Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

SELECTION OF ESSENTIAL SPECTRA TO IMPROVE THE MULTIVARIATE CURVE RESOLUTION OF MINOR COMPOUNDS IN COMPLEX PHARMACEUTICAL FORMULATIONS

Laureen Coic[a], Pierre-Yves Sacré[a], Amandine Dispas[a,b], Charlotte De Bleye[a], Marianne Fillet[b], Cyril Ruckebusch[c], Philippe Hubert[a], Éric Ziemons[a]

[a] *University of Liege (ULiege), CIRM, Vibra-Santé Hub, Laboratory of Pharmaceutical Analytical Chemistry, Avenue Hippocrate 15, 4000, Liege, Belgium*

[b] *University of Liege (ULiege), CIRM, MaS-Santé Hub, Laboratory for the Analysis of Medicines, Avenue Hippocrate 15, 4000, Liege, Belgium*

[c] *Univ. Lille, CNRS, UMR 8516 - LASIRE - Laboratory of Advanced Spectroscopy, Interactions, Reactivity, and Environment, Cité Scientifique, Bâtiment C5, 59000, Lille, France*

Keywords:

Raman; FT-IR; Hyperspectral imaging; MCR-ALS; Data reduction; Essential spectral pixels (ESPs); Falsified medicines

Abstract

Multivariate curve resolution unmixing of hyperspectral imaging data can be challenging when low sources of variance are present in complex samples, as for minor (low-concentrated) chemical compounds in pharmaceutical formulations. In this work, it was shown how the reduction of hyperspectral imaging data matrices through the selection of essential spectra can be crucial for the analysis of complex unknown pharmaceutical formulation applying Multivariate Curve Resolution e Alternating Least Squares (MCRALS). Results were obtained on simulated datasets and on real FT-IR and Raman hyperspectral images of both genuine and falsified tablets. When simulating the presence of minor compounds, different situations were investigated considering the presence of single pixels of pure composition as well as binary and ternary mixtures. The comparison of the results obtained applying MCR-ALS on the reduced data matrices with those obtained on the full matrices revealed unequivocal: more accurate decomposition could be achieved when only essential spectra were analyzed. Indeed, when analyzing the full dataset, MCR-ALS failed resolving minor compounds even though pure spectra were provided as initial estimation, as shown for Raman hyperspectral imaging data obtained on a medicine sample containing 7 chemical compounds. In contrast, when considering the reduced dataset, all minor contributions (down to 1 pixel over 17,956) were successfully unmixed. The same conclusion could be drawn from the results obtained analysing FT-IR hyperspectral imaging data of a falsified medicine.

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

1. Introduction

The advantages of hyperspectral imaging techniques have no longer to be introduced. Indeed, for a couple of decades, these spectroscopic techniques have known a growing interest essentially because they provide both organic and inorganic information. Hyperspectral imaging has been used for different applications in the agri-food [[1], [2], [3]], biological [[4], [5], [6]] and pharmaceutical fields [[7], [8], [9], [10], [11]]. Interestingly, Raman hyperspectral imaging of pharmaceutical tablets has been widely used for quality control (QC) purpose, which requires the most exhaustive and best quality results, and has now been included in the general chapters of the European Pharmacopeia [12]. However, some critical issues remain when analyzing hyperspectral imaging data since pharmaceutical formulations can be complex matrices, due to their chemical composition and physical properties/form (granulometry, galenic form, …) [13]. Sample size can also be a critical aspect for microimaging [9].

Unmixing of vibrational hyperspectral imaging data of pharmaceutical formulations (spectral identification and spatial distribution of pure chemical compounds) remains a difficult task because of the presence of many different sources of spectral variation. Indeed, depending on the sample composition, abundance and spectral signature of the different chemical compounds, their individual contribution to the global variance may be very different. Minor compounds may have a very small contribution to the total signal variation and can thus be difficult to extract by methods based on the minimization of the sum of squared error, as for factor analysis techniques [14]. Dedicated strategies are then required [[15], [16], [17], [18]]. In particular, complete elucidation of pharmaceutical tablets could be achieved from the analysis of Raman hyperspectral imaging data by coupling iterative approaches and database matching [16,19]. Multiset Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) [20,21] has also revealed to be a useful alternative to extract the contributions of minor compounds. However, these different strategies share the same limitations. First, they can only be applied when both the data and the database have a reasonable size. Second, they cannot be fully applied to the analysis of falsified medicines, i.e. samples of unknown composition and for which unusual raw materials may be present in very low abundance [9].

In recent work, full chemical elucidation of complex pharmaceutical formulations could be achieved by using a pixel-based data analysis approach coupled with databased matching [10]. For this purpose, essential spectral pixels (ESPs) were selected first and the method was applied on the resulting reduced (compressed) data matrix [22,23]. It was possible to detect impurities with concentrations down to 0.1% (w/w) even for very large data matrices (reducing the computational time by two orders of magnitude) [10]. However, the proposed approach requires the presence of selective (pure) pixels, i.e., pixels for which only one chemical compound is observed.

In contrast, for the analysis of samples of unknown composition, MCR-ALS [[24], [25], [26]] algorithm can be used, providing both distribution maps and pure spectra of all chemical compounds present in complex mixtures. This algorithm solves the matrix factorization problem by using a two-step iterative optimization problem. At each step, constraints such as non-negativity or image processing algorithm can be applied on both spectral and spatial information [22,25,[27], [28], [29], [30], [31], [32]].

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

In this work, we propose to overcome the aforementioned issues related to the presence of minor compounds, in situations where very large vibrational hyperspectral imaging datasets are dealt with, by coupling ESP selection with MCR-ALS unmixing. To validate this approach, FT-IR and Raman hyperspectral imaging data were first simulated dealing with various situations corresponding to complex pharmaceutical formulations characterized by the presence of very minor contributions, considering both the presence and the absence of pure pixels. In a second step, the methodology was applied on real images acquired by analyzing market pharmaceutical tablets and suspected falsified medicines. Overall, the results obtained on the ESP-reduced data matrices were more accurate than those obtained on the full matrices.

2. Materials and methods

2.1. SIMULATED DATASETS

A first simulated hyperspectral imaging dataset was built (dataset 1, Fig. 1A) for which the FT-IR spectra of 8 pure compounds were taken from our in-house database containing 34 compounds referenced compounds. Dataset 1 was built with in mind to mimic a realistic pharmaceutical formulation contaminated by unknown compounds only present in very minor quantities (see Fig. 1A–B). For this purpose, pure pixels of lactose monohydrate (48% of the pixels), sodium bicarbonate (25%), acetaminophen (10%), citric acid (9.3%), ascorbic acid (4.5%) and starch (2.9%) were first distributed. In a second step, talc and ibuprofen contributions were added in very minor quantities, corresponding to 0.012% and 0.006% of the pixels, respectively, and resulting in binary- and ternary-mixed pixels. The map size of dataset 1 corresponds to 134 × 134 pixels and the spectral range consists of 1600 wavenumbers. The spatial distribution of each raw material is provided in Figure S1 of the supporting information. The full spectral data are provided in Figure S2A and it was checked that the influence of minor compounds on the singular value decomposition (SVD) of the data matrix was not significant. For this purpose, SVD were calculated without considering the addition of talc and ibuprofen and compared to the ones obtained on the full dataset. As seen from the values in Table S1, no significant change can be noticed. The second dataset was simulated to benchmark our approach using a semi-artificial Raman hyperspectral imaging example (dataset 2, Fig. 1C). We started from a real pharmaceutical Raman image corresponding to a map of 1000 × 1000 pixels and 1600 Raman shifts measured on a ternary mixture composed of lactose monohydrate, piroxicam polymorph α2 and piroxicam polymorph β, where the β form was present at trace level (0.1% w/w) (see Ref. [10]). We selected a patch of 134 × 134 pixels and artificially added the contributions of four compounds, namely microcrystalline cellulose, sodium croscarmellose, magnesium stearate and glucose (see Fig. 1D, spectra were taken from the in-house database.). Binary and ternary mixture pixels were introduced as follows: two pixels were made of magnesium stearate and piroxicam α2 mixtures, in 0.8 and 0.2 proportions, and 10 pixels were composed of mixtures of microcrystalline cellulose, sodium croscarmellose and glucose (to mimic contaminant), in random proportions. It should be noted that magnesium stearate and glucose cannot be found as pure compound in the data set (no pure spectral pixel). The full spectral data and the individual distribution maps of each compound are provided in

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

Figures S4A and S5. To recap, the spectral pixels in dataset 2 consist of lactose monohydrate (59% of the pixels, no pure spectra), microcrystalline cellulose (25%, pure), sodium croscarmellose (5%, pure), piroxicam α2 form (9%, no pure spectra), piroxicam β form (2%, no pure spectra), magnesium stearate (0.01%, only observed in binary mixtures) and, finally, glucose (0.005%, only observed in ternary mixtures).

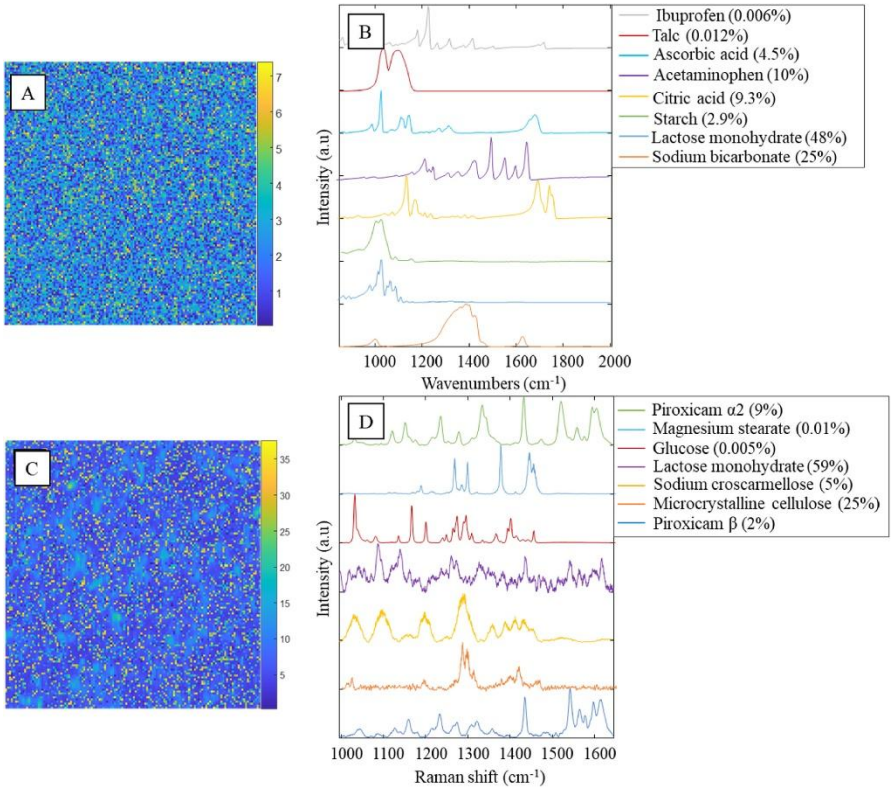

Fig. 1. Representation of the simulated datasets. A) Mean image for dataset 1. B) Pure FT-IR spectra and their distribution over pixels (e.g., in parenthesis, 1% means that the compound can be found in 1% of the pixels). C) Mean image for dataset 2. D) Pure Raman spectra their distribution over pixels.

2.2. REAL PHARMACEUTICAL SAMPLES

2.2.1. FT-IR HYPERSPECTRAL IMAGING

A pharmaceutical tablet was collected from the official supply chain and analyzed by FT-IR hyperspectral imaging (dataset 3). The FT-IR image was acquired with a Cary 620/670 Agilent series microscope (Agilent Technologies, Resolution Pro acquisition software) equipped with a 15x infrared objective, with a numerical aperture (NA) of 0.62 and a FPA detector (64 × 64 pixels) providing a pixel size of 5.5 µm.

Data acquisition was performed in reflection mode, with a resolution of 8 cm−1 over the spectral range 850–2000 cm−1 and 16 co-added scans. The FT-IR spectra are represented in Figure S7A. Only a part of the 2 cm diameter medicine was analyzed (12% of the initial surface), resulting in a map size of 448

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

× 448 pixels and a spectral range covering 1600 wavenumbers. The composition of the medicine tablet is given in Table S2. The analysis of dataset 3 can be considered challenging for two main reasons: i) FT-IR spectra are noisy and show strong spectral overlap in the range considered, and ii) the size of the dataset was quite large even though it represents only a fraction of the initial sample size.

2.2.2. RAMAN HYPERSPECTRAL IMAGING

A falsified chloroquine medicine tablet seized during the COVID-19 pandemic was analyzed using Raman hyperspectral imaging [8] (dataset 4). Image acquisition was performed with a Labram HR Evolution (Horiba scientific) equipped with an EMCCD detector (1600 × 200 pixels, Andor Technology Ltd.), a 300 mm/grating, a confocal pinhole of 20 µm, a Leica 50x Fluotar LWD objective and a 785 nm laser with a power of 45 mW at sample (XTRA II single frequency diode laser, Toptica Photonics AG). The spectra were collected with the LabSpec 6 software (Horiba Scientific).

The tablet was glued on a microscope slide and its surface was milled using a Leica EM Rapid milling system equipped with a tungsten carbide miller (Leica Microsystems GmbH) before Raman mapping. A 150 × 150 mapping was done, generating a data cube of 150 × 150 × 1600 (from 550 to 1850 cm−1). The Duoscan mode was used to obtain a pixel size of 50 µm. The spectra are provided in Figure S10. As can be seen in Figure S10A, the raw Raman data are strongly affected by fluorescence and even after preprocessing (see section 2.3), the signal to noise ratio remained quite low (see Figure S10B).

2.3. DATA ANALYSIS

2.3.1. PREPROCESSING

For FT-IR hyperspectral imaging, the first step was to perform a baseline correction by an Asymmetric Least Squares approach [33] (λ = 3.104, p = 1.10−5). Then a Kubelka Munk transformation [34] was applied on the reflectance spectra to minimize the diffuse reflection effect.

For Raman hyperspectral imaging, the first step was applying a Savitzky & Golay smoothing [35] (window size: 15). The baseline was finally removed by an automatic Whittaker filter (λ = 3.104, p = 1.10−5).

2.3.2. SELECTION OF ESSENTIAL SPECTRA

To select the most relevant spectral information for linear unmixing, i.e. the most linearly dissimilar spectra, we applied convex hull calculation in the PCA score space of the initial data matrix [22]. Geometrically, the convex hull of a planar set of points may be described as a convex polygon whose vertices are actual data points and that contains the whole set of points. Essential spectra correspond to points that can be found on the envelope of score cloud and, specifically, at its vertices, where the purest information is found [29]. For this purpose, a SVD was first applied on the unfolded hyperspectral data matrix. Scores were then reconstructed by considering the most meaningful principal components (default value was set to 5). The selection of the most relevant spectra was then performed applying the essential convex hull calculation as described in Ref. [17]. The resulting ESP-reduced is highly compressed and contains the most linearly relevant spectral features for MCR-ALS analysis. It has to be noticed that convex hull calculation is not based on variance but on a geometric

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

approach. This means that the dimensionality of the PCA subspace in which convex hull calculation is performed does not have to match the one expected for the linear unmixing problem. The main step of the procedure is provided in the equations below:

1) Unfold the spectral imaging data cube $\underline{\boldsymbol{D}}_{I\times J\times K}$ into a two-dimensional data matrix $\boldsymbol{D}_{I.J\times K}$;

2) Calculate the singular value decomposition (SVD) of $\boldsymbol{X}_{I.J\times K}$ as in equation (1):

$$\boldsymbol{D}_{I.J\times K} = \boldsymbol{U}_{I.J\times n}\boldsymbol{\Sigma}_{n\times n}\boldsymbol{V}^{T}_{n\times K} + \widetilde{\boldsymbol{E}}_{I.J\times K} \quad (1)$$

where $\boldsymbol{U}_{I.J\times n}$ and $\boldsymbol{V}^{T}_{n\times K}$ are the left and right singular vectors of $\boldsymbol{D}_{I.J\times K}$, $\boldsymbol{\Sigma}_{n\times n}$ is the diagonal matrix of singular values, n is the number of factors, and $\widetilde{\boldsymbol{E}}_{I.J\times K}$ is the residual matrix. The scores can be calculated as follows:

$$\boldsymbol{X}_{I.J\times n} = \boldsymbol{U}_{I.J\times n}\boldsymbol{\Sigma}_{n\times n} \quad (2)$$

3) Apply CH calculation on the scores $\boldsymbol{X}_{I.J\times n}$ to select ESPs as in equation (3).

$$conv\ \boldsymbol{X}_{I.J\times n} = \left\{ (\boldsymbol{x}_{i.j.n} \in \boldsymbol{X}_{I.J\times n}) \left| \begin{array}{l} \sum_{(i.j)}^{P} \alpha_{(i.j)}\boldsymbol{x}_{i.j.n};\ \alpha(i.j) \geq 0 \\ and\ \sum_{(i.j)}^{P} \alpha(i.j) = 1 \end{array} \right. \right\} \quad (3)$$

4) where ($X_{(I.J\times n)}$) is now normalized with unit first singular value, for the sake of simplicity the notation is left unchanged), and P denotes the number of ESPs found. The convex hull of $\boldsymbol{X}_{I.J\times n}$, denoted $conv\ \boldsymbol{X}_{I.J\times n}$,. is the set of all convex combinations of points in $\boldsymbol{X}_{I.J\times n}$.

2.3.3. MULTIVARIATE CURVE RESOLUTION – ALTERNATING LEAST SQUARES

MCR-ALS [36,37] can be used to solve Equation (1) following the Beer Lambert law. For all the computation, the non-negativity constraint was applied on both concentration and spectral matrix. To evaluate the quality of the fit of the MCR-ALS model, the lack of fit (LOF) was calculated as in Equation (2).

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E}\,(1)$$

$$LOF = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(d_{i,j} - \hat{d}_{i,j}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}(d_{i,j})^2}} \times 100\,(2)\text{Where:}$$

- corresponds to the original data **D** or to the ESP
- is the calculated d by the MCR model
- LOF is the lack-of-fit in %

When applied on simulated data (datasets 1 and 2), MCR-ALS was initialized with pure spectra. In contrast, for the analysis of real samples (datasets 3 and 4), SVD was applied to evaluate the number

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

of contributions and SIMPLe-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) [38] was used to extract initial estimation and to find with.

For each dataset, MCR-ALS models were built on both the reduced and the full data matrices. The correlation coefficient (r) [10] in Equation (3) was calculated in order to evaluate the match between pure component spectra and database spectra.

$$rs, Db = \frac{cov(S,Db)}{\sigma_S \sigma_{Db}} \quad (3)$$

Where:

- cov(S,Db)=covariance of variables S (pure component) and Db (database spectrum)
- σS=standard deviation of S
- σRef=standard deviation of Ref

Components for which r was found above 0.60 were considered relevant and the corresponding spectra were gathered in a matrix called S of dimensions n x 1600. In some cases, several pure spectra were assigned to the same compound (r slightly different, but the corresponding spectra are visually the same), their contributions were summed, and the resulting spectra was implemented in S. To obtain the concentration distribution maps of individual compound, the full images of the components are generated in a non-negative least-squares step.

2.4. SOFTWARE

The proposed strategy was developed on a workstation with Intel® Core™ i7-7820X CPU @ 3.60 GHz, 8 cores with 128 GB of RAM. For some key steps, the Matlab Parallel Computing Toolbox™ was used to improve the speed of algorithms. All computations were carried out with Matlab R2019b (The Mathworks) with the PLS Toolbox (version 8.6.2, Eigenvector Research Inc) and in-house routines.

3. Results and discussion

3.1. SIMULATED PHARMACEUTICAL MIXTURES

The results obtained on the simulated FT-IR hyperspectral imaging data (dataset 1) are reported first. This dataset consists of a realistic eight-component system, as described in section 2.1 (see also, see Figures S1 and S2A). The pure spectra corresponding to the results obtained by applying MCR-ALS on the full dataset are provided in Fig. 2B. These results were obtained taking the reference spectra (used for the simulation, see Fig. 2A) as initial estimation. The computed LOF was 7.61%. To get better insight into the reliability of these results, the pure spectra obtained from the MCR-ALS analysis were matched with the in-house database and the spectra in Fig. 2B were colored according to the identification found. It appears clearly that the pure spectra corresponding to talc and ibuprofen (very minor compounds, marked red and grey in Fig. 2, respectively), could not be "resolved" by the eight-

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

component MCR-ALS model, even though pure spectra were provided as initial estimate. This can be explained by the fact that, in absence of additional constraints, the contributions of talc and ibuprofen signals to the total residual error are too low to drive the optimization to the expected results. The goal of this work was to consider and overcome this pitfall. Moreover, the very similar spectra of lactose monohydrate (marked dark blue) and sodium bicarbonate (orange) could not be resolved either. In conclusion, only 4 out of the 8 compounds could be extracted applying MCR-ALS to the full dataset, i.e., lactose monohydrate, citric acid, acetaminophen, and starch.

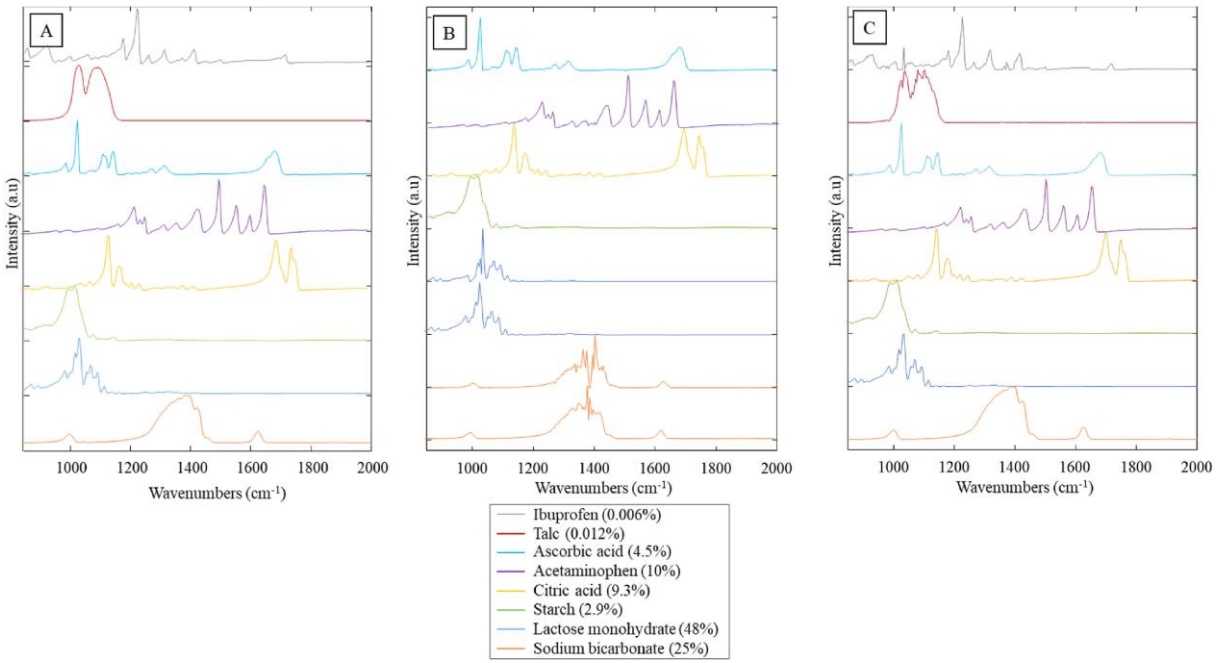

Fig. 2. FT-IR spectra of the 8 compounds used for the simulation of dataset 1. A) Reference spectra. B) Pure spectra obtained from the MCR-ALS analysis of the full dataset (17,956 spectra). C) Pure spectra obtained from the MCR-ALS analysis of the ESP-reduced dataset (151 spectra).

In Fig. 2C the results obtained applying MCR-ALS on the reduced dataset, consisting of 151 spectra (ESPs) selected by applying convex hull calculation (see section 2.4.2), are provided. These spectra are shown in Figure S2B of the supporting information. Once again, reference spectra were used as initial estimation for the MCR-ALS analysis. The computed LOF was 7.41% which is similar to the one obtained with the full dataset. However, as it can be clearly seen from Fig. 2C, the 8 expected components could now be resolved. In our opinion, this example illustrates the striking effect of working with the ESP-reduced dataset in order to extract the contributions of very minor compounds. In Figure S3, the component distribution maps associated to these pure spectra are provided (see section 2.3.3) and, as can be seen, they also match the simulated ones very-well.

The results obtained on a semi-artificial Raman hyperspectral imaging dataset containing 17,956 spectra (dataset 2, see Figures S4A and S5) are now reported. The procedure applied was the same as for dataset 1 and the results obtained applying a seven-component MCR-ALS on the full dataset are provided in Fig. 3B (only the spectra are shown). The computed LOF was 7.32%. When comparing the spectra in Fig. 3A and B, only 4 compounds could be identified, i.e piroxicam α2, piroxicam β,

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

microcrystalline cellulose and sodium croscarmellose. The 3 remaining spectral signatures (spectra in dark grey, Fig. 3B) could not be assigned. It should be recall here that dataset 2 was built starting from measured data for which additional sources of variance exist. Applying convex hull calculation on dataset 2; 151 essential spectra could be identified, as shown in Figure S4B. By comparison to the full dataset (Figure S4A), it is striking that the principal variability was kept while most of the spectral redundancy and uncontrolled variation could be removed. MCR-ALS was then applied on the ESP-reduced dataset and the spectra obtained from a seven-components model are shown in Fig. 3C (the computed LOF was 4.1%). Once again, initial estimates were provided by using the pure spectra (Fig. 3A). Although no pure pixel was present for those compounds, the spectral signatures of glucose and magnesium stearate could now be resolved and identified. In addition, it is interesting to note that lactose monohydrate could this time be clearly resolved (it could not be when working with the full dataset, even though this compound is present in 59% of all pixels). Still, it should also be noted that some characteristic spectral features of lactose monohydrate can be found in the dark grey spectra in Fig. 3B. Overall, working with ESP-reduced data can here again be a clear asset, and similar distribution maps than the reference ones could be obtained (see Figures S5 and S6).

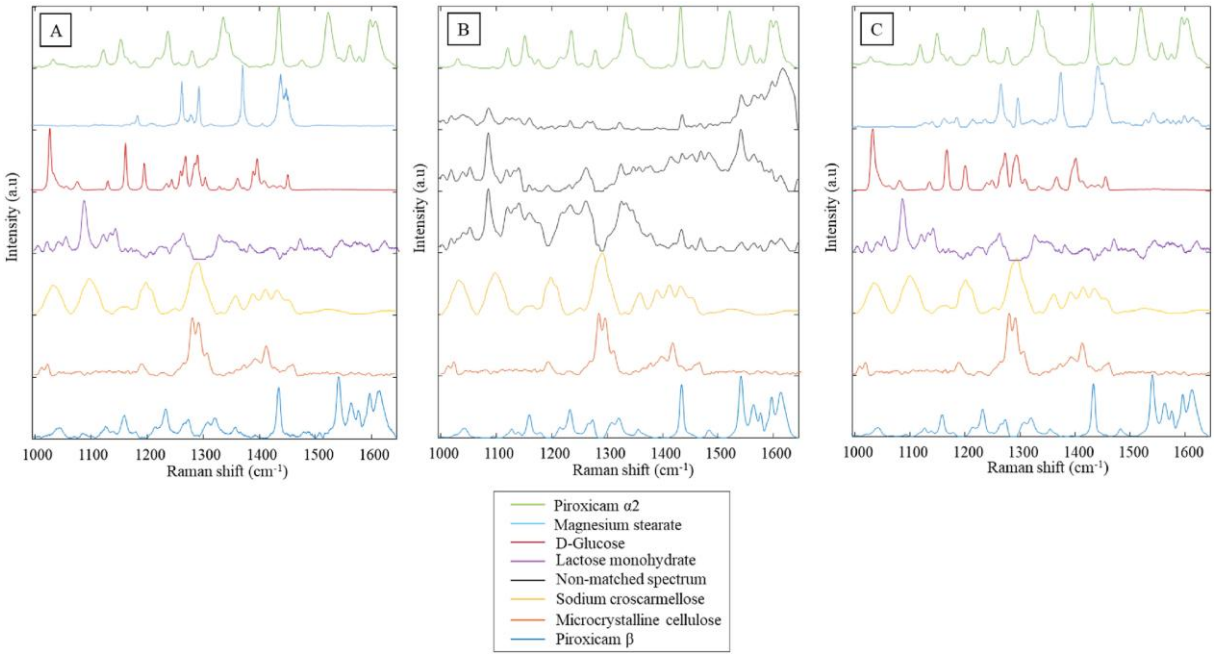

Fig. 3. Raman spectra of the 7 compounds used for the simulation of dataset 2. A) Reference spectra. B) Pure spectra obtained from the MCR-ALS analysis of the full dataset. C) Pure spectra obtained from the MCR-ALS analysis of the ESP- reduced dataset.

3.2. APPLICATION OF THE STRATEGY ON REAL SAMPLES

3.2.1. GENUINE PHARMACEUTICAL SAMPLE (FT-IR HYPERSPECTRAL DATA)

A pharmaceutical tablet obtained from the official supply chain was analyzed by FT-IR hyperspectral imaging (dataset 3, see Figure S7A). The resolution was very challenging since the formulation was very

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

complex, with nine chemical compounds found in different proportions (see Table S2 and Figure S7A) and one main source of spectral variation corresponding to sodium bicarbonate (effervescent tablet). In addition, the penetration depth of the IR signal being only a few microns, it might be difficult to detect all the compounds. The resolution of this dataset was tricky and it was decided to start by considering a nine PCA components. This resulted in the selection of 17,139 essential spectra were selected out of the 200,704 spectra constituting the full data (see Figures S7 and S8). We then built a nine-component MCR-ALS model. The results obtained on the full dataset (LOF 6.51%) are reported in Fig. 4A. SIMPLISMA was used for initial estimation and the presence of 4 compounds could here be detected, acetaminophen (marked green in Fig. 4), citric acid (red), acetylsalicylic acid (grey) and sodium bicarbonate (blue) whose spectral signature could be identified in 4 components. The spectra corresponding to the 2 remaining MCR components (black) correspond to mixed spectral signatures. Moreover, citric acid was found (red) and acetaminophen (green). In a second step, MCR-ALS was applied to the ESP-reduced dataset and the results are shown in Fig. 4B (LOF 8.92%). The noticeable difference with respect to the previous situation is that ascorbic acid, which is present in high proportion (300 mg), can now be identified (second component, yellow).

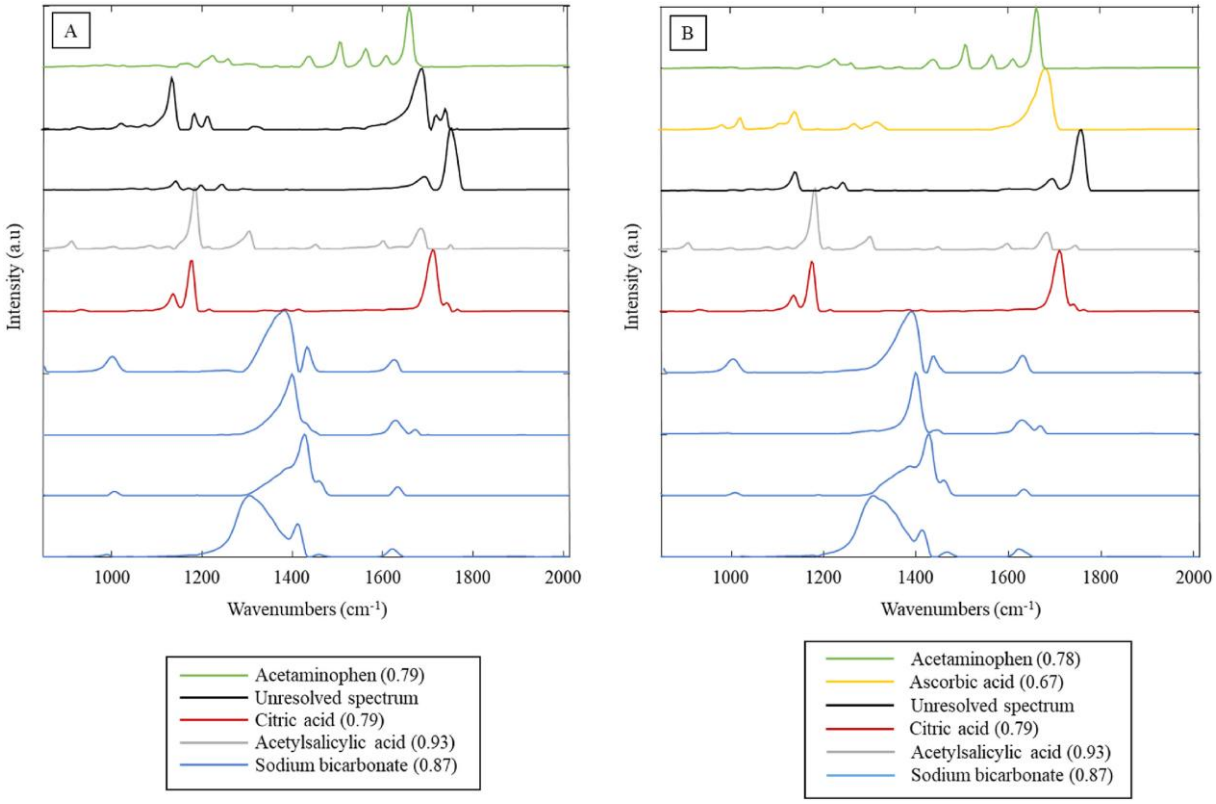

Fig. 4. MCR resolution of data set 3. A) Pure spectra obtained from the MCR-ALS resolution on the full dataset. B) Pure spectra obtained from the MCR-ALS resolution for the ESP-reduced matrix.

Sodium bicarbonate could be identified in 4 of the MCR components extracted. This can be explained by the fact that the sodium bicarbonate participates a lot to the variance of data, requiring to add supplementary components to find the minor compounds. For the sake of interpretation, and as a way to provide the concentration distribution maps (as described in section 2.3.3) corresponding to the

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

different compounds identified, these 4 spectral signatures were averaged. The results are provided in Fig. 5. Overall, we could identify 5 out of the 9 compounds present in the tablet. This result was the best we could achieve and was found satisfying considering the limited spectral resolution and the fact that only a small region of the sample was analyzed.

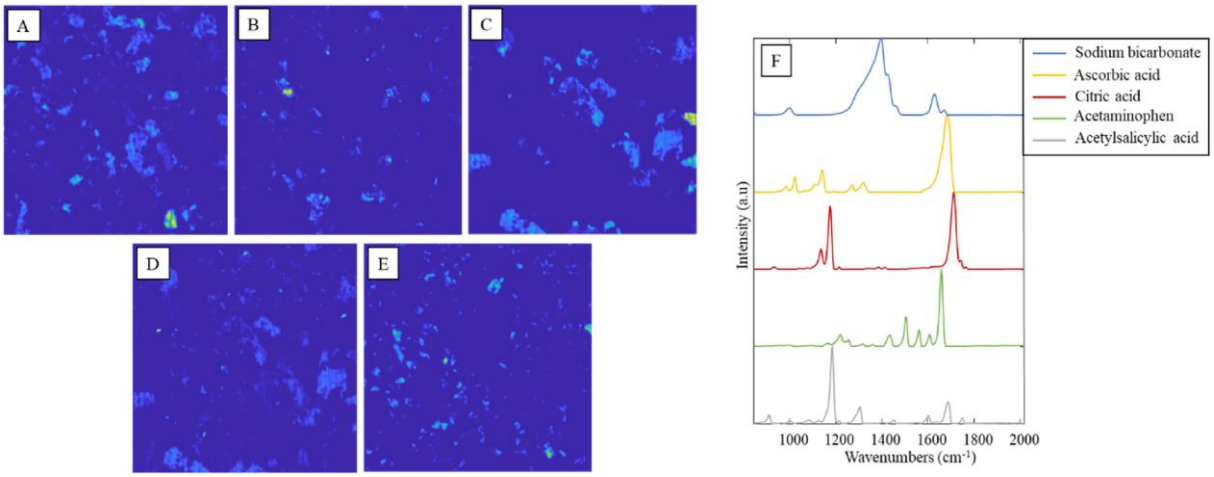

Fig. 5. MCR-ALS results of FT-IR dataset. Concentration distribution maps of A) acetylsalicylic acid, B) acetaminophen, C) citric acid, D) ascorbic acid and E) sodium bicarbonate. F) Corresponding pure spectra.

3.2.2. FALSIFIED MEDICINES (RAMAN HYPERSPECTRAL DATA)

The last example dealt with the analysis of a falsified chloroquine tablet seized during the COVID-19 pandemic [8] (dataset 4, Figure S10). It was not only the most realistic but also the most challenging situation because of the relatively poor quality of the Raman signal. This can be explained by the fact that the data acquisition procedure can hardly be optimal for the analysis of falsified medicines, which are poor quality medicines containing impurities and dusts, requiring low laser power translating into very low signal to noise ratio.

When analyzing a falsified tablet, the analyst does not have a priori information regarding the number of chemical compounds and SVD is the only insight one can get into data complexity. However, as explained before, in presence of minor contributions and impurities, interpretation of the obtained results can be very misleading. We applied SVD analysis on both the full (22,500 spectra) and ESP-resolved (95 spectra) data matrices. Results are provided in Figure S11 for the sake of comparison. From the values obtained on the reduced data, 9 components were considered here. The results obtained applying MCR-ALS on the full dataset are shown in Fig. 6A (LOF 15.2%). As can be seen, most of the extracted pure spectra could not be identified. Only starch (marked dark blue), calcium phosphate (light blue) and metronidazole (yellow) could be found. The results obtained applying MCR-ALS on the ESP-reduced dataset are shown in Fig. 6B (LOF 13.4%). In this case, 6 compounds could be identified, namely metronidazole (marked yellow), acetaminophen (dark red), magnesium stearate

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

(green), calcium phosphate (light blue) and starch (dark blue). In addition, some unidentified components were also extracted for both the reduced and the full dataset (Figure S9).

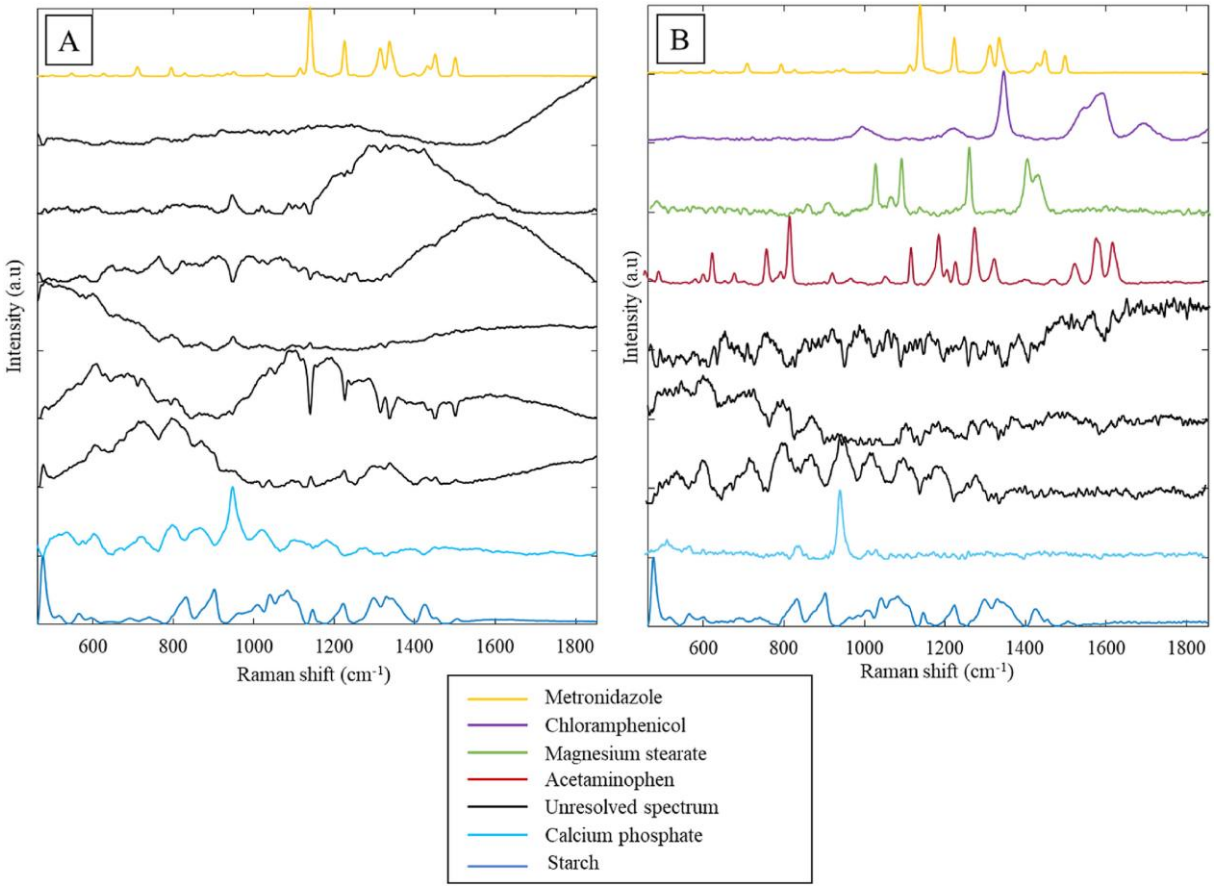

Fig. 6. MCR resolution of data set 4. A) Pure spectra obtained from the MCR-ALS resolution on the full dataset. B) Pure spectra obtained from the MCR-ALS resolution for the ESP-reduced matrix.

The concentration distribution maps of the 6 compounds extracted from the ESP-reduced dataset are provided in Fig. 7. It was not only possible to detect even chemicals that correspond to a low source of variance as for acetaminophen and magnesium stearate (Fig. 7D and E, respectively) but to extract a single pixel contribution for chloramphenicol (Fig. 7F). The results obtained were confirmed by comparison to previous works [8,10].

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

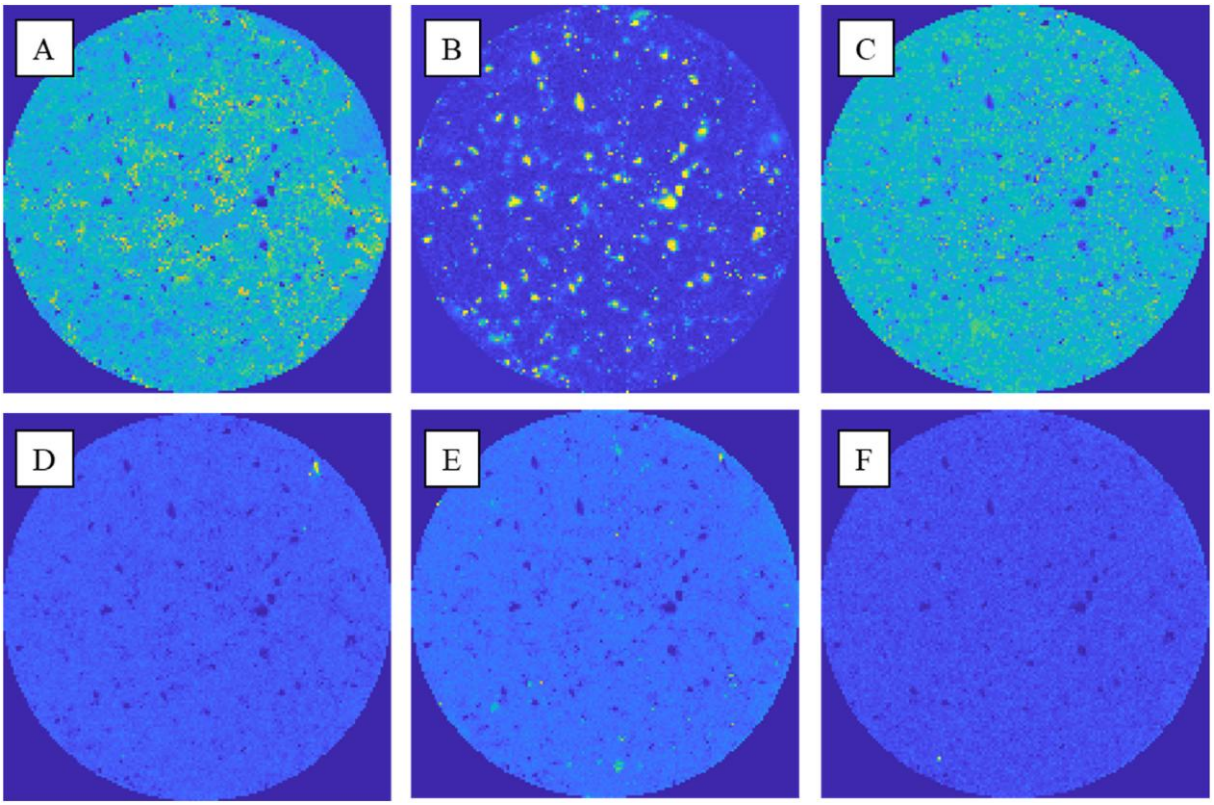

Fig. 7. Raman distribution maps obtained by the new constrained MCR-ALS. Concentration distribution maps of A) starch, B) metronidazole, C) calcium phosphate, D) acetaminophen, E) magnesium stearate and. F) chloramphenicol.

4. Conclusion remarks

We proposed a new strategy to analyze complex pharmaceutical formulations by hyperspectral imaging techniques and MCR-ALS analysis. We emphasized on the importance to perform an ESP-data reduction step before applying the curve resolution algorithm. This allowed resolving the chemical signatures (distribution maps and spectra) of very minor compounds. For both FT-IR and Raman imaging data, good analytical performance could be achieved., even when working with very large data sets and low signal-to-noise spectra. For FTIR analysis of pharmaceutical tablet, it was possible to achieve an almost complete elucidation of the medicine, disregarding the fact that we had a low representativeness of the sample. For the Raman analysis of falsified chloroquine tablet, it was possible to achieve a complete elucidation of the medicine. Deeper investigations of the convex hull selection of essential spectra with noisy data remains to be performed in the future, along with the potential assessment of the approach for quantitative purpose.

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

CRediT authorship contribution statement

Laureen Coic: Writing – original draft, Writing – review & editing, Software, Formal analysis, Visualization. **Pierre-Yves Sacré:** Conceptualization, Writing – review & editing. **Charlotte De Bleye:** Writing – review & editing. **Marianne Fillet:** Conceptualization, Funding acquisition, Writing – review & editing. **Cyril Ruckebusch:** Conceptualization, Funding acquisition, Writing – review & editing. **Philippe Hubert:** Supervision, Funding acquisition, Project administration, Writing – review & editing. **Éric Ziemons:** Supervision, Funding acquisition, Project administration, Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project has been supported by the European funds of regional development (FEDER) and by Walloon Region of Belgium as part of the operational program "Walloon-2020.EU" (L. Coic and A. Dispas).

The financial support of this research by the Walloon Region of Belgium in the framework of the Vibra4Fake project (convention n°:7517) was gratefully acknowledged.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2022.339532.

References

[1] S. Bureau, D. Cozzolino, C.J. Clark, Contributions of Fourier-transform mid infrared (FT-MIR) spectroscopy to the study of fruit and vegetables: a review, Postharvest Biol. Technol. 148 (2019) 1e14, https://doi.org/10.1016/ J.POSTHARVBIO.2018.10.003.

[2] T. Yaseen, D.W. Sun, J.H. Cheng, Raman imaging for food quality and safety evaluation: fundamentals and applications, Trends Food Sci. Technol. 62 (2017) 177e189, https://doi.org/10.1016/j.tifs.2017.01.012.

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

[3] J. Qin, M.S. Kim, K. Chao, S. Dhakal, B.-K. Cho, S. Lohumi, C. Mo, Y. Peng, M. Huang, Advances in Raman spectroscopy and imaging techniques for quality and safety inspection of horticultural products, Postharvest Biol. Technol. 149 (2019) 101e117, https://doi.org/10.1016/ j.postharvbio.2018.11.004.

[4] S. Gupta, S. Mittal, A. Kajdacsy-Balla, R. Bhargava, C. Bajaj, A fully automated, faster noise rejection approach to increasing the analytical capability of chemical imaging for digital histopathology, PLoS One 14 (2019), e0205219, https://doi.org/10.1371/journal.pone.0205219.

[5] J.L. Xu, K.V. Thomas, Z. Luo, A.A. Gowen, FTIR and Raman imaging for microplastics analysis: state of the art, challenges and prospects, TrAC Trends Anal. Chem. (Reference Ed.) 119 (2019), https://doi.org/10.1016/j.trac.2019.115629.

[6] V. Castiglione, P.Y. Sacre, E. Cavalier, P. Hubert, R. Gadisseur, E. Ziemons, Raman chemical imaging, a new tool in kidney stone structure analysis: casestudy and comparison to Fourier Transform Infrared spectroscopy, PLoS One 13 (2018), https://doi.org/10.1371/journal.pone.0201460.

[7] H. Rebiere, M. Martin, C. Ghyselinck, P.-A. Bonnet, C. Brenier, Raman chemical imaging for spectroscopic screening and direct quantification of falsified drugs, J. Pharm. Biomed. Anal. 148 (2018) 316e323, https://doi.org/10.1016/ j.jpba.2017.10.005.

[8] C.A. Waffo Tchounga, P.Y. Sacre, P. Ciza, R. Ngono, E. Ziemons, P. Hubert, R.D. Marini, Composition analysis of falsified chloroquine phosphate samples seized during the COVID-19 pandemic, J. Pharm. Biomed. Anal. 194 (2021) 113761, https://doi.org/10.1016/j.jpba.2020.113761.

[9] L. Coic, P.-Y. Sacre, A. Dispas, A.K. Sakira, M. Fillet, R.D. Marini, P. Hubert, E. Ziemons, Comparison of hyperspectral imaging techniques for the elucidation of falsified medicines composition, Talanta 198 (2019) 457e463, https://doi.org/10.1016/j.talanta.2019.02.032.

[10] L. Coic, P.Y. Sacre, A. Dispas, C. De Bleye, M. Fillet, C. Ruckebusch, P. Hubert, E. Ziemons, Pixel-based Raman hyperspectral identification of complex pharmaceutical formulations, Anal. Chim. Acta 1155 (2021) 338361, https:// doi.org/10.1016/j.aca.2021.338361.

[11] J. Cailletaud, C. De Bleye, E. Dumont, P.-Y. Sacre, Y. Gut, L. Bultel, Y.-M. Ginot, P. Hubert, E. Ziemons, Towards a spray-coating method for the detection of low-dose compounds in pharmaceutical tablets using surface-enhanced Raman chemical imaging (SER-CI), Talanta 188 (2018) 584e592, https:// doi.org/10.1016/j.talanta.2018.06.037.

[12] EDQM - European Directorate for the Quality of Medicines |, (n.d.). https:// www.edqm.eu/(accessed December 4, 2020).

[13] R. Spectroscopy, P.-Y. Sacre, L. Netchacovitch, E. Dumont, J. Cailletaud, C. De Bleye, M. Boiret, P. Hubert, E. Ziemons, Raman Hyperspectral Imaging: an Essential Tool in the Pharmaceutical Field Application Note Pharmaceutical RA-66, (n.d.).

[14] A. Nardecchia, C. Fabre, J. Cauzid, F. Pelascini, V. Motto-Ros, L. Duponchel, Detection of minor compounds in complex mineral samples from millions of spectra: a new data analysis strategy in LIBS imaging, Anal. Chim. Acta 1114 (2020) 66e73, https://doi.org/10.1016/j.aca.2020.04.005.

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

[15] M. Boiret, N. Gorretta, Y.M. Ginot, J.M. Roger, An iterative approach for compound detection in an unknown pharmaceutical drug product: application on Raman microscopy, J. Pharm. Biomed. Anal. 120 (2016) 342e351, https://doi.org/10.1016/j.jpba.2015.12.038.

[16] M. Boiret, A. de Juan, N. Gorretta, Y.M. Ginot, J.M. Roger, Distribution of a low dose compound within pharmaceutical tablet by using multivariate curve resolution on Raman hyperspectral images, J. Pharm. Biomed. Anal. 103 (2015) 35e43, https://doi.org/10.1016/j.jpba.2014.10.024.

[17] L. Duponchel, Exploring hyperspectral imaging data sets with topological data analysis, Anal. Chim. Acta 1000 (2018) 123e131, https://doi.org/10.1016/ j.aca.2017.11.029.

[18] S. Moncayo, L. Duponchel, N. Mousavipak, G. Panczer, F. Trichard, B. Bousquet, F. Pelascini, V. Motto-Ros, Exploration of megapixel hyperspectral LIBS images using principal component analysis, J. Anal. At. Spectrom. 33 (2018) 210e220, https://doi.org/10.1039/c7ja00398f.

[19] M. Boiret, N. Gorretta, Y.M. Ginot, J.M. Roger, An iterative approach for compound detection in an unknown pharmaceutical drug product: application on Raman microscopy, J. Pharm. Biomed. Anal. 120 (2016) 342e351, https://doi.org/10.1016/j.jpba.2015.12.038.

[20] C. Ruckebusch, A. De Juan, L. Duponchel, J.P. Huvenne, Matrix augmentation for breaking rank-deficiency: a case study, Chemometr. Intell. Lab. Syst. 80 (2006) 209e214, https://doi.org/10.1016/J.CHEMOLAB.2005.06.009.

[21] A.T. Badaro, J.M. Amigo, J. Blasco, N. Aleixos, A.R. Ferreira, M.T.P.S. Clerici, D.F. Barbin, Near infrared hyperspectral imaging and spectral unmixing methods for evaluation of fiber distribution in enriched pasta, Food Chem. 343 (2021) 128517, https://doi.org/10.1016/J.FOODCHEM.2020.128517.

[22] M. Ghaffari, N. Omidikia, C. Ruckebusch, Essential spectral pixels for multivariate curve resolution of chemical images, Anal. Chem. 91 (2019)

10943e10948, https://doi.org/10.1021/acs.analchem.9b02890.

[23] M. Ghaffari, N. Omidikia, C. Ruckebusch, Joint selection of essential pixels and essential variables across hyperspectral images, Anal. Chim. Acta 1141 (2021) 36e46, https://doi.org/10.1016/J.ACA.2020.10.040.

[24] T. Fearn, Multivariate curve resolution, NIR News 22 (2011) 18e19, https:// doi.org/10.1255/nirn.1229.

[25] A. de Juan, Multivariate curve resolution for hyperspectral image analysis, in: Data Handl. Sci. Technol., Elsevier Ltd, 2020, pp. 115e150, https://doi.org/ 10.1016/B978-0-444-63977-6.00007-9.

[26] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem e a review, Anal. Chim. Acta 1145 (2021) 59e78, https://doi.org/10.1016/j.aca.2020.10.051.

[27] M. Ghaffari, S. Hugelier, L. Duponchel, H. Abdollahi, C. Ruckebusch, Effect of image processing constraints on the extent of rotational ambiguity in MCRALS of hyperspectral images, Anal. Chim. Acta 1052 (2019) 27e36, https:// doi.org/10.1016/j.aca.2018.11.054.

Published in : Analytica Chimica Acta (2022), 1198, p. 339532
DOI: 10.1016/j.aca.2022.339532
Status : Postprint (Author's version)

[28] A. de Juan, R. Tauler, Multivariate curve resolution-alternating least squares for spectroscopic data, in: Data Handl. Sci. Technol., Elsevier Ltd, 2016, pp. 5e51, https://doi.org/10.1016/B978-0-444-63638-6.00002-4.

[29] C. Ruckebusch, R. Vitale, M. Ghaffari, S. Hugelier, N. Omidikia, Perspective on essential information in multivariate curve resolution, TrAC Trends Anal. Chem. (Reference Ed.) 132 (2020), https://doi.org/10.1016/j.trac.2020.116044.

[30] S. Hugelier, O. Devos, C. Ruckebusch, A smoothness constraint in multivariate curve resolution-alternating least squares of spectroscopy data, in: Data Handl. Sci. Technol., Elsevier Ltd, 2016, pp. 453e476, https://doi.org/10.1016/ B978-0-444-63638-6.00014-0.

[31] A. Malik, R. Tauler, Ambiguities in multivariate curve resolution, in: Data Handl. Sci. Technol., Elsevier Ltd, 2016, pp. 101e133, https://doi.org/10.1016/ B978-0-444-63638-6.00004-8.

[32] S. Hugelier, O. Devos, C. Ruckebusch, On the implementation of spatial constraints in multivariate curve resolution alternating least squares for hyperspectral image analysis, J. Chemom. 29 (2015) 557e561, https://doi.org/ 10.1002/cem.2742.

[33] H.F.M.B. Paul, H.C. Eilers, Baseline Correction with Asymmetric Least Squares Smoothing - PDF Free Download, 2005. https://technodocbox.com/3D_

Graphics/76845483-Baseline-correction-with-asymmetric-least-squaressmoothing.html. (Accessed 18 August 2021).

[34] H.G. Hecht, The interpretation of diffuse reflectance spectra, J. Res. Natl. Bur. Stand. Chem. 80 (1976).

[35] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (2002) 1627e1639, https://doi.org/ 10.1021/AC60214A047.

[36] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem e a review, Anal. Chim. Acta 1145 (2021) 59e78, https://doi.org/10.1016/J.ACA.2020.10.051.

[37] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Http://Dx.Doi.Org/10.1080/ 10408340600970005. 36, 2007, pp. 163e176, https://doi.org/10.1080/ 10408340600970005.

[38] A. Bogomolov, M. Hachey, Application of SIMPLISMA purity function for variable selection in multivariate regression analysis: a case study of protein secondary structure determination from infrared spectra, Chemometr. Intell. Lab. Syst. 88 (2007) 132e142, https://doi.org/10.1016/J.CHEMOLAB.2006.07.006.