



Exploring local spatial features in hyperspectral images

Journal:	<i>Journal of Chemometrics</i>
Manuscript ID	CEM-20-0053.R1
Wiley - Manuscript type:	Short Communication
Date Submitted by the Author:	n/a
Complete List of Authors:	Ahmad, Mohamad; University of Modena and Reggio Emilia, Chemical and Geological Sciences; Univ. Lille, CNRS, LASIRE Vitale, Raffaele; KU Leuven, Department of Chemistry; Univ. Lille, CNRS, LASIRE Silva, Carolina; Federal University of Pernambuco, Department of Chemical Engineering Ruckebusch, Cyril; Univ. Lille, CNRS, LASIRE Cocchi, Marina; University of Modena and Reggio Emilia, Chemical and Geological Sciences
Keyword:	hyperspectral images, spatial features, wavelet transform, grey-level co-occurrence matrix, multivariate image analysis

SCHOLARONE™
Manuscripts

Exploring local spatial features in hyperspectral images

Mohamad Ahmad^{1,2}, Raffaele Vitale^{2,3}, Carolina S. Silva⁴, Cyril Ruckebusch², [Marina Cocchi](#)¹

¹*Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi 103, 41125 Modena, Italia*

²*Univ. Lille, CNRS, LASIRE, F-59000 Lille, France*

³*KU-Leuven, Department of Chemistry, Molecular Imaging and Photonics Unit, Celestijnenlaan 200F, B-3001 Leuven, Belgium*

⁴*Department of Chemical Engineering, Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235, Cidade Universitaria, Recife, Brazil*

Abstract

We propose a methodological framework to extract spatial features in hyperspectral imaging data and establish a link between these features and the spectral regions, capturing the observed structural patterns. The proposed approach consists of five main steps: i) two dimensional Stationary Wavelet Transform (2D-SWT) is applied to a hyperspectral data cube, decomposing each single-channel image with a selected wavelet filter up to the maximum decomposition level; ii) a grey-level co-occurrence matrix is calculated for every 2D-SWT image resulting from stage i); iii) distinctive spatial features are determined by computing morphological descriptors from each grey-level co-occurrence matrix; iv) the morphological descriptors are rearranged in a two dimensional data array; v) this data matrix is subjected to Principal Component Analysis (PCA) for exploring the variability of the aforementioned descriptors across spectral channels. As a result, groups of spectral wavelengths associated to specific spatial features can be pointed out yielding a better understanding and interpretation of the data. In principle, this information can also be further exploited, e.g. to improve the separation of pure spectral profiles in a multivariate curve resolution context.

Keywords: hyperspectral images; spatial features; wavelet transform; grey-level co-occurrence matrix; multivariate image analysis.

1. Introduction

Hyperspectral imaging (HSI) has numerous possible applications that, depending on the instrumentation and the spectral domain covered, can range from environmental surveillance to cellular monitoring [1-4]. HSI data consist of three-dimensional arrays with two spatial dimensions and one spectral dimension, providing an image for each scanned spectral channel. When HSI is concerned, one is usually interested in retrieving both the pure spectra of the individual components constituting the image and their respective spatial distribution.

In chemistry, one of the most used approaches for achieving this aim is Multivariate Curve Resolution – Alternating Least Squares (MCR–ALS). With MCR–ALS, a hyperspectral image is first unfolded pixel-wise, and afterwards a bilinear model is fitted to the unfolded data using an ALS approach under appropriate constraints. This permits to unravel the distribution maps and the pure spectral signatures of the physico-chemically meaningful constituents of the image [5]. However, unfolding the data results in losing the information on the local spatial features within the dataset. A solution to this issue can be the implementation of spatial constraints as described in [6]. Nonetheless, if different constituents exhibit distinct spatial features (i.e., textural patterns), and/or two (or more) of these different constituents show a significant overlap along both the spectral and the spatial domain, this approach may not be actionable. Multivariate Image Analysis (MIA) is a very useful alternative to investigate spatial features in greyscale, RGB and, to a lesser extent, hyperspectral images [7-9]. The basic principle of MIA is **to analyze the unfolded data** by means of multivariate tools like Principal Component Analysis (PCA) or Partial Least Squares regression (PLS). Spatial features are captured considering the relationships between each pixel and its neighbors in the unfolding step (see references [7, 8] for details). MIA has also been coupled to Wavelet Transform (WT) multiresolution analysis [10] and this combination has been proven effective in resolving spatial features in multispectral [11] and Raman hyperspectral images [12]. In addition, other

1
2
3 strategies, like co-clustering [13] and grey-level co-occurrence matrices (GLCM) [14], have
4 recently been applied to examine texture in HSI datasets. Textural features in HSI have also
5 been explored by using three-dimensional discrete WT [15] or by fusion of the two-
6 dimensional discrete WT decomposition images obtained from each spectral channel [16]. All
7 these approaches are capable of enhancing spatial features in HSI data and of establishing some
8 link between these features and the spectral regions responsible for the observed textural
9 patterns. However, their performance has not been satisfactory in situations where both
10 textural/spatial and spectral information are highly mixed, i.e. when pure chemical components
11 and/or distinct physical contributions are overlapped both in the investigated spectral range and
12 in their distribution across the image [17].

13
14 Aiming at facing this issue, we propose in this short communication a methodological
15 framework that relies on the capability of two-dimensional Stationary Wavelet Transform (2D-
16 SWT [18-19]) to capture distinct spatial features in disjoint subspaces (different wavelet
17 images can be extracted for every spectral channel) and on the versatility of multivariate data
18 analysis tools to explore the information these spatial features carry. The approach consists of
19 five consecutive steps: i) 2D-SWT decomposition of the HSI data cube resulting in wavelet
20 sub-images; ii) from these sub-images, computation of the GLCM; iii) calculation of
21 morphological descriptors from each GLCM; iv) rearrangement of the obtained descriptors into
22 a matrix; and v) PCA modelling of this matrix for the exploration of the variability of the
23 morphological descriptors across spectral channels.

24
25 Such a workflow would allow, for example, spectral wavelengths mostly capturing specific
26 spatial features to be pointed out by the simultaneous investigation of PCA loadings and scores.
27 PCA can also be coupled to other statistical approaches for addressing particular tasks
28 depending on the study at hand. Here, for example, we used the k-means algorithm to cluster
29 the loadings obtained from the PCA analysis of the descriptors matrix and determine the
30

1
2
3 spectral regions that show similar variation patterns in terms of spatial descriptors. These
4 regions can be selective for different pure components underlying HSI data and highlighting
5 them can help users to, e.g., improve the quality of MCR-ALS solutions.
6
7
8
9

10 11 **2. Methods**

12
13 The proposed methodology consists of five main steps which are illustrated in figure 1 and
14 detailed below.
15
16
17

18 19 Step 1: 2D-SWT decomposition (figure 1.2)

20
21 A low-pass and a high-pass filter are applied to every spectral channel of the analyzed
22 hyperspectral image (see figure 1.1) to obtain four distinct sets of wavelet coefficients, denoted
23 H, V, D and A. Each set corresponds to a decomposition block and will be referred to as a
24 wavelet sub-image. The horizontal set (H) corresponds to the application of both a low-pass
25 filter, row-wise, and high-pass filter, column-wise; the vertical set (V) to the application of a
26 high-pass filter, row-wise, and a low-pass filter, column-wise; the diagonal set (D) to the
27 application of a high-pass filter, **applied** both row-wise and column-wise; and the
28 approximation set (A) to the application of a low-pass filter, **applied** both row-wise and
29 column-wise. This scheme is iterated on the approximation sub-image until a certain
30 decomposition level has been reached (see figure 1.2). In 2D-SWT, at each iteration of the
31 decomposition, the wavelet filter is up-sampled, contrary to standard Discrete Wavelet
32 Transform (DWT) [20] where the wavelet coefficients are down-sampled. In this way,
33 congruent wavelet sub-images are yielded and the position of the spatial patterns in the image
34 is preserved.
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54 For the objective of this short communication, the decomposition level is set to the maximum
55 compatible with the size of the original image. Furthermore, the Haar filter was utilized here
56 even though different decomposition filters exist and can be exploited for the same purpose
57
58
59
60

[21]. ~~To the best of our knowledge, this is the first attempt of extending wavelet transform multiresolution analysis to HSI datasets.~~

< figure 1 about here >

Step 2: GLCM calculation (figure 1.3)

A GLCM is computed for each slab corresponding to a specific wavelet sub-image i.e., for a given block of coefficients (H, V, D, A) associated to a specific decomposition level and ~~to a~~ ~~specific~~ spectral wavelength. A GLCM maps the local textures of a given image by counting how often pairs of pixels with ~~certain~~ normalized integer intensity values occur at a particular distance [22, 23]. ~~The type of~~ normalization, and the direction along which the distance is calculated, need to be set *a priori*. Here, we used a 64-integer intensity range and different directions for each wavelet sub-image, matching the spatial pattern every wavelet decomposition image highlights: vertical direction for the horizontal coefficients image (the information retained after the decomposition, in fact, reflects the local spatial changes in that direction); horizontal for the vertical coefficients image; top-left to bottom-right direction for the diagonal coefficients image and the summation of the previous directions for the approximation coefficients image.

Step 3: Descriptors calculation (figure 1.4)

A set of eight descriptors (Energy, Contrast, Correlation, Variance, Inverse difference moment, Sum entropy, Information Measure of Correlation 1, and Maximal correlation coefficient [23]) is computed for each GLCM (see figure 1.4). ~~A brief description of each of these features and their respective formulas are included as supplementary material (Table S1).~~ These descriptors were chosen because they summarize most of the local spatial features one can find in an image. However, depending on the case-study, distinct descriptors can be selected based on prior knowledge or on the necessity of specific image features to be highlighted.

Step 4: Descriptors matrix rearrangement (figure 1.5)

All the morphological descriptors values estimated for every spectral channel are organized into individual column vectors subsequently gathered in a single data matrix. Thus, the **descriptors** matrix (see figure 1.5) features a number of rows equal to the number of descriptors (8) times the number of decomposition levels (which depends on the size of the hyperspectral image) times the number of wavelet sub-images per decomposition level (4). **The** number of columns **corresponds** to the number of **sampled** spectral **variables** (wavelengths). ~~From now on we will refer to this matrix as the descriptors matrix.~~

Step 5: Multivariate analysis

The descriptors matrix is subjected to PCA for the exploration of the information it carries and, more specifically, for establishing a link between the spatial and spectral information captured by the variation of the morphological descriptors within the investigated spectral range. In this work, a possible pathway to establishing this link in a more systematic way is also explored, i.e. the application of k-means clustering to the resulting PCA loadings to get an idea about the spectral channels associated to similar spatial features.

3. Results and Discussion

The aim of this communication is to show how local spatial features extracted with the use of the procedure outlined in the previous section can provide valuable information for HSI data analysis and exploration. For this purpose, the results of two case-studies are presented.

Oil-in-water emulsion

The first case-study [24, 25] relates to a Raman HSI dataset of an oil-in-water emulsion, which illustrates a situation where the spatial and spectral information are both somehow selective in their respective domains, i.e. no severe overlap of the spatial and spectral features is observed. More specifically, the different individual chemical components of the image (featured in the

1
2
3 spectral domain) are associated to clearly distinguishable shapes/spatial structures. This dataset
4 is relatively simple and will serve as a proof of concept for the proposed methodology. The
5 Raman imaging system **by which these data were collected** has a spatial resolution of around 1
6 μm and the image is 60×60 pixels. The spectral resolution is 3.6 cm^{-1} and the investigated
7 **spectral range goes** from 953.6 to 1861 cm^{-1} (**253 wavelenghts**). In figures 2a and 2b, the mean
8 image, averaged over all the spectral channels, and the mean spectrum, averaged over all the
9 pixels, are shown, respectively.

10
11
12
13
14
15
16
17
18
19
20 < figure 2 about here >

21
22
23 We applied our approach considering eight descriptors (Section 2) and up to five
24 decomposition levels. The outcomes resulting from the PCA modelling of the descriptors
25 matrix are shown in figures 2c and 2d. They display the scores and loadings plots of the two
26 first principal components, respectively. The scores plot provides a graphical representation of
27 the variation of the morphological descriptors across the wavelet sub-images whereas the
28 loadings plot accounts for their variation across the spectral channels. As it can be assumed
29 that the most extreme score values are the most informative, **as recently pointed out in MCR**
30 **context [26]** we will focus on the ten most extreme scores values along PC_1 and PC_2 in figure
31 2c. The corresponding points are labelled according to their respective descriptor name and to
32 the wavelet sub-image from which such a descriptor has been calculated. In figure 2d, the
33 loadings lying in the same quadrant along the direction determined by each one of these points
34 and the origin of the PCA subspace (± 9 degrees) are highlighted accordingly (same
35 colors/symbols). This way, ten different groups of spectral channels were identified (loadings
36 too close to the origin and, thus, contributing very little to the definition of PC_1 and PC_2 were
37 excluded from this assessment). One wavelet sub-image can be associated to each group which
38 corresponds to one of the labelled descriptors in the scores plot, as represented in figure 3 (a
39 maximum number of three wavelenghts per group is considered here).

< figure 3 about here >

The comparative inspection of figures 2 and 3 enables the simultaneous exploration of the spatial and spectral characteristics of the investigated HSI data. The loadings plot gives insights into the spectral domain while the scores plot together with the wavelet sub-images does so for the spatial domain.

From figure 3 overall, four main spatial contributions are discernible:

- two small droplets (see figures 3.1 and 3.4): the descriptors capturing these spatial features were Energy-A2 (i.e. energy calculated on the wavelet sub-image corresponding to the approximations coefficients at decomposition level two) and Variance-H3;
- a larger droplet-like structure (see figures 3.3 and 3.8), associated to descriptors Variance-A3 and Sum-Entropy-A2;
- a circular border around the larger droplet-like structure (see figures 3.2 and 3.5), associated to descriptors Energy-V2 and Variance-V3. It was expected that the vertical details would be able to capture the border shape;
- a background effect (see figures 3.7 and 3.10), associated to descriptors Variance-A5 and Information Mean Correlation-V5. Actually, the deepest decomposition level typically captures very smooth textural features. Overall, we may regard the fifth level of decomposition as the one that captures background and/or illumination effects.

On the other hand, figures 3.6 and 3.9 seem to result from the overlap of some of these contributions.

To summarize, the proposed approach enabled to identify and distinguish four different components within the oil-in-water scene, spatially unraveled in the wavelet sub-images (figure 3) and spectrally identified in the loadings plot (figure 2d). These results are in good agreement

1
2
3 with previous findings [24, 25], which discussed the interior droplet as due to oil and the border
4 structure being the oil/water interface, while matching the smaller droplet to oil with a different
5 composition. However, to complement the results of this preliminary visual inspection ~~the~~
6 ~~descriptors matrix was subjected to~~ k-means clustering ~~was exploited~~. For this purpose, the
7 descriptors matrix was compressed by PCA (17 PCs, explaining 90.12 % variance) and
8 clustering was applied on the estimated PCA loadings (4 clusters of wavelengths were
9 retrieved).

10
11
12
13
14
15
16
17
18
19
20 < figure 4 about here >

21
22
23 As highlighted in figure 4c, averaging the hyperspectral image across the four extracted clusters
24 of spectral channels led to the isolation of the different structures previously unraveled. It is
25 then clear that for the data at hand contributions showing a distinctive spatial distribution are
26 also associated to rather selective spectral signatures within different wavelength ranges. These
27 spectral signatures can be inspected in figure 4a:

- 28 • Cluster # 1, which is associated to the small droplets, coincides with the minor peaks
29 at 1300, 1684 and 1789 cm^{-1} ;
- 30 • Cluster # 2, which is associated to the border structure, corresponds to three major
31 peaks at 1044, 1126 and 1317 cm^{-1} ;
- 32 • Cluster # 3 mainly encompasses the peaks at 1483 and 1508 cm^{-1} . It corresponds to the
33 interior droplet;
- 34 • Cluster # 4 corresponds to the baseline regions observed in the mean spectrum.

35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52 A point to take note of is the overall correspondence of the spectral channels belonging to the
53 groups identified by exploratory PCA of the descriptors matrix with the clusters found through
54 k-means, as can be seen by comparing figures 2d and 4b. In addition, the spatial features
55
56
57
58
59
60

1
2
3 revealed by the wavelet sub-images in figure 3 mostly match those highlighted in the clustered
4
5 mean images in figure 4c.
6
7

8 Due to the straightforward nature of the results, the dataset has been used to assess the
9
10 relevance of each individual step of the proposed workflow (Figure 1). Analysis have been
11
12 performed by removing some of these steps, i.e. applying k-means on the PCA of unfolded
13
14 HSI data or excluding the Wavelet decomposition step and carrying out the GLCM and
15
16 descriptors calculations directly on the raw images. The obtained results (Supplementary
17
18 material, Figure S1) showed that the information obtained was less significant than when
19
20 applying the full original workflow.
21
22
23
24

25 Semen droplet on cotton tissue

26
27
28 The second case-study regards a 222×220 pixels HSI-near infrared (NIR) image (acquired in
29
30 the wavelength range 1268.8-2456.2 nm with a spectral resolution of 6.3 nm) of a semen
31
32 droplet on a piece of white cotton. Further details on the data acquisition are given in [15]. The
33
34 mean image is represented in figure 5a. A quite complex structure is observed which is
35
36 characterized, at least at a first glance, by: i) a distinct horizontal pattern across the entire image
37
38 that is due to the rough surface of the cotton fabric (texture); ii) an almost indiscernible shadow
39
40 of the oval-shaped border of the semen droplet; and iii) a spurious fiber filament in the lower
41
42 middle area of the image.
43
44
45

46
47 It is worth noting that this case-study exhibits a much higher complexity than the previous one:
48
49 the cotton contribution is present everywhere across the image, thus, there exists no spatial area
50
51 selective for semen. In addition, the spectral profiles of the different constituents of the
52
53 captured scene are severely overlapped and semen might be not homogeneously distributed
54
55 over the cotton sample.
56
57

58
59 In order to explore these data, we applied our approach as detailed in Section 2.
60

1
2
3 < figure 5 about here>
4
5

6 The three first PCs of the descriptors matrix were inspected here. Figure 5 displays the resulting
7 outcomes. Different clusters of wavelength channels were identified within both the PC₁/PC₂
8 and PC₂/PC₃ subspaces. The **ten** most extreme score values in figures 5c and 5e were taken into
9 consideration, **following the same procedure outlined for the emulsion data set**. For the sake of
10 simplicity not all the corresponding points in the loadings plot (figure 5d) were considered to
11 extract the related wavelet sub-images, but all were investigated. Among them, only those
12 associated to the wavelet sub-images showing easily interpretable spatial patterns were
13 isolated. These wavelet sub-images are shown in figure 6.
14
15
16
17
18
19
20
21
22
23
24

25 < figure 6 about here>
26
27

28 The images shown in figure 6 can be categorized into three groups, **each showing one of the**
29 main spatial contributions **underlying this dataset**:
30
31
32

- 33 • the semen stain associated to descriptors Energy-A1 (figure 6.1) and Variance-A4
34 (figures 6.4 and 6.5). Two sub-groups seem to be present, as two spatially distinct forms
35 of the stain seem to exist. In figure 6.5, for the images taken at 1700 and 1500 nm, the
36 two forms are apparent. This is in good agreement with previous findings showing how
37 the spatial distribution observed ~~for such a cluster~~ might be generated by
38 complementary semen compounds exhibiting a distinct behavior during the drying
39 process of the biological fluid on the cotton fabric [17];
40
41
42
43
44
45
46
47
48
49 • the background (figure 6.2) captured by the descriptor Contrast-H1. This textural
50 pattern that is observed across the entire image is most likely caused by the reflection
51 of light on the cotton fibers;
52
53
54

55 The corresponding loadings (purple stars in figure 5d) are associated to wavelengths
56 1287.5, 1306.2 and 1318.8 nm. These are **located** in the range where cotton absorbs
57
58
59
60

(1268.8 nm - 1362.5 nm) [27], and would most likely be related to the 2nd overtones and combination of C-H stretching and C-H deformation;

- the fiber filament (figures 6.1 and 6.3) captured by descriptor Contrast-V3. The associated spectral wavelengths are in the range between 2000 nm and 2243.8 nm, where O-H bending and C-O stretching contributions are expected from cotton.

Notice that the fiber filament and the semen stain appear to be slightly overlapped in various wavelet sub-images. Another important point to consider is that the best separation of the two semen stain forms resulted from the A4 wavelet sub-image (figure 6.5). This highlights the fact that wavelet decompositions can provide a much greater spatial resolution, as they have the ability to **unravel** distinct spatial structures. This is observed in figure 6.2 and 6.4, where the separation between semen and the horizontal spatial interference is evident (such a separation cannot be visualized in any of the single channel images of the original HSI data, not shown). Considering the isolation of the horizontal details (figure 6.2), the different forms of the semen stain (figure 6.5), the fiber (figures 6.1 and 6.3) and the “mask” that excludes the semen stain (figure 6.5, see at 1875 nm), the wavelet sub-images feature a high potential for further investigation of the data. **For example, considering methods such as MCR-ALS, on one hand these isolated images can furnish the number of components to use, on the other hand can** serve as initial estimates and could, e.g., increase the spatial unmixing capability of the current methodology [28]. However, this could come with some ambiguity, as with higher decomposition levels, the wavelet sub-images will only retain low frequency signals. This has to be considered carefully and will be explored in a future publication.

In order to corroborate the conclusion drawn after this visual inspection, k-means was applied, as previously explained **(in this case the descriptor matrix was compressed to an 18 PCs model, explaining 90.61% variance).**

< figure 7 about here >

Figure 7 represents the obtained results. A more ambiguous clustering was obtained here compared to the previous case-study, most likely due to the increased complexity and spatial overlap of the different components underlying the HSI dataset. Yet, the previously determined components are discernable in figure 7d. However, they are considerably more mixed. For example, 'Cluster # 1' encompasses both the fiber and (to a lesser extent) the semen stain. Also, in 'Cluster # 3' and 'Cluster # 4' the semen stain, the fiber and the background are not completely **separated** from one another. Nonetheless, despite being mixed with the textural background, the two different spatial forms of semen were **isolated** in Clusters # 2 and Cluster # 3.

The spectral channels corresponding to Cluster # 2 and Cluster # 3 (figure 7a), include the wavelengths regions around 1700 nm, and in the range 1500 - 1575 nm, which are selective for the two forms of semen when wavelet sub-image A4 is considered (figure 6.4 and 6.5). It must be emphasized that compared to the emulsion dataset, the results of the clustering are not satisfactory in terms of isolation of the different components (figure 7d). However, it has been shown (in figure 6) that the wavelet sub-images estimated at specific spectral wavelengths and recovered by the PCA analysis of the descriptor matrix do have the ability to isolate the different components.

Concluding remarks

In this short communication a methodological framework based on combining both spatial features and spectral information for the analysis of HSI data is proposed. The method decomposes every single-wavelength image of a three-dimensional HSI array by 2D-SWT and computes individual GLCM for every resulting wavelet sub-image. Morphological descriptors are afterwards estimated from all the GLCMs. In this way, spatial and spectral information is

1
2
3 enhanced and conveyed in a **single** features matrix, which is finally processed by multivariate
4 data analysis **tools** like PCA. Depending on the specific tasks the user must address, different
5
6 multivariate statistical tools can be exploited at this point.
7
8
9

10 **Although the proposed workflow combines different computational steps, which translates into**
11 **higher complexity, every one of them is a necessary link in the chain. In fact, when some of**
12 **these steps were skipped, the information obtained was insufficient to unravel all the distinct**
13 **spatial features underlying the dataset under study and relate them to specific spectral regions.**
14
15
16
17
18
19

20
21 According to the results obtained in two different case-studies, it can be concluded that the
22 proposed strategy is capable of consistently recovering the main spatial features of a HSI
23 dataset and of highlighting the distinctive spectral regions accounting for them. In particular,
24 the outcomes related to the investigation of the semen droplet dataset were found to be
25 particularly promising considering the extreme physico-chemical complexity of the examined
26 image. **In fact, even though the semen and cotton components show highly overlapped spectra,**
27 **and cotton fabric is present everywhere in the sample, it was possible to highlight and localize**
28 **the two different forms of the semen stain as well as the spectral wavelengths at which this**
29 **spatial separation is effective.**
30
31
32
33
34
35
36
37
38
39
40
41

42 Nonetheless, further developments can be foreseen. GLCM is just one of the possible ways to
43 compress the spatial information encoded in wavelet sub-images, and others will be explored
44 in future research. The possibility of **utilizing** other multivariate data analysis tools in the
45 developed framework will also be assessed. Finally, the improved and at least preliminarily
46 disentangled spatial/spectral information returned by the described approach might constitute
47 a valuable starting point for the design of new constraints to be applied in the context of MCR-
48
49
50
51
52
53
54
55
56
57
58
59
60
ALS. Additional work is currently in progress towards this direction.

Acknowledgements

Dr. C.S. Silva acknowledges financial support from: NUQAAPE-FACEPE (APQ-0346-1.06/14), Núcleo de Estudos em Química Forense (NEQUIFOR; CAPES AUXPE 3509/2014, Call PROFORENSE 2014), FACEPE (BFP-0800-1.06/17 and APQ-0576-1.06/17)

References

- [1] Guolan L, Baowei F. Medical hyperspectral imaging: a review. *J. of Biomedical Optics* 2004; **19** (1): 010901.
- [2] Gowenl AA, O'Donnell CP, Cullen PJ, Downey G, Frias JM. Hyperspectral imaging – an emerging process analytical tool for food quality and safety control. *Trends in Food Sc. & Tech.* 2007; **18** (12): 590-598.
- [3] Liang, H. Advances in multispectral and hyperspectral imaging for archaeology and art conservation. *Appl. Phys. A* 2012; **106**: 309–323.
- [4] Goetz AFH, Curtiss B. Hyperspectral imaging of the earth: Remote analytical chemistry in an uncontrolled environment. *Field analytical chemistry and technology* 1996; **1** (2): 67–76.
- [5] De Juan A, Tauler R, Dyson R, Marcolli C, Rault M, Maeder M. Spectroscopic Imaging and Chemometrics: A Powerful Combination for Global and Local Sample Analysis. *Tr. AC* 2004; **23** (1): 70–79.
- [6] Hugelier S, Devos O, Ruckebusch C. On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis. *J. Chemom* 2015; **29**: 557–561.
- [7] Bharati MH, Liu JJ, MacGregor JF, Image texture analysis: methods and comparisons. *Chemom. Intel. Lab. Syst.* 2004; **72**: 57–71.
- [8] Prats-Montalbán JM, de Juan A, Ferrer A, Multivariate image analysis: A review with applications. *Chemom. Intel. Lab. Syst.* 2011; **107**: 1–23.
- [9] Jamme F, Duponchel L. Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis. *J. Chemom* 2017; **31**: e2882.
- [10] Liu J, MacGregor J. On the extraction of spectral and spatial information from images. *Chemom Intel Lab Syst* 2007; **85**: 119-130.
- [11] Li Vigni M, Prats-Montalbán JM, Ferrer A, Cocchi M. Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA). *J. Chemom.* 2018; **32**: e2970.
- [12] Gosselin R, Rodrigue D, Gonzalez-Nunez R, Duchesne C. Potential of Hyperspectral Imaging for Quality Control of Polymer Blend Films. *Ind. Eng. Chem. Res.* 2009; **48**: 3033–3042.

- 1
2
3 [13] Jacques K, Ruckebusch C. Model-based co-clustering for hyperspectral images. *J.*
4 *Spectral Imaging* 2016; **5**: a3
5
6 [14] Xu JL, Gowen A. Spatial-spectral analysis method using texture features combined with
7 PCA for information extraction in hyperspectral images. *J. Chemometrics* 2019; **34**: e3132.
8
9 [15] Guo, Xian, Xin Huang, and Liangpei Zhang. Three-dimensional wavelet texture feature
10 extraction and classification for multi/hyperspectral imagery. *IEEE Geoscience and remote*
11 *sensing letters* 2014; **11.12**: 2183-2187
12
13 [16] Beauchemin M. Spatial pattern discovery for hyperspectral images based on
14 multiresolution analysis. *International Journal of Image and Data Fusion* 2012; **3:1**: 93-110
15
16 [17] Silva CS, Pimentel MF, Amigo JM, Honorato RS, Pasquini C. Detecting semen stains on
17 fabrics using near infrared hyperspectral images and multivariate models. *Trends in*
18 *Analytical Chemistry* 2017; **95**: 23e35.
19
20 [18] Nason GP, Silverman BW. The stationary wavelet transform and some statistical
21 applications. *Wavelets and Statistics* 1995, A. Antoniadis, Ed. Springer-Verlag, New York,
22 Lecture Notes in Statistics.
23
24 [19] Juneau P, Garnier A, Duchesne C. The undecimated wavelet transform—multivariate
25 image analysis (UWT-MIA) for simultaneous extraction of spectral and spatial information.
26 *Chemom Intel Lab Syst.* 2015; **142**: 304-318.
27
28 [20] Mallat S. A theory for multi-resolution signal decomposition: the wavelet
29 representation. *IEEE Trans Pattern Anal Mach Intell.* 1989; **11**: 674-693.
30
31 [21] Prats-Montalbán JM, Ferrer A, Cocchi M. N-way modeling for wavelet filter
32 determination in multivariate image analysis. *J Chemom.* 2015; **29**: 379-388.
33
34 [22] Haralick RM. Statistical and structural approaches to texture. *Proceedings of the IEEE*
35 1979; **67** (5): 780–803.
36
37 [23] Haralick RM, Shanmugan K, Dinstein I. Textural Features for Image Classification.
38 *IEEE Transactions on Systems, Man, and Cybernetics* 1973; **SMC-3**: 610-621.
39
40 [24] Andrew JJ, Browne MA, Clark IE, Hancewicz TM, Millichope AJ. Raman Imaging of
41 Emulsion Systems. *Applied Spectroscopy* 1998; **52** (6).
42
43 [25] De Juan A, Maeder M, Hancewicz T, Tauler R. Use of local rank-based spatial
44 information for resolution of spectroscopic images. *J. Chemometrics* 2008; **22**: 291–298
45
46 [26] Ghaffari M, Omidikia N., Ruckebusch C. Essential Spectral Pixels for Multivariate
47 Curve Resolution of Chemical Images. *Anal. Chem.* 2019; **91**: 10943–10948.
48
49 [27] Burns DA, Ciurezak EW. *Handbook of Near-Infrared Analysis III edition* 2008;
50 Chapter 25: Table 25.2.
51
52 [28] Vitale R., Hugelier S., Cevoli D., Ruckebusch C. A spatial constraint to model and
53 extract texture components in Multivariate Curve Resolution of near-infrared hyperspectral
54 images. *Anal. Chim. Acta* 2020; **1095**: 30-37.
55
56
57
58
59
60

Legend of figures

Figure 1. The methodological framework, the spectral dimension is colored after every step for the sake of a better and easier visualization. (1.1) a generic hyperspectral data cube; (1.2) the data structure obtained after the 2D-SWT decomposition; (1.3) GLCM obtained from each slab of the wavelet coefficient three-dimensional arrays depicted in 1.2; (1.4) the descriptors calculated from the GLCM; (1.5) rearrangement of the descriptors matrix.

Figure 2. Oil in water data set. (a) Mean hyperspectral image (mean taken across spectral dimension). (b) Mean spectrum (mean taken across pixels after unfolding). (c) PC1 vs. PC2 scores plot (PCA of descriptors matrix). The most extreme values in the scores space are highlighted by different colored symbols and labelled. (d) PC1 vs PC2 loadings plot. The points in the loadings plot that correspond (i.e. are on the same direction with respect to origin) to the points highlighted in the scores plot are shown with the same symbols.

Figure 3. Oil in water data set. The wavelet sub-images, at decomposition block and level as reported in the points label in Figure 2c and at the spectral wavelengths corresponding to the ones showing the same symbols in Figure 2d, are shown in separate frames, numbered from 1 to 10. In each frame is reported on top the name of the descriptor-block-level, e.g. in frame 1 (named figure 3.1 in the text) are shown the Approximation images at the second decomposition level for the three wavenumbers 1695.2, 1760 and 1763.6 cm^{-1} , and so on for 3.2 to 3.10. Above each image is reported the corresponding wavenumber.

Figure 4. Oil in water data set. Results of the k-means clustering algorithm on the loadings matrix. (a) mean spectrum, with the point at each wavelength colored according to the cluster's number they were assigned to; (b) PC1 vs. PC2 loadings plot, points colored according to clusters; (c) mean images for each cluster, mean taken across the spectral channels belonging to the same cluster.

Figure 5. Semen data set. (a) mean image; (b) mean spectrum; (c) PC1 vs PC2 scores plot resulting from the application of PCA on the descriptors matrix (the ten most extreme values are labelled with the descriptor's name and the wavelet sub-image on which it has been calculated). (d) PC1 vs PC2 loadings plot. The points highlighted in the scores plot are shown with the same symbols; (e) PC2 vs PC3 score plot; (f) PC2 vs PC3 loadings plot.

Figure 6. Semen data set. The selected wavelet sub-images from PCA analysis of descriptors matrix, shown in separate frames labelled by their respective number, e.g. in frame 1 (named figure 6.1 in the text) are shown the Approximation images at the first decomposition level for the three wavelength 2256.2, 2262.5 and 1612.5 nm, and so on for 6.2 to 6.5.

1
2
3 **Figure 7.** Semen data set. Results of the k-means clustering algorithm on the loadings matrix.
4 (a) mean spectrum, with clustering highlighted on the spectral channels; (b) PC1 vs. PC2
5 loadings plot of, points colored according to clusters; (c) PC2 vs PC3 loadings plot; (d) mean
6 image (across the spectral dimension) for each cluster.
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

Supplementary material:

Comparative analysis of the different aspects within the workflow

Some minor investigations have been made to assess the need of the several steps of the procedure. In the first case (figure S1a), both the wavelet decomposition and GLCM steps are excluded. In the second case (figure S1b) only the wavelet decomposition step is excluded. The clustering of the spectral channels obtained by the analysis of the emulsion dataset (figure 4 in the paper) was taken as a reference, as it showed a clear result that is consistent with literature [refs. 24, 25 in the paper].

These results show mixed image maps and/or spectral regions. For example, on figure S1a in cluster #2 all components seem to be present and looking at the mean spectrum with the clusters highlighted, the background spectral channels are mixed with the main peaks. On the other hand, figure S1b shows that the big droplet and ring are not separated whatsoever.

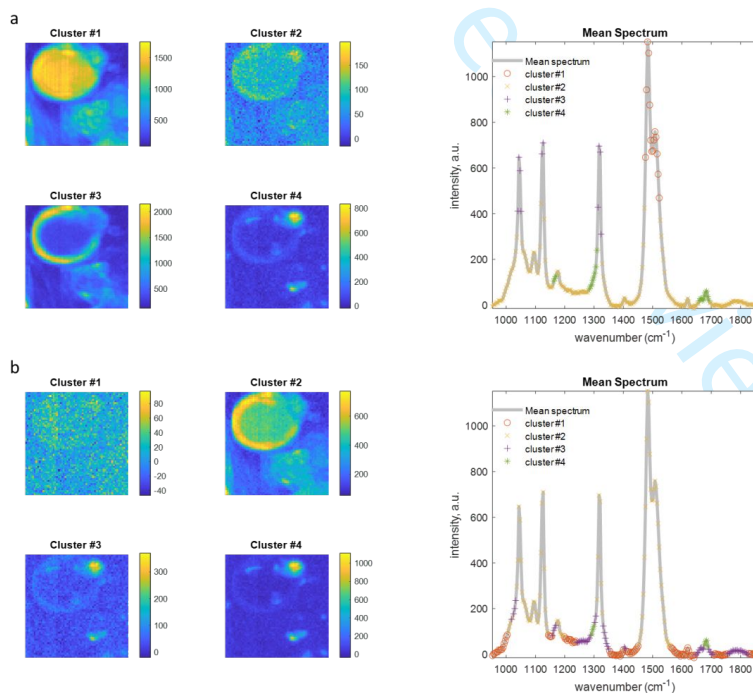


Figure S1. Results obtained by clustering the PCA loadings for a) the original HSI; b) the descriptors matrix of the GLCM obtained excluding the wavelet decomposition step.

Table S1: Definition of the Haralick features used in the paper. The GLCM is normalized and its elements, denoted as $p(i,j)$, correspond to the probability estimate of a given pair of gray intensity levels (i,j) to be observed in the image.

<p>Energy equals the squared sum of all the elements of the GLCM, giving a measure of uniformity of the original image. It reaches a maximum when the image is constant.</p>	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)^2$ <p>Where N_g is the number of distinct gray levels in the quantized image, and $p(i,j)$ is the (i,j)th entry in the GLCM.</p>
<p>Contrast represents a measure of the intensity contrast between a pixel and its neighbor over the whole image. Its value is zero for a constant image and increases along with the intensity difference between neighboring pixels.</p>	$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{\substack{j=j \\ n= i-j }}^{N_g} p(i,j) \right\}$
<p>Correlation is a measure of how correlated a pixel is to its neighbor over the whole image. Correlation is a measure of grey tone linear dependency in the image. It attains values of 1 or -1 for a perfectly positively or negatively correlated image, respectively. It is undefined for a constant image.</p>	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$ <p>μ_x, μ_y and σ_x, σ_y are the means and standard deviations of p_x and p_y, respectively. Where:</p> $p_x = \sum_{j=1}^{N_g} p(i,j), p_y = \sum_{i=1}^{N_g} p(i,j)$ $\mu_x = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i \cdot p(i,j), \mu_y = \sum_{j=1}^{N_g} \sum_{i=1}^{N_g} j \cdot p(i,j)$ $\sigma_x^2 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu_x)^2 \cdot p(i,j)$ $\sigma_y^2 = \sum_{j=1}^{N_g} \sum_{i=1}^{N_g} (j - \mu_y)^2 \cdot p(i,j)$

<p>Variance corresponds to the variance of the probability distribution p.</p>	$\sigma^2 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu)^2 p(i,j)$ <p>Where for a symmetric GLCM: $\mu_x = \mu_y = \mu$</p>
<p>The inverse difference moment measures how similar the GLCM is to a diagonal matrix. It is also known as homogeneity and equals 1 for a diagonal GLCM.</p>	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1 + (i - j)^2} p(i,j)$
<p>Sum Entropy uses the same formula as Entropy* but considers, instead of single elements of the GLCM, the sum of elements in its diagonals.</p> $\text{Entropy}^* (HXY) = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \log(p(i,j))$ <p>It represents a measure of disorder related to the grey-level distribution of the image and it is large when the image is not texturally uniform and many GLCM elements have very small values.</p>	$- \sum_{k=2}^{2N_g} p_{x+y}(k) \log(p_{x+y}(k))$ <p>Where:</p> $p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \quad k = i + j$
<p>The Information Measure of Correlation 1 equals the ratio of the difference between overall entropy and joint entropy (across rows and columns) to the max entropy across rows/columns. It can be interpreted as a measure of texture complexity.</p>	$\frac{HXY - HXY1}{\max(HX, HY)}$ <p>HXY is Entropy. $HX(HY)$ are rows (columns) Entropy, respectively, e.g.:</p> $HX = - \sum_{j=1}^{N_g} p_x(i) \log p_x(i)$ <p>and:</p> $HXY1 = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \{ \log p_x(i) p_y(j) \}$

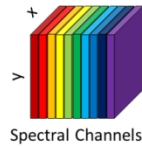
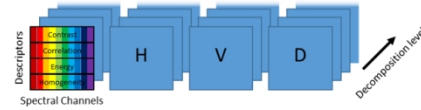
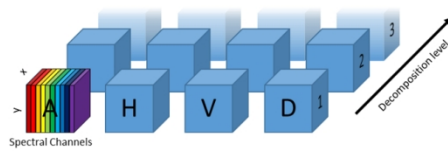
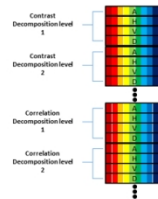
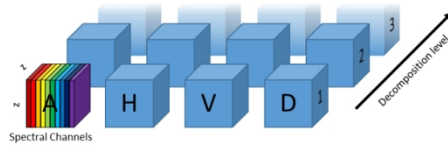
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

The **maximal correlation coefficient** is defined as the square root of the second largest eigenvalue of matrix \mathbf{Q} .

\mathbf{Q} is defined as:

$$Q(i,j) = \sum_m \frac{p(i,m)p(j,m)}{p_x(i)p_y(m)}$$

For Peer Review

1.1 Hyperspectral data**1.4 Descriptors matrix****1.2 Wavelet decomposition****1.5 Descriptors matrix (rearranged)****1.3 Grey-level co-occurrence distribution**

**Multivariate analysis
solutions**

Figure 1. The methodological framework, the spectral dimension is colored after every step for the sake of a better and easier visualization. (1.1) a generic hyperspectral data cube; (1.2) the data structure obtained after the 2D-SWT decomposition; (1.3) GLCM obtained from each slab of the wavelet coefficient three-dimensional arrays depicted in 1.2; (1.4) the descriptors calculated from the GLCM; (1.5) rearrangement of the descriptors matrix.

255x190mm (149 x 149 DPI)

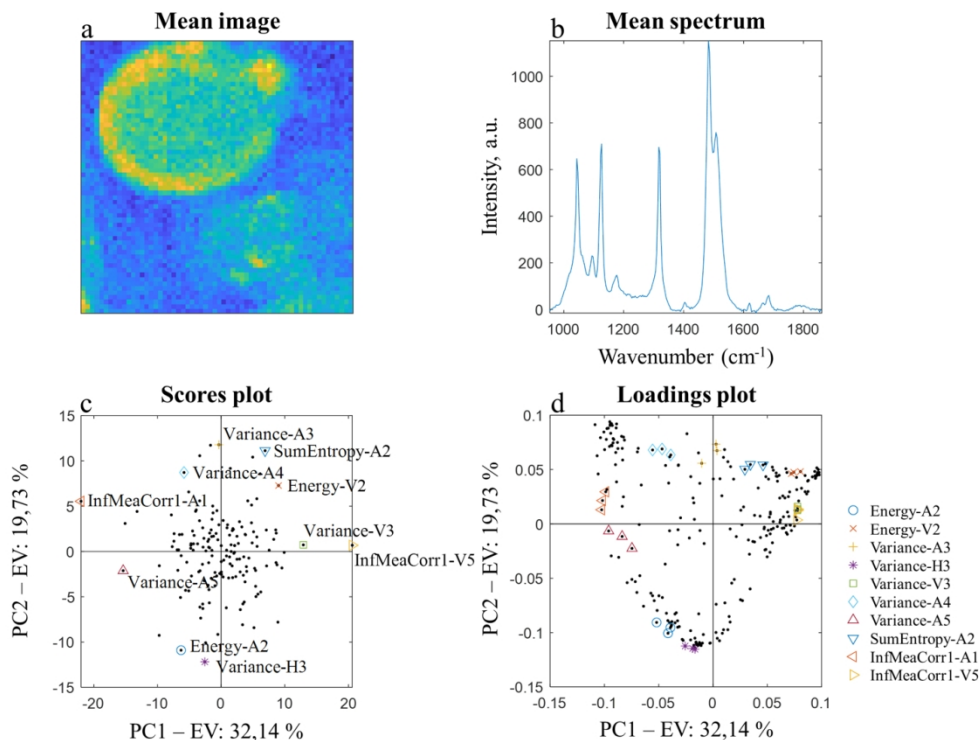


Figure 2. Oil in water data set. (a) Mean hyperspectral image (mean taken across spectral dimension). (b) Mean spectrum (mean taken across pixels after unfolding). (c) PC1 vs. PC2 scores plot (PCA of descriptors matrix). The most extreme values in the scores space are highlighted by different colored symbols and labelled. (d) PC1 vs PC2 loadings plot. The points in the loadings plot that correspond (i.e. are on the same direction with respect to origin) to the points highlighted in the scores plot are shown with the same symbols.

251x190mm (149 x 149 DPI)

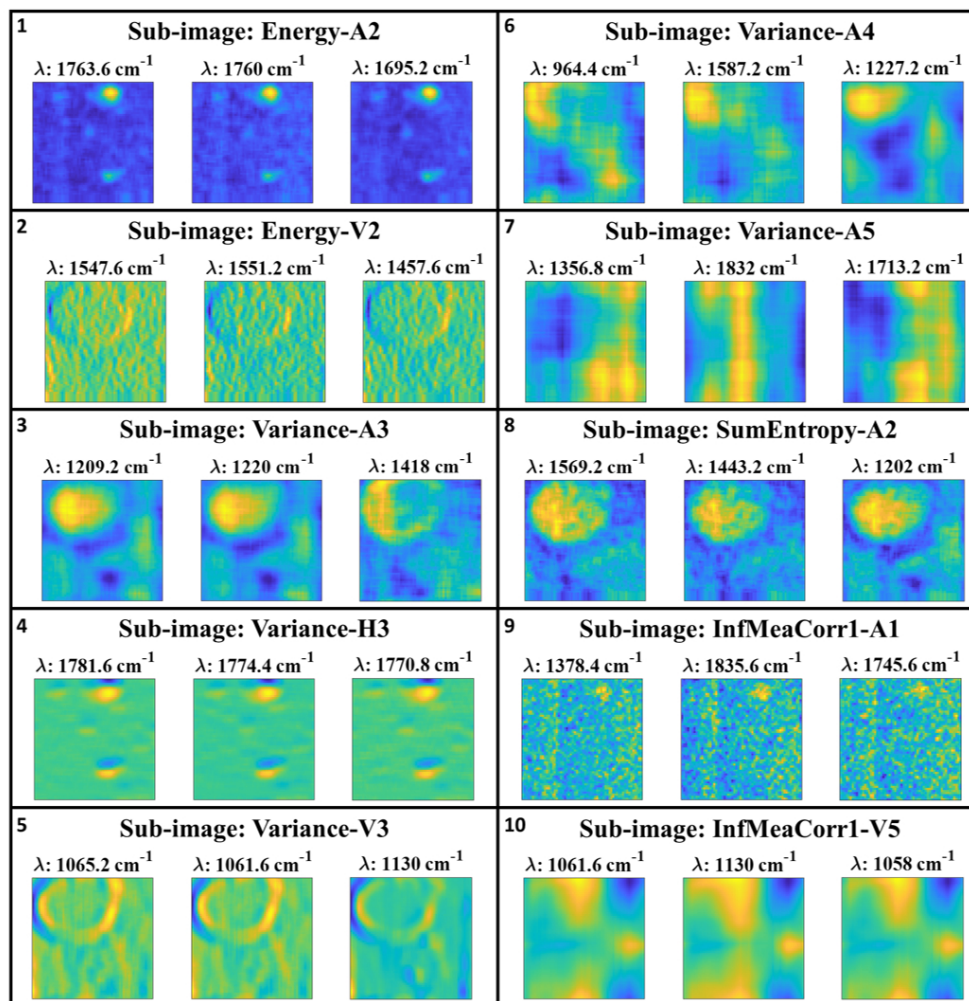


Figure 3. Oil in water data set. The wavelet sub-images, at decomposition block and level as reported in the points label in Figure 2c and at the spectral wavelengths corresponding to the ones showing the same symbols in Figure 2d, are shown in separate frames, numbered from 1 to 10. In each frame is reported on top the name of the descriptor-block-level, e.g. in frame 1 are shown the Approximation images at the second decomposition level for the three wavenumbers 1695.2, 1760 and 1763.6 cm^{-1} . Above each image is reported the corresponding wavenumber.

180x183mm (149 x 149 DPI)

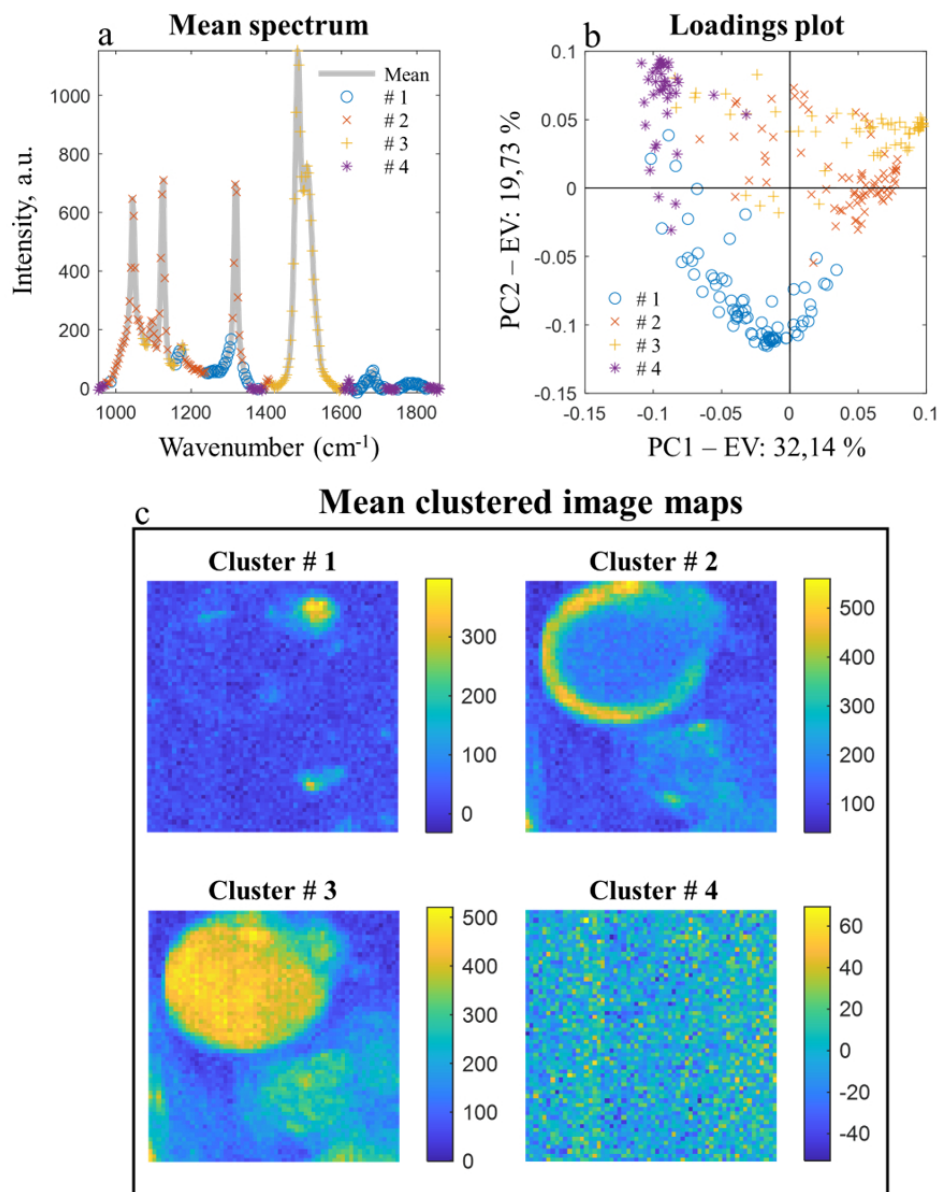


Figure 4. Oil in water data set. Results of the k-means clustering algorithm on the loadings matrix. (a) mean spectrum, with the point at each wavelength colored according to the cluster's number they were assigned to; (b) PC1 vs. PC2 loadings plot, points colored according to clusters; (c) mean images for each cluster, mean taken across the spectral channels belonging to the same cluster.

150x190mm (149 x 149 DPI)

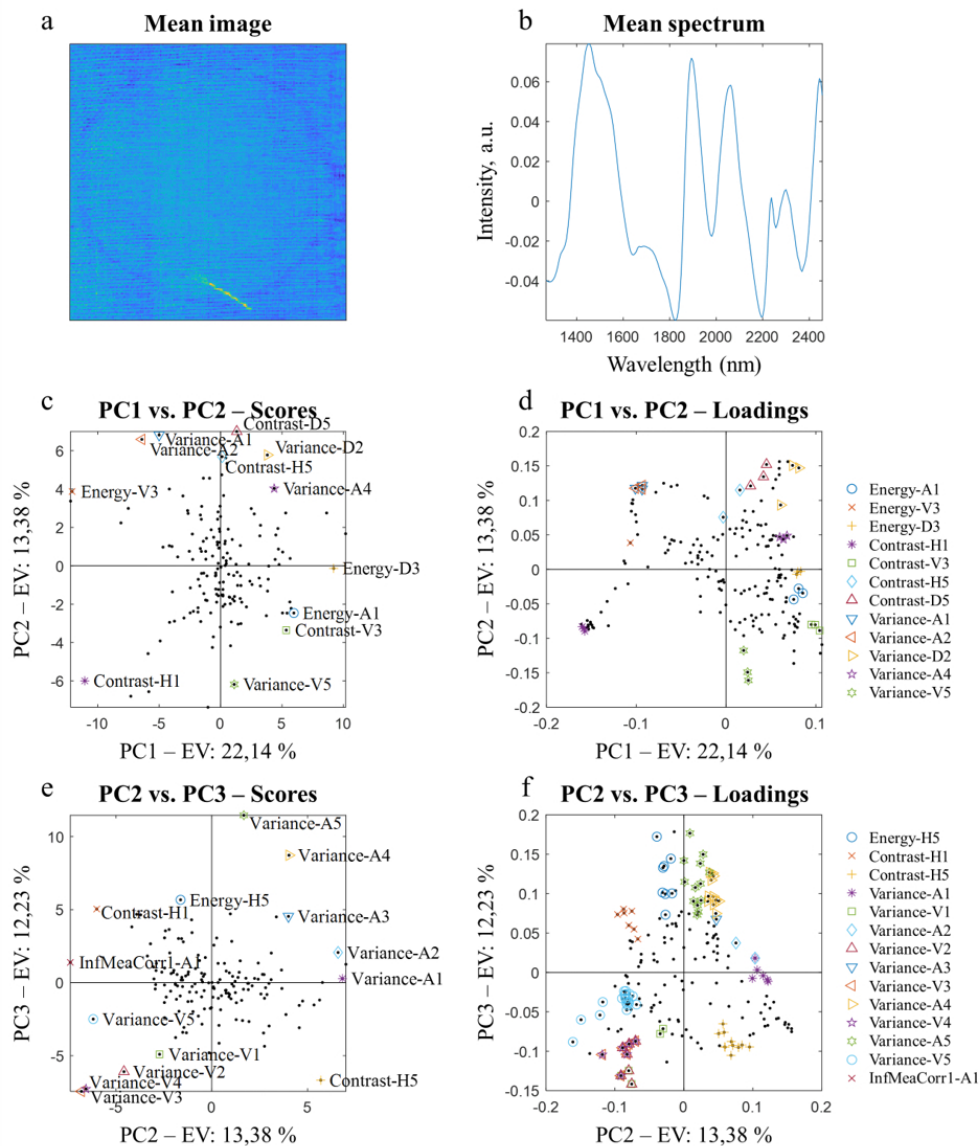


Figure 5. Semen data set. (a) mean image; (b) mean spectrum; (c) PC1 vs PC2 scores plot resulting from the application of PCA on the descriptors matrix (the ten most extreme values are labelled with the descriptor's name and the wavelet sub-image on which it has been calculated). (d) PC1 vs PC2 loadings plot. The points highlighted in the scores plot are shown with the same symbols; (e) PC2 vs PC3 score plot; (f) PC2 vs PC3 loadings plot.

162x190mm (149 x 149 DPI)

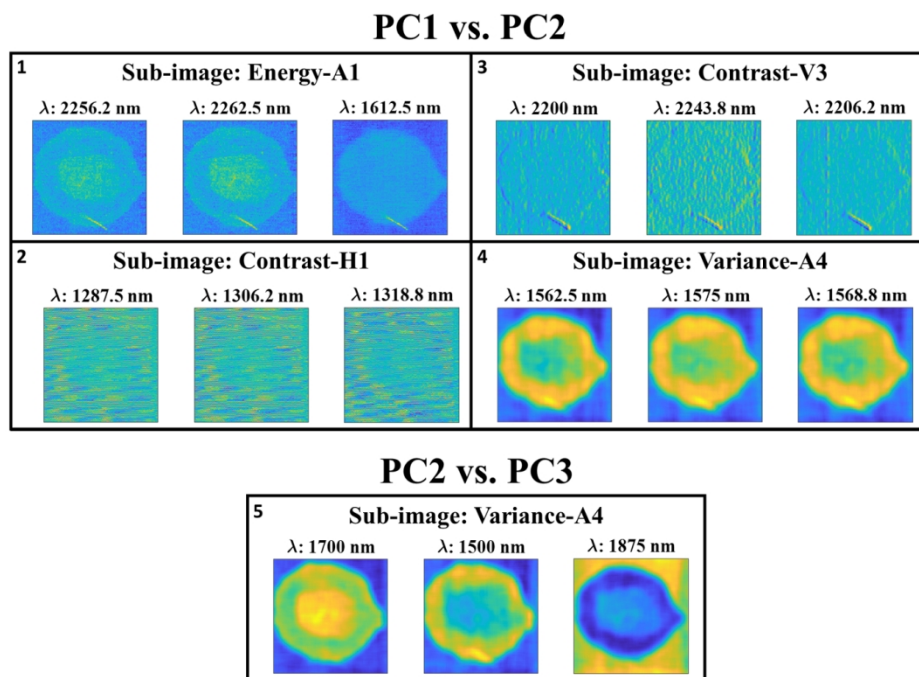


Figure 6. Semen data set. The selected wavelet sub-images from PCA analysis of descriptors matrix, shown in separate frames labelled by their respective number, i.e. in 6.1 are shown the Approximation images at the first decomposition level for the three wavelength 2256.2, 2262.5 and 1612.5 nm.

256x190mm (149 x 149 DPI)

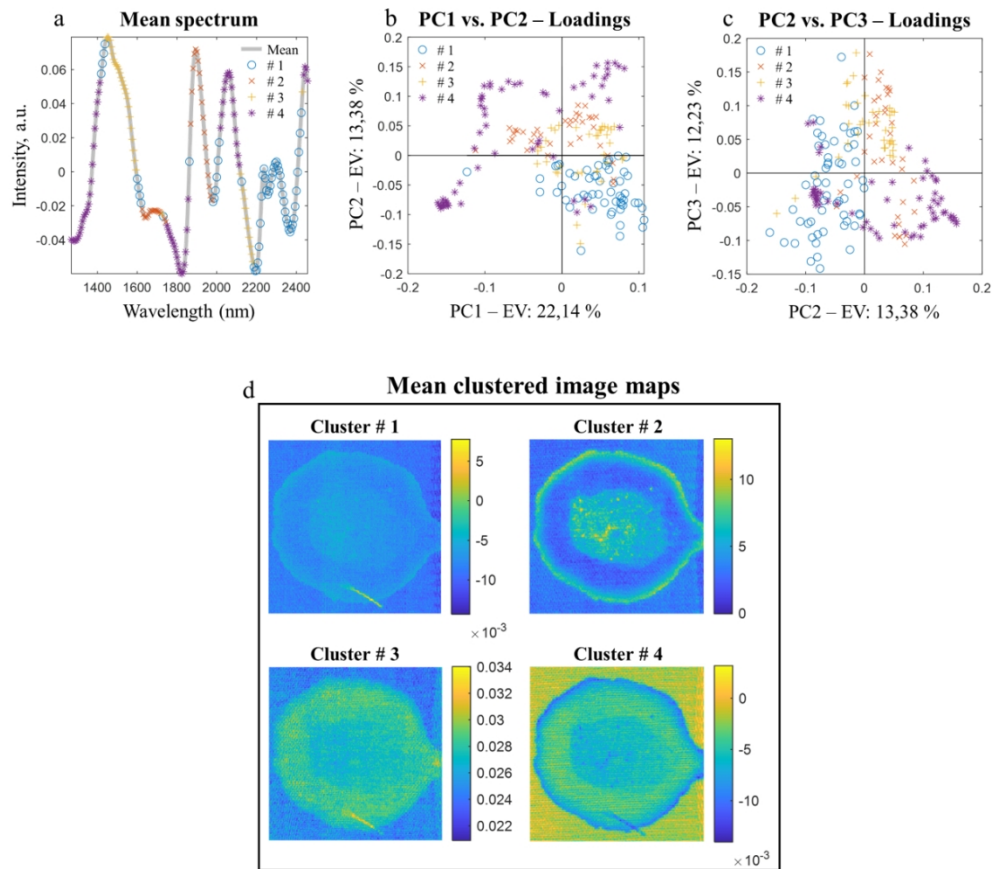


Figure 7. Semen data set. Results of the k-means clustering algorithm on the loadings matrix. (a) mean spectrum, with clustering highlighted on the spectral channels; (b) PC1 vs. PC2 loadings plot of, points colored according to clusters; (c) PC2 vs PC3 loadings plot; (d) mean image (across the spectral dimension) for each cluster.

214x190mm (149 x 149 DPI)