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

Serge Cohen, Gilles Celeux, Agnès Grimaud, Pierre Gueriau, Sajjad Mahdavi

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

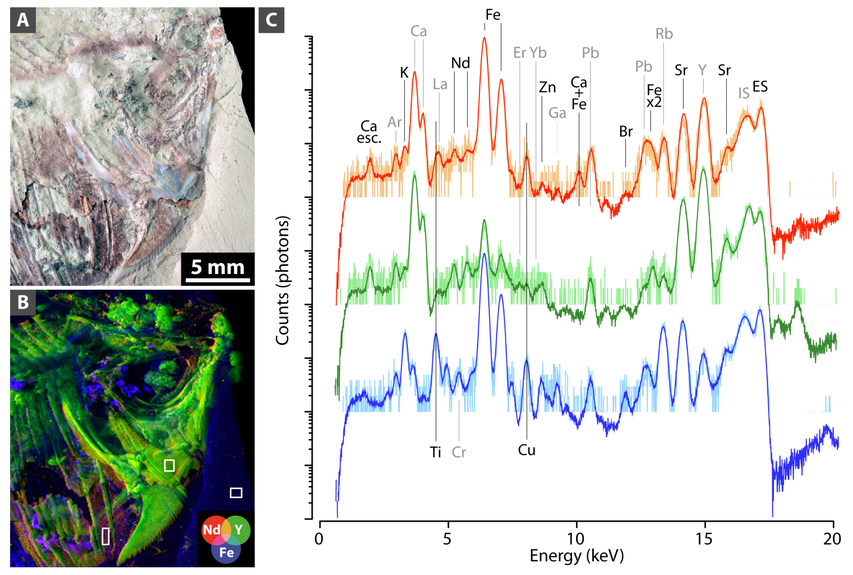
Gilles Celeux[2], Serge X Cohen[1], Agnés Grimaud[3], Pierre Gueriau[1], Sajjad Mahdavi[1]

[1] IPANEMA, UAR 3461 CNRS / Ministère de la Culture / MNHN / Université de Versailles Saint-Quentin, Gif-sur-Yvette, France
[2] IMO, UMR 8628 Inria, CNRS / Université Paris-Saclay, Orsay, France
[3] LMO, UMR 8100, CNRS / Université de Versailles Saint-Quentin, Versailles, France

## 01 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 and compositional dimensions, 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 radation dose density to which the samples are subjected.

## 02 Typical dataset




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. Acquisition parameters: 100 x 100 µm² scan step, 50,851 pixels. (c): Mean (dark colored; 90 pixels) and central individual (light colored) spectra from the boxes in b, corresponding to fossilized muscles (red), bone (green) and the sedimentray matrix (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.

## 03 Statistically oriented processing of XRF spectra scans

We consider each XRF spectra as a random sample of the population of photon that are emitted under X-ray excitation of the elementary volume. We, hence, use tools adapted to the comparison of population samples. we propose is based on the χ2 as a tool to assess homogeneity between two samples.

$$d^2_{\chi^2}(\mathcal{S}_i, \mathcal{S}_j) = \sum_{p=1}^{P} \frac{(t^i_p - t^j_p)^2}{f_{\bullet p}}$$

This is naturally extended to comparing two sets of pixels :

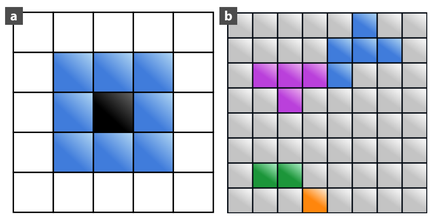
$$\delta_{\chi^2}(C, C') = \frac{\mu_C \mu_{C'}}{\mu_C + \mu_{C'}} d^2_{\chi^2}(S_{g_C}, S_{g_{C'}})$$

A aggregation of two classes leads to :

$$S_{g_{C\cup C'}} = \frac{\mu_C S_{g_C} + \mu_{C'} S_{g_{C'}}}{\mu_C + \mu_{C'}} \qquad \mu_{C\cup C'} = \mu_C + \mu_{C'}$$

## 04 Imposing spatial constraints

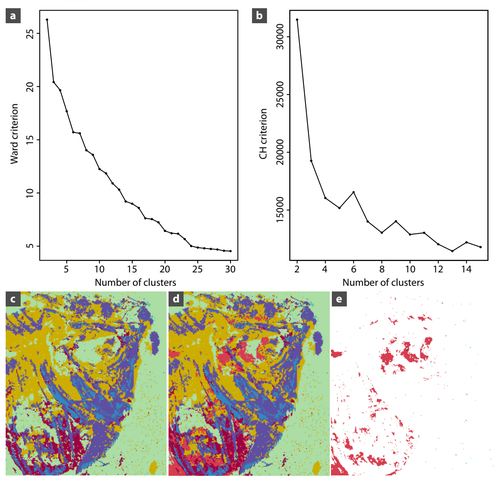
Schematic representation of the second-order neighborhoods approach. (a): neighbors for a pixel that is not located on an edge or at a corner. (b): example of clusters spatially neighboring; on the top, the *blue* and *purple* clusters are spatially neighboring, while on the bottom left the *green* and *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.




## 05 Algorithm

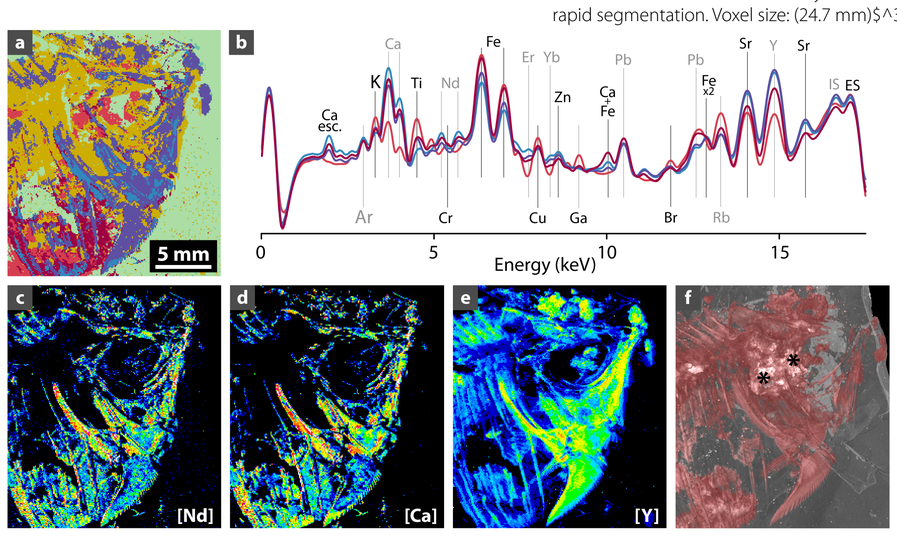
Initialization : computes the $\chi^2$ distances between two spectra for neighboring pixels
Define $L := 1$
**while** $L < N$ **do**
  Aggregates the two neighboring clusters with the smallest Ward criterion value (or $\chi^2$ distances at the first step)
  Updates the neighborhoods of clusters.
  Updates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

## 06 Produced segmentation

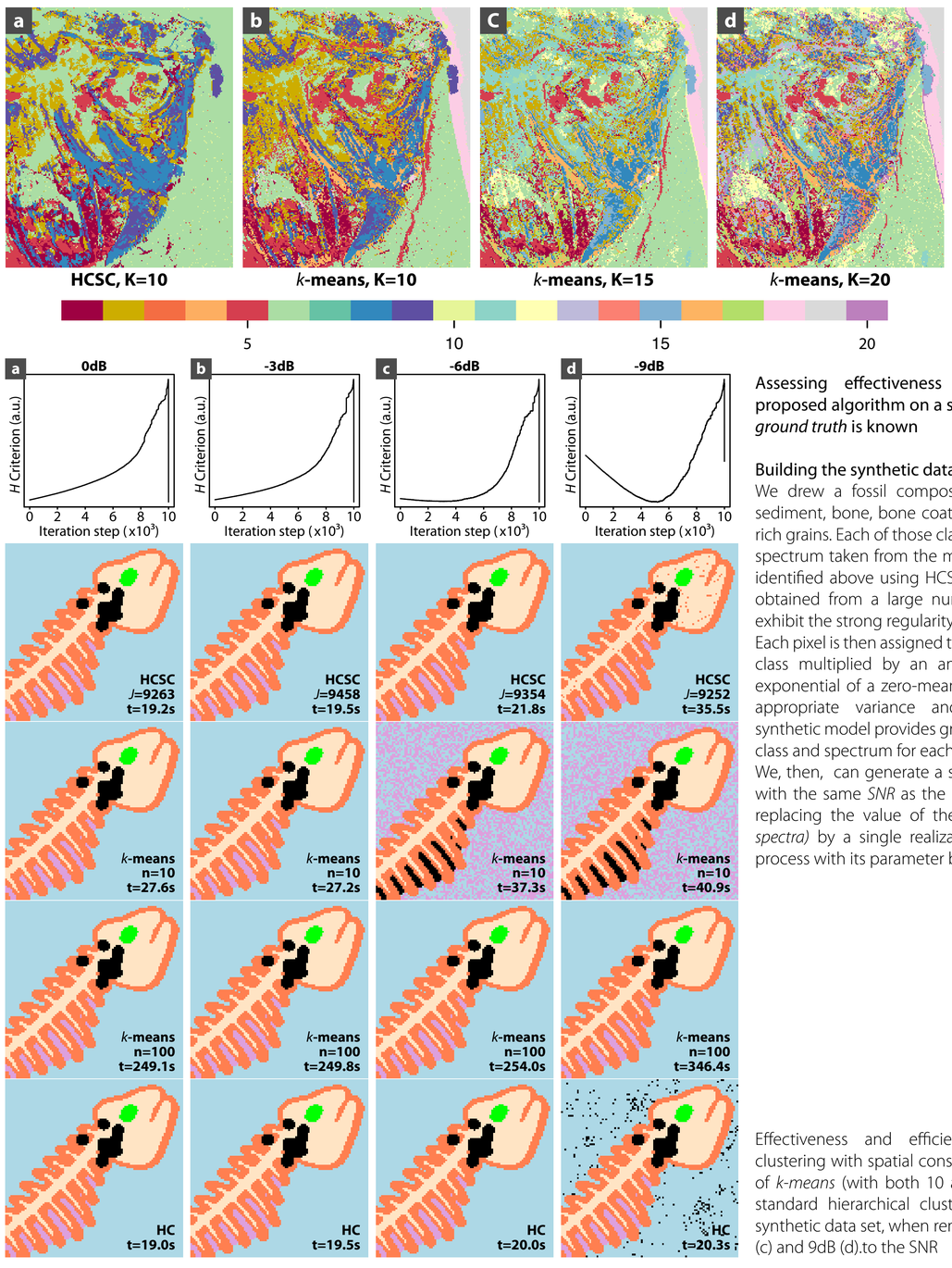



Defining the number of clusters used for hierarchical segmentation
(a, b): Ward (a) and Calinski and Harabasz (b) criteria against the number of clusters (starting with 2 clusters).
(c–e): False color distributions obtained for 5 (c) and 10 (d) clusters, and difference (e)

(a): Segmentation results when 10 clusters are selected with the proposed algorithm, disabling spatial constraint at step J=44175.
(b): mean spectra from 4 of the 10 clusters 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)$^3$




## 07 Comparing to more classical clustering methods

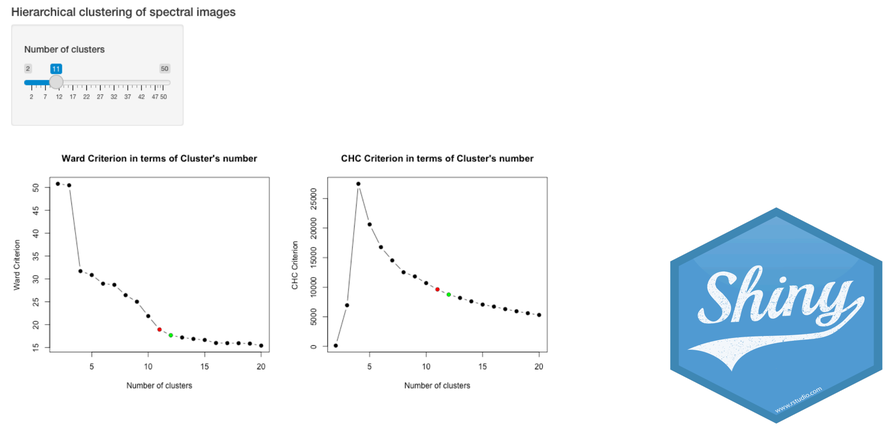



Assessing effectiveness and efficiency of the proposed algorithm on a synthetic dataset for which *ground truth* is known

Building the synthetic dataset:
We drew a fossil composed of six classes, namely sediment, bone, bone coating, muscles, eye and iron-rich grains. Each of those classes is assigned a reference spectrum taken from the mean spectra of the clusters identified above using HCSC (these mean spectra are obtained from a large number of pixels and hence exhibit the strong regularity of noiseless spectra).
Each pixel is then assigned the reference spectrum of its class multiplied by an amplitude factor being the exponential of a zero-mean Gaussian random field of appropriate variance and spatial regularity. This synthetic model provides ground truth both in terms of class and spectrum for each pixel.
We, then, can generate a simulated observed spectra with the same *SNR* as the raw observation, by simply replacing the value of the above model (*zero noise spectra)* by a single realization of a Poisson random process with its parameter being the model's value.

Effectiveness and efficiency of our hierarchical clustering with spatial constraints as compared to that of *k-means* (with both 10 and 100 initializations) and standard hierarchical clustering (HC) on the purely synthetic data set, when removing 0dB (a), 3dB (b), 6dB (c) and 9dB (d).to the SNR

## 08 A graphical exploratory interface




*E-mail: serge.cohen@ipanema-remote.fr