Infrared mapping of inorganic materials: a supervised method to select relevant spectra

T. Bonnal, E. Prud'Homme, S. Tadier, G. Foray

Version of Record: https://www.sciencedirect.com/science/article/pii/S0169743918301977
Manuscript_d27db1941c5b85cfb004b8cc621144fa

Infrared mapping of inorganic materials: a supervised method to select relevant spectra

Bonnal Thomas[a], Elodie Prud'homme[a], Solene Tadier[a,*], Genevieve Foray[a]

[a] *Univ. Lyon, INSA Lyon, UCBL, MATEIS UMR CNRS 5510, Bt. Blaise Pascal, 7 Av. Jean Capelle, F-69621 Villeurbanne, France*

Abstract

Spectroscopic imaging is expanding thanks to new instrumental concepts, technological developments, easier use and lower costs. This leads to collect and to handle huge hyperspectral databases. Among data providing useful information, there may also be unwanted information (i.e., either biased or irrelevant with respect to the information one wants to retrieve from the analysis). The classical approach aims at building a spectra library, which contains both informative spectra and outliers, so as to train learning algorithms. However, reference spectra cannot always be acquired, especially for samples prone to ageing or changes when exposed to humidity, temperature... Thus, building a library with reference spectra is not always possible. To handle this issue, a new supervised method (SSMS for Supervised Selective Method based on SIMPLISMA) has been designed and is described in this article, to identify and exclude unwanted spectra from the resolution process when no library is available. SSMS relies on a supervised exclusion of the unwanted spectra. It ensures both a quick treatment and an accurate

*Corresponding author:
Email address: solene.tadier@insa-lyon.fr (Solene Tadier)

analysis of the data (reduced number of representative spectra to supervise). This new method is applicable to any type of hyperspectral database. In this work, its efficiency is demonstrated on a database acquired using a FT-IR microscope. To avoid issues arising from the acquisition of maps with classic ATR crystals (cross-contamination and successive residual imprints on the material), the use of a new set-up called static-ATR is explored. In addition, the combined use of SSMS and of a physical model permits to identify the origins of the outlier spectra. Thus, it becomes possible to improve the experimental method (sample preparation, acquisition parameters...). Finally, with a constrained Alternating Least Squares method (ALS), relevant chemical information is obtained. The robust method developed here permits to achieve chemical maps at the micron scale for inorganic materials.
Keywords: Iterative Boundary Least Squares Method, Infrared Spectroscopy, Multivariate Curve Resolution, Inorganic Materials, Hyperspectral Imaging, Chemical Imaging

1. Introduction

Recent developments in advanced techniques for material analysis, such as characterizations in $3D$ [1, 2] or in complex environments [3], and an easier access to cutting-edge techniques lead to deal with a larger and larger amount of data (large scanned areas and volumes, monitoring of the evolution of a sample vs. time, large number of replicates...). In particular, the use of spectroscopy techniques (e.g., Raman [4], NMR [5] or terahertz [6]) requires to handle hyperspectral databases. More specifically, Fourier-Transform Infrared (FT-IR) microscopy is employed to obtain chemical mapping of ma-

terials and has known very interesting developments in the past few years [7, 8, 9].

Along with spectra providing useful information on the samples composition, these hyperspectral databases may also contain biased data (out-of-measurement ranges, noise, artifacts...) or unwanted information (e.g., linked to mounting resin or stand). Therefore, such big data analysis requires a thorough pre-processing stage to exclude non useful information and keep only the relevant spectra. Data, which are out-of-range or characterized by a too low signal-to-noise ratio, are relatively easily removed from databases. However, the identification of misinformative or unwanted information is based on preconceptions on their spectral characteristics or on their detection by learning algorithms (e.g. Random Forest classifier, Support Vector Machines, artificial Neural Networks). The latter approach is highly efficient but requires the algorithms to be trained on an available database. Such IR spectral libraries including relevant spectra and outliers are scarce, especially for inorganic materials.
In this work, a home made selective robust method, called SSMS (Supervised Selective Method based on SIMPLISMA), has been developed to eliminate non useful spectra based on statistical considerations without needing any training database. SSMS enables to identify irrelevant spectra without any *a priori* knowledge of the material. Such an automatic treatment becomes essential to enable the analysis of large databases.

SSMS algorithm is detailed and illustrated on the obtaining of chemi-

cal maps issued from FT-IR microscopy acquisitions. The use of a particular set-up, called static attenuated total reflection (s-ATR), is shown to be an efficient tool for the mapping of inorganic materials. A granular composite used for construction applications, containing a very sensitive ettringite binder, was successfully studied avoiding cross-contamination and surface degradation issues. SSMS permitted to exclude spectra related to artifacts linked to the s-ATR set-up without any preconceptions on the material. Chemical maps were obtained using a Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) procedure. It is shown how SSMS can be used as an efficient pre-processing step in the analysis of any type of hyperspectral database.

2. Theoretical background

2.1. FT-IR spectroscopy

Infrared spectra can be obtained from different acquisition modes measuring reflected beam, transmitted beam or a combination of both [10]. The intensity I of these measured beams is then compared to the intensity I_0 of the incident beam. To do so, absorbance A and reflectance R are defined as $A(\omega) = log_{10}\left(I_0(\omega)/I(\omega)\right) = -log_{10}(R(\omega))$, ω being the wavenumber.

The transmission mode is not applicable for most minerals since they are dense and absorbent. **Focusing on the reflective mode**, three different optical phenomena have been exploited for FT-IR spectra acquisitions: specular, diffuse and total reflections, the latter one occurring when the incident angle is larger than the critical angle of the studied material.

The amount of light detected by the sensor is the key parameter to obtain exploitable spectra (i.e., bands relative to covalent bonds are predominant in the spectra). It can be influenced by material parameter as 1) the roughness of material surface and 2) the material nature. 1) An increase of the sample roughness leads to an increase of the scattered light and to a decrease of the signal-to-noise ratio [11]. 2) The imaginary refractive index influences the spatial distribution of the light scattered by the material [12, 10]. This may result in modulations and distortions in spectra [13]. Moreover, in a quantitative study, a minimal number of photons has to be detected to obtain a measurement included in the linearity range of the sensor. This may be achieved by increasing the number of scans or by minimizing the roughness of the material.

Acquiring a spectrum using the **specular reflection** is the easiest technique to implement, as it does not require any specific set-up: the infrared beam is directly projected through the ambient air at the surface of the sample. A sensor placed above the sample then records the reflected beam (Fig.1.a).

Crystals, which are used to achieve Attenuated Total Reflection **ATR**, have no absorption in the infrared range and a much higher refractive index than the studied material. The incident infrared beam travels through the ambient air and then through the ATR crystal before reaching the surface of the sample beyond the critical angle. Thus, total internal reflection occurs, maximizing the amount of reflected light. Practically, a vertical load is ap-

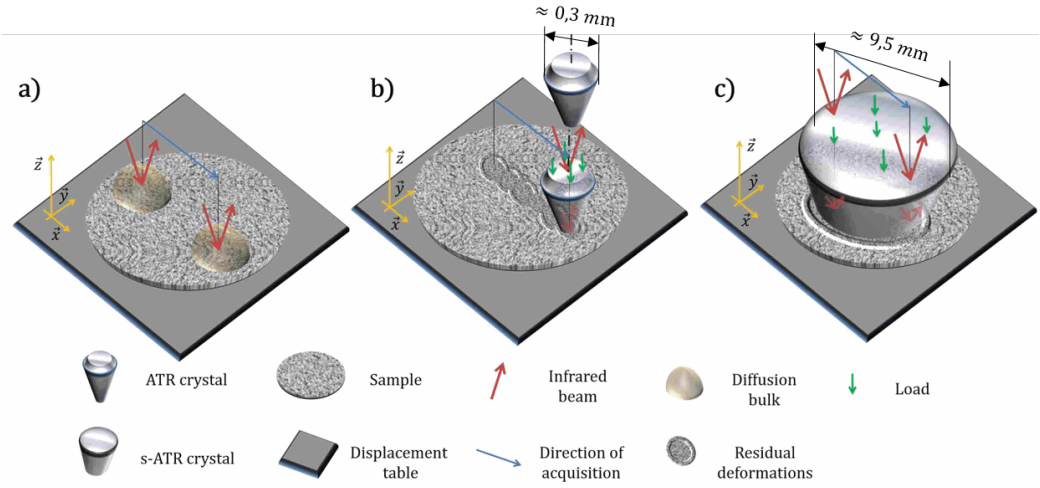

Figure 1: Schematic view of FT-IR microscopy set-ups: a) Specular Reflection, b) ATR Crystal and c) Static ATR.

plied on the crystal to ensure a neat material/crystal interface. This induces stresses at the interface (Fig.1.b).

2.2. FT-IR microscopy

To obtain a $2D$ FT-IR map ($(\vec{x}, \vec{y})$ plane), the sample is placed on a micrometric (x, y, z) motorized stage to scan its surface, which enables the acquisition of FT-IR spectra at different locations.

The spatial resolution of the resulting map is theoretically limited by the diffraction limit described by the Rayleigh criterion r:

$$r = \frac{0.61.\lambda}{NA} \tag{1}$$

Where λ is the wavelength and NA is the numerical aperture of the optical system. The numerical aperture is defined by $NA = n.\sin\theta$ and depends on the incident medium (n: refractive index of incident medium: air or ATR

crystal) through which the light travels. It also depends on θ, the half angle of the cone of incident light [14]. The spatial resolution is the minimal distance for which two adjacent points can be separated and is theoretically equal to $2r$. In practice, optical aberrations lead to a worse spatial resolution [15]. Thus, the actual spatial resolution has to be experimentally measured. Methods based on optical reference targets are often used [7, 16] but Kazarian [17] recently suggested a more accurate method when ATR crystals are involved in the acquisition process.
Finally, a distinction has to be made between the spatial discretization arisen from the step of the motorized stage and the actual spatial resolution [14].

To acquire a FT-IR map in **specular reflection**, no additional set-up is required (Fig.1.a).

For **ATR**, the crystal, located above the sample, is motionless. The stage moves vertically up (z direction) to make the contact between the sample and the crystal. Then, the stage keeps going up until a specified stress ($\sim 10\ N$) ensures a good contact between the sample and the crystal: a spectrum is recorded. Then the stage moves down before switching to the next location ($(\vec{x}, \vec{y})$ plane) (Fig.1.b).
The crystal focuses the incident beam and ensures a higher spatial resolution. However, cross-contamination may be induced by the (x, y) displacements. Moreover, with a crystal diameter of few hundreds of microns, localized loads of about 80 MPa are generated. They may lead to residual imprints in case of sample with weak mechanical properties [18]. Caution should be taken for

sensitive material to avoid or check for such artifacts.

The **static-ATR** (s-ATR) set-up enables performing ATR without applying successive localized loads onto the sample. The crystal has a conical shape with a half-sphere on its top (Fig.1.c). The bottom flat surface is about ten times larger than usual ATR crystals and is put in contact with the sample only once for the whole acquisition process. The resulting block {crystal + sample} is fixed on the motorized stage. When the stage moves, the location of the incident beam on the upper base of the crystal is changed (Fig.1.c), enabling the acquisition of FT-IR microscopy maps. This minimizes damage at the surface, avoids potential cross contamination issues and prevents any pressure loss between the crystal and the sample. However, this technique is seriously affected by surface imperfections: a high roughness can lead to signal loss for some areas. In addition, a lack of parallelism between the two sample faces can hinder the contact between the crystal and the sample. This contact is ensured by applying a localized load, which may lead to one residual imprint (Fig.2).
Finally, some of the spectra acquired with this set-up are aberrant (i.e., to the authors' knowledge, not corresponding to the actual material). These artifacts will be further denoted 's-ATR artifacts' and discussed later in the manuscript.

Spatial resolutions and other comparative items of the different set-ups are summarized in Table 1, while their schematic views are displayed in Fig.1. Actual spatial resolutions have been provided by the manufacturer of the mi-

croscope used in this study.

Table 1: Summary of FT-IR microscopy set-ups to obtain FT-IR maps in reflection

FT-IR microscopy set-ups	Spatial Resolutions ($2100 - 2000\ cm^{-1}$)	Advantages	Drawbacks
Specular Reflection	25 μm (25 μm)[a]	Fast acquisition No contact with sample No CC [b]	Distortions in spectra[c]
ATR Crystal	10 μm (6.25 μm)[a]	Spectral accuracy Point specific zones Access to recessed areas	Spectral range cut-off Slow acquisition Successive residual imprints[d] Possible CC[b]
static ATR (s-ATR)	10 μm (6.25 μm)[a]	Spectral accuracy Sample compressed only once No CC[b]	Spectral range cut-off Need for lower roughness Need for more even surface Limited mapping area 's-ATR' artifacts[e]

[a] Pixel size; [b] CC: Cross-Contamination; [c] Perturbations due to heterogeneous scattering; [d] May disturb the structure or deform samples; [e] 's-ATR' artifacts will be discussed in section 5 of this article.

2.3. Obtaining a FT-IR map from acquired spectra: bilinear modelling

First of all, the following hypotheses have to be made to obtain a FT-IR map: (i) the studied material is made of p compounds, (ii) each compound is isotropic and (iii) its characteristics remain unchanged versus time. It is important to note that a compound may be composed by a mixture of several chemical phases.

The Beer-Lambert law links absorbance values $A(\omega)$ for a wavenumber ω with the molar attenuation coefficient ϵ, the concentration of absorbing species c

and the path length l: $A(\omega) = \epsilon(\omega).l(\omega).c$.

Working on relative quantities of a mixture of p compounds, this law can also be expressed by:

$$A(\omega) = \sum_{i=1}^{p} s_i(\omega).c_i \tag{2}$$

with s_i and c_i the spectrum and the relative concentration related to each compound $i \in [1; p]$, respectively. Acquiring one FT-IR spectrum per location, a $3D$ data matrix D_{3D} of dimensions $[x \times y \times m]$ is constructed [19]. x and y are the numbers of acquisitions in the two spatial directions $(\vec{x}, \vec{y})$. m is the number of wavenumbers for which absorbance values are recorded. As most chemometrics tools are not designed to directly handle $3D$ matrices, the two spatial indexes are combined to store the data in a $2D$ matrix:

$$[x, y, m] \rightarrow [n, m], \ n \in [1; x \times y] \tag{3}$$

$$\underset{[x \times y \times m]}{D_{3D}} \rightarrow \underset{[n \times m]}{D} \tag{4}$$

Finally, the application of the Beer-Lambert law to the matrix of data leads to consider the following bilinear problem:

$$\underset{[n \times m]}{D} = \underset{[n \times p]}{C} \, . \, \underset{[p \times m]}{S^t} + \underset{[n \times m]}{E} \tag{5}$$

S is the matrix of spectra relative to each compound, C the matrix of concentrations and E the matrix of residual errors [20].

3. Material and methods

3.1. Material

A construction material has been fabricated to illustrate the method developed in this work. A cylinder (diameter: 30 mm, height: 50 mm) of

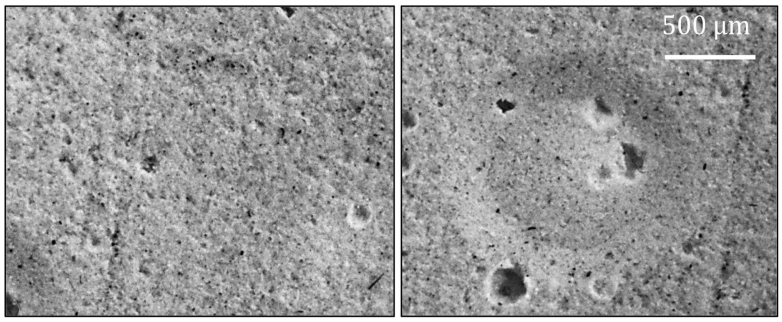

Figure 2: Optical micrography of the granular composite with ettringite binder: a) surface state before acquisition, and b) s-ATR crystal imprint in the material after s-ATR acquisition.

granular composite has been prepared with equal weights of ground silica and limestone filler. Their granular size distribution was comprised in the range $[0.3; 300]$ μm, the carbonate filler being smaller than the siliceous one. They were binded together with an ettringite binder, based on calcium aluminate cement and anhydrite. This granular composite with ettringite binder was elaborated according to Michel et al. [21]. This material includes non stable ettringite hydrates. Carbonate groups may be incorporated over time, which drastically influences the composition and the properties of the material. The phases, which are formed in small amounts during this process, are poorly crystalline. Thus, their identification and characterization are challenging.

For FT-IR microscopy observations, a $4 \times 4 \times 4$ mm^3 cube was extracted from the core of the cylindrical sample. One surface of the cube was resurfaced using a target surfacing system EM TCP (*Leica microsystems*, Germany)

with a milling step of $1\mu m$, followed by a polishing (Fig.2.a).

3.2. FT-IR microscopy equipment and acquisition parameters

A FT-IR microscope *Nicolet*^{TM} *iN*10 from *ThermoFisher Scientific* (USA) has been used. The only set-up used in this study was the static-ATR, using a Germanium crystal (refractive index: 4, top diameter: 9.525 mm, bottom diameter: 0.350 mm).

The aperture was 100 $\mu m \times 100$ μm. The focus of the beam due to combined use of the Cassegrain reflector and of the *Ge* crystal led to a projected size of the IR beam of 25 $\mu m \times 25$ μm. A Mercury Cadmium Telluride (MCT) array sensor cooled with liquid nitrogen was used, leading to an effective pixel size of 6.25 $\mu m \times 6.25$ μm. Its inferior limit of linearity corresponds to 5 % of the incident beam. Actual spatial resolutions of the set-ups (Table 1) were evaluated by the manufacturer using a USAF 1951 spatial resolution target. The stage moved by a ($(\vec{x}, \vec{y})$ plane) step of 25 μm. The scanned area was about 300×515 μm^2 leading to a map composed of 48×82 pixels.

Spectra were collected with a spectral resolution of 8 cm^{-1} in the Infrared range between 717 and 4000 cm^{-1}, with 16 scans per spectrum. *Omnic Picta*^{TM} software was used to acquire FT-IR maps.

4. Description of the resolution process

The resolution process developed in the following sections aims at obtaining relevant chemical maps describing the spatial distribution of each compound at the surface of the material.

A schematic view of the preprocessing steps is shown in figure 3. As detailed in section 2.3, once one FT-IR spectrum has been acquired per location, a

$3D$ data matrix of dimensions $[x \times y \times m]$ is built (Fig.3, Step (1)). Finally, the bilinear problem is described by equation (5):

$$\underset{[n\times m]}{D} = \underset{[n\times p]}{C} \cdot \underset{[p\times m]}{S^t} + \underset{[n\times m]}{E}$$

S is the matrix of spectra relative to each compound, C the matrix of concentrations and E the matrix of residual errors.

In our application, the number of compounds (and their related spectra) is *a priori* unknown leading to an under-determined system of equations. Thus, a first step is needed to determine the number of compounds and their relative spectra. The resolution of this step will be described in section 4.1. It permits to obtain an over-determined system, which can then be solved to find the maps of relative concentrations, as detailed in section 4.2.

In this work, all data preprocessing and resolution techniques were implemented in a home-made algorithm encoded with *Matlab R2015a* using *Curve Fitting*, *Global Optimization* and *Statistics & Machine Learning* toolboxes.

4.1. Initialization of the iterative process

- Data preprocessing

Before solving the bilinear problem, outlier spectra had to be excluded from the resolution process in order to avoid chemically non-meaningful solutions. Spectra with values out of the linearity range of the MCT sensor (linearity threshold of 1.3 in absorbance) were therefore removed from the data (Fig.3, Step (2) : deep blue). Spectra with a too low signal-to-noise ra-

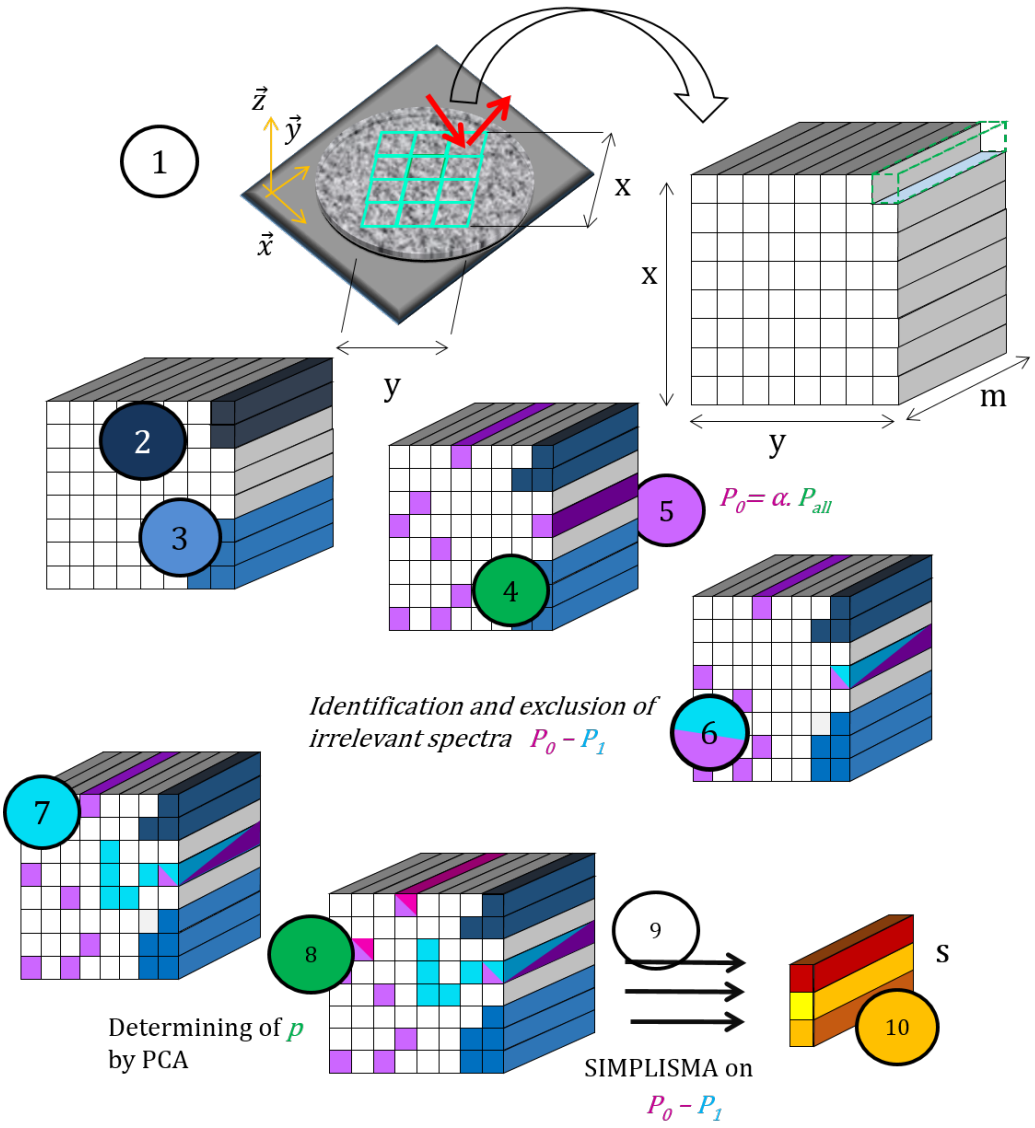

Figure 3: Schematic view of data preprocessing (steps 1 to 3) and initialization with SSMS (steps 4 to 10). x, y, m and p are the numbers of acquisitions in the two spatial directions $(\vec{x}, \vec{y})$, the number of wavenumbers and the number of compounds respectively. Step numbers and color code defined here are recalled in the text to explain the whole process. The figure shows how irrelevant spectra (in color) are excluded from the database, step after step. The color code is consistent with colors used in Fig.4 & Fig. 5.

tio (highest absorbance value lower than 0.4 for this study) were also removed (Fig.3, Step (3): blue).

- Supervised Selection Method for Spectra (SSMS)

After the exclusion of outlier spectra by the previous data preprocessing stage, a classical approach would consist in two more steps: (i) determining of p, the number of compounds, and (ii) first guess for the matrix S.
However, the use of the s-ATR set-up induces artifacts (Table 1), resulting in biased spectra, which would not be identified using this classical approach. Due to their singularity, these biased spectra would be assimilated to fictive compounds. This could lead to misjudge the number of compounds p and hence the matrix S of spectra.
To deal with this problem, the algorithm Supervised Selection Method for Spectra (SSMS) was developed to solve the whole issue at once: solving the two steps (i) and (ii) all together including the detection of s-ATR artifacts and the exclusion of related spectra.
SSMS uses SIMPLISMA, an iterative algorithm designed by Windig et Guilment [22] to select the most singular spectra out of a database. In SIMPLISMA, at each iteration, the purest spectrum is selected as the one which forms the "most orthogonal" hypercube with the spectra already selected. For more details about the SIMPLISMA algorithm, see [22, 23, 24].

SSMS consists in the following steps:

i First evaluation of the number of compounds P_{all} based on a Principal Component Analysis (PCA). Then, a threshold has to be set to determine the number of compounds. In this work, the Kaiser's criterion

(intensity of components higher than 1) was used to give a first estimation of P_{all} [25, 26](Fig.3, Step (4): green). As a validation for this estimation of P_{all}, a second criterion was based on the slope change in the plot of component intensity vs. component index (obtained from PCA). Segregation methods could also have been used but, as for PCA, they would also require a criterion or a threshold to be fixed.

Since spectra containing s-ATR artifacts are singular, they may be considered as pure spectra by SIMPLISMA and contained in P_{all}.

ii Application of SIMPLISMA to select a number $P_0 \approx \alpha P_{all}$ of spectra out of the preprocessed database (i.e., initial database - outlier spectra) (Fig.3, Step (5): purple). The coefficient α ensures that the number P_0 of selected spectra is small enough to be easily and accurately checked by the operator, but large enough to contain at least one spectrum representative of each actual compound and of each unwanted spectrum.

iii Examination of the P_0 spectra, supervised selection and exclusion of the P_1 spectra presenting s-ATR artifacts among P_0 (Fig.3, Step (6): half light blue, half purple symbolizing the spectra first selected within P_0 (purple ones) and then excluded by the operator). At this stage, it is also possible to select and exclude spectra, which do not present any interest for a specific study (e.g., those of a mounting resin).

iv Back to the preprocessed data matrix (i.e., initial database - outlier spectra): all unwanted spectra, presenting a correlation coefficient $>$ 0.95 with the aforementioned P_1 spectra are identified by statistical considerations in the preprocessed data matrix (i.e., initial database -

outlier spectra) and excluded from the bilinear problem (Fig.3, Step (7): light blue). At this stage, a reduced matrix is obtained, which only contains relevant spectra with respect to the study.

v PCA on the reduced database (i.e., obtained after Step (7)) to find p (Fig.3, Step (8): green).

vi Selection of the p most different spectra using SIMPLISMA on the remaining $P_0 - P_1$ spectra (Fig.3, Step (9)).

vii The p spectra selected previously form the first estimation of the matrix of spectra S ((Fig.3, Step (10): red, yellow and orange).

At this point, SSMS gives a first guess of p, the number of compounds, and of S their relative spectra. Like for every other statistical methods, the process of resolution has then to be applied with different numbers of compounds around this first estimation ($[...p-1, p, p+1, ...]$) (see [2]: chapter 2, section 3 for further details).

4.2. Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS)

With a first estimation of the matrix S, it is now possible to solve the bilinear problem with a regular ALS method. Constraints and iterative process are detailed thereafter.

4.2.1. Constraints:

To obtain chemically acceptable solutions, different kinds of constraints can be taken into account:

Physical constraints Non-negativity of spectra and concentrations, normalization of spectra or sum of relative concentrations equal to one (closure) are

quite easy to handle with an ALS process [27, 20].

Model constraints When an evolutive process is studied with a known physico-chemical model, specific constraints arising from this model can be applied [28].

Statistical constraints From statistical considerations, such as local rank analysis [29, 30], the number of compounds can be imposed in the resolution process.

Spectra normalization and closure cannot be used together [27]. In this study, spectra have been normalized ($\frac{D_i}{\|D_i\|}$) before ALS process but closure was applied during ALS to remain in a linear resolution. In addition, tolerances ϵ_k were included in all the constraints, leading to the following inequalities:

- All absorbance values in the range $[0, L_{sensor}]$,

$$0 - \epsilon_1 \leq S_{ij} \leq L_{sensor} + \epsilon_1 \tag{6}$$

With L_{sensor} the upper limit of linearity of the sensor (the upper limit in absorbance corresponds to the inferior limit in transmittance).

- Non-negativity of concentrations and local rank constraints,

$$0 - \epsilon_2 \leq C_{ij} \leq C_{local\ rank,ij} + \epsilon_2 \tag{7}$$

Fixed Size Image Window-EFA algorithm (FSIW-EFA) [30] was used to evaluate $C_{local\ rank}$, which was a matrix only composed of 1 and 0 (absence of a compound was marked by a 0).

- Closure.

$$1 - \epsilon_3 \leq \sum_{j=1}^{p} C_{ij} \leq 1 + \epsilon_3 \tag{8}$$

As the number of compounds was considered to be well-defined, the sum of relative concentrations was closed to one in all acquisitions.

4.2.2. Iterative process:

Under these constraints, the iterative process consisted in successive evaluations of C and S until convergence was achieved:

- Evaluation of C:

$$\min_{C} \parallel D - C.S^t \parallel \quad \begin{cases} 0 - \epsilon_2 \leq C_{ij} \leq C_{local\ rank,ij} + \epsilon_2 \\ 1 - \epsilon_3 \leq \sum_{j=1}^{p} C_{ij} \leq 1 + \epsilon_3 \end{cases} \tag{9}$$

- Evaluation of S:

$$\min_{S} \parallel D - C.S^t \parallel \quad \left\{ 0 - \epsilon_1 \ll S_{ij} \ll L_{sensor} + \epsilon_1 \right. \tag{10}$$

- Evaluation of the quality of the reconstructed data (Lack of Fit):

$$LoF = \frac{\parallel D - C.S^t \parallel^2}{\parallel D \parallel^2} \tag{11}$$

In this study, the function *lsqlin* in *Matlab R2015a* was used to deal with the constrained linear least-squares problem. If spectra normalization had been chosen instead of imposing closure, the function *fmincon* could have been used to deal with non linear constraints.

5. Results and discussion

5.1. SSMS applied on a granular composite with an ettringite binder

The new methodology developed in this work is illustrated step by step on a s-ATR map (48 X 82 pixels, i.e. 3936 spectra) acquired on a mineral

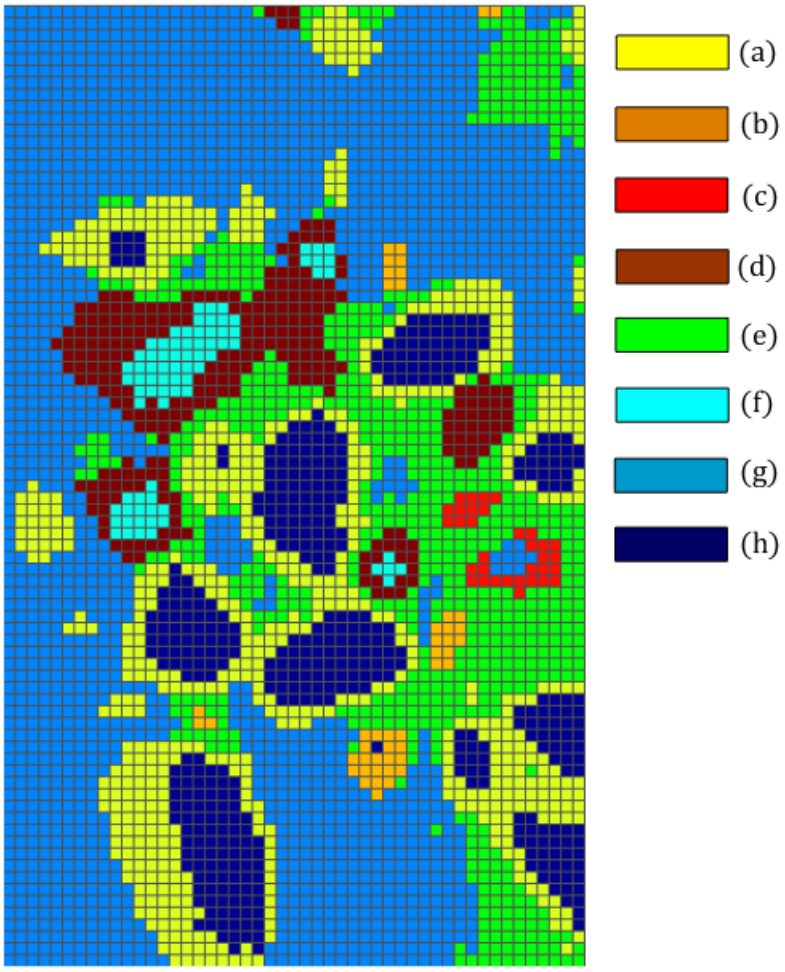

Figure 4: Spatial distribution after SSMS method of spectra assigned to: a) compound 1 in yellow, b) compound 2 in orange, c) compound 3 in red, d) compound 4 in brown, e) a mixture of several compounds in green, f) s-ATR artifacts in light blue. Spectra with: g) a too low signal to noise ratio and h) values above linearity range are represented in blue and deep blue respectively. A pixel is assigned to one of the previous items if their spectra match (correlation coefficient > 0.95). The projected pixel size is about $6.25 \times 6.25\ \mu m^2$ per pixel, the map is composed of 48×82 pixels.

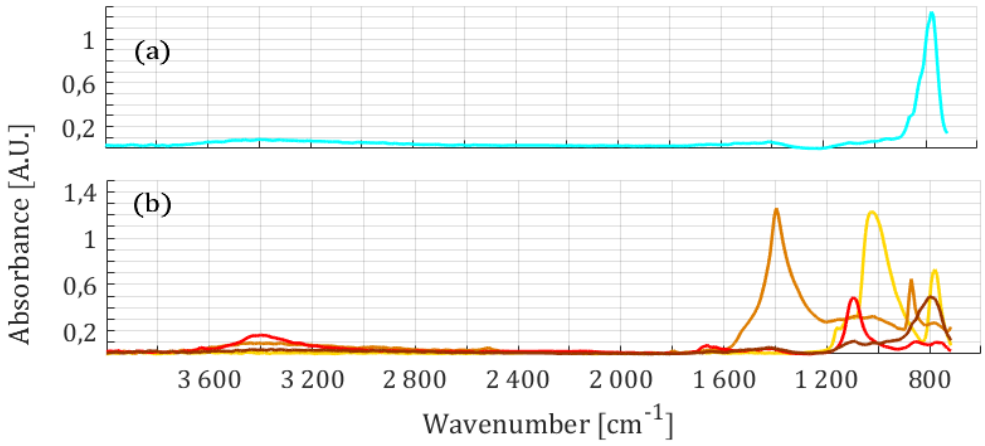

Figure 5: Infrared spectra of the granular composite with s-ATR set-up. a) Spectrum identified by the operator among the P_0 spectra as corresponding to s-ATR artifacts (step 6; in this example only $P_1 = 1$ spectrum was excluded from the P_0 spectra) and b) first estimation of the matrix S obtained by SSMS prior to MCR-ALS, as schemed in Fig.3, Step 10: compound 1 in yellow, compound 2 in orange, compound 3 in red and compound 4 in brown.

granular composite with an ettringite binder.

Applying SSMS has first led to exclude 452 spectra (11.48% of the database) due to values out of the linearity range of the sensor (Fig.3 step (2), Fig.4.h). It has also led to exclude 1654 spectra (42.02% of the database) because of a too low signal to noise ratio (Fig.3 step (3), Fig.4.g). Then, using a PCA study on the remaining spectra, P_{all} was found to be equal to 4 (Fig.3 step (4)). To make sure that at least one spectrum for each actual compound and one spectrum for each possible artifact was selected, α was set to 4. Hence, using SIMPLISMA, $P_0 = 16$ spectra have been selected in the database (Fig.3 step (5)). Thus, with α equal 4 (same order of magnitude as P_{all}), it became possible for the operator to accurately examine each of the $P_0 = 16$ spectra out of the 3936 initial spectra. As a result, $P_1 = 1$ spectrum presenting s-ATR artifacts has been identified by the operator among the P_0 spectra (Fig.3 step (6), Fig.5.a). 54 spectra (1.37% of the database) were correlated to the previous spectrum and were therefore excluded from the bilinear problem (Fig.3 step (7), Fig.4.f). Then, a PCA was used to find the number of compounds on the reduced database. Similarly to the previous estimation, $p = 4$ compounds were found (Fig.3 step (8)). Thus, the number of compounds defined before and after SSMS (i.e., before and after excluding the irrelevant spectra from the database) was the same (Fig.6.a). In this case, SSMS led to exclude only 1.37% of the database, which explains why it did not impact on the number of compounds found by a PCA. However, if SSMS had been used to remove more spectra (e.g., all the spectra corresponding to a mounting resin), the number of compounds found after SSMS

would have been different than the one found before. In addition, more than the determined number of compounds, SSMS brought a real benefit in the determining of the spectra of S: without SSMS, the spectrum of the second compound obtained with SIMPLISMA was an artifact spectrum (data not shown); whereas using SSMS, the first four spectra corresponding to the $p = 4$ compounds were all chemically meaningful (Fig.5.b).

To save time, SIMPLISMA was performed on the $P_0 - P_1 = 15$ purest spectra rather than on the reduced database (Fig.3 step (9)). Finally, the matrix S was built (Fig.3 step (10), Fig.5.b).
In a first approximation, the spatial distribution of the four compounds was found with a correlation coefficient higher than 0.95 (Fig.4.a,.b,.c & .d).

5.2. Chemical maps: computation and analysis

The linearity limit of the MCT sensor has been taken equal to $L_{sensor} = 1.3$ in absorbance in the ALS resolution. To release the constraints, ϵ_1 ϵ_2 and ϵ_3 were fixed equal to 10^{-4}. The evolution of the iterative process is illustrated by the Lack of Fit criterion. As shown in Fig.6.b, convergence of the iterative process was achieved after 13 iterations. During the convergence, spectra selected by SSMS evolved (band shifts, band intensities, smoothing). After the 13^{th} iteration, only non-significant modifications occurred.

Once the bilinear problem was solved, matrices S and C were obtained. The matrix S contained spectra corresponding to the $p = 4$ compounds (Fig.7).

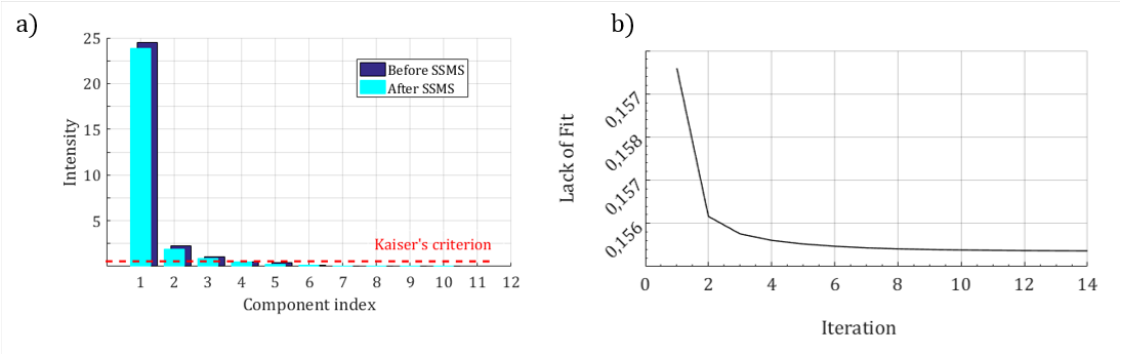

Figure 6: a) Intensities of the first components (PCA) before and after SSMS b) evolution of the Lack of Fit criterion vs. ALS iteration.

In an IR spectrum, each band is representative of the vibration of a covalent bond within its environment in the material. Thus, the four compounds could be clearly identified thanks to their characteristic bands, well described in literature (Table 2). The first compound was identified as silica (Fig.7.a). The second compound was identified as crystalline calcium carbonate with the presence around 1400 cm^{-1} and around 740 cm^{-1} of characteristic vibration bands of carbonate group (Fig.7.b). The combination of compounds 1 and 2 formed the granular structure. The third compound was associated to the ettringite binder thanks to the OH bonds around 3630 cm^{-1} and to the sulfate bonds at 1088 cm^{-1} [34]. The small shoulder at 1020 cm^{-1} (Fig.7.c) may be assigned to $Al-O$ bonds, due to the presence of gibbsite in the material.

Gibbsite is a classical phase formed simultaneously with ettringite. Due to its poor crystallinity, very few techniques are able to track gibbsite. This highlights the important role played by infrared analysis in phase identification.

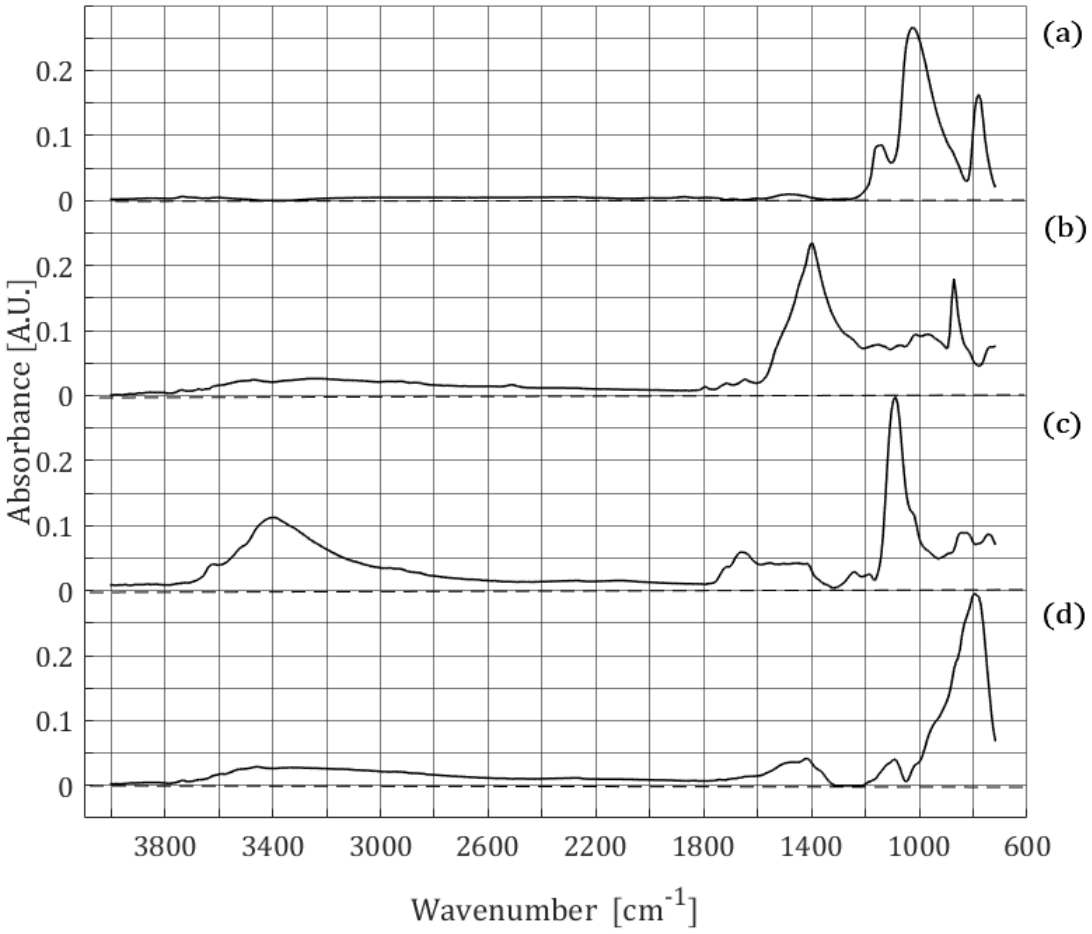

Figure 7: Normalized spectra obtained after SSMS coupled with ALS resolution: a) compound 1, b) compound 2, c) compound 3, and d) compound 4.

Compound 4 was identified as ettringite containing carbonate groups (Fig. 7.d). The combination of compounds 3 and 4 formed the binder phase.

From the matrix C, chemical mapping of each compound may also be displayed (Fig.8). The maps highlight the presence of different areas: zones where one compound was highly predominant (relative concentration close

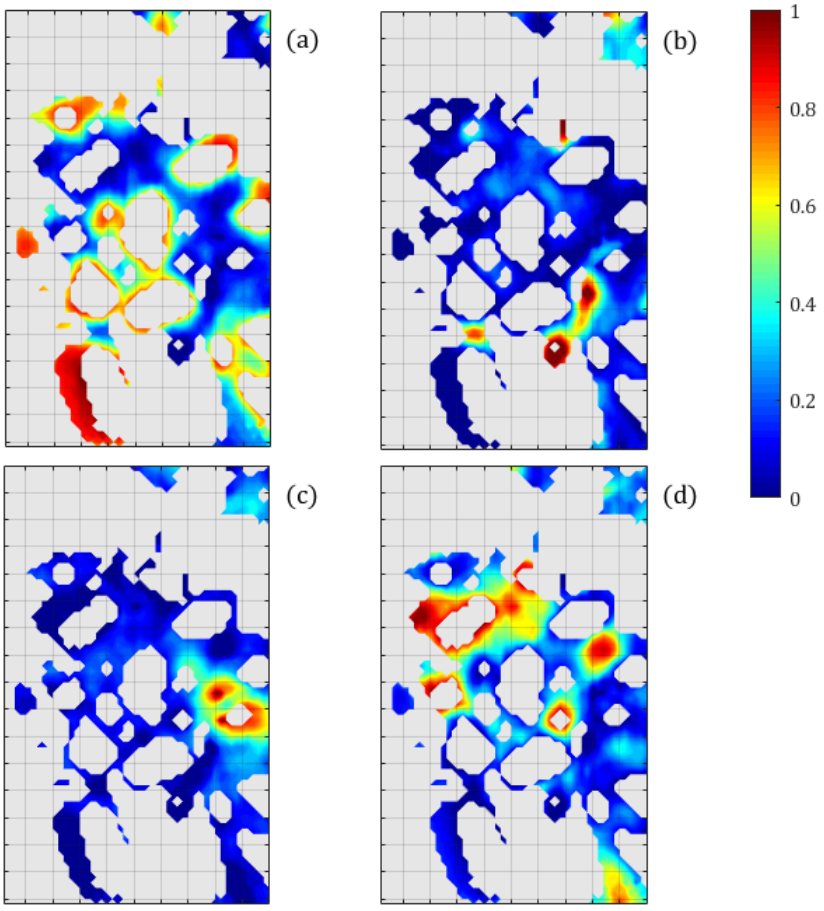

Figure 8: Maps of relative concentrations obtained by a s-ATR acquisition coupled with SSMS-ALS resolution for the granular composite with an ettringite binder: a) compound 1 assigned to silica, b) compound 2 - calcium carbonate, c) compound 3 - silicate ettringite + gibbsite, and d) compound 4 - carbonated ettringite. Analysed surface area was 300×515 μm^2. *contourf* function on *Matlab R2015a* has been used to obtain smoothed representations in the mapping.

to 1, red to brown colours), zones where several compounds were present (green to yellow colours) and zones where spectra have been excluded from the resolution by SSMS (grey).

The main part of these grey zones, blue areas in Fig.4.g, corresponds to spectra with a too low signal to noise ratio. This was due to an insufficient contact between the s-ATR crystal and the sample. This may have arisen from an insufficient planarity of the sample surface, from the sample roughness (including open pores) or from the presence of phases with different mechanical properties. In addition, the possible creep of the sample during long acquisitions should not completely be excluded. Indeed, as summarized in Table 1, the use of a s-ATR crystal requires a very flat and even surface.

Surprisingly, the remaining grey areas, deep and light blue in Fig.4, are located in the middle of closed zones where one compound was highly predominant. Comparing Fig.4 and Fig.8, spectra in these areas were excluded from the analysis for two different reasons:

- absorbance values out of the linearity range of the sensor (Fig.4.h, 6.06% of the database)

A high absorbance, thus a low reflectance ($A \sim 1/R$) implies that there was an insufficient number of photons hitting the sensor during the acquisition. This occurred in zones assigned to granular silica (compound 1). For such absorbent materials, a higher number of scans would permit to reduce the signal to noise ratio; a longer duration of acquisition (per scan) would permit to collect more photons and perform a reliable measurement.

- occurrence of s-ATR artifacts (Fig.4.f, 1.37% of the database)

Spectra assigned to these artifacts were only found right in the middle of zones where carbonate ettringite (compound 4) was predominant. Looking

at one of the spectra excluded from the resolution process by SSMS (Fig.5.a) and at the spectrum corresponding to carbonate ettringite (compound 4, Fig.7.d), the unique band of the first one (at 786 cm^{-1}) matches perfectly with the main band of the latter ($Si-O-Si$ vibration).

However, all the values displayed by the spectrum presented in Fig.5.a are: (i) within the linearity range of the sensor and (ii) above the defined threshold for the signal-to-noise ratio (higher absorbance value > 0.4). This implies that (i) the acquisition parameters (duration and scans number) were adapted and (ii) the local contact between the s-ATR crystal and the material was properly done. Therefore, the relative intensities displayed by the s-ATR artifacts spectrum can be considered as correct: there is no other hidden band in the spectrum (for instance due to a scale effect). Thus, it was considered that this spectrum really contained only one band, which is physically non possible for any material. Finally, it was hard to determine the physical origin of such s-ATR artifacts spectra. One explanation could arise from the occurrence of distortions in spectra acquired in reflection modes due to Mie scattering effect as shown by Bassan et al. [10]. Another comprehensive study on distortion effects occuring in reflection [2] showed that they also may stem from material dispersion.

5.3. Implications of the use of SSMS and s-ATR acquisitions

All in all, a new supervised method has been developed in this work: SSMS combined with ALS. This method permitted to handle a set of 3936 FT-IR spectra to obtain a relevant chemical mapping of compounds present in an ettringite binder.

In this work, a PCA on the hyperspectral database led to the same number of compounds with or without the exclusion of the s-ATR artifacts spectra thanks to SSMS. But, without SSMS, the 4 selected spectra and their related maps had no chemical meaning. Moreover, as already discussed, SSMS can also be used to exclude spectra of no interest for a specific study. In this case, the number of compounds found by a PCA would be changed after SSMS.

In SSMS, thresholds and a correlation coefficient have to be adjusted for each study. Since the operator has to intervene in the process, SSMS requires the operator to have enough hindsight and expertise for his study. However, such a requirement holds true for all statistical methods applied on physical problems.
On the other hand, the main interest of SSMS lies precisely in the fact that it is based on statistical considerations. Thus, SSMS can be used to identify all kinds of artifacts in all kinds of database without any restriction on material nature and characterization mean; more generally, SSMS may be used for all types of problems with few adaptations.

The s-ATR crystal is an interesting set-up for the acquisition of valuable data to map the relative concentrations of compounds in a material. Thanks to this set-up, no relative motion occurred between the s-ATR crystal and the sample all along the acquisition: cross-contamination and successive stamps, experienced in previous studies on materials with similar properties [18], were avoided here. However, the s-ATR crystal left one unique imprint of around 500 μm in diameter in the sample (Fig.2.b). Noteworthy, this mark could

be used to easily find the imaged zone for complementary characterizations (such as SEM-EDX).

In addition, and as discussed in the previous section, some of the acquired spectra presented s-ATR artifacts, which remain unexplained until now.

As highlighted in Table 1, using the s-ATR set-up also requires the sample to have a very low roughness and a very flat surface for the contact between the crystal and the sample to be perfect on a relatively large surface area (around 1 mm^2). This requires a meticulous preparation and polishing of the surface of the sample prior to analysis. On this aspect, one very interesting feature of the SSMS approach is illustrated on the granular composite with ettringite binder in Fig.4: this new method enables to highlight zones of the sample, where the surface state has to be improved (Fig.4.g), as well as zones where the acquisition parameters have to be optimized (Fig.4.h) to map more completely the sample. In other words, SSMS proved to be very efficient to identify issues in specific zones: in addition to permit to obtain relevant chemical maps of a material, SSMS can also be used as a powerful tool to adapt one's method (sample preparation, acquisition parameters) for the specific zones one wants to map.

6. Conclusions

A new supervised method, called SSMS, has been developed to deal with the analysis of hyperspectral data.

Experimental acquisitions may lead to unexpected biased information, which can have different origins, such as the acquisition technique, the environment, the analyzed sample. To obtain relevant information from the database, it is

fundamental to be able to exclude all of these biased data without impacting on the informative ones. This selection is easily carried out with classical filtering methods when the origin and the characteristics of the biased data are known. SSMS is a powerful tool to handle this problem in the general case and exclude biased data when all their sources cannot be predicted.
Moreover, SSMS is a preprocessing method, and thus can be coupled with a reliable physical model to obtain quantitative chemical information. In addition, the combined use of SSMS and of a physical model permits to identify the sources of the biased data. Thus, it becomes possible to improve the experimental method (sample preparation, acquisition parameters...).
All steps involved in SSMS are described and justified in this article. The interest and the gain provided by SSMS is illustrated on the analysis of data acquired with a FT-IR microscope using a disregarded set-up (static-ATR). Relevant chemical maps of a complex inorganic material are obtained coupling FT-IR microscopy data with the Beer Lambert law. Maps are analyzed, sources of biased data are identified thanks to SSMS and discussed.
Finally, such a supervised preprocessing opens a new route to obtain more reliable quantitative information and to improve the experimental protocols.

Funding: This work was supported by the French Ministry of Higher Education and Research and ENS Paris-Saclay.

Acknowledgments: The authors would like to thank R. Pujol (Univ. Lyon, INSA Lyon) for fruitful discussions.

[1] E. Maire, N. Gimenez, V. Sauvant-Moynot, H. Sautereau, X-ray tomography and three-dimensional image analysis of epoxy-glass syntactic foams, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 364 (1838) (2006) 69–88. doi:10.1098/rsta.2005.1691.

[2] T. Bonnal, Development of optical models and non-supervised methods to solve bilinear problems: application to vibrationnal mapping, Ph.D. thesis Univ. Lyon 1 (2018).
URL tel.archives-ouvertes.fr/tel-01818822/document

[3] J. Xiao, G. Foray, Masenelli-Varlot, Analysis of liquid suspensions using Scanning Electron Microscopy in transmission: estimation of the water film thickness using Monte Carlo simulations, Journal of Microscopy.

[4] A. Bannerman, R. L. Williams, S. C. Cox, L. M. Grover, Visualising phase change in a brushite-based calcium phosphate ceramic., Scientific reports 6 (August 2015) (2016) 32671. doi:10.1038/srep32671.

[5] F. Puig-Castellví, I. Alfonso, B. Piña, R. Tauler, 1H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis, Scientific Reports 6 (1) (2016) 30982. doi:10.1038/srep30982.

[6] S. Yamaguchi, Y. Fukushi, O. Kubota, T. Itsuji, T. Ouchi, S. Yamamoto, Brain tumor imaging of rat fresh tissue using terahertz spectroscopy, Scientific Reports 6 (1) (2016) 30124. doi:10.1038/srep30124.

[7] M. Offroy, Y. Roggo, P. Milanfar, L. Duponchel, Infrared chemical imaging: Spatial resolution evaluation and super-resolution concept, Analytica Chimica Acta 674 (2) (2010) 220–226. doi:10.1016/j.aca.2010.06.025.

[8] M. J. Nasse, M. J. Walsh, E. C. Mattson, R. Reininger, A. Kajdacsy-Balla, V. Macias, R. Bhargava, C. J. Hirschmugl, High-resolution Fourier-transform infrared chemical imaging with multiple synchrotron beams., Nature methods 8 (5) (2011) 413–6. doi:10.1038/nmeth.1585.

[9] S. Amarie, P. Zaslansky, Y. Kajihara, E. Griesshaber, W. W. Schmahl, F. Keilmann, Nano-FTIR chemical mapping of minerals in biological materials, Beilstein Journal of Nanotechnology 3 (1) (2012) 312–323. doi:10.3762/bjnano.3.35.

[10] P. Bassan, H. J. Byrne, J. Lee, F. Bonnier, C. Clarke, P. Dumas, E. Gazi, M. D. Brown, N. W. Clarke, P. Gardner, Reflection contributions to the dispersion artefact in FTIR spectra of single biological cells., The Analyst 134 (6) (2009) 1171–1175. doi:10.1039/b821349f.

[11] S. A. Macdonald, C. R. Schardt, D. J. Masiello, J. H. Simmons, Dispersion analysis of FTIR reection measurements in silicate glasses, Journal of Non-Crystalline Solids 275 (2000) 72–82. doi:10.1016/S0022-3093(00)00121-6.

[12] P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, P. Gardner, Resonant Mie scattering in infrared spectroscopy of biological materials–understanding the 'dispersion artefact'., The Analyst 134 (8) (2009) 1586–1593. doi:10.1039/b904808a.

[13] M. Miljković, B. Bird, M. Diem, Line shape distortion effects in infrared spectroscopy., The Analyst 137 (17) (2012) 3954–64. doi:10.1039/c2an35582e.

[14] K. L. A. Chan, S. G. Kazarian, Attenuated total reflection Fourier-transform infrared (ATR-FTIR) imaging of tissues and live cells, Chem. Soc. Rev. 45 (2016) 1850–1864. doi:10.1039/C5CS00515A.

[15] K. L. A. Chan, S. G. Kazarian, New opportunities in micro-and macro-attenuated total reflection infrared spectroscopic imaging: spatial resolution and sampling versatility, Applied spectroscopy 57 (4) (2003) 381–389.
URL http://www.opticsinfobase.org/abstract.cfm?id=118140

[16] M. Offroy, M. Moreau, S. Sobanska, P. Milanfar, L. Duponchel, Pushing back the limits of Raman imaging by coupling super-resolution and chemometrics for aerosols characterization, Scientific Reports 5 (1) (2015) 12303. doi:10.1038/srep12303.
URL http://www.nature.com/articles/srep12303

[17] S. G. Kazarian, K. L. A. Chan, Micro- and Macro- Attenuated Total Reflection Fourier Transform Infrared Spectroscopic Imaging, Applied Spectroscopy 64 (5) (2010) 135A–152A.

[18] T. Bonnal, G. Foray, E. Prud'homme, S. Tadier, Towards chemical imaging: Fourier transform infrared mapping on organo-mineral materials, European Journal of Environmental and Civil Engineering 8189 (April) (2017) 1–11. doi:10.1080/19648189.2017.1304278.

[19] D. A. Burns, E. W. Ciurczak, Handbook of near-infrared analysis, 3rd ed. (2009). arXiv:arXiv:1011.1669v3, doi:10.1021/ja015320c.

[20] A. De Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Analytical Methods 6 (14) (2014) 4964. doi:10.1039/c4ay00571f.

[21] M. Michel, J. Ambroise, P. Hamelin, Développement de composites à matrice minérale et à renfort textile, in: Treizième édition des Journées Scientiques du (RF)2B, 2012, pp. 57–66.

[22] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Analytical Chemistry 63 (14) (1991) 1425–1432. doi:10.1021/ac00014a016.

[23] W. Windig, C. Heckler, F. Agblevor, R. Evans, Self-modeling mixture analysis of categorized pyrolysis mass spectral data with the SIMPLISMA approach, Chemometrics and Intelligent Laboratory Systems 14 (1-3) (1992) 195–207. doi:10.1016/0169-7439(92)80104-C.

[24] F. Sánchez, D. Massart, Application of SIMPLISMA for the assessment of peak purity in liquid chromatography with diode array detection, Analytica Chimica Acta 298 (3) (1994) 331–339. doi:10.1016/0003-2670(94)00283-5.

[25] A. C. Rencher, Principal Component Analysis, in: Methods of Multivariate Analysis, Second Edition, 2nd Edition, A JOHN WILEY & SONS, INC. PUBLICATION, New York, 2002, Ch. 12, p. 727. doi:10.1080/07408170500232784.

[26] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems 2 (1-3) (1987) 37–52. arXiv:arXiv:1011.1669v3, doi:10.1016/0169-7439(87)80084-9.

[27] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, Journal of Chemometrics 15 (8) (2001) 627–646. doi:10.1002/cem.654.

[28] A. De Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Critical Reviews in Analytical Chemistry 36 (3-4) (2006) 163–176. arXiv:arXiv:1011.1669v3, doi:10.1080/10408340600970005.

[29] A. De Juan, M. Maeder, T. Hancewicz, R. Tauler, Local rank analysis for exploratory spectroscopic image analysis. Fixed Size Image Window-Evolving Factor Analysis, Chemometrics and Intelligent Laboratory Systems 77 (2005) 64–74. doi:10.1016/j.chemolab.2004.11.006.

[30] A. De Juan, M. Maeder, T. Hancewicz, R. Tauler, Use of local rank-based spatial information for resolution of spectroscopic images, Journal of Chemometrics 22 (5) (2008) 291–298. doi:10.1002/cem.1099.

[31] M. T. Tognonvi, S. Rossignol, J. P. Bonnet, Effect of alkali cation on irreversible gel formation in basic medium, Journal of Non-Crystalline Solids 357 (1) (2011) 43–49. doi:10.1016/j.jnoncrysol.2010.10.003.

[32] J. D. Rodriguez-Blanco, S. Shaw, L. G. Benning, The kinetics and mech-

anisms of amorphous calcium carbonate (ACC) crystallization to calcite, via vaterite, Nanoscale 3 (1) (2011) 265–271. doi:10.1039/c0nr00589d.

[33] M. Muroya, Infrared Spectra of SiO2 Coating Films Prepared from Various Aged Silica Hydrosols, Bulletin of the Chemical Society of Japan 64 (1991) 1019–1021.

[34] H. V. Olphen, J. J. Fripiat, Data Handbook for Clay Materials and Other Nonmetallic Minerals, 1979.

[35] W. Haynes, CRC Handbook of Chemistry and Physics, 95th Edition, 2014-2015, Vol. 54, 2014. doi:10.1136/oem.53.7.504.

[36] R. Frost, J. Kloprogge, S. Russell, J. Szetu, Dehydroxylation and structure of alumina gels prepared from trisecbutoxyaluminium, Thermochimica Acta 329 (1) (1999) 47–56. doi:10.1016/S0040-6031(98)00663-7.

Table 2: Identification of the 4 compounds using FTIR bands described in the literature.

Wavenumber[a]	Bond	Reference
	Compound 1	
779	$Si-O-Si$	[31]
1026	$Si-O-Si$	[31]
1150	$Si-O-Si$	[31]
	Compound 2	
740	CO_3^{2-}	[32]
872	CO_3^{2-}	[32]
1400	CO_3^{2-}	[32]
$3070-3730$	OH	[33]
	Compound 3	
1088	SO_4^{2-}	[34]
1242	$C-O$	[35]
1450	CO_3^{2-}	[32]
1660	H_2O	[33]
3630	OH[b]	[36]
$3070-3730$	OH	[33]
	Compound 4	
786	$Si-O-Si$	[31]
1100	$Si-O-Si$	[31]
1088	SO_4^{2-}	[34]
$1350-1450$	CO_3^{2-}	[32]
$3070-3730$	OH	[33]

[a] Wavenumber in cm^{-1}; [b] $Al(OH)_6$ octahedral environment