MIA and NIR Chemical Imaging for pharmaceutical product characterization

JoséM.Prats-Montalbán[1], Jackeline I. Jerez-Rozo[2], Rodolfo J. Romañach[2], and Alberto Ferrer[1]

[1]Universidad Politécnica de Valencia Dpto. Estadística e I.O. Aplicadas y Calidad.

[2]Department of Chemistry, University of Puerto Rico, Mayagüez, PR.

Abstract

This paper presents a three step methodology based on the use of chemical oriented models (MCR and CLS) for extracting out the chemical distribution maps (CDM's) from hyperspectral images, afterwards performing multivariate image analysis (MIA) on the CDM's, and finally extracting "channel" and textural features from the score images related to quality characteristics These features show complementary properties to those directly obtained from the CDM's, since they take advantage of their internal correlation structure. The approach has been successfully applied to the evaluation of homogeneity and cluster presence of API in a novel formulation developed to improve the dissolution of poorly soluble drugs.

INTRODUCTION

Pharmaceutical regulations, such as The Current Good Manufacturing Practices (CGMPs) as described in 21CFR211.10 require sampling and testing of in-process materials and drug products to evaluate the adequacy of mixing to assure uniformity and homogeneity. The optimal determination of the distribution of the drug and excipients affects blend homogeneity, content uniformity, and may also affect dissolution [1]. These issues are related, not only to the manufacturing process, but also to the solubility of the Active Pharmaceutical Ingredient (API).

It is estimated that over 40% of all possible new active drug candidates have very low solubility [2]. One possible approach for improving the solubility, and hence dissolution, of these drugs is to disperse them in a polymeric film, reducing surface tension, and thus, preventing them from aggregating again [3]. Therefore, the ability to visualize and assess the compositional heterogeneity and structure of the end products is extremely important for the design, development and manufacture of polymeric films. Add to this, process understanding and product design are of main importance from the process analytical technology (PAT) and quality by design (QbD) points of view.

Before the advent of chemical imaging methods, the evaluation of the adequacy of mixing was limited to an inter-unit definition determined through the standard deviation and average in drug concentration between different unit doses of the formulation. Spectroscopic chemical imaging methods, such as Near Infrared Chemical Imaging (NIR-CI) now permit an intra-unit definition of drug distribution by providing reliable chemical and spatial information on the distribution of drug and excipients. Chemical imaging methods provide knowledge on how the drug and excipients are mixed together, providing an understanding of the microstructure of the formulation.

A number of approaches have been developed to extract information from the hyperspectral images obtained [1, 4]. In some cases, simple univariate approaches have been used [5]. However, it is preferable to work with the entire data array linked to the hyperspectral image to study the distribution of the compounds in the mixture, and determine the abundance of each chemical compound at each pixel location. This task has been mainly performed by Multivariate Curve Resolution (MCR) [6, 7], or by Classical Least Squares (CLS) [8, 9] if the constituents in a mixture are known *a priori*.

These methods create chemical distribution maps (CDM's) [10] by taking into account the natural spectral correlation in the hyperspectral data cube. Nevertheless, CDM´s are afterwards analyzed in a univariate way. Thus, the correlation structure between and within the CDM's segregated chemical compounds, both in terms of chemical and spatial (textural and physicochemical mixture properties) information is not used.

When using MCR or CLS, what we do is to segregate the information linked to each of the chemical compounds in the mixture, by using the (hyper)spectral correlation, in a chemical sense. This way, we obtain full chemical interpretable images, where the chemical compounds of the mixture appear in each separated “chemical channel (or band)” and distributed according to their corresponding chemical concentration, converted into a grey level intensity.

However, this information might not be sufficient for process monitoring or final quality prediction purposes, since it is not only the different distribution and concentration of the chemical compounds in the image what counts, but also the way they combine.

In order to analyze these correlation structures of the mixtures, multivariate image analysis (MIA) [10, 11] may be useful, taking advantage of the use of multivariate statistical modeling on the CDM’s, and unraveling the different behaviors in separate PC’s. Properties may depend on the correlation structure of the distribution of the different chemical compounds. Performing this way, we can assess the importance of each type of information, and link it to the final quality properties (*i.e.* maybe for some final property we have the separation zones gathered by one PC in its score image as being the most important one; whereas for another property it is the mixing information what matters). Thus, the former commented PAT and QbD goals (process understanding, process and product design and final quality) can be better achieved, hence obtaining better monitoring and predictive models.

The present work reports a three-step methodology to analyze the chemical composition and the spatial relationships between API and different excipients. The first step consists of the application of resolution models [12] (Classical Least Squares (CLS) [9] or Multivariate Curve Resolution (MCR) [6]) in order to properly separate the chemical information in the hyperspectral images. The second step applies MIA to obtain meaningful and complementary improved information from the images related to each chemical compound in the mixtures, i.e. to explore the spectral and spatial relationship between the API and different excipients. Finally, a third step uses these score images obtained from MIA to extract out features able to characterize quality properties of the images. These features will be compared to those provided by features directly extracted from the CDM’s in a univariate way.

Section 2, devoted to materials and methods, presents the type of formulation used, as well as the procedure employed for their preparation. Afterwards, the NIR spectral acquisition system and the data pretreatment applied to the spectra are commented. In Section 3, the chemometrics models used: CLS and MCR, and the MIA methodology are briefly explained. Section 4 presents the results obtained. Finally, Section 5 provides the conclusions.

2. MATERIALS AND METHODS

Materials

HPMC (Hydroxypropylmethylcellulose), 2% viscosity in aqueous solution (20^{o}C), and griseofulvin ((2*S*,6'*R*)- 7-chloro- 2',4,6-trimethoxy- 6'-methyl- 3*H*,4'*H*-spiro [1-benzofuran- 2,1'-cyclohex[2]ene]- 3,4'-dione)were obtained from Aldrich Chemical (Milwaukee, WI).

Procedure

The films were prepared using two different procedures. The first procedure was followed to obtain large drug agglomerates; the diameter of these agglomerates is approximately 0.3mm. A total of 100 mL of distilled water were heated to 70^{o}C, and then 4 g of HPMC and 4 g of griseofulvin were added. The dispersion was mixed with a spatula and was poured on a smooth plastic surface and left at room temperature until all the solvent evaporated. These films with large agglomerates were prepared to facilitate method development and to better understand the spectral changes of the HPMC-griseofulvin film.

The method of preparation of the second set is described in [5]. Micronized drug was dispersed in HPMC at 80°C with constant agitation for 12 h. This set consisted of four polymeric thin films that ranged from 36.4% - 57% (w/w) drug concentration.

NIR Chemical Imaging

Hyperspectral images are images where a full spectrum per pixel is obtained. Near infrared hyperspectral images were acquired using the Malvern SyNIRgi Near Infrared Chemical Imaging System (Olney, MD). The images were obtained in diffuse reflectance mode by placing the polymeric thin films over a white ceramic disk with a diameter of 28 mm. This disk was used as a reference for the acquisition system. The spectra of the HPMC and Griseofulvin were obtained as pure powders also in the diffuse reflectance mode. Spectra were collected with the system's focal plane array detector that has 256×320 pixel elements, with a total collection time of about 2 minutes, if we do not consider the acquisition of dark and reference measurements. Images were acquired using a 10 µm per pixel objective, providing images of an area of approximately 3.2×2.6 mm. Spectra were obtained with 1 scan using a spectral range of 1200-2400 nm. The Pixis ® CI software from Malvern Instrument was used for data acquisition.

Data treatment for NIR spectra

The logarithm, log10 (1/R), was first applied to the data cube to convert the spectra to absorbance units. Bad pixels [13] were removed and replaced by the average value of the intensities of the surrounding pixels and a low-pass Fourier filter was applied; these pretreatments were made using ISys 5.0 software package version 1.0.4. The spectra were then normalized using the Standard Normal Variate method and Savitzky-Golay second derivative (filter order 3, filter width 9). These pretreatment were applied with the purpose to eliminate multiplicative and additive effects, and baseline differences between the spectra. Finally, a Db4 discrete wavelet transform function was applied as a compression tool in the spectrum domain, for efficiently reducing the size of the third dimension, maintaining the useful information and at the same time producing some more de-noising of the signal (although this was not the goal). The reason was the need for memory space in the computer. For this purpose, we used MATLAB 7.5 (The MathWorks, Natick, MA). This way, the original 10 nm resolution was reduced to 20nm, remaining the 1200-2400 nm range the same.

These pretreatments were applied carefully, trying to deal at each point with the problem at hand.

3. MIA & NIR CHEMICAL IMAGING

In this section, the proposed three-step methodology is introduced. It is summarized in Fig. 1.

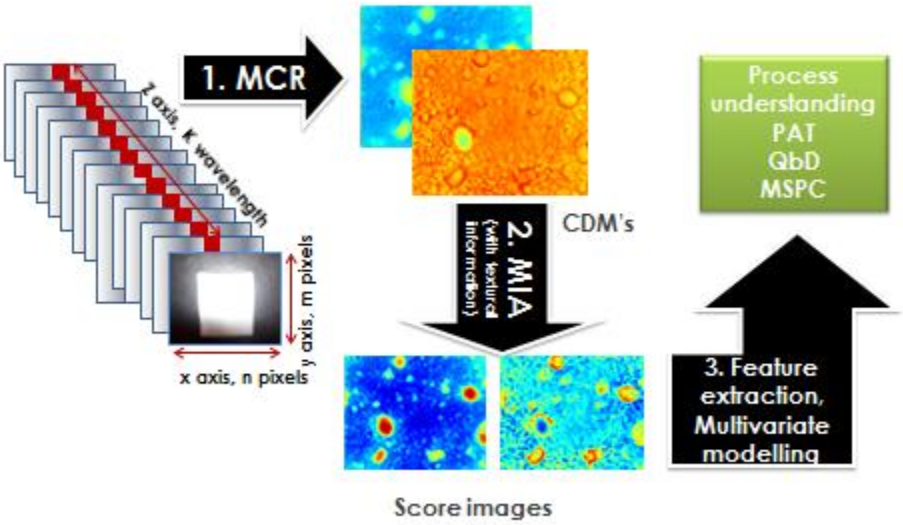


Figure 1. Scheme of the three step method proposed. Red squares relate to one same pixel at each wavelength used as an example.

3.1 First step: Resolution methods

The first step of the methodology searches for the distribution of chemical compounds in a mixture. It is based on the use of chemometric resolution models that incorporate the Beer-Lambert law (e.q. 1)

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad (1)$$

Where **X** relates to the unfolded image data matrix, $\mathbf{S}^{\mathbf{T}}$ is the matrix of pure spectra and **C** are the stretched concentration profiles. This law states that the spectrum of any sample (pixel in this case) can be represented by the concentration-weighted sum of the contributions of the pure spectra present in the mixture. In this work, Multivariate Curve Resolution (MCR) [6, 14] and Classical Least Squares (CLS) [8, 9] have been used.

These models are briefly presented, and the reader is directed to the references for further details. Moreover, it must be pointed out that, in the case of a hyperspectral image, prior to applying these models it is necessary to unfold it in a multivariate image analysis fashion [10], hence obtaining an **X** matrix where the rows gather the pixels spectra and the columns their intensities at each wavelength.

Classical Least Squares (CLS)

When the original constituents of the mixture and their spectra are *a priori* known, application of CLS models [8, 9] is an option in this case. CLS regression consists of projecting each sample spectrum forming an **X** matrix on the pure spectra, hence obtaining the concentration directly related to the chemical compounds in it, by using eq (2)

$$\mathbf{C}= \mathbf{X}(\mathbf{S}^{T})^{+} + \mathbf{R} \qquad (2)$$

Where $(\mathbf{S}^{\mathbf{T}})^{+}$ is the pseudoinverse of the $\mathbf{S}^{\mathbf{T}}$ matrix related to the NIR pure spectra. By refolding matrix **C** into the original spatial dimensions on the images, the chemical distribution maps (CDM's) are obtained, then being able to go to the second step of the methodology.

Multivariate Curve Resolution (MCR)

MCR methods [6, 14] allow for the resolution of individual contributions when the spectra of pure components are not available, under certain specific constraints that can be introduced in the model. Different algorithms can be used for obtaining **C** and **S**. In this work, the MCR-ALS has been used [6, 7]. The MCR results are, regarding these image applications, the concentration matrix **C** and the pure spectra of the image constituents gathered in matrix $\mathbf{S}^{\mathbf{T}}$ in eq 1. The basic steps of MCR-ALS are:

1. Determination of the rank of the data matrix **X** (in this case, *a priori* known).
2. Generation of initial estimates (**C**-type or $\mathbf{S}^{\mathbf{T}}$-type).

3. Given **X** and **C**, constrained least squares calculation of $\mathbf{S}^\mathbf{T}$.
4. Given **X** and $\mathbf{S}^\mathbf{T}$, constrained least squares calculation of **C**.
5. Reproduction of **X** from calculated $\mathbf{CS}^\mathbf{T}$. If reproduction is satisfactory, end of the process. If not, go back to 3.

As for CLS, by refolding into the original spatial dimensions on the images, the CDM's are obtained.

It is important to stress that resolution [12] can be applied to one or more images together (see Figure 2). Multiimage analysis is the option to be used when a multilayer image from a single sample or a series of images with related chemical composition, *e.g.* groups of samples (polymeric films in this case) imaged as a function of time, temperature or any other variable, are encountered [10, 15]. This way, by stacking different sample groups (pixels) related to different images formed by the same chemical compounds, one below the other, one single **S** matrix is forced to be obtained. Thus, all the concentration profiles are then related to the same spectra; since MCR-ALS is performed on the new **X** multiimage data set (Fig. 2).

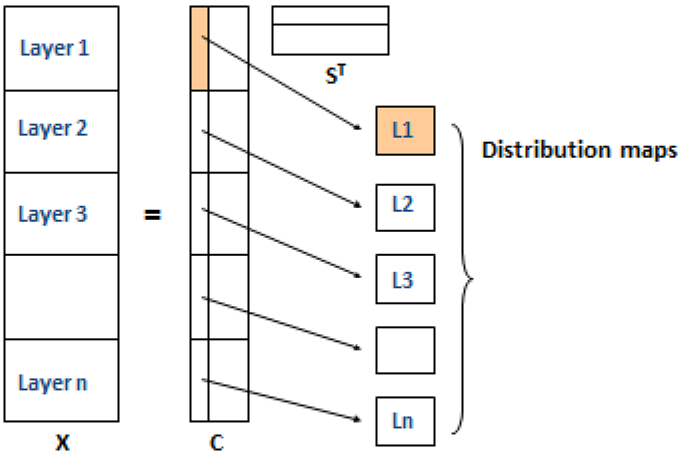


Figure 2. Application of MCR on a multiimage data set

In this study, only non-negativity constraints in the concentrations were considered, due to the pre-treatment applied to the spectra. In order to give all the pixels the same *a priori* weight, equal length scaling was also applied.

Applications of additional methods for the analysis of chemical images have been published by A. de Juan *et al*. for improving these methods, in order to provide local rank information (presence or absence of compounds in pixels) that can be later on used as an additional constraint [15, 16].

3.2 Second step: Multivariate Image Analysis

Once the chemical information has been resolved, it is appropriate to arrange the CDM's into a single multivariate image, formed by the components determined in the analysis. This is fundamental since the final spatial distribution of the concentrations is not only determined by each chemical compound itself, but also by the internal correlation structure between and within the different chemical compounds in the mixture.

Thus, a methodology able to gather all this chemical information, providing new meaningful chemical information is needed. This methodology comes from the application of MIA [10], which creates new images called score images, combination of the original ones, and vectors of loadings that provide the type of information provided in these score images (i.e. the internal correlation structure). This is achieved by applying principal component analysis (PCA) [17] on to the unfolded images [10, 11].

PCA models compress the image information into a reduced number of uncorrelated (orthogonal) variables, called principal components (PCs). PCs are linear combinations of the original variables and describe the most important information of the image (data set variance) in decreasing order. The general PCA model can be written as: $\hat{\mathbf{X}} = \mathbf{TP^T}$. The pixel coordinates in the space of principal components (scores) are in the score matrix **T**. The relevance of the original variables in the principal component space and their internal correlation structure is retained by the loadings matrix **P**, which provides information on how the channel bands combine in order to form the new score images with uncorrelated information, obtained by refolding the score vectors **t** into the original 2D image spatial structure. Finally, the difference between the original data structure and the predicted by the PCA model, computed as $\mathbf{TP}^{\mathbf{T}}$, is stored in the residual matrix [10].

In MIA, the **X** matrix is formed by unfolding each channel band of the image into one single column, and juxtaposing each unfolded channel, *i.e.* each column, one beside the other. This way, the **X** matrix has as many rows as pixels has the image, and as many columns as channel bands. During this unfolding stage, the textural (spatial) information is lost because each row is linked to each pixel of the image. To avoid this limitation, the Bharati and MacGregor approach [18] can be applied for each chemical channel concentrations following [19] to preserve the textural (spatial) information into the model. Thus, the **X** matrix is sized (nr. pixels × (nr. of compounds × nr. of pixel neighbors)). Each row of **X** contains now the concentration values of the different chemical compounds for a given pixel and for all the neighboring pixels.

As commented before, the final spatial distribution of the concentrations depends on the internal correlation structure between and within the different chemical compounds in the mixture. And this fact, *i.e.* the way how the chemical compounds interact (depending on the proportion of the mixture, process conditions, etc.), will of course have an influence on the final quality properties of the pharmaceutical product. This also means that process conditions can be determined by inspecting these latent variables (score images), related to specific behaviors shown up by the loadings, in a QbD framework.

Furthermore, one could be able to understand how process conditions influence the segregation or mixture between some chemical compounds, taking a look at the loadings and their percentage of variance in the model. Thus, it would also be possible to design processes that favor one or some other desired property.

3.3 Third step: feature extraction

Therefore, a further third step is to use PCA or some multivariate regression model, *e.g.* Partial Least Squares, PLS [20], to relate the extracted features with final quality properties and/or with process conditions. This third step consists of characterizing the score images by different types of features, in order to take advantage of the meaningful information provided by them when looking at the loading plots. Thus, we can characterize chemical separation and mixing behaviors, spatial characteristics related to some specific compound (*e.g.* API if this adds some value to the final quality), etc.

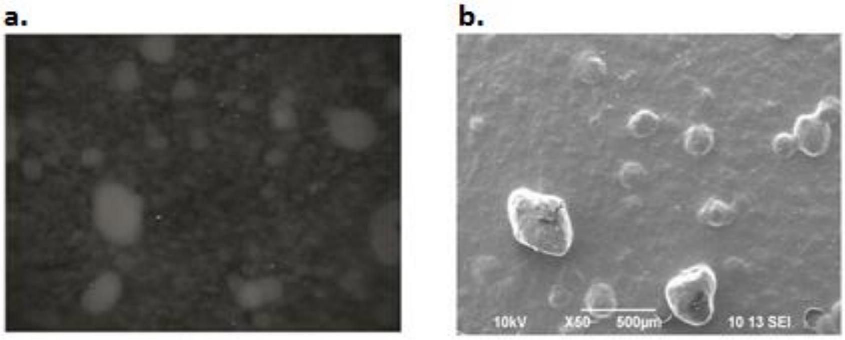


Figure 3. Image of films with larger agglomerates of drug. **a.** Image obtained with a microscope coupled to NIR-CI using a magnification of 10 μm per pixels. **b.** Image obtained with the SEM.

The characterization of these score images can be done as for any type of image. One common way to do this is by texture feature extraction, which tries to analyze and summarize the spatial distribution of the intensities in one gray level image (as is the case for the score values in a score image). This way, each set of score images related to one image is converted into a row of features of each of the score images analyzed, hence forming a feature vector. We have to do this because we are analyzing the spatial distribution characteristics of different kind of phenomena explained by the PCA loadings. In this study, the spatial distribution of the chemical segregation zones, and the spatial distribution of the chemical mixing zones.

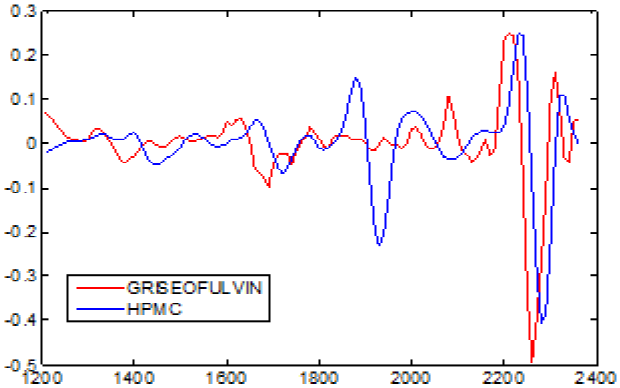


Figure 4. Spectra of pure components.

Many texture analysis techniques exist in the literature, such as co-occurrence matrices, structural models, wavelets, etc. The reader is referred to [10], where several texture analysis techniques are presented and referenced.

Table 1

Percentage of variation explained by PCs.

PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
39.5%	28.2%	10.0%	9.25%	6.92%	6.17%

4. RESULTS

In this section, the results of applying this methodology are shown, first on a large agglomerates image, in order to properly illustrate the procedure; afterwards on real process conditions formulations, introducing the difficulties that may arise when dealing with many images, maybe different suppliers or variability in the illumination conditions.

4.1 Analysis of large agglomerates

Figure 3 shows the large drug clusters that characterize the first data set commented in Section 2. The system's microscope easily provided an image of these clusters (Fig. 3a), and the clusters were then confirmed by SEM (Fig. 3b). The images shown in Fig. 3 are from the same sample, although the sample areas observed with the two microscopes are not identical. The large clusters in these films facilitated method development. The NIR spectral data was then used to obtain CDM’s for the two components of the film, using both CLS and MCR-ALS.

As commented before, the use of CLS algorithms requires the spectra of the pure compounds, or an approximation if performing MCR-ALS. In this case spectrum of each the pure compound was obtained from the first principal component of a PCA model fitted from spectra of the pure compounds. The obtained pure spectra for API and HPMC are shown in Figure 4, after applying SNV and second derivative transforms. This was also used as an initial approximation of the final solution for the MCR-ALS, which in fact rapidly rotated to the same CLS solution.

In this case, the selection of CLS was based on spectral similarity, which, even though different experimental conditions between the reference spectrum and the ones registered in the images exist, is an objective indicator of the quality of the final results. However, as shown later on, this criterion may change depending on the images analyzed and the problem at hand.

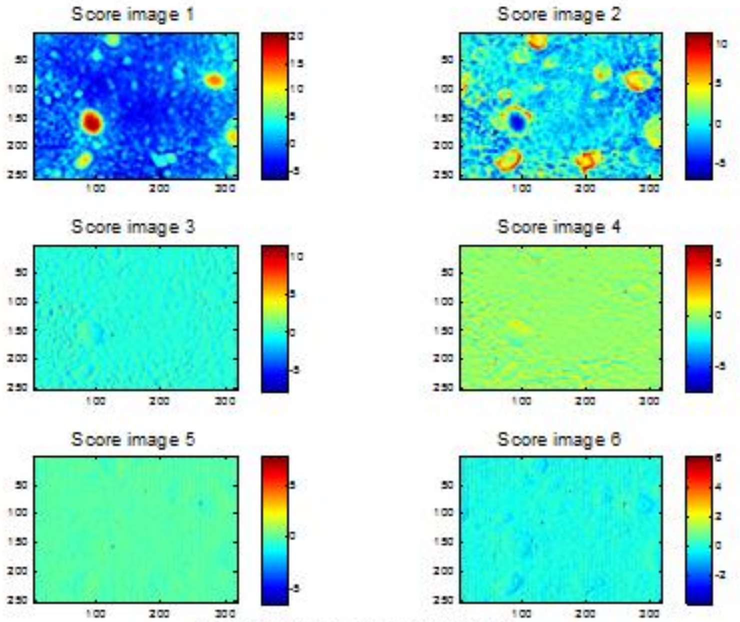


Figure 5. Score Images from PCA.

MIA was performed on the CDM’s provided by the CLS model following [19]. This procedure was applied for each chemical channel (griseofulvin and HPMC) concentrations. The reason for applying this approach is that spatial information specific to each chemical compound might exist and have an influence in the model close to that provided by the chemical correlation structure gathered by other PC’s. In these cases, it should be included as another color (concentration profile) band.

It may be useful to apply MIA on CDMs (without textural augmentation) when many compounds are present and a simple overlapped RGB map would not reflect all possible mixing phenomena, or it does not provide clear enough information, as will be shown later. Enlarging the **X** matrix by neighboring pixel channels needs to be done only when textural information can be potentially relevant, regardless of the number of chemical compounds in the image. This can be the case when changes in process conditions may affect these physical spatial phenomena, giving them more importance in the model, and hence becoming worthy to be included in further analyses, as the one proposed as a third step in section 3.

A window size of 3×3 was used, which means that, for each pixel in the image, we consider not only its own CDM intensity, but also the intensities of the eight neighboring pixels. This size was used because the sizes of the agglomerates are supposed to be small, so the window size to use should also be small. This is done for each of the CDM’s considered. Thus, the first nine columns of the **X** matrix refer to one chemical compound (HPMC in this case), whereas the last nine columns relate to API. A PCA model with six PC’s (Principal Components) was fitted on to the resulting **X** matrix. The first two principal components explain 67.66% of the variation in the data. Table 1 presents the percentages of variation explained by each PC. Figure 5 presents the corresponding score images, which show the value of the corresponding PC at each pixel location. The loading plot further explains the information conveyed by the score images as shown in Fig. 6. In this case, the first PC is providing the concentration difference map between the API and the excipient (HPMC). Pixels with high levels of API and low levels of HPMC will have high positive values for PC 1 (red color in score image 1). Pixels with high levels of HPMC and low levels of API will have high negative values for PC 1 (blue color in score image 1). On the other hand, the second PC shows an average of the concentrations of any of the two chemical compounds (API or excipient). PC 2 is providing very interesting information, since the average value seems to correspond to those areas where API and excipient are mixed. This is explained by the positive values observed in score image 2. Thus, the first two principal components are providing information related to the mixing of the HPMC and API. The rest of the PC's provide spatial information of each chemical compound separately, which is also very interesting, since it means that the textural information of each chemical compound evolves independently of the other one. PC's 3 and 4 are gathering the textural aspects of the HPMC, whereas PC's 5 and 6 are mainly related to textural aspects of the API.

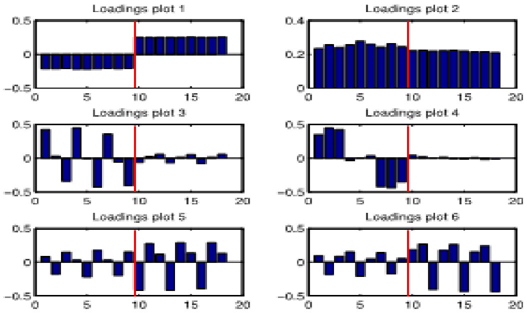


Figure 6. Loading Plots from PCA. Loadings of HPMC are on the left side of each Loading plot, and API of the right (separated by the red line)

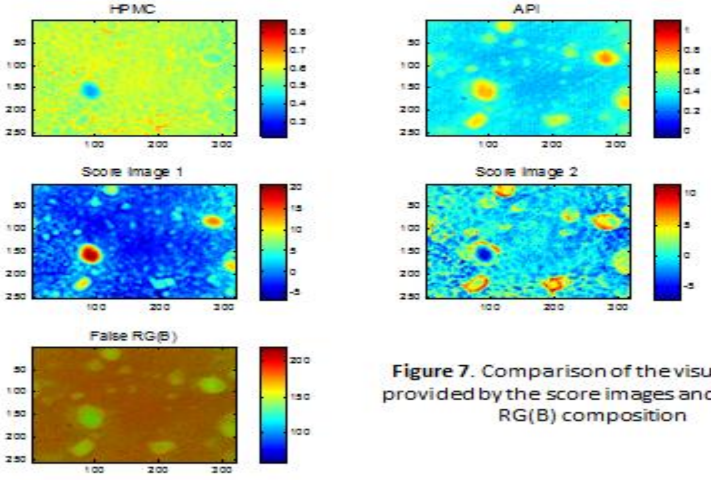


Figure 7. Comparison of the visualization provided by the score images and the false RG(B) composition

The most important conclusion from this first study is the benefit of using both types of models: the chemically oriented ones (CLS or MCR) for compressing, transforming the spectral information into separated chemical distribution maps; and MIA for improving the interpretation when separating the joint variability into orthogonal information maps (the score images) with a clear physical interpretation (chemical compounds blend and separation zones).

Thus, this two-step methodology analyzes the chemical and spatial relationships between API and different excipients simultaneously; and takes advantage over other approaches by reporting which type information is gathered individually for each component, *i.e.* improving the process understanding task.

It would be also possible, in this case, to form a false RG(B) image from the HPMC and API CDM's, in order to assess these blending and separation zones, as shown in Figure 7, bottom left. However, when inspecting the false RG(B) composition, the blending areas (in yellow) do not appear as well defined and graded as in score image 2 (Fig. 7b, middle right). These blending areas are mainly related to the surroundings of the clusters, which in the false RG(B) are not so well detected. Even more, when inspecting score image 1, the well defined separation areas related to API (in red color) do not show up so clearly in the false RG(B) image. Summarizing, the information provided by the score images seems to be richer than that provided by the false RG(B) image.

Finally, since only one image prepared in these conditions was provided, the third step could not be applied. Moreover, the purpose of this preparation was not to reproduce real conditions, but to form clear agglomerates for validating the methodology.

Table 2

Correlations between pure spectra and optimized.

1.0000	0.2511	0.9495	0.3537
0.2511	1.0000	0.3421	0.9183
0.9495	0.3421	1.0000	0.3725
0.3537	0.9183	0.3725	1.0000

4.2. Analysis of real process conditions agglomerates

The next set of images was formulated as explained in section 2, using real process conditions. Eight images related to two locations (a and b) of four different thin films were analyzed by the proposed methodology. The films API concentrations were 36.4%, 44%, 50% and 57% (w/w). In this case, the question was to investigate if is it possible to extract out valid features to be related to final quality characteristics.

CLS and MCR were applied on each of the eight provided images. However, residuals from CLS showed higher values than for the first formulation. This could be due to the fact that pure reference spectra were obtained in powder form for HPMC, so there was probably much more scattering with them, or because the pure spectra were not coming from the exact same batch as the one used for creating the formulations.

Using MCR for each image provided poor results too, because correlations between the pure and optimized API spectra ranged from 57.84% for the 34.6% API to 80.14% for the 57% (w/w) API; *i.e.* an increasing correlation between the API pure and optimized spectra with its concentration in the mixture. This is according to previous studies where, the more concentration of any chemical compound was in the image the better the related spectra were predicted.

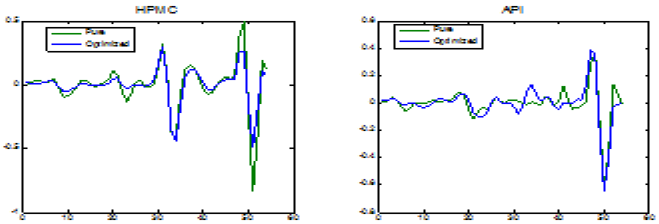


Figure 8. Pure and MCR-ALS optimized spectra for HPMC (left) and API (right).

When having several images with common chemical compounds, a possible approach is to apply MCR on a multiimage data set [10, 12, 15], *i.e.* to use all them in the MCR model, hence forcing a unique optimized spectra data set. This is really applicable in this case, since we have the same chemical compounds in the formulations. Moreover, this is one way to eliminate different illumination variations from image to image, and at the same time permits to use different providers, since the pure spectra are introduced just as an approximation.

In our case, for computational restrictions, it was necessary to take subsets of pixels from the images, instead of the whole images. Nevertheless, the results obtained, in terms of correlation between the pure spectra and the optimized ones (Table 2 and Figure 8) show that this approximation is valid. The CDM's obtained are shown in Fig. 9.

However, still some illumination effects could be observed in the images. In order to improve these images for MIA analysis, it was decided to apply some background elimination image analysis procedure, such as morphological opening [21-23]. The improved images show better aspect, as shown in Fig. 10 for 44% (w/w) formulation, loc. a).

Once images have been enhanced, they were stacked in a 2 chemical bands image, so MIA could be applied, following again [19]. PC's 1 and 2 information, associated to the chemical correlation structure between HPMC and Griseofulvin, are shown in Fig 11. Inspecting the loadings, it can be stated again that PC1 reflects the blend distribution map (score image 1), whereas PC2 is related to the difference distribution map (score image 2). The same results were obtained for the rest of images.

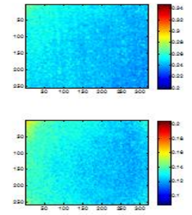


Figure 9. CDM's for HPMC (top) and API (bottom) for the 44% (w/w) formulation. The model explains a 99.52% of the data.

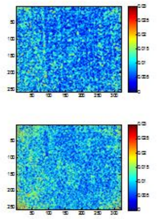


Figure 10. Improved CDM's images: HPMC (top) and API (bottom) for the 44% (w/w) formulation.

The fact of using just two score images (PC's) for this example (real conditions images) is to make the paper simpler and provide a fair comparison between the information extracted from the two CDM's available and the information derived from the score images.

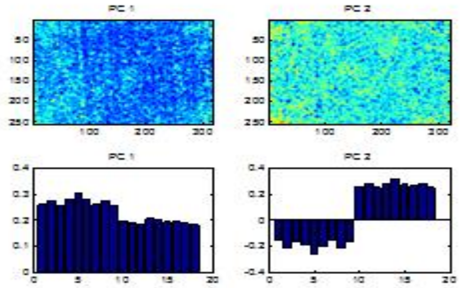


Figure 11. Score images and loadings plots for the PCA model

Note that the number of PC's to use depends on the final goal (e.g. process understanding, monitoring, prediction, missing data recovery, …) searched. In this study, since few samples were available and films were prepared under similar process conditions, information on possible changes in process conditions or quality properties was not available, and a sound study on the appropriate number of PC's was not possible.

Until this point, the proposed methodology has shown how to gain in process understanding by inspecting the main sources of variation in the hyperspectral images. However, there is still much to know, as for instance relating these internal structures to quality parameters, and even how to create, produce these by process manufacturing conditions determination/design. One reasonable way to do this is by converting these images that inform about the final spatial distribution of each chemical compound (coming from many possible causes, chemical and not chemical, as commented before), into textural and color (concentration profiles) characteristics and analyze them by PCA [17] or, when DOE (Design of Experiments) parameters or quality variables are available, by PLS [20].

4.3 Score images based feature extraction

In this case, no DOE or quality variables were available. Even though, it was still possible to extract different characteristics from the score images, afterwards applying PCA as an unsupervised method, and see how the different locations and concentrations images relate.

In order to do this, different color-texture descriptors may be applied. In this study, the texture techniques applied have been:

- Soft-Color Texture Descriptors (SCTD)
- Covariances extracted from 4 scales from a DWT decomposition
- MR-MIA I (because texture can be important within each score image)
- Log SVD values

Descriptions of these methods can be found in [24]. The best results were found for MR-MIA I (Figs. 12 and 13). It must be stressed that all methods provide very similar score values for the first PC. The score values show a clear evolution of the images with the concentration, even for an unsupervised method such as PCA. But even more, when inspecting score plot 1*vs*2 (Fig. 13), the different locations of each concentration cluster very well. This means that underlying phenomena not only related to the concentration is being caught by the characteristics extracted from the score images. So, promising results when quality variables are available, such a drug delivery, are expected.

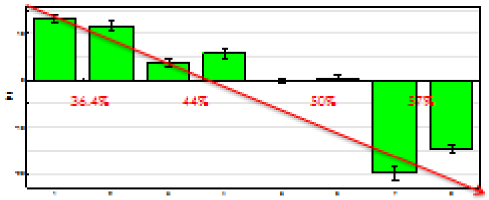


Figure 12. Score column plot for MR-MIA I PCA model.

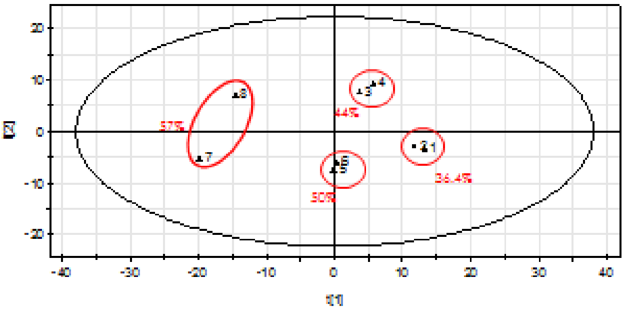
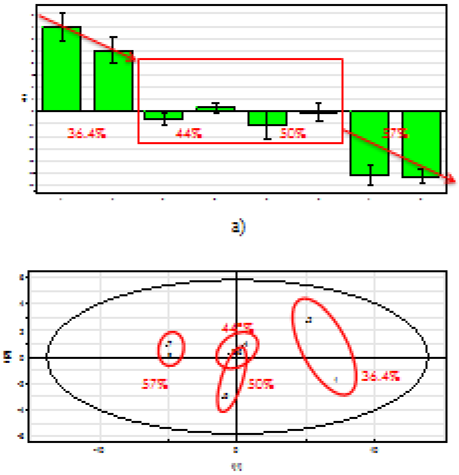


Figure 13. PC's 1 and 2 score plot for MR-MIAI PCA model.

4.4 Comparison with the characteristics extracted from the CDM's

In order to validate the benefits of the methodology, the same PCA models on the four types of features presented, but directly extracted from the CDM's were built. In this case, no evolution with the concentration could be observed. When repeating all these analyses on Multiplicative Scatter Correction [25] pre-processed data (instead of SNV and Savitzky-Golay), better results were found for CDM's in relation with the API concentration. This could be due to the fact this way non negativity constraints could be applied on the MCR modeling stage. However, the clustering ability of the score images was not reached by the CDM's. Results are shown in Fig. 14 for the best results obtained by the CDM's on MSC preprocessed data, which still show some overlapping in both PC's 1 and 2.



a)

b)

Figure 14. Score column plot (a) and PC's 1 and 2 score (b) plot for MR-MIA I PCA model.

Summarizing, the concentration prediction ability for the score images is better using SNV-SG preprocessing, while for the CDM's results are better when applying the MSC preprocessing. Anyway, the rest of the remaining relevant information in the images, which can be of main importance for the desired goals, is only gathered by the score images.

5. CONCLUSIONS

The present work introduces a methodology for efficiently analyzing hyperspectral images from pharmaceutical formulations by means of chemometrics resolution models and multivariate image analysis, in a three-step approach.

The methodology not only provides information on the relative concentration and spatial distribution of API and excipient in the formulation, but also on the internal correlation structure of the mixture, allowing the extraction of features related to quality characteristics of the formulation. This has clear benefits for process understanding and real time release, critical issues for PAT and QbD.

This methodology has been successfully applied on a novel pharmaceutical formulation, designed to disperse poorly soluble drug particles in polymeric thin films and keep the API particles from agglomerating.

Acknowledgements

Research in this study was partially supported by the Spanish Ministry of Science and Innovation and FEDER funds from the European Union through grant DPI2011-28112-C04-02, and also by NSF-Engineering Research Center for Structured Organic Particulate Systems (ERC-SOPS, EEC-0540855) and the program NSF-Major Research Instrumentation grant 0821113.

References

[1] C. Gendrin, Y. Roggo, C. Collet, Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review, *Journal of Pharmaceutical and Biomedical Analysis*, 48 (2008) 533-553.

[2] G. Verreck, I. Chun, J. Peeters, J. Rosenblatt, M.E. Brewster, Preparation and Characterization of Nanofibers Containing Amorphous Drug Dispersions Generated by Electrostatic Spinning, *Pharmaceutical Research*, 20 (2003) 810-817.

[3] X. Meng, Y. Chen, S.R. Chowdhury, D. Yang, S. Mitra, Stabilizing dispersions of hydrophobic drug molecules using cellulose ethers during anti-solvent synthesis of micro-particulates, *Colloids and Surfaces B: Biointerfaces*, 70 (2009) 7-14.

[4] J.M. Amigo, J. Cruz, M. Bautista, S. Maspoch, J. Coello, M. Blanco, Study of pharmaceutical samples by NIR chemical-image and multivariate analysis, *TrAC Trends in Analytical Chemistry*, 27 (2008) 696-713.

[5] J.I. Jérez Rozo, A. Zarow, B. Zhou, R. Pinal, Z. Iqbal, R.J. Romañach, Complementary near-infrared and raman chemical imaging of pharmaceutical thin films, J*ournal of Pharmaceutical Sciences*, (2011).

[6] R. Tauler, Multivariate curve resolution applied to second order data, *Chemometrics and Intelligent Laboratory Systems*, 30 (1995) 133-146.

[7] A. de Juan, M. Maeder, T. Hancewicz and R. Tauler, Use of local-rank based spatial information for resolution of spectroscopic images, *Journal of Chemometrics*, 22 (2008) 291-298.

[8] J. Kalivas, Calibration Methodologies, in: S. Brown, Tauler, R, Walczak, B. (Ed.) *Comprehensive Chemometrics: Chemical and Biochemical data analysis*, Tauler & Walczak Eds. Amsterdam: Elsevier, 2009.

[9] D.M. Haaland, D.K. Melgaard, New augmented classical least squares methods for improved quantitative spectral analyses, *Vibrational Spectroscopy*, 29 (2002) 171-175.

[10] J.M. Prats-Montalbán, A. de Juan, A. Ferrer, Multivariate image analysis: A review with applications, *Chemometrics and Intelligent Laboratory Systems*, 107 (2011) 1-23.

[11] P. Geladi, H, Grahn, *Multivariate Image Analysis,* John Wiley & Sons Ltd.Chichester, England, 1996.

[12] A. de Juan, M. Maeder, T. Hancewicz, L. Duponchel, R. Tauler, Chemometric Tools for Image Analysis, in: *Infrared and Raman Spectroscopic Imaging*, Wiley-VCH Verlag GmbH & Co. KGaA, 2009, pp. 65-109.

[13] Burger, J. Bad pixel detection in hyperspectral staring camera systems, *NIR News*, 20 (1) (2009), 9-12.

[14] J. Jaumot, R. Gargallo, A. Dejuan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, *Chemometrics and Intelligent Laboratory Systems*, 76 (2005) 101-110.

[15] A. de Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault, M. Maeder, Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis, *TrAC Trends in Analytical Chemistry*, 23 (2004) 70-79.

[16] 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.

[17] J.E. Jackson. *A User's Guide to Principal Components*, John Wiley & Sons, Inc., 1991.

[18] M.H. Bharati, J.F. MacGregor, Texture analysis of images using principal component analysis, *SPIE/Photonics Conference on Process Imaging for Automatic Control*, Boston (2000), pp. 27–37, 2001.

[19] J.M. Prats-Montalbán, A. Ferrer, Integration of colour and textural information in multivariate image analysis: defect detection and classification issues, *Journal of Chemometrics*, 21 (2007) 10-23.

[20] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, *Analytica Chimica Acta*, 185 (1986) 1-17.

[21] P. Soille, *Morphological Image Analysis: Principles and Applications*, Springer, NY, 1999.

[22] R.M. Haralick, S.R. Sternberg, X. Zhuang, Image analysis using mathematical morphology, *IEEE Trans. Pattern Anal. Mach. Intell*, 9 (1987) 532-550.

[23] W. Yin, T. Chen, S.X. Zhou, A. Chakraborty, Background correction for cDNA microarray images using the TV+L1 model, *Bioinformatics*, 21 (2005) 2410-2416.

[24] J.M. Prats-Montalbán, A. Ferrer, R. Bro, T. Hancewicz, Prediction of skin quality properties by different Multivariate Image Analysis methodologies, *Chemometrics and Intelligent Laboratory Systems*, 96 (2009) 6-13.

[25] Rinnan, A; van der Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry* (2009) 28 (10) 1201-1222