular way, the Embedded K-Means (EKM), hereafter explained. KM clustering is a well-known unsupervised classification method vastly used in the chemometric society for the investigation of many different types of samples and spectroscopies. The limitation of this algorithm, such as for many other chemometric tools, is that its outcomes can be influenced by the total explained variance of the spectra in the matrix. This means that in a very big dataset where some small components are easily available, this kind of approach would be biased, leading to the clustering of only the major compounds, while the minor elements and traces would be easily missed. EKM is an interesting alternative that overcomes these limitations, allowing the classification of also minor compounds and traces, as it is following described.

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Detection of minor compounds in complex mineral samples from millions of spectra: A new data analysis strategy in LIBS imaging

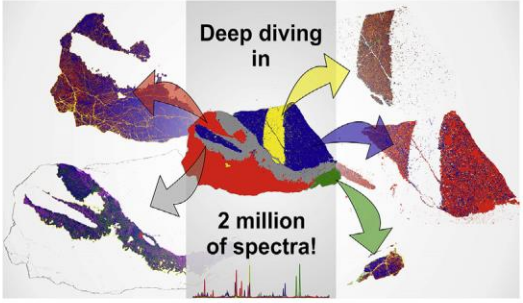
Alessandro Nardecchia [a], Cécile Fabre [b], Jean Cauzid [b], Frédéric Pelascini [c], Vincent Motto-Ros [d], [**], Ludovic Duponchel [a], [*]

[a] Univ. Lille, CNRS, UMR 8516 – LASIRe – LAboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, F-59000, Lille, France
[b] Université de Lorraine, Laboratoire GeoRessources, UMR CNRS, 7359, France
[c] Cetim Grand Est, Illkirch-Graffenstaden, France
[d] Institut Lumière Matière, UMR 5306, Université Lyon 1 - CNRS, Université de Lyon, 69622, Villeurbanne, France

HIGHLIGHTS

- A new strategy for detecting and localizing simultaneously major and minor compounds from millions of LIBS spectra.
- Introduction of the embedded k-means clustering approach.
- The first use of the PBM (Pakhira–Bandyopadhyay–Maulik) index in LIBS imaging.
- The first clustering technique applied to millions of LIBS spectra.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 5 December 2019
Received in revised form
2 April 2020
Accepted 3 April 2020
Available online 6 April 2020

Keywords:
Laser-induced breakdown spectroscopy (LIBS)
Big data
Hyperspectral imaging
Clustering

ABSTRACT

Today, Laser-Induced Breakdown Spectroscopy (LIBS) imaging is in full change. Indeed, always more stable instrumentations are developed, which significantly increases the signal quality and naturally the analytical potential of the technique for the characterization of complex and heterogeneous samples at the micro-scale level. Obviously, other intrinsic features such as a limit of detection in the order of ppm, a high field of view and high acquisition rate make it one of the most complete chemical imaging techniques to date. It is thus possible in these conditions to acquire several million spectra from one single sample in just hours. Managing big data in LIBS imaging is the challenge ahead. In this paper, we put forward a new spectral analysis strategy, called embedded k-means clustering, for simultaneous detection of major and minor compounds and the generation of associated localization maps. A complex rock section with different phases and traces will be explored to demonstrate the value of this approach.

1. Introduction

Laser-induced breakdown spectroscopy (LIBS) imaging is actually becoming an essential tool to characterize complex samples in many scientific domains [1–5]. In this spectroscopic technique, a pulse laser beam focused on the sample surface generates a plasma from a small amount of vaporized material. Due to the electronic

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: vincent.motto-ros@univ-lyon1.fr (V. Motto-Ros), ludovic.duponchel@univ-lille.fr (L. Duponchel).

https://doi.org/10.1016/j.aca.2020.04.005
0003-2670/© 2020 Elsevier B.V. All rights reserved.

relaxation of excited atoms and ions, an emission spectrum characteristic of the elemental composition of the sample can be acquired using an optical spectrometer. In LIBS imaging experiments, the sample surface is explored in a raster scanning mode (i.e. acquisition of one spectrum for each spatial position of a predefined grid) covering the region of interest. An elemental image can then be generated from the acquired data set using a simple signal integration of a given emission line. The richness of this imaging approach lies in its many advantages that cannot be observed simultaneously in any other spectroscopic technique. Indeed, LIBS imaging has multi-elemental capabilities, a high acquisition rate ($\geq$100 spectra/s), full compatibility with optical microscopy and ease of use on samples without almost any size restriction (up to several tens of cm^2), all under atmospheric conditions. On top of that, this technique has a high field of view and a spatial resolution around 10 μm coupled with a limit of detection in the order of weight ppm. It is thus very convenient to explore a sample at the micronic scale by acquiring several million spectra in just hours.

Concerning data analysis in LIBS, we see today big differences between the two frameworks of bulk analysis and imaging. Indeed, researchers have quickly learned that multivariate data analysis could bring valuable tools for qualitative and quantitative explorations of samples at the bulk level, for instance by developing regression or classification models [6–10]. At the imaging level, there is a relatively limited number of papers dealing with the use of multivariate data analysis in the LIBS community. Indeed, elemental images are, in general, generated from single emission wavelengths, even though the whole spectral domain could be used. The application of chemometric approaches to imaging data sets is in fact more complex, both from a conceptual and practical point of view. Although a large part of the LIBS community is increasingly sensitive to the use of chemometric tools, understanding the concept of hyperspectral imaging, finding appropriate tools for data exploration, and finally interpreting their outputs represent a big task for non-expert researchers. In addition, it is clear that managing millions of spectra increases the difficulty of this task even if they know the great potential of chemometrics. This is not just about the availability of computational resources, but also, the development of new data exploration tools able to manage such big data structures.

In this paper, the idea is obviously not to systematically apply a well-known unsupervised classification method to a LIBS imaging data set. Indeed, it would be totally inefficient in detecting minor compounds because most chemometric algorithms exploit explained variances. As a consequence, we will introduce a new data processing strategy, that we call embedded k-means clustering, in order to detect and localize simultaneously major and minor compounds in a complex mineral sample from a data set of more than 2 million spectra.

2. Experimental section

2.1. Sample description and preparation

In order to demonstrate the potential of our strategy of spectroscopic exploration, we have selected a complex mineral sample from the polymetallic W–Au–Pb–Zn–Ag (Sb–Ba) district of Tighza (Central Morocco). More specifically, it is related to the Sidi Ahmed hydrothermal event [11]. This district has been mined for centuries for Pb and Ag, Pb–Zn–Ag mineralization being formed of sulfides in gangues of carbonates. Naturally, we can expect the simultaneous presence of major and minor compounds but also traces in such mineralization [3]. The size of the selected rock section is approximately 3.2 cm × 1.6 cm and 1 cm thick. Prior to LIBS analysis, the surface of the sample has been finely polished using polisher as it is usually done in other techniques such as Scanning Electron Microscopy (SEM) and Energy Dispersive X-Ray Spectroscopy (EDS).

2.2. Experimental setup and spectral data acquisition

The LIBS instrumental setup used in this work is based on a homemade optical microscope and a Nd:YAG laser (Centurion GRM, Quantel by Lumibird) with an 8 ns pulse duration operating at 100 Hz. The laser beam is focused on the sample surface using a 15× magnification objective (LMM-15X–P01, Thorlabs). The rock section is placed on a three axes XYZ motorized stage in order to move precisely the sample during the mapping experiment. Atomic force microscopy (AFM) has been used in order to check that the crater size after ablation was smaller than the distance between two consecutives acquisition positions on the sample which is 15 μm. An autofocus system is also used during the analysis in order to keep the objective-to-sample distance from changing. Thus, we always have the same distance between the objective and the plasma emission regardless of the sample flatness. Every spectra in the data set have been acquired from single laser pulses at each spatial position of the sample. The plasma emission has been collected by a quartz lens and focused onto the entrance of a round-to-linear fiber bundle (19 fibers with a 200-μm core diameter) connected to a Czerny-Turner spectrometer (Shamrock 500, Andor Technology). This spectrometer is equipped with a 600 l/mm grating blazed at 300 nm and an intensified charge-coupled device (ICCD) camera (iStar, Andor Technology). The camera is synchronized with the Q-switch of the laser, and spectra are acquired with a delay of 500 ns and a gate of 3000 μs, in full vertical binning mode. Moreover, a servo control loop based on a power meter and a computer-controlled attenuator (ATT1064, Quantum Composers) is used to control the laser power. A homemade software, developed under LabVIEW® environment, has allowed the automation of scanning sequence as well as the spectral acquisition. All measurements have been performed at room temperature under ambient pressure conditions.

The hyperspectral LIBS data set has been acquired considering a 15 μm spatial resolution and a 0.15 nm spectral resolution. The 251.38–339.99 nm spectral domain (2048 spectral channels) has been selected to cover the main emission lines of all elements of interest. In these conditions, we have obtained a data cube of size 2100 pixels × 1090 pixels x 2048 wavelengths (i.e. 2.289.000 acquired spectra for a 515 mm^2 field of view). The total acquisition time was approximately 6 h, which is finally not so long regarding the richness of the chemical information. It is then easy to understand that a specific data analysis strategy must be implemented if we really want to extract information about major and minor compounds from such a big data set.

2.3. Multivariate data exploration

In this work, the main idea is to propose a method able to explore megapixel LIBS data set without prior knowledge about the sample composition and to highlight simultaneously the presence of major, minor compounds, and even traces. In the multivariate data analysis framework, this task corresponds to the development of an unsupervised classification model. In other words, such techniques try to find natural groupings of spectra in the considered data set, which will represent different chemical compounds. Even if the chemometric community has developed different tools for unsupervised classification of spectra, we can say without hesitation that the well-known k-means [12] clustering (KM) is certainly the most popular one. Indeed, behind the apparent simplicity of this method, it has been proved effective for many different kinds of data sets and spectroscopies. To the best of our knowledge, as this algorithm has never been used in the framework of LIBS imaging, a short description of the algorithm is provided below. Like any other chemometric method, a spectrum is considered as a point (denoted $\boldsymbol{x}_i$) in a multidimensional space. Let X = {$\mathbf{x}_i$, $i = 1, \ldots, n$} be a dataset composed of n points (i.e. spectra) with $\mathbf{x}_i \in \mathbb{R}^w$, w being the number of spectral variables in a spectrum. For illustrative purposes, let's consider a small LIBS imaging data set. This data cube of size 5 pixels × 5 pixels x 2 wavelengths consists of 25 pseudo-spectra. Fig. 1a illustrates the successive steps of the k-means algorithm applied to this toy example. In a first step, k initial points called centroids (in this example $k = 3$) are randomly generated within the data domain (shown in color in Fig. 1a). In the second step, one calculate distances between all points of the data set and the generated centroids. In fact, the distance is used as a measure of similarity between spectra. In this work, the cosine distance has been preferred to the Euclidean one, the latter being sensitive to global intensity changes in spectra. However if the Euclidean distance had been selected, then it would have been necessary to use a signal normalization commonly used in the LIBS community. The cosine distance $d_{i,j}$ between spectra $\mathbf{x}_i$ and $\mathbf{x}_j$ is given in equation (1) considering a point as a vector in a multidimensional space:

$$d_{i,j} = 1 - \frac{\overrightarrow{x}_i . \overrightarrow{x}_j}{\|\overrightarrow{x}_i\| . \|\overrightarrow{x}_j\|} \tag{1}$$

As we can see, this distance corrects for global intensity variations by dividing each spectrum i and j by its norm. Given all the distances, each point (i.e. spectrum) is associated with the nearest centroid and now belongs to one of the k classes. In a third step, the mean spectrum of each class is calculated and will represent the k new centroids. In the fourth step, spectra in the data set are again unassigned. Then steps 1–3 are repeated in a loop in order to refine the position of the k centroids. Calculations are stopped when convergence is observed, i.e. when no further changes are observed in the spectra class memberships. In the last step, the knowledge of the class membership of each spectrum and its localization in the pixel space allow us to generate a clustering map using a color-coding. At the same time, the centroid corresponding to each class is a spectrum used for chemical interpretation.

Behind the simplicity and ease of use of KM, there is an important issue which we have to address, namely, how to select the optimal number of clusters or classes. Unfortunately far too often in the literature, authors select with *a priori* this value of k, which is definitely the ultimate negative choice. Indeed, no one can know the whole chemical complexity of the considered sample. In general, the most reasonable way is to use a criteria called index in order to automatically choose this value. This index is a mathematical function that measures the quality of a partition. The idea is then to perform a KM clustering for different values of k ($2 \leq k \leq k_{max}$) and to calculate this index for each partition. The highest index value indicates the optimal number of clusters for the considered data set. One of the best index in the literature is PBM (Pakhira–Bandyopadhyay–Maulik) [13]. It is defined as the square ratio between the largest normalized inter-cluster distance ER and the normalized sum of intra-cluster distances RA:

$$PBM(k) = \left(\frac{ER}{RA}\right)^2 \tag{2}$$

with $ER = \frac{max_{l,m-1,\ldots,k}\|\boldsymbol{c}_l \quad \boldsymbol{c}_m\|}{k}$, $RA = \frac{\sum_{i-1}^{k}\sum_{j-1}^{n_i}\|\boldsymbol{x}_{i(j)} \quad \mathbf{c}_i\|}{\sum_{i-1}^{n}\|\boldsymbol{x}_i \quad \overline{\boldsymbol{x}}\|}$, $\mathbf{c}_i$ the centroid of the ith cluster ($i = 1 \ldots k$), $\boldsymbol{x}_{i(j)}$ the jth spectrum of the cluster i, n_i the total number of spectra in the cluster i, and $\overline{\boldsymbol{x}}$ the mean point of the considered dataset. The PBM index will be used in this work in order to select the optimal number of clusters.

We could obviously explore directly the proposed data set with KM in these conditions, but we should not lose sight of our main goal, which is the simultaneous detection of major and minor compounds. Indeed, this inquiry about the intrinsic data structure is very important because KM algorithm (and most of the clustering methods) can fall into a trap under two specific conditions (Fig. 1b). The first problematic situation is observed when classes in the data set are unbalanced, that is to say when a big difference in the number of spectra between classes is observed. This is precisely the case for major and minor chemical compounds present in an imaging data set. As a consequence, small populations of spectra would not be detected and wrongly associated with the nearest big clusters. The second problematic situation arises when subpopulations of spectra are observed in a given cluster. In this case, only a global cluster is generated and small spectroscopic details are lost during this exploration. To address these issues, we have developed a new strategy, which we call embedded k-means clustering (EKM). We were inspired by the way our brain works when we are looking at a picture. We first extract the main features of the image (i.e. the main classes of objects) and, then, we extract details about sub-zones of it. Thus, in the EKM strategy, a first k-means clustering will be applied to the whole data set and the second round of clusterings will be applied to each previously calculated cluster (Fig. 1c). Obviously, the PBM index will be used at each step of the way.

All calculations in this work have been performed under the Matlab 2016b environment (The Mathworks, Inc., Natick, Massachusetts) using homemade codes.

3. Results and discussion

To better understand the strengths of our data analysis strategy, it is essential to open this section with the exploration of the considered imaging data set using the state-of-the-art method to generate chemical maps [4,14]. First, a single emission line is selected for an element of interest. Then a baseline correction is applied on every single spectrum of the data set in order to extract corresponding net intensities at the given wavelength. Lastly, color-coding is used in order to generate a colored elemental map from these extracted values, the intensity of the chosen color being correlated with abundance. Of course, this procedure can be successively repeated for all elements of interest in the sample, with the possibility to observe them simultaneously in overlay mode on the same image. Nevertheless, despite this operational simplicity, this traditional method imposes two constraints which should be considered for the generation of unbiased chemical maps. First, each selected emission line should be the strongest one in the spectral domain for each element. But what is more important, a selected emission line should not present potential interferences with other lines. Due to the natural complexity of the samples we usually explore, we quickly see that it is a strong hypothesis, which, for each element of interest, could be difficult to hold in relation to the very high number of lines in a spectrum. Fig. 2 illustrates the use of this conventional approach to the rock section. More specifically, Fig. 2a shows the mean spectrum calculated from all spectra of the imaging data set. From this spectrum, it is always simple and fast to identify major elements by matching the observed emission lines with an atomic spectra database. Thus it is easy to see, without being exhaustive, the presence of different elements such as Pb, Ag, Fe, Ca, Mg, Mn, Cu, and Si. Fig. 2b presents the global intensity image of the sample generated from the integration of the emission signal for each pixel on the whole spectral domain. Of course, we are losing elemental information with this

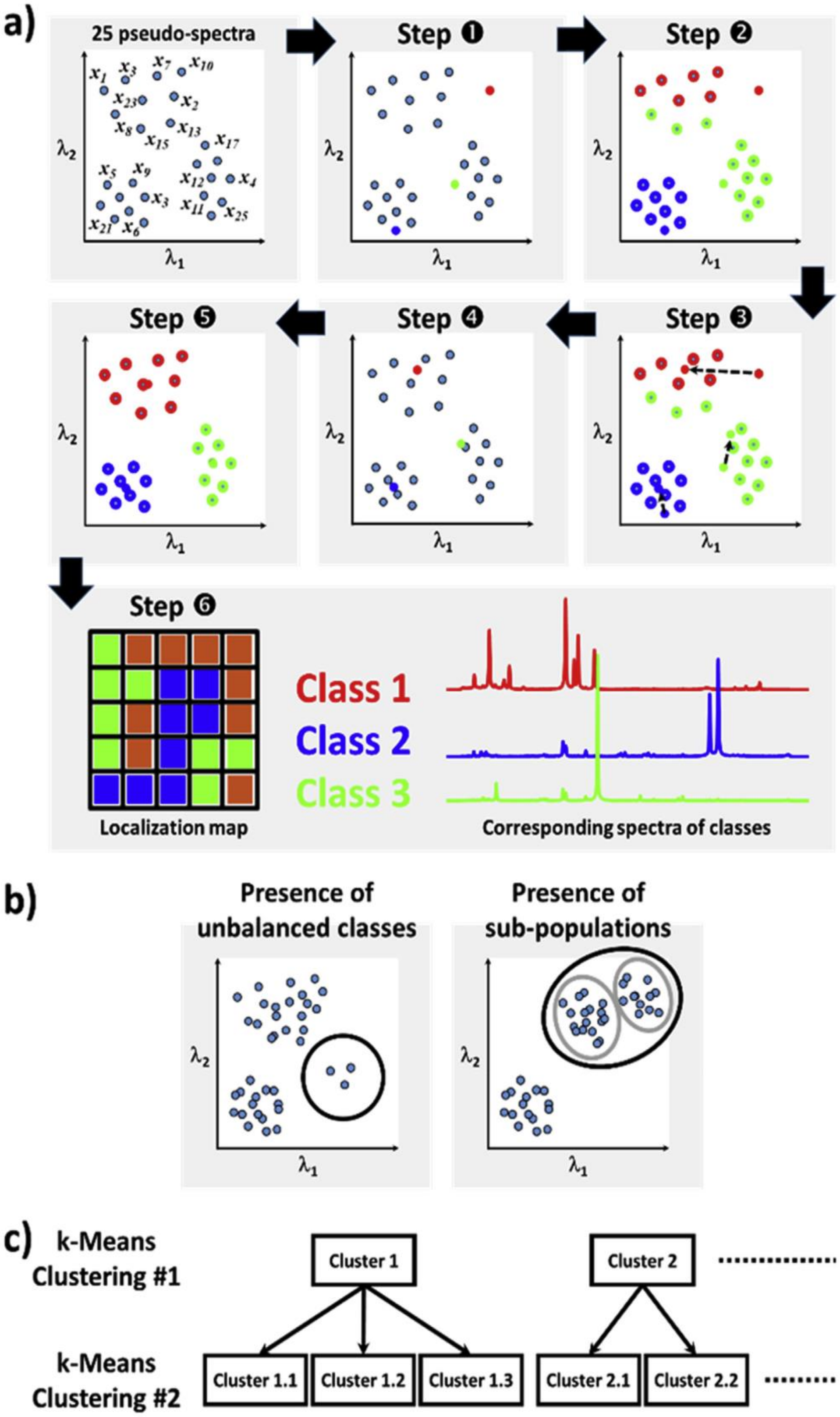


Fig. 1. a) The k-means algorithm applied to spectroscopic imaging. b) Problematic data structures hardly managed by k-means. c) The proposed method called embedded k-means clustering.

observation but different zones of the samples can nevertheless be highlighted in this image. It is even possible to observe different levels of homogeneity, textures, and sub-structures on the sample. By contrast, Fig. 2c and d give elemental images generated with the conventional approach using single integrations described above. At first glance, we notice that many elements are localized in specific areas. Although it is possible to observe the colocalization of element pairs such as Ag/Pb, Si/Al, Si/Ti, and Zn/Cu for example, finding a correlation between all elements in this data set is a hard task. Yet, we have to remember that such correlations should allow a trace-back to molecular information i.e. mineral phases in this particular case. A further point concerns the detection of potential anti-correlation

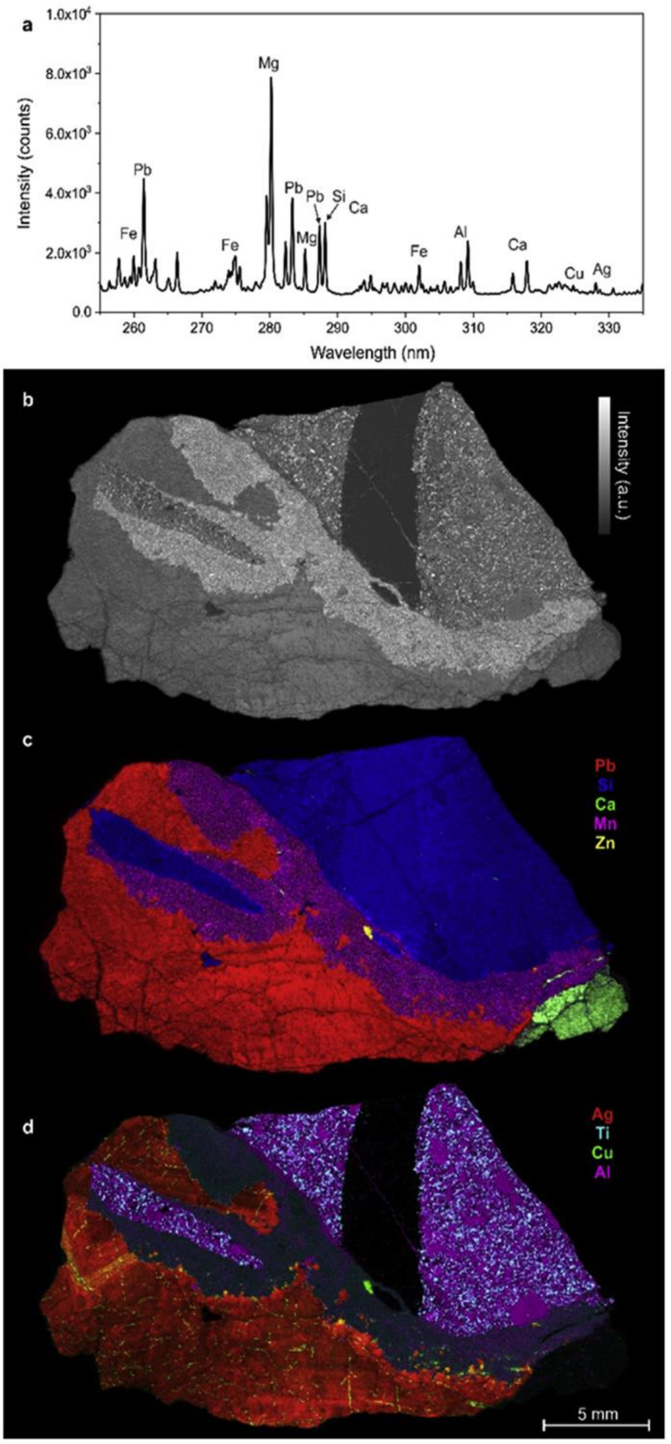


Fig. 2. a) The mean spectrum of the LIBS data set. b) The global intensity image. c) and d) Elemental images generated with the conventional approach. A high-resolution version of this image can be downloaded from supplementary materials.

between elements, which is especially difficult to achieve by just comparing elemental images. It is indeed very interesting to know if a specific element is present in a zone when another one is systematically absent or has a low concentration, and vice versa. In conclusion of this section, while the usual procedure allows us to generate consistent elemental images most of the time, we can clearly see that we are still not harnessing all the information contained in the data set, minor compounds and minor phases not being particularly highlighted.

In this new section, the idea is to apply the strategy of embedded k-means clustering on the considered data set and assess its interest for the simultaneous detection of major and minors compounds. As explained previously, the initial step of this approach consists of the application of a first k-mean clustering on the whole data set (i.e. all spectra). Fig. 3a shows the evolution of the PBM index according to the number of cluster *k* used in this first partitioning of pixels. Here it can be seen clearly that an optimal number of five clusters has to be considered. Using this consideration as a starting point, Fig. 3b provides a classification map from which we can observe the localization of the five compounds. The percentage of pixels in a class for the total number of pixels in the data set is also given. We can see, therefore, that classes 1,4 and 5 correspond to major compounds with 37%, 27% and 23% of pixels respectively. Nevertheless, at this point, we cannot say that classes 2 and 3 correspond to minor compounds with 3% and 10% of pixels respectively. In fact, they are only somewhat less present. As regards the dispersion of compounds in the sample, classes 1 et 2 are strictly observed in well-delimited and continuous areas. It is almost the case for class 5, which is nevertheless also located around the area of class 1. More heterogeneous distributions are observed for classes 3 and 4. Fig. 3c gives the corresponding spectra of the centroids for each class. These representative spectra are naturally used for chemical interpretation. Despite the fact that LIBS spectroscopy is an elemental one, the use of the whole spectral domain and some prior knowledge about the genesis of rocks allow us to identify potential mineral phases. Thus, class 1 is associated with galena (PbS) with traces of copper, silver, antimony, and tin. The mineral phase corresponding to class 2 is calcite ($CaCO_3$) with traces of manganese, magnesium, silicon, and aluminum. Class 3 is linked to quartz (SiO_2) with traces of magnesium, aluminum, calcium, titanium and iron. The next mineral phase with class 4, is potentially an aluminosilicate (SiO_2/Al_2O_3) or kinds of clays with traces of magnesium, calcium, iron, manganese and titanium. Finally, class 5 is associated with ankerite ($Ca(Fe, Mg,Mn)(CO_3)_2$) with traces of titanium.

To go deeper into the exploration of previous mineral phases, we shall apply the second step of the embedded k-means strategy. Therefore, for each class, a new k-means clustering is applied only to associated spectra. In other words, five k-means clustering are calculated in parallel considering the five different sub-populations of spectra contained in the five classes. Obviously, the PBM index is used again to optimize the number of clusters of each k-means clustering. The five graphs representing the evolution of the PBM index according to the number of clusters *k* are supplied in the supplementary material (Fig. S1). We then discover that all mineral phases exhibit sub-populations of spectra. The galena (class 1) contains 3 sub-classes of compounds, the calcite (class 2) has 4, the quartz (class 3) has 5, the aluminosilicate phase (class 4) has 3 and ankerite (class 5) has 6. Fig. 4 gives classification maps for each phase and corresponding spectra of sub-classes. For galena, classes 1.1 and 1.3 (in blue and red respectively) are the two major compounds of the galena phase with 64% and 26% of pixels respectively. These two sub-classes exhibit different ratios of elements such as Cu, Sb, Ag, and Sn. In this case, it is difficult to see any particular geographic locations of the two. Class 1.2 (in yellow) constitutes the minor compound of the phase with 10% of pixels for the total number of pixels in class 1. It takes the form of fine veins containing the highest concentrations of Cu, Fe and Al compared to the two other sub-classes. For the calcite phase, classes 2.4 and 2.1 (respectively in blue and red) are the most abundant with 50% and 30% of pixels respectively. They are distributed rather homogeneously and are very close in terms of element concentrations except for Y and La. They form the purest calcites, Ca and Mn being their major elements. The situation is very different for classes 2.3 and 2.2 (respectively in yellow and green), which are concentrated in small areas mainly at the borders of class 2. These minor

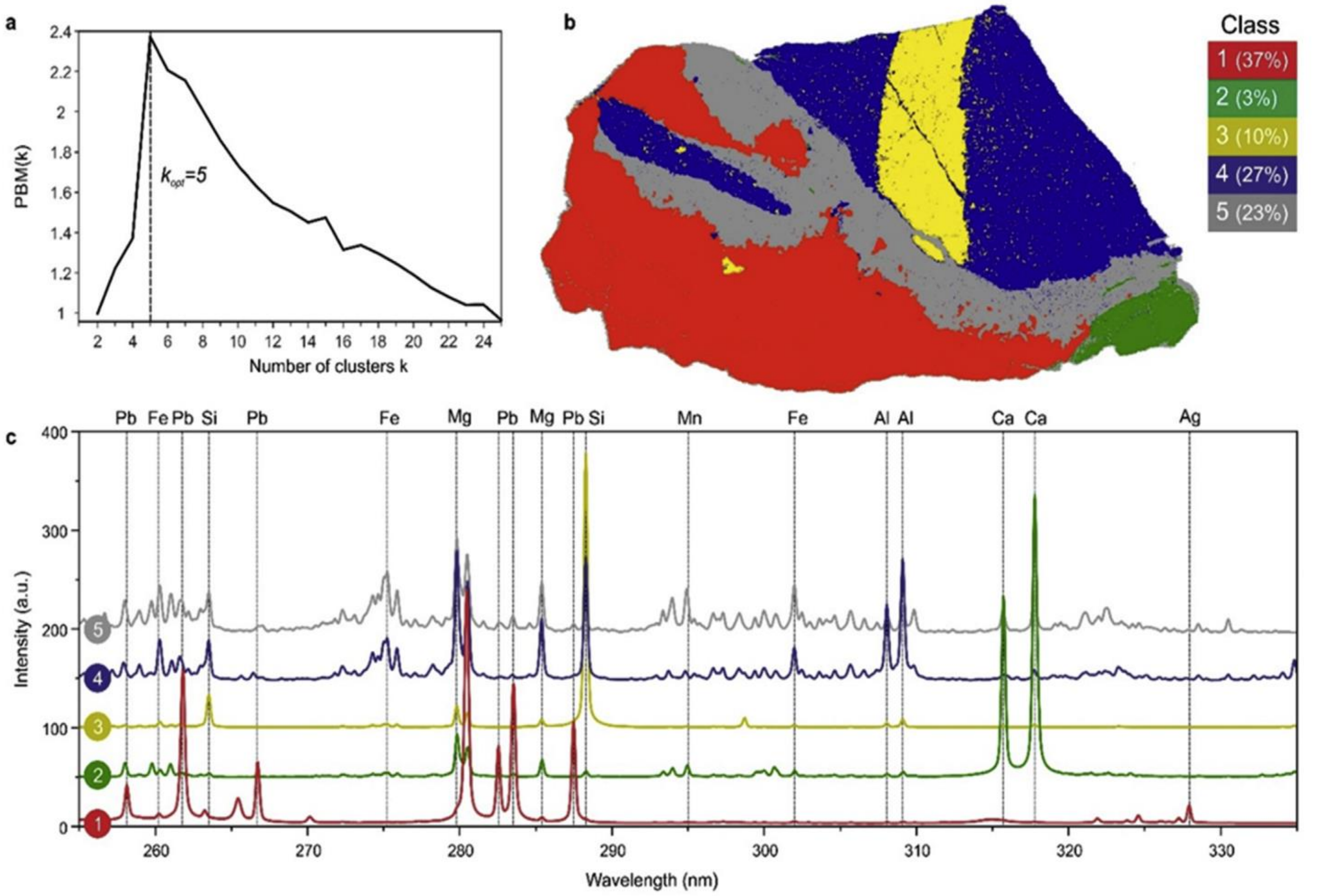


Fig. 3. a) Evolution of the PBM index according to the number of clusters *k*. b) The classification map considering an optimal number of clusters equal to 5. c) Representative spectra of each class.

compounds correspond to 14% and 6% of pixels respectively. It is also remarkable that class 2.2 has the highest concentration of Mg, Si, Fe, and Mn. Moreover, very small contributions of Y and La are now particularly detected in the class 2.3, while being almost totally indectectable from the raw data set. The quartz phase is slightly more complex with 5 sub-classes. However, a more balanced split can be observed between the percentage of pixels of sub-classes. Classes 3.1 and 3.4 (respectively in red and blue) are the most abundant. They are regularly distributed over a trapezoidal area such as class 3.3 (in yellow). For its part, class 3.5 (in grey) is spread all over the class 3 area mostly in the form of tiny clusters. This quartz is really particular because it has by far the highest concentration of Mg, Ca, Fe, Al, and Ti. Class 3.2 (in green) is a minor compound with 9% of pixels. It is mainly observed along a vein through the trapezoidal area. It contains less Si than the classes 3.1, 3.2, 3.3 and 3.4 but more Mg, Ca, Fe, Al, and Ti. The aluminosilicate phase seems less complex with 3 sub-classes. However, from a spectroscopic point of view, they are well-contrasted. Class 4.3 (in red) is the major compound with 71% of pixels, followed by class 4.1 (in blue) with 25%. They are both spread all over the class 4 area. They show high concentrations of Si, Mg, Fe, and Al but also different ratios between them. Class 4.2 (in yellow) is the minor compound of this phase with only 4% of pixels. It is spread all over the area in the form of small clusters. At the same time, it has by far the highest Ti concentration and the lowest concentrations for all other elements. The fifth and last phase i.e. ankerite is certainly the most complex case with six sub-classes and the most contrasted element concentrations. Classes 5.4 (in blue), 5.1 (in green) and 5.5 (in pink) are the most abundant with 34%, 33%, and 21% of pixels respectively. They are distributed rather homogeneously with rather high concentrations of Mg, Ca, and Fe. The last three sub-classes are minor compounds. Class 5.6 (in grey) with 7% of pixels is mainly located at the border of the rock section. It has medium concentrations of Ca and Si, a medium one for Mg and contains neither Fe nor Zn. Class 5.3 (in yellow) with 4% of pixels is only located on one side of the area defined by classes 5.1, 5.4, and 5.5. It has also concentrations of Fe, Mg, Ca and Si comparable to those three previous classes. However, small variations of concentration ratios are observed between them. For its part, class 5.2 (in red) is the less abundant compound with 0.2% of pixels. It is presented in the form of a single cluster. It is the only compound containing Zn and a small concentration of Fe. The other elements are absent. Readers interested in a global representation of the 21 sub-classes in overlay mode should refer to Fig. S2 in the supplementary material. As we have just seen, our strategy allows us to deeply explore LIBS data sets of complex samples providing simultaneously the localization and the identification of major and minor compounds. Class 5.2 is certainly the perfect example of the potential of this approach because it corresponds to the detection of only 730 specific spectra of a given compound over the 2.289.000 present in the considered data set. In a natural way, the PBM index was also used on each cluster of the second levels of clustering demonstrating that there was no more possible discrimination at this level thus ending the exploration of this megapixel LIBS imaging data set.

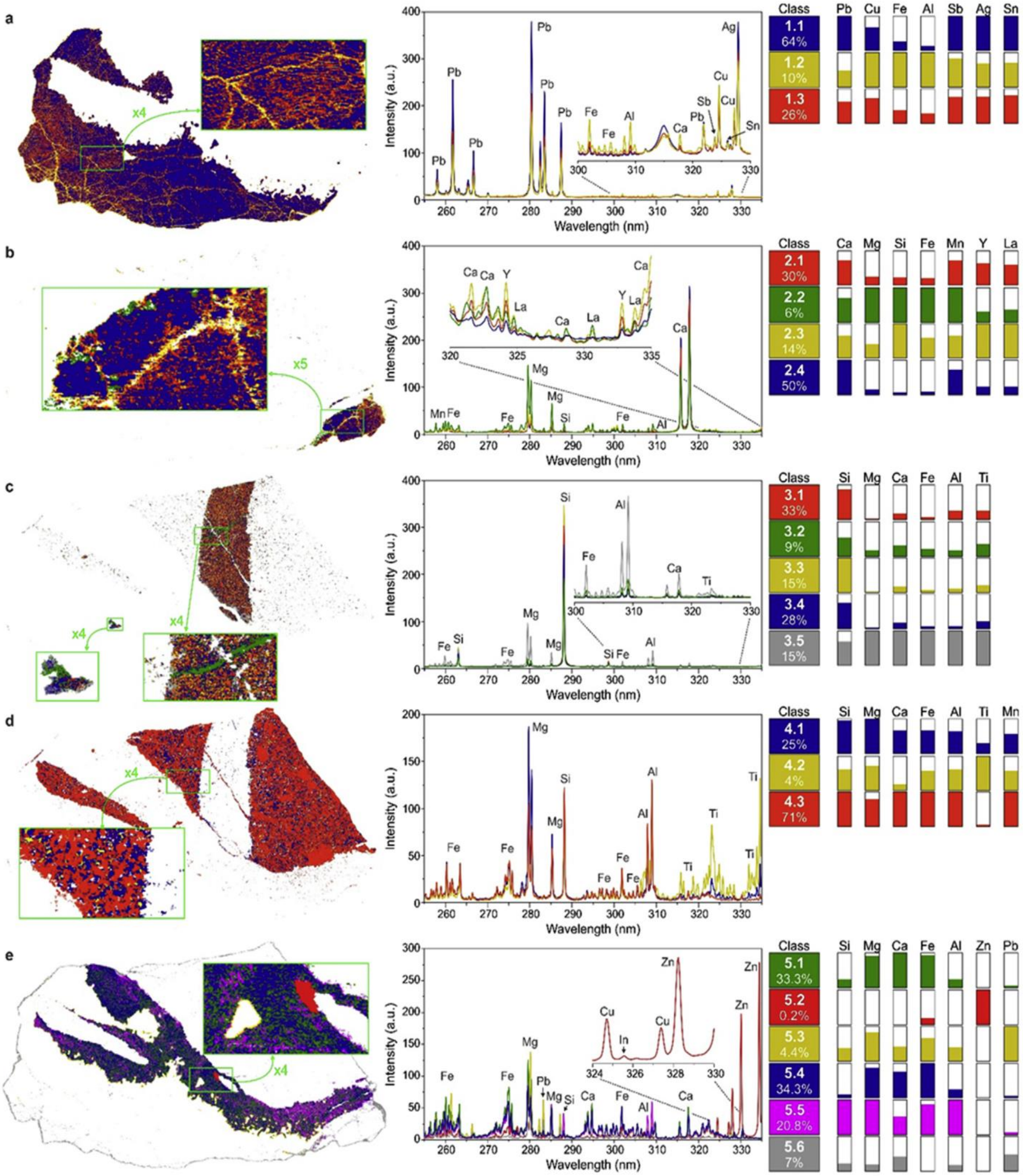


Fig. 4. Classification maps obtained for each phase with corresponding spectra of sub-populations and relative concentrations of elements.

4. Conclusion

The main objective of this work was to evaluate an original strategy called embedded k-means clustering in order to explore a big LIBS imaging data set acquired from a complex mineral sample. More specifically, the idea was to propose a simultaneous identification and localization of both major and minor compounds. From the very start of this work, we have quickly observed that while the traditional signal integration method generates unbiased elemental images most of the time, it remains especially tricky if the objective is to obtain information at the phase level, for the highest as well as the lowest concentrations. Generally speaking, we have demonstrated that multivariate data analysis is an efficient complementary tool to explore LIBS imaging data sets in this particular framework. Indeed, the k-means algorithm has allowed us to group similar pixels (i.e. spectra) without any prior knowledge of class memberships. We have also highlighted the importance of using an index in order to select the right number of clusters, with no *a priori* about the considered sample, which to our knowledge has never been done in the LIBS framework. Lastly, we have shown that our approach based on successive k-means clustering provides a deeper exploration of the sample from major to minor compounds with great sensitivity, without compromise on the detection of both.

Supporting information

A high-resolution image of Fig. 2, the PBM index plots used in the second step of EKM, high-resolution images of Figs. 3 and 4, a global representation of the 21 sub-classes.

Author contributions

The manuscript was written through the contributions of all authors. Moreover, they have given approval to the final version of the manuscript.

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.

CRediT authorship contribution statement

Alessandro Nardecchia: Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Cécile Fabre:** Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Jean Cauzid:** Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Frédéric Pelascini:** Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Vincent Motto-Ros:** Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Ludovic Duponchel:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

Acknowledgment

This work was partially supported by Pulsalys (#L0978-L1294), the French region Grand Est, and the French region Rhônes Alpes Auvergne (Optolyse, CPER2016). In addition, we gratefully acknowledge Tristan Mantoy and the CMT company (Compagnie Minière de Touissit) for providing the sample. Lastly, we would like to thank Ms Astrid Marissal for proofreading.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2020.04.005.

References

[1] L. Jolivet, M. Leprince, S. Moncayo, L. Sorbier, C.-P. Lienemann, V. Motto-Ros, Review of the recent advances and applications of LIBS-based imaging, Spectrochim. Acta Part B At. Spectrosc. 151 (2019) 41–53, https://doi.org/10.1016/j.sab.2018.11.008.

[2] R. Gaudiuso, N. Melikechi, Z.A. Abdel-Salam, M.A. Harith, V. Palleschi, V. Motto-Ros, B. Busser, Laser-induced breakdown spectroscopy for human and animal health: a review, Spectrochim. Acta Part B At. Spectrosc. 152 (2019) 123–148, https://doi.org/10.1016/j.sab.2018.11.006.

[3] C. Fabre, D. Devismes, S. Moncayo, F. Pelascini, F. Trichard, A. Lecomte, B. Bousquet, J. Cauzid, V. Motto-Ros, Elemental imaging by laser-induced breakdown spectroscopy for the geological characterization of minerals, J. Anal. At. Spectrom. 33 (2018) 1345–1353, https://doi.org/10.1039/C8JA00048D.

[4] J.O. Cáceres, F. Pelascini, V. Motto-Ros, S. Moncayo, F. Trichard, G. Panczer, A. Marín-Roldán, J.A. Cruz, I. Coronado, J. Martín-Chivelet, Megapixel multi-elemental imaging by Laser-Induced Breakdown Spectroscopy, a technology with considerable potential for paleoclimate studies, Sci. Rep. 7 (2017) 1–11, https://doi.org/10.1038/s41598-017-05437-3.

[5] F. Trichard, F. Gaulier, J. Barbier, D. Espinat, B. Guichard, C.-P. Lienemann, L. Sorbier, P. Levitz, V. Motto-Ros, Imaging of alumina supports by laser-induced breakdown spectroscopy: a new tool to understand the diffusion of trace metal impurities, J. Catal. 363 (2018) 183–190, https://doi.org/10.1016/j.jcat.2018.04.013.

[6] J. El Haddad, L. Canioni, B. Bousquet, Good practices in LIBS analysis: review and advices, Spectrochim. Acta Part B At. Spectrosc. 101 (2014) 171–182, https://doi.org/10.1016/j.sab.2014.08.039.

[7] J.-B. Sirven, B. Bousquet, L. Canioni, L. Sarger, Laser-induced breakdown spectroscopy of composite Samples: comparison of advanced chemometrics methods, Anal. Chem. 78 (2006) 1462–1469, https://doi.org/10.1021/ac051721p.

[8] I. Gaona, J. Serrano, J. Moros, J.J. Laserna, Range-adaptive standoff recognition of explosive fingerprints on solid surfaces using a supervised learning method and laser-induced breakdown spectroscopy, Anal. Chem. 86 (2014) 5045–5052, https://doi.org/10.1021/ac500694j.

[9] N.C. Dingari, I. Barman, A.K. Myakalwar, S.P. Tewari, M. Kumar Gundawar, Incorporation of support vector machines in the LIBS toolbox for sensitive and robust classification amidst unexpected sample and system variability, Anal. Chem. 84 (2012) 2686–2694, https://doi.org/10.1021/ac202755e.

[10] T. Zhang, H. Tang, H. Li, Chemometrics in laser-induced breakdown spectroscopy, J. Chemom. 32 (2018) e2983, https://doi.org/10.1002/cem.2983.

[11] M. Bouabdellah, J.F. Slack, Mineral Deposits of North Africa, Springer, 2016.

[12] 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, Statistics, 1967.

[13] 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.

[14] Y. Gimenez, B. Busser, F. Trichard, A. Kulesza, J.M. Laurent, V. Zaun, F. Lux, J.M. Benoit, G. Panczer, P. Dugourd, O. Tillement, F. Pelascini, L. Sancey, V. Motto-Ros, 3D imaging of nanoparticle distribution in biological tissue by laser-induced breakdown spectroscopy, Sci. Rep. 6 (2016) 29936, https://doi.org/10.1038/srep29936.

2.3.2. Conclusions and future perspectives

KM clustering has shown to be one of the most interesting methods for classification, particularly when it is not possible to have many information about the nature of the specimen, due to the iterative nature that is the core of how this approach works. In fact, being an unsupervised method, KM tries to find natural groupings of spectra in the considered dataset, which will represent different chemical compounds. One of the main tasks of this PhD has been the use of chemometrics focusing on LIBS imaging with the aim of finding interesting approaches that can help the operator in solving different limitations related to this spectroscopy. The main purpose of this chapter, and specifically this paragraph, has been to show an alternative method to the standard approach based on the use of KM clustering, aiming to the detection, identification and localization of not only the major, but also the minor compounds and traces into a big dataset, leading to promising results. The limitation of the presented EKM approach is that it can be nevertheless complex, because the clustering has to be applied several times. In fact, in a first step it is used to identify the main different regions of the map, containing the different elements (or compounds). Then, again, the same procedure is repeated, this time on each single subregion, in order to find new details coming from minor compounds and traces. Having said that, more chemometric tools may be applied to LIBS analysis, concerning various and different tasks and approaches, with the general goal of proposing always innovative data analyses that could replace the routine methodologies. Particularly, next chapters will focus in LIBS image analysis, for which different aspects regarding especially, but not exclusively, this spectroscopy will be faced and discussed more in detail.

CHAPTER 3

3. CORRECTION OF THE SPECTROSCOPIC INFORMATION: THE IMPORTANCE OF USING RELIABLE DATA

3.1. Why it is important to have good data before any analysis and/or use of chemometric approaches

As briefly stated in Chapter 1, one of the most important tasks in any analytical investigation, no matter the research field, is the use of reliable spectral information [109]. In fact, by way of example, if the right preprocessing or analysis settings are not used, there is a real risk to extract incorrect details that would drive the interpretation of the data in the wrong direction. Among the possible acquisition mistakes that can be faced in this scenario, one in particular has been studied during this PhD. During an acquisition, many spectroscopies can generate the phenomenon of a saturated signal. This kind of response, also known as clipping, is a distortion of the signal, when it exceeds a certain threshold. As a consequence of this, saturated bands with their characteristic plateau present numerical values that do not correspond to the analytical reality of the analyzed sample. Clearly, this kind of response cannot be used to generate any result, because it is far to be reliable and accurate. Therefore, it is fundamental to find a way to deal with this particular artifact. If saturation is observed on a spectrum acquired in a bulk analysis of a single sample, the situation is easily solved. In fact, the specimen can be again acquired, changing the preparation modality or the instrumental characteristics in order to avoid the presence of this phenomenon. Very different is the scenario in which many samples are acquired. Obviously, it is mandatory to set unique experimental conditions to be applied to all samples. Nevertheless, this step can be more challenging than expected, particularly when the nature of the acquired matrix is very heterogeneous, such as normally happens when the used technique is the hyperspectral imaging. In fact, for given acquisition conditions it is possible to acquire in the region of interest of the sample thousands (or even more) spectra. Since each spectrum corresponds to a specific micro-surface of the sample with potentially different molecular distributions, it is quite likely that some of them are saturated, no matter the acquiring conditions. In addition, some instruments are destructive, meaning that the acquisition can be carried out only once. This is for example a LIBS-related problem. This kind of spectroscopy, as already described in the manuscript, shows very interesting characteristics such as the fast acquisition rate and its resolution limits. Nevertheless, using a laser as excitation source, it generates the ablation of the surface of the sample, reason why on occasions it can be impossible to repeat the acquisition. A practical solution exploited over the years to overcome the instrumental limitation represented by the saturation is here briefly explained. The cited approach is based on the exclusion of the saturated signals instead of correcting them. In this kind of scenario, two different strategies can be chosen. In the first case, known as row-wise deletion, the samples corresponding to some saturated signals will be suppressed. This means that the final quantity of investigated spectra will be lower than the initial one. Clearly, this can be a solution, but at the same time, this kind of approach can be a problem if the total number of deleted samples is too significant. It is also important to consider the fact that using this strategy, especially in hyperspectral imaging, there is a real chance to lose some pure and particular information related to the investigated sample. On the other hand, a second solution is represented by the idea of removing only the spectral variables related to saturated signals, the column-wise deletion. In this case, all the samples will be kept, but at the same time some variables will be deleted, the ones that present saturation for at least one specimen. It goes without saying, part of the total spectral information that can be related to very fundamental details of the analyzed matrix, is completely lost in this way. In other words, both the strategies can be applied to remove saturated signals, but neither the first nor the second method clearly correspond to a good solution when an exhaustive chemical study is required. Also in this case, chemometrics can be used as an interesting alternative to the routine approaches. Specifically in this kind of situation, in which the best suggested solution is represented by the removal of the artifacts, and so the possible loss of important details related to the studied matrix, here it is proposed another approach. Specifically, it is based on the use of the multivariate information contained in the sample to generate a prediction of the missing values in order to correct the saturated signals by the use of statistical imputation, as following explained.

3.2. Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed phenomenon?

3.2.1. Introduction

Among the different spectroscopies, LIBS is for sure one of the most interesting that in the last years obtained an increasing importance due to its suitable characteristics. Nevertheless, setting the experimental conditions can be sometimes a hard task, particularly due to the complex nature of the different matrices that can be analyzed. Therefore, a real risk is the one of generating saturated signals, which would lead to incorrect results, as previously explained. Despite the fact that saturation is a phenomenon that can be experienced within many spectroscopies, the present work focuses particularly in the resolution of this artifact in LIBS imaging. This is due to various reasons. First of all, as already stated, this kind of spectroscopy is a very interesting instrumentation that can be used in many research areas for the elemental analysis. The con is that, using a laser as excitation source and ablating the matter, it is a potential destructive technique. This leads to the fact that, potentially, an analysis cannot be executed a second time on the same surface also if required, such as in the case that some artifacts (e.g., saturated signals) are present. Also important is that LIBS can generate from thousands to millions of spectra in a very short time, leading to very interesting and heterogeneous acquired data cubes. Therefore, the possibility of obtaining some clipped signals is very likely. Last but not least, LIBS is related to very fine spectral features if compared with other spectroscopies in which broadened bands are generated. Furthermore, each element is normally related to more bands, distributed in the whole spectral domain. It means that if a particular signal is saturated, others (with a lower intensity) related to the same element will be probably available. These aspects are very suitable for the here proposed method. The general idea of this work, that has been published in Analytica Chimica Acta, Volume 1157 (2021) [213], is to generate multivariate regressions using all the spectral variables that are not saturated in order to predict the right values related to the clipped signals. As a first step, the data containing saturations will be considered as missing values, to be then replaced with new calculated ones. The used approach is the imputation, a field of statistics in which the gaps in the data are filled with plausible values that are calculated within the data themselves. In this way, it is possible to keep the initial dimensions of the dataset, but correcting the artifacts coming from the acquisition.

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed phenomenon?

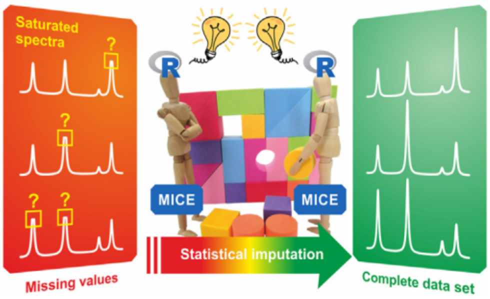
Alessandro Nardecchia [a], Vincent Motto-Ros [b], Ludovic Duponchel [a, *]

[a] Univ. Lille, CNRS, UMR 8516 – LASIRE – LAboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille, F-59000, France
[b] Institut Lumière Matière, UMR 5306, Université Lyon 1 - CNRS, Université de Lyon, Villeurbanne, 69622, France

HIGHLIGHTS

- Important to correct saturated signals in a spectroscopic data sets at the risk of having a biased view of the sample.
- Row-wise and column-wise deletion are not good strategies to manage saturated signals in spectroscopic imaging.
- Considering satured signals as missing values is an original approach.
- Statistical imputation is a good strategy in order to recover the information lost during the measurement.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 January 2021
Received in revised form
23 February 2021
Accepted 1 March 2021
Available online 12 March 2021

Keywords:
Saturated signal
Imaging spectroscopy
Statistical imputation

ABSTRACT

We have all been confronted one day by saturated signals observed on acquired spectra, whatever the technique considered. A saturation, also known as clipping in signal processing, is a form of distortion that limits a signal once it exceeds a threshold. As a consequence, clipped or saturated bands with their characteristic plateau present numerical values that do not correspond to the analytical reality of the analyzed sample. Of course, analysts know that they cannot consider these erroneous values and therefore reconsider either sample preparation or instrument settings. Unfortunately, there are many experiments today (and this is the case in spectroscopic imaging) for which we will not be able to fight against the saturation effect that will undeniably be observed on the acquired spectra. The aim of this article is first to show why it is important to correct these saturation effects at the risk of having a biased view of the sample and more specifically in the context of multivariate data analysis. In a second step, we will look at strategies for managing saturated bands. An original concept will then be presented by considering saturated values as missing ones. A statistical imputation strategy will then be implemented in order to recover the information lost during the measurement.

1. Introduction

Saturation is a phenomenon regularly observed in spectroscopy. Its presence can be linked to various factors such as sample preparation, specific photon-matter interactions or even instrumental

* Corresponding author.
E-mail address: ludovic.duponchel@univ-lille.fr (L. Duponchel).

https://doi.org/10.1016/j.aca.2021.338389

limitations in the detection chain. For example, too high extinction coefficients and/or too high pathlengths in the mid-infrared spectral range reduce so much the number of non-absorbed photons arriving at the detector that the absorbance levels are infinitely high, values that cannot of course be transcribed in a spectrum, or by default in the form of a plateau. On the other hand, in the case of scattering or emission measurements as in Raman, fluorescence or LIBS (laser-induced breakdown spectroscopy), the number of photons collected by the measurement chain can be sometimes so important that it cannot be transcribed into the spectrum. Again in this case, a clipping effect is observed which does not allow to observe the real values to be measured. Generally speaking, we can say that a saturation may occur when a signal is recorded by a detector that has constraints on the range of data it can measure. This can therefore be the case when a signal is digitized using an analog-to-digital converter, or any other time an analog or digital signal is transformed, particularly in the presence of gain.

When a saturation phenomenon is observed on a spectrum acquired for bulk analysis of a single sample, the analysts know that they have has to reconsider the preparation of their samples or the acquisition parameters depending of course on the constraints related to the experiment under consideration. The newly acquired spectrum then has every chance this time to present values that are representative of the analytical reality of the sample. The situation is quite different when we have to do bulk analyses on a set of samples with the final objective of comparing their spectra. For this specific purpose, we must then set unique experimental conditions that will be applied to all samples. We could then easily observe perfectly exploitable spectra next to others that are potentially saturated. This is a situation that often occurs in spectroscopic imaging when exploring a single and complex heterogeneous sample. Indeed, for given acquisition conditions, hundreds, thousands or even hundreds of thousands of spectra are acquired in a region of interest of the sample. Since each spectrum corresponds to a specific micro-surface of the sample with a potentially different molecular distribution, it is quite likely that some of them are saturated. If we are lucky we might be able to find experimental conditions that remove these saturations. Nevertheless, we must not lose sight of the fact that it is not always possible to reproduce the experiment a second time, for example when the technique is destructive as in LIBS.

In general, we can say that we always try to avoid the saturation phenomenon as much as possible. Unfortunately, it is observed in many cases and it is necessary to deal with these data as they are. The question that then arises is the following: what should we do with saturated values that we know to be systematically erroneous? Fig. 1a gives a schematic representation of a dataset with six spectra of which three contain saturations highlighted in red. We can notice first that it is not always the same bands that are saturated in this dataset used as a toy example. Second, the number of saturations in a given spectrum is quite variable. Fig. 1b presents the two strategies typically used to manage potential spectral saturation in a dataset. The idea is finally very simple since knowing that saturated values do not represent the true values, it seems logical to remove them from the acquired dataset. We then have a first possibility which is to remove all the spectra as soon as they contain at least one saturated spectral variable also known as row-wise deletion. This strategy might seem satisfactory because it is simple to implement, but it is not flawless. Indeed, we could then remove a spectrum made up of several hundreds of spectral variables and thus potentially a very large amount of molecular/atomic information just because a single variable would be saturated, for example. We would then have a significant loss of chemical information as in the present case study where only 50% of the spectra would be kept for multivariate analysis. In the specific case of spectroscopic imaging, we would then end up with areas of the image without defined chemical information. From a more statistical point of view, we would also have a biased analysis since we would no longer have the initial population of acquired spectra. In a second strategy known as column-wise deletion, we could suppress a spectral variable in the data set as soon as at least one of the spectra of the dataset presents a saturation on this same variable. This strategy is no more satisfactory because a significant loss of information would still be observed. In the case of the presented example, we notice that such a strategy would remove almost all the spectral information from the dataset. Thus even if these two strategies are regularly exploited in spectroscopy, we see that they are unsatisfactory on different aspects.

Starting from the observation that a saturated value in a spectrum is an erroneous one, we propose in this work to consider it as a missing value. It is indeed more relevant to say that a value could not be measured than to exploit a value that finally does not represent a reality. Thus in the matrix representation of the data set in Fig. 1b, red boxes that were initially saturated values will become missing ones. In statistics, the art of dealing with missing values in a matrix is called imputation [1]. It is in fact the process of replacing missing data with substituted values. By approaching the problem of saturation in this way, we see that we can then work on a data set while keeping its initial dimensions, i.e. with the initial number of spectra and spectral variables resulting from the acquisition. Thus in this work, three different spectroscopic imaging datasets will first be used to show the need to manage saturations present in the spectra at the risk of seeing many artifacts during multivariate analyses generating biased chemical images and extracted spectral profiles. The principle of imputation will of course be explained and the analysis of the corrected datasets will allow us to demonstrate the benefits of this concept to find chemical images and corresponding spectroscopic information representative of the analytical reality of complex samples.

2. Material and methods

2.1. Imputation

Imputation is a field of statistics. The great idea in imputation is to fill gaps in the data with plausible values, the uncertainty of which is coded in the data itself. There are many ways of doing data imputation today [1]. However, we will use in this work the so-called ‘multiple imputation’ now considered as the best general method to deal with incomplete data (i.e. containing missing values) in many scientific domains [2–4]. Our goal here is of course not to redo a whole development of the theory of imputation but to explain some general principles in order to understand the results presented in this work. Readers who would nevertheless like to have all the details on this topic are invited to read other works specifically dedicated to statistic [1,3]. The two main approaches for imputing multivariate data are called joint modeling [5] (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE) [6]. As the JM approach is often more constraining from a statistical point of view to be applied, the MICE method has been considered in this work. MICE specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable (i.e. containing missing values). In this work, a regression model is developed using the complete variables of the matrix as input and a given incomplete variable as output. Once this model is established, we can then use it to predict missing values of spectra at this specific spectral variable based on known values in the matrix. We will thus have as many regression models developed as spectral variables containing missing values in the

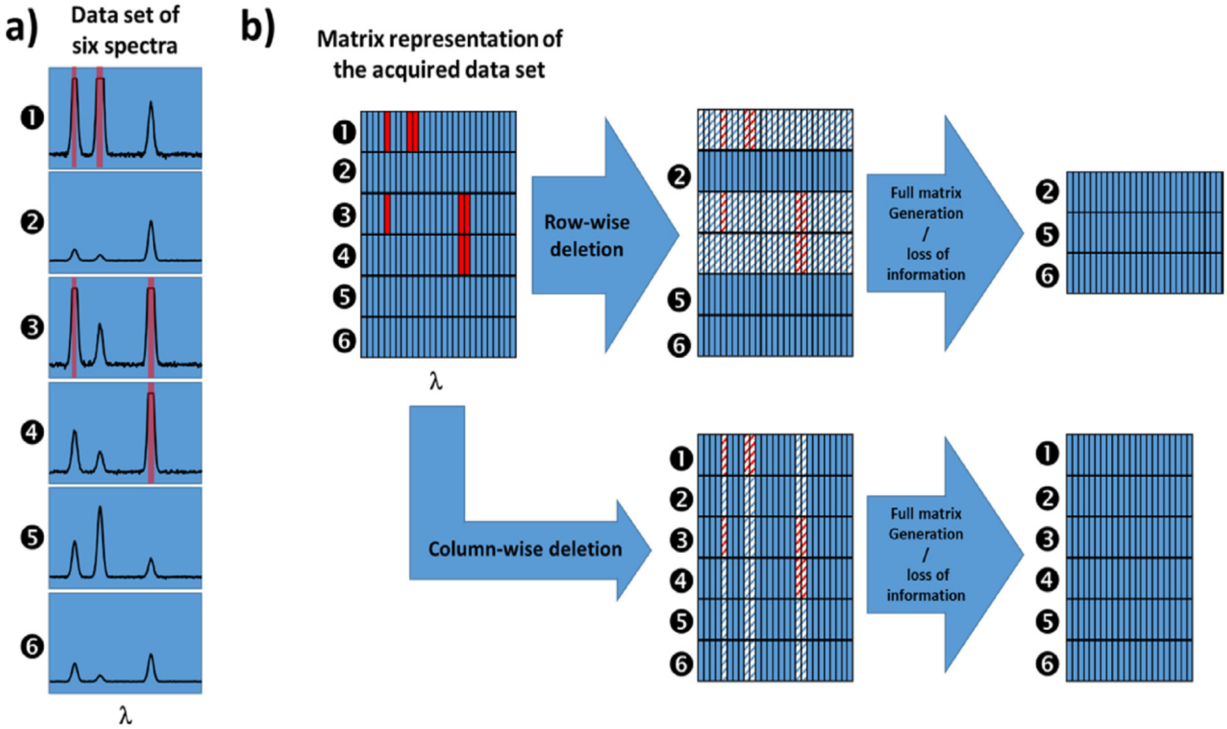

a)
Data set of six spectra
❶
❷
❸
❹
❺
❻
λ
b)
Matrix representation of the acquired data set
λ
Row-wise deletion
Full matrix Generation / loss of information
Column-wise deletion
Full matrix Generation / loss of information

Fig. 1. a) A schematic representation of a toy example with six spectra containing saturations highlighted in red. b) The two conventional strategies to manage saturated signals in a dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

considered dataset. At the end of this imputation procedure, we then find a full matrix free of missing values that we can then explore with usual multivariate methods. In this work, all imputation calculations have been done under the R environment using MICE, an open source R package. Source code and documentation can be found at https://github.com/amices/mice.

2.2. Multivariate data analysis

Principal component analysis (PCA) is one of the most flexible and effective chemometric method for exploratory data analysis applied to hyperspectral imaging [7]. It is indeed very sensitive and thus allows the detection of very low variance levels. Its use will first of all allow to see artifacts generated during a direct use of saturated spectral datasets but also to estimate in a second time the efficiency of the corrections brought by our imputation strategy. All PCA calculations were performed under the Matlab 2016b environment.

2.2.1. Dataset #1: a simulated sample

The first hyperspectral imaging dataset consists of synthetic spectra that could have been acquired using LIBS. The advantage of using such simulations lies in the fact that all the parameters potentially influencing a given problem are under controlled. In this way, we often have a less biased view of the phenomena and a real generalization is possible. As we will see further on, it will also be a way to vary the importance of the saturation phenomenon. On the basis of the spectroscopic information given by the Kurucz database [8], we first simulated the emission spectra of silver, aluminum and arsenic by considering a typical plasma temperature and electron density (9000 K and 5.1016 cm^{-3} respectively). In order to be the most faithful with the spectral reality, we then applied a Lorentzian profile with a linewidth of 0.15 nm to each emission line, corresponding to the resolution of classical spectrometer used in LIBS [9]. In the considered spectral range (250–350 nm), several emission lines of Ag, Al, and As were observed with various intensity ranges. On the basis of these three pure spectra, it was then possible to generate by linear combination 62,880 spectra of mixtures in percentages ranging from 100 to 0 for each of them. A white noise of 5% has also been added to each spectrum. In this way, we obtained a hyperspectral data cube defined by 131 pixels × 480 pixels x 2018 wavelengths. Fig. 1S in supplementary material presents the three element spectra, all the generated spectra of mixtures in overlay mode as well as the spatial distribution of the different elements in this synthetic sample.

2.2.2. Dataset #2: a lung biopsy

The second dataset used in this work corresponds to a LIBS imaging experiment conducted on a lung biopsy of a patient with severe emphysema [10]. Note that the patient signed informed consent, and the clinical procedure was approved by the local ethics committee. The LIBS imaging has been conducted with a protocol dedicated to paraffin-embedded tissues, as described in a previous work [11]. The aim of such application was to characterize the distribution of metallic particles (from nanometric to micrometric size) in tissue biopsies, which represent a precious help for clinicians to diagnose the cause of the exposition (i.e. environmental and/or occupational). This spectroscopic experiment is a good example of a case where saturated spectra cannot be avoided. Since the concentration, composition, location and size of the particles are not known prior the experiment, the measurement system requires an extremely large dynamic in term of detection, typically from few ppm to a few percent in mass. Despite our efforts to set up the experimental parameters as optimized as possible, it is not uncommon to have a significant number of spectra showing saturations on a LIBS image as in this case. The size of the analyzed area of the biopsy was 5.42 mm long by 3.18 mm wide with a spatial resolution of 20 μm. We have thus acquired 43,089 spectra over a spectral range from 282.01 to 310.03 nm and an approximate spectral resolution of 0.04 nm. The hyperspectral data cube was therefore defined by 271 pixels × 159 pixels x 644 wavelengths.

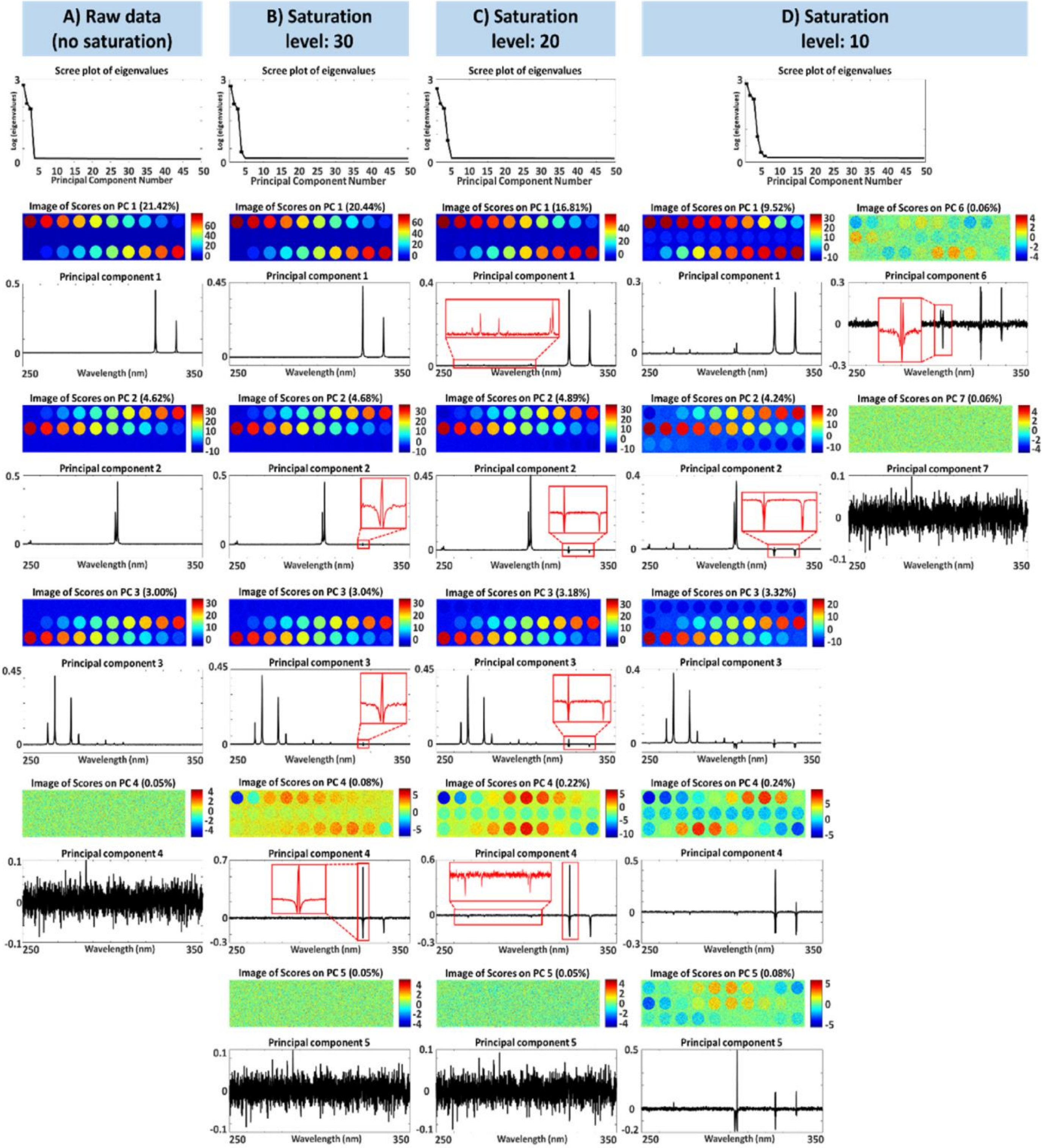


Fig. 2. Principal component analysis on a) raw data, b) on the saturated dataset with a 30 saturation level, c) with a 20 saturation level and, d) with a 10 saturation level.

2.2.3. Dataset #3: a rock section

The third dataset was also acquired using a LIBS imaging instrument on a banded iron formation rock consisting of alternating layers of iron oxides and silicates. We selected this sample because it somehow allowed us to find approximately the same chemical distributions for two contiguous zones of the sample on which we could set different acquisition parameters. This procedure of selection of analysis area was necessary because we must not forget that LIBS is a destructive technique. Therefore, we could not analyze the same area several times. As a consequence, two successive zones each having a size of 20 mm long and 2 mm wide were analyzed following the protocol already used for other work [12]. A spatial resolution of 20 μm and a spectral range from 245.85 to 334.03 nm were considered for these acquisitions generating two hyperspectral datasets defined each by 1000 pixels × 100 pixels x 2048 wavelengths. The laser pulse energy and the detection gate were adjusted for these two sub-zones in order to control the saturation level. Indeed, our aim was to obtain no saturated lines on the first zone of the sample (gate: 1 μs; energy: 1.2 mJ) and saturation of Si emission lines on the second zone (gate: 5 μs; energy:

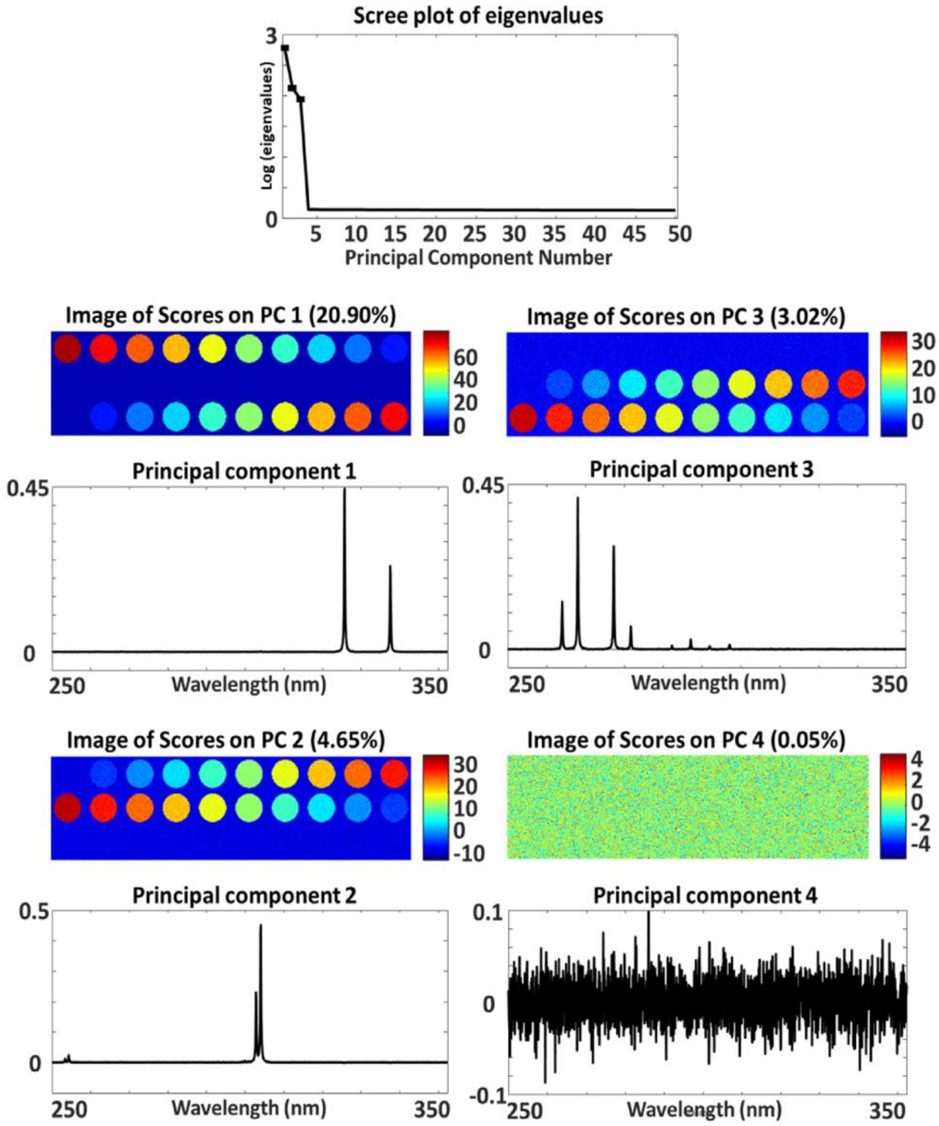


Fig. 3. Principal component analysis on the imputed dataset with an initial saturation level equal to 20.

1.2 mJ). All the other acquisition parameters such as the delay and detector gain was kept constant.

3. Results and discussion

Before tackling the problem of correcting saturation in the spectra, it is important to understand how this is necessary to manage it, at the risk of giving a completely biased vision of the analyzed sample. For this purpose, we will use the first dataset of simulated spectra. Fig. 2a thus presents the results of a first principal component analysis applied to the raw data (i.e. without saturated signals). Unsurprisingly, we note first of all that there are three significant eigenvalues in the scree plot that correspond to three spectral contributions. More specifically, the first three principal components respectively extract the spectra of the three pure elements Ag, Al and As. This is quite logical since there is no correlation between these elements in the considered dataset. The scores images then perfectly reproduce the distributions of the three elements given in SI Figure 1S. In a natural way, the fourth principal component extracts the noise variance. From the raw data, we then simulate a first level of saturation by clipping all emissions above 30, knowing that the maximum emission observed on the initial spectra is around 43. SI Figure 2S gives the location of the saturated signals at the spatial and spectral levels. 5047 spectra thus present saturations, i.e. 8% of all the spectra. We notice that saturations are present in areas where silver is the most concentrated with a percentage higher than 80%. From a spectral point of view, it is the most intense line of silver which is naturally saturated. Fig. 2b shows again the PCA results on these new saturated data. The consequences are not long in coming since a fourth significant component is already detected in the eigenvalues scree plot. Of course, this is not normal because we know that only three elements are present. Compared to the initial results (in Fig. 2a), both the first principal component and the first scores map are no different. Nevertheless, there is a small decrease in the expressed variance from 21.42 to 20.44%. As far as the second and third scores maps are concerned, they are quite comparable to those observed from the non-saturated dataset. On the other hand, the corresponding principal components show small artifacts in the spectral region of the saturated Ag band (highlighted by red boxes in the corresponding figure). Finally, the fourth principal component specifically reflects the saturation phenomenon observed on silver for an expressed variance of 0.08%. We see many structures in the corresponding scores map but we know that they do not reflect any analytical reality. We observe even more the typical W-shaped artifact in this principal component. This shape can be explained

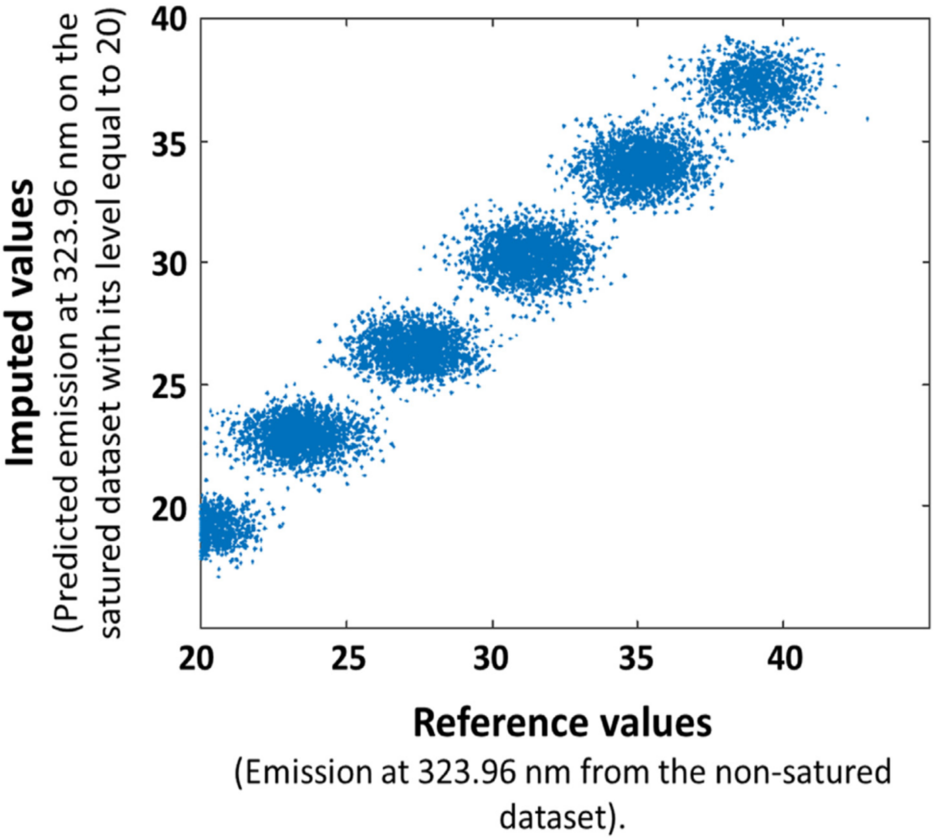


Fig. 4. Emission of a silver line predicted by the imputation model as a function of known values at the wavelength 323.96 nm from non-saturated data.

quite well. Indeed, when saturations are potentially present at a given wavelength, principal components have to express the variance precisely at this wavelength for the unsaturated spectra in the dataset but other ones also have to express variances specifically localized on the feet of this same peak. In other words, a clipped band at a certain wavelength of a given spectrum is no longer homothetic to an unsaturated band of another spectrum. So if we do not have any prior information about this dataset, we see that even a limited saturation level can induce the extraction of erroneous information at the spatial and spectral levels about the sample being explored. It is then interesting to amplify the saturation effect by considering this time a saturation level equal to 20. Under these new conditions, 10,119 spectra are saturated, i.e. 16% of all the spectra. Once again, saturations are present in areas where Ag is the most concentrated, but this time for values higher than 50% (SI Figure 3S). We are also starting to see saturated spectra for pure Al pixels. Fig. 2c show PCA results of this new dataset. Four significant contributions are still observed, but with an even greater influence of artifacts on all components. We notice thus on the first principal component which should be specific to Ag that contributions from Al and As are now easily observed. We note here that the presence of saturations can also create spurious correlations. The second and third components have even greater expressed variances and ever more pronounced ‘W-shapes’. While the scores maps are relatively little changed under these new conditions for the first three principal components, this is not at all the case for the fourth one. This high-contrast, low-noise scores image could indeed lead us to believe that real chemical compounds are present, which is of course not the case. In a final step, the saturation is further increased by considering this time a level equal to 10. This situation is extreme since 27,870 spectra are now saturated, i.e. 44% of the dataset. What is more, the saturated pixel location map shows that this percentage is even underestimated since almost all of the areas that should contain the three elements are almost all saturated, the unsaturated areas being mainly the background (SI Figure 4S). At the same time, we observe that almost all emission lines show saturation over the entire spectral range. PCA results of this new dataset is given in Fig. 2d. Under these conditions where saturation is omnipresent, six significant contributions are now detected. The first principal components are more and more perturbed. They are now undeniably different from pure spectra extracted on unsaturated spectra. As examples, the first principal component contains distinct contributions from all three elements and the following ones, which contain more and more artifacts, have equally increasing explained variances. Two new principal components 5 and 6 are also extracted in these conditions with quite singular scores maps. The presence of these additional principal components is explained by the fact that the variance of all the saturated peaks must of course be explained, but also that they are not necessarily saturated at the same time in all the spectra of the dataset. Generally speaking, we can say that the more spectral variables containing saturations, the more parasitic principal components and biased scores maps are extracted. From this first experiment, it is obvious that we cannot directly process saturated spectra with multivariate tools at the risk of making very hazardous exploration of unknown and complex samples for which we have no a priori. Based on this observation, we know that we must now absolutely manage these saturations. Thus, if we wanted to implement a row- or a column-wise strategy that is simple to set up in order to eliminate these saturations, we would quickly observe too many deleted pixels or a particularly small explored spectral domain. It is in this sense that the proposed imputation strategy makes sense by first considering saturated signals as missing values and then applying the MICE approach to make statistical estimates of the latter, i.e. retrieve lost spectral information and consequently a full data matrix. Imputation was therefore applied to the previous

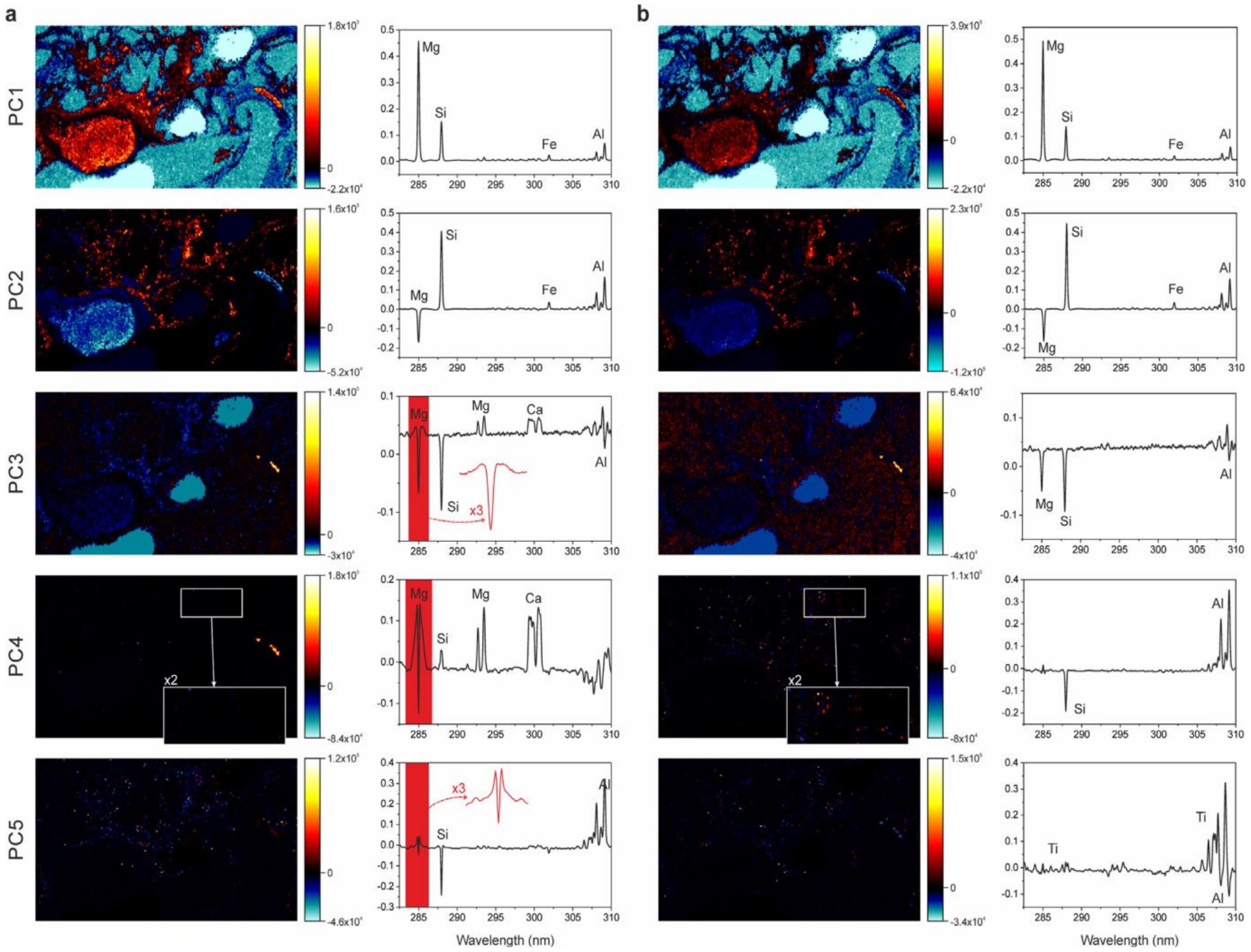


Fig. 5. a) Principal component analysis on raw spectral data of the lung biopsy. (b) Same analysis on spectra corrected with imputation.

datasets by considering the three levels of saturation. In order to appreciate the quality of the data reconstruction, we simply reapplied PCA on these three new datasets. Fig. 3 shows the results concerning the intermediate saturation level equal to 20. The results for the other levels are presented in the supplementary material (SI Figure 5S). By comparing these new extractions with those obtained on unsaturated data, we observe rather spectacular results. First of all, we recover the three significant components on the eigenvalues scree plot, which is consistent with the initial results. Moreover, principal components and corresponding scores maps are also very comparable to the initial ones. These good results can be explained by the fact that the multivariate regressions used in the MICE approach predict missing values rather well. By way of illustration, Fig. 4 shows the emission predicted by the imputation model as a function of known values at the wavelength 323.96 nm from non-saturated data, this silver emission line being the most often saturated for a saturation level equal to 20. Looking specifically at the results concerning the most saturated dataset (SI Figure S5, saturation level equal 10), some readers might say that despite the three significant contributions detected on the scree plot, it is possible to observe information related to a fourth component at both spectral and spatial levels. This would be quite commendable but we must not lose sight of the fact that these contributions are very close to the noise level. Moreover, these results were obtained from a dataset for which almost all the spectra were saturated, which could not be more challenging.

In this second part, we propose to explore a lung biopsy sample. This sample is particularly interesting because the analyzed area of lung presents a certain diversity of materials since we have naturally biological tissues but also mineral phases and metal particles localized in specific sub-areas. In these conditions, we quickly understand that it is almost impossible to find acquisition conditions allowing us to avoid saturation over the entire surface analyzed. In a way, one always make a bet before such an analysis because the considered spectroscopy is destructive and it is not possible to return to this sample area with new acquisition parameters. Location of the spectra containing saturations on the surface of this sample is shown in SI Figure 6S. Although only 1014 spectra out of the 43,089 in total are saturated (i.e. 2.35%), this phenomenon is finally observable almost everywhere, mainly on two well-localized areas (denoted A and B in this figure) but also in the form of single pixels scattered over almost the entire surface of the sample. Additionally, Figure 6S shows that saturations are observed for almost all emission bands in the considered spectral range. Fig. 5a and b shows PCA results on raw spectral data and spectra corrected with imputation respectively. Differences are noticed very quickly if we look at the contributions of each principal component in these two conditions two by two. So even though the first principal component is quite comparable in both cases with the main spectral contributions observed for Mg and Si but also smaller ones for Al and Fe, the associated scores maps are very different. Indeed, there is an overestimation of this first contribution for raw data on zones A and B of the sample but also widely around zone B. For its part, the first scores map associated with the corrected data mainly

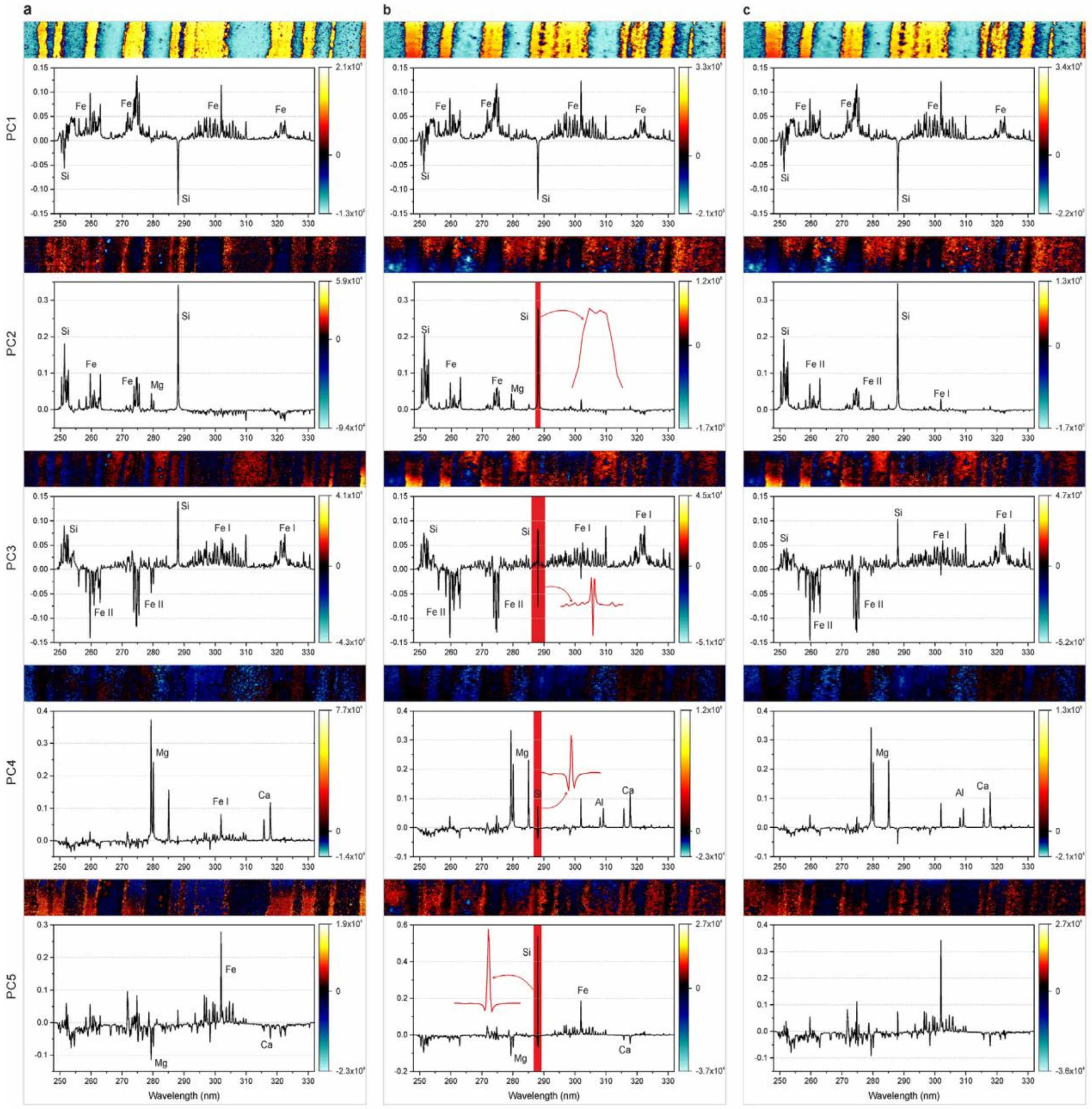


Fig. 6. Three principal component analysis calculated on, a) the first sample area (i.e. with no saturation), b) the raw data of the second sample area (i.e. with saturations) and c) the imputed data (i.e. corrected ones) of the same area.

locates this contribution on the periphery of zones A and B or on specific pixels scattered outside these zones. Another way to observe these differences is to compare the histograms of positive scores for this first component in the two conditions (SI Figure 7S). The saturation effect thus limits the range of scores values that should be observed and profoundly changes the structure of the distribution and therefore the visual perception one might have of it. For the second component, we are in much the same situation as before. We therefore have very comparable second principal components on the raw and imputed data. The Mg contribution is now anticorrelated to the Si, Fe and Al ones. On the other hand, once again there are differences on the scores maps for this component in the two conditions. Negative scores (blue color scale) are thus distributed more homogeneously in areas A and B when the spectral data are imputed. It is from the third principal components that we observe the largest spectral differences between the two conditions. Thus for raw data, typical W-shaped artifacts are observed (in red in Fig. 5) around the Mg contribution with correlations or anticorrelations with other elements. We observe on this occasion that the third and fourth principal components are extracted from raw data to express the saturation of pixels mainly located in the B zone of the sample. Even more specifically, we can see on the third principal component that the W-shaped artifact on Mg is positively correlated with another Ca contribution around 300 nm. This component thus just testifies to the simultaneous saturation of emission bands associated with the Mg and Ca elements on specific pixels according to the information given in SI Figure 6S. This example shows a very good example of spurious correlation created by the saturation phenomenon, which no longer exists once the data are corrected by imputation. Finally, the imputation strategy allows the appearance of dispersed particles opposing the Si and Al elements for the third principal component and the Ti et Al elements for the fourth one. It is obvious that such potentially less biased observations of particles represent a precious help for clinicians to diagnose the causes of the patient's exposure. From a general point of view, it is very interesting to see how a small percentage of saturated spectra can have an influence on a multivariate exploration method as sensitive as principal component analysis. This experience shows again here the necessity not to neglect the saturation phenomenon by setting up an adapted correction method such as imputation prior any chemometric analysis.

This last part of this work is dedicated to the analysis of the rock sample. As a reminder, two contiguous regions of the sample were analyzed considering two acquisition settings. In this way, we analyzed the sample by ensuring the absence of saturation for a first area but also its presence in the second one. Figure 8S shows the location of pixels containing saturations on the surface of the second sample area. It can be said that in this case saturation is omnipresent since it is observed on 32.5% of the analyzed surface. On the other hand, the same figure shows that this time these saturations are only found on the specific contribution of an element, namely silicon around 288 nm. Fig. 6 presents the three principal component analysis calculated on the first sample area (i.e. with no saturation), on the raw data of the second sample area (i.e. with saturations) and on the imputed data (i.e. corrected ones) of the same area. By comparing the principal components two by two in Fig. 6a and b, we observe very quickly the impact of saturation since we find the typical W-shaped artifact around 288 nm for components 3 and 4. The situation is even more critical for the fifth principal component with completely different profiles between Fig. 6a and b. In fact, it is above all the saturation effect that is expressed here for the second area of the sample. Finally, by comparing the results on the imputed data in Fig. 6c and the unsaturated data of the first sample area, we observe a perfect agreement between extracted profiles demonstrating the capacity of our approach to correct the saturated spectral data.

4. Conclusion

As we have seen in the work, it is crucial to consider the phenomenon of saturation present in the spectra. Through different datasets we have indeed shown that its presence quickly induces artifacts on spectral profiles but also on generated images when multivariate tools are used for their exploration. Make no mistake, even the presence of a limited percentage of saturated spectra in a given dataset can have an impact on the veracity of the chemometric results. It is obvious that it is absolutely necessary to avoid the presence of saturation in the acquired spectroscopic data whenever possible by modifying, for example, the sample preparation or the acquisition parameters. Unfortunately, there are many situations where this phenomenon is observed as in LIBS imaging and we have to find solutions to exploit these acquired data anyway. The usual column- or row-wise deletion is not a satisfactory solution because it can be accompanied by a large loss of spectral information in the dataset. As a consequence, we would have a partial or even biased view both at the spectral and spatial level of the sample. All the originality of our work was to consider the saturated signals as values that had not really been measured and by extension as missing values. The goal being to preserve all the spectral and spatial dimensions of the dataset, statistical imputation allowed us to retrieve complete data cubes consistent with the analytical reality of the samples considered as shown in the results. With this new approach, we will potentially have a chance to explore all those datasets that we think are being lost due to saturated signals.

Author contributions

The manuscript was written through contributions of all authors.

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.

Acknowledgment

The authors would like to acknowledge the French ANR research program for their financial support. In addition, the authors are grateful to B. Busser, V. Bonneterre and G. Panczer for helpful discussions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.aca.2021.338389.

References

[1] S. van Buuren, Flexible Imputation of Missing Data, Chapman and Hall/CRC, 2012, https://doi.org/10.1201/b11826.

[2] F. Scheuren, Multiple imputation: how it began and continues, Am. Statistician 59 (2005) 315–319, https://doi.org/10.1198/000313005X74016.

[3] Multiple imputation for nonresponse in surveys, in: D.B. Rubin (Ed.), Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 1987, https://doi.org/10.1002/9780470316696. Hoboken, NJ, USA.

[4] D.B. Rubin, Multiple imputation after 18+ years, J. Am. Stat. Assoc. 91 (1996) 473–489, https://doi.org/10.2307/2291635.

[5] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman and Hall/CRC, 1997, https://doi.org/10.1201/9780367803025.

[6] S. van Buuren, K. Groothuis-Oudshoorn, Mice: multivariate imputation by chained equations in R, J. Stat. Software 45 (2011), https://doi.org/10.18637/jss.v045.i03.

[7] 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.

[8] R. Kurucz, B. Bell, Atomic line data, atomic line data (R.L. Kurucz and B. Bell) Kurucz CD-ROM No. 23. Cambridge, mass, Smithsonian Astrophysical Observatory, 1995, http://adsabs.harvard.edu/abs/1995KurCD23K, 1995. accessed October 7, 2020, 23.

[9] V. Motto-Ros, S. Moncayo, F. Trichard, F. Pelascini, Investigation of signal extraction in the frame of laser induced breakdown spectroscopy imaging, Spectrochim. Acta B Atom Spectrosc. 155 (2019) 127–133, https://doi.org/10.1016/j.sab.2019.04.004.

[10] B. Busser, V. Bonneterre, L. Sancey, V. Motto-Ros, LIBS imaging is entering the clinic as a new diagnostic tool, Spectroscopy 35 (2020) 29–31.

[11] S. Moncayo, F. Trichard, B. Busser, M. Sabatier-Vincent, F. Pelascini, N. Pinel, I. Templier, J. Charles, L. Sancey, V. Motto-Ros, Multi-elemental imaging of paraffin-embedded human samples by laser-induced breakdown spectroscopy, Spectrochim. Acta B Atom Spectrosc. 133 (2017) 40–44, https://doi.org/10.1016/j.sab.2017.04.013.

[12] J.O. Cáceres, F. Pelascini, V. Motto-Ros, S. Moncayo, F. Trichard, G. Panczer, A. Marín-Roldán, J.A. Cruz, I. Coronado, J. Martín-Chivelet, Megapixel multi-elemental imaging by Laser-Induced Breakdown Spectroscopy, a technology with considerable potential for paleoclimate studies, Sci. Rep. 7 (2017) 1–11, https://doi.org/10.1038/s41598-017-05437-3.

3.2.2. Methodological perspective

Imputation is, as already mentioned, a field of statistics in which the selected data are used to generate new values that will replace the ones coming from, in this case, the saturated signals, that will be considered in a first step as missing data. In this way, the original dimensionality of the dataset will be maintained, meaning that no information will be missed at the end in the analysis. Nowadays, many different data imputation approaches are available. Despite that, the one that has been selected for this work is related to the 'multiple imputation', one of the best strategies to deal with incomplete data [214]. The main approaches for imputing multivariate data are the Joint Modeling (JM) and the Fully Conditional Specification (FCS), also known with the name of Multivariate Imputation by Chained Equations (MICE) [215]. Due to the constraints shown by the JM method, here MICE has been used to correct the saturation of the signals. This approach specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. It is also important to highlight the fact that the imputation model should account for the process that created the missing data, and preserve both the relations in the data and the uncertainty about these relations. The main concept behind MICE methodology is that multiple imputation is best done as a sequence of small steps, each of them requiring diagnostic checking. It is possible to briefly resume the main steps of multiple imputation into three parts, as reported in Fig. 20: imputation, analysis and pooling.

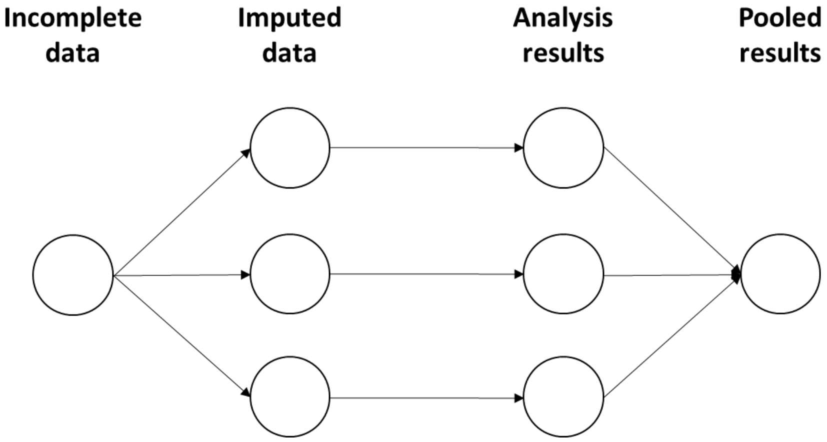


Fig. 20 – General scheme showing the main steps of the multiple imputation approach MICE.

As previously introduced, the data used at the starting point of MICE are the ones presenting the missing values (the incomplete data). The main point here is that clearly, it is not possible to estimate the missing values without making unrealistic assumptions about the unobserved data.

In the imputed data step, multiple imputation versions of the data will be used replacing the missing values by plausible data. Fundamental is that these values must be reasonable. This is the reason why they are drawn from a distribution specifically modelled for each missing entry. Naturally, due to the fact that the original value is missing, it is reasonable to consider that a degree of uncertainty is inevitable. This component leads to the fact that multiple values to impute will be generated. In the shown scheme for example, from one missing value, an amount of $t = 3$ different imputed data are generated. The three imputed sets are identical for the non-missing data entries, but differ in the imputed values. The higher the magnitude of the difference between the calculated values and the higher the uncertainty about what value to impute. Then, in a second step (the analysis results), each imputed dataset will be used to calculate the outcomes and estimate their robustness. It is important to highlight again that the differences among the various estimations found with the MICE algorithm are caused due to the uncertainty about which value to impute. Normally, the analysis results are at the end collectively stored as a multiply imputed repeated analysis. Finally, in the last step, the $t = 3$ estimates obtained with MICE are pooled together into a single value, and its variance (within- and between-imputation variance) is calculated. In other words, a regression model is developed, and the results of the function are stored as a multiple imputed pooled outcomes object. Once the model is established, it can be used to predict missing values of spectra at this specific spectral variable based on known values in the matrix. As a consequence of this, at the end a dataset free of missing values will be generated, which will be used for the multivariate analysis, as usual.

3.2.3. Conclusions and future perspectives

Saturation of the signals is clearly a very important spectroscopic problem that should be faced before any data analysis. The challenge is that it is not easy to find a solution when the value to be replaced is not available within the matrix, or at least not directly related to the given information, such as in the case of clipped bands. Despite that, solving this problem is mandatory, due to the fact that using a strategy such as the exclusion of these values can result the worst approach, leading for example to the loss of important information. The presented method, based on the use of a multivariate data imputation, more precisely the MICE algorithm, represents an interesting approach that can be used to overcome the limitations related to the standard methods. Nevertheless, this kind of strategy can show a certain degree of drawbacks. First of all, MICE has been tested on LIBS spectra, which are represented by very narrow peaks and bands, each of them potentially selective to specific elements. It means that if for example a peak is saturated, others will be anyway potentially correlated to the same element, leading to the chance of finding further signals that could make it possible to build a model able to impute the missing values. Also important is that LIBS can easily generate from thousands to millions of spectra, increasing the possible number of values to be used to create a reasonably robust and accurate MICE prediction model. The situation can be different when other spectroscopies are used, which are related to other characteristics. For example, Raman spectroscopy, in which broadened bands can be observed, can hardly collect the same amount of data of LIBS analysis. In addition, despite the use of fingerprints to recognize different molecules, it is off the table the fact that Raman and LIBS spectra cannot be compared, from an interpretation point of view. Therefore, it is clear that when LIBS is used, it is easier to create a regression model that makes it possible to predict missing values. Nevertheless, it would be interesting to implement the use of imputation in order to generate model robust enough to be used for other spectroscopies. Lastly, it is important to highlight another limitation related to this method. In fact, no matter the computational calculations, if a saturated signal is related to an element (or compound) that is that pure to be completely absent from the rest of the matrix, naturally it will be impossible to generate a model able to impute that value, leading to the creation of artifacts that would correspond to wrong data analysis outcomes. Nevertheless, this is a very extreme scenario. Another extension of this work would be to use the same concept to specifically correct the self-absorption phenomena often observed in LIBS. However, in this case a first difficult has to be faced. In fact, it would be necessary at first to detect automatically the zones presenting this phenomenon before their correction by an imputation stage. To conclude, it can be said that from a general point of view using this approach, finally it is possible to use the totality of the information related to a dataset, both the spectral and spatial components, leading to the potential possibility of exploring datasets that otherwise would not be analyzed due to their quantity of saturated signal.

CHAPTER 4

4. LIBS IMAGING AND CHEMOMETRICS: HOW TO EXPLOIT MULTIVARIATE DATA ANALYSIS TO MAXIMIZE THE OBTAINABLE RESULTS FOR THIS SPECTROSCOPY

4.1. A general overview of LIBS imaging framework and possible chemometric approaches

At this point, it is clear that an important part of the present PhD has been dedicated to the investigation of chemometric tools and methodologies linked in particular to LIBS analysis. This imaging spectroscopy, rapidly developed in the last decade, is related to very interesting characteristics. Due to this, it is nowadays considered an essential instrumentation in many research areas. From a general point of view, it is very likely to obtain images made of millions of pixels associated to thousands of spectral channels. Of course, this means that very interesting images can be generated, but at the same time, despite the general development from the spectroscopic point of view and especially, from the computational instrumentation perspective, it is still really complicated to find a reasonable way to deal with and investigate this huge amount of produced data. The typical routine approach used to study this kind of data cube is based on the integration of the acquired signal at a particular wavelength (i.e., an emission line of a given element), leading to the generation of a distribution image of the considered element present in the sample. Naturally, this can be a limitation. In fact, first of all the operator needs to have a general idea of the various elements present in the sample of interest in order to obtain the corresponding images. Then it is clear that, particularly in a scenario in which a big dataset is acquired, different elements will be present as minor compounds or even traces, maybe showing very small intensity signals compared with the rest of the compounds in the specimen. All these things correspond to the possibility of losing very important information related to the sample, and so to carry out an incomplete investigation of its real heterogeneity. It is also very difficult to really detect correlations between elements with this classical integration method. Due to all these reasons, it is obvious that chemometrics might be used as an important alternative to the routine approaches in order to study in deep the outcomes related to this kind of spectroscopy and so, lead to new ways to interpret the results. In order to do this, the main chemometric tool used to show and prove the robustness of the present work is MCR-ALS [129,183]. In fact, this algorithm is nowadays one of the most important milestones in the spectral unmixing framework, a technique used with the purpose of selecting and showing the pure contribution related to the different components of a given investigated sample. Nevertheless, it is important to highlight the fact that also using chemometrics, and so MCR-ALS, it could be challenging to use the raw data, due to the huge amount of information related to the acquired sample. This is true from two different points of view. First, again, it is a very easy scenario, when the used matrix corresponds to a big dataset, the one of losing information, particularly related to minor compounds and trace elements. Second, despite the very interesting characteristics of MCR-ALS, some calculation problems may arise, due to the size of the data cube. So, as a first step, it is always important to consider the possibility of selecting only a part of the total information (the most relevant one), and work exclusively with that. In other words, it is very plausible the idea that in millions of spectra, only a small percentage will be related to very pure information. This means that, with the right approach, it would be possible to enormously reduce the quantity of data to be used to obtain final accurate outcomes. From a certain point of view, this kind of idea has been already previously discussed in this manuscript. Clearly, we are referring to the randomised SIMPLISMA [180] and the Embedded K-Means (EKM) [212] approaches. So, which is the reason of developing a new strategy for this kind of spectroscopy? As previously explained, randomised SIMPLISMA is a very useful method, and it can be applied to select the rank and generate the initial estimates to be used in the MCR-ALS approach. Nevertheless, if the investigated image is too big and complex, it might be a real challenge the selection of the right randomised SIMPLISMA input values (i.e., the number of subsets to be generated and the percentage of pixels to be selected for each subset). Another important thing is that, also if it would be possible to obtain good results, anyway MCR-ALS could be not useable due to the dimensions of the dataset from a computational point of view. Therefore, it is fundamental to find an alternative way in which MCR-ALS can be applied without using the totality of the data acquired by LIBS. Considering EKM clustering, as previously stated it is a very interesting tool to deal with a huge amount of data, in order to find the contribution of not only the major, but also the minor compounds, when it is impossible to have an adequate knowledge of the investigated sample. Nevertheless, it is important to remember the fact that this kind of approach uses the totality of the pixels of the matrix. Therefore, the presented methods have both some very interesting aspects, but naturally also some constraints. In other words, it would be important to find an alternative method that, in situations in which the sample is very big and heterogeneous, could drive the operator with an automatic approach to accurate results. The core of this chapter is to provide a data analysis pipeline capable to drastically decrease the amount of imaging data (both the spectral channels and the pixels) used for the investigation of a sample in order to perform a simpler unmixing analysis exploiting only the essential information selected from the LIBS image. For informational purposes, it is important here to inform the reader that a first part of this chapter, also published in Analytica Chimica Acta, Volume 1192 (2022) [216], will focus on this aspect, and an interesting data analysis pipeline will be described to reach this goal.

4.2. LIBS data fusion: the importance of fusing different spectroscopic techniques

Another fundamental part of this chapter regards the use of chemometrics in the framework of the data fusion of different hyperspectral imaging systems, focusing particularly in the use of LIBS spectra and other spectroscopic responses. As previously stated, LIBS is an elemental spectroscopy. From an analytical point of view, it is very important to study a given sample from the elemental perspective, to deepen the comprehension of its chemical composition. Despite this, the information related to the elemental point of view can be not sufficient, and it is always interesting to merge together different kind of data to obtain a more complete overview of the characteristics of the investigated matrix. On the other hand, sometimes it can be very challenging to give the right interpretation to the data and understand the molecular composition of a given component (if for example an exhaustive library for a particular spectroscopy/technique is not available). In this kind of scenario, an elemental analysis such as the one obtainable with LIBS can be decisive. In fact, using the elemental information, it would be for example possible to carry out details from the elemental and also molecular points of view that are not clear when LIBS is not involved in the analysis. Furthermore, another important aspect is the LIBS resolution. Making the most of it, it may be possible to obtain better spatial details coming from other spectroscopies. Lastly, LIBS can be directly coupled with other spectroscopic responses. It means that it is possible to acquire at the same time different spectral domains, without any necessity of changing the used instrument and platform. In this way, it is possible to obtain more data cubes that are represented by the same spatial dimensions, leading to a faster and easier data fusion between the considered spectroscopies. In detail, during this PhD it has been possible to investigate two different techniques related to LIBS. The first one is PIL [91], a luminescence effect that can be generated in particular situations. As previously described, the general principle of LIBS is the use of a laser that will produce a plasma able to ablate the surface of the sample. Some elements can keep the excess of energy coming from the LIBS excitation source and then release it with a delay of some milliseconds in the form of a luminescence effect. Despite the relative simplicity in acquiring these additional PIL spectra, the interpretation of such signals remains uncertain [90]. This is a perfect scenario in which the fusion can be used with the purpose of leading to an easier explanation of the data, due to the absence of a real library for PIL phenomena. These results have also been reported in the aforementioned paper published in Analytica Chimica Acta, Volume 1192 (2022) [216]. On the other hand, another interesting example regarding the data fusion is the combination of LIBS with Raman spectroscopy. Naturally, here the scenario is very different. If PIL is a technique that from a chemical point of view shows very limited information, Raman is related to very important aspects of the molecular composition of the studied sample. Nevertheless, it can be sometimes complicated to give the right interpretation to some Raman bands, as well as it can result challenging to recognize some LIBS spectra, if the composition of the sample is too heterogeneous, or if the intensity of the corresponding signal is too weak. In addition, it is plausible that some information may be related only to one or the other instrument, if not coupled. Clearly, fusing the two techniques, it would be possible in this way to extract more details, correlations and anticorrelations between these different spectroscopic responses, obtaining both an elemental and molecular investigation of a given sample. In conclusion, without any doubt a fusion strategy is a very interesting approach that can be used in order to deepen the knowledge of the chemical composition of a specimen. The limitation in this scenario is anew represented by the fact that an enormous quantity of data can be easily acquired (considering in this case the generation of millions of spectra for both LIBS and the supplementary coupled spectroscopy). Again, the approach proposed in this chapter based on the selection of only the most relevant spectral and spatial information before the data analysis can represent a good solution to this kind of problem. Even so, it is important to consider some aspects in order to avoid inaccurate outcomes. First of all, in the moment that two different datasets are fused together, the right normalization has to be applied, in order to obtain comparable spectral magnitudes. Second point, it is fundamental to apply the aforementioned data reduction approach at the right moment of the pipeline. This aspect is much related to the kind of spectroscopic response coupled with LIBS. Here is explained the reason. Due to the few chemical information linked to PIL, the use of the data reduction after the fusion has to be preferred. This is due to the fact that PIL spectra alone are not very easily interpretable, as a consequence of their corresponding broad signals. Instead, applying the reduction of the data after the fusion, the selected information will depend on both the spectroscopies, leading to the observation of the correlations between the various elements identified by the use of LIBS and their luminescence effect, linked to PIL phenomenon. Contrariwise, using Raman spectroscopy, it is recommended to merge the datasets after the selection of the spectra used on the separated matrices. The reason of using this approach is related to the very specific and different details linked to LIBS and Raman spectra (respectively, elemental and molecular information of the acquired matrix). Therefore, in a first step it is necessary to skim the data, in order to select only their most relevant part. Then, it is possible to fuse the two reduced datasets and observe the interaction between LIBS and Raman spectral responses, in order to maximize the obtainable information. Due to the complexity of material explained in the present chapter, only some results concerning the Raman spectroscopy coupled with LIBS will be shown, while the PIL investigation is reported in detail into the already cited publication [216].

4.3. Data fusion of LIBS and PIL hyperspectral imaging: Understanding the luminescence phenomenon of a complex mineral sample

4.3.1. General aspects related to the data reduction in LIBS analysis

As described in the introduction of this chapter, one of the main points of the present work has been the development of a strategy able to select the most important information related to LIBS, and not only this spectroscopy. One of the principal limitations in LIBS is that an enormous amount of data can be easily generated. In fact, millions of spectra can be acquired in a reasonable time, due to the very interesting instrumental characteristics of this spectroscopy. Despite this, it is important to consider some challenges. For example, the raw data can be hardly analyzed obtaining good outcomes, if not correctly treated. Therefore, find an adequate pipeline able to reduce the amount of used data, but at the same time be sure to consider the most important information, no matter if related to major compounds (easily identifiable) or, and particularly, minor components and traces (represented by a small quantity of spectra) is a mandatory task. The main problem is that normally the identification of the most relevant information is related to factors such as the total explained variance. This means that, in the case in which a big dataset is analyzed, minor compounds will be usually represented by small values, and for this reason easily skipped. Nevertheless, other algorithms can be used in order to select the information based not on the total explained variance, but on the purity of the spectra. This scenario is clearly recommendable, in order to generate better results. One of the most common used techniques for this purpose is SIMPLISMA [192], as it has been vastly discussed into this manuscript. Again, one of the main limitations concerning this algorithm is related to the fact that, in order to be applied, some inputs have to be insert by the operator, with the purpose of selecting the right number of pure components to be used, for example in the framework of the spectral unmixing.

In a different way, here SIMPLISMA has been used to select the purest spectra, but with the intention of only skimming the total information, in order to finally work with a reduced dataset. In this way, as a main point, instead of working with millions of spectra, only a small percentage will be taken into consideration for the chemometric approach (in this case, MCR-ALS). Second, and more importantly, the purpose of using the presented pipeline is the selection of not only a small quantity of spectra, but the most interesting and purest ones, in order to be sure of obtaining at the end results that will be representative of the heterogeneity of the matrix and so, of the existence of eventual traces, no matter its complexity, that otherwise would be easily missed. In order to give a general idea about this procedure to the reader, it is important to stress again the fact that this part of the chapter focuses only on the description of the used approach for the selection of the most important pixels and variables, while the part related to the chemometric interpretation of the data and the fusion with PIL spectra is well described into the reported published paper [216]. Also important is to understand the use of SIMPLISMA in this kind of scenario. As introduced, this algorithm is normally applied in the spectral unmixing framework in order to select a precise number of pure contributions that are used as initial estimates in the, for example, MCR-ALS procedure. In the present approach, its use is slightly different. In fact, SIMPLISMA is firstly applied on the spectra of LIBS, in order to select only some of the variables, the purest in the whole spectral domain. This is possible due to the very fine spectral features of this spectroscopy. In addition, one should consider the fact that multiple peaks can refer to the same element. Using this approach, it is then possible to select only a part of them, the ones that are stronger related to a certain information. Contrariwise, considering PIL and Raman responses, another spectral reduction procedure has been applied. Regarding PIL, due to the fact that only two big band signals are available, the rest of the spectral range that is related to the baseline has been removed prior the analysis. Instead, for Raman spectra, considering the fact that generally the peaks are broadened, there was selected only one point each three, in order to reduce the total amount of spectral variables. This approach could be considered as an attempt that would lead to inaccurate outcomes, due to the possible lack of information. This is not true. In fact, MCR-ALS results are calculated using only the selected information. Then, as a last part of the procedure, a suitable single least-squares step is applied, in order to use the solutions obtained with MCR-ALS to reconstruct the full spectral signatures. Lastly, considering again the selection procedure, once that only a part of the spectral range has been selected, SIMPLISMA is newly applied, this time in order to reduce the total final number of used pixels for the spectral unmixing, considering the fact that only few of them will be related to a pure information coming from a specific element or compound. Naturally, the aforementioned least-squares procedure is at the end applied also on the pixels, in order to reobtain the full distribution maps of the initial data cube used in the analysis. As a last, but essential point, it is also important to highlight another aspect. Using this SIMPLISMA-based procedure, it is not possible a priori to know the number of pure variables and pixels to be selected in order to reduce the total amount of data, but at the same time to have at the end the certainty of conserving also the minor information. This is the reason why normally the selected value is an overestimation of the possible real one. In this way on one side the idea of enormously reducing the total amount of data will be anyway performed, and on the other, the operator can be relieved of the fact that any information, no matter if coming from major or minor compounds, will not be missed.

Analytica Chimica Acta 1192 (2022) 339368

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Data fusion of LIBS and PIL hyperspectral imaging: Understanding the luminescence phenomenon of a complex mineral sample

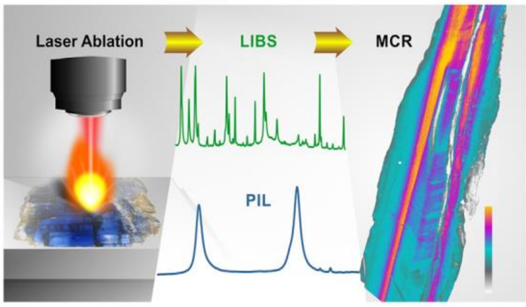
Alessandro Nardecchia [a], Anna de Juan [b], Vincent Motto-Ros [c], Michael Gaft [d], Ludovic Duponchel [a, *]

[a] Univ. Lille, CNRS, UMR 8516, LASIRE, Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille, F-59000, France
[b] Chemometrics Group, Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, Diagonal 645, 08028, Barcelona, Spain
[c] Institut Lumière Matière, UMR 5306, Université Lyon 1, CNRS, Université de Lyon, Villeurbanne, 69622, France
[d] Ariel University, Department of Physics, Ariel, 40700, Israel

HIGHLIGHTS

- A new data fusion strategy to manage big hyperspectral data sets.
- A data compression approach to keep relevant chemical information.
- Better understanding of the luminescence phenomenon thanks to LIBS/PIL fusion.
- The first simultaneous use of PIL and LIBS imaging to characterize complex samples.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 October 2021
Received in revised form
7 December 2021
Accepted 8 December 2021
Available online xxx

Keywords:
Hyperspectral imaging
LIBS
PIL
Data fusion
Chemometrics

ABSTRACT

Laser-induced breakdown spectroscopy (LIBS) imaging is an innovative technique that associates the valuable atomic, ionic and molecular emission signals of the parent spectroscopy with spatial information. LIBS works using a powerful pulse laser as excitation source, to generate a plasma exhibiting emission lines of atoms, ions and molecules present in the ablated matter. The advantages of LIBS imaging are potential high sensitivity (in the order of ppm), easy sample preparation, fast acquisition rate (up to 1 kHz) and μm scale spatial resolution (weight of the ablated material in the order of ng). Despite these positive aspects, LIBS imaging easily provides datasets consisting of several million spectra, each containing several thousand spectral channels. Under these conditions, the current chemometric analyses of the raw data are still possible, but require too high computing resources. Therefore, the aim of this work is to propose a data compression strategy oriented to keep the most relevant spectral channel and pixel information to facilitate, fast and reliable signal unmixing for an exhaustive exploration of complex samples. This strategy will apply not only to the context of LIBS image analysis, but to the fusion of LIBS with other imaging technologies, a scenario where the data compression step becomes even more mandatory. The data fusion strategy will be applied to the analysis of a heterogeneous kyanite mineral sample containing several trace elements by LIBS imaging associated with plasma induced luminescence (PIL) imaging, these two signals being acquired simultaneously by the same microscope. The association of compression and spectral data fusion will allow extracting the compounds in the mineral sample associated with a fused LIBS/PIL fingerprint. This LIBS/PIL association will be essential to interpret the PIL

* Corresponding author.
E-mail address: ludovic.duponchel@univ-lille.fr (L. Duponchel).

https://doi.org/10.1016/j.aca.2021.339368

spectral information, which is nowadays very complex due to the natural overlapped signals provided by this technique.

1. Introduction

Laser-induced breakdown spectroscopy (LIBS) imaging is nowadays a very powerful technique for the elemental analysis of complex samples used in many different scientific fields [1–7]. This technique uses a pulse laser beam focused on the sample surface to generate a plasma that atomizes and excites the ablated matter. As a consequence, the excited atoms, ions and molecules release the excess of energy with electronic relaxations, and a characteristic emission spectrum for each element present in the matrix can be acquired using an optical microscope coupled with a spectrometer. In LIBS imaging, the sample surface is usually explored in a scanning configuration mode, acquiring one spectrum at a time for each spatial position of a predefined grid. Then, using a classical integration of the acquired signal at a particular wavelength (i.e. an emission line of a given element), it is possible to generate a distribution image of the considered element present in the sample. LIBS technique shows many advantages, such as multi-elemental capabilities including light elements (<Mg), a high acquisition rate (up to 1000 spectra/s), high sensitivity most of the time, high dynamic range (major elements to traces can be observed), and compatibility with optical microscopy. Nevertheless, even if the high acquisition rate of LIBS imaging allows analyzing large sample areas of several cm^2 in a very reasonable time, this advantage becomes a major limitation because a huge amount of data is naturally produced due to both the many spectral channels explored by LIBS and the massive number of sampling points – the pixels – scanned. In fact, it is nowadays common to get images with millions of pixels associated with thousands of spectral channels [8,9]. Another important aspect in the LIBS exploration of a sample is the possibility to obtain an additional plasma induced luminescence (PIL) [10] response using the same instrument. Indeed, the plasma generated by the LIBS laser shot acts as an excitation source and produces the emission of a luminescence response for specific elements present on the sample surface [11]. Nevertheless, despite the relative simplicity of acquiring these additional PIL spectra, the interpretation of such signals remains uncertain [12]. Chemometrics and multivariate data analysis are very suitable approaches for the exploration of this complex kind of imaging datasets. However, the use of these tools for the study of LIBS and/or PIL images is nowadays still limited. Understanding the concept of hyperspectral imaging, finding appropriate tools for data exploration to deal with millions of spectra and able to provide interpretable outputs is still a very complex task, which can be of invaluable help for the LIBS community members.

The central point of this work is to provide a data analysis pipeline capable to drastically decrease the amount of imaging data (both the spectral channels and the pixels) used for the investigation of a complex and heterogeneous sample in order to perform a simpler unmixing analysis of the essential information selected for LIBS images or for fused LIBS/PIL data configurations. To do the unmixing task, Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) analysis [13–15] will be applied on the selected small amount of data coming from the previous compression step. We will demonstrate that applying MCR-ALS on such compressed dataset is sufficient to reconstruct high quality full maps and spectral signatures of the compounds in the imaged sample without losing the initial spectral and spatial resolution [16]. The methodology proposed will be tested to study a heterogeneous kyanite mineral sample containing several trace elements analyzed by LIBS and PIL imaging. The results of the analysis of fused LIBS and PIL datasets will provide the identification and distribution of the different elements present in the sample but, most importantly, will shed light for a better understanding of the luminescence phenomenon in this kind of complex samples. To the best of our knowledge, this is the first time that this data analysis pipeline (data compression and fusion) is used on LIBS/PIL imaging platforms.

2. Material and methods

2.1. Experimental setup and spectral data acquisition

The LIBS experimental setup has been already described elsewhere [3,11,17,18]. It included a Nd:YAG laser source operating at 100 Hz and emitting at the fundamental wavelength (i.e. 1064 nm) with an 8 ns pulse duration (Centurion, Quantel laser by Lumibird). The laser beam was focused onto the sample using a 15 × magnification objective as shown in Fig. 1a. All the measurements were conducted in ambient atmosphere with an argon flow of 0.8 l/min acting on the plasma region. A laser line scanning was performed in raster scan mode with the use of a motorized XYZ stage. In this configuration and considering the laser frequency rate, about 360,000 laser shots were produced in 1 h. The ablation craters, observed afterward with optical microscopy, were less than 8 μm in diameter. Two spectrometers (Shamrock 500 and Shamrock 303, Andor Technology) equipped with intensified charge-coupled device (ICCD) cameras (iStar, Andor Technology) were used to probe simultaneously two spectral ranges in different temporal domains. The Shamrock 500 was used for LIBS experiments and was equipped with a 2400 l/mm grating (Holographic, peak at 220 nm) covering the 425–440 nm spectral range with a resolution of ~0.04 nm. This range was selected to detect primarily iron (Fe), chromium (Cr), vanadium (V) and titanium (Ti), although calcium (Ca) and zirconium (Zr) lines could also be detected (as can be seen on the mean LIBS spectrum in Fig. 1b). The Shamrock 303 was used for PIL experiments. It was equipped with a 1200 l/mm grating and setup in the 680–720 nm spectral range, where intense luminescence lines were detected. The mean PIL spectrum is presented in Fig. 1c. Both ICCD cameras were synchronized to the Q-switch of the laser. The LIBS acquisition was performed with a delay of 1200 ns and a gate of 4000 ns, while the PIL acquisition was performed with a delay of 100 μs and a gate of 2.2 ms. The light emitted by the plasma was collected by two quartz lenses and focused onto the entrance of round-to-linear fiber bundles connected to each spectrometer. Each fiber bundle was formed by 19 fibers, each with a 200 μm core diameter. Spectra were acquired in full vertical binning mode for the two spectrometers. The laser energy was stabilized throughout the experiment and was fixed to 1 mJ per pulse. Finally, a homemade software developed under the LabVIEW environment was used to control the entire setup, allowing automatic sequences of any selected regions of interest with a preset lateral resolution.

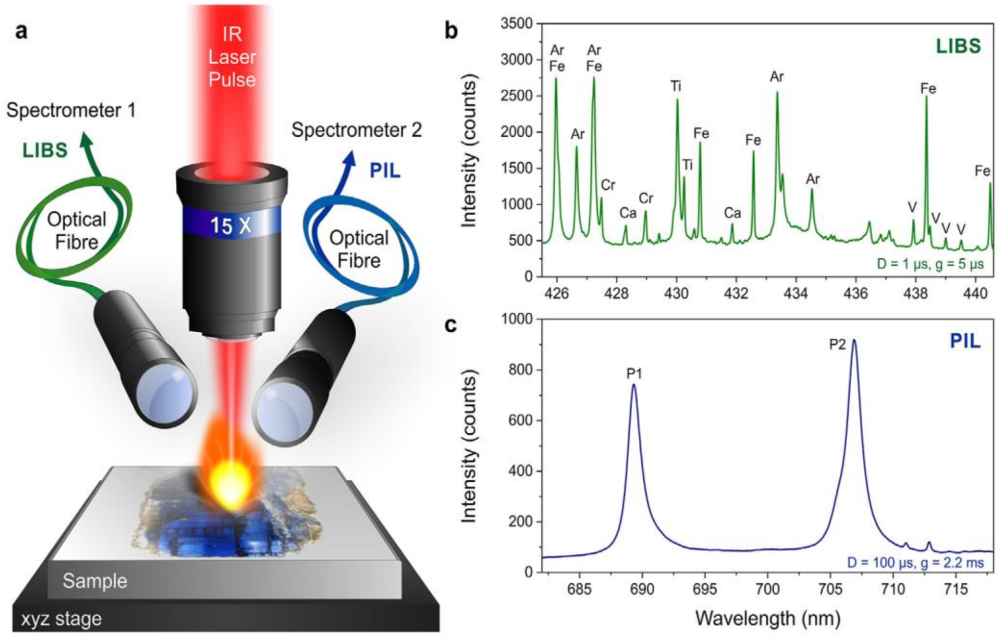


Fig. 1. a) Experimental setup. b) Mean LIBS spectrum of the considered sample. c) Mean PIL spectrum of the considered sample.

2.2. Sample and dataset description

The sample selected for this study is a section of blue kyanite cristal (also called disten or cyanite) approximately 3 by 1.5 cm in size, which isa low temperature - high pressure metamorphic phase mainly formed by Al_2SiO_5 with many heterogeneities and several trace elements (mainly iron, calcium, vanadium, titanium and chromium), collected in Siberia. For LIBS and PIL imaging, the characterized cross-section was embedded in epoxy resin, cut, and finally polished with SiC paper under water to obtain a clean flat surface ready to be scan. The LIBS and PIL images acquired from this sample are sized each 1100 × 2000 pixels (i.e. a total of 2,200,000 spectra) x 2048 spectral channels with a spatial resolution of 20 μm per pixel. The size occupied in terms of storage by the two datasets is equal to 8 gigabytes for the LIBS dataset and more than 6 gigabytes for the PIL one.

A first idea of the chemical information related to this sample can be obtained observing the mean spectra of the two datasets using LIBS and PIL (Fig. 1b and c respectively). Thus the observation of the mean PIL spectrum shows a first line near 706 nm that looks not symmetric and evidently has a shoulder at its left side. In fact, under UV excitation, the literature indicates that two close lines with similar intensities can be observed depending on the orientation of the sample at 706.2 and 704.6 nm respectively. This statement is quite surprising in our case because the line at 704.6 nm is supposed to have a much shorter decay than the one at 706.2 nm (75 μs and 1.2 ms, respectively) and therefore our acquisition parameters should not allow us to see the former. On the other hand, the line near 689 nm seems to be symmetrical and also corresponds to UV excitation. Once again, the literature indicates the presence of two lines this time even closer at 688.9 and 690.1 nm, the second being of very low intensity which finally explains the observation of a single line at first sight. Comparing the two mean spectra, it is clear that LIBS is represented by a larger number of emission signals compared to PIL, which is mainly formed by two broad spectral contributions. Fig. 2 shows distribution images of elements present in the sample obtained from the signal integration method classically used on specific wavelengths of the LIBS spectra. Thus, even if this approach takes only into account individual wavelengths and not the full spectra, the spatial distribution of the elements presents some interesting aspects. For instance, while elements Cr and V seem to be strongly correlated and distributed in a very large area, Fe, Ca and Ti are more specific to small zones of the mineral. Note that no distribution image of Al and Si are proposed, although kyanite is an aluminosilicate because this element does not present LIBS emission in the considered spectral range. The two PIL images generated from the two spectral contributions at 689.315 (P1) and 706.865 nm (P2) also show slight spatial locations. However, from a general point of view, it can be seen that it is rather difficult to find spatial correlations between all the LIBS and PIL images by simple visual inspection. The only thing clearly seen is that the luminescence at 689 and 706 nm does not come from the Ti element since we observe some inverse correlation among LIBS and PIL images. Even if the generation of these integration images remains an easy first step to observe the chemical contributions of the sample, it is obvious that the characterization of the correlations between images remains delicate. Unfortunately, this is not the only constraint since the information of elements that can coexist to form mineral phases is lost or, in the best of the cases, incomplete using this univariate approach. Similarly, minor compounds associated with weak signal and small localized areas may be missed. For all these reasons, we propose in this work a multivariate data processing pipeline combining compression, fusion, and signal unmixing for an exhaustive and simultaneous exploration of LIBS and PIL spectroscopies.

3. Data treatment

3.1. Data compression and signal unmixing

The massive nature of the datasets to be analyzed demands a mandatory step of data compression to save computational resources and analysis time. The whole process of compression and multivariate resolution (signal unmixing) is described below in several successive steps. For convenience, this section of the text will illustrate the application of the methodology to the LIBS dataset, due to the complexity and rich information provided by this measurement. However, the same procedure was applied to both LIBS and PIL datasets, except for some specificities, that will be described, discussed and shown in Fig. 3.

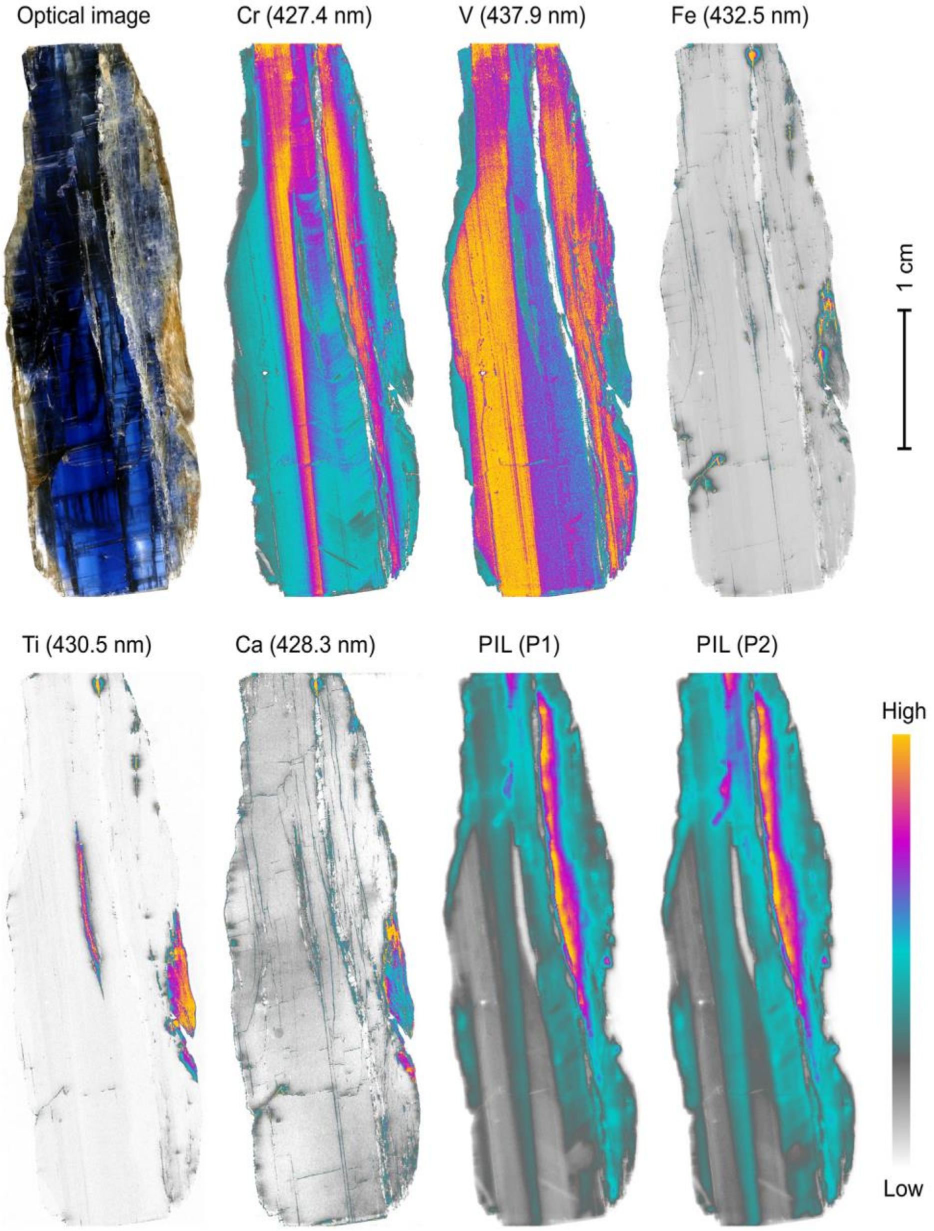


Fig. 2. Distribution images of elements generated from the classical signal integration method from LIBS and PIL spectra.

Thus the proposed data processing pipeline is divided into the following steps:

1) *Image blocking*: first of all, the whole dataset was divided into 16 subimages. In this way, every single subimage had a reduced size of 780 × 125 pixels and provides an unfolded submatrix $\mathbf{D_i}$ of 97500 spectra. Blocking allows performing data analysis tasks much faster in each $\mathbf{D_i}$ subset of 97500 spectra than if the work was carried out in the initial **D** matrix of 2 million spectra of the full image. Each subsequent step described was applied separately to each of the $\mathbf{D_i}$ submatrices (Fig. 3a).
2) *Spectral and spatial preprocessing*: heterogeneous samples analyzed by LIBS imaging often contain saturated spectra in the acquired datasets. Recent works have shown the importance of correcting these signals prior to any data processing by proposing effective but relatively complex correction strategies [19,20]. In this work, we have proposed a simple procedure for managing the saturated spectra. First, the use of a threshold on the spectra (set by the maximum A/D converter dynamic range) was used to identify and locate the saturated pixels in the image spectra. Second, each saturated spectrum was replaced by the average of the non-saturated neighboring pixels in the image. For the LIBS dataset, a baseline correction was afterward applied using the asymmetric least squares (AsLS) algorithm based on the Whittaker smoother [21], with $\lambda = 104$ and an asymmetry parameter of 0.0003. For the PIL dataset, which had a radically lower signal-to-noise ratio, it was first necessary to apply a Savitzky-Golay smoothing (filter width = 20; polynomial order = 2) [22], and then, again, the AsLS algorithm ($\lambda = 10^7$ and asymmetry parameter of 0.0003) for baseline correction. Finally, the use of a threshold on the global spectral intensity of the pixels helped to generate a mask to separate the spectra from the mineral sample, used for further analysis, from the spectra of the epoxy resin surrounding it, discarded in all the following data treatments steps. This image cropping step reduced the total amount of spectra to be analyzed from more than 2 million to around 1 million. The comparison between the starting raw spectra, and the ones with this first spectral and spatial pretreatment are represented in the supplementary material (Fig. S1).

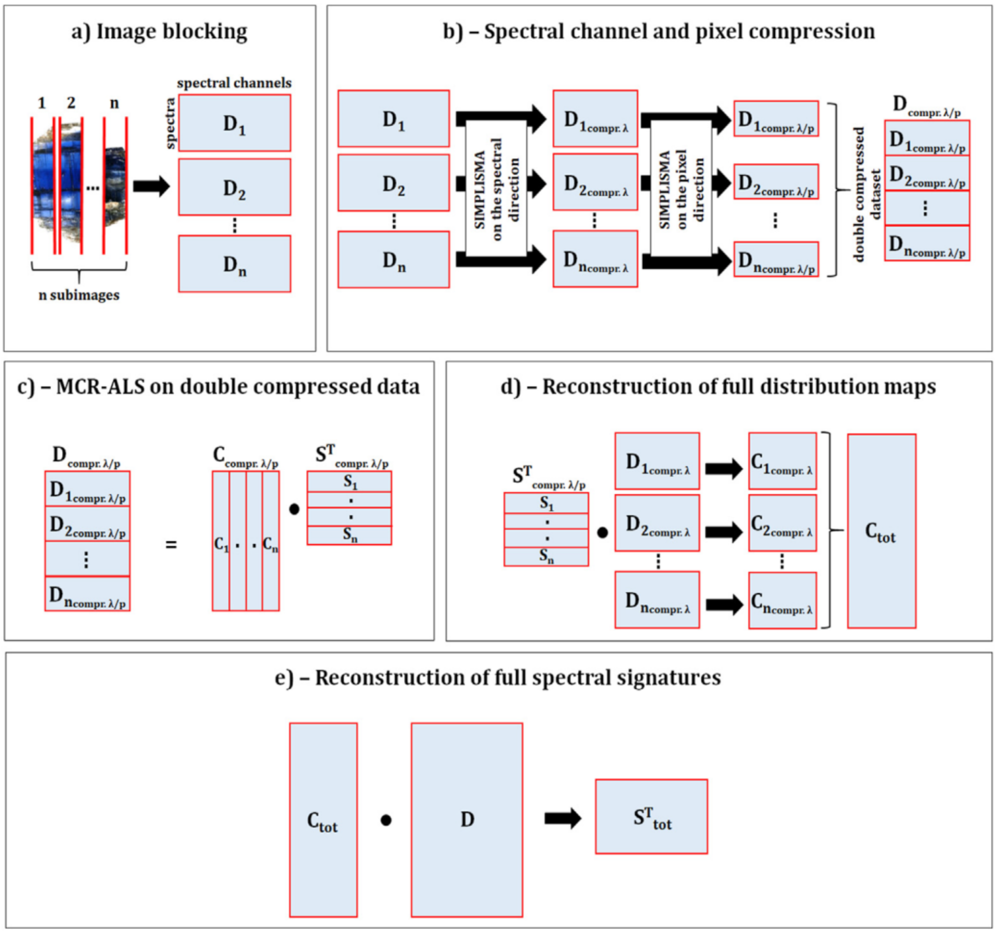


Fig. 3. The data fusion and signal unmixing procedure.

3) *Double data compression*: this is a key central step in all the proposed chemometric strategy. In fact, despite the already performed signal corrections and the massive reduction of spectra, the dimensions of the datasets (both LIBS and PIL) were still very huge. In chemometrics, and particularly in the signal unmixing framework, it is common to use methods oriented to the purest selection of variables (understood as pixels or spectral channels) to generate initial estimates for iterative unmixing methods, such as MCR-ALS. In the present work, a SIMPLe-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) [23–26] based-method was first used to select only the purest image information and drastically compress the number of selected pixels and spectral channels prior to the final MCR-ALS analysis. For each $\mathbf{D_i}$ image submatrix in Fig. 3b, taking advantage of the fine spectral features of LIBS, the selection of purest information was applied first on the spectral channels. In this way, only the most important spectral variables (i.e. wavelengths) related to different chemical elements in the mineral were selected, discarding the redundant information. It is also important to underline the fact that at this stage of the procedure, as the final number of purest spectral variables needed is unknown, it was decided to overestimate this value. Thus, each $\mathbf{D_i}$ image submatrix was first divided into blocks of 200 spectra from which the first 20 purest spectral variables were selected. At the end, a small list of spectral channels considering all the selected channels in the 16 $\mathbf{D_i}$ submatrices was used to generate the spectral-compressed $\mathbf{D_{i,compr.\lambda}}$ image submatrices, which had a much lower number of spectral channels than the original $\mathbf{D_i}$ blocks, but had all image pixels. In a next step, the same approach based on SIMPLISMA was used in each of the spectral-compressed $\mathbf{D_{i,compr.\lambda}}$ image submatrices to select the purest pixels. So, each of the $\mathbf{D_{i,compr.\lambda}}$ image submatrices was divided into blocks of 500 spectra, and the first 40 purest pixels of each block were selected, providing a spectral- and pixel-compressed $\mathbf{D_{i,compr.\ \lambda/p}}$ submatrix with a much lower number of pixels and spectral channels than the initial $\mathbf{D_i}$ related block. The extracted information of all the $\mathbf{D_{i,compr.\ \lambda/p}}$ submatrices was fused together in order to create a final, compressed, $\mathbf{D_{compr.\ \lambda/p}}$ dataset with selected information from the full initial image (Fig. 3b). The sequential use of a purest variable selection method on small blocks of the initial image not only helps in speeding up the data analysis process but, most importantly, ensures that even minor compounds present in local zones of the raw dataset will be kept in the image compressed version. At this point it is important to note that the selection of information is driven by the difference in spectral and spatial features and not by the percentage of variance expressed by the different compounds in the considered sample. Here, a difference between LIBS and PIL spectra has to be pointed out. While the LIBS spectra show numerous and very fine peaks, only two broad bands are observed with PIL spectroscopy. So, in order to further compress the PIL spectral domain, the baseline part between the two bands was suppressed. In the rest of the dataset, the selection of the purest pixels was applied as in the LIBS dataset.

4) *Unmixing MCR-ALS analysis on the double compressed image data*: once the compression is accomplished, the MCR-ALS process can be started to retrieve the spectral signatures and related distribution maps of the compounds in the image. Using the double-compressed dataset $\mathbf{D_{compr.\ \lambda/p}}$, the first step was to newly apply the SIMPLISMA-based method to extract the purest

spectra to be used as initial estimates for the signal unmixing technique, as normally done in the routine approach, and, most importantly, to assist in the selection of the number of components for the unmixing step. Indeed, selecting the optimal number of components for an MCR model is always a challenge on complex imaging samples, particularly if they contain minor compounds [27], since the presence of these contributions may not be detected by methods based on analyzing the variance explained to decide the MCR model size, such as Singular Value Decomposition (SVD) does. The innovation in this work is that instead of using SVD to estimate the number of components, the purity of the selected spectra, p_i, as defined by SIMPLISMA, will be the adopted criterion [28]. Hence, a graphical representation plotting purity vs. nr. of components is used to set the threshold for component selection. For the sake of a better interpretability, the y-axis represents p_i/p_1, being p_1 the purity of the first selected spectrum and p_i the purity of the following *i* selected spectra. In this way, the y-axis goes from 0 to 1, i.e., a value of 0.9 will mean that the purity of a certain selected spectrum, meaning a new component in the model, is 90% the value of the first selected spectrum. Therefore, it is possible to observe the difference in purity of any selected spectrum with the first one, related to the difference in spectral shape between them and not to explained variance. Of course, the purity value decreases from component to component and a threshold expressed as, e.g., 1% of purity with respect to the first spectrum selected, can be set. In this manner, it is possible to estimate the right interval of components to consider in the MCR-ALS analysis, avoiding to include either too few or too many possible chemical contributions. An example of this graphical representation is reported in the supplementary material (Fig. S2), where a number of components around 8–10 seems a reasonable estimate. Different MCR-ALS resolutions will then be calculated for different values of rank within the range estimated but a single solution will finally be adopted, based on the quality of the extracted pure spectra compared to LIBS database spectra (Fig. 3c). The MCR model obtained can be expressed as: $\mathbf{D}_{\text{compr. }\lambda/p} = \mathbf{C}_{\text{compr. }\lambda/p}\,\mathbf{S}^{\mathbf{T}}_{\text{compr. }\lambda/p}$ where $\mathbf{C}_{\text{compr. }\lambda/p}$ and $\mathbf{S}^{\mathbf{T}}_{\text{compr. }\lambda/p}$ are compressed versions of the information in the distribution maps and spectral fingerprints, respectively. In these analyses, only the constraint of non-negativity was used in the concentration and spectral mode, respectively.

5) *Reconstruction of full distribution maps and spectral signatures*: in this last step, the full spectral signatures and complete distribution maps of the initial image were recovered with two suitable single least-squares steps. Note that the MCR model that corresponds to the analysis of full image would be $\mathbf{D} = \mathbf{C}_{\mathbf{tot}}\,\mathbf{S}^{\mathbf{T}}_{\mathbf{tot}}$, meaning $\mathbf{D}$ the matrix containing all the spectra of the original image, $\mathbf{S}^{\mathbf{T}}_{\mathbf{tot}}$ the pure complete spectra of the compounds in the image and $\mathbf{C}_{\mathbf{tot}}$ the matrix of related concentration profiles that conveniently refolded provide the complete distribution maps. First, the double compressed spectral signatures ($\mathbf{S}^{\mathbf{T}}_{\text{compr. }\lambda/p}$, where λ represents the compressed spectral channels and p the compressed pixels) obtained from the previous MCR results were combined with each of the *i* subimages ($\mathbf{D}_{\mathbf{i},\text{compr. }\lambda}$), where the spectral dimension was compressed but the pixel dimension was as in the original image, to rebuild the concentration profiles ($\mathbf{C}_{\mathbf{i}\text{ compr. }\lambda}$), as represented in Eq. (1):

$$\mathbf{C}_{\mathbf{i}\text{ compr. }\lambda} = \mathbf{D}_{\mathbf{i},\text{compr. }\lambda}\,(\mathbf{S}^{\mathbf{T}}_{\text{compr. }\lambda/p})^{+} \tag{1}$$

where $(\mathbf{S}^{\mathbf{T}}_{\text{compr. }\lambda/p})^{+}$ is the pseudoinverse of matrix $\mathbf{S}^{\mathbf{T}}_{\text{compr. }\lambda/p}$. All the ($\mathbf{C}_{\mathbf{i}\text{ compr. }\lambda}$) matrices were appended together to reobtain the matrix of concentration profiles corresponding to the whole initial image, $\mathbf{C}_{\mathbf{tot}}$ (Fig. 2d), the profiles of which can be conveniently refolded to give distribution maps. Finally, by combining the obtained concentration profile $\mathbf{C}_{\mathbf{tot}}$ matrix with the original data matrix $\mathbf{D}$ (represented by the whole spectral domain), the full spectral signatures ($\mathbf{S}^{\mathbf{T}}_{\mathbf{tot}}$) were finally obtained (Fig. 2e), as represented in Eq. (2):

$$\mathbf{S}^{\mathbf{T}}_{\mathbf{tot}} = (\mathbf{C}_{\mathbf{tot}})^{+}\,\mathbf{D} \tag{2}$$

3.2. *Data fusion strategy*

As already described above, at first the LIBS and PIL datasets will be separately investigated by the use of the MCR-ALS analysis, in order to obtain a general idea about the information coming from the different techniques. This will form the first part of the results section. Without revealing these results, the observation of the average spectra acquired in LIBS and PIL already allow us to think that the exploitation of the latter will certainly be very limited due to the low number of spectral contributions when it is used alone. The interpretation of PIL spectral data acquired on complex natural samples remains a real challenge today. So, the main idea of this data fusion approach was to increase the possibility to better understand the luminescence phenomenon, investigating potential correlation or anticorrelation between the two spectroscopic techniques. Different fusion strategies have already been proposed in the literature [29–31]. In this work, the low-level data fusion was used for this study, which means concatenating the LIBS and PIL spectrum related to each image pixel to form a multiset configuration. Note that this step is easily done because the instrumental setup of the LIBS/PIL system ensures a perfect spatial congruency between LIBS and PIL pixel spectra. As a first step, all the LIBS and PIL pixel spectra, already compressed in the spectral dimension were concatenated together. It is important to stress here that only the LIBS and PIL matrices from the compression on the spectral variables were used for the fusion, not those from the double compression. The reason for using this strategy is that PIL spectra alone, as already explained, have much poorer information than LIBS. Using all the pixels initially in the fused structure ensures that the subsequent pixel selection will be carried out in such a way that the relevant correlation between the two techniques is appropriately captured. Finally, the extraction of the initial estimates with SIMPLISMA and the MCR-ALS analysis were carried out as previously described, but using the fused LIBS/PIL dataset. Again, the full spectral signatures and complete distribution maps were recovered using two single least-squares steps combining the MCR results and the information in the extended pixel and spectral dimensions.

4. Results and discussion

4.1. *LIBS dataset analysis. MCR-ALS results*

Due to its complexity, the LIBS dataset was the first one to be analyzed and investigated by the use of MCR-ALS. With the proposed data analysis pipeline, only 89400 over the more than 2 million of spectra (the 4% of the whole information) and only 489 spectral channels over the initial 2048 (the 24% of the initial value) were used. Considering the double compression, only 1% of the initial image information was used to perform MCR-ALS analysis and finally extract the full size maps and resolved LIBS spectra of each pure chemical contribution. The number of components needed to describe this compressed information, as estimated from the purity-based graphical method then suggested to select 6 significant chemical contributions. Fig. 4 shows that the first four contributions from the MCR-ALS model on the compressed data are

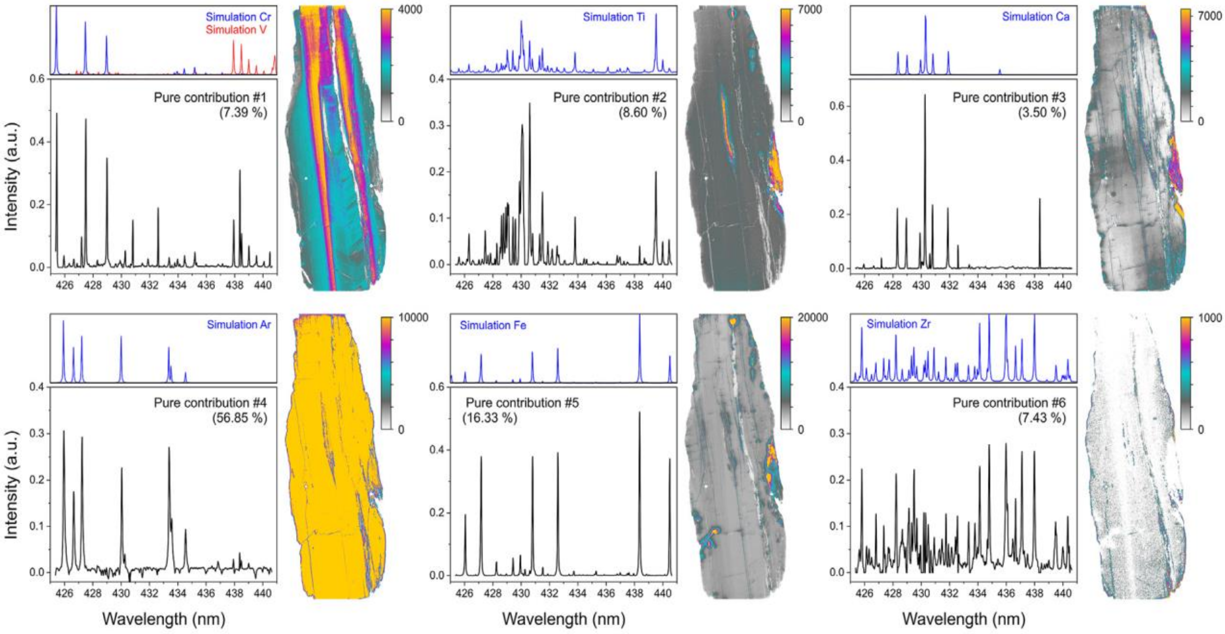


Fig. 4. MCR-ALS (Multivariate Curve Resolution - Alternating Least Squares) results on the LIBS dataset of the considered kyanite sample.

related to elements already observed in Fig. 2. Comparison of the resolved LIBS features with simulated spectra confirmed the identity of these compounds. However, the results offer as additional information the evidence that Cr and V are spatially correlated because they are present in the same MCR pure component. It is then interesting to see that the other 5 pure contributions correspond to single elements which is quite unusual in the context of LIBS signal unmixing. If we look at the distributions of these elements, we could first say that Ca, Fe and Ti are approximately located in the same areas. However, a more detailed analysis shows specific sub-zones of the sample for each of them. Additionally, it should be noted that Ca and Fe are often present along cracks of the mineral. The Ti contribution is also very interesting because it highlights areas of the sample for which no PIL signal was observed in Fig. 2. The last two components deserve a separate discussion. The first one seems to be equally distributed in all the mineral (signal contribution #4). The extracted spectrum is undeniably that of argon. However, the reader should not misunderstand the location of this element, which is not in fact part of the sample but comes from the gas flow above it, used to stabilize the plasma. As for the last contribution #6, its distribution image seems at first sight to be mainly related to noise. However, the extracted pure spectrum shows that the LIBS signature can be unambiguously attributed to the emission spectrum of zirconium. This is a very interesting result because this trace element could not be detected by classical single band LIBS integration because of the low signal-to-noise ratio of the related signal and the strong spatial and spectral overlap with other major elements.

4.2. PIL dataset analysis. MCR-ALS results

As previously explained, the interpretation of only the PIL image can be very challenging because of the broad spectral features provided by this technique, and this sample is not an exception. Although the luminescence of kyanite has been studied for more than 80 years [12], the interpretation of the emission characteristics is still not clear; hence the interest in proposing an original spectroscopic setup and an associated data processing approach. Researchers agree that the emissions observed in the luminescence spectrum are attributed to different Cr^{3+} centers in the aluminosilicate mineral. They also often associate the differences in luminescence behavior of Cr^{3+} with its substitution in different positions of Al^{3+} inside the kyanite structure. Nevertheless, these positions are so similar that it is challenging to explain significant differences in luminescence properties. It is therefore time to see whether the MCR-ALS approach can help us in this exploration of the PIL dataset exploited alone. The pure spectra and the corresponding distribution maps extracted from this signal unmixing method are shown in Fig. 5. First, it is interesting to see that the number of components estimated for the PIL dataset was two.

We were far from observing this with the classical integration method since the two PIL images at 689 and 706 nm presented in Fig. 2 were at first sight very similar to each other. Thus, even if the two spatial distributions extracted with MCR-ALS are very close, they still show areas with some variations in intensity. From a spectral point of view, we see that the two most intense emission bands of the dataset are now separated in these two contributions. In the pure contribution #1, an intense luminescence signal at 706.87 nm is accompanied by a doublet of low intensity at 711.03 and 712.87 nm. In addition, a luminescence signal is observed at 688.93 nm. It is in this same spectral zone that we can find the maximum of luminescence in the pure component #2 centered on 689.38 nm. Thus we now understand that the luminescence initially observed around 689 nm from the raw data was in fact at least coming from two distinct signals. Finally, a last observation of the pure component #2 could make us think of the presence of a doublet at 706.19 and 707.48 nm. This is not the case since it is in fact the representation of a broadening of the emission band observed on the component #1 at 706.87 nm. With the results obtained from the MCR-ALS analysis of PIL data, the complexity of the PIL spectra can be confirmed. However, the very strong overlap between the components both in the spectral and spatial directions limits the signal unmixing power of the method and the ambiguity in the solutions obtained hinders the proper understanding of the luminescence phenomenon. To improve this situation, the fusion of the PIL and LIBS datasets will be mandatory to reliably understand which kind of elements are located in a specific zone of the mineral and associated with the luminescence effect.

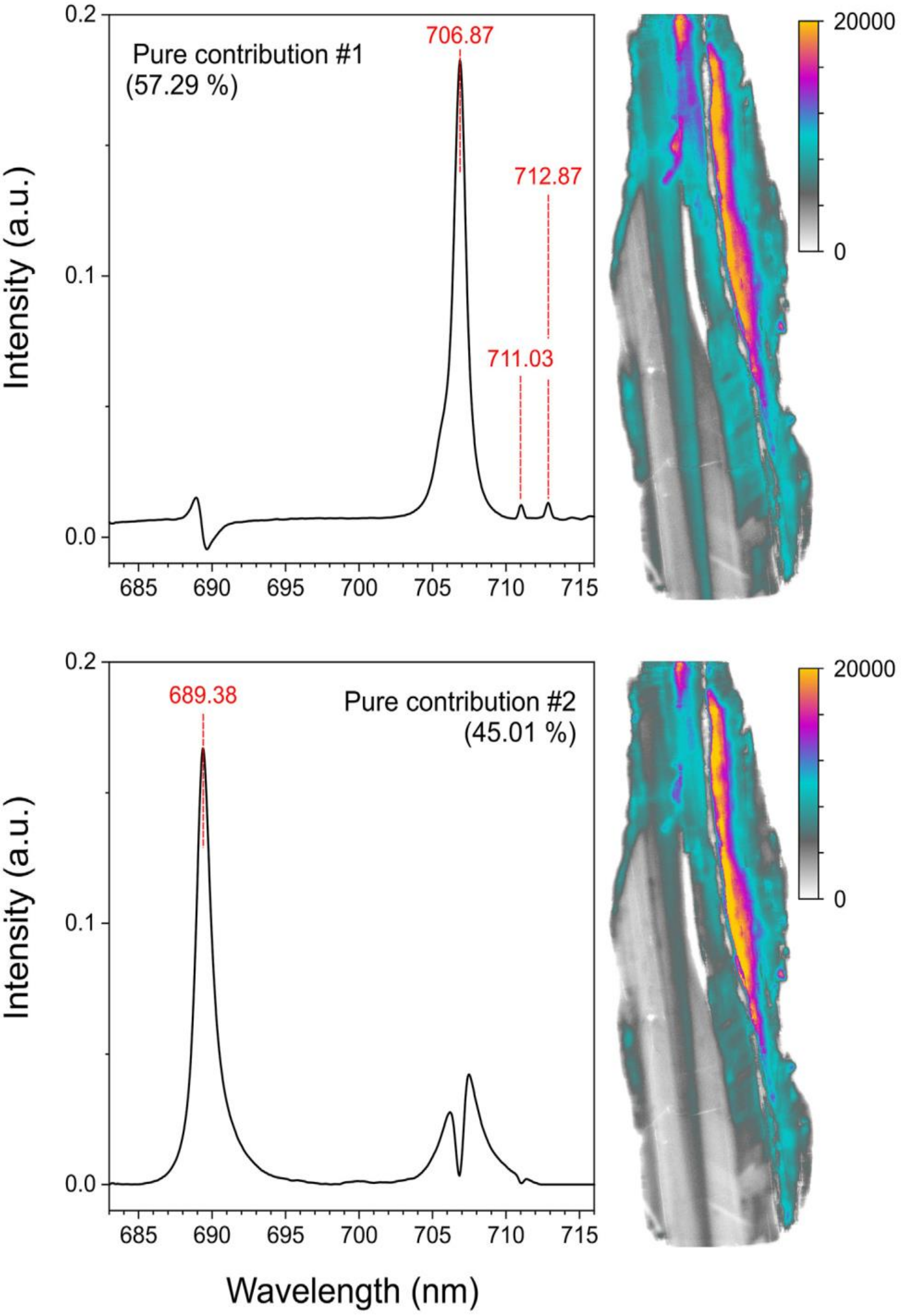


Fig. 5. MCR-ALS results on the PIL dataset of the considered kyanite sample.

4.3. LIBS/PIL data-fusion. MCR-ALS results

After the observation of the individual MCR-ALS results for the LIBS and the PIL images separately, the fusion of the two datasets was carried out. The main idea of this strategy was to deepen the chemical information from both techniques by finding correlations among the different spectroscopic signals and components. In particular, by combining the finer spectroscopic features of LIBS with the broad PIL signals, it now seems possible to give a suitable chemical interpretation to the luminescence signals. Using the proposed compression pipeline, only 89400 pixels and 1165 spectral channels were selected from the fused dataset. More precisely, considering the initial size of the merged dataset (more than 2 million spectra and about 4000 channels for both spectroscopies), the amount of information resulting from the double compression and thus finally used for the signal unmixing constituted only 1% of the original data. The MCR-ALS results on the fused dataset are given in Fig. 6 after the estimation of the number of components suggested seven significant chemical contributions. As a reminder, the pure spectra extracted by MCR-ALS on the fused dataset contain simultaneously a LIBS part and a PIL one. Thus, the extended LIBS/PIL fingerprint of the extracted components will directly differentiate the elements that present a LIBS signal associated with a PIL luminescence phenomenon from those that only have a LIBS signal and no luminescence induced. The observation of the first three pure contributions in Fig. 6 clearly show that the elements Ca and Fe present a very low luminescence phenomenon (689.5 nm), which is completely absent for Ti. Moreover, there is no significant luminescence for the Ar element either (pure contribution #6), which it is not part of the sample but of the atmosphere above its

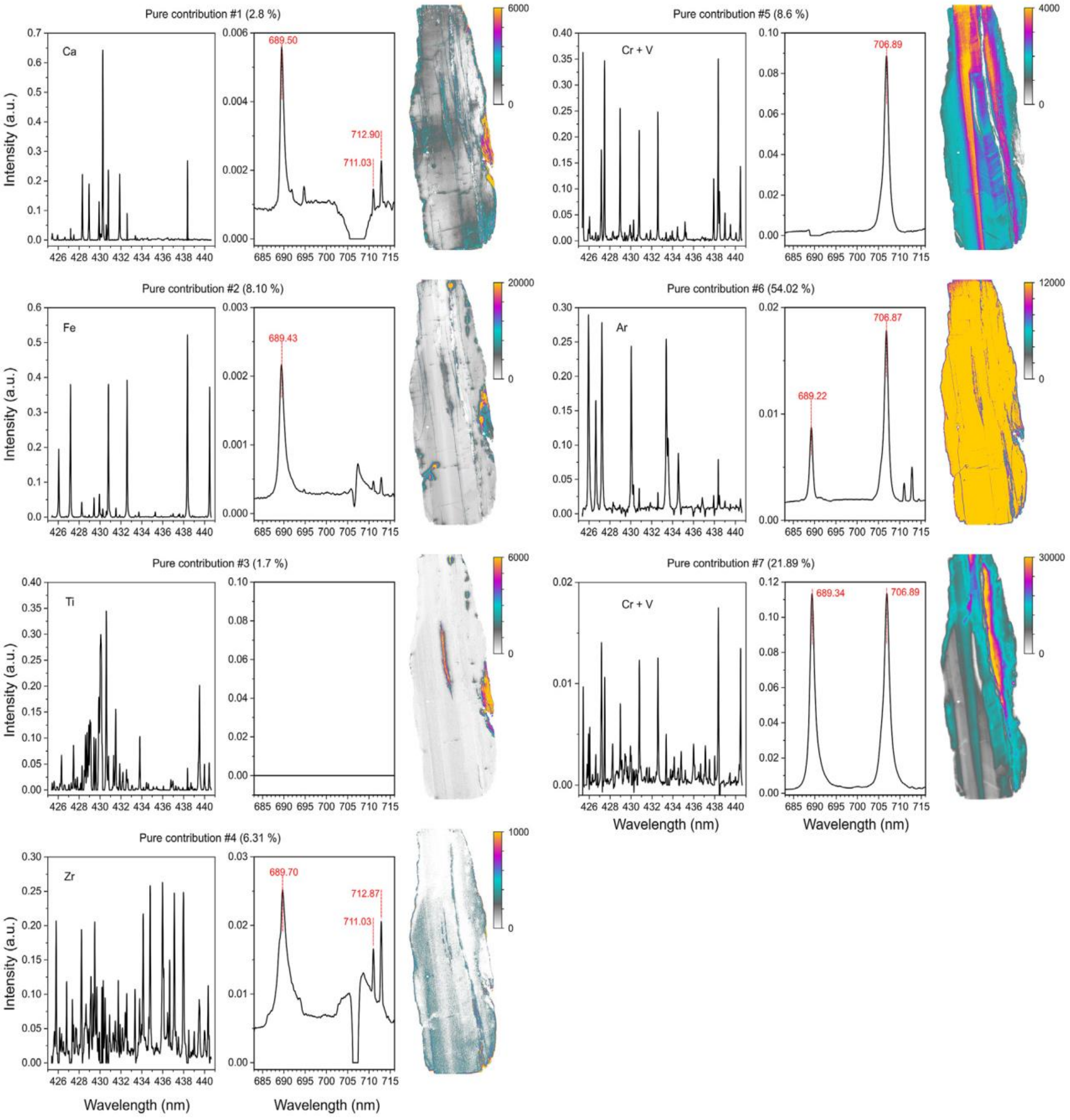


Fig. 6. MCR-ALS results on the fused LIBS/PIL dataset of the considered kyanite sample.

surface. The strongest luminescence signals are observed for pure contributions #5 and #7 corresponding to a mixture of Cr and V followed by a weaker but nevertheless significant luminescence for Zr in pure contribution #4. This last contribution is very interesting because Zr does not form luminescence center. We can therefore say that the luminescence observed on contribution #4 is certainly due to an indirect correlation of another element with Zr. If we look at these three PIL contributions at the spectral level, they are quite singular. First, the pure contribution #7 has two peaks at 689.34 and 706.89 nm. On the other hand, the pure contribution #5 shows only one peak at 706.89 nm. Finally, pure contribution #4 has a peak at 689.70 nm and an unclear very low signal at 711.03 and 712.87 nm. These last results show the power of the MCR-ALS approach associated with LIBS/PIL data fusion since the luminescence is not due to two intense contributions, but potentially several contributions slightly shifted in wavelength that can be differentiated. In conclusion, we can state that the luminescence of this sample comes mainly from the simultaneous presence of Cr and V. These results are fully consistent with previous work suggesting that luminescence could potentially originate from the Cr^{3+} and V^{2+} centers [12].

However, even if we can state this, there is still the end of the story to write. Indeed, even though the spatial distributions of the pure #5 and #7 contributions are relatively close and partially overlapped, they both potentially show singular chemical information. The same statement can also be made at the spectral level. Thus, the joint exploitation of new LIBS and PIL spectral domains and even an extension of the fusion concept to other spectroscopic imaging techniques should help further in elucidating the whole nature of the complex luminescence phenomenon.

5. Conclusions

LIBS imaging is now clearly a tool of choice for the elemental characterization of complex samples with applications in many fields. Nevertheless, its high acquisition rate, which is an undeniable advantage, is also constrained by the millions of spectra (each containing thousands of wavelengths) acquired from a single sample that require the use of powerful multivariate data analysis tools. This difficult aspect leads to the interest in proposing a data analysis procedure capable of extracting the most distinct information, i.e., purest variables, at the spectral and spatial level, both for major and minor compounds, to facilitate the unmixing analysis without losing quality in the spatial and spectral definition of the imaged components. With this simple procedure, it was possible to reduce the initial amount of data and keep the best and most unmixed 1% of the total information. The size and quality of the selected information allowed not only speeding up the analysis, but obtaining extremely reliable spectral fingerprints and distribution maps for the extracted MCR-ALS components. The data analysis pipeline has been tested on the LIBS/PIL dataset, but can be used in any other kind of large imaging dataset coming from an individual platform or from the fusion of several of them.

The study of the kyanite dataset showed that each resolved component was potentially related to one or two elements present in the mineral. Last but not least, another important aspect of this work was fusing together LIBS and PIL datasets to provide a chemical interpretation for the PIL bands and better understand which elements were related to this luminescence effect. This work represents the first published work on the fusion of LIBS and PIL imaging data and their simultaneous exploitation in a signal unmixing approach such as MCR-ALS. Thus even if we have demonstrated for this particular kyanite sample that the luminescence phenomenon was mainly associated with the Cr and V elements, our next work could be focused on the exploitation of new spectral domains or the addition of another spectroscopy such as Raman and Laser-Induced Time-Resolved Luminescence to the first two in a fusion process.

Notes

The authors declare no competing financial interest.

CRediT authorship contribution statement

Alessandro Nardecchia: Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft. **Anna de Juan:** Conceptualization, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Vincent Motto-Ros:** Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – review & editing. **Michael Gaft:** Validation, Visualization, Writing – review & editing. **Ludovic Duponchel:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

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.

Acknowledgment

The authors warmly thank Rodrigo Rocha de Oliverira and Adrián Gómez-Sánchez, PhD students in the chemometric group of the department of Chemical Engineering and Analytical Chemistry (Universitat de Barcelona) for fruitful discussions about chemometrics strategies. The authors also thank the University of Lille for providing a mobility grant (called MOB-LIL-EX) to carry out this research at the Universitat de Barcelona. A.J. acknowledges financial support from the Spanish govern-ment through the research project PID 2019-1071586B-IOO.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2021.339368.

References

[1] L. Jolivet, M. Leprince, S. Moncayo, L. Sorbier, C.-P. Lienemann, V. Motto-Ros, Review of the recent advances and applications of LIBS-based imaging, Spectrochim. Acta B Atom Spectrosc. 151 (2019) 41–53, https://doi.org/10.1016/j.sab.2018.11.008.

[2] V. Motto-Ros, S. Moncayo, F. Trichard, F. Pelascini, Investigation of signal extraction in the frame of laser induced breakdown spectroscopy imaging, Spectrochim. Acta B Atom Spectrosc. 155 (2019) 127–133, https://doi.org/10.1016/j.sab.2019.04.004.

[3] T.A. Labutin, V.N. Lednev, A.A. Ilyin, A.M. Popov, Femtosecond laser-induced breakdown spectroscopy, J. Anal. At. Spectrom. 31 (1) (2016) 90–118, https://doi.org/10.1039/C5JA00301F.

[4] R.R.V. Carvalho, J.A.O. Coelho, J.M. Santos, F.W.B. Aquino, R.L. Carneiro, E.R. Pereira-Filho, Laser-induced breakdown spectroscopy (LIBS) combined with hyperspectral imaging for the evaluation of printed circuit board composition, Talanta 134 (2015) 278–283, https://doi.org/10.1016/j.talanta.2014.11.019.

[5] R. Gaudiuso, N. Melikechi, Z.A. Abdel-Salam, M.A. Harith, V. Palleschi, V. Motto-Ros, B. Busser, Laser-induced breakdown spectroscopy for human and animal health: a review, Spectrochim. Acta B Atom Spectrosc. 152 (2019) 123–148, https://doi.org/10.1016/j.sab.2018.11.006.

[6] F. Trichard, F. Gaulier, J. Barbier, D. Espinat, B. Guichard, C.-P. Lienemann, L. Sorbier, P. Levitz, V. Motto-Ros, Imaging of alumina supports by laser-induced breakdown spectroscopy: a new tool to understand the diffusion of trace metal impurities, J. Catal. 363 (2018) 183–190, https://doi.org/10.1016/j.jcat.2018.04.013.

[7] A. Nardecchia, C. Fabre, J. Cauzid, F. Pelascini, V. Motto-Ros, L. Duponchel, Detection of minor compounds in complex mineral samples from millions of spectra: a new data analysis strategy in LIBS imaging, Anal. Chim. Acta 1114 (2020) 66–73, https://doi.org/10.1016/j.aca.2020.04.005.

[8] J.O. Cáceres, F. Pelascini, V. Motto-Ros, S. Moncayo, F. Trichard, G. Panczer, A. Marín-Roldán, J.A. Cruz, I. Coronado, J. Martín-Chivelet, Megapixel multi-elemental imaging by laser-induced breakdown spectroscopy, a Technology with considerable potential for paleoclimate studies, Sci. Rep. 7 (1) (2017) 1–11, https://doi.org/10.1038/s41598-017-05437-3.

[9] C. Fabre, D. Devismes, S. Moncayo, F. Pelascini, F. Trichard, A. Lecomte, B. Bousquet, J. Cauzid, V. Motto-Ros, Elemental imaging by laser-induced breakdown spectroscopy for the geological characterization of minerals, J. Anal. At. Spectrom. 33 (8) (2018) 1345–1353, https://doi.org/10.1039/C8JA00048D.

[10] M. Gaft, L. Nagli, Y. Groisman, Plasma induced luminescence (PIL), Opt. Mater. 34 (2) (2011) 368–375, https://doi.org/10.1016/j.optmat.2011.05.024.

[11] M. Gaft, Y. Raichlin, F. Pelascini, G. Panzer, V. Motto Ros, Imaging rare-earth elements in minerals by laser-induced plasma spectroscopy: molecular emission and plasma-induced luminescence, Spectrochim. Acta B Atom Spectrosc. 151 (2019) 12–19, https://doi.org/10.1016/j.sab.2018.11.003.

[12] M. Gaft, R. Reisfeld, G. Panczer, Modern Luminescence Spectroscopy of Minerals and Materials, Springer Mineralogy; Springer International Publishing, Cham, 2015, https://doi.org/10.1007/978-3-319-24765-6.

[13] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate Curve resolution in matlab, Chemometr. Intell. Lab. Syst. 76 (1) (2005) 101–110, https://doi.org/10.1016/j.chemolab.2004.12.007.

[14] A. de Juan, R. Tauler, Multivariate Curve resolution (MCR) from 2000: progress in concepts and applications, Crit. Rev. Anal. Chem. 36 (3–4) (2006) 163–176, https://doi.org/10.1080/10408340600970005.

[15] 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.

[16] 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.

[17] L. Bassel, V. Motto-Ros, F. Trichard, F. Pelascini, F. Ammari, R. Chapoulie, C. Ferrier, D. Lacanette, B. Bousquet, Laser-induced breakdown spectroscopy for elemental characterization of calcitic alterations on cave walls, Environ. Sci. Pollut. Res. 24 (3) (2017) 2197–2204, https://doi.org/10.1007/s11356-016-7468-5.

[18] V. Motto-Ros, E. Negre, F. Pelascini, G. Panczer, J. Yu, Precise alignment of the collection fiber assisted by real-time plasma imaging in laser-induced breakdown spectroscopy, Spectrochim. Acta B Atom Spectrosc. 92 (2014) 60–69, https://doi.org/10.1016/j.sab.2013.12.008.

[19] A. Nardecchia, V. Motto-Ros, L. Duponchel, Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed

phenomenon? Anal. Chim. Acta 1157 (2021) 338389, https://doi.org/10.1016/j.aca.2021.338389.

[20] C. D'Angelo, D. Diaz Pace, D. Bertuccelli, G. Bertuccelli, Spectroscopic analysis of signals from LIBS experiments, in: O.A. Marcano, J.L. Paz (Eds.), 2004, pp. 1037–1042, https://doi.org/10.1117/12.591209.

[21] E.T. Whittaker, On a new method of graduation, Proc. Edinb. Math. Soc. 41 (1922) 63–75, https://doi.org/10.1017/S0013091500077853.

[22] Abraham Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (8) (1964) 1627–1639, https://doi.org/10.1021/ac60214a047.

[23] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (14) (1991) 1425–1432, https://doi.org/10.1021/ac00014a016.

[24] W. Windig, J.L. Lippert, M.J. Robbins, K.R. Kresinske, J.P. Twist, A.P. Snyder, Interactive self-modeling multivariate analysis, Chemometr. Intell. Lab. Syst. 9 (1) (1990) 7–30, https://doi.org/10.1016/0169-7439(90)80050-G.

[25] S. Gourvénec, D.L. Massart, D.N. Rutledge, Determination of the number of components during mixture analysis using the durbin–watson criterion in the orthogonal projection approach and in the SIMPLe-to-use interactive self-modelling mixture analysis approach, Chemometr. Intell. Lab. Syst. 61 (1–2) (2002) 51–61, https://doi.org/10.1016/S0169-7439(01)00172-1.

[26] A. Nardecchia, L. Duponchel, Randomised SIMPLISMA: using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging, Talanta 217 (2020) 121024, https://doi.org/10.1016/j.talanta.2020.121024.

[27] M. Boiret, A. de Juan, N. Gorretta, Y.-M. Ginot, J.-M. Roger, Distribution of a low dose compound within pharmaceutical tablet by using multivariate Curve resolution on Raman hyperspectral images, J. Pharmaceut. Biomed. Anal. 103 (2015) 35–43, https://doi.org/10.1016/j.jpba.2014.10.024.

[28] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (3) (1936) 211–218, https://doi.org/10.1007/BF02288367.

[29] A. de Juan, A. Gowen, L. Duponchel, C. Ruckebusch, Image fusion, in: Data Handling in Science and Technology, vol. 31, Elsevier, 2019, pp. 311–344, https://doi.org/10.1016/B978-0-444-63984-4.00011-9.

[30] S. Piqueras, C. Bedia, C. Beleites, C. Krafft, J. Popp, M. Maeder, R. Tauler, A. de Juan, Handling different spatial resolutions in image fusion by multivariate Curve resolution-alternating least squares for incomplete image multisets, Anal. Chem. 90 (11) (2018) 6757–6765, https://doi.org/10.1021/acs.analchem.8b00630.

[31] A. Gómez-Sánchez, M. Marro, M. Marsal, P. Loza-Alvarez, A. de Juan, 3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images, Anal. Chem. 92 (14) (2020) 9591–9602, https://doi.org/10.1021/acs.analchem.0c00780.

4.4. LIBS and Raman spectroscopy data fusion analysis

4.4.1. The importance of choosing the right pretreatment and analysis pipeline to obtain good results fusing different spectroscopic responses

As previously mentioned, an interesting aspect related to LIBS is that this kind of instrumentation can be easily coupled with other spectroscopic techniques, without the necessity of changing the acquisition setup, leading to a fast acquisition of different spectral responses in a very reasonable time. The main limitation, as described in the previous paragraph, is that a huge amount of data is easily generated, leading to problems related to the analysis of the raw data. Nevertheless, the possibility of using an approach able to reduce the dimensionality of the data cube and, at the same time, to conserve only the most important information is fundamental to dig in any analytical domain, no matter the spectroscopy used for the acquisition. This seems to be particularly true when the data fusion is the main purpose of the study, due to the fact that a huge amount of spatial and spectral information will be generated, coming from different instrumental responses. If any pretreatment is not applied a priori, also if it would be possible to analyze the data, hardly the outcomes will represent the reality of the information contained. Clearly, this situation can be faced using a very interesting spectroscopy such as Raman, which can be linked to complex signals, considering the molecular information that can be extrapolated from the matrix using this instrument. In order to show the interesting aspects of a data fusion using these two spectroscopies, a specific data cube will be here described and investigated by the use of chemometric methodologies. It is important to highlight here the fact that a work reporting these aspects of data fusion between LIBS and Raman has been recently submitted to the journal Spectrochimica Acta Part B: Atomic Spectroscopy. The selected data cube is a subzone of the same sample used and well described into another work of this PhD, and previously explained [212]. The dimensions of the selected area of the matrix are 500 pixels by 316 pixels considering 1044 variables for Raman, in the spectral domain between 118 and 2000 cm^{-1}, 2048 variables for LIBS, between 251 and 335 nm, and a spatial resolution of 25 µm. Briefly, the observed hyperspectral image is a complex mineral sample containing traces of various elements and different phases. Therefore, if only LIBS is used for the analysis, in this case there is a real chance to miss some important information coming from a molecular point of view. This is why it is also interesting to fuse LIBS and Raman spectra to obtain finally a more complete idea of the heterogeneous nature of the mineral of interest. A first idea of the complexity of the given sample is here reported in Fig. 21, in which the main information related respectively to LIBS and Raman are considered:

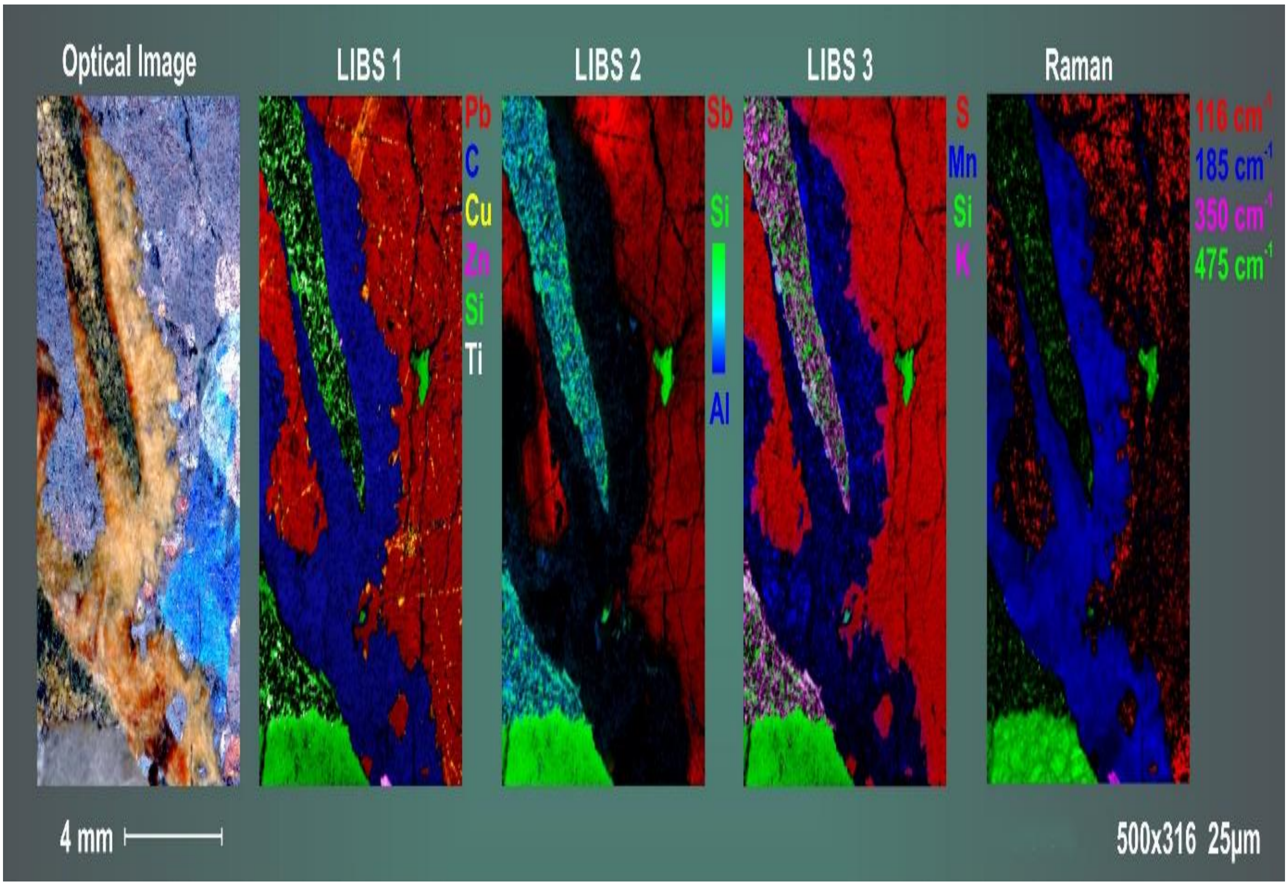


Fig. 21 – Respectively: optical image, LIBS and Raman first information of the section of the investigated mineral sample.

The aforementioned figure shows from the left to the right the optical image of the corresponding mineral used for the analysis, three integration images linked to LIBS spectra, describing the distribution of various elements in the sample and, lastly, the main Raman bands used to obtain the molecular distribution of different components present in the rock. It is also important to consider that these first outcomes have been obtained using a conventional procedure, meaning that each component in the images is observed by using the integration at a specific wavenumber (for LIBS) or at a particular wavelength (for Raman) that has to be peculiar for a given element or mineral phase. Clearly, this means that a first investigation of the sample of interest, a general knowledge of its composition, and an observation of the most important spectral bands for both the spectroscopies is mandatory to obtain a first idea of the complex nature of the specimen. Differently, what is proposed in this part of the manuscript is the use of various chemometric tools in order to obtain at the end a general comprehension of the given sample, also considering the aspect of not having a priori knowledge of its composition. Another important point that has to be considered is that when different spectral responses are acquired using the same instrument, some issues can be faced. Particularly, it is a very easy scenario the one in which some instrumental problems will be generated due to, for example, different technical settings. Clearly, it is mandatory to correct these artifacts before any further analysis in order to generate the right outcomes. For informational purposes, a fast and useful procedure used to correct LIBS problems developed during this PhD work has been already described in the same aforementioned published paper [216] and so it will not be discussed again in this chapter. More interesting is the approach applied on Raman spectra. Due to the acquisition settings, observing the corresponding spectra (reported in Fig. 22a), it is clear how the raw signals are related to very extreme values, which seem to be impossible to be investigated without any previous pretreatment. Chemometrics can be also in this case a useful tool able to generate better initial data to be used for the final investigation. As a starting point, it is noticeable that many pixels (i.e., many spectra) show saturated signals and fluorescence. Therefore, the first thing to be done is to set a threshold on a maximum intensity to remove from the data cube these extreme spectra. Another problem in the dataset is that a negative peak around 1606 cm^{-1} is observable, probably due to instrumental issues. To correct this spectral imperfection, an interpolation with the neighboring points of this spectral zone was applied. Finally, at this stage, it was possible to apply a baseline correction, using the AsLS algorithm ($\lambda = 10^4$ and asymmetry parameter of 0.001). As final step, a normalization based on the ratio of each spectrum of the dataset for its own norm was used. This step is necessary to highlight the presence of possible spikes coming from the acquisition. In order to simplify the procedure for managing these spectra, these values were replaced by the average spectrum of the neighboring pixels of that particular pixel. As a proof of the significant difference between the raw and the final data, the pretreated spectra are reported in Fig. 22b. To remark it, it is out of the question the fact that the initial raw data cannot be used in any analysis, due to the saturated signals and the absence of a baseline, reasons why many aspects of this dataset would be missed. Contrariwise, using the proposed approach, finally it is possible to give an interpretation to many different spectral bands, as it will be better described in the following paragraphs.

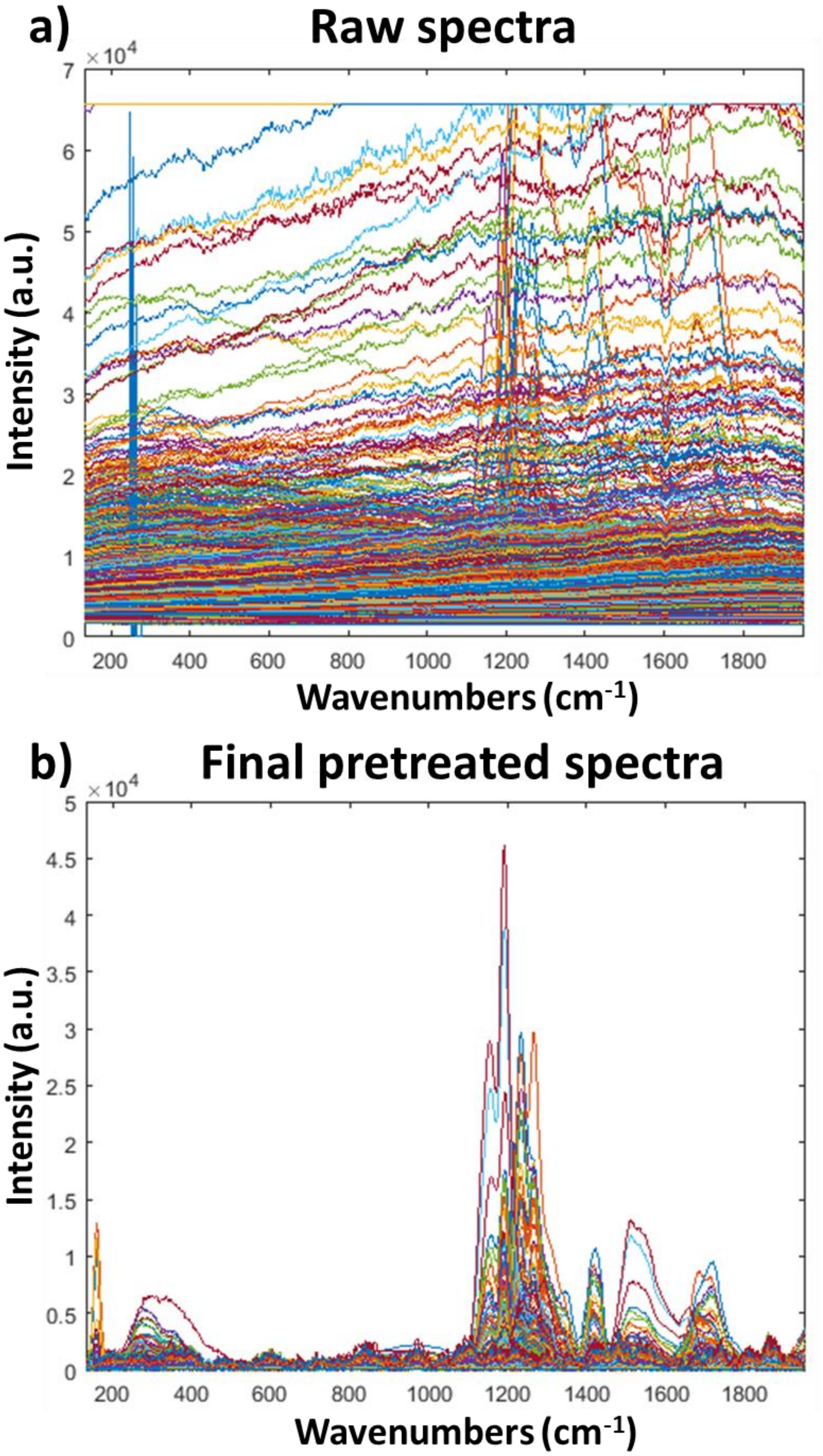


Fig. 22 – Comparison between a) raw spectra and b) final pretreated spectra for Raman data.

4.4.2. First outcomes and interpretation by the use of the spectral unmixing approach

Here are reported the first results obtained using MCR-ALS on this complex mineral sample. It is important to stress here the fact that in this study has been used the same procedure of the previous work. For this reason, it will not be described in detail in the present section. Nevertheless, some useful information needs to be recalled. First of all, it is important at the beginning to separately investigate the two different matrices. This is a very obvious step. In fact, as previously stated, LIBS and Raman spectroscopies are related to different responses. In this way, it is possible to obtain a general idea from, respectively, the elemental and the molecular points of view, to dig the knowledge regarding both the elements present in the mineral, and its different compounds. Then, it is possible to investigate the results coming from the data fusion. The interesting aspect is that, as it will be shown in the results, some components are related exclusively to only one or the other spectroscopy. Therefore, in a second step, fusing the two spectral ranges, it is possible to obtain some new information (from the elemental and/or the molecular perspective) that otherwise would be not present if the fusion strategy is not applied. Another important aspect to be considered is that it is mandatory to normalize the data, in order to give them the same weight. In fact, without the right approach, the two spectral data would be probably represented by different scale intensities and variances, and so, it would be impossible to obtain the right contribution coming from LIBS and Raman simultaneously. Finally, one of the most important things to recall is that in the present approach, the first step is represented by the reduction of the data (both regarding the variables and the pixels), in order to use only a small percentage (and especially the purest details) of the initial information of the datasets, to obtain more precise and satisfactory results. Therefore, it is important to apply the fusion in the right moment of the analysis pipeline. This aspect mainly depends on the type of investigated data. Regarding the two used spectroscopies for this study, it is more convenient to apply the reduction of the data before the fusion step. The reason is that both the techniques are related to very important details from a spectroscopic point of view. As a consequence, it is better at first to skim the information, in order to keep only the most interesting part of the data, and then to use the fusion, to see how correlations and anticorrelations between different elements and compounds are carried out during the analysis. For informational purposes, the illustrated sample is represented by only a limited number of spectra of the original image. In fact, as already discussed, this data cube is a subzone of a bigger hyperspectral image. Precisely, the corresponding dimensions are 500 pixels by 316 pixels, for a total of 158000 spectra. LIBS was acquired using 2048 variables, and 1044 for Raman. So, the use of the data reduction in this case is not mainly related to the necessity of use a smaller quantity of spectra for a computational problem. Instead, it is used with the only purpose of selecting, to work properly, the most interesting spectra of the original dataset. Below are reported the first outcomes coming from separately LIBS, Raman and finally the data fusion. It is important to consider the fact that this is just a first interpretation of the results, which will need a deeper investigation in order to at last understand which are the elements and/or minerals phases related to each found pure contribution image.

4.4.2.1. Spectral unmixing results for LIBS data

One of the most interesting aspects regarding the interpretation of the results in LIBS analysis, as stated in the previous work, is that being this spectroscopy an elemental technique, it is possible using MCR-ALS to observe the contribution related to the different elements of the investigated matrix. Furthermore, due to the very selective and characteristic spectral information that this technique can show, it is possible to use libraries to drive the interpretation of the data and so, characterize the different elements and compounds contained in the analyzed mineral. The first results using the proposed approach are here reported, in Fig. 23. From a first observation, it is possible to easily recognize some specific mineral phases and compounds, very well distributed in the different areas of the mineral. For example, the first pure component is mainly related to silicon, and so recognized as quartz (SiO_2). The second and the fourth components show bands linked mainly to lead, but also traces of copper, silver and antimony with different contribution intensities. Probably, the corresponding mineral is galena (PbS), shown as different mineral phases. The third component seems to be a compound coming from the aluminosilicate class, showing peaks related to the presence of silicon, iron, aluminum and traces of magnesium. Fifth and ninth pure contributions show very characteristics bands of titanium (probably the second image is related to saturation signals of this element) and silicon. They can correspond to anatase, a metastable mineral form of composition TiO_2. Sixth, seventh and eighth components are related to iron, manganese, calcium, magnesium and traces of silicon. These elements are normally found in ankerite, a class of carbonate minerals. Lastly, the tenth contribution seems to be very interesting. In fact, it is related to a very specific distribution of traces of the corresponding mineral phase. Particularly, it is possible to observe bands related to sodium, copper, silver, iron and zinc. From a general point of view, it is then possible to confirm the fact that using the proposed approach, it has been possible to obtain a fast and global identification of different compounds of the heterogeneous nature of the mineral.

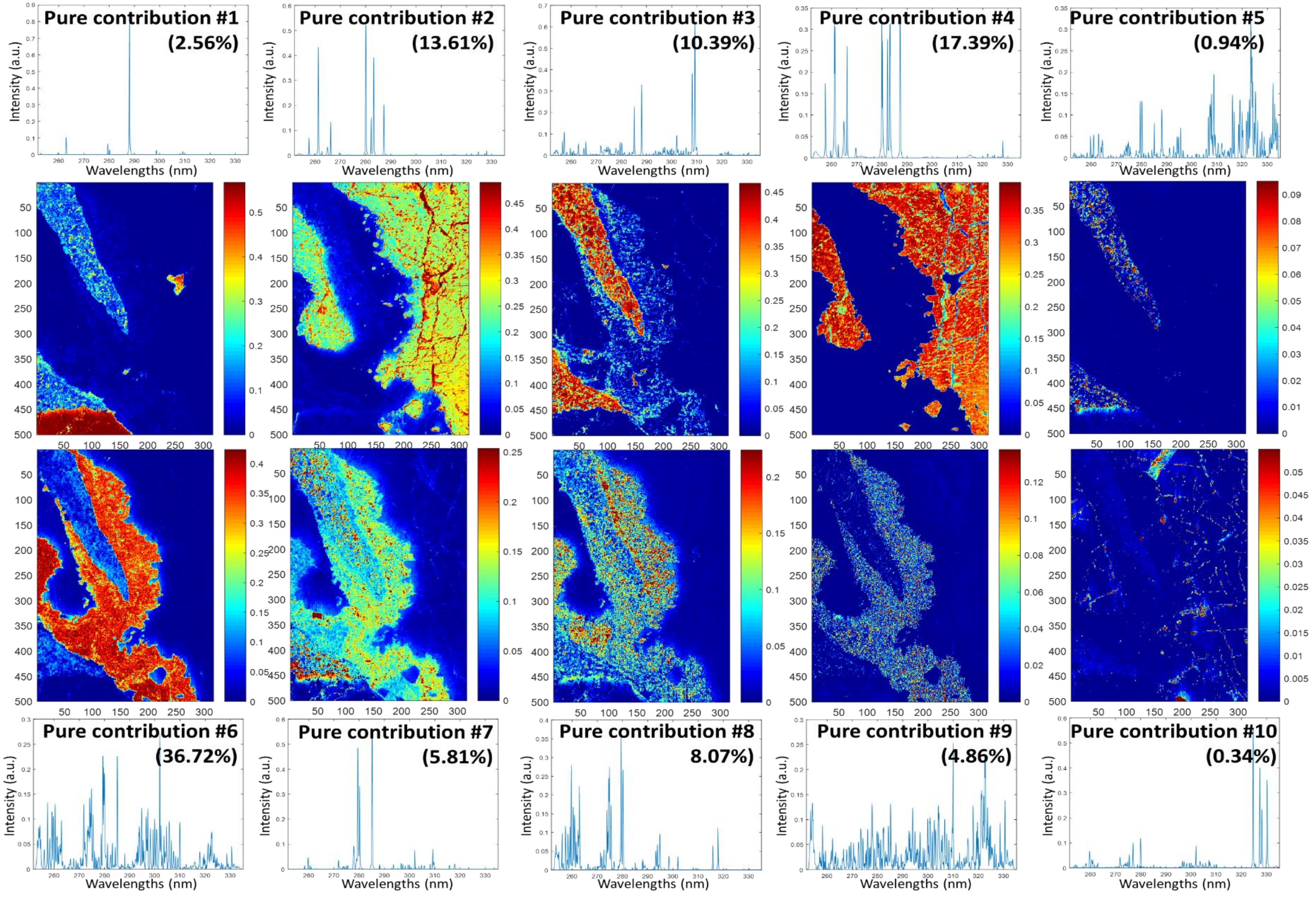


Fig. 23 – MCR-ALS results on the LIBS dataset of the considered mineral sample.

4.4.2.2. Spectral unmixing results for Raman data

In the same way, it is possible to observe the results coming from the Raman dataset. It is important to remember the fact that the original raw spectra, as previously shown in Fig. 22, were completely saturated. This is due to, normally, acquisition problems (i.e., different technical settings needed) using the same instrument to obtain the data of the two different spectroscopies. Naturally, it would be challenging to use this kind of data before applying a chemometric approach in order to correct, as much as possible, these imperfections. At the same time, it is understandable that some atypical signals can be anyway observed, particularly using the spectral unmixing, in which the main purpose is to find the pure contributions corresponding to the different purest signals in the dataset. This is the reason why, regarding the first outcomes of Raman, reported in Fig. 24, the interpretation has been more complicated compared with the LIBS results shown in the previous paragraph. Immediately, it is possible to observe that while some pure contributions are related to very fine spectral information, others are more complicated to be interpreted, due to their very noisy signals. Therefore, the different images were also compared with the already discussed results of LIBS, in order to have a better general idea of which compounds can be observed using the Raman spectroscopy. The first pure contribution is related to very good spectral signatures, as well as it is possible to notice a certain correlation between this map and the first one observed for LIBS results. In fact, this mineral phase is again linked to quartz. Second and fourth components are represented by very noisy signals, hard to be interpreted. This is the reason why a direct comparison using the previous pure contributions of LIBS has been used, leading to the hypothesis that they are related to the presence of galena. A similar situation is represented by the fifth component, which shows some interesting peaks, but complicated to be identified. Therefore, comparing this component with again the LIBS results, it is conceivable that this mineral phase is related to anatase due to the similarities in the distribution maps of this component and the fifth one of LIBS. The sixth component is very interesting. In fact, it is represented by very good spectral information that corresponds to the presence of ankerite. In addition, comparing the distribution maps of this compound for Raman and LIBS results, it is possible to observe that this mineral is related to a very vast area, but that its concentration intensity widely changes in the different zones of the rock. This is an information that can be obtained mainly from Raman spectra and not from the LIBS ones. Finally, third, seventh, and in part eighth components are related to very specific spectral signatures. Another important aspect is that some of the observable areas of these maps are related to zones of the mineral that are not highlighted using LIBS, and so, they are specifically related to Raman signals. Nevertheless, it has been impossible to give a chemical interpretation to these components, and therefore, to give a name to the corresponding minerals present in these areas. This is the perfect scenario in which a data fusion is mandatory, in order to finally obtain a better interpretation of the data, when separate analyses can show some limitations, as described in the next paragraph.

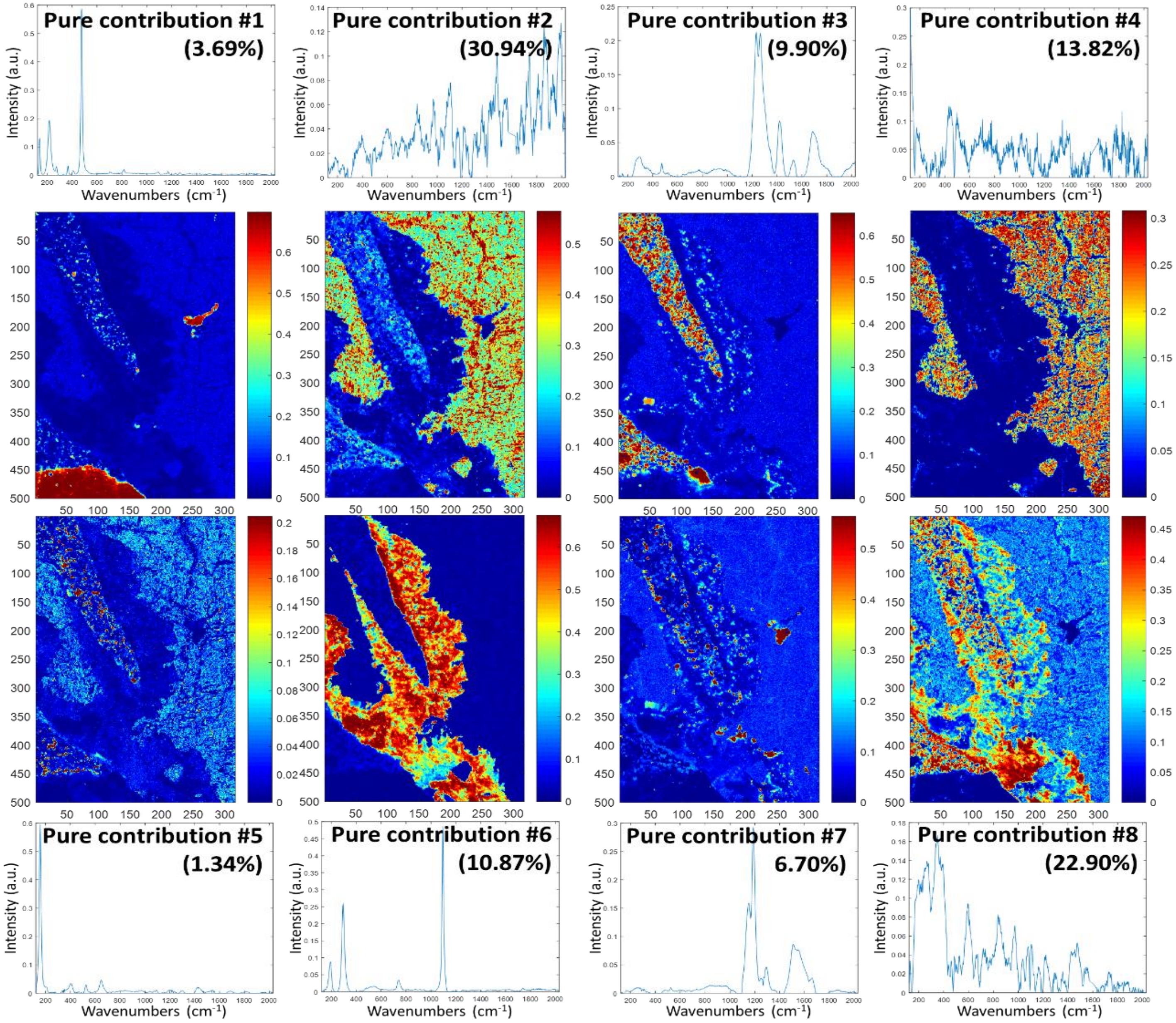


Fig. 24 – MCR-ALS results on the Raman dataset of the considered mineral sample.

4.4.2.3. Spectral unmixing results for the data fusion

Finally, here are reported the results fusing LIBS and Raman spectra, using the precautions previously explained. As stated, the importance in using a fusion strategy lies in the possibility of obtaining a more complete interpretation of the data. Particularly, as previously observed in the already shown results, it is not always possible to identify some components using only one or the other technique, due to for example libraries limitations. In addition, some pure contributions are related to only one or the other spectroscopy. This is why, fusing the data, the main purpose is to try to understand the existent correlations between the two spectral domains, and exceed the limits shown separately by LIBS and Raman. Here, are reported the first outcomes of the MCR-ALS approach, as shown in Fig. 25. Observing the results, it is possible to give some initial interpretations. Clearly, the first component is related to quartz, as shown by the spectral signals coming from both the techniques. It is also interesting to observe how, using the data fusion, a better discrimination of the pixels containing this mineral is possible, compared with LIBS and Raman spectra separately. In the same way, third and eighth maps are related to galena. Again, it is possible to observe the typical LIBS signals related to this mineral. In addition, using the fusion it is possible to assume that, as previously supposed, the noisy Raman signals are also correlated to this compound. Second pure contribution is also very interesting, comparing the two different spectral information. In fact, as a proof of the interpretation given in the previous paragraphs, it is possible to observe that both the spectral ranges are correlated to anatase. The sixth component too is related to the same mineral, except that some new Raman bands are here observable. Both the fifth, the seventh and the tenth components seem to be related to ankerite, also if the corresponding images show some distribution differences. Also noticeable is that while the map of the seventh contribution has very fine spectral signals for both the techniques, the ones related to the other two images show some broadened bands related to Raman spectra. In addition, comparing these maps with the ones of LIBS and Raman when the data fusion approach is not used, it is possible to notice that here a better distribution of the concentration intensity of this compound in the different areas of the mineral is noticeable. Lastly, fourth and ninth components deserve a separated and more detailed description. Referring again to Raman outcomes, the spectra corresponding to these two images are the ones previously observed that anyway could not be identified from a chemical point of view. Finally, using also the LIBS spectra, it is possible to give a first interpretation to these maps. The fourth contribution in fact seems to be related to the same typical LIBS signals of lead (for more details, refer to the second pure contribution of LIBS outcomes before the data fusion). Therefore, this specific image is

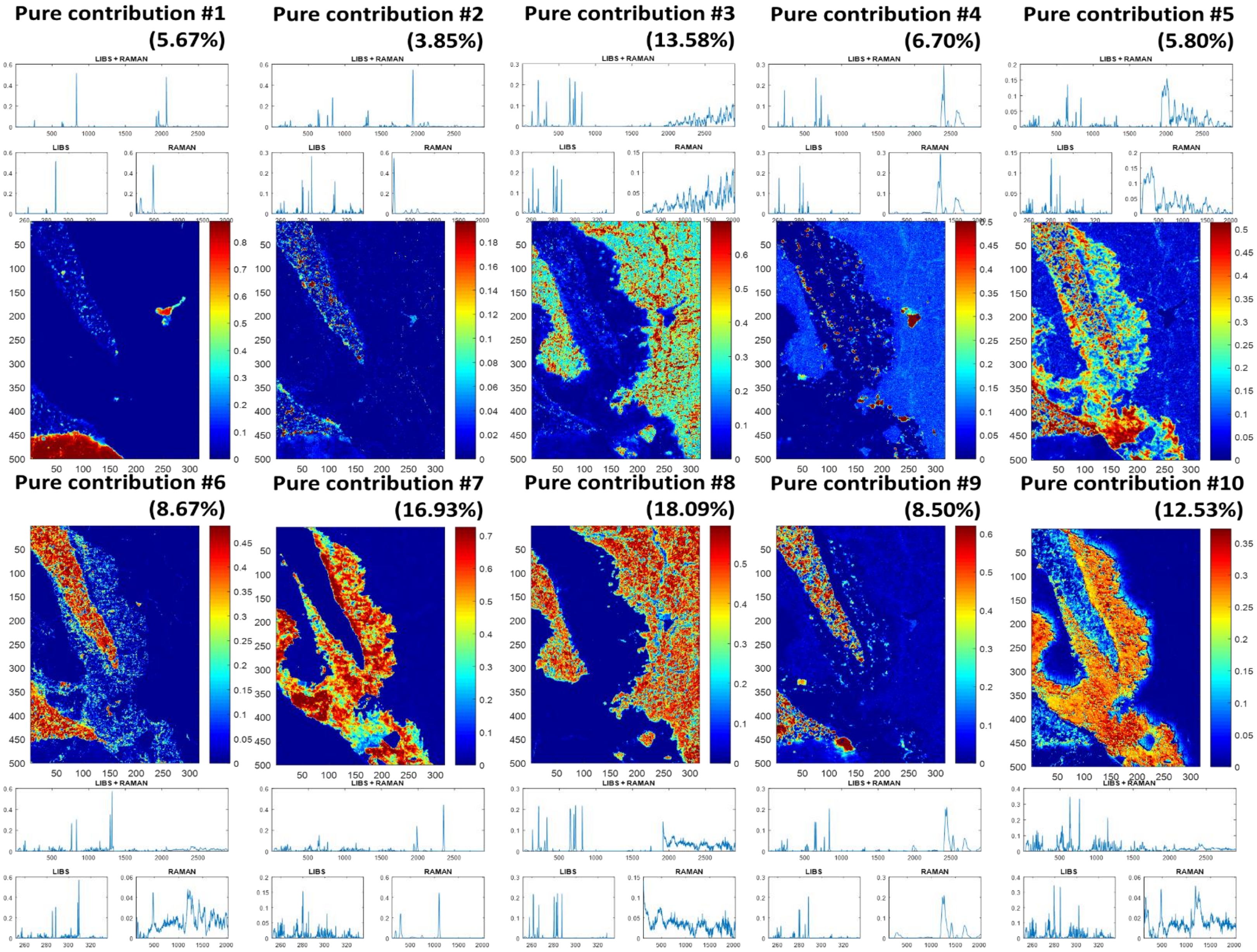


Fig. 25 – MCR-ALS results on the LIBS and Raman fused datasets of the considered mineral sample.

probably related to a particular mineral phase of the same compound previously identified, galena. Instead, ninth component shows LIBS signals that can be traced back to the aluminosilicate class (for more details, refer to the third pure contribution of LIBS outcomes before the data fusion). The most interesting aspect of these two maps is that, compared with LIBS and Raman spectra taken separately, here it has been finally possible to identify some pure components that were observed in Raman results, but that could not be recognized from a chemical point of view. In addition, fusing the two datasets, it has been possible to observe some areas that were highlighted by the use of exclusively Raman spectroscopy and that instead seemed to be invisible to the LIBS instrument. Therefore, in conclusion, it is undeniable that a data fusion is a mandatory approach to be used from a general point of view, with the purpose of generating better and more interesting details for investigational purposes in analytical chemistry.

4.5. General conclusions in the framework of the use of chemometrics applied to LIBS analysis and future perspectives

As vastly stated, LIBS imaging is clearly a very important spectroscopy that is obtaining an always increasing interest in many research areas. Nevertheless, chemometrics shows still limited applications to this instrumental data analysis. As described, a relevant part of this PhD project has focused on this aspect, trying to overcome the use of the routine approaches in LIBS analysis and find new ways to exploit this kind of spectral information. Particularly, the main goal described in this chapter is related to the application of an interesting analysis pipeline able to exceed some of the main problems related to LIBS. Due to the fact that an enormous quantity of data is easily generated, it is not always possible, or neither recommendable, to work with the raw data. Therefore, find a way to select the most important and purest information (from the spectral and spatial points of view) is mandatory, in order to obtain adequate outcomes. SIMPLISMA has been also in this case used with the purpose of accomplishing this complicated task. It is important to understand the fact that this algorithm has been chosen due to its particular benefits. In fact, SIMPLISMA is based on the selection of the purest information and not, compared with other techniques, values such as the total explained variance. This is a cardinal point in LIBS such as in other spectroscopies. Generating millions of spectra, it is plausible the fact that only a small percentage will be related to pure information, while the rest of the spectra are a combination of different elements. Also important is that, investigating a heterogeneous matrix such as, for example, a mineral, some pure components will be present in small and very specific areas of the data cube. Therefore, the use of the total explained variance would probably lead to the loss of some important information. In addition, the main purpose in using this approach is not to select a priori the information in order to use a spectral unmixing analysis. It is more related to the idea of reducing the total number of spectra, but being sure at the same time to keep the whole heterogeneity that can be related to the original dataset, an aspect that has not to be underestimated, to obtain at the end reliable results. Another concept related to the use of this strategy is that, from a computational point of view, the calculation of the results will be faster, due to the reduced amount of used data. Clearly, this kind of approach might be used for other spectroscopies. Therefore, an important aspect that has to be considered is the idea of implementing the proposed data analysis pipeline in order to use it for other instrumental responses. Another important point of this chapter is obviously the data fusion approach. In fact, as described, a further good aspect of LIBS is that this device can be used to obtain simultaneously different spectral ranges responses. The same strategy can be applied to the different datasets, in order to reduce the quantity of data and obtain a general interpretation from a chemometric point of view of the chemical complexity of the investigated sample. Using the information related to different spectral ranges is fundamental, as previously highlighted. LIBS can be used to drive an easier interpretation of the spectral information related to other spectroscopies, due to for example a limited knowledge of the corresponding spectral data (e.g., PIL phenomena). At the same time, it can be possible the contrary. In fact, despite the very interesting spectral features related to LIBS, it is not always easy to identify some specific mineral phases using this technique. In addition, in some cases, LIBS cannot extract the signal from some areas of the sample. So, fusing this spectroscopy with other responses (e.g., Raman spectroscopy), it can be possible to deep the total amount of information that the operator can obtain, compared with the use of only one or the other dataset. Here in this manuscript and during this PhD thesis, only two different spectral responses were fused to the ones obtained with LIBS instrument (i.e., PIL and Raman). Naturally, it would be interesting to use the same pipeline applied to new data coming from further spectroscopies and clearly, use the same data reduction and fusion procedures not only to LIBS, but extend this idea to other fields of analysis.

CHAPTER 5

5. SPECTRAL AND SPATIAL FUSION STRATEGIES: HOW TO COMBINE THESE TWO FUNDAMENTAL HYPERSPECTRAL IMAGING INFORMATION

5.1. Spatial information: the importance of using it and what is the best strategy to apply

Finally, this chapter is dedicated to another fundamental aspect that until now has not be really taken into consideration in the present manuscript. As discussed in the introduction, one of the main limitations related to the use of chemometrics in the framework of the hyperspectral imaging is that from a general point of view, no matter the used technique (e.g., PCA, MCR-ALS, PLS-DA, etc.), an intermediate step in which the three-dimensional data cube is unfolded in its corresponding two-dimensional dataset is required. Naturally, this procedure leads to the complete loss of the spatial information related to the investigated sample. This represents a real problem related to the use of this kind of data. In fact, it is undeniable that using only the spectral information, but not the spatial details related to an image, is a very big limitation in any data analysis. Particularly, imaging spectroscopy is obtaining an increasing importance in many research areas. The modern instruments can acquire very interesting hyperspectral images made of thousands, hundreds of thousands or even millions of spectra related to not only spectral, but also and mainly, important spatial information. Nevertheless, if it is not possible to deal directly with the original data cube, these details cannot be really investigated. Different chemometric approaches have been exploited in the last years, but they are unfortunately almost always focusing only on the spectral information. Some methods in which additional steps are used during the analysis with the aim of using the spatial information were proposed [17,96,103,209]. Nevertheless, in order to integrate the spatial information, these methods involve the use of particular constraints and/or the observation of only one pixel and its neighborhoods per time, which will lead to a longer and less fluent analysis. On the other hand, one particular algorithm is nowadays in the spotlight regarding the concept of extracting the spatial information from the studied hyperspectral image, before any further analysis, and so before the unfolding step. This approach is based on the use of wavelet transform that from a general point of view is a digital signal processing [198,217]. The concept behind the idea of using this algorithm is here briefly described. In particular, it is important to highlight the fact that over the years, many improvements in the use of wavelet transform were accomplished. Despite this, in the present manuscript will be taken into consideration only one particular kind of wavelet transform algorithm, the 2-D Stationary Wavelet Transform (SWT 2-D), which peculiarities have been discussed in Chapter 1. Moreover, this particular kind of wavelet shows very interesting features when linked to hyperspectral imaging, leading to a real extraction of spatial features, as it will be better described during this chapter. From a technical point of view, this algorithm uses some filters that extract the frequency contents of the considered signal. In this way, four distinct sets of wavelet coefficients can be obtained. Namely, they are the approximation (A) coefficients, and the horizontal (H), vertical (V), and diagonal (D) detail ones. The particularity of this algorithm is that it can be used directly on the image, without the necessity of unfolding it. In this way, the extracted details will be genuinely related to the spatial information and not to the spectral one. So, once that this part of the data is finally obtained, it is possible to merge it with the initial dataset in order to effectively observe simultaneously the information coming from both the spectral and the spatial details of the original matrix, which can be at this stage unfolded. Another important aspect regarding wavelet transform is that different families can be used, each of them related to particular signal decompositions [204]. By way of example, some of the most important families that are nowadays used are the Daubechies (the most commonly used), the biorthogonal and reverse biorthogonal wavelets (which are very interesting in the framework of image analysis), and the Gaussian wavelets. For informational purposes, in this manuscript at first it will be introduced a general description regarding the use of SWT 2-D. This method has been investigated in a first attempt using PCA, in order to show the effective necessity of using not only the spectral part of the data, but also the spatial one, and how wavelets can be used for this purpose. The corresponding outcomes have been published in Talanta, Volume 224 (2021) [218]. Then, some further ideas related to the use of wavelet transform will be introduced and discussed in the classification framework.

5.2. Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images

5.2.1. General aspects using the SWT 2-D for the spectral and spatial fusion in the framework of hyperspectral image analysis

So far, the main aspect of this chapter has been the one of explaining the necessity of using the right approach in order to exploit simultaneously both the spectral and the spatial information related to the same hyperspectral image. It has been pointed the attention on the use of a specific kind of approach based on the wavelet transform. By way of example, considering a complex sample, it is reasonable the possibility that some specific compounds can be not easily observed in the heterogeneity of the specimen, due to different reasons. For example, a specific component can be present in very small concentrations, and so hardly observable within the rest of the spectral information. Without any doubt, in this kind of scenario, having the possibility of using not only the spectral information, but also the spatial one is fundamental in order to observe also these additional compounds. Nevertheless, it is important to consider an aspect of the wavelet transform. In fact, wavelets are represented by very complex signals that are not always easy to be used correctly and so, interpreted. By way of example, the aforementioned extractable details (approximation, horizontal, vertical and diagonal coefficients) are all related to orthogonal and so, not correlated information. Furthermore, depending on the chosen decomposition level, it is possible to extract always more and more details from a spatial point of view. This means that clearly, it is possible to dig deep into the information related to the studied hyperspectral image, but at the same time, an increasing amount of data will be generated and so the interpretation will be progressively more complicated. It is important to consider these aspects before to use the wavelet transform, in order to select the right approach able to, despite the enormous amount of generated data, extract the useful information to finally find new interesting results. This is the main reason why some of the outcomes obtained during this PhD are based on the use of, as it will be also shown in the following published work, simulated images. This aspect is related to the fact that to get a better and easier interpretation of the obtained results, a general knowledge of the real composition of a given sample is required. In other words, using a simulated dataset means that the operator knows a priori all the information related to the structure of the matrix, and any interpretation error can be obtained. Furthermore, it is important to consider the reason why the first work focuses on the use of PCA. Considering again the complexity of wavelet results, the choice of using this exploratory analysis is evident. As stated, PCA is clearly one of the most exploited chemometric approaches over the others. So, use this kind of algorithm can lead to interesting results in order to better understand how wavelet transform can be used to obtain more interesting results in hyperspectral imaging. Also important is that PCA is based on the total explained variance of the information related to the studied sample. The fact that the different PCs are orthogonal among them, such as the different extractable coefficients obtained using the wavelets, is another interesting aspect. In this way, it is possible to skim the data, and so try to understand which are the most important factors related to the SWT algorithm to consider in the analysis.

Talanta 224 (2021) 121835

Contents lists available at ScienceDirect

Talanta

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

Short communication

Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images

Alessandro Nardecchia, Raffaele Vitale, Ludovic Duponchel *

Univ. Lille, CNRS, UMR 8516 - LASIRe – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, F-59000, Lille, France

ARTICLE INFO

Keywords:
Hyperspectral imaging
Spectral and spatial fusion
Stationary wavelet transform 2-D (SWT 2-D)
PCA

ABSTRACT

Nowadays, it is clear that there is an increasing importance in spectroscopic imaging in all fields of science. Obviously, one bulk analysis can no longer be satisfactory, as the interest focuses more on the chemical nature and the location of the compounds present within a given complex matrix. This is, evidently, due to the fact that for a more comprehensive exploration of complex samples, one single acquired hyperspectral data cube can provide both spectral and spatial information simultaneously. Although many techniques were proposed by the chemometric community in explorations of these specific datasets, unfortunately, they are almost always focusing on spectral information, even if chemical images were ultimately observed. In other words, spatial information is not well exploited, and therefore lost during the actual chemometric calculation phase. The goal of this short communication is to present a very simple and fast spectral/spatial fusion approach based on 2-D stationary wavelet transform (SWT 2-D) which is able to improve the obtainable information, compared with a classical data analysis, in which the spatial domain would not be considered nor used.

1. Introduction

Nowadays, hyperspectral imaging is a powerful tool. It is also out of the question how hyperspectral image analysis is broadening the horizons in different domains. The principle behind this technique is the acquisition of the whole referring spectrum for every single pixel of the image. This means each pixel is a column vector whose dimensions are equal to the number of spectral bands. As a result, the final data cube will lead to a data set of several thousands of spectra or even more, which allows a new and much deeper investigation of the sample. Rapidly, the interest in this discipline has been spread in many fields of analytical chemistry. For instance, food quality and control [1–5], and other branches have investigated the use of hyperspectral imaging for their purposes [6–13]. Furthermore, this technique has been applied for medical tasks too [14–16], in which hyperspectral images were used mainly for tumour diagnostics [17–20]. Certainly, the great interest that image analysis is obtaining is owing to the continuous overcomes of its limitations, which in turn will constantly leading to the analysis of more complex and meaningful matrices obtained by various kinds of experiments and spectroscopic techniques. On the one hand, this development leads to the possibility to obtain more details of the data set from the spectral perspective views, but on the other hand, the spatial domain remains a non-used part of the global information that can be extracted from a hyperspectral image. Nowadays in fact, the main studies on images are closely related to the use of multivariate statistical methods such as Principal Component Analysis (PCA) [21–23] and Multivariate Curve Resolution – Alternate Least Squares (MCR-ALS) approaches [24–30]. The peculiarity of these methods is that they are based only on the exploration of spectral contribution of the selected data set, no matter the dimensions of the matrix. In the meantime, the widely accepted tools by the scientific community, developed with the goal of exploring images exclusively, as the multivariate image analysis (MIA) [31,32] show the same issue. In fact, an essential step in which the acquired three-dimensional data cube is unfolded into a two-dimensional data set in order to be analyzed is always involved in the classical approaches for spectroscopic image analysis. This procedure leads to a completely absence of the spatial information correlated to a particular pixel and its neighbourhoods. Then, it is manifest that the importance of these details should not be underestimated. In fact, with the use of powerful instruments of today and the growing interests in multicomponent matrices, certain spatial information could lead to obtaining a deeper understanding in many research fields. Clearly, with the goal of exploring a complex matrix deeply, it would be interesting to observe not only the part related to the spectra, but also the new details from the

* Corresponding author.
E-mail address: ludovic.duponchel@univ-lille.fr (L. Duponchel).

https://doi.org/10.1016/j.talanta.2020.121835
Received 20 September 2020; Received in revised form 25 October 2020; Accepted 27 October 2020
Available online 30 October 2020

spatial point of view and merge the two parts of information together during the data analysis stage. In order to prove it, different procedures in which some additional steps are used during the analysis were investigated, as shown in few works [33–36]. Despite this, some limitations still exist. Particularly, in order to integrate the spatial information, these methods involve the use of particular constraints and/or the observation of only one pixel and its neighbourhoods per time, which will lead to a longer and less fluent analysis. Particularly, one algorithm, the wavelet transform [37,38], is in the spotlight regarding the importance of the spatial information nowadays, and its peculiarities have been recently investigated [39–41]. The main con here is that till now only a small part of the wide information provided by the use of the wavelet is used in order to merge the spectral and the spatial features together of an image.

The aim of this work is to overcome this lack in the use of both spectral and spatial information together within a unified analysis. In order to achieve this, an interesting and new procedure that fuses spectral and spatial details is shown. In detail, 2-D stationary wavelet transform (SWT 2-D) is applied to the raw data cube with the intention to enhance the whole extractable information from a matrix to deepen its understanding. The goal of this short communication is to broaden the horizons of new ways to analyse hyperspectral data cubes by fully exploiting all their dimensions.

2. Material and methods

2.1. 2-D stationary wavelet transform algorithm and spectral/spatial data fusion approach

In this work, stationary wavelet transform algorithm (SWT) has been exploited to develop a new approach for the simultaneous fusion of spectral and spatial data. From a technical point of view, SWT uses particular low- (*g*) and high-pass (*h*) filters in order to extract the frequency contents of the considered signal (i.e. its identity) and obtain four distinct sets of wavelet details, namely approximation (A), horizontal (H), vertical (V), and diagonal (D) coefficients [37], respectively. The SWT mechanism is shown in Fig. 1. Another important aspect of the SWT algorithm is that various wavelet families can be used, with the aim to better fit the type of signals to be explored [42,43]. By way of example, the first and simplest one, the well-known Haar wavelet, which is represented by a discontinuous and step-size function. Readers who are interested are invited to read other works specifically dedicated to this concept [44].

To illustrate the proposed fusion approach, it could be useful to introduce a scheme of the single algorithms used for the investigation of the hyperspectral image. Fig. 2 shows the well-known PCA [45] used in the framework of spectroscopic imaging. As highlighted in the *Introduction*, the limit of this method is the obligated image unfolding step before the analysis, leading to the loss of spatial information contained in the data cube. In other words, the spatial information is not used during this classical analysis. As in many other multivariate methods, only a reshaping procedure of the final results being done to finally produce an image representation. Differently, the proposed fusion approach based on wavelet algorithm (Fig. 3) operates directly on the non-unfolded image i.e. the raw data cube. Especially, as shown in the scheme, SWT 2-D is applied to each single wavelength image (from $\lambda = 1$ to $\lambda = n$) of the cube and decomposes the whole information into the contribution of the four different coefficients (approximation, horizontal, vertical and diagonal details). The objective of this first step is, thus, to capture spatial information. Once the calculation is finished, all the single results will be anew augmented and four different hyperspectral images of the same dimensionality of the original data cube are obtained. Finally, the original data set is fused with the wavelet results to obtain an augmented hyperspectral image. Then a multivariate data analysis method such as PCA can be conducted on this new data cube after an unfolding step. Clearly, the use of this procedure could lead to the idea that the problem remains unsolved since there is always an unfolding stage. Despite of this, the spectral and spatial information were intrinsically included in the same augmented data cube after applying this method.

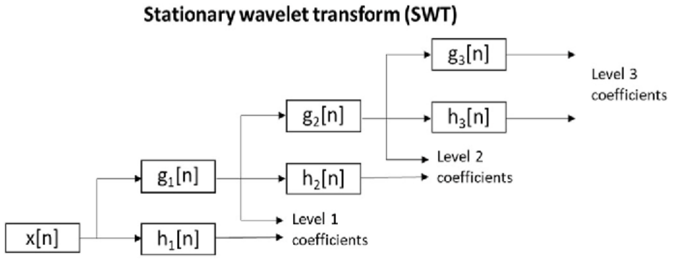


Fig. 1. Stationary wavelet transform (SWT) mechanism scheme. *x[n]* is the signal analyzed by the use of the wavelet algorithm. $g_m[n]$ and $h_m[n]$ are respectively the low- and high-pass filters, where *m* represents a particular decomposition level.

It is fair to point out that, despite the existence of different wavelet families, that were also investigated during the working line of this project, good results were obtained by the use of the most common wavelet family called Daubechies, particularly the Daubechies-1, which is the same as the Haar wavelet, the simplest wavelet [38]. As an initial step, the second decomposition level was used to investigate the results. Further levels were then used, in order to extract more information. The used approach in order to explore the further information carried by the fusion of the original spectra with the extracted wavelet coefficients can be summarized into simple and ordered steps, as following:

1) The original data cube was analyzed by the use of the stationary wavelet transform in order to extract the information regarding A, H, V, and D coefficients. As shown in Fig. 3, the SWT method was applied to each image at a given wavelength allowing to generate 4 other images of wavelet coefficients for the first level of decomposition. It is important to underline the fact that, depending on the chosen decomposition level, the total number of coefficient images changes, hence the total number of the variables in the augmented data cube. For instance, the second decomposition level will lead to the extrapolation of more blocks, in such order: A1, H1, V1 and D1 corresponding to the first decomposition level and A2, H2, V2 and D2 for the second one. For the third decomposition level, also A3, H3, V3 and D3 coefficients will be generated and so on. In order to obtain an easier interpretation of the loadings during the analysis, the blocks were organized in this order: firstly, the spectra and then each single coefficient, lastly grouping together the various selected decomposition levels. As an example, considering a second decomposition level, then the variables in the matrix will be organized in spectra, A1, A2, H1, H2, V1, V2, D1 and D2. The augmented data cube was unfolded for further processing. Notice that, regardless of this unfolding step, the final data structure contains both spectral and spatial information, the latter was encoded by the wavelet coefficients.
2) Naturally, these two series of data (spectra and wavelet coefficients) present different properties and so, they need to be pre-processed separately and differently. Especially, while the spectra were pretreated by the use of the mean centering, wavelet coefficients were auto-scaled [46].
3) An additional normalization step is necessary, in order to give the same importance to spectra and wavelet coefficients. This procedure is necessary due to the fact that the variance of a particular block could result bigger and so bias the PCA results. In order to avoid this problem, Normalization by Frobenius norm, which is defined as the square root of the sum of the absolute squares of the matrix elements, was used to obtain a unit norm for each block [47]. In detail, the wavelet coefficients (A1, A2, …, H1, H2, …, V1, V2, …, D1, D2, …)

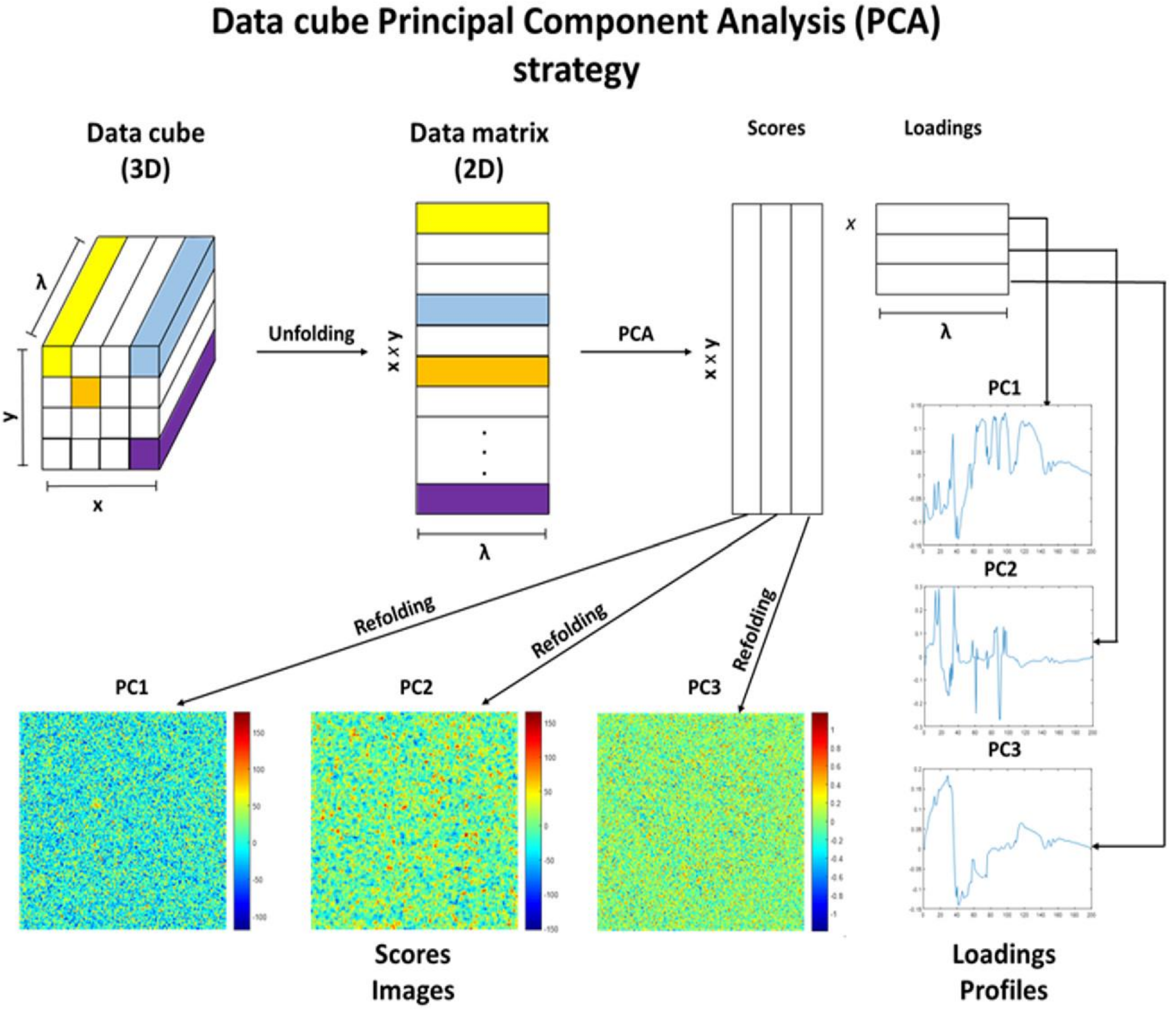


Fig. 2. Typical PCA procedure scheme. In order to apply it, a data cube needs to be unfolded in a two-dimensional dataset.

have been separately normalized before merging them together into only one block.

4) Last step was the PCA on the unfolded augmented data cube matrix. An important aspect to be remarked is that to obtain an easier interpretation regarding to the loadings in the paper, different colours were applied, coefficient per coefficient. For this reason, while the spectra-related loadings are black, the approximations are red, the horizontals blue, the verticals green and the diagonals cyan in the showed results.

It should be noted that PCA is not an integral part of the fusion approach. It has been employed in the present work just because it is generally used for unbiased exploration of spectral data. In other words, other chemometric exploration techniques could be, of course, applied to the fused data cube, but that is beyond the scope of this short communication.

2.2. Dataset description

The dataset used to show the obtainable supplementary information by fusing together the spatial and the spectral domains corresponds to a simulated hyperspectral image. The reason to use an artificial image lies in the fact that nowadays the use of wavelets in hyperspectral imaging is still a new, limited knowledge open field with thousands of possibilities to be explored. Knowing exactly the real nature and structure of the image would bring to an easier interpretation of the obtained results. Knowledge of the ground truth is indeed a good way of objectively assessing what such an approach can bring. It should be also emphasized that the aim of this work and therefore of this short communication is, above all, to present the concept with preliminary results that will be extended in future work. In detail, the dataset was obtained by a linear combination of three different images, which were used to generate the hyperspectral data cube for the exploration (Fig. 4a). Each image was built to have a unique geometrical shapes, and a specific Gaussian-distribution spectral intensity and domain. In detail, the first image is composed by three concentric circles and three rectangles with different orientations. The second is made of seven circles of different diameters. The third and last shows an oblique cross and four oblique rectangles. Furthermore, the geometric figures of the first two components show different intensities in order to obtain a higher variability in the spectra. By a spectral point of view, the two first components are represented by two Gaussians of the same intensity that don't overlap. Differently, the third spectral contribution corresponds to a Gaussian that is located between the first ones. In particular, with the aim of showing the interesting contribution carried by fusing spectra and wavelet coefficients, the third pure image is represented by a lower spectral intensity compared with the other two components, leading to hiding its contribution. In detail, when boosting the spectral noise, it is clear that the third component is invisible by a visual point of view, as confirmed also by the observation of the global integration image in Fig. 4b. The reason to stress the spectra of this dataset was intentionally chosen with the aim to demonstrate whether the use of wavelets could add some new

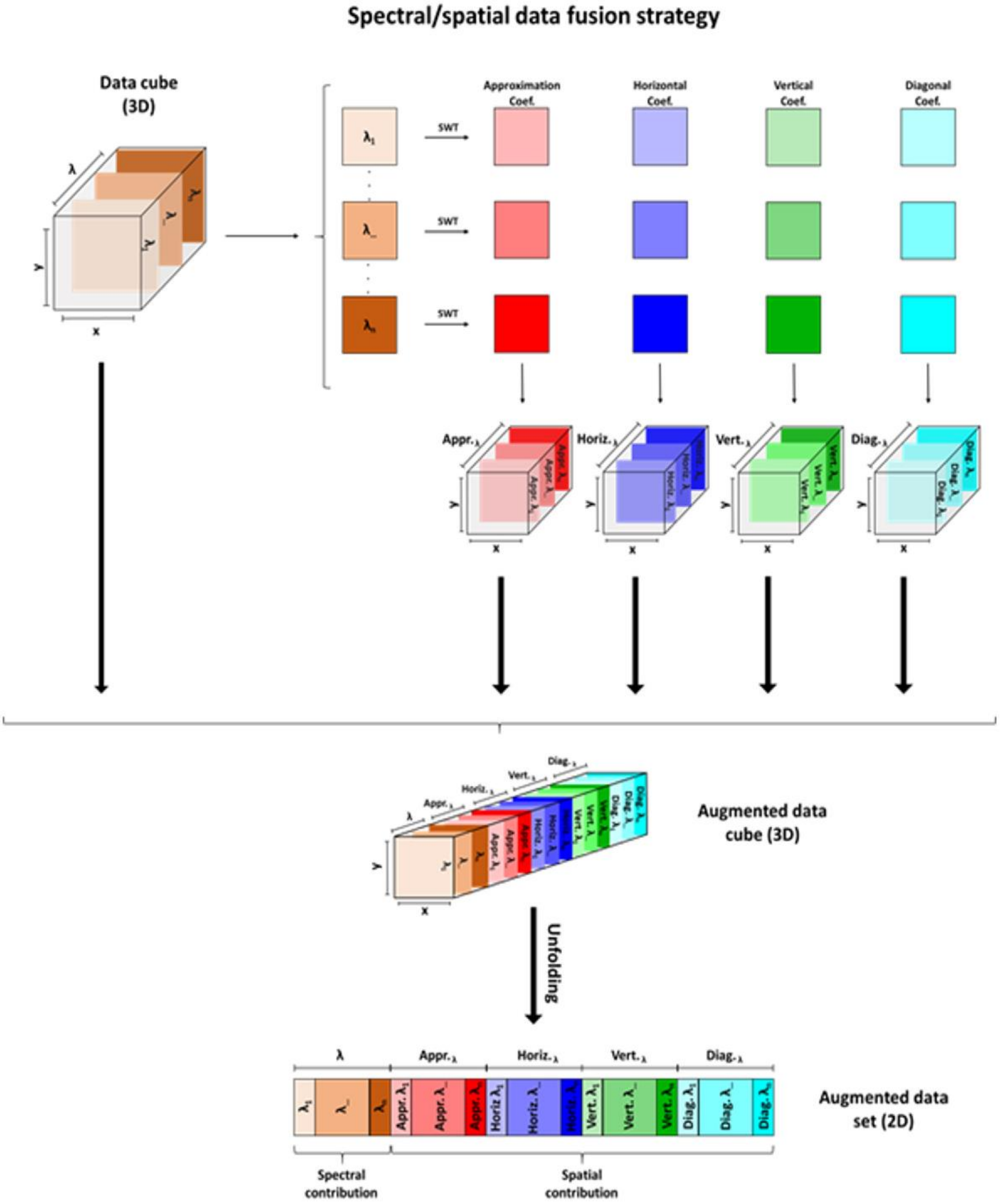


Fig. 3. Spectral/spatial fusion approach scheme. At first, every single variable is analyzed with the SWT algorithm. Then, the five (original spectra, approximation, horizontal, vertical, and diagonal coefficients) data cubes are augmented in a new, extended image that will be unfolded in order to use the PCA.

information obtained from the spatial part of the hyperspectral image, leading to a deeper knowledge of the dataset. The size of the image is 208 pixels by 208 pixels, for a total of 43,264 spectra and 100 spectral variables.

3. Results and discussion

As already discussed above, investigating the global integration image is not possible to obtain any information regarding to the third pure component, due to the fact that from a spectral point of view, boosting the noise level, its relative variance is enormously lower compared with the other two. More interesting is the fact that neither PCA, one of the mainstay tools used in exploratory analysis, seems to show the ability of extracting hardly any useful information regarding to the contribution of this component. As shown in Fig. 5, in fact, having a look at the eigenvalues scree plot and particularly at the various principal components (PCs), both scores and loadings, the whole information is divided into the first two principal components (PC1 and PC2) with 57.23% and 13.25% of explained variance, respectively. The rest of the information is spread into the other PCs, 0.33% of the total explained variance per PC. Knowing the exact composition of the dataset, some details regarding to the third pure image can be vaguely glimpsed. Despite this, it is out of discussion that the shown details are too feeble to be considered in a real exploratory data analysis. To investigate the results obtained by the use of the proposed fusion approach, a second decomposition level of the wavelet decomposition was initially considered leading to a total number of 900 variables (100 from the spectra and 100 for each wavelet coefficient, considering the selected decomposition level). So, applying the proposed fusion approach to the same dataset, it is possible to observe some interesting aspects investigating

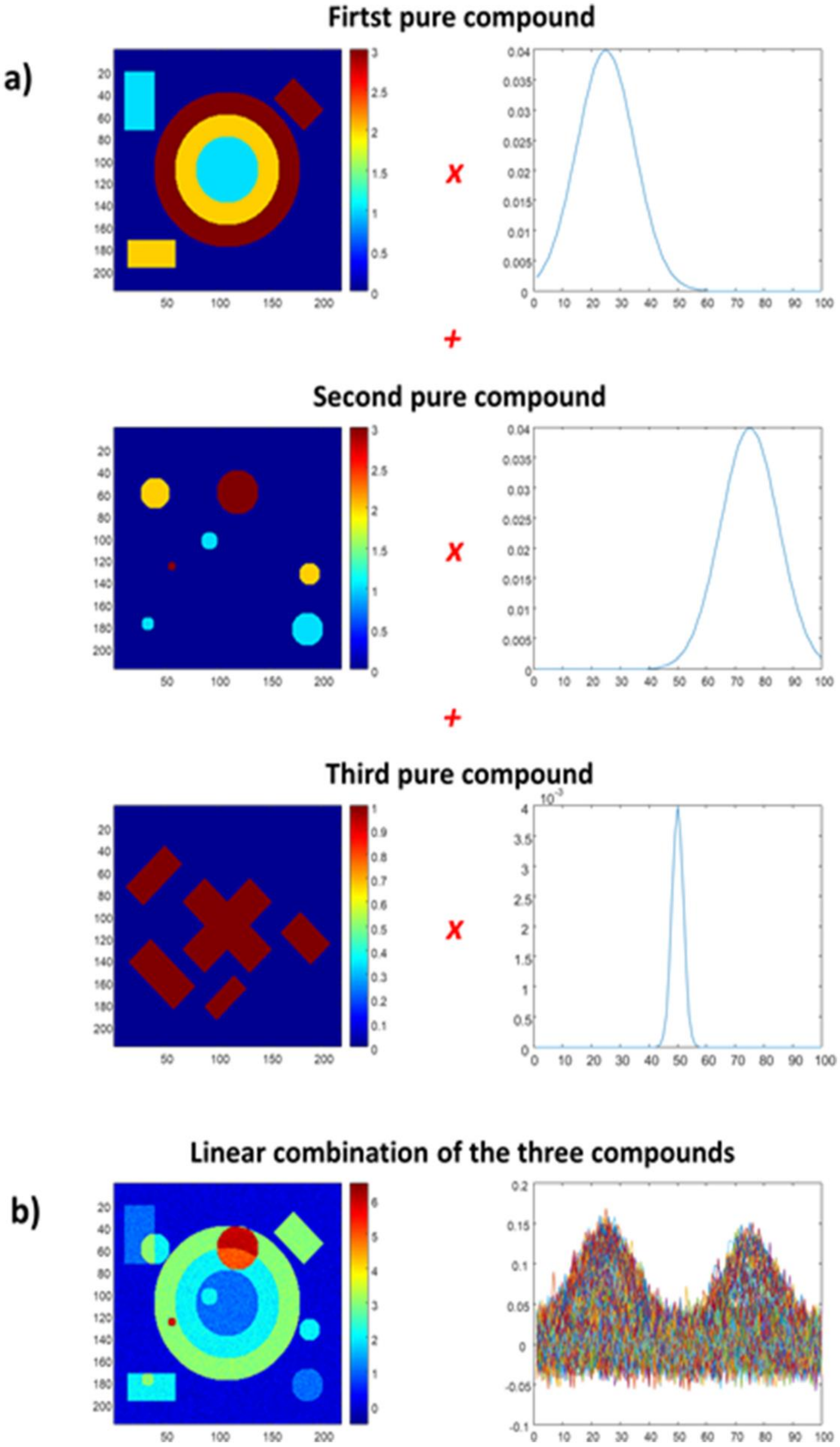


Fig. 4. A – The three simulated images and the Gaussian-distribution used for generate the artificial data cube. **B –** Global integration image and spectra of the final data cube.

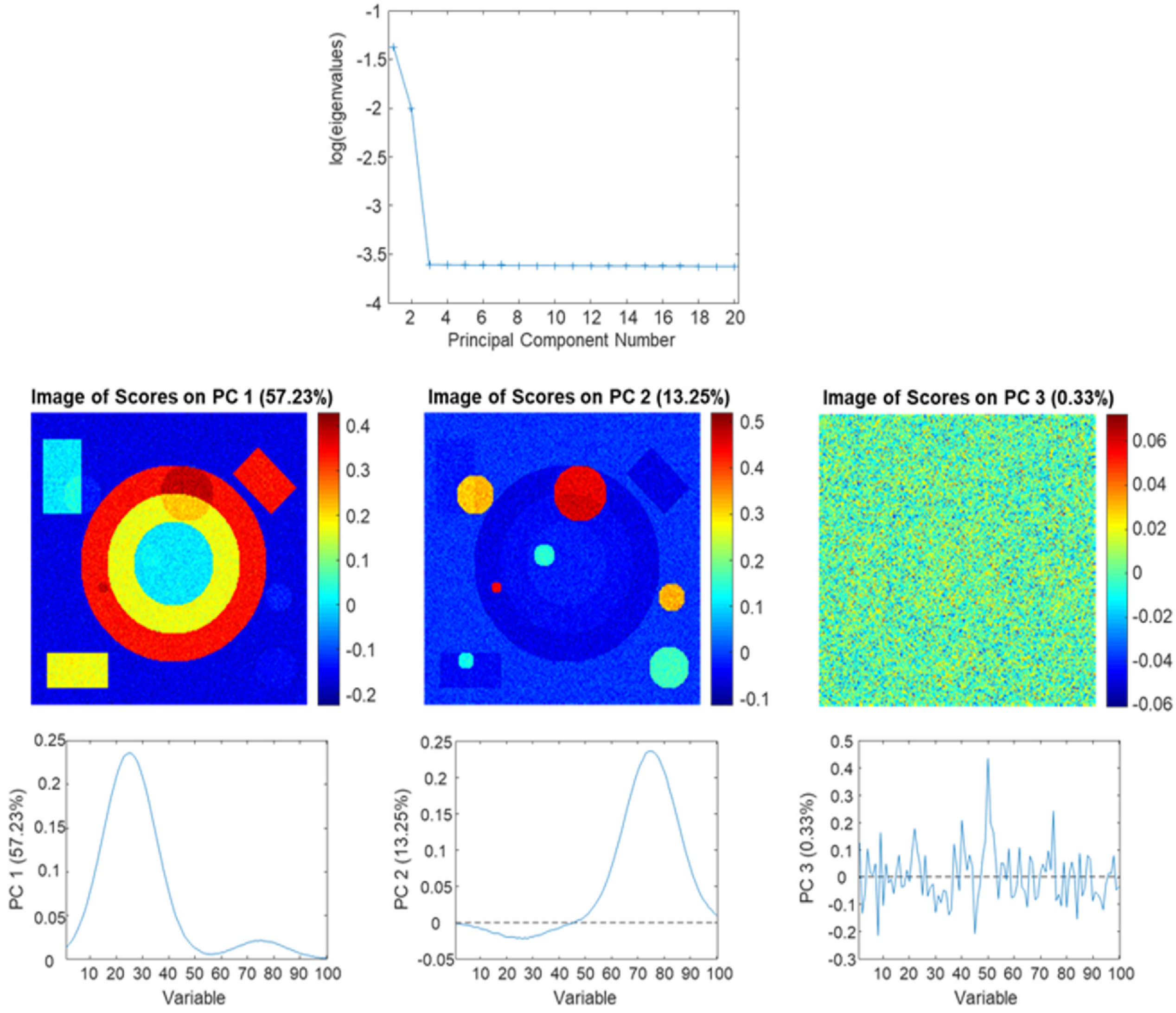


Fig. 5. PCA results of the simulated data cube, before the use of the proposed spectral/spatial fusion method.

both the scores and the loadings of the final PCA (Fig. 6). In fact, while the first two PCs show the same information obtained in the original dataset (with a total explained variance value of 15.46% for PC1 and 8.70% for PC2, respectively), digging deeper it is possible now to observe the shape of the third pure component, particularly thanks to the fact that the PCA is driven by the use of both spectral and spatial variables. Precisely, the information related to the third component is observable in PC17 (0.17% of explained variance). Investigating the loadings, it is evident that this phenomenon is because the variables related to the wavelets add an important contribution to the interpretation of the third object. Specifically, the approximation coefficients show a clear peak corresponding to the missing component, giving it the possibility to glimpse its shape from the background noise in the scores. It is also interesting to mention that the same peak shape is also observable in the variables of the second decomposition levels related to horizontal, vertical and diagonal coefficients, albeit their intensities are lower compared with the rest of the loadings. This leads to the idea that higher is the used decomposition level, better is the ability of the proposed fusion method ability to show the missing information of the original data cube. As prove of this, specific contributions of the third compound are shown in Fig. 7, comparing to the second, the third and the fourth decomposition levels. Respectively, when the third decomposition level is considered, the missing object is observable in PC22 (0.14% of the total explained variance), while for the fourth decomposition level, it is visible in PC30 (0.13% of the total explained variance). Observing the scores is then clear that the use of more decomposition levels leads to a better visibility of all contributions. Comparing this factor with the loadings, it seems obvious that most of this information is related to the approximation coefficients, whose intensities are higher compared with the horizontal, vertical and diagonal details. Finally, it is undeniable that it is possible to obtain details regarding to the third component that were not visible using the usual approach on the raw data cube. Furthermore, it is obvious that fusing the wavelet coefficients with the spectra, an enhancement of the total information is obtained and particularly the approximation coefficients can lead to gain details that are able to show aspects which could be hidden in a direct approach to the raw data cube.

In conclusion it can be assumed that while the first two components show higher loading intensities related to the spectral region, the third is associated to high-contribution loadings in the spatial (wavelet) domain. Furthermore, it is noteworthy that, when merging the spectral and the spatial information, it is necessary to investigate more PCA components. This is explained by the fact that wavelets carry a large amount of orthogonal, thus, independent information. It is also true that most of the information is related to the approximation coefficients, as mentioned above. Removing the rest of the wavelet information could lead to the investigation of less PCs in order to retrieve the missing component (results not shown) . Despite this, due to the fact that the approach is applied for an exploratory analysis, it has been decided to use all the wavelet coefficients in order to obtain a global view of the data obtained by the contribution of both spectral and spatial information. Last but not the least, in order to validate the method, it has been

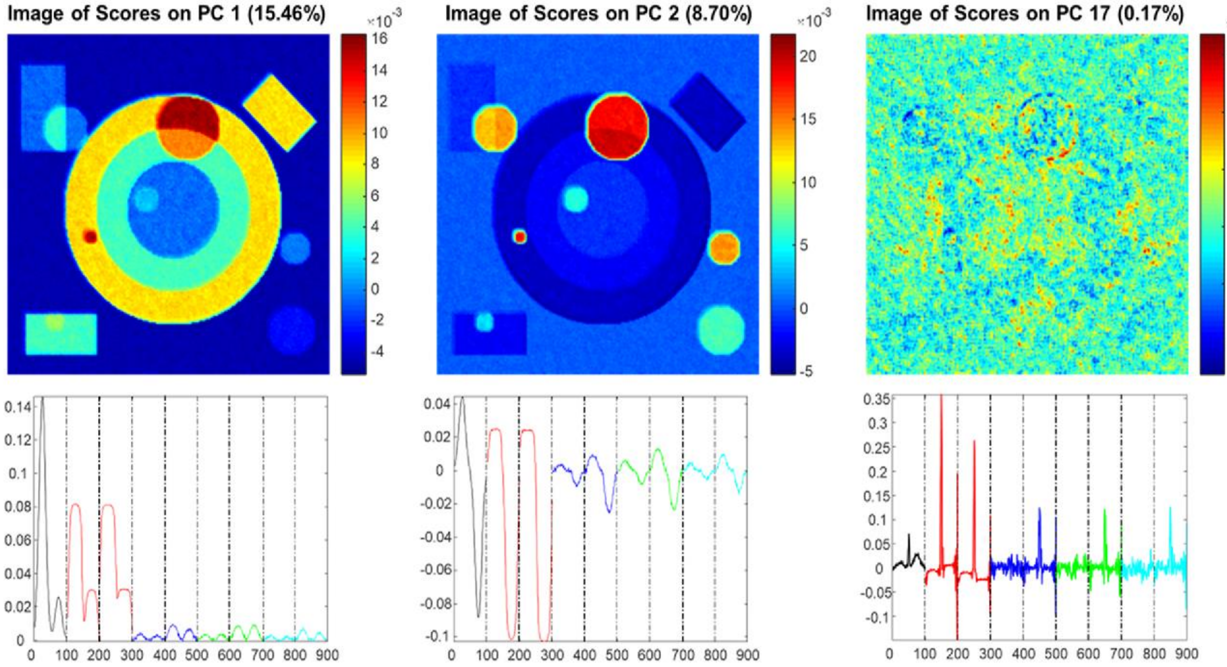


Fig. 6. PCA results of the simulated data cube, after the use of the proposed spectral/spatial fusion method. Particularly, regarding the wavelet coefficients, a second decomposition level has been applied. In the loadings, black variables are related to the original spectra, red to the approximation, blue to the horizontal, green to the vertical and cyan to the diagonal coefficients. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

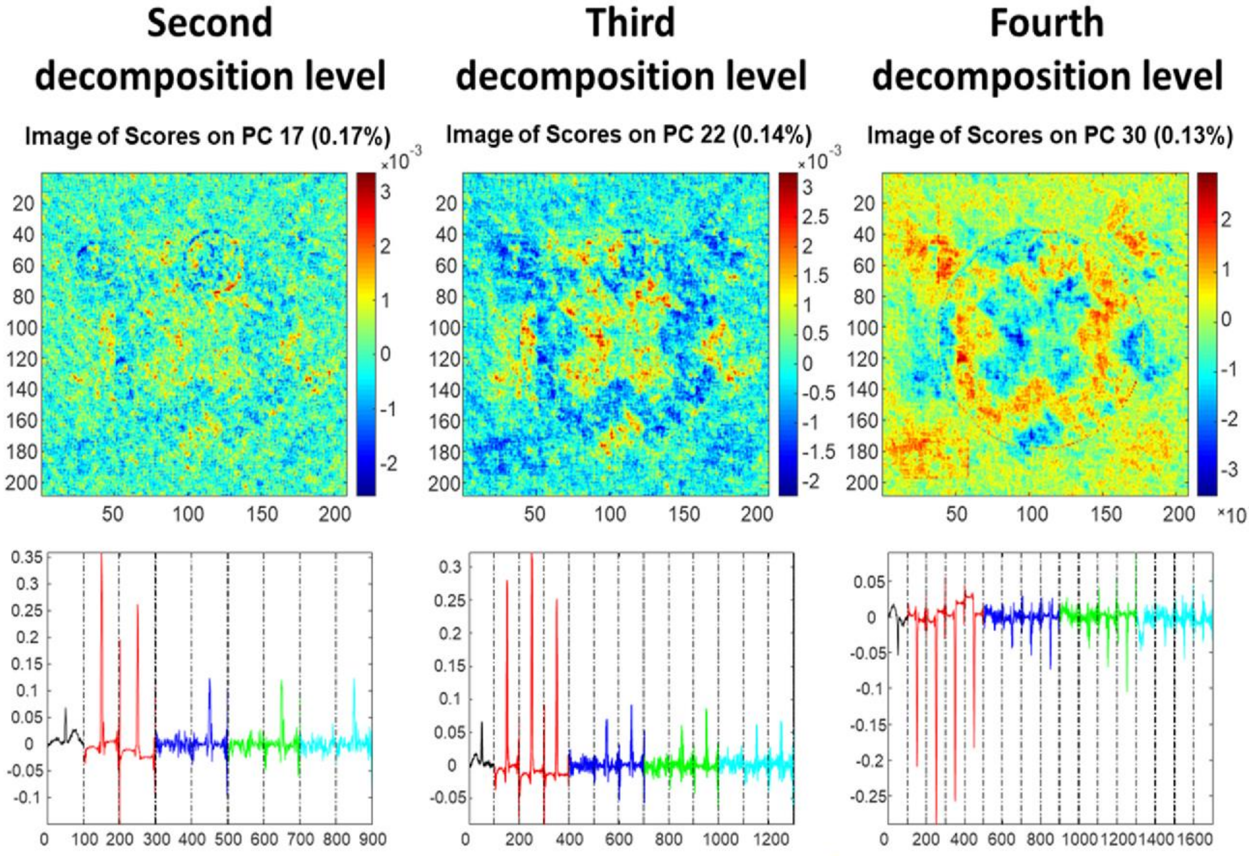


Fig. 7. Comparison of the third component when, respectively, a second, third and fourth decomposition level is used to extrapolate the wavelet coefficients. In the loadings, black variables are related to the original spectra, red to the approximation, blue to the horizontal, green to the vertical and cyan to the diagonal coefficients. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

compared with the use of the Bharati-MacGregor approach [34,48]. As expected, being this another method applied for the extraction and the use of the spatial information of a data cube, it is possible to observe the third component. Comparison of the two approaches proved that the contrast obtained by the use of the present method is higher with the Root Mean Square (RMS) contrast [49], particularly when more decomposition levels were selected (outcomes not shown) . Hence the Bharati-MacGregor approach can be used for the extraction of the spatial information, the method in this work seems to show a good alternative for the investigation of hyperspectral images.

4. Conclusions

The presented short communication was developed with the aim of showing the possibility to fuse together spectral and spatial information of a spectroscopic image data cube and so overcome the lack in the most common standard procedures. The typical used approaches nowadays in fact show an unfolding step drawback, that implies the impossibility to use spatial information. Particularly, just including in the default exploring procedure (typically PCA) the use of the wavelet algorithm to extract the spatial-related information seems to lead in obtaining new interesting hyperspectral image details. Particularly, the results of this work show that the use of the augmented dataset (obtained fusing spectral variables and wavelet coefficients) leads to retrieve new information that correspond to weak spectral fingerprints, but are related to strong spatial details. As shown in the presented case, this approach can be used to reveal new details that are difficult to be noticed relying only on the spectral information, thanks to the ability to extract spatial information and the well-known denoising ability of wavelets. To conclude, it can be asserted that using this spectral and spatial fusion is possible to show new information for a deeper interpretation of a spectroscopic imaging data set. Despite this, this work represents only the first step for an alternative methodology that could broaden the horizons of new ways to drive a more complete hyperspectral image analysis exploration with potentially multiple extensions for clustering, classification and regression methods.

CRediT author statement

All authors participated equally in this work.

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.

References

[1] D. Wu, D.-W. Sun, Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: a review — Part I: fundamentals, Innovat. Food Sci. Emerg. Technol. 19 (2013) 1–14, https://doi.org/10.1016/j.ifset.2013.04.014.

[2] A. Gowen, C. Odonnell, P. Cullen, G. Downey, J. Frias, Hyperspectral imaging – an emerging process analytical tool for food quality and safety control, Trends Food Sci. Technol. 18 (2007) 590–598, https://doi.org/10.1016/j.tifs.2007.06.001.

[3] J.M. Amigo, I. Martí, A. Gowen, Hyperspectral imaging and chemometrics, in: Data Handling in Science and Technology, Elsevier, 2013, pp. 343–370, https://doi.org/10.1016/B978-0-444-59528-7.00009-0.

[4] J. Nogales-Bueno, F.J. Rodríguez-Pulido, F.J. Heredia, J.M. Hernández-Hierro, Comparative study on the use of anthocyanin profile, color image analysis and near-infrared hyperspectral imaging as tools to discriminate between four autochthonous red grape cultivars from La Rioja (Spain), Talanta 131 (2015) 412–416, https://doi.org/10.1016/j.talanta.2014.07.086.

[5] Y.-Z. Feng, D.-W. Sun, Near-infrared hyperspectral imaging in tandem with partial least squares regression and genetic algorithm for non-destructive determination and visualization of Pseudomonas loads in chicken fillets, Talanta 109 (2013) 74–83, https://doi.org/10.1016/j.talanta.2013.01.057.

[6] R. Gosselin, D. Rodrigue, C. Duchesne, A hyperspectral imaging sensor for on-line quality control of extruded polymer composite products, Comput. Chem. Eng. 35 (2011) 296–306, https://doi.org/10.1016/j.compchemeng.2010.07.020.

[7] J. Cruz, M. Bautista, J.M. Amigo, M. Blanco, Nir-chemical imaging study of acetylsalicylic acid in commercial tablets, Talanta 80 (2009) 473–478, https://doi.org/10.1016/j.talanta.2009.07.008.

[8] P. Mishra, A. Nordon, J. Tschannerl, G. Lian, S. Redfern, S. Marshall, Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products, J. Food Eng. 238 (2018) 70–77, https://doi.org/10.1016/j.jfoodeng.2018.06.015.

[9] W. Fortunato de Carvalho Rocha, G.P. Sabin, P.H. Março, R.J. Poppi, Quantitative analysis of piroxicam polymorphs pharmaceutical mixtures by hyperspectral imaging and chemometrics, Chemometr. Intell. Lab. Syst. 106 (2011) 198–204, https://doi.org/10.1016/j.chemolab.2010.04.015.

[10] P. Tatzer, M. Wolf, T. Panner, Industrial application for inline material sorting using hyperspectral imaging in the NIR range, R. Time Imag. 11 (2005) 99–107, https://doi.org/10.1016/j.rti.2005.04.003.

[11] A.A. Gowen, Y. Feng, E. Gaston, V. Valdramidis, Recent applications of hyperspectral imaging in microbiology, Talanta 137 (2015) 43–54, https://doi.org/10.1016/j.talanta.2015.01.012.

[12] R.R.V. Carvalho, J.A.O. Coelho, J.M. Santos, F.W.B. Aquino, R.L. Carneiro, E. R. Pereira-Filho, Laser-induced breakdown spectroscopy (LIBS) combined with hyperspectral imaging for the evaluation of printed circuit board composition, Talanta 134 (2015) 278–283, https://doi.org/10.1016/j.talanta.2014.11.019.

[13] M.R. Almeida, L.P.L. Logrado, J.J. Zacca, D.N. Correa, R.J. Poppi, Raman hyperspectral imaging in conjunction with independent component analysis as a forensic tool for explosive analysis: the case of an ATM explosion, Talanta 174 (2017) 628–632, https://doi.org/10.1016/j.talanta.2017.06.064.

[14] B. Fei, Hyperspectral imaging in medical applications, in: Data Handling in Science and Technology, Elsevier, 2020, pp. 523–565, https://doi.org/10.1016/B978-0-444-63977-6.00021-3.

[15] G. Lu, B. Fei, Medical hyperspectral imaging: a review, J. Biomed. Optic. 19 (2014), 010901, https://doi.org/10.1117/1.JBO.19.1.010901.

[16] M.A. Calin, S.V. Parasca, D. Savastru, D. Manea, Hyperspectral imaging in the medical field: present and future, Appl. Spectrosc. Rev. 49 (2014) 435–447, https://doi.org/10.1080/05704928.2013.838678.

[17] S.G. Kong, Z. Du, M. Martin, T. Vo-Dinh, in: T. Vo-Dinh, W.S. Grundfest, D. A. Benaron, G.E. Cohn (Eds.), Hyperspectral Fluorescence Image Analysis for Use in Medical Diagnostics, 2005, p. 21, https://doi.org/10.1117/12.596463. San Jose, CA.

[18] S.V. Panasyuk, S. Yang, D.V. Faller, D. Ngo, R.A. Lew, J.E. Freeman, A.E. Rogers, Medical hyperspectral imaging to facilitate residual tumor identification during surgery, Canc. Biol. Ther. 6 (2007) 439–446, https://doi.org/10.4161/cbt.6.3.4018.

[19] Z. Liu, H. Wang, Q. Li, Tongue tumor detection in medical hyperspectral images, Sensors 12 (2011) 162–174, https://doi.org/10.3390/s120100162.

[20] L. Zhi, D. Zhang, J. Yan, Q.-L. Li, Q. Tang, Classification of hyperspectral medical tongue images for tongue diagnosis, Comput. Med. Imag. Graph. 31 (2007) 672–678, https://doi.org/10.1016/j.compmedimag.2007.07.008.

[21] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52, https://doi.org/10.1016/0169-7439(87)80084-9.

[22] K. Pearson, On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dubl. Phil. Mag. J. Sci. 2 (1901) 559–572, https://doi.org/10.1080/14786440109462720.

[23] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441, https://doi.org/10.1037/h0071325.

[24] W.H. Lawton, E.A. Sylvestre, Self modeling curve resolution, Technometrics 13 (1971) 617–633, https://doi.org/10.1080/00401706.1971.10488823.

[25] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemometr. 9 (1995) 31–58, https://doi.org/10.1002/cem.1180090105.

[26] 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.

[27] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods. 6 (2014) 4964–4976, https://doi.org/10.1039/C4AY00571F.

[28] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemometr. Intell. Lab. Syst. 76 (2005) 101–110, https://doi.org/10.1016/j.chemolab.2004.12.007.

[29] 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.

[30] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: a review of advanced and tailored applications and challenges, Anal. Chim. Acta 765 (2013) 28–36, https://doi.org/10.1016/j.aca.2012.12.028.

[31] K. Esbensen, P. Geladi, Strategy of multivariate image analysis (MIA), Chemometr. Intell. Lab. Syst. 7 (1989) 67–86, https://doi.org/10.1016/0169-7439(89)80112-1.

[32] J.M. Prats-Montalbán, A. de Juan, A. Ferrer, Multivariate image analysis: a review with applications, Chemometr. Intell. Lab. Syst. 107 (2011) 1–23, https://doi.org/10.1016/j.chemolab.2011.03.002.

[33] M.H. Bharati, J.J. Liu, J.F. MacGregor, Image texture analysis: methods and comparisons, Chemometr. Intell. Lab. Syst. 72 (2004) 57–71, https://doi.org/10.1016/j.chemolab.2004.02.005.

[34] F. Jamme, L. Duponchel, Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis, J. Chemometr. 31 (2017), e2882, https://doi.org/10.1002/cem.2882.

[35] J.J. Liu, J.F. MacGregor, On the extraction of spectral and spatial information from images, Chemometr. Intell. Lab. Syst. 85 (2007) 119–130, https://doi.org/10.1016/j.chemolab.2006.05.011.

[36] S. Hugelier, O. Devos, C. Ruckebusch, On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image

analysis: spatial constraints in HSI-MCR-ALS, J. Chemometr. 29 (2015) 557–561, https://doi.org/10.1002/cem.2742.

[37] G.P. Nason, B.W. Silverman, The stationary wavelet transform and some statistical applications, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, Springer New York, New York, NY, 1995, pp. 281–299, https://doi.org/10.1007/978-1-4612-2544-7_17.

[38] S.G. Mallat, A Wavelet Tour of Signal Processing: the Sparse Way, third ed., Elsevier/Academic Press, Amsterdam ; Boston, 2009.

[39] M. Li Vigni, J.M. Prats-Montalban, A. Ferrer, M. Cocchi, Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA): coupling 2D-WT to multivariate image analysis (2D WT-MIA), J. Chemometr. 32 (2018), e2970, https://doi.org/10.1002/cem.2970.

[40] P.-M. Juneau, A. Garnier, C. Duchesne, The undecimated wavelet transform–multivariate image analysis (UWT-MIA) for simultaneous extraction of spectral and spatial information, Chemometr. Intell. Lab. Syst. 142 (2015) 304–318, https://doi.org/10.1016/j.chemolab.2014.09.007.

[41] M. Ahmad, R. Vitale, C.S. Silva, C. Ruckebusch, M. Cocchi, Exploring local spatial features in hyperspectral images, J. Chemometr. 34 (10) (2020) e3295, https://doi.org/10.1002/cem.3295.

[42] J. Kovacevic, W. Sweldens, Wavelet families of increasing order in arbitrary dimensions, IEEE Trans. Image Process. 9 (2000) 480–496, https://doi.org/10.1109/83.826784.

[43] Cé Vonesch, T. Blu, M. Unser, Generalized Daubechies wavelet families, IEEE Trans. Signal Process. 55 (2007) 4415–4429, https://doi.org/10.1109/TSP.2007.896255.

[44] L. Debnath, F.A. Shah, Wavelet Transforms and Their Applications, Birkhäuser Boston, Boston, MA, 2015, https://doi.org/10.1007/978-0-8176-8418-1.

[45] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, (n.d.) 16.

[46] J.E. Jackson, A User's Guide to Principal Components, tenth ed., John Wiley & Sons, 1991 https://doi.org/10.1002/0471725331.

[47] G.H. Golub, C.F. Van Loan, Matrix Computations, fourth ed., The Johns Hopkins University Press, Baltimore, 2013.

[48] M.H. Bharati, J.F. MacGregor, in: H. McCann, D.M. Scott (Eds.), Texture Analysis of Images Using Principal Component Analysis, 2001, p. 27, https://doi.org/10.1117/12.417179. Boston, MA.

[49] E. Peli, Contrast in complex images, J. Opt. Soc. Am. A 7 (1990) 2032, https://doi.org/10.1364/JOSAA.7.002032.

5.3. The use of wavelet transform for a classification analysis (PLS-DA)

5.3.1. Introduction

Wavelet transform can be clearly coupled not only with PCA, but also with other chemometric approaches. One of the most interesting ideas investigated during this PhD has been the one of exploiting the spatial information extracted with the SWT 2-D in order to obtain better results in the framework of classification analysis, using one of the most known algorithms in the chemometric community, i.e., PLS-DA. By way of example, considering again a complex sample, it is possible that different compounds are related to very similar spectral information, but different physical shapes or even different localizations in the sample. Naturally, using only the spectral domain, these different aspects of the image would be erroneously classified into a single family. Another example can be the one in which objects that have very thin borders are acquired. In this case, depending on the used spectroscopy, there is a real chance that these borders will show a signal very close to the noise level, or that the corresponding spectra can be very similar to the ones of the hyperspectral image background. Using not only the spectral domain, but also exploiting the spatial information of the data cube would be in these cases very convenient to obtain at the end better classification outcomes. Here will be reported the first results obtained for these two different situations. Again, it is important to highlight the fact that, using this kind of algorithm to extract the spatial information, it is not always easy to give the right interpretation to the results. This is why the first shown dataset is a simulated one. In this way, it has been possible to generate a complex case, which structure is known a priori in order to give a better evaluation of how wavelets can be used for the classification analysis. Thus, it has been conceivable to rapidly investigate different wavelet families, and understand which one can be the best to be used for this kind of analysis. Lastly, for informational purposes, the present work has been recently submitted to Talanta, in order to better diffuse the found outcomes regarding the fusion of the spectral and spatial information of a hyperspectral image in the framework of the classification analysis.

5.3.2. A simulated dataset investigation

The simulated dataset is here described and shown in Fig. 26. The corresponding data cube is represented by dimensions equal to 300 pixels by 300 pixels and 300 spectral points. It is made by four distinct compounds, using two different geometric shapes. First and second components are represented by circles, while third and fourth are rectangles.

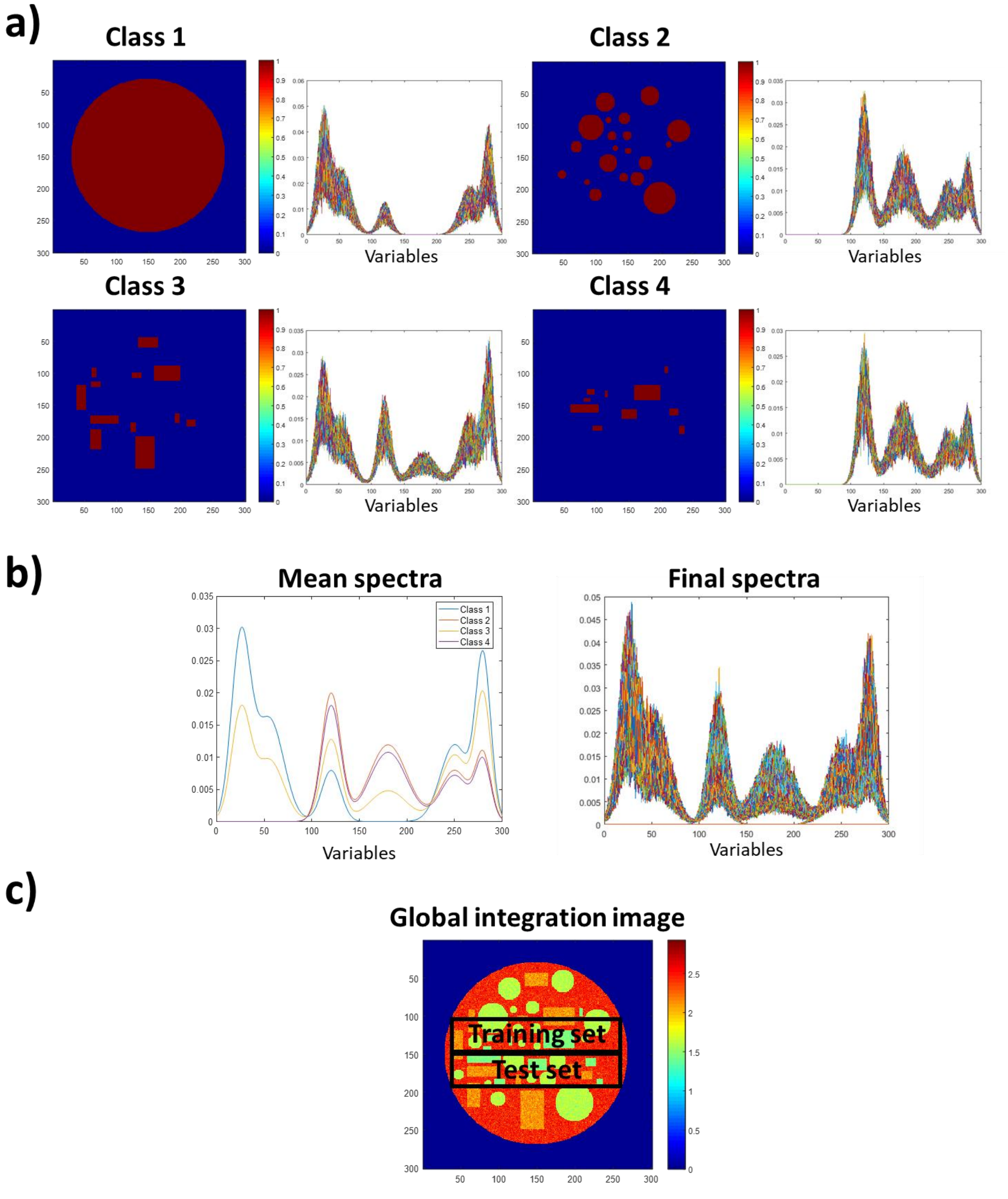


Fig. 26 – Representation of the simulated dataset. a) The four classes. b) Spectra of the four classes. c) Global integration image of the final simulated dataset.

As previously explained, in the case in which different classes (for example, from a physical point of view) are linked to similar spectral information, clearly it would be impossible using any approach to correctly classify them exploiting only the spectra. This is the reason why in the simulated dataset it is possible to distinguish specific spectral and spatial information, as reported in Fig. 26a. The first class shows very specific bands that are only partially shared with the second component. Furthermore, both of these classes are made by circles, as previously stated. Anyway, the use of a robust method such as PLS-DA would not face problems to correctly classify these two different compounds, if no other information was present. In fact, the interesting aspect that is related to the necessity of using wavelets to extract and exploit the spatial information comes out when also the other two components are taken into consideration. The third class is a linear combination of the spectral information of the already two described components. This means that using only the spectral details, in this case the operator could experience some issues in obtaining good outcomes. The only difference from the first two classes is that this one is made by rectangles, meaning that the spatial information in this case is a fundamental aspect in order to discern it from the other classes. Lastly, fourth component corresponds to spectra completely equal to the ones of the second class, except for a slightly lower signal intensity. In this case too, the only way to differentiate the two classes is that this last one is made by rectangles and not circles such as the second one. In addition, some noise has been added to the spectra in order to generate a better differentiation of the pixels and obtain an image closer to the one of a real case (Fig. 26b). Then, in order to use PLS-DA, two subregions of the final image have been selected in order to create a training set and a test set, as represented in Fig. 26c. At this point, it has been possible to use the classification method in order to compare the outcomes related to the use of only the spectra, and the simultaneous use of both the spectral and the spatial information extracted by the use of the SWT 2-D. For informational purposes, it is important to highlight the fact that the same kind of pipeline to merge the two parts of the data (spectral and spatial details) used in the previous work has been applied also here. In brief, before the unfolding step, wavelet transform has been used to extract the spatial information from the image, and then, the corresponding coefficients used as an extension of the variables merging them with the original dataset, once it has been finally unfolded for the chemometric analysis. Four decomposition levels have been selected. Despite the use of different families, reverse biorthogonal wavelets have shown the best classification values. The first results are shown in Fig. 27. It corresponds to the prediction of the validation dataset comparing the use of only the spectral part of the data and also the corresponding spatial information. From the top to the bottom, the four different classes are reported. From the left to the right, it is possible to compare the ground truth (which structure is known due to the fact that the dataset is simulated) and respectively, the best outcomes using only the spectral information and the ones when also wavelets are used. It is fundamental to highlight here an aspect. In fact, comparing the PLS-DA predictions using only the spectral information and the ones when also wavelets are applied to extract the spatial details, the number of required LVs to obtain good outcomes increases. This is due to the fact that wavelet coefficients are orthogonal, and so a higher number of LVs is fundamental to explain the data and correctly classify the different compounds of the dataset. At the same time, it can be said that the increase of this value is a way to prove that additional information (related to the spatial details) is really taken into account, using this procedure.

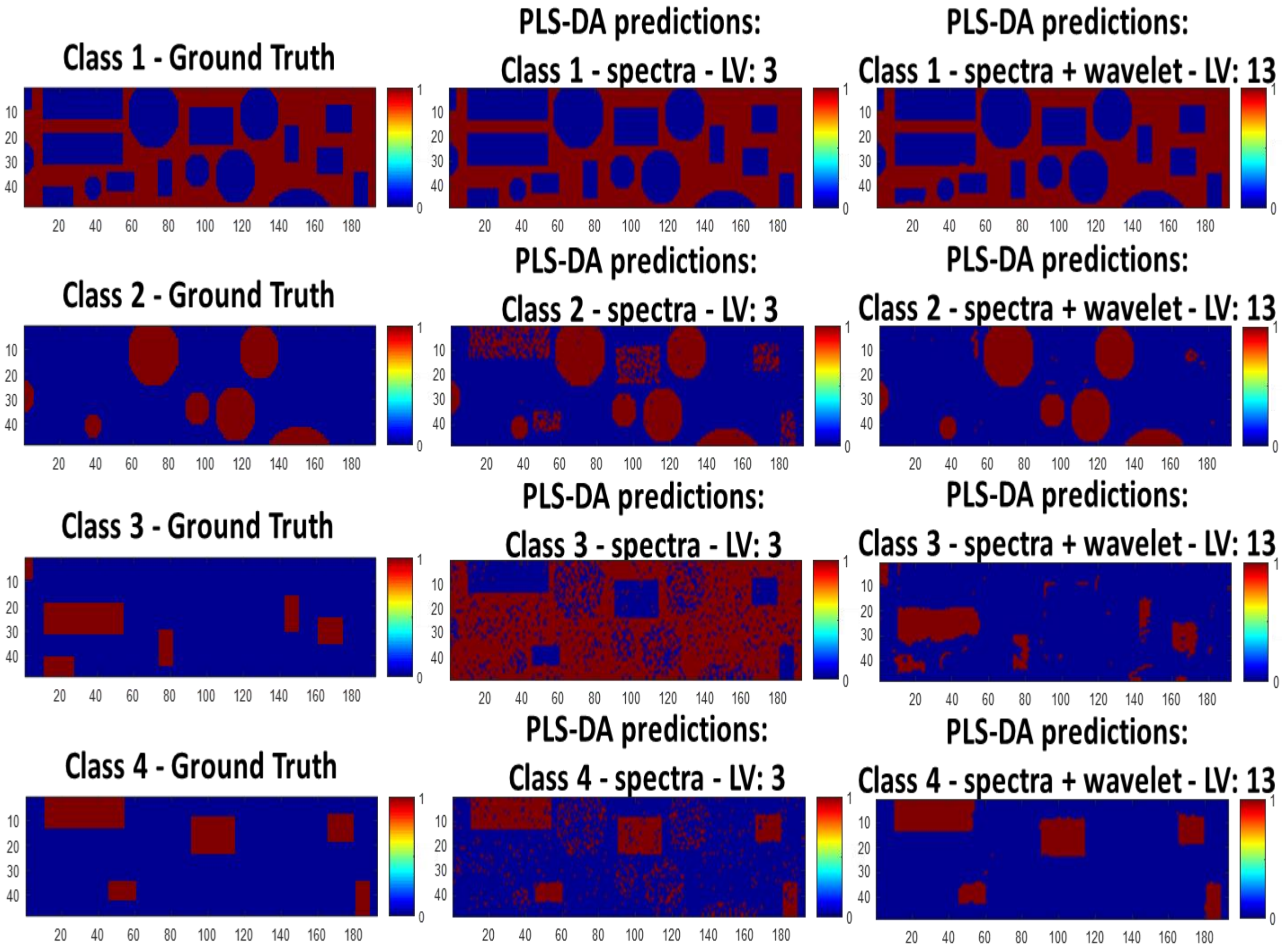


Fig. 27 – From the left to the right, the ground truth, and the PLS-DA predictions using respectively only the spectral and also the spatial information of the given sample for the four different classes.

As discussed at the beginning of this paragraph, using only the spectral information is expectable that some classes, linked to similar spectra, would lead to not precise results. In fact, except for the first class (the purest one), second and fourth classes, linked to the same spectral information, show very similar outcomes in which some pixels of the rectangles and the circles are clearly misclassified. Particularly interesting is the third class. Indeed, due to the fact that it is related to spectra obtained as a linear combination of the other classes, it is completely impossible to carry out a good classification of this particular compound when only the spectral information is used. The general situation completely changes when also the spatial details are finally exploited. Despite the fact that some pixels are still misclassified, it is now possible to distinguish second and fourth classes. Also, regarding the third class, although some imperfections, it is clear how the use of the wavelets has been fundamental to, at the end, obtain better classification outcomes. To conclude, in the Table 1 are reported the main classification figures of merit, in order to evaluate the accuracy of the obtained results compared with the typical approach in which it is possible to notice a general improvement of the predictions when the spatial information is taken into account:

<table>
<tr><th colspan="4">Only spectral information – 3 LVs in the PLS-DA model</th></tr>
<tr><th>Class</th><th>Specificity (%)</th><th>Sensitivity (%)</th><th>Accuracy (%)</th></tr>
<tr><td>1</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>2</td><td>0.8932</td><td>0.9892</td><td>0.9400</td></tr>
<tr><td>3</td><td>0.4508</td><td>0.6599</td><td>0.5454</td></tr>
<tr><td>4</td><td>0.9140</td><td>0.9344</td><td>0.9241</td></tr>
<tr><th colspan="4">Spectral + spatial information (wavelets) – 13 LVs in the PLS-DA model</th></tr>
<tr><th>Class</th><th>Specificity (%)</th><th>Sensitivity (%)</th><th>Accuracy (%)</th></tr>
<tr><td>1</td><td>0.9919</td><td>0.9996</td><td>0.9957</td></tr>
<tr><td>2</td><td>0.9910</td><td>0.9989</td><td>0.9949</td></tr>
<tr><td>3</td><td>0.9754</td><td>0.8185</td><td>0.8935</td></tr>
<tr><td>4</td><td>0.9943</td><td>0.9863</td><td>0.9903</td></tr>
</table>

Table 1 – PLS-DA outcomes of specificity, sensitivity and accuracy of the four classes using, respectively, only spectral and both spectral and spatial information.

5.3.3. A real dataset investigation

In a second step, the same approach has been used, this time exploring a real dataset, obtained in the framework of a collaboration with Prof. José M. Amigo from University of Basque Country, Spain. Particularly, for this study the corresponding matrix is a subregion (128 pixels by 128 pixels) of a more complex case composed of different microplastics, acquired with a µ-FTIR instrumentation, investigated and discussed in another work [219]. Only a particular class of plastics (i.e., polyamides) has been used for the first outcomes discussed here. As previously stated, a common problem is represented by the fact that real samples often have very detailed contours. This scenario can correspond to spectral signals close to the noise level, and so leading to the possibility of confusing parts related to the sample with the background. This may seem insignificant and trivial, but many applications require rigorous estimations of the particle size distribution. This misclassification of the spectra on the edges of the particles has a direct impact on the final estimates, especially for the smallest fragments, which are often the most interesting. The main idea here is that using not only the spectral, but also the spatial information, it would be possible to obtain better classification results, particularly related to the borders of the investigated objects. As in the previous example, for the analysis, a part of the matrix has been used as training set and the corresponding calculated model has been applied to another region of the image used as test set, in order to predict the outcomes to understand the quality of the classification analysis. The same approach described in the previous paragraph has been used and the best classification results have been obtained once again with reverse biorthogonal wavelets, this time using a decomposition level equal to three. Fig. 28 shows the first results, comparing the outcomes using only the spectra and then exploiting also the spatial information obtained with the wavelet transform:

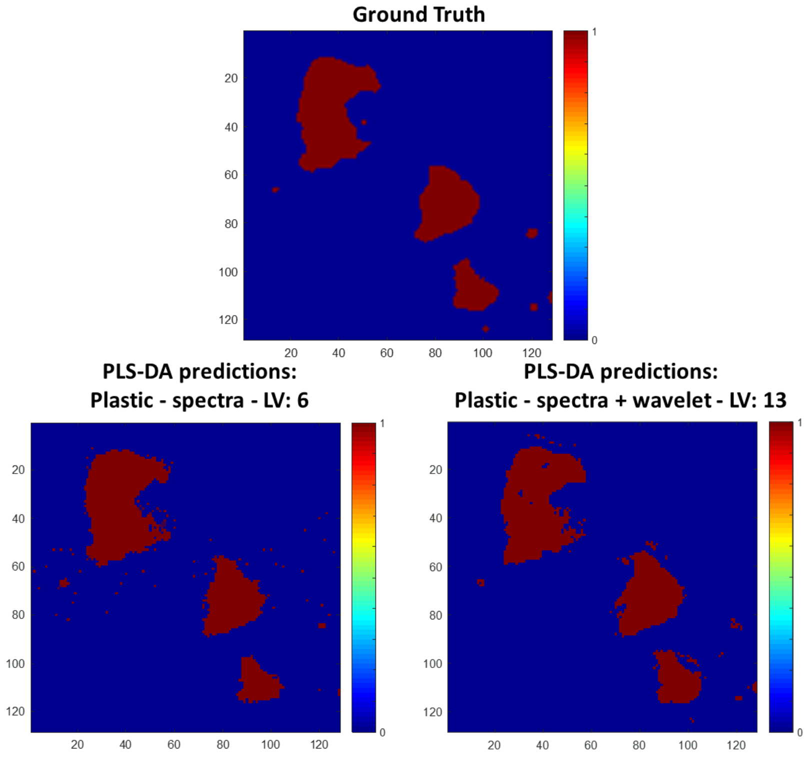


Fig. 28 – On the top, the ground truth. On the bottom, from the left to the right, the PLS-DA predictions using respectively only the spectral and also the spatial information of the given sample.

Compared to the ground truth, it is possible to notice that using the wavelets some borders of the considered plastic objects seem to be better defined. Also important is that using the proposed approach, some random pixels that have been misclassified as plastic using only the spectra are, taking into account also the spatial information, not observable in the results. Lastly, in order to obtain a better idea of the differences of the outcomes using the two approaches, here in Table 2 are again reported the PLS-DA figures of merit when the full field of view is considered:

<table>
<tr><th colspan="4">Only spectral information – 6 LVs in the PLS-DA model</th></tr>
<tr><th>Class</th><th>Specificity (%)</th><th>Sensitivity (%)</th><th>Accuracy (%)</th></tr>
<tr><td>1</td><td>0.9719</td><td>0.9010</td><td>0.9358</td></tr>
<tr><td>2</td><td>0.9010</td><td>0.9719</td><td>0.9358</td></tr>
<tr><th colspan="4">Spectral + spatial information (wavelets) – 13 LVs in the PLS-DA model</th></tr>
<tr><th>Class</th><th>Specificity (%)</th><th>Sensitivity (%)</th><th>Accuracy (%)</th></tr>
<tr><td>1</td><td>0.9709</td><td>0.9760</td><td>0.9735</td></tr>
<tr><td>2</td><td>0.9760</td><td>0.9709</td><td>0.9735</td></tr>
</table>

Table 2 – PLS-DA outcomes of specificity, sensitivity and accuracy of the two classes (the first, the plastic and the second, the background) using, respectively, only spectral and both spectral and spatial information.

At first sight, one could say that the improvement of the classification results is not spectacular. This is mainly due to the fact that the number of particles observed is particularly low, but at the same time, they are also relatively large in size. To demonstrate the incomings using wavelets, these figures of merit were recalculated on a subset of small areas containing some smaller details of the dataset. These areas are identified in Fig. 29 and the results are given in Table 3:

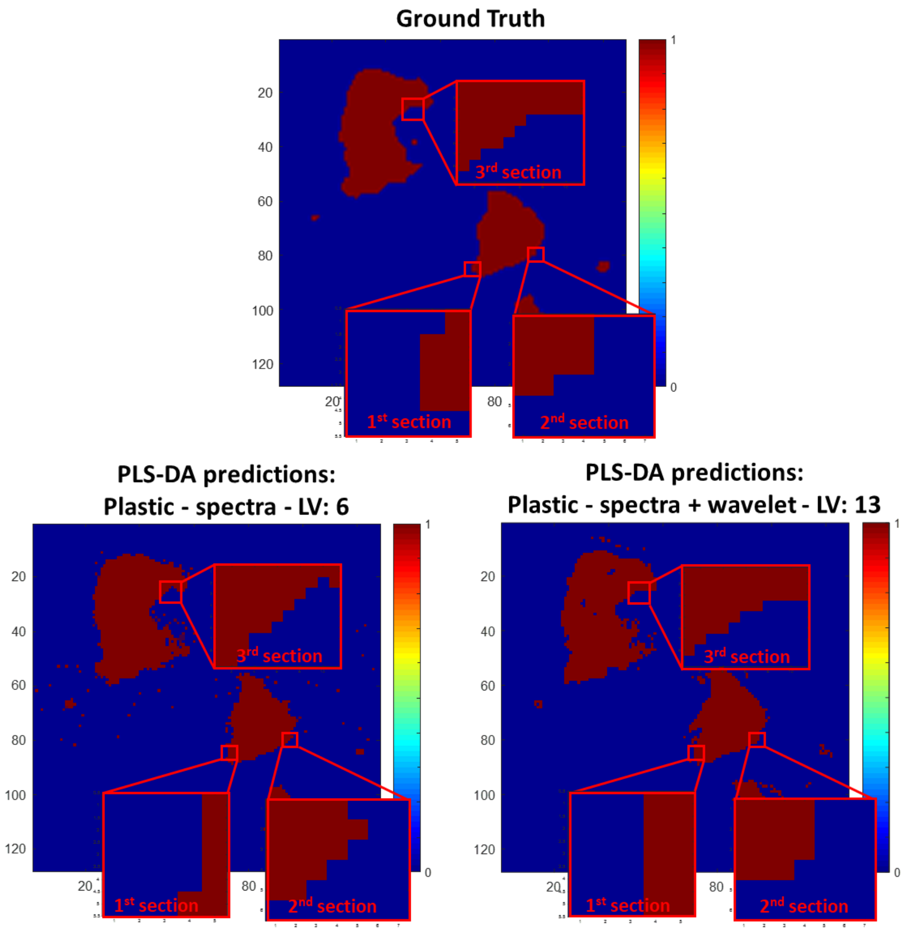


Fig. 29 – The selected subareas used to recalculate the figures of merit of PLS-DA predictions.

Only spectral information – 6 LVs in the PLS-DA model			
Selected subarea	**Specificity (%)**	**Sensitivity (%)**	**Accuracy (%)**
1	0.8947	0.5714	0.7150
2	0.8571	1	0.9258
3	0.8333	0.9216	0.8763
Spectral + spatial information (wavelets) – 13 LVs in the PLS-DA model			
Selected subarea	**Specificity (%)**	**Sensitivity (%)**	**Accuracy (%)**
1	0.8333	1	0.8333
2	0.9643	1	0.9643
3	0.9792	1	0.9895

Table 3 – PLS-DA outcomes of specificity, sensitivity and accuracy of the plastic class for the three selected subareas using, respectively, only spectral and both spectral and spatial information.

5.4. General conclusions in the framework of the use of wavelet transform and future perspectives

Wavelet transform is a very powerful algorithm that can be used to overcome the limitations related to the investigation of a hyperspectral image in chemometrics, due to its particularity in extracting the spatial information, which otherwise would be completely lost during the analysis pipeline. In this way, as it has been shown in the present chapter, it would be possible to obtain better results in complicated situations in which clearly it is important to study not only the spectral, but also the spatial part of the data. Nevertheless, the use of this method is undoubtedly very challenging, due to the complexity shown by this approach. In fact, it is a very complicated task to choose for example the right wavelet family to use, in order to find good outcomes. Also hard is the task of selecting the right decomposition level value to work in the framework of wavelet transform. Lastly, a massive quantity of variables is generated, when this algorithm is used. This is clearly a limitation due to the fact that it is not always easy to find good outcomes when too much information is given. As previously stated, all the obtained coefficients using the wavelet transform (approximation, horizontal, vertical and diagonal details) are orthogonal, and so related to different aspects of the studied image. Despite this, it is assumable that not all of them will be related to very important and fundamental information. An interesting aspect of working with this approach would be to find the way of weighting the real importance of the different variables related to the wavelets (several approximation, horizontal, vertical and diagonal coefficients might be in fact related to futile information for the classification analysis) and thus, in a first step skim them in order to proceed with the analysis reducing the redundant and useless parts of the data. Another important point would be to find a method to easily select the right wavelet family and decomposition level to be used for a particular data cube. In fact, different families can give different outcomes, as well as use different decomposition levels can lead to better or worst results, depending on the complexity and structure of the investigated sample. In order to obtain at the end good PLS-DA predictions is normally necessary to compare the outcomes obtained using various wavelets inputs. Find a way to directly select the right decomposition level and wavelet family for a given matrix without being forced to this intermediate step would solve many problems related to time consuming in the calculations. Nevertheless, it is undeniable that wavelets show very promising results in order to finally deal with the spatial information in hyperspectral imaging, a fundamental aspect that is nowadays one of the main purposes of different research areas.

GENERAL CONCLUSIONS AND PERSPECTIVES

Analytical chemistry research area is developing very rapidly. The use of always more advanced instrumentations, which are linked to new investigation possibilities, is not an option, but the reality of the situation. Hyperspectral imaging is just one of the various ways in which nowadays the study of complex matrices can be faced. Nevertheless, it is mandatory to consider the many limitations and challenges those new techniques can experience. As vastly described into the present manuscript, just to mention again a few of the current problems related to this discipline, hyperspectral image acquisition is normally related to the generation of big datasets, made of hundreds of thousands to millions of spectra. The nature of the explored samples compared with a bulk analysis can be more complex and heterogeneous. At the same time, it is important to consider that many components can be present as traces and so irretrievably lost during the data analysis, if the right investigation approach is not used. Naturally, this is a scenario to avoid, because in many situations the smallest information of a given sample is also the most interesting. Nowadays, find a way to deal with the vastness of details carried out by the acquisition of a hyperspectral image is a real challenge, but mandatory. It is not conceivable the option of not exploiting completely the possibilities related to this technique, also considering the obtainable spatial information. Chemometrics has shown in many research areas to be one of the most interesting methodologies that can currently be used to overcome many limitations linked to different routine analyses. Of course, the interest in exploiting this tool in the hyperspectral image analysis is not an exception and recently a real progress has been shown, considering many aspects. Nevertheless, the possibility of always overcoming the new limits encountered using the new developed methodologies is what research should aim at. This is the main goal of the present doctorate project, as vastly described into this manuscript: the exploration of the already existent chemometric methodologies that are normally used for data analysis, particularly in the domain of the hyperspectral imaging investigation, in order to provide new and useful approaches for the analysis of complex matrices. The present project tried to follow a precise path in which different but correlated aspects and limitations concerning the use of hyperspectral images could be faced and new chemometric methods exploited to overcome the main challenges in this domain. A very important part of this thesis has focused in the investigation of big datasets. Find a way to manage a huge quantity of information and anyway keep the most important part of the data is fundamental. Over other algorithms, SIMPLISMA has been mainly exploited for this task. Using its peculiarity, based on the selection of the purest details of a dataset, it has been demonstrated that it is a useful tool with the purpose of reducing the total quantity of information, in order to work with smaller data matrices, made of only the most important part of the data. The main limitation of SIMPLISMA is that some inputs have to be set in order to work properly. This algorithm is normally used for the generation of initial estimates for the spectral unmixing methods, such as MCR-ALS. Using SIMPLISMA, the operator needs to choose the right rank, in order to select the right number of initial estimates and obtain good outcomes in the spectral unmixing analysis. Nevertheless, this task can be challenging. In fact, it is not always very easy to select the right rank. Some components of the sample can be, for example, represented by few pixels, hard to be observed by the use of methods such as an exploratory analysis (e.g., PCA), due to the small explained variance related to these compounds, and so irretrievably lost. The used method in this work is related to the implementation of a more automatic way to deal with this kind of situation in order to help the operator using a more intuitive graphical approach. In another work of this doctorate (described in Chapter 4 of the present manuscript), SIMPLISMA has been also applied, this time with the main purpose of reducing the total quantity of spectra, with the only intention of working with datasets made of a very small percentage compared with the initial data dimensions. In the present PhD project, SIMPLISMA has been applied on data of different nature acquired using distinct instruments (Raman, EDX, UV, LIBS, PIL, other spectroscopic techniques from synchrotron beamline, etc.), showing the potentiality of being applied in various contexts obtaining good results. In the same way, KM clustering can be used to deal with big datasets. In this situation as well, the choice of the right number of clusters and some initial inputs are fundamental steps to obtain good results. Another important limitation using this algorithm is again related to the fact that big datasets can be linked to classes represented by a limited number of spectra. In this kind of scenario, the chance of losing some important information is very reasonable. In a very intuitive way, it has been shown in the present manuscript an alternative way to deal with this situation, leading at the end to better results, in which not only the major, but also the minor compounds and the traces can be observed and correctly classified. Within the many spectroscopic techniques that are normally used, LIBS is for sure one of the most interesting nowadays and many research groups use this spectroscopy for the elemental analysis, also in the hyperspectral imaging domain. Nevertheless, the investigation of this kind of data is still very limited and only routine analyses are generally used. Considering the complexity of the generated information and the huge quantity of spectra that can be acquired in very reasonable times, chemometrics is for sure an interesting methodology that can be exploited to overcome the general limitations related to this spectroscopy. This is the reason why a part of this PhD has focused particularly on the possible implementations related to this specific technique. In a first step, instrumental artifacts were faced. In fact, a very common problem related to LIBS spectra is the generation of saturated signals, which would lead to a wrong interpretation of the data.

Statistical imputation has been used for this purpose, in order to generate good resolved peaks when the device faces this kind of problem. Then, as previously explained, another important aspect considering the huge quantity of produced data is to find a way to go through the computational problems that can be experienced due to this common scenario. A general analysis pipeline has been developed during this work, in order to facilitate the use of the enormous amount of obtained data, also correcting possible artifacts generated during the acquisition. Finally, another important part of the same explored procedure is related to the possibility of fusing LIBS data with other spectroscopies, due to the capability of this instrumentation in obtaining the response coming from different spectral regions (in the present case, PIL and Raman). In fact, nowadays it is a very important aspect the one of fusing different spectral responses in order to obtain a better and more complete investigation of a given sample, to deep the knowledge related to the complexity of the investigated specimen. Using the proposed approach, it has been possible in a very easy way to fuse the datasets, considering only the most important information related to the different spectroscopies and at the end obtain very interesting outcomes. Lastly, as stated, the analysis of a hyperspectral image without taking into consideration also the spatial information corresponds to an incomplete investigation. From a certain point of view, it is unconceivable the concept of not exploiting also this part of the data, considering the fact that a hyperspectral image is first thing made of spatial details, which differentiate it from a classical dataset containing unordered spectra. Nevertheless, extracting and using in the right way this kind of information is still very challenging, also in chemometrics. Wavelet transform has been here used in order to deal with this kind of problem. In fact, this kind of algorithm can look at the spatial part of the data without unfolding a given data cube in the corresponding two-dimensional dataset, leading to the observation of new details. The limitation is represented by the fact that wavelets are complex signals, hard to be interpreted. Chemometrics has been used also in this scenario, to manage the extractable information and obtain the most of the spatial details related to hyperspectral image analysis. Particularly, exploratory analysis (PCA) and classification analysis (PLS-DA) were exploited to demonstrate the value of using wavelets in order to obtain at the end more interesting results, not based only on the spectral part of the data, but also the spatial one, when an image is investigated.

Finally, it can be concluded that this PhD project has faced various arguments related to the limitations and issues linked to hyperspectral imaging, particularly considering big datasets and multimodality, using different chemometric strategies to overcome these aspects. Nevertheless, it has to be considered the fact that the presented work is only a step forward this interesting research area and that many other things can be done in order to constantly obtain new results and ideas to develop advanced methodologies. Exploiting the already considered chemometric tools, it would be possible to broaden the horizons to new concepts. It is a very important task, by way of example, to automatize the data analysis pipeline. Due to their different but correlated characteristics, SIMPLISMA and KM clustering could be used in the same procedure, in order to at first select the most important spectra of an investigated sample and then divide the obtained information into different clusters, based on the similarities among the various extracted spectral details. In the same way, SIMPLISMA (or other chemometric strategies pointed to the selection of the most important information in a given dataset) could be exploited in order to filter the redundant data extracted by the use of the wavelets, with the purpose of facilitating the use of the complex information obtained with this algorithm in the framework of the exploration of the spatial details linked to a hyperspectral image. Wavelet transform is nowadays for sure one of the most important algorithms that can be used for the interpretation of modern hyperspectral image analysis. In fact, due to the development of always more sophisticated instruments, the quality of the acquired images is constantly increasing and more spatial details are obtainable. More work has to be done on this, and new ideas in order to really exploit this powerful aspect have to be carried out, such as the application of wavelet transform with other chemometric tools. LIBS analysis, as widely described in this manuscript, is very interesting and many facets related to the use of the acquired datasets using this instrumentation have been faced. Nevertheless, more chemometric methods could be used in order to obtain more precise outcomes using this elemental analysis technique. In particular, the data fusion of different spectral regions has been here described. One example would be to merge together not only two different instrumental responses (LIBS and PIL or LIBS and Raman), but exploiting methods that could reduce the dimensionality of the data such as SIMPLISMA, observing the outcomes of these three spectroscopies, or even more, at the same time. In the same way, the proposed methodologies linked to LIBS analysis might be used for other spectroscopies, leading to better results.

Beyond the concepts here proposed, these such as many other ideas could be used in the investigation of a hyperspectral image. Nowadays there are not limitations from both the instrumental and the chemometric methodologies points of view. Hyperspectral imaging is a very interesting research area and it is out of the question that it is obtaining an always increasing importance in many research domains, particularly when coupled with chemometric approaches. The work carried out during these three years of doctorate are linked to the expectation that this manuscript will be useful material for who will be from now on dealing with hyperspectral imaging, and that these results will be the inspiration to find new possible project lines in chemometrics and spectroscopy.

REFERENCES

[1] G. Gauglitz, T. Vo-Dinh, eds., Handbook of Spectroscopy, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG, 2005. https://doi.org/10.1002/3527602305.fmatter.

[2] D.L. Pavia, G.M. Lampman, G.S. Kriz, J.R. Vyvyan, Introduction to Spectroscopy, 2013. https://doi.org/10.1021/ed056pA323.2.

[3] J.M. Hollas, Modern spectroscopy, 4th ed, J. Wiley, Chichester; Hoboken, NJ, 2004. https://doi.org/10.1021/ed082p43.1.

[4] V. Baeten, P. Dardenne, Spectroscopy: Developments in instrumentation and analysis, Grasas y Aceites. 53 (2002) 45–63. https://doi.org/10.3989/gya.2002.v53.i1.289.

[5] J. Sugiyama, Visualization of Sugar Content in the Flesh of a Melon by Near-Infrared Imaging, Journal of Agricultural and Food Chemistry. 47 (1999) 2715–2718. https://doi.org/10.1021/jf981079i.

[6] P. Tauler, E. Casassas, Application of principal component analysis to the study of multiple equilibria systems, Analytica Chimica Acta. 223 (1989) 257–268. https://doi.org/10.1016/S0003-2670(00)84089-1.

[7] M. Boiret, D.N. Rutledge, N. Gorretta, Y.M. Ginot, J.M. Roger, Application of independent component analysis on Raman images of a pharmaceutical drug product: Pure spectra determination and spatial distribution of constituents, Journal of Pharmaceutical and Biomedical Analysis. 90 (2014) 78–84. https://doi.org/10.1016/j.jpba.2013.11.025.

[8] B.M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review, Postharvest Biology and Technology. 46 (2007) 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024.

[9] O. Piot, J.C. Autran, M. Manfait, Spatial Distribution of Protein and Phenolic Constituents in Wheat Grain as Probed by Confocal Raman Microspectroscopy, Journal of Cereal Science. 32 (2000) 57–71. https://doi.org/10.1006/jcrs.2000.0314.

[10] L. Ravikanth, D.S. Jayas, N.D.G. White, P.G. Fields, D.W. Sun, Extraction of Spectral Information from Hyperspectral Data and Application of Hyperspectral Imaging for Food and Agricultural Products, Food Bioprocess Technol. 10 (2017) 1–33. https://doi.org/10.1007/s11947-016-1817-8.

[11] P.S. Thenkabail, J.G. Lyon, A. Huete, eds., Hyperspectral Remote Sensing of Vegetation, 2nd ed., CRC Press, 2018. https://doi.org/10.1201/9781315164151.

[12] G. ElMasry, D.W. Sun, Principles of Hyperspectral Imaging Technology, in: Hyperspectral Imaging for Food Quality Analysis and Control, Elsevier, 2010: pp. 3–43. https://doi.org/10.1016/B978-0-12-374753-2.10001-2.

[13] B. Fei, Hyperspectral imaging in medical applications, in: Data Handling in Science and Technology, Elsevier, 2020: pp. 523–565. https://doi.org/10.1016/B978-0-444-63977-6.00021-3.

[14] H.F. Grahn, P. Geladi, eds., Techniques and Applications of Hyperspectral Image Analysis, John Wiley & Sons, Ltd, Chichester, UK, 2007. https://doi.org/10.1002/9780470010884.

[15] M. Imani, H. Ghassemian, An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges, Information Fusion. 59 (2020) 59–83. https://doi.org/10.1016/j.inffus.2020.01.007.

[16] J. Ma, D. W. Sun, H. Pu, J. H. Cheng, Q. Wei, Advanced Techniques for Hyperspectral Imaging in the Food Industry: Principles and Recent Applications, Annu. Rev. Food Sci. Technol. 10 (2019) 197–220. https://doi.org/10.1146/annurev-food-032818-121155.

[17] M.H. Bharati, J.J. Liu, J.F. MacGregor, Image texture analysis: methods and comparisons, Chemometrics and Intelligent Laboratory Systems. 72 (2004) 57–71. https://doi.org/10.1016/j.chemolab.2004.02.005.

[18] L. Vincent, Morphological grayscale reconstruction: definition, efficient algorithm and applications in image analysis, Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (1992) 633–635. https://doi.org/10.1109/CVPR.1992.223122.

[19] L. Vincent, Morphological grayscale reconstruction in image analysis: applications and efficient algorithms, IEEE Trans. on Image Process. 2 (1993) 176–201. https://doi.org/10.1109/83.217222.

[20] J. Dengler, H. Bertsch, J.F. Desaga, M. Schmidt, New Trends of Image Analysis in the Medical Field, Methods Inf Med. 27 (1988) 53–57. https://doi.org/10.1055/s-0038-1635520.

[21] A. Rosenfeli, Image analysis: problems, progress and prospects, Readings in Computer Vision. (1987) 3–12. https://doi.org/10.1016/B978-0-08-051581-6.50006-4.

[22] C. Costa, F. Antonucci, F. Pallottino, J. Aguzzi, D.-W. Sun, P. Menesatti, Shape Analysis of Agricultural Products: A Review of Recent Research Advances and Potential Application to Computer Vision, Food Bioprocess Technol. 4 (2011) 673–692. https://doi.org/10.1007/s11947-011-0556-0.

[23] S. Cubero, N. Aleixos, E. Moltó, J. Gómez-Sanchis, J. Blasco, Advances in Machine Vision Applications for Automatic Inspection and Quality Evaluation of Fruits and Vegetables, Food Bioprocess Technol. 4 (2011) 487–504. https://doi.org/10.1007/s11947-010-0411-8.

[24] H.M. Ghadirli, A. Nodehi, R. Enayatifar, An overview of encryption algorithms in color images, Signal Processing. 164 (2019) 163–185. https://doi.org/10.1016/j.sigpro.2019.06.010.

[25] A. Giraudo, R. Calvini, G. Orlandi, A. Ulrici, F. Geobaldo, F. Savorani, Development of an automated method for the identification of defective hazelnuts based on RGB image analysis and colourgrams, Food Control. 94 (2018) 233–240. https://doi.org/10.1016/j.foodcont.2018.07.018.

[26] K.A. Bakeev, 2nd ed., Process analytical technology: spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries, Blackwell Pub, Oxford, UK ; Ames, Iowa, 2005. https://doi.org/10.1002/9780470689592.

[27] C.I. Chang, Hyperspectral Imaging: techniques for spectral detection and classification, Springer US, Boston, MA, 2003. https://doi.org/10.1007/978-1-4419-9170-6.

[28] M. Kamruzzaman, D.W. Sun, Introduction to Hyperspectral Imaging Technology, in: Computer Vision Technology for Food Quality Evaluation, Elsevier, 2016: pp. 111–139. https://doi.org/10.1016/B978-0-12-802232-0.00005-0.

[29] J.M. Amigo, I. Martí, A. Gowen, Hyperspectral Imaging and Chemometrics, in: Data Handling in Science and Technology, Elsevier, 2013: pp. 343–370. https://doi.org/10.1016/B978-0-444-59528-7.00009-0.

[30] D. Manolakis, R. Lockwood, T. Cooley, Hyperspectral Imaging Remote Sensing: Physics, Sensors, and Algorithms, Cambridge University Press, Cambridge, 2016. https://doi.org/10.1017/CBO9781316017876.

[31] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, J. Sousa, Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry, Remote Sensing. 9 (2017) 1110. https://doi.org/10.3390/rs9111110.

[32] J. Zabalza, J. Ren, M. Yang, Y. Zhang, J. Wang, S. Marshall, J. Han, Novel Folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing, ISPRS Journal of Photogrammetry and Remote Sensing. 93 (2014) 112–122. https://doi.org/10.1016/j.isprsjprs.2014.04.006.

[33] X. Zhang, R. Tauler, Application of Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) to remote sensing hyperspectral imaging, Analytica Chimica Acta. 762 (2013) 25–38. https://doi.org/10.1016/j.aca.2012.11.043.

[34] Q.L. Ma, V. Motto-Ros, W.Q. Lei, M. Boueri, L.J. Zheng, H.P. Zeng, M. Bar-Matthews, A. Ayalon, G. Panczer, J. Yu, Multi-elemental mapping of a speleothem using laser-induced

breakdown spectroscopy, Spectrochimica Acta Part B: Atomic Spectroscopy. 65 (2010) 707–714. https://doi.org/10.1016/j.sab.2010.03.004.

[35] F.A. Kruse, Identification and mapping of minerals in drill core using hyperspectral image analysis of infrared reflectance spectra, International Journal of Remote Sensing. 17 (1996) 1623–1632. https://doi.org/10.1080/01431169608948728.

[36] R. Fandrich, Y. Gu, D. Burrows, K. Moeller, Modern SEM-based mineral liberation analysis, International Journal of Mineral Processing. 84 (2007) 310–320. https://doi.org/10.1016/j.minpro.2006.07.018.

[37] T.A. Labutin, V.N. Lednev, A.A. Ilyin, A.M. Popov, Elemental imaging by laser-induced breakdown spectroscopy for the geological characterization of minerals, Journal of Analytical Atomic Spectrometry. 31 (2016) 90–118. https://doi.org/10.1039/C8JA00048D.

[38] D.W. Sun, Hyperspectral Imaging for Food Quality Analysis and Control, Elsevier, 2010. https://doi.org/10.1016/C2009-0-01853-4.

[39] B. Park, R. Lu, Hyperspectral Imaging Technology in Food and Agriculture, Springer New York, New York, NY, 2015. https://doi.org/10.1007/978-1-4939-2836-1.

[40] X. Li, R. Li, M. Wang, Y. Liu, B. Zhang, J. Zhou, Hyperspectral Imaging and Their Applications in the Nondestructive Quality Assessment of Fruits and Vegetables, in: A.I.L. Maldonado, H.R. Fuentes, J.A.V. Contreras (Eds.), Hyperspectral Imaging in Agriculture, Food and Environment, InTech, 2018. https://doi.org/10.5772/intechopen.72250.

[41] M.Á. Fernández de la Ossa, J.M. Amigo, C. García-Ruiz, Detection of residues from explosive manipulation by near infrared hyperspectral imaging: A promising forensic tool, Forensic Science International. 242 (2014) 228–235. https://doi.org/10.1016/j.forsciint.2014.06.023.

[42] K.B. Ferreira, A.G.G. Oliveira, A.S. Gonçalves, J.A. Gomes, Evaluation of Hyperspectral Imaging Visible/Near Infrared Spectroscopy as a forensic tool for automotive paint distinction, Forensic Chemistry. 5 (2017) 46–52. https://doi.org/10.1016/j.forc.2017.06.001.

[43] L.N. Brewer, J.A. Ohlhausen, P.G. Kotula, J.R. Michael, Forensic analysis of bioagents by X-ray and TOF-SIMS hyperspectral imaging, Forensic Science International. 179 (2008) 98–106. https://doi.org/10.1016/j.forsciint.2008.04.020.

[44] G.J. Edelman, E. Gaston, T.G. van Leeuwen, P.J. Cullen, M.C.G. Aalders, Hyperspectral imaging for non-contact analysis of forensic traces, Forensic Science International. 223 (2012) 28–39. https://doi.org/10.1016/j.forsciint.2012.09.012.

[45] H. Akbari, L.V. Halig, D.M. Schuster, A. Osunkoya, V. Master, P.T. Nieh, G.Z. Chen, B. Fei, Hyperspectral imaging and quantitative analysis for prostate cancer detection, J. Biomed. Opt. 17 (2012) 0760051. https://doi.org/10.1117/1.JBO.17.7.076005.

[46] Z. Liu, H. Wang, Q. Li, Tongue Tumor Detection in Medical Hyperspectral Images, Sensors. 12 (2011) 162–174. https://doi.org/10.3390/s120100162.

[47] M.A. Calin, S.V. Parasca, D. Savastru, D. Manea, Hyperspectral Imaging in the Medical Field: Present and Future, Applied Spectroscopy Reviews. 49 (2014) 435–447. https://doi.org/10.1080/05704928.2013.838678.

[48] H. Akbari, Y. Kosugi, K. Kojima, N. Tanaka, Detection and Analysis of the Intestinal Ischemia Using Visible and Invisible Hyperspectral Imaging, IEEE Trans. Biomed. Eng. 57 (2010) 2011–2017. https://doi.org/10.1109/TBME.2010.2049110.

[49] M.A. Calin, S.V. Parasca, R. Savastru, D. Manea, Characterization of burns using hyperspectral imaging technique – A preliminary study, Burns. 41 (2015) 118–124. https://doi.org/10.1016/j.burns.2014.05.002.

[50] P.Y. Sacré, C. De Bleye, P.F. Chavez, L. Netchacovitch, Ph. Hubert, E. Ziemons, Data processing of vibrational chemical imaging for pharmaceutical applications, Journal of Pharmaceutical and Biomedical Analysis. 101 (2014) 123–140. https://doi.org/10.1016/j.jpba.2014.04.012.

[51] Y. Roggo, A. Edmond, P. Chalus, M. Ulmschneider, Infrared hyperspectral imaging for qualitative analysis of pharmaceutical solid forms, Analytica Chimica Acta. 535 (2005) 79–87. https://doi.org/10.1016/j.aca.2004.12.037.

[52] W. Fortunato de Carvalho Rocha, G.P. Sabin, P.H. Março, R.J. Poppi, Quantitative analysis of piroxicam polymorphs pharmaceutical mixtures by hyperspectral imaging and chemometrics, Chemometrics and Intelligent Laboratory Systems. 106 (2011) 198–204. https://doi.org/10.1016/j.chemolab.2010.04.015.

[53] J.M. Amigo, Practical issues of hyperspectral imaging analysis of solid dosage forms, Anal Bioanal Chem. 398 (2010) 93–109. https://doi.org/10.1007/s00216-010-3828-z.

[54] M. Al Ktash, M. Stefanakis, B. Boldrini, E. Ostertag, M. Brecht, Characterization of Pharmaceutical Tablets Using UV Hyperspectral Imaging as a Rapid In-Line Analysis Tool, Sensors. 21 (2021) 4436. https://doi.org/10.3390/s21134436.

[55] A. Rehman, S.A. Qureshi, A review of the medical hyperspectral imaging systems and unmixing algorithms' in biological tissues, Photodiagnosis and Photodynamic Therapy. 33 (2021) 102165. https://doi.org/10.1016/j.pdpdt.2020.102165.

[56] V. Studer, J. Bobin, M. Chahid, H.S. Mousavi, E. Candes, M. Dahan, Compressive fluorescence microscopy for biological and hyperspectral imaging, Proceedings of the National Academy of Sciences. 109 (2012) E1679–E1687. https://doi.org/10.1073/pnas.1119511109.

[57] R. Vejarano, R. Siche, W. Tesfaye, Evaluation of biological contaminants in foods by hyperspectral imaging: A review, International Journal of Food Properties. (2017) 1–34. https://doi.org/10.1080/10942912.2017.1338729.

[58] M.D.P.S. Peña, A. Gottipati, S. Tahiliani, N.M. Neu-Baker, M.D. Frame, A.J. Friedman, S.A. Brenner, Hyperspectral imaging of nanoparticles in biological samples: Simultaneous visualization and elemental identification: Hyperspectral Mapping in Biological Samples, Microsc. Res. Tech. 79 (2016) 349–358. https://doi.org/10.1002/jemt.22637.

[59] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, Journal of Pharmaceutical and Biomedical Analysis. 44 (2007) 683–700. https://doi.org/10.1016/j.jpba.2007.03.023.

[60] B. Pavoni, N. Rado, R. Piazza, S. Frignani, FT-IR Spectroscopy and Chemometrics as a Useful Approach for Determining Chemical-Physical Properties of Gasoline, by Minimizing Analytical Times and Sample Handling, Annali Di Chimica. 94 (2004) 521–532. https://doi.org/10.1002/adic.200490066.

[61] J.J. Workman, P.R. Mobley, B.R. Kowalski, R. Bro, Review of Chemometrics Applied to Spectroscopy, Applied Spectroscopy Reviews. 31 (1996) 73–124. https://doi.org/10.1080/05704929608000565.

[62] P. Geladi, E. Dåbakk, An Overview of Chemometrics Applications in near Infrared Spectrometry, Journal of Near Infrared Spectroscopy. 3 (1995) 119–132. https://doi.org/10.1255/jnirs.63.

[63] T. Fearn, Chemometrics: An Enabling Tool for NIR, NIR News. 16 (2005) 17–19. https://doi.org/10.1255/nirn.856.

[64] H. Akbari, L.V. Halig, H. Zhang, D. Wang, Z.G. Chen, B. Fei, Detection of cancer metastasis using a novel macroscopic hyperspectral method, in: San Diego, California, USA, 2012: p. 831711. https://doi.org/10.1117/12.912026.

[65] S. Mahesh, A. Manickavasagan, D.S. Jayas, J. Paliwal, N.D.G. White, Feasibility of near-infrared hyperspectral imaging to differentiate Canadian wheat classes, Biosystems Engineering. 101 (2008) 50–57. https://doi.org/10.1016/j.biosystemseng.2008.05.017.

[66] F. Rosi, C. Miliani, R. Braun, R. Harig, D. Sali, B.G. Brunetti, A. Sgamellotti, Noninvasive Analysis of Paintings by Mid-infrared Hyperspectral Imaging, Angew. Chem. Int. Ed. 52 (2013) 5258–5261. https://doi.org/10.1002/anie.201209929.

[67] C.V. Raman, K.S. Krishnan, A new class of spectra due to secondary radiation, Indian J. Phys. 2 (1928) 399–419. Retrieved from URL (Accessed the 01/03/2022): http://dspace.rri.res.in/bitstream/2289/2134/1/1928%20IJP%20V2%20p399-419.pdf.

[68] C.V. Raman, K.S. Krishnan, A new type of secondary radiation, A New Type of Secondary Radiation. 121 (1928) 501–502. https://doi.org/10.1038/121501c0.

[69] R. Salzer, H.W. Sielser, Infrared and Raman Spectroscopic Imaging. Edited by Reiner Salzer and Heinz W. Siesler., 1st ed., Wiley-VCH, 2010. Retrieved from URL (Accessed the 01/03/2022): https://onlinelibrary.wiley.com/doi/10.1002/anie.200906567.

[70] X. Dong, M. Jakobi, S. Wang, M.H. Köhler, X. Zhang, A.W. Koch, A review of hyperspectral imaging for nanoscale materials research, Applied Spectroscopy Reviews. 54 (2019) 285–305. https://doi.org/10.1080/05704928.2018.1463235.

[71] Y.Z. Feng, D.W. Sun, Application of Hyperspectral Imaging in Food Safety Inspection and Control: A Review, Critical Reviews in Food Science and Nutrition. 52 (2012) 1039–1058. https://doi.org/10.1080/10408398.2011.651542.

[72] W.H. Su, D.W. Sun, Fourier Transform Infrared and Raman and Hyperspectral Imaging Techniques for Quality Determinations of Powdery Foods: A Review, Comprehensive Reviews in Food Science and Food Safety. 17 (2018) 104–122. https://doi.org/10.1111/1541-4337.12314.

[73] D. Wu, D.W. Sun, Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part I: Fundamentals, Innovative Food Science & Emerging Technologies. 19 (2013) 1–14. https://doi.org/10.1016/j.ifset.2013.04.014.

[74] D. Fu, G. Holtom, C. Freudiger, X. Zhang, X.S. Xie, Hyperspectral Imaging with Stimulated Raman Scattering by Chirped Femtosecond Lasers, J. Phys. Chem. B. 117 (2013) 4634–4640. https://doi.org/10.1021/jp308938t.

[75] Y. Shao, Y. Li, L. Jiang, J. Pan, Y. He, X. Dou, Identification of pesticide varieties by detecting characteristics of Chlorella pyrenoidosa using Visible/Near infrared hyperspectral imaging and Raman microspectroscopy technology, Water Research. 104 (2016) 432–440. https://doi.org/10.1016/j.watres.2016.08.042.

[76] 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.

[77] S. Piqueras, L. Duponchel, R. Tauler, A. de Juan, Monitoring polymorphic transformations by using in situ Raman hyperspectral imaging and image multiset analysis, Analytica Chimica Acta. 819 (2014) 15–25. https://doi.org/10.1016/j.aca.2014.02.027.

[78] A.A. Gowen, Y. Feng, E. Gaston, V. Valdramidis, Recent applications of hyperspectral imaging in microbiology, Talanta. 137 (2015) 43–54. https://doi.org/10.1016/j.talanta.2015.01.012.

[79] S. Schlücker, Infrared and Raman Spectroscopic Imaging. Edited by Reiner Salzer and Heinz W. Siesler., Angewandte Chemie International Edition. 49 (2010) 1192–1192. https://doi.org/10.1002/anie.200906567.

[80] J.I. Goldstein, D.E. Newbury, J.R. Michael, N.W.M. Ritchie, J.H.J. Scott, D.C. Joy, Scanning Electron Microscopy and X-Ray Microanalysis, Springer New York, New York, NY (2018). https://doi.org/10.1007/978-1-4939-6676-9.

[81] J. Ofner, J. Kirschner, E. Eitenberger, G. Friedbacher, A. Kasper-Giebl, H. Lohninger, C. Eisenmenger-Sittner, B. Lendl, A novel substrate for multisensor hyperspectral imaging, Journal of Microscopy. 265 (2017) 341–348. https://doi.org/10.1111/jmi.12506.

[82] J. Ofner, K.A. Kamilli, E. Eitenberger, G. Friedbacher, B. Lendl, A. Held, H. Lohninger, Chemometric Analysis of Multisensor Hyperspectral Images of Precipitated Atmospheric Particulate Matter, Anal. Chem. 87 (2015) 9413–9420. https://doi.org/10.1021/acs.analchem.5b02272.

[83] P.R. Edwards, R.W. Martin, M.R. Lee, Combined cathodoluminescence hyperspectral imaging and wavelength dispersive X-ray analysis of minerals, American Mineralogist. 92 (2007) 235–242. https://doi.org/10.2138/am.2007.2152.

[84] R.A. Ruane, K.A.J. Doherty, R. Dorrepaal, B. Twomey, A. Gowen, J. Flanagan, D. de Faoite, K.T. Stanton, Hyperspectral imaging with unsupervised pattern recognition: A novel surface characterization technique for thermal control coatings, Materials Letters. 254 (2019) 273–277. https://doi.org/10.1016/j.matlet.2019.07.085.

[85] L. Jolivet, M. Leprince, S. Moncayo, L. Sorbier, C.P. Lienemann, V. Motto-Ros, Review of the recent advances and applications of LIBS-based imaging, Spectrochimica Acta Part B: Atomic Spectroscopy. 151 (2019) 41–53. https://doi.org/10.1016/j.sab.2018.11.008.

[86] T.A. Labutin, V.N. Lednev, A.A. Ilyin, A.M. Popov, Femtosecond laser-induced breakdown spectroscopy, J. Anal. At. Spectrom. 31 (2016) 90–118. https://doi.org/10.1039/C5JA00301F.

[87] M. Gaft, Y. Raichlin, F. Pelascini, G. Panzer, V. Motto Ros, Imaging rare-earth elements in minerals by laser-induced plasma spectroscopy: Molecular emission and plasma-induced

luminescence, Spectrochimica Acta Part B: Atomic Spectroscopy. 151 (2019) 12–19. https://doi.org/10.1016/j.sab.2018.11.003.

[88] V. Motto-Ros, S. Moncayo, F. Trichard, F. Pelascini, Investigation of signal extraction in the frame of laser induced breakdown spectroscopy imaging, Spectrochimica Acta Part B: Atomic Spectroscopy. 155 (2019) 127–133. https://doi.org/10.1016/j.sab.2019.04.004.

[89] V. Motto-Ros, S. Moncayo, C. Fabre, B. Busser, LIBS imaging applications, in: Laser-Induced Breakdown Spectroscopy, Elsevier, 2020: pp. 329–346. https://doi.org/10.1016/B978-0-12-818829-3.00014-9.

[90] M. Gaft, R. Reisfeld, G. Panczer, Modern luminescence spectroscopy of minerals and materials, Springer, Berlin; New York, 2005. https://doi.org/10.1007/978-3-319-24765-6.

[91] M. Gaft, L. Nagli, Y. Groisman, Plasma induced luminescence (PIL), Optical Materials. 34 (2011) 368–375. https://doi.org/10.1016/j.optmat.2011.05.024.

[92] E. Clavé, M. Gaft, V. Motto-Ros, C. Fabre, O. Forni, O. Beyssac, S. Maurice, R.C. Wiens, B. Bousquet, Extending the potential of plasma-induced luminescence spectroscopy, Spectrochimica Acta Part B: Atomic Spectroscopy. 177 (2021) 106111. https://doi.org/10.1016/j.sab.2021.106111.

[93] M.C. Martin, C. Dabat-Blondeau, M. Unger, J. Sedlmair, D.Y. Parkinson, H.A. Bechtel, B. Illman, J.M. Castro, M. Keiluweit, D. Buschke, B. Ogle, M.J. Nasse, C.J. Hirschmugl, 3D spectral imaging with synchrotron Fourier transform infrared spectro-microtomography, Nat Methods. 10 (2013) 861–864. https://doi.org/10.1038/nmeth.2596.

[94] G. Capobianco, M.P. Bracciale, D. Sali, F. Sbardella, P. Belloni, G. Bonifazi, S. Serranti, M.L. Santarelli, M. Cestelli Guidi, Chemometrics approach to FT-IR hyperspectral imaging analysis of degradation products in artwork cross-section, Microchemical Journal. 132 (2017) 69–76. https://doi.org/10.1016/j.microc.2017.01.007.

[95] E. Stavitski, R.J. Smith, M.W. Bourassa, A.S. Acerbo, G.L. Carr, L.M. Miller, Dynamic Full-Field Infrared Imaging with Multiple Synchrotron Beams, Anal. Chem. 85 (2013) 3599–3605. https://doi.org/10.1021/ac3033849.

[96] F. Jamme, L. Duponchel, Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis, Journal of Chemometrics. 31 (2017) e2882. https://doi.org/10.1002/cem.2882.

[97] S.D. Keyes, K.R. Daly, N.J. Gostling, D.L. Jones, P. Talboys, B.R. Pinzer, R. Boardman, I. Sinclair, A. Marchant, T. Roose, High resolution synchrotron imaging of wheat root hairs growing in soil and image based modelling of phosphate uptake, New Phytol. 198 (2013) 1023–1029. https://doi.org/10.1111/nph.12294.

[98] A. Kyrieleis, M. Ibison, V. Titarenko, P.J. Withers, Image stitching strategies for tomographic imaging of large objects at high resolution at synchrotron sources, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 607 (2009) 677–684.
https://doi.org/10.1016/j.nima.2009.06.030.

[99] L. Théron, A. Vénien, F. Jamme, X. Fernandez, F. Peyrin, C. Molette, P. Dumas, M. Réfrégiers, T. Astruc, Protein Matrix Involved in the Lipid Retention of Foie Gras during Cooking: A Multimodal Hyperspectral Imaging Study, J. Agric. Food Chem. 62 (2014) 5954–5962. https://doi.org/10.1021/jf5009605.

[100] J. Burger, A. Gowen, Data handling in hyperspectral image analysis, Chemometrics and Intelligent Laboratory Systems. 108 (2011) 13–22.
https://doi.org/10.1016/j.chemolab.2011.04.001.

[101] L. Duponchel, Exploring hyperspectral imaging data sets with topological data analysis, Analytica Chimica Acta. 1000 (2018) 123–131. https://doi.org/10.1016/j.aca.2017.11.029.

[102]C. Shi, L. Wang, Incorporating spatial information in spectral unmixing: A review, Remote Sensing of Environment. 149 (2014) 70–87. https://doi.org/10.1016/j.rse.2014.03.034.

[103] J.J. Liu, J.F. MacGregor, On the extraction of spectral and spatial information from images, Chemometrics and Intelligent Laboratory Systems. 85 (2007) 119–130.
https://doi.org/10.1016/j.chemolab.2006.05.011.

[104] D.G. Blumberg, Subpixel hyperspectral target detection using local spectral and spatial information, J. Appl. Remote Sens. 6 (2012) 063508.
https://doi.org/10.1117/1.JRS.6.063508.

[105] N. Mobaraki, J.M. Amigo, HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis, Chemometrics and Intelligent Laboratory Systems. 172 (2018) 174–187. https://doi.org/10.1016/j.chemolab.2017.11.003.

[106] J.M. Amigo, H. Babamoradi, S. Elcoroaristizabal, Hyperspectral image analysis. A tutorial, Analytica Chimica Acta. 896 (2015) 34–51. https://doi.org/10.1016/j.aca.2015.09.030.

[107] A. de Juan, Multivariate curve resolution for hyperspectral image analysis, in: Data Handling in Science and Technology, Elsevier, 2020: pp. 115–150.
https://doi.org/10.1016/B978-0-444-63977-6.00007-9.

[108] P. Geladi, Chemometrics in spectroscopy. Part 1. Classical chemometrics, Spectrochimica Acta Part B: Atomic Spectroscopy. 58 (2003) 767–782.
https://doi.org/10.1016/S0584-8547(03)00037-5.

[109] P.R. Mobley, B.R. Kowalski, J.J. Workman, R. Bro, Review of Chemometrics Applied to Spectroscopy: 1985-95, Part 2, Applied Spectroscopy Reviews. 31 (1996) 347–368. https://doi.org/10.1080/05704929608000575.

[110] S. Roussel, S. Preys, F. Chauchard, J. Lallemand, Multivariate Data Analysis (Chemometrics). Process Analytical Technology for the Food Industry, Springer New York, New York, NY, 2014: pp. 7–59. https://doi.org/10.1007/978-1-4939-0311-5_2.

[111] M.J. Adams, 2nd ed., Chemometrics in analytical spectroscopy, Royal Society of Chemistry, Cambridge [England], 1995. https://doi.org/10.1021/ja040928h.

[112] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, 0 ed., CRC Press, 2016. https://doi.org/10.1201/9781420059496.

[113] I.E. Frank, J.H. Friedman, A Statistical View of Some Chemometrics Regression Tools, Technometrics. 35 (1993) 109–135. https://doi.org/10.1080/00401706.1993.10485033.

[114] S.J. Haswell, A.D. Walmsley, Chemometrics: the issues of measurement and modelling, Analytica Chimica Acta. 400 (1999) 399–412. https://doi.org/10.1016/S0003-2670(99)00620-0.

[115] A.J. Burnham, J.F. MacGregor, R. Viveros, Latent variable multivariate regression modeling, Chemometrics and Intelligent Laboratory Systems. 48 (1999) 167–180. https://doi.org/10.1016/S0169-7439(99)00018-0.

[116] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[117] R.G.D. Steel, R.A. Bottenberg, J.H. Ward Jr., Applied Multiple Linear Regression. Biometrics. 20 (1964) 652. https://doi.org/10.2307/2528505.

[118] M. Sjöström, S. Wold, W. Lindberg, J.Å. Persson, H. Martens, A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables, Analytica Chimica Acta. 150 (1983) 61–70. https://doi.org/10.1016/S0003-2670(00)85460-4.

[119] F. Marini, R. Bucci, A.L. Magrì, A.D. Magrì, Artificial neural networks in chemometrics: History, examples and perspectives, Microchemical Journal. 88 (2008) 178–185. https://doi.org/10.1016/j.microc.2007.11.008.

[120] F. Marini, Classification Methods in Chemometrics, Curr. Anal. Chem. 6 (2010) 72–79. https://doi.org/ 10.2174/157341110790069592.

[121] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics. 7 (1936) 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[122] B. Park, S. C. Yoon, K. C. Lawrence, W. R. Windham, Fisher Linear Discriminant Analysis for Improving Fecal Detection Accuracy with Hyperspectral Images, Transactions of the ASABE. 50 (2007) 2275–2283. https://doi.org/10.13031/2013.24080.

[123] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemometrics. 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[124] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, R. Nescatelli, F. Marini, Classification and Class-Modelling, in: Data Handling in Science and Technology, Elsevier, 2013: pp. 171–233. https://doi.org/10.1016/B978-0-444-59528-7.00005-3.

[125]S. De Luca, R. Bucci, A.D. Magrì, F. Marini, Class Modeling Techniques in Chemometrics: Theory and Applications, in: R.A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd, Chichester, UK, 2018: pp. 1–24. https://doi.org/10.1002/9780470027318.a9578.

[126] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recognition. 8 (1976) 127–139. https://doi.org/10.1016/0031-3203(76)90014-5.

[127] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. https://doi.org/10.1039/C3AY41907J.

[128] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (1999) 264–323. https://doi.org/10.1145/331499.331504.

[129] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review, Analytica Chimica Acta. 1145 (2021) 59–78. https://doi.org/10.1016/j.aca.2020.10.051.

[130] G.P. Nason, B.W. Silverman, The Stationary Wavelet Transform and some Statistical Applications, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, Springer New York, New York, NY, 1995: pp. 281–299. https://doi.org/10.1007/978-1-4612-2544-7_17.

[131] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J.M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools, Anal Bioanal Chem. 409 (2017) 5891–5899. https://doi.org/10.1007/s00216-017-0517-1.

[132] M.B. Seasholtz, B.R. Kowalski, The effect of mean centering on prediction in multivariate calibration, J. Chemometrics. 6 (1992) 103–111. https://doi.org/10.1002/cem.1180060208.

[133] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra, Appl Spectrosc. 43 (1989) 772–777. https://doi.org/10.1366/0003702894202201.

[134] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–1639. https://doi.org/10.1021/ac60214a047.

[135] E.T. Whittaker, On a New Method of Graduation, Proceedings of the Edinburgh Mathematical Society. 41 (1922) 63–75. https://doi.org/10.1017/S0013091500077853.

[136] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems. 2 (1987) 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

[137] E. Saccenti, J. Camacho, Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods, Chemometrics and Intelligent Laboratory Systems. 149 (2015) 99–116. https://doi.org/10.1016/j.chemolab.2015.10.006.

[138] R.B. Cattell, The Scree Test For The Number Of Factors, Multivariate Behavioral Research. 1 (1966) 245–276. https://doi.org/10.1207/s15327906mbr0102_10.

[139] E. Saccenti, A.K. Smilde, J.A. Westerhuis, M.M.W.B. Hendriks, Tracy-Widom statistic for the largest eigenvalue of autoscaled real matrices: TW statistic for autoscaled real matrices, J. Chemometrics. 25 (2011) 644–652. https://doi.org/10.1002/cem.1411.

[140] J. Camacho, A. Ferrer, Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects: Cross-validation in PCA with ekf algorithm, J. Chemometrics. 26 (2012) 361–373. https://doi.org/10.1002/cem.2440.

[141] S. Dray, On the number of principal components: A test of dimensionality based on measurements of similarity between matrices, Computational Statistics & Data Analysis. 52 (2008) 2228–2237. https://doi.org/10.1016/j.csda.2007.07.015.

[142] Ting Su, J. Dy, A deterministic method for initializing K-means clustering, in: 16th IEEE International Conference on Tools with Artificial Intelligence, IEEE Comput. Soc, Boca Raton, FL, USA, 2004: pp. 784–786. https://doi.org/10.1109/ICTAI.2004.7.

[143] M.E. Celebi, H.A. Kingravi, P.A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications. 40 (2013) 200–210. https://doi.org/10.1016/j.eswa.2012.07.021.

[144] J.M. Peña, J.A. Lozano, P. Larrañaga, An empirical comparison of four initialization methods for the K-Means algorithm, Pattern Recognition Letters. 20 (1999) 1027–1040. https://doi.org/10.1016/S0167-8655(99)00069-0.

[145] S.S. Khan, A. Ahmad, Cluster center initialization algorithm for K-means clustering, Pattern Recognition Letters. 25 (2004) 1293–1302. https://doi.org/10.1016/j.patrec.2004.04.007.

[146] M.C. Naldi, A. Fontana, R.J.G.B. Campello, Comparison Among Methods for k Estimation in k-means, in: 2009 Ninth International Conference on Intelligent Systems Design and Applications, IEEE, Pisa, Italy, 2009: pp. 1006–1013. https://doi.org/10.1109/ISDA.2009.78.

[147] R. Bro, Multivariate calibration, Analytica Chimica Acta. 500 (2003) 185–194. https://doi.org/10.1016/S0003-2670(03)00681-0.

[148] M. Haenlein, A.M. Kaplan, A Beginner's Guide to Partial Least Squares Analysis, Understanding Statistics. 3 (2004) 283–297. https://doi.org/10.1207/s15328031us0304_4.

[149] L. Breiman, J.H. Friedman, Predicting Multivariate Responses in Multiple Linear Regression, J Royal Statistical Soc B. 59 (1997) 3–54. https://doi.org/10.1111/1467-9868.00054.

[150] A. Höskuldsson, PLS regression methods, J. Chemometrics. 2 (1988) 211–228. https://doi.org/10.1002/cem.1180020306.

[151] A. Höskuldsson, Variable and subset selection in PLS regression, Chemometrics and Intelligent Laboratory Systems. 55 (2001) 23–38. https://doi.org/10.1016/S0169-7439(00)00113-1.

[152] S. Chevallier, D. Bertrand, A. Kohler, P. Courcoux, Application of PLS-DA in multivariate image analysis, J. Chemometrics. 20 (2006) 221–229. https://doi.org/10.1002/cem.994.

[153] U.G. Indahl, H. Martens, From dummy regression to prior probabilities in PLS-DA, J. Chemometrics. 21 (2007) 529–536. https://doi.org/10.1002/cem.1061.

[154] O.M. Kvalheim, T.V. Karstang, SIMCA - Classification by Means of Disjoint Cross Validated Principal Components Models, in: Data Handling in Science and Technology, Elsevier, 1992: pp. 209–248. https://doi.org/10.1016/S0922-3487(08)70207-7.

[155] M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, Chemometrics and Intelligent Laboratory Systems. 93 (2008) 132–148. https://doi.org/10.1016/j.chemolab.2008.05.003.

[156] P. Nomikos, J.F. MacGregor, Multivariate SPC Charts for Monitoring Batch Processes, Technometrics. 37 (1995) 41–59. https://doi.org/10.1080/00401706.1995.10485888.

[157] W.H. Lawton, E.A. Sylvestre, Self Modeling Curve Resolution, Technometrics. 13 (1971) 617–633. https://doi.org/10.1080/00401706.1971.10488823.

[158] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods. 6 (2014) 4964–4976. https://doi.org/10.1039/C4AY00571F.

[159] E.R. Malinowski, Window factor analysis: Theoretical derivation and application to flow injection analysis data, J. Chemometrics. 6 (1992) 29–40. https://doi.org/10.1002/cem.1180060104.

[160] R. Manne, H. Shen, Y. Liang, Subwindow factor analysis, Chemometrics and Intelligent Laboratory Systems. 45 (1999) 171–176. https://doi.org/10.1016/S0169-7439(98)00101-4.

[161] O.M. Kvalheim, Y.Z. Liang, Heuristic evolving latent projections: resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution, Anal. Chem. 64 (1992) 936–946. https://doi.org/10.1021/ac00032a019.

[162] R. Tauler, Multivariate curve resolution applied to second order data, Chemometrics and Intelligent Laboratory Systems. 30 (1995) 133–146. https://doi.org/10.1016/0169-7439(95)00047-X.

[163] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemometrics and Intelligent Laboratory Systems. 76 (2005) 101–110. https://doi.org/10.1016/j.chemolab.2004.12.007.

[164] P.J. Gemperline, A priori estimates of the elution profiles of the pure components in overlapped liquid chromatography peaks using target factor analysis, J. Chem. Inf. Model. 24 (1984) 206–212. https://doi.org/10.1021/ci00044a004.

[165] B.G.M. Vandeginste, W. Derks, G. Kateman, Multicomponent self-modelling curve resolution in high-performance liquid chromatography by iterative target transformation analysis, Analytica Chimica Acta. 173 (1985) 253–264. https://doi.org/10.1016/S0003-2670(00)84962-4.

[166] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, Journal of Chemometrics. 9 (1995) 31–58. https://doi.org/10.1002/cem.1180090105.

[167] A. de Juan, Y. Heyden, R. Tauler, D. Massart, Assessment of new constraints applied to the alternating least squares method, Analytica Chimica Acta. 346 (1997) 307–318. https://doi.org/10.1016/S0003-2670(97)90069-6.

[168] M. Sawall, A. Jürß, H. Schröder, K. Neymeyr, On the Analysis and Computation of the Area of Feasible Solutions for Two-, Three-, and Four-Component Systems, in: Data Handling in Science and Technology, Elsevier, 2016: pp. 135–184. https://doi.org/10.1016/B978-0-444-63638-6.00005-X.

[169] A. de Juan, M. Maeder, M. Martínez, R. Tauler, Combining hard- and soft-modelling to solve kinetic problems, Chemometrics and Intelligent Laboratory Systems. 54 (2000) 123–141. https://doi.org/10.1016/S0169-7439(00)00112-X.

[170] J. Diewok, A. de Juan, M. Maeder, R. Tauler, B. Lendl, Application of a Combination of Hard and Soft Modeling for Equilibrium Systems to the Quantitative Analysis of pH-Modulated Mixture Samples, Anal. Chem. 75 (2003) 641–647. https://doi.org/10.1021/ac026248j.

[171] O.S. Borgen, B.R. Kowalski, An extension of the multivariate component-resolution method to three components, Analytica Chimica Acta. 174 (1985) 1–26. https://doi.org/10.1016/S0003-2670(00)84361-5.

[172] R. Tauler, A. Izquierdo-Ridorsa, E. Casassas, Simultaneous analysis of several spectroscopic titrations with self-modelling curve resolution, Chemometrics and Intelligent Laboratory Systems. 18 (1993) 293–300. https://doi.org/10.1016/0169-7439(93)85006-3.

[173] P.J. Gemperline, Target transformation factor analysis with linear inequality constraints applied to spectroscopic-chromatographic data, Anal. Chem. 58 (1986) 2656–2663. https://doi.org/10.1021/ac00126a018.

[174] J.C. Hamilton, P.J. Gemperline, Mixture analysis using factor analysis. II: Self-modeling curve resolution, J. Chemometrics. 4 (1990) 1–13. https://doi.org/10.1002/cem.1180040103.

[175] S. Bijlsma, A.K. Smilde, Estimating reaction rate constants from a two-step reaction: a comparison between two-way and three-way methods, J. Chemometrics. 14 (2000) 541–560. https://doi.org/10.1002/1099-128X(200009/12)14:5/63.0.CO;2-1.

[176] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbuehler, Quantification of a known component in an unknown mixture, Analytica Chimica Acta. 193 (1987) 287–293. https://doi.org/10.1016/S0003-2670(00)86160-7.

[177] K. Pearson, On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 2 (1901) 559–572. https://doi.org/10.1080/14786440109462720.

[178] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology. 24 (1933) 498–520. https://doi.org/10.1037/h0070888.

[179] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Analytical Chemistry. 63 (1991) 1425–1432. https://doi.org/10.1021/ac00014a016.

[180] A. Nardecchia, L. Duponchel, Randomised SIMPLISMA: Using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging, Talanta. 217 (2020) 121024. https://doi.org/10.1016/j.talanta.2020.121024.

[181] A. Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures, Analytica Chimica Acta. 500 (2003) 195–210. https://doi.org/10.1016/S0003-2670(03)00724-4.

[182] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: New features and applications, Chemometrics and Intelligent Laboratory Systems. 140 (2015) 1–12. https://doi.org/10.1016/j.chemolab.2014.10.003.

[183] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Critical Reviews in Analytical Chemistry. 36 (2006) 163–176. https://doi.org/10.1080/10408340600970005.

[184] J. Jaumot, Multivariate curve resolution: a powerful tool for the analysis of conformational transitions in nucleic acids, Nucleic Acids Research. 30 (2002) e92. https://doi.org/10.1093/nar/gnf091.

[185] L. Duponchel, W. Elmi-Rayaleh, C. Ruckebusch, J.P. Huvenne, Multivariate Curve Resolution Methods in Imaging Spectroscopy: Influence of Extraction Methods and Instrumental Perturbations, J. Chem. Inf. Comput. Sci. 43 (2003) 2057–2067. https://doi.org/10.1021/ci034097v.

[186] X. Zhang, R. Tauler, Application of Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) to remote sensing hyperspectral imaging, Analytica Chimica Acta. 762 (2013) 25–38. https://doi.org/10.1016/j.aca.2012.11.043.

[187] W.F.J. Vermaas, J.A. Timlin, H.D.T. Jones, M.B. Sinclair, L.T. Nieman, S.W. Hamad, D.K. Melgaard, D.M. Haaland, In vivo hyperspectral confocal fluorescence imaging to determine pigment localization and distribution in cyanobacterial cells, Proceedings of the National Academy of Sciences. 105 (2008) 4050–4055. https://doi.org/10.1073/pnas.0708090105.

[188] H.D.T. Jones, D.M. Haaland, M.B. Sinclair, D.K. Melgaard, A.M. Collins, J.A. Timlin, Preprocessing strategies to improve MCR analyses of hyperspectral images, Chemometrics and Intelligent Laboratory Systems. 117 (2012) 149–158. https://doi.org/10.1016/j.chemolab.2012.01.011.

[189] J.J. Andrew, M.A. Browne, I.E. Clark, T.M. Hancewicz, A.J. Millichope, Raman Imaging of Emulsion Systems, Appl Spectrosc. 52 (1998) 790–796. https://doi.org/10.1366/0003702981944472.

[190] W. Windig, J.L. Lippert, M.J. Robbins, K.R. Kresinske, J.P. Twist, A.P. Snyder, Interactive self-modeling multivariate analysis, Chemometrics and Intelligent Laboratory Systems. 9 (1990) 7–30. https://doi.org/10.1016/0169-7439(90)80050-G.

[191] A.P. Snyder, W. Windig, J.P. Toth, Interactive self-modeling multivariate analysis of thermolysis mass spectra, Chemometrics and Intelligent Laboratory Systems. 11 (1991) 149–160. https://doi.org/10.1016/0169-7439(91)80062-U.

[192] W. Windig, C.E. Heckler, F.A. Agblevor, R.J. Evans, Self-modeling mixture analysis of categorized pyrolysis mass spectral data with the SIMPLISMA approach, Chemometrics and Intelligent Laboratory Systems. 14 (1992) 195–207. https://doi.org/10.1016/0169-7439(92)80104-C.

[193] S. Gourvénec, D.L. Massart, D.N. Rutledge, Determination of the number of components during mixture analysis using the Durbin–Watson criterion in the Orthogonal Projection Approach and in the SIMPLe-to-use Interactive Self-modelling Mixture Analysis approach, Chemometrics and Intelligent Laboratory Systems. 61 (2002) 51–61. https://doi.org/10.1016/S0169-7439(01)00172-1.

[194] B.M. Wise, P. Geladi, A Brief Introduction to Multivariate Image Analysis (MIA), Physics, (2000). Retrieved from URL (Accessed the 01/03/2022): http://www.eigenvector.com/Docs/MIA_Intro.pdf.

[195]P. Geladi, H.F. Grahn, Multivariate Image Analysis, (2016). https://doi.org/10.1002/9780470027318.a8106.pub3.

[196] K. Esbensen, P. Geladi, Strategy of multivariate image analysis (MIA), Chemometrics and Intelligent Laboratory Systems. 7 (1989) 67–86. https://doi.org/10.1016/0169-7439(89)80112-1.

[197] J.M. Combes, A. Grossmann, P. Tchamitchian, eds., Wavelets: Time-Frequency Methods and Phase Space, Springer Berlin Heidelberg, Berlin, Heidelberg, 1989. https://doi.org/10.1007/978-3-642-97177-8.

[198] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Trans. Inform. Theory. 36 (1990) 961–1005. https://doi.org/10.1109/18.57199.

[199] R.E.A.C. Paley, N. Wiener, Fourier Transforms in the Complex Domain. (1934). https://doi.org/10.1007/BF01699343.

[200] G. Strang, Wavelet transforms versus Fourier transforms, Bull. Amer. Math. Soc. 28 (1993) 288–306. https://doi.org/10.1090/S0273-0979-1993-00390-2.

[201] M.J. Shensa, The discrete wavelet transform: wedding the a trous and Mallat algorithms, IEEE Trans. Signal Process. 40 (1992) 2464–2482. https://doi.org/10.1109/78.157290.

[202] G.P. Nason, B.W. Silverman, The Stationary Wavelet Transform and some Statistical Applications, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, Springer New York, New York, NY, 1995: pp. 281–299. https://doi.org/10.1007/978-1-4612-2544-7_17.

[203] J. Kovacevic, W. Sweldens, Wavelet families of increasing order in arbitrary dimensions, IEEE Trans. on Image Process. 9 (2000) 480–496. https://doi.org/10.1109/83.826784.

[204] C. Vonesch, T. Blu, M. Unser, Generalized Daubechies Wavelet Families, IEEE Trans. Signal Process. 55 (2007) 4415–4429. https://doi.org/10.1109/TSP.2007.896255.

[205] H. Liang, Advances in multispectral and hyperspectral imaging for archaeology and art conservation, Appl. Phys. A. 106 (2012) 309–323. https://doi.org/10.1007/s00339-011-6689-1.

[206] A. de Juan, Hyperspectral image analysis. When space meets Chemistry, Journal of Chemometrics. 32 (2018) e2985. https://doi.org/10.1002/cem.2985.

[207] K. Kumar, Principal component analysis: Most favourite tool in chemometrics, Reson. 22 (2017) 747–759. https://doi.org/10.1007/s12045-017-0523-9.

[208] D.T. Pham, S.S. Dimov, C.D. Nguyen, Selection of K in K-means clustering, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 219 (2005) 103–119. https://doi.org/10.1243/095440605X8298.

[209] S. Hugelier, O. Devos, C. Ruckebusch, On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis: Spatial constraints in HSI-MCR-ALS, J. Chemometrics. 29 (2015) 557–561. https://doi.org/10.1002/cem.2742.

[210] J.O. Cáceres, F. Pelascini, V. Motto-Ros, S. Moncayo, F. Trichard, G. Panczer, A. Marín-Roldán, J.A. Cruz, I. Coronado, J. Martín-Chivelet, Megapixel multi-elemental imaging by Laser-Induced Breakdown Spectroscopy, a technology with considerable potential for paleoclimate studies, Sci Rep. 7 (2017) 5080. https://doi.org/10.1038/s41598-017-05437-3.

[211] 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.

[212] A. Nardecchia, C. Fabre, J. Cauzid, F. Pelascini, V. Motto-Ros, L. Duponchel, Detection of minor compounds in complex mineral samples from millions of spectra: A new data analysis strategy in LIBS imaging, Analytica Chimica Acta. 1114 (2020) 66–73. https://doi.org/10.1016/j.aca.2020.04.005.

[213] A. Nardecchia, V. Motto-Ros, L. Duponchel, Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed phenomenon?, Analytica Chimica Acta. 1157 (2021) 338389. https://doi.org/10.1016/j.aca.2021.338389.

[214] F. Scheuren, Multiple Imputation: How It Began and Continues, The American Statistician. 59 (2005) 315–319. https://doi.org/10.1198/000313005X74016.

[215] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, J. Stat. Soft. 45 (2011). https://doi.org/10.18637/jss.v045.i03.

[216] A. Nardecchia, A. de Juan, V. Motto-Ros, M. Gaft, L. Duponchel, Data fusion of LIBS and PIL hyperspectral imaging: Understanding the luminescence phenomenon of a complex mineral sample, Analytica Chimica Acta. 1192 (2022) 339368. https://doi.org/10.1016/j.aca.2021.339368.

[217] M. Li Vigni, M. Cocchi, Multiresolution Analysis and Chemometrics for Pattern Enhancement and Resolution in Spectral Signals and Images, in: Data Handling in Science and

Technology, Elsevier, 2016: pp. 409–451. https://doi.org/10.1016/B978-0-444-63638-6.00013-9.

[218] A. Nardecchia, R. Vitale, L. Duponchel, Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images, Talanta. 224 (2021) 121835. https://doi.org/10.1016/j.talanta.2020.121835.

[219] V.H. da Silva, F. Murphy, J.M. Amigo, C. Stedmon, J. Strand, Classification and Quantification of Microplastics (<100 μm) Using a Focal Plane Array–Fourier Transform Infrared Imaging System and Machine Learning, Anal. Chem. 92 (2020) 13724–13733. https://doi.org/10.1021/acs.analchem.0c01324.

LIST OF PUBLICATIONS AND CONFERENCES

This thesis has been a good opportunity to present the work carried out during these three years at national and international levels. Here below are reported the published papers (and the ones submitted) related to this PhD project and the conferences and posters in which the different works had the possibility to be presented.

PUBLICATIONS

Already published:

1) A. Nardecchia, L. Duponchel, *Randomised SIMPLISMA: Using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging*, Talanta. 217 (2020) 121024.
https://doi.org/10.1016/j.talanta.2020.121024.
2) A. Nardecchia, C. Fabre, J. Cauzid, F. Pelascini, V. Motto-Ros, L. Duponchel, *Detection of minor compounds in complex mineral samples from millions of spectra: A new data analysis strategy in LIBS imaging*, Analytica Chimica Acta. 1114 (2020) 66–73.
https://doi.org/10.1016/j.aca.2020.04.005.
3) A. Nardecchia, V. Motto-Ros, L. Duponchel, *Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed phenomenon?*, Analytica Chimica Acta. 1157 (2021) 338389.
https://doi.org/10.1016/j.aca.2021.338389.
4) A. Nardecchia, R. Vitale, L. Duponchel, *Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images,* Talanta. 224 (2021) 121835.
https://doi.org/10.1016/j.talanta.2020.121835.
5) A. Nardecchia, A. de Juan, V. Motto-Ros, M. Gaft, L. Duponchel, *Data fusion of LIBS and PIL hyperspectral imaging: Understanding the luminescence phenomenon of a complex mineral sample*, Analytica Chimica Acta. 1192 (2022) 339368.
https://doi.org/10.1016/j.aca.2021.339368.

Submitted/to be submitted:

6) A. Nardecchia, A. de Juan, V. Motto-Ros, C. Fabre, L. Duponchel, *LIBS and Raman data fusion: an interesting analysis approach based on the use of chemometric methodologies,* To be submitted to: Spectrochimica Acta Part B: Atomic Spectroscopy (2022).
7) A. Nardecchia, R. Vitale, L. Duponchel, *Application of the 2-D stationary wavelet transform (SWT 2-D) for the use of spectral and spatial hyperspectral image information in the framework of classification analysis*, To be submitted to: Talanta (2022).

CONFERENCES

For informational purposes, the speaker presenting the work in the given conferences is here reported in the list underlining its name.

International conferences:

1) *Exploring hyperspectral imaging data sets with pixel resampling. The randomized SIMPLISMA approach*, L. Duponchel, A. Nardecchia, TIC 2019 (Topics In Chemometrics 2019), Szeged, Hungary (15-18 May 2019).
2) *Randomised SIMPLISMA: facilitating rank evaluation and initial guesses generation in the framework of spectral unmixing and spectroscopic imaging*, A. Nardecchia, L. Duponchel, 10TH COLLOQUIUM CHEMIOMETRICUM MEDITERRANEUM, Menorca, Spain (12-14 June 2019).
3) *Randomization as a way to explore hyperspectral imaging data sets*, L. Duponchel, A. Nardecchia, 10TH COLLOQUIUM CHEMIOMETRICUM MEDITERRANEUM, Menorca, Spain (12-14 June 2019).
4) *Classification in NIR hyperspectral imaging: The importance of using both spectral and spatial information*, A. Nardecchia, R. Vitale, L. Duponchel, SWIIMS 2021 (Short Wave Infrared IMaging and Spectroscopy 2021), Amsterdam, Netherlands, online conference due to COVID pandemic (24-26 March 2021).
5) *Chemometric methods and strategies for laser-induced breakdown spectroscopy (LIBS) hyperspectral images analysis: compression and data fusion*, A. Nardecchia, A. de Juan, V. Motto-Ros, M. Gaft, L. Duponchel, Road to CAC 2022 (Chemometrics in Analytical Chemistry 2021), online conference due to COVID pandemic (20-21 July 2021).
6) *Compression and data fusion strategies for laser-induced breakdown spectroscopy (LIBS) and plasma induced luminescence (PIL) hyperspectral images*, A. Nardecchia, A. de Juan, V.

Motto-Ros, M. Gaft, L. Duponchel, SCIX 2021 (SCIentific eXchange 2021), Providence, USA, online conference due to COVID pandemic (26 September-01 October 2021).

7) *Saturated signals in LIBS imaging: why and how we should correct for these artifacts in the framework of multivariate analysis?*, L. Duponchel, A. Nardecchia, V. Motto-Ros, SCIX 2021 (SCIentific eXchange 2021), Providence, USA, online conference due to COVID pandemic (26 September-01 October 2021).
8) *Compression and data fusion in the framework of hyperspectral LIBS imaging*, A. Nardecchia, A. de Juan, V. Motto-Ros, M. Gaft, L. Duponchel, EMSLIBS 2021 (Euro-Mediterranean Symposium on Laser-Induced Breakdown Spectroscopy 2021), Gijón, Spain (29 November-02 December 2021).
9) *Saturated signals in LIBS imaging: why and how we should correct for these artifacts in the framework of multivariate analysis?*, L. Duponchel, A. Nardecchia, V. Motto-Ros, EMSLIBS 2021 (Euro-Mediterranean Symposium on Laser-Induced Breakdown Spectroscopy 2021), Gijón, Spain (29 November-02 December 2021).

National conferences:

10) *Chemometrics exploration in hyperspectral imaging in the framework of multimodality and big data*, A. Nardecchia, L. Duponchel, Journée des doctorants 2019, Lille, France (04 July 2019).
11) *Imageries multiéchelles et multispectrales en fluorescence dans l'UV-visible et l'UV profond pour le suivi des couches externes du grain de blé au cours du développement* (poster presentation), C. Alvarado, M.F. Devaux, A. L. Chateigner-Boutin, F. Guillon, L. Hélary, S. Durand, F. Jamme, A. Nardecchia, L. Duponchel, 9eme JST du RMUI (9^{e} Journées Scientifiques et Techniques du Réseau des Microscopistes de l'Inra), Nantes, France (27-29 Novembre 2019).
12) *Spectral and spatial fusion: An interesting approach for classification in hyperspectral imaging*, A. Nardecchia, R. Vitale, L. Duponchel, e-Chimiométrie 2021, online conference due to COVID pandemic (2-3 February 2021).