Hierarchical clustering of spectral images with spatial constraints for the rapid processing1of large and heterogeneous datasets from ancient material studies

Gilles Celeux, Serge X. Cohen, Agnès Grimaud, Pierre Gueriau

HAL Id: hal-03104488
https://hal.uvsq.fr/hal-03104488v1

Preprint submitted on 8 Jan 2021 (v1), last revised 10 Apr 2022 (v4)

Hierarchical clustering of spectral images with spatial constraints for the rapid processing of large and heterogeneous datasets from ancient material studies

Gilles Celeux[*], Serge X. Cohen[†], Agnès Grimaud[‡], and Pierre Gueriau[§]

Abstract. The study of very complex and heterogeneous materials, such as those encountered in the science of ancient materials, benefits from the wealth of information provided by the acquisition and exploitation of full spectrum images, *i.e.* spectral images. In order to obtain a high dynamic range in both the spatial dimensions and composition, great efforts have made it possible to considerably accelerate data collection and increase the average size of a single data set, each image reaching up to several tens of GB. Rapid processing is now required to allow feedback during data collection, within the short time available for instruments and samples. Here we propose an approach combining hierarchical clustering and spatial constraint. Spatial constraints allow both a significant reduction in the computational cost of segmentation and a certain level of robustness with respect to the signal-to-noise ratio: the *prior* knowledge injected by the spatial constraint partially compensates for the increase in noise level; hierarchical clustering provides a statistically sound and known framework that allows accurate reporting of the instrument noise model. We illustrate the proposed algorithm on a X-ray fluorescence spectral image collected on an *ca.* 100 Myr fossil fish, as well as on simulated data to assess the sensitivity of the results to the noise level. It can be foreseen how such an approach could simultaneously lead to an increase in the spatial definition of the collected spectral image and to a reduction in the potentially harmful radiation dose density to which the samples are subjected.

Key words. Spectral image segmentation, Ward criterion, spatial constraint, ancient material, X-ray fluorescence

1. Introduction

Ancient materials, studied by archaeology, paleontology or as part of the cultural heritage research, are very diverse but share the particularity of being composite and heterogeneous on several scales [6]. Moreover, they are the results of multiple processes at various time scales, inducing strong constraints in terms of handling and physico-chemical characterization whilst often having limited *a priori* certainties concerning them [5, 4]. In this context, spectral imaging, *i.e.* images for which each pixel is characterized by a full spectrum (see *e.g.* Figure 1), is a tool of

[*] Inria Saclay-Île-de-France, IMO campus d'Orsay 91405 Orsay.

[†] IPANEMA, CNRS, ministère de la Culture, UVSQ, MNHN, USR3461, Université Paris-Saclay, 91192 Gif-sur-Yvette, France (serge.cohen@ipanema-remote.fr).

[‡] Université Paris-Saclay, UVSQ, CNRS, Laboratoire de Mathématiques de Versailles, 78000 Versailles, France.

[§] Institute of Earth Sciences, University of Lausanne, Géopolis, CH-1015 Lausanne, Switzerland.

choice for simultaneously obtaining physico-chemical information (*e.g.* elemental, chemical or mineralogical composition), and the morphological information essential for understanding the behavior of the material over long periods of time. Such datasets make several GB or even tens of GB, depending on the type of detection used. Indeed, while 1D detectors typically record thousands of values (*i.e.* a few KB) per pixel, 2D array detectors record images of several MB per pixel (*e.g.* [1, 3, 16]). As such, these datasets are too massive to be timely exploited with standard algorithms and we need to develop algorithms able to analyze such images in a timeframe compatible with the data collection time to provide feedback possibilities on the measurements (*e.g.* [1, 3]). One should also consider that on those measurements involving a probe, increasing the signal to noise ratio (SNR) comes at a cost: increasing probe/material interaction indeed most often leads to longer measurement times and always to a higher radiation dose deposited in the material. In such a framework one has to find a balance between SNR and dose/time, so that the experiment is conclusive without producing alteration of the samples during the analysis (*e.g.* [14]).

In this article, we focus on the question of image segmentation when the dataset comes from X-ray fluorescence (XRF) mapping, a technique by which each individual pixel is characterized by its XRF spectrum, providing elemental composition information on that pixel (Figure 1). The classical approach to plot quickly or even *live* elemental distributions recorded by XRF mapping consists of integrating the signal (*i.e.* photons counted by the detector) in spectral regions of interest (ROI) corresponding to targeted element peaks. This does not, however, hold true elemental distribution images as such ROI integrations additionally include significant contributions from other elements or phenomena (namely scattering, and sum and escape peaks); these overlapping biases can only be circumvented by applying slower approaches allowing a spectral decomposition of the dataset (*e.g.* [15, 1]). Here, we propose a *hierarchical segmentation* algorithm combining the characteristics of hierarchical clustering with the imaging properties of a composite material. In other words, we aim at proposing a hierarchical classification procedure of spectral dissimilarities allowing to take into account the spatial proximities between the pixels.

It is important to understand the nature of the signal measured in such experiment. In XRF, we measure the energy of the photons emitted by the material when it is subjected to monochromatic incident radiation. Because this re-emission phenomenon is a stochastic process, the measured spectrum is an empirical sampling of the law of this process. Instead of analyzing the signal using *generic tools* for Euclidean spaces, such as the ℓ_2 distance, it is therefore more relevant to use tools adapted to the comparison of population samples. On this respect, the algorithm we propose is based on the χ^2 as a tool to assess homogeneity between two samples, in the present case two pixels of different compositions.

After defining the terms and notations used throughout this article, we expose the general framework of our dissimilarity measure (using a Ward criteria based on χ^2, subsection 2.2), and then propose an approach to impose spatial constraint upon the agglomerative process of the hierarchical clustering in subsection 2.3. Then, in section 3, we concentrate on the proper steps at which the spatial constraint should be released to properly account for non connex domains made of the same material. We further consider the appropriate number of classes at which the agglomeration process should be stopped in section 4. To illustrate our approach, we apply the proposed algorithm on a true dataset corresponding to the XRF mapping of a fossil teleost fish, including both the analysis of the experimental dataset in section 5, and, in section 6, the analysis of a synthetic dataset resembling the experimental one but providing the possibility to simulate various signal to noise ratio and giving insight into the robustness of the proposed algorithm to the noise level.

2. The proposed hierarchical clustering method

2.1. Notations and definitions

A spectral image of N pixels is considered. Each pixel $i \in \{1, ..., N\}$ is characterized by a spectrum $\mathcal{S}_i = (s_i(p))_{p \in \{1,...,P\}}$, where $s_i(p)$ is the number of photon counts for pixel i in energy canal p.

For $i \in \{1, ..., N\}$ and $p \in \{1, ..., P\}$, let

$$f_{i,p} = \frac{s_i(p)}{s_{\bullet\bullet}} \text{ and } t_p^i = \frac{s_i(p)}{s_{i\bullet}} \text{ where } s_{i\bullet} = \sum_{p=1}^{P} s_i(p) \text{ and } s_{\bullet\bullet} = \sum_{i=1}^{N} s_{i\bullet}.$$

The aim is to propose a hierarchical classification procedure of spectra $(\mathcal{S}_i)_{i \in \{1,...,N\}}$ using the conditional distributions (or profiles) of pixels $((t_p^i)_{p \in \{1,...,P\}}, i \in \{1, ..., N\})$, the pixel i being weighted by $f_{i\bullet} = s_{i\bullet}/s_{\bullet\bullet}$ $(i \in \{1, ..., N\})$. Since these profiles are probability distributions, they are compared using the χ^2 distance.

For two pixels i and j, let

$$d_{\chi^2}^2(\mathcal{S}_i, \mathcal{S}_j) = \sum_{p=1}^{P} \frac{(t_p^i - t_p^j)^2}{f_{\bullet p}}$$

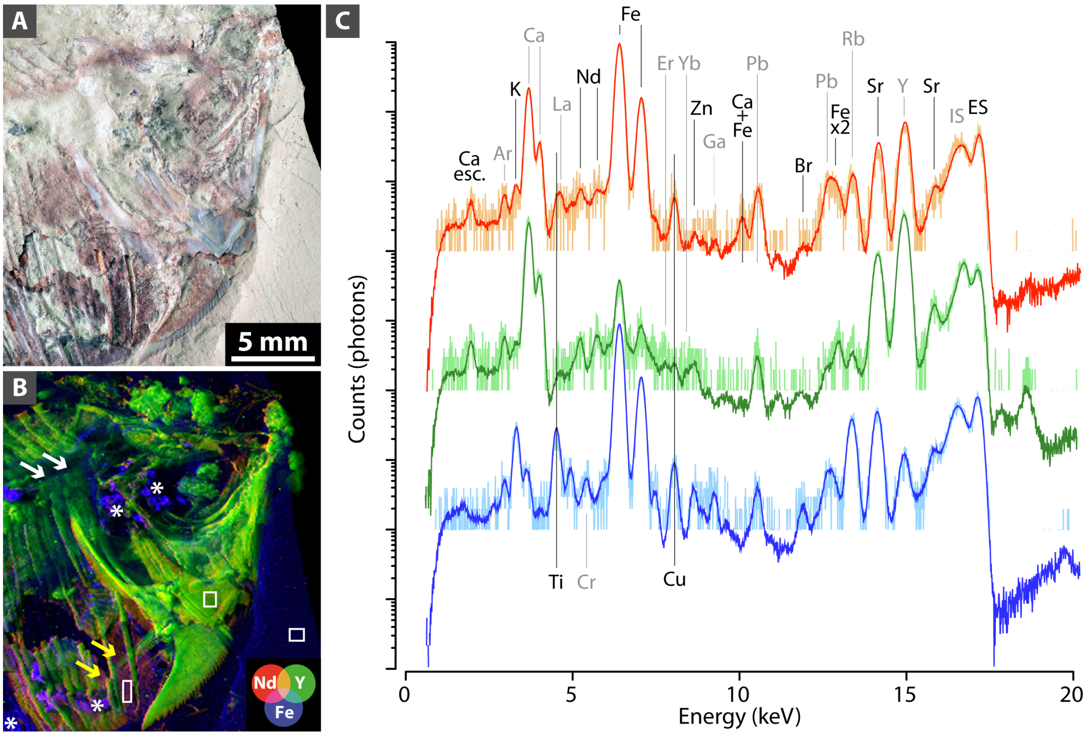

Figure 1. *Synchrotron XRF mapping of major-to-trace elements of the anterior part (skull on the right) of the yet undescribed fish MHNM-KK-OT 03a from the Jbel Oum Tkout Lagerstätte (Upper Cretaceous, 100 Myr, Morocco). (A): optical photograph. (B): false color overlay of the distributions of two rare earth elements, neodymium (red) and yttrium (green), and of iron (blue), reconstructed from a full spectral decomposition of the data (modified from [15]). Acquisition parameters: 100 x 100 μm^2 scan step, 50,851 pixels. Lighter tones indicate higher concentrations. Arrows and asterisks in B are discussed in the text. (C): Mean (dark colored; 90 pixels) and central individual (light colored) spectra from the boxes in B, corresponding to fossilized muscles (red and orange), bone (dark and light green) and the sedimentray matrix (dark and light blue), respectively. Spectra are shown using a logarythmic scale, vertically shifted for clarity. Main peaks are labelled. Abbreviations: esc., escape peak; ES, elastic scattering; IS, inelastic scattering; x2, sum (double) peak. Note that the Ar-peak does not arise from the sample but is due to excitation of Ar in the air (ca. 0.93 %) between the sample and the detector.*

with $f_{\cdot p} = \sum_{i=1}^{N} f_{i,p} = \frac{1}{s_{\cdot\cdot}} \sum_{i=1}^{N} s_i(p)$.

Remarks:

- It is assumed that $f_{\cdot p} \neq 0$ for all p. If there is a canal p such that $f_{\cdot p} = 0$, then $s_i(p) = 0$ for all $i \in \{1, ..., N\}$, hence $t_p^i = t_p^j = 0$ for all $(i, j) \in \{1, ..., N\}^2$. Thus, such canals are removed beforehand.
- It is assumed that $s_{i\cdot} \neq 0$ for all pixel i (otherwise it would mean that the detector did not received any photon for the corresponding pixel).

2.2. The Ward criterion

Using the χ^2 distance as the proximity measure between the spectra of pixels, the hierarchical clustering is designed with the agglomerative Ward criterion δ_{χ^2} [22], which consists of minimizing the increase of the within-cluster inertia at each step. This agglomerative criterion for two clusters C and C' is:

$$\delta_{\chi^2}(C, C') = \frac{\mu_C \mu_{C'}}{\mu_C + \mu_{C'}} d^2_{\chi^2}(S_{g_C}, S_{g_{C'}}) \tag{2.1}$$

where $\mu_C = \sum_{i \in C} f_{i.}$ is the weight of cluster C and S_{g_C} the gravity center of cluster C; $S_{g_C} = (g_c(p))_{p \in \{1,\ldots,P\}}$ with $g_c(p) = \frac{1}{\mu_C} \sum_{i \in C} f_{i.} t^i_p$.

Note that the gravity center of the union of two clusters is $S_{g_{C \cup C'}} = \frac{\mu_C S_{g_C} + \mu_{C'} S_{g_{C'}}}{\mu_C + \mu_{C'}}$.

Usually, the dissimilarity matrix between clusters is updated with a special occurrence of the general Lance and Williams formula, see [11, 17] for example. The dissimilarity between the possible aggregation $C_i \cup C_j$ of two clusters C_i and C_j and any other cluster C_k can be expressed by:

$$\delta_{\chi^2}(C_k, C_i \cup C_j) = \frac{(\mu_{C_k} + \mu_{C_i})\delta_{\chi^2}(C_k, C_i) + (\mu_{C_k} + \mu_{C_j})\delta_{\chi^2}(C_k, C_j) - \mu_{C_k}\delta_{\chi^2}(C_i, C_j)}{\mu_{C_i} + \mu_{C_j} + \mu_{C_k}}.$$

2.3. Taking the spatial constraint into account

With spectral images, the dimension of the dissimilarity matrix at the start is too large ($\mathcal{O}(N^2) \approx 20$ GB), hence it is computationally too expensive ($\mathcal{O}(N^2 P) \approx 5 \times 10^{12}$ operations to design directly a hierarchical clustering with the Ward criterion described above). Moreover, it is desirable that the clusters form unions of spatially connected sub-clusters.

For these two reasons, following [18], we propose a first hierarchical clustering algorithm that only aggregates two spatially neighboring clusters. More precisely, two clusters C and $C^{'}$ are spatially neighboring if there exists $(i, i^{'}) \in C \times C^{'}$ such that i and $i^{'}$ are neighboring pixels.

In our application, we will consider second-order neighborhoods (Figure 2). It implies that most of the pixels have eight neighbors (Figure 2A), while the pixels on an edge or at the corners have only five and three neighbors, respectively.

The advantage of this algorithm is, at each step, that for each cluster only a few

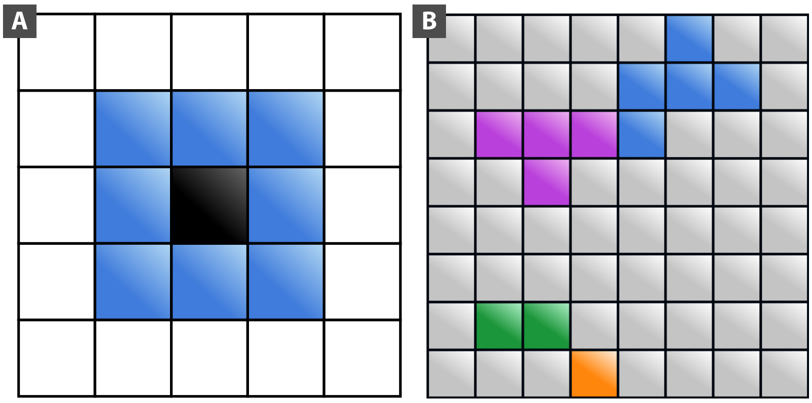

Figure 2. *Schematic representation of the second-order neighborhoods approach. (A): neighbors for a pixel that is not located on an edge or at a corner. (B): example of clusters spatially neighboring; on the top, the blue and purple clusters are spatially neighboring, while on the bottom left the green and orange clusters are spatially neighboring. All are spatially neighboring to the grey cluster. On another hand, for example, the purple and orange clusters are not spatially neighboring.*

dissimilarities have to be computed. Nevertheless, for this reason, it is not possible to use the Lance and Williams formula [11, 17] to update the dissimilarities. Therefore, equation (2.1) is used to compute the dissimilarities needed to design the hierarchy.

The hierarchical algorithm with the spatial constraint operates following the steps below:

Algorithm 2.1 Hierarchical spatial clustering

```
Initialization : computes the $\chi^2$ distances between two spectra for neighboring pixels
Define $J := 1$
**while** $J < N$ **do**
    Aggregates the two neighboring clusters with the smallest Ward criterion value (or $\chi^2$ distances at the first step)
    Updates the neighborhoods of clusters.
    Updates the dissimilarity matrix (for spatially neighboring clusters).
    $J := J + 1$
**end while**
```

If this algorithm is run until it remains only two clusters, we get a hierarchy where at each step the clusters are spatially connected. However, it is not desirable to impose such clusters connexion during the final steps. Indeed, from the point of view of the application domain scientist/specialist, the relevant clusters, while connected at fine scale, have no reason to be spatially connected at large scale.

As a consequence, the proposed algorithm taking the spatial constraint into account is run for J steps leading to spatially connected clusters, J being large (the choice of the switching step J will be discussed hereafter in section 3). It leads to $(N-J)$ spatially connected clusters, hereafter called *patches*. Then from these $(N-J)$ patches, unconstrained agglomerative hierarchical clustering algorithm with the Ward criterion is used. Thus, the proposed final clusters are union of the J spatially connected patches. Obviously, a relevant number of final clusters is to be chosen; this point is discussed in section 4.

3. Selecting the switching step J

3.1. The proposed criterion

In order to select the switching step J in the proposed hierarchical algorithm, a criterion balancing the between-cluster inertia with a regularization term measuring the spatial homogeneity of the clusters is proposed. This criterion to be maximized has the form:

$$H(J) = B(J) + \alpha G(J),$$

where $\alpha \in \mathbb{R}_+$, $B(J)$ is the between-cluster inertia of a partition of the pixels into J patches and $G(J)$ is a measure of the spatial homogeneity of this partition. Following [2], we consider

$$G(J) = \frac{1}{2}\sum_{k=1}^{J}\sum_{i=1}^{N}\sum_{j=1}^{N} c_{ik}c_{jk}v_{ij}$$

where $v_{ij} = 1$ if i and j are neighbors, and 0 otherwise (with $v_{ii} = 0$ by convention), and $c_{ik} = 1$ if $i \in C_k$ and 0 otherwise.

The important point is to choose the scalar α to get an equilibrium between $B(J)$ and $G(J)$. In the extreme situation of a partition into N patches, we have $G(N) = 0$ and in the opposite extreme situation $G(1) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} v_{ij} \approx \frac{1}{2}\sum_{j=1}^{N} 8 = 4N$. (For simplicity, we consider here improperly that each pixel has 8 neighbors.)

Assuming an equilibrium $H(N) = H(1)$ between these two extreme situations leads to $\alpha = \frac{T}{4N}$, T being the total inertia of the whole set of pixels. Thus, the criterion to be maximized is

$$H(J) = B(J) + \frac{T}{4N} G(J).$$

However, as it will be apparent in the case study in section 5, this choice of α leads to the selection of a too large number of patches $J_{\max}$. In order to select a more relevant number of patches, J, from which to release spatial constraints in the clustering, we propose to make use of the "one standard deviation" procedure proposed in [7] to cut a decision tree. This procedure consists of computing $H(K)$ for $K = N, \dots, 1$, then to compute the standard error sd(H) of the resulting $(H(K))_{K=N,\dots,1}$ and choosing the smallest $\hat{J}$ such that

$$H(\hat{J}) \geq H(J_{\max}) - \text{sd}(H).$$

The rationale for this procedure is to determine the value of $\hat{J}$ that corresponds to a large number of clusters and provides a good compromise between the between-cluster inertia and the spatial homogeneity. Note that, while it is expected that $H(J)$ increases from $H(N)$ to $H(J_{\max})$ and then decreasing back to $H(1)$, there is no guarantee for such a behaviour. Such an unexpected behaviour of the H criterion is obtained for low signal to noise ratio image in Figure 6E and G. In order to also address these types of behaviour of H we express the choice of $\hat{J}$ in a different way :

$$\hat{J} = \max\{J | (J < J_{\max}) \text{ and } (H(J) < H(J_{\max}) - \text{sd}(H))\}$$

3.2. In practice

In practice, depending on the image size, it can be too long or impossible to compute $H(K)$ for all $K \in \{N, ..., 1\}$. In this case, the following heuristic approach is proposed to determine $\hat{J}$:

As a first step, the idea is to compute $(H(\ell \times by))_l$, with $by \geq 1$ chosen to have a reasonable computing time and $\ell \in \mathbb{N}^*$ such that $\ell \times by \leq N$. However, after testing on the studied dataset described in section 5, we noticed that the obtained $\hat{J}$ can change significantly according to the value by. Hence, by must be small enough to obtain a correct value for $\hat{J}$ (but we do not know its order of magnitude).

On the other hand we noticed that the obtained values for the ratios $\beta_k = \frac{\hat{J}_k}{N_k}$ ($\hat{J}_k$ is computed with the criterion described in subsection 3.1) are similar for sub-images $(I_k)_k$ of size $(N_k)_k$ having a similar type of morphology.

Therefore we propose to evaluate the “constant” $\beta = \dfrac{\hat{J}}{N}$, which seems linked to the morphology of the studied image, by cutting the image into q sub-images $(I_k)_{k\in\{1,...,q\}}$ with size $(N_k)_{k\in\{1,...,q\}}$). Then, for each sub-image, $\hat{J}_k$ is computed with either the criterion described in subsection 3.1 or the introduction of a small value for by, and the terms $\beta_k = \hat{J}_k/N_k$ are computed for all $k \in \{1, ..., q\}$.

Finally, $\hat{J}$ is chosen as the nearest integer to $\bar{\beta} \times N$, with $\bar{\beta} = \dfrac{1}{q}\sum_{k=1}^{q} \beta_k$.

4. Selecting the number of clusters

4.1. Statistical heuristics

A first and simple way to properly assess the number of clusters from a dendrogram is to select the numbers of clusters producing the greater jumps in the plot of the cluster criterion values (*i.e.* here the Ward criterion), against the number of clusters. We refer to this strategy as the *jump heuristic*.

An another natural and popular criterion for choosing a relevant number of clusters K in a hierarchy designed with the Ward criterion is to use the value of K corresponding to the maximum value of the Calinski and Harabasz criterion (CHC, [8])

$$\mathrm{CHC}(K) = \frac{\mathrm{Tr}(B_K)}{(K-1)} \Big/ \frac{\mathrm{Tr}(W_K)}{(N-K)},$$

where B_K and W_K are respectively the between-cluster matrix and the within-cluster matrix of the partition $C_1, \ldots, C_K$. In the present context, we have

$$\mathrm{Tr}(B_K) = \sum_{k=1}^{K} \mu_{C_k} d^2_{\chi^2}(S_{g_{C_k}}, S_g)$$

and

$$\mathrm{Tr}(W_K) = \sum_{k=1}^{K} \sum_{i\in C_k} f_{i.} d^2_{\chi^2}(S_{g_{C_k}}, \mathcal{S}_i)$$

where S_g is the gravity center of the N pixels, and for $k = 1, \ldots, K$, $S_{g_{C_k}}$ is the gravity center of cluster C_k.

This criterion has been shown to perform well in practical situations (see [19]).

4.2. Particular considerations in the case of spectral images from Ancient materials science

Although we listed above several ways to statistically determine the number of clusters to retain, it is however not recommended to choose a unique number of clusters with a formal technique in the case of ancient material studies. Instead, we here prefer to use the following strategy:

- Preselect several number of clusters using the jump heuristic and the CHC.
- Analyze the preselected clusterings with the help of a specialist of the application domain. Having this purpose in mind, it is desirable to provide the specialist with the mean spectra of the preselected clusters, which represent complementary information to those obtained from usual spectral image processing (*e.g.* ROI integration and full spectral decomposition, see section 1) and are, as such, critical to assess the robustness and benefits of the approach.
- Select with this person the clustering(s) to be interpreted.

The present paper exemplifies in the following section this way of assessing ancient material clusterings.

5. Application to a real world dataset

5.1. Data description

The proposed algorithm has been applied to a spectral image dataset collected on a yet undescribed *ca.* 100-million-year-old new teleost fish from Morocco (Figure 1, [15]). The information embedded in this dataset is a synchrotron micro-X-ray fluorescence (μXRF) major-to-trace-elemental map, where a full XRF spectrum has been recorded for each pixel, over a $22.5 \times 22.5\ mm^2$ area using a scan step of 100×100 μm^2 and a 500 ms counting time ($211 \times 241 = 50,851$ pixels in total; Figure 1B, C). The experiment was performed at the DiffAbs beamline (SOLEIL synchrotron, Gif-sur-Yvette, France) using a 17.2 keV incident beam focused down to a diameter of $10 \times 7\ \mu m^2$.

Very interestingly, the distribution of strontium and yttrium Kα lines, which substitute for calcium in calcium phosphates such as bone apatite [15, 13] and whose information depths under hard X-rays reach 200-300 μm in pure apatite with the used geometry, revealed previously indiscernible anatomical features in this peculiar new fish (Figure 1B, [15]). They particularly unveil the morphology of the first vertebrae (white arrows in Figure 1B), the neurocranium that extends into a sharp supraoccipital at the top of the skull, the metapterygoid, and the hyomandibular that appears dorsally flared. These new information help deciphering the affinities of this new fossil species (in preparation). The other main outcome of this work was that a false color overlay of the distribution of different rare earth elements (REEs; e.g. neodymium and yttrium, red and green distributions in Figure 1B, respectively) discriminates phosphatized muscles (yellow arrows in Figure 1B) and bone [15].

5.2. Resulting hierarchical spatial clustering

In the following, the proposed algorithm has been implemented with R [20] on this image of $N = 211 \times 241 = 50,851$ pixels, for which at each pixel i the spectrum $\mathcal{S}_i$ has $P = 1780$ values. The size of the file containing the dataset is 1.7 GB.

For such a dataset, it is too long to compute the criterion H described in section 3 for all $K \in \{N, ..., 1\}$ in order to determine the switching step $\hat{J}$. Hence, the image has been cut into 4 sub-images with size 100×115 and $\hat{J}$ has been computed as described in subsection 3.2. As explained in subsection 3.2, to limit the computational cost of the complete algorithm, the $(\beta_k)_k$ values are estimated on the four sub-images with H computed every 5 agglomerative steps ($by = 5$). We obtained $(\beta_k)_k$ values of 0.84087, 0.93043, 0.85565 and 0.85522 respectively for top right, bottom right, bottom left and top left sub-images leading to $\bar{\beta}_4 \simeq 0.87054$ and consequently $\hat{J} = 44268$. Note that the $(\beta_k)_k$ values are close for the three sub-images having a similar type of morphology, whilst the bottom-right sub-image consist mostly of sediment and is more homogeneous than other sub-images.

To select the number of clusters, we plotted the jump heuristic and the CHC against the number of clusters (starting with two clusters) (Figure 3). These criteria are here to complement the knowledge of the application domain specialist, in the case of the present example a paleontologist (PG). The jump heuristic leads to propose 6 clusters, whilst the CHC leads to propose either 6 or 9 clusters, corresponding to the two local maxima of the curve.

Looking more precisely at the differences between 6 and 9 clusters, it can be seen that only the light green cluster of the 6-cluster solution is modified (Figure 3C–E). The difference image (Figure 3E) illustrates how the light green cluster incorporates three smaller clusters: one with 5 black isolated pixels, one green cluster with 50 isolated pixels, and a light-red cluster with 3014 pixels spread out it several patches. If the additional clusters with 5 and 50 isolated pixels seem of little interest, the light-

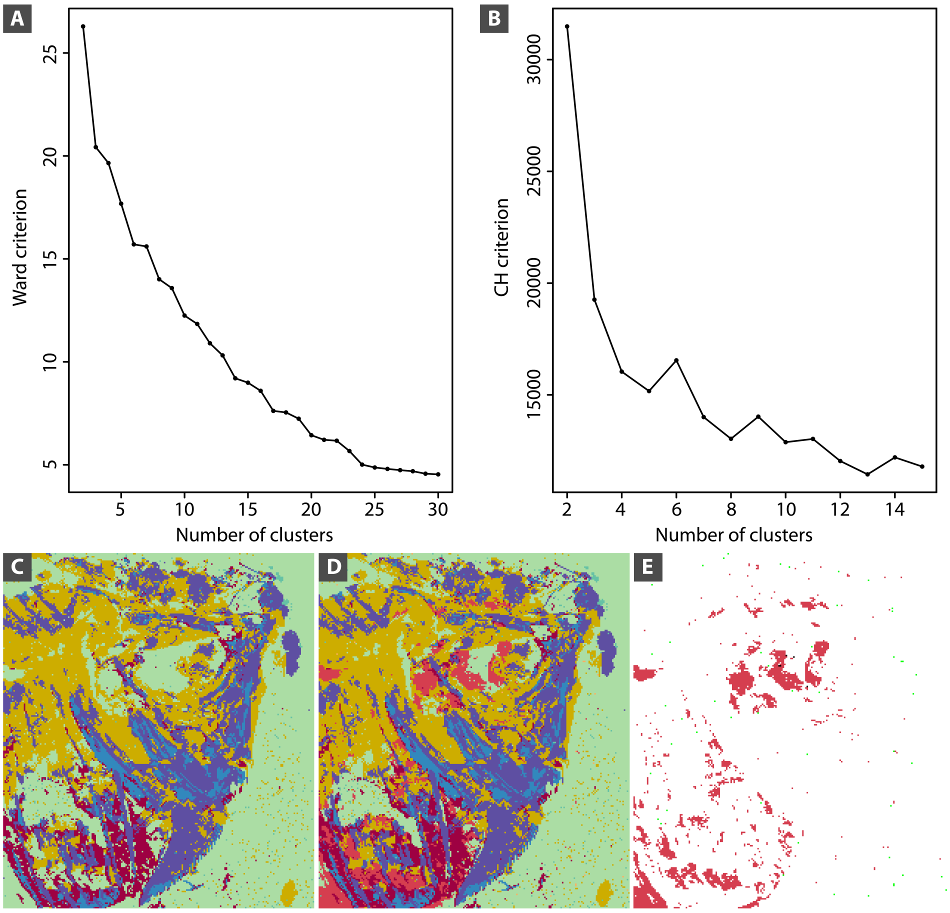

Figure 3. *Defining the number of clusters used for hierarchical segmentation.(A, B): Ward (A) and Calinski and Harabasz (B) criteria against the number of clusters (starting with 2 clusters). (C–E): False color distributions obtained for 6 (C) and 9 (D) clusters, and difference (E).*

red cluster appears to be interesting, as it highlights areas richer in iron (asterisks in Figure 1B) that are not clearly obvious in the μXRF elemental maps obtained using ROI integration or spectral decomposition.

From a paleontological point of view, the segmentation offered by the selected clustering (Figure 4A) does not improve the visualization of hidden anatomical details, but provides new insights into the chemical composition of the different tissues and materials present in the sample through the mean spectra of the clusters (Figure 4B). While individual elemental distributions show no strong contrast in the incorporation of light REEs between bone and muscles (Figure 4C), following the distribution of calcium, which they substitute and that originates from a comparable depth (Figure 4D), the yttrium distribution shows strong enrichment in the bone as compared to the muscles (Figure 4E). In fact, in the muscles area (yellow arrows in Figure 1B), rather than following the type of tissue the yttrium distribution largely follows the thickness of the material as shown by X-ray microtomography where most of the muscles region appear to be very thin or not discernible (Figure 4F). Consequently, thickness and information depth were likely responsible for the apparent REE contrast. Nevertheless, the selected clustering clearly discriminates bone from phosphatized muscles (blue/purple and dark red clusters in Figure 4A, respectively) on the basis on the full μXRF spectra. The muscles dark red cluster appears richer in Fe and Pb (Figure 4B), which come from a reddish fossil biofilm made of iron hydroxides and covering the phosphatized muscles [14, 10, 12] rather than the phosphatic material itself. In turn, the bone blue and purple clusters contain much higher contents in heavier REEs (Lβ1 emission lines from erbium and ytterbium particularly stand out in Figure 4B, as they do not fall in the same energy domain as major elements [15]). This is most likely again an effect of information depths and thickness of the tissues.

On another hand, the selected clustering isolates well the large, highly absorbing iron grains situated posteriorly to the orbit (asterisks in Figure 1B; light red cluster in Figure 4A, B) that prevent segmentation of the first vertebrae and posterior part of the head from the X-ray tomography data (asterisks in Figure 4F). These grains are particularly rich in Fe, Ti, Cu and Ga, but not so much in Pb (Figure 4B) and are therefore, besides their larger size, a different material than the reddish thin film of iron hydroxides covering most of the fossil.

By providing a global discrimination of the different materials composing the fossil much faster than a full spectral decomposition (approximatively two hours here for a computer with specifications i5-4590 @ 3.30GHz, 4 Core, four days using the freeware PyMCA [21]), the proposed clustering methodology provides a robust and quick way to extract, "live" at the beamline, chemical information not hampered by local heterogeneity or contamination for further higher resolution mapping of areas of interest, or point analyzes using, e.g., X-ray absorption spectroscopy.

5.3. Regarding the chosen switching step J

One may wonder if the value of the switching step, J, has an influence on the results for the choice of the number of clusters and for the clusters shape. In this section we tackle this question by applying the algorithm using switching steps equal to $J = 43000$ and $J = 46000$. In Figure 5 are the graphic representations of the jump

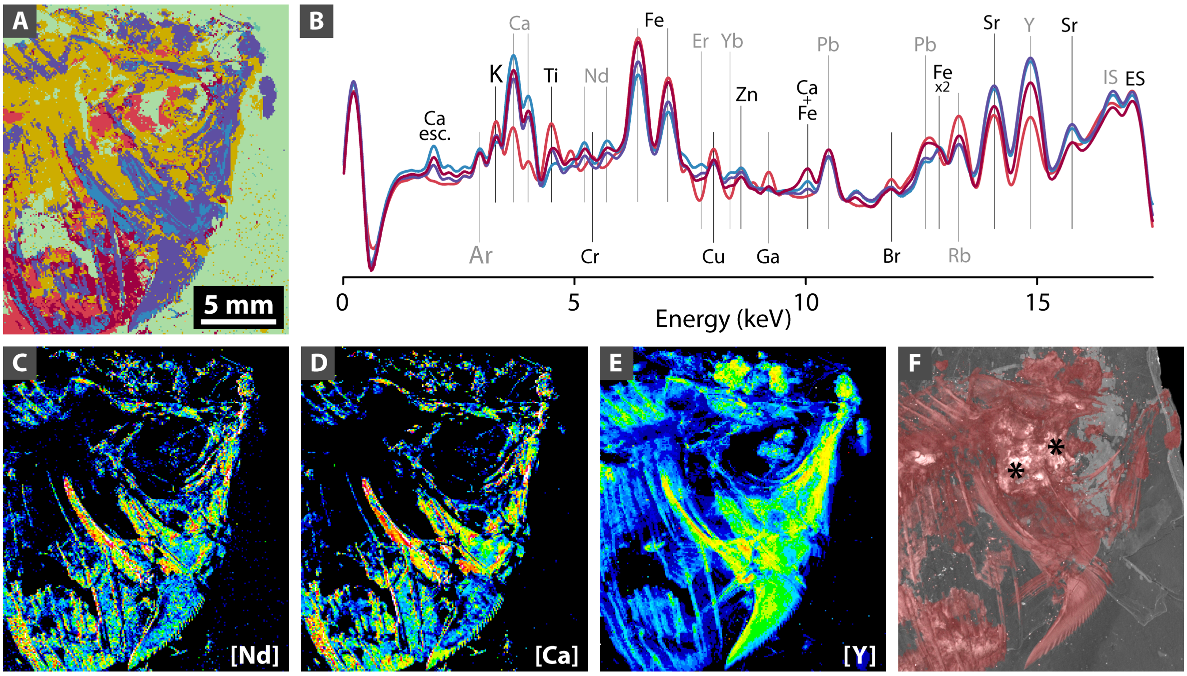

Figure 4. *Hierarchical segmentation of the synchrotron μXRF spectral image dataset of the yet undescribed fish (MHNM-KK-OT 03a) from the Jbel Oum Tkout Lagerstätte (Upper Cretaceous, 100 Myr, Morocco). (A): Segmentation results when 9 classes are selected with the proposed algorithm, disabling spatial constraint at agglomerative step $\hat{J} = 44268$. (B): mean spectra from 4 of the 9 classes visible in (A). (C-E): concentration maps of neodymium (C), calcium (D) and yttrium (E). The color scale goes from dark blue (for low concentration) to red (high concentration) going through green and yellow. (F): micro-computed tomography 3D rendering of the fossil within the sedimentary matrix after rapid segmentation. Voxel size: (24.7 mm)^3.*

heuristic and the CHC for $J = 43000$ and $J = 46000$, respectively.

For $J = 43000$, the jump heuristic plot leads to propose 6 or 10 clusters while the CHC leads to 3, 10 or 12 clusters (Figure 5A,B). For $J = 46000$, the jump heuristic plot leads to propose 5 clusters or maybe 9, and the CHC leads to propose 9 clusters (first local maximum) or more (Figure 5E,F). These results show that the value of the switching step has an influence on the result of the hierarchical clustering. Comparisons of the graphic representations for $J = 43000$, 44268 and 46000 (Figure 5 C,D,G) clearly identify the segmentation resulting from the latter as absolutely unsatisfactory as many fossil areas are found mixed up with the surrounding sediment (Figure 5G). Graphic representations for $J = 43000$ and 44268 appear in turn very similar. Nevertheless, representation for $J = 44268$ (the computed $\hat{J}$ value, see subsection 5.2) more accurately reflects elemental distributions (Figure 1B), particularly regarding the iron-rich phase located around the fish orbit.

6. Assessing robustness of the segmentation to signal to noise ratio

To assess the robustness of the proposed segmentation method in regard of the signal-to-noise ratio (SNR) we prepared simulated data having features close to the one of the experimental dataset used in the previous section. Starting with a single realistic model, we generated a family of simulated observation with a decreasing SNR. Performing the segmentation on this family of simulated data, which are all originating from the same generative model, enabled us to assess the effect of SNR levels on the proposed segmentation results. To achieve consistency we generated this simulated dataset in two steps: (i) we constructed a *zero noise model* that would correspond to a likely observed object; (ii) from this *zero noise model* we generated simulated observation by applying a noise generation process that mimics the physical observation process while providing control on the noise level of the simulated data. We will present the two steps of this procedure, then the results in terms of segmentation.

6.1. Building a *zero noise model* and simulating data with controlled SNR

We based our *zero noise model* on the above studied experimental dataset that we regularized using local polynomial regression smoothing, through the `loess` function in R [20, 9]. To account for the nature and the dynamic of the signal on the observed X-ray fluorescence data, the weight was set to the reciprocal square root of the observation when the observation is not 0, and to 1 otherwise. The second important parameter was the *span* of the filter that we set to 0.02 in order to account for the approximate width of the fluorescence bands on such spectra. Finally, a thresholding was performed on the regularized form so that its value is never lower than 0.001.

While this procedure is producing a realistic *zero noise* X-ray fluorescence spectra in each pixel of our image, it has to be noted that this should not be considered as a *ground truth* version of the observation. Indeed, since each spectrum is dealt with independently from its neighbors, there is no spatial regularization and the *estimations* performed are far from optimal for detector channels that have measured a low level of photons.

The noise present in the observation is mostly due to the counting statistic of each channel of the detector. Hence, we can generate a simulated observed spectra with the same SNR as the raw observation, by simply replacing the value of the *zero*

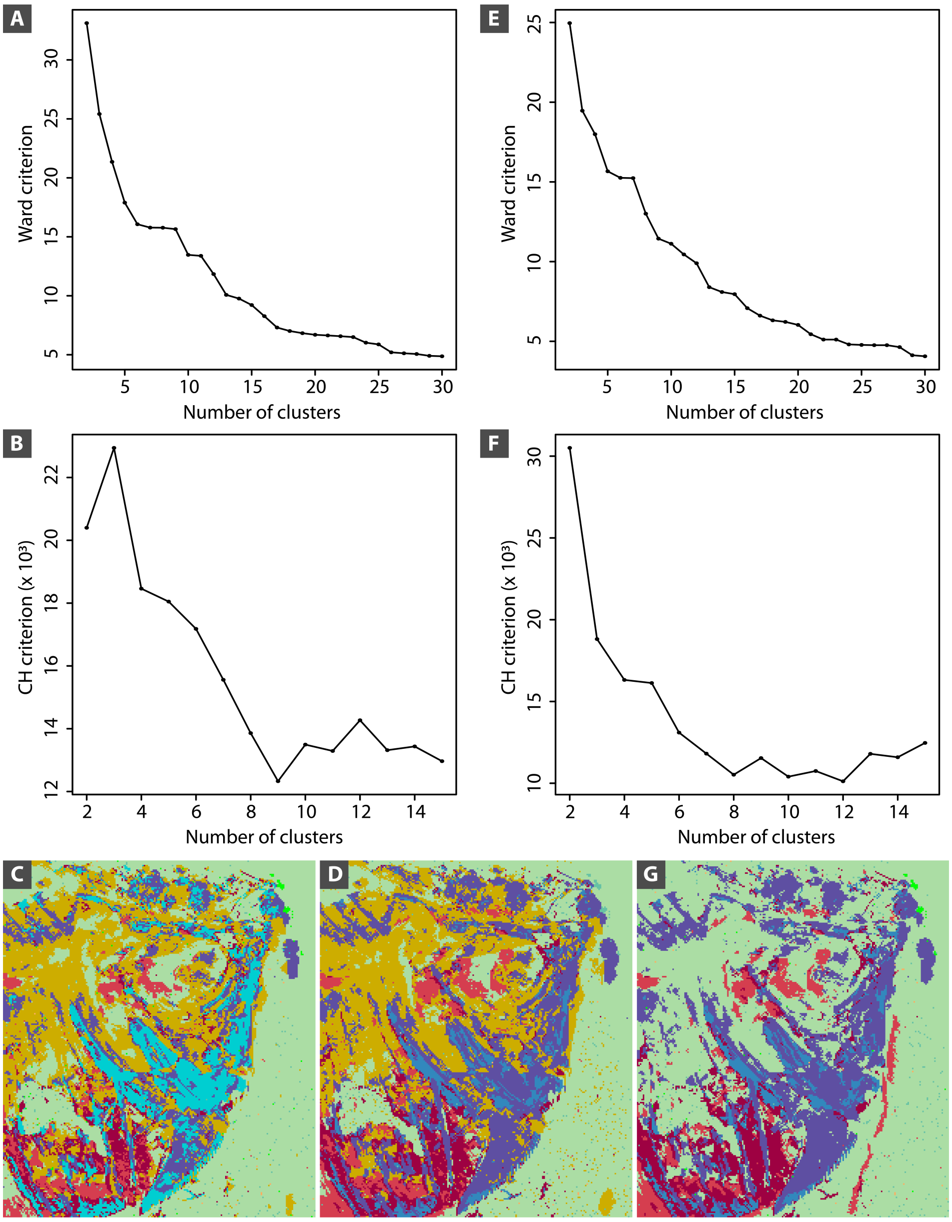

Figure 5. *Hierarchical segmentation for different choices of switching step $J = 43000$ and $J = 46000$. (A, B): Ward (A) and Calinski and Harabasz (B) criteria against the number of clusters (starting with 2 clusters) for $J = 43000$. (C): False color distributions obtained for $J = 43000$ (10 clusters). (D): False color distributions obtained for $J = 44268$ (9 clusters). (E, F): Ward (E) and Calinski and Harabasz (F) criteria against the number of clusters (starting with 2 clusters) for $J = 46000$. (G): False color distributions obtained for $J = 46000$ (9 clusters).*

noise spectra by a single realization of a Poisson random process with its parameter being the *zero noise spectra*'s value. We generated such a dataset, for which we have, by construction, the *ground truth* and a SNR equal to the one of the raw dataset. This simulated dataset is later on referred as a *plus 0db dataset* (p0db in short).

Starting from the same *zero noise model*, we also generated simulated observation with lower SNR. Since each *theoretical value* is replaced by a Poisson realization, dividing the model by a factor of 2 would decrease the SNR by a factor of $\sqrt{2}$, which correspond to removing 3db to the SNR. This simulated dataset is later on referred as a *minus 3db dataset* (m3db in short). Repeating this procedure two more times enabled us to generate a *minus 6db dataset* (m6db) and finally a *minus 9db dataset* (m9db).

Each of these datasets resembles what could have been measured if the exposure time was divided by two incrementally. In other words, obtaining for the m3db dataset a spatial clustering similar to that obtained for the p0db dataset would lead to the conclusion that the experiment could have been done twice faster without significant loss in term of the explained morphology of the fossil. A shorter exposure time also means a lower radiation dose for the sample and correspondingly lower risk of alteration during and due to the measurements.

One has to note that the protocol we use here to smooth the data is not valid as a denoising algorithm since it has some advert effects on the concentration of the trace elements, and in particular the REEs. Still, while the obtained spectra are not properly estimating the *ground truth* of this particular fossil, they have all the features making them likely to be present in a fossil. Hence, the generated dataset should be considered as the XRF spectral image of a purely *phantom fossil*, enabling us to test the proposed hierarchical clustering algorithm on totally controlled data.

6.2. Impact of noise on hierarchical spatial clustering results

Following the same process as in subsection 5.2, we use the criterion H to determine the switching step $\hat{J}$. In Figure 6, we can see that the curve of H has the expected shape for datasets p0db (Figure 6A) and m3db (Figure 6C), while the shape begins to change for dataset m6db (Figure 6E) and is significantly different in m9db (Figure 6G) (the curve correspond to the top left quadrant sub-image but the same behaviour is observed for the three other sub-images). The values obtained for these dataset are: $\hat{J} = 44063$ for p0db, $\hat{J} = 45871$ for m3db, $\hat{J} = 48988$ for m6db. Such choice is not possible for dataset m9db for the exact reason explained at the end of subsection 3.1, hence for each of the four sub-image we took the smallest $\hat{J}_k$ on the right (greater than 8000) such that $H(\hat{J}_k) \geq H(J_{\max}) - sd(H)$ (where here $J_{\max}$ and $sd(H)$ are values computed for the associated sub-image I_k). This led to choose $\hat{J} = 50387$ for m9db. The higher the noise, the higher the $\hat{J}$, getting closer to the total number of pixels N in the image.

According to the plot of the jump heuristic, to the CHC and to clusters appearing to be interesting from a paleontological point of view, the selected number of clusters is 11 for p0db (Figure 6B), 10 for m3db (Figure 6D) and 9 for m6db (Figure 6F). Concerning the m9db dataset, no fossil morphology can be seen when the selected number of clusters is 10 or lower, hence we have decided to represent the 11-cluster segmentation for this dataset (Figure 6H).

As expected, similarity of the graphic representations as compared to the original and simulated datasets quickly degrades with increasing noise, and most morphological information is lost for dataset with a SNR greater than or equal to 6db from the original data. Increasing further the level of noise leads to totally unexploitable data, with which the morphology of the sample could hardly be observed, as demonstrated on the m9db simulation (Figure 6H). Note that, in the m3bd representation (Figure 6D), the pale yellow triangular area that “appears” on the top right of the image and clusters with some of the fossil corresponds to air (there is no sample there, see Figure 1A); it otherwise clusters with the sediment in the other noise models, which can be explained by the geometry used during the experiment where the beam came from the right of the sample with a 45° angle, leading for pixels in that area to record the X-rays-sediment interaction below the fossil surface.

7. Discussion

In this article we propose a spatially constrained hierarchical clustering method to be applied on spectral images, in particular on energy resolved X-ray fluorescence images. The first aspect of the method is to choose an agglomerative criteria based on a dissimilarity measure that is consistent with the noise model of the measured spectra. Then, the main aspect of this method is to apply constraints during the agglomerative process such that only spectra belonging to neighboring pixels could be clustered together. While this constraint is meaningful as long as the classes form small clusters on the image, it is obvious that when the number of classes is small this spatial constraint should not be applied anymore, bringing the problem of the *proper step* at which the spatial constraint should be released. To address this

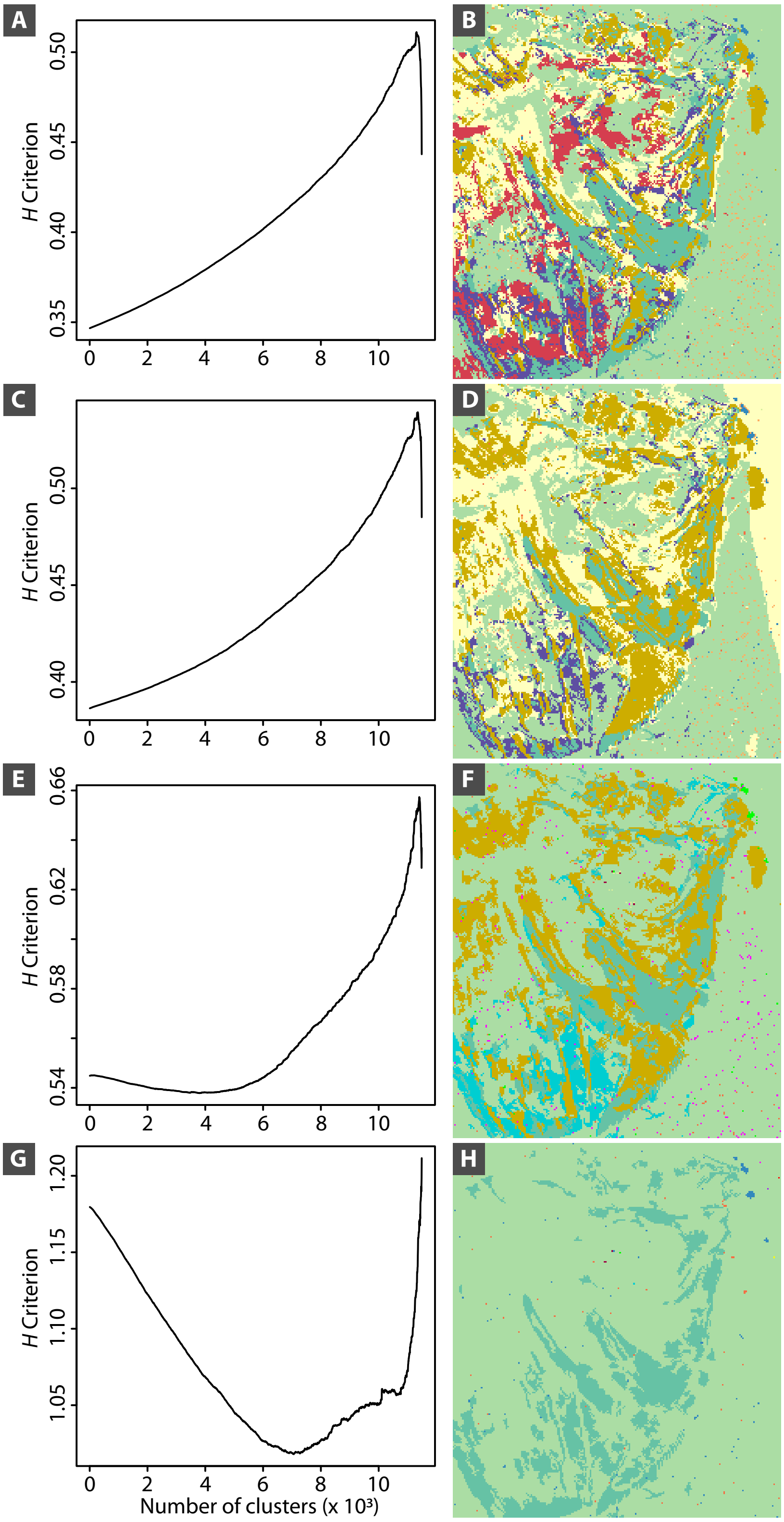

Figure 6. *Behavior of the H criterion for selecting the switching step J while adding noise to the* zero noise model *simulated dataset. The H criterion computed in the top left quadrant sub-image (as explained in subsection 3.2 and subsection 5.2) and resulting* hierarchical spatial clustering *image for the* zero noise model *(adding 0db) (A,B), and after removing 3db (C,D), 6db (E,F) and 9db (G,H).*

problem, we proposed a heuristic that balances the spatial coherence of the proposed segmentation, as measured through the G penalization, and the *between-cluster inertia* deriving from the Ward agglomerative criteria. The outcome of this algorithm is a hierarchy of possible segmentations that the practitioner should choose from. To aid this final selection step, the Ward and Calinski and Harabasz criteria are both computed to determine the most significant segmentation within the full hierarchy.

The advantages of such a *simple minded* algorithm is two-fold: first the general principles of the method do not require deep knowledge of statistical methods and as such can be grasped by the application domain specialist, the paleontologist in the presented example. Second, the computational cost of the segmentation is relatively low, even for a rather large dataset, and the processing time is on par with the typical measurement time for such spectral images. Hence, this method can be applied to the data while the experiment is still ongoing and used for a rapid diagnostic and experimental feedback within the global data acquisition strategy.

As a diagnostic tool, this method helps at finding a balance between a higher signal to noise ratio of individual spectra and the measurement time and radiation dose to which the sample is subjected. In such μXRF imaging modality, the SNR is inversely proportional to the square root of the radiation dose. Increasing the SNR increases the risk of producing radiation-induced damages to the sample during the experiment, but also often leads to increased measurement time and fewer (or smaller) samples being characterized in the allocated time slot. In such a situation it is therefore important to quickly and properly assess the optimal exposure parameters (mostly time, but possibly also beam intensity), which need to be sufficient to produce exploitable spectra while avoiding any risk of radiation-induced damages to the simple and enabling large maps to be collected. Using simulated data, we have shown that the algorithm is robust to an increased level of measurement noise and as such is not only helpful in asserting an optimal measurement time but also in reducing it and lowering the radiation dose.

In our SNR test application, it seems indeed that the behavior of H is a good early indicator of the quality of the observed data, providing insight into the discrimination power of the collected spectra. Indeed the curve in Figure 6G illustrates a behavior significantly different from the ones of Figure 6A,C,E, which we link to the fact that the segmentation obtained in Figure 6H is not very informative. In other words, the behavior of the H criterion as classes get aggregated is a good predictor of the usefulness of the segmentation that will be attained with the data.

Note that while the simulated data tested herein demonstrate that the behavior of the H criterion depends on the SNR, it seems equally likely that this behavior also depends on the type of morphology of the sample being imaged. This is somehow evidenced in our real data test when comparing the H criterion found in the four quadrants of the image, three of which have a very similar morphology and H criterion curve, while the fourth bottom-right quadrant, with mostly sediment and very little fossil features, produces a slightly different H criterion curve.

As a continuation of the present work, one could assess how the H criterion depends on the morphology of the image. From our currently limited experience, it seems likely that if the studied dataset exhibits a similar type of morphology in all the image, a possibility is to choose a sub-image of size N_0 representative of the image morphology. The parameter β can be then evaluated by $\beta_0 = \hat{J}_0/N_0$ (and $\hat{J}$ is chosen as the nearest integer to $\beta_0 \times N$), leading to a drastic reduction of the computational cost of the evaluation of this parameter. Furthermore, this would promote β to be a scalar descriptor of the image's morphology.

Last but not least, this method provides to the practitioner a *complete* view of the information contained in a given spectral image dataset. When such data are collected, the *prior knowledge* on the chemistry of the sample often leads to the selection of very specific features of the spectra to be analyzed. Moreover, although entire μXRF spectra mostly contain XRF elemental information they also include additional, non-elemental signal including escape and sum peaks, as well as inelastic and elastic scattering and peaks from elements present in the air between the sample and the detector such as Ar (Figure 1C). Depending on the sample, some of these peaks can carry interesting signal and one could need to keep them in the analysis. However, it is often preferred to remove them from the analysis and crop the spectra to the “true” elemental signal only, or only a few peaks, prior to the analysis. This can simply be done at the practitioner's discretion prior to applying the algorithm.

Conversely, we here propose to confront the result of such *focused* analysis with an analysis based on the full spectra. Indeed, both the *focused* and *complete* analysis could be performed using the same algorithm but selecting for each one either a subset or the fullset of the spectral channels of the image. Using such an approach the application scientist could both use the data in a *prior knowledge* directed approach, verifying pre-existing hypothesis on the nature of the signal to be detected in the spectral image, as well as a *unsupervised discovery* approach where the full spectral dataset is subjected to the segmentation with *a priori* on which channel is of importance to exploit the image. Finally, this algorithm might even be used as a *post-hoc* analysis to test *a posteriori* the importance of unexpected features of the spectra as *discovered* discriminant features of the sample, as exemplified herein with the iron-rich phase located around the fossil fish orbit for which the cluster mean spectrum provided complementary and new information to decipher its chemistry.

Contribution of authors.

This work arose from discussions between GC, SXC and AG. SXC proposed the exploitation of spatial constraints and the use of χ^2 as an adapted dissimilarity measure for XRF spectra. GC proposed the heuristic rule to stop applying spatial constraint on the segmentation. AG proposed a version of the χ^2 metric consistent between the spatially constrained initial steps and the unconstrained agglomerative steps, so that Lance and William formulae could be used in this latter part. SXC and AG implemented the algorithm and its result representations in R. SXC proposed and implemented the *zero noise model*. PG performed all the experimental measurements and interpretations on the fossil, and oriented the algorithm design to ensure results are valuable for the practitioner. All authors contributed to the writing of this manuscript.

BIBLIOGRAPHY

[1] M. Alfeld and K. Janssens, *Strategies for processing mega-pixel x-ray fluorescence hyperspectral data: a case study on a version of caravaggio's painting supper at emmaus*, Journal of analytical atomic spectrometry, 30 (2015), pp. 777–789.

[2] M. Ambroise and G. Govaert, *Convergence of an EM-type algorithm for spatial clustering*, Pattern recognition letters, 19 (1998), pp. 919–327.

[3] A. Bergamaschi, K. Medjoubi, C. Messaoudi, S. Marco, and A. Somogyi, *Mmx-i: data-processing software for multimodal x-ray imaging and tomography*, Journal of synchrotron radiation, 23 (2016), pp. 783–794.

[4] L. Bertrand, M. Cotte, M. Stampanoni, M. Thoury, F. Marone, and S. Schöder, *Development and trends in synchrotron studies of ancient and historical materials*, Physics Reports, 519 (2012), pp. 51–96, https://doi.org/10.1016/j.physrep.2012.03.003.

[5] L. Bertrand, L. Robinet, M. Thoury, K. Janssens, S. X. Cohen, and S. Schöder, *Cultural heritage and archaeology materials studied by synchrotron spectroscopy and imaging*, Applied physics. A, Materials science & processing, 106 (2012), pp. 377–396, https://doi.org/10.1007/s00339-011-6686-4.

[6] L. Bertrand, M. Thoury, and E. Anheim, *Ancient materials specificities for their synchrotron examination and insights into their epistemological implications*, Journal of Cultural Heritage, 14(4) (2013), pp. 277–289.

[7] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, Taylor & Francis, 1984.

[8] T. Calinski and A. Harabasz, *A dendrite method for cluster analysis*, Communications in Statistics, 3 (1974), pp. 1–27.

[9] W. Cleveland, E. Grosse, and W. M. Shyu, *Statistical Models in S*, Wadsworth & Brooks/Cole, New York, 1992, ch. Chapter 8 : Local Regression Models.

[10] D. Davesne, P. Gueriau, D. Dutheil, and L. Bertrand, *Exceptional preservation of a cretaceous intestine provides a glimpse of the early ecological diversity of spiny-rayed fishes (acanthomorpha, teleostei)*, Scientific Reports, 8 (2018), p. 8509.

[11] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis, 5th edition*, Wiley, 2010.

[12] P. Gueriau, S. Bernard, F. Farges, C. Mocuta, D. B. Dutheil, T. Adatte, B. Bomou, M. Godet, D. Thiaudière, S. Charbonnier, et al., *Oxidative conditions can lead to exceptional preservation through phosphatization*, Geology, (2020).

[13] P. Gueriau, C. Jauvion, and M. Mocuta, *Show me your yttrium, and i will tell you who you are: implications for fossil imaging*, Palaeontology, 61(6) (2018), pp. 981–990.

[14] P. Gueriau, C. Mocuta, and L. Bertrand, *Cerium anomaly at microscale in fossils*, Analytical Chemistry, 87(17) (2015), pp. 8827–88367.

[15] P. Gueriau, D. Mocuta, C.and Dutheil, S. Cohen, D. Thiaudière, the OT1 Consortium, S. Charbonnier, G. Clément, and L. Bertrand, *Trace elemental imaging of rare earth elements discriminates tissues at microscale in flat fossils*, PLoS One, 9(1) (2014), p. e86946.

[16] P. Gueriau, S. Réguer, N. Leclercq, C. Cupello, P. Brito, C. Jauvion, S. Morel, S. Charbonnier, D. Thiaudière, and C. Mocuta, *Visualizing mineralization processes and fossil anatomy using synchronous synchrotron X-ray fluorescence and X-ray diffraction mapping*, Journal of the Royal Society Interface, 17 (2020), p. 20200216, https://doi.org/10.1098/rsif.2020.0216.

[17] G. N. Lance and W. T. Williams, *A general theory of classificatory sorting strategies: II. Clustering systems*, The Computer Journal, 10 (1967), pp. 271–277, https://doi.org/10.1093/comjnl/10.3.271.

[18] L. Lebart, *Programme d'agrégation avec contrainte*, Cahiers de L'analyse des Données, 3 (1978), pp. 275–287.

[19] G. Milligan and M. Cooper, *An examination of procedures for determining the number of clusters in a data set*, Psychometrika, 50 (1985), pp. 159–179.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org/.

[21] V. A. Solé, E. Papillon, M. Cotte, P. Walter, and J. Susini, *A multiplatform code for the analysis of energy-dispersive x-ray fluorescence spectra*, Spectrochimica Acta B, 62 (2007), pp. 63–68.

[22] J. H. Ward, *Hierarchical grouping to optimize an objective function*, Journal of the American Statistical Association, 58 (1963), pp. 236–244.