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

Alessandro Nardecchia, Cécile Fabre, Jean Cauzid, Frederic Pelascini, Vincent Motto-Ros, Ludovic Duponchel

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

Alessandro Nardecchia[†], Cécile Fabre[⊥], Jean Cauzid[⊥], Frédéric Pelascini[§], Vincent Motto-Ros[‡*] and Ludovic Duponchel[†*]

[†] Univ. Lille, CNRS, UMR 8516 – LASIRe – LAboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France.

[⊥] Université de Lorraine, Laboratoire GeoRessources, UMR CNRS 7359, France.

[§] Cetim Grand Est, Illkirch-Graffenstaden, France.

[‡] Institut Lumière Matière, UMR 5306, Université Lyon 1 - CNRS, Université de Lyon 69622 Villeurbanne, France.

Corresponding Authors

* ludovic.duponchel@univ-lille.fr and vincent.motto-ros@univ-lyon1.fr

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.

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 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 (≥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 wave-

lengths, 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.

EXPERIMENTAL SECTION

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 x 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) .

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 15x 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 x 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 six hours,

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.

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 x 5 pixels x 2 wavelengths consists of 25 pseudo-spectra. Figure 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 figure 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 co-

sine 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:

$$d_{i,j} = 1 - \frac{\vec{x}_i.\vec{x}_j}{\|\vec{x}_i\|.\|\vec{x}_j\|} \quad (1)$$

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:

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

with $ER = \frac{max_{l,m=1,\ldots,k}\|\boldsymbol{c}_l - \boldsymbol{c}_m\|}{k}$, $RA = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\|\boldsymbol{x}_{i(j)} - \boldsymbol{c}_i\|}{\sum_{i=1}^{n}\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|}$, $\boldsymbol{c}_i$ the centroid of the i^{th} cluster (i=1…k), $\boldsymbol{x}_{i(j)}$ the j^{th} spectrum of the cluster i, n_i the total number of spectra in the cluster i, and $\bar{\boldsymbol{x}}$ 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 (figure 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 (figure 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.

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 gen-

erate 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. Figure 2 illustrates the use of this conventional approach to the rock section. More specifically, figure 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. Figure 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 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, figures 2c and 2d 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 po-

tential anti-correlation 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). Figure 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, figure 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. Figure 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 (Figure S1). We then discover that all mineral phases exhibit subpopulations 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. Figure 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 exbibit 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 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 figure 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.

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.

AUTHOR INFORMATION

ORCID

Ludovic Duponchel: 0000-0002-7206-4498

Cécile Fabre: 0000-0001-8627-4050

Jean Cauzid: 0000-0001-5587-9874

Frédéric Pelascini: 0000-0002-3779-6685

Vincent Motto-Ros: 0000-0002-6955-3098

Author Contributions

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

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 Mininère de Touissit) for providing the sample. Lastly, we would like to thank Ms Astrid Marissal for proofreading.

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 La-

ser-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, Volume 1: Statistics, 1967.

[13] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recognit. 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.

FIGURES

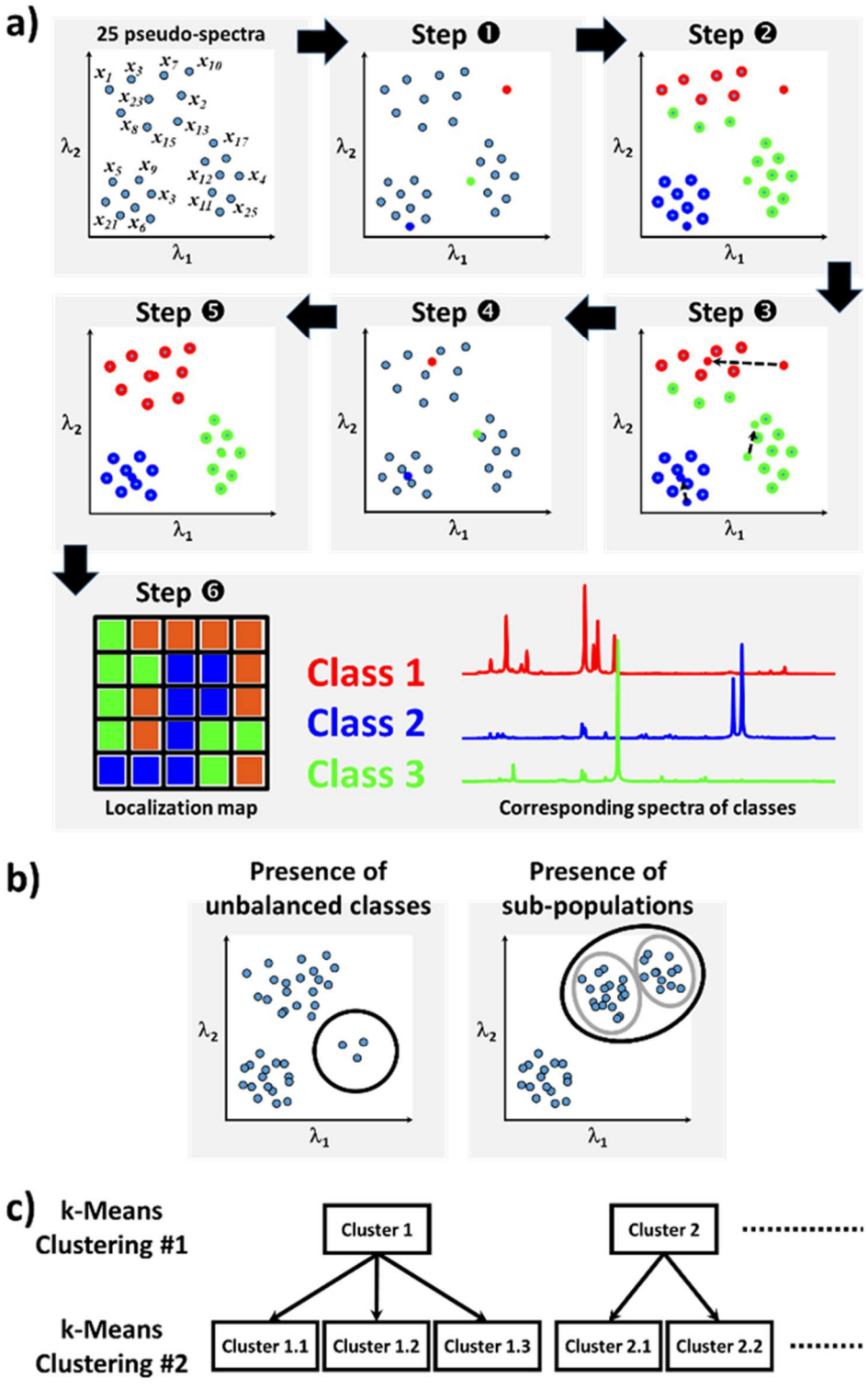


Figure 1. a) The k-means algorithm applied to spectroscopic imaging. b) Problematic data structures hardly managed by k-means. c) The proposed method called embedded k-means clustering.

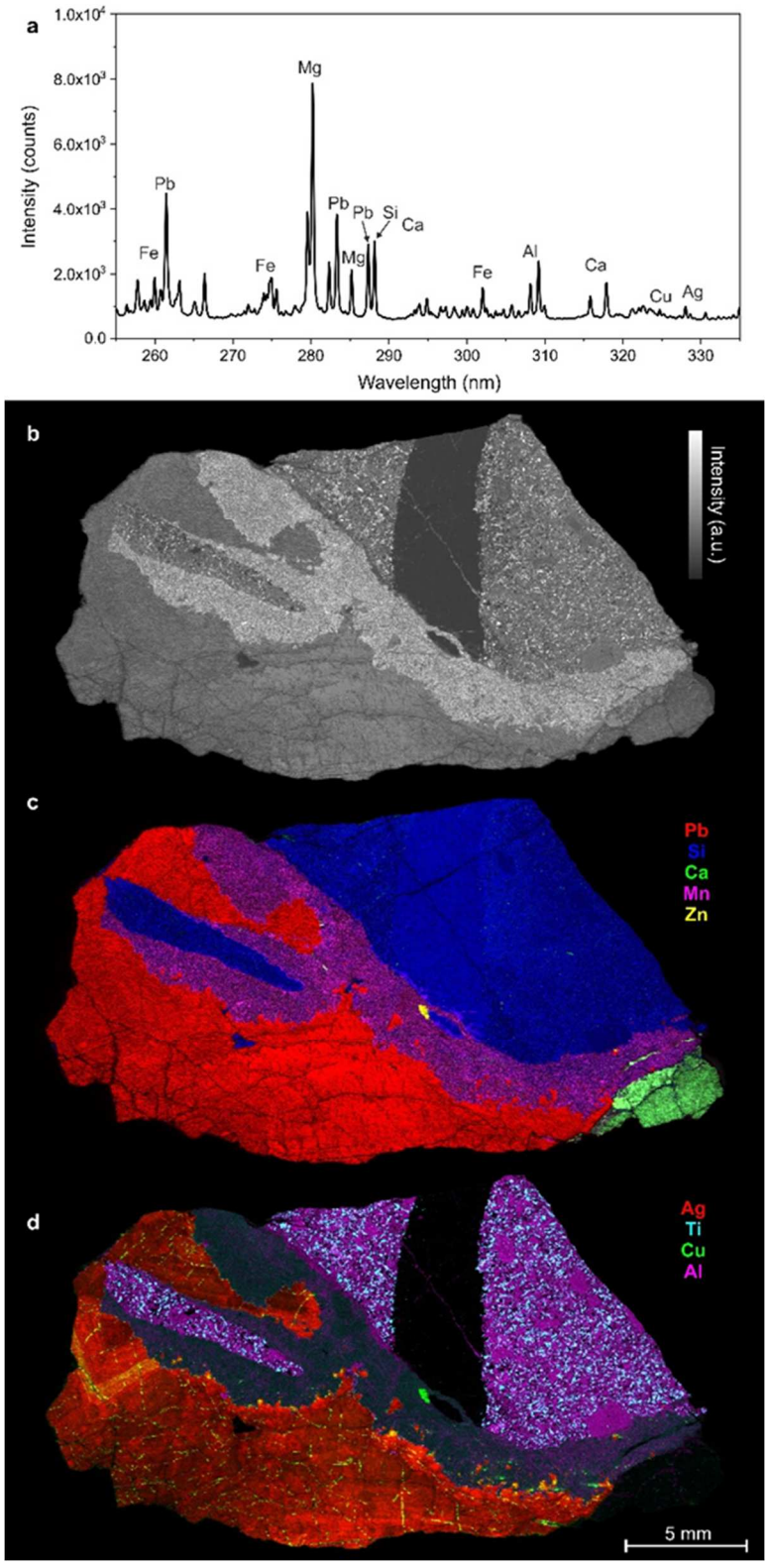
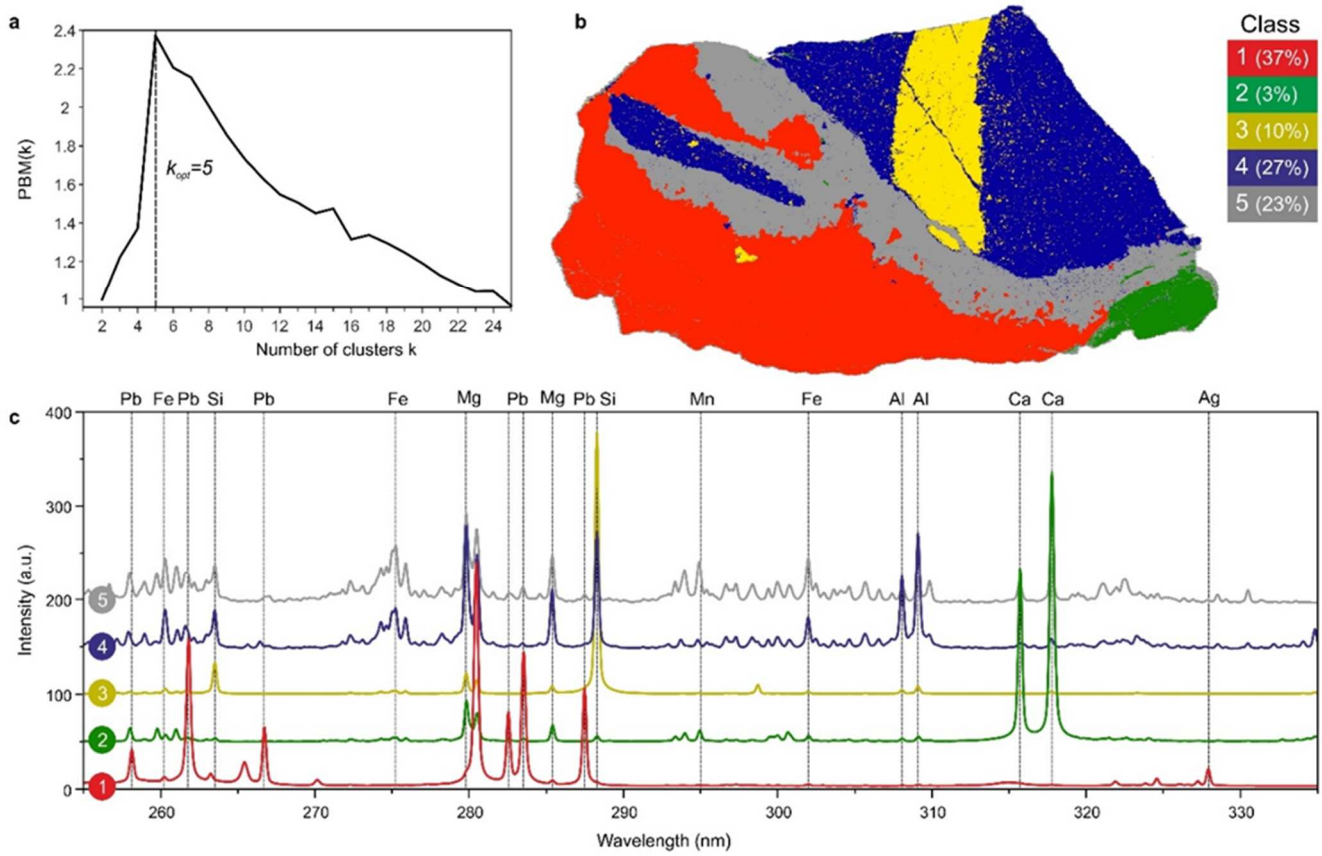


Figure 2. a) The mean spectrum of the LIBS data set. b) The global intensity image. c) and d) Elemental images generated with the conventional approach. A high-resolution version of this image can be downloaded from supplementary materials.



383

384 **Figure 3.** a) Evolution of the PBM index according to the number of clusters *k*. b) The classification

385 map considering an optimal number of clusters equal to 5. c) Representative spectra of each class.

386

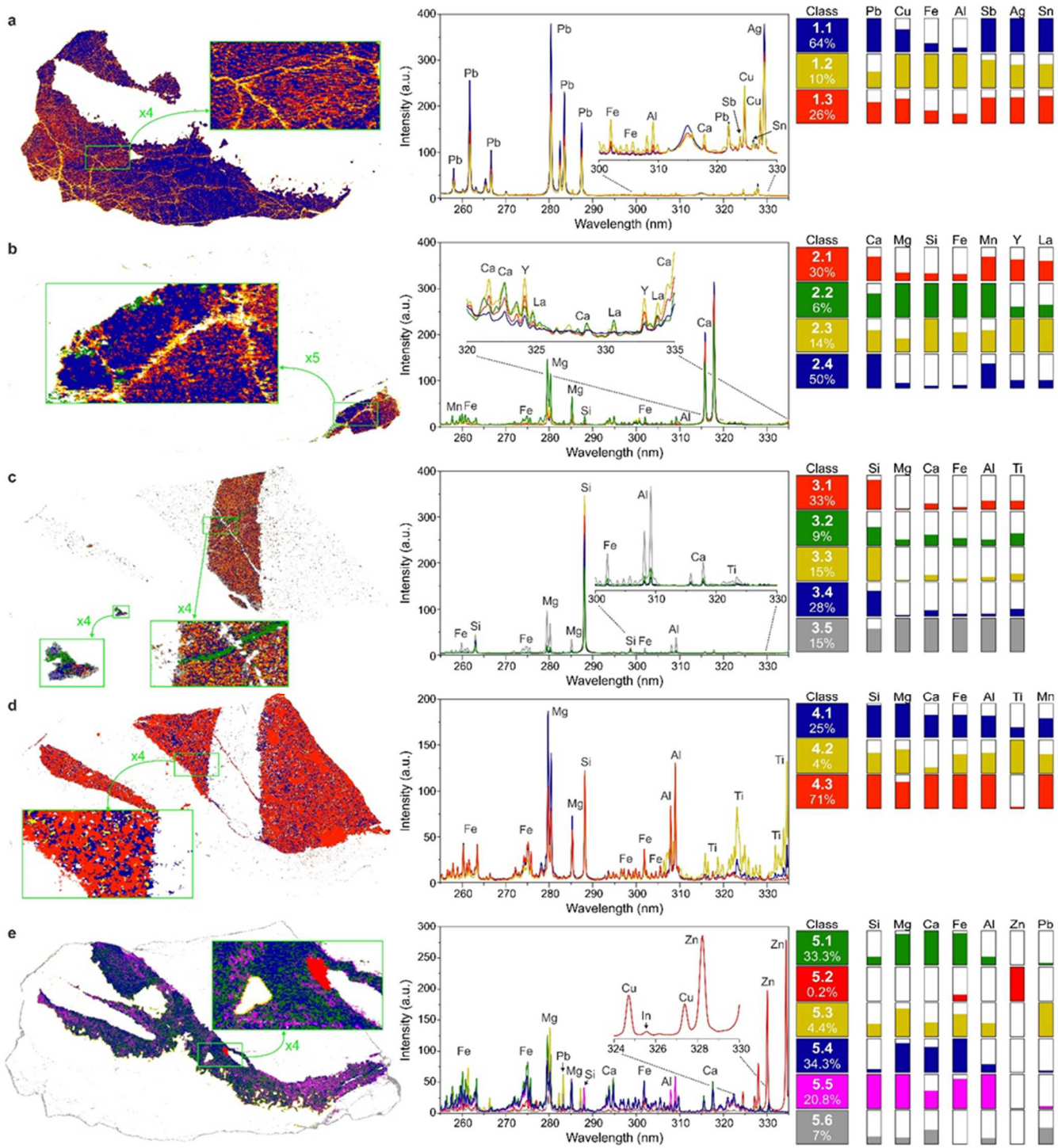
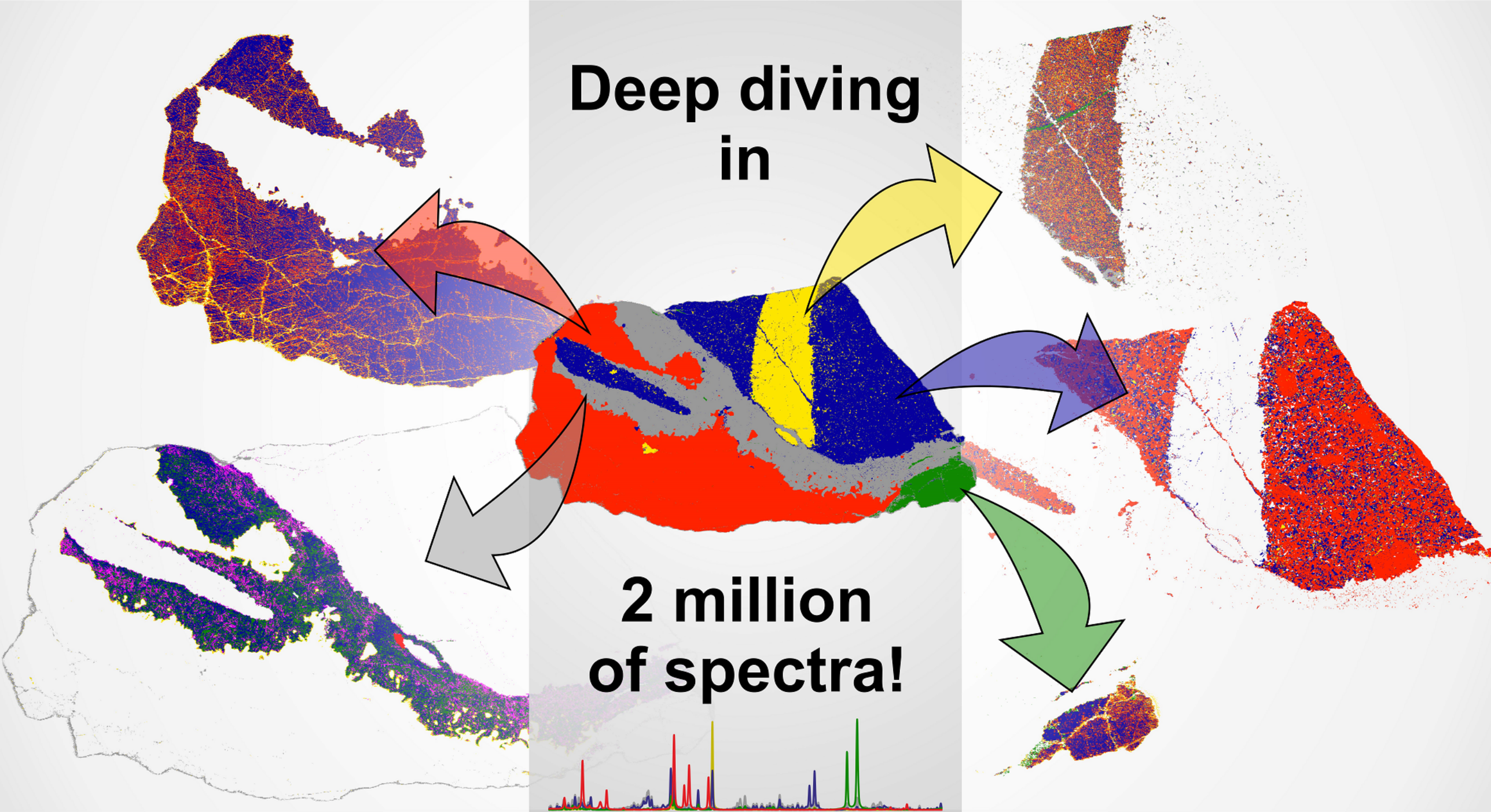


Figure 4. Classification maps obtained for each phase with corresponding spectra of sub-populations and relative concentrations of elements.


Deep diving
in
2 million
of spectra!