HANDLING LARGE DATASETS OF HYPERSPECTRAL IMAGES: REDUCING DATA SIZE WITHOUT LOSS OF USEFUL INFORMATION

Carlotta Ferrari, Giorgia Foca, Alessandro Ulrici*

Department of Life Sciences - Interdepartmental Research Centre for Agri-Food Biological Resources Improvement and Valorization - University of Modena and Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122 Reggio Emilia, Italy

alessandro.ulrici@unimore.it

Abstract

HyperSpectral Imaging (HSI) is gaining increasing interest in the field of analytical chemistry, since this fast and non-destructive technique allows one to easily acquire a large amount of spectral and spatial information on a wide number of samples in very short times. However, the large size of hyperspectral image data often limits the possible uses of this technique, due to the difficulty of evaluating many samples altogether, for example when one needs to consider a representative number of samples for the implementation of on-line applications. In order to solve this problem, we propose a novel chemometric strategy aimed to significantly reduce the dataset size, which allows to analyse in a completely automated way from tens up to hundreds of hyperspectral images altogether, without losing neither spectral nor spatial information. The approach essentially consists in compressing each hyperspectral image into a signal, named *hyperspectrogram*, which is created by combining several quantities obtained by applying PCA to each single hyperspectral image. Hyperspectrograms can then be used as a compact set of descriptors and subjected to blind analysis techniques. Moreover, a further improvement of both data compression and calibration/classification performances can be achieved by applying proper variable selection methods to the hyperspectrograms. A visual evaluation of the correctness of the choices made by the algorithm can be obtained by representing the selected features back into the original image domain. Likewise, the interpretation of the chemical information underlying the selected regions of the hyperspectrograms related to the loadings is enabled by projecting them in the original spectral domain. Examples of applications of the hyperspectrogram-based approach to hyperspectral images of food samples in the NIR range (1000-1700 nm) and in the Vis-NIR range (400-1000 nm), facing a calibration and a defect detection issue respectively, demonstrate the effectiveness of the proposed approach.

Keywords

Hyperspectral Imaging; Data compression; Variable selection; Multivariate Image Analysis

1. Introduction

HyperSpectral Imaging (HSI), also known as hyperspectral chemical imaging (HCI), represents an emerging technique that provides both spatial information of imaging systems and spectral information of spectroscopy [1]. HSI techniques are based on the acquisition of spectral data not only from a single point but at each pixel of an image, to form a three-dimensional multivariate array of data (also called *hypercube*) with two spatial dimensions (x, y) and one wavelength dimension (λ). Therefore, compared with traditional spectroscopic methods, HSI allows not only to achieve identification and quantification of the chemical components within the analysed sample, but also to map their spatial distribution. Thanks to the possibility that this technique offers in describing heterogeneous samples by taking into account also spatial-related features, HSI has found a wide range of applications in several fields [2-6], in particular in pharmaceutical industry [7, 8] and in food industry [9, 10]. In these two fields, several studies have been carried out in order to address calibration [11-13], classification [14-16] as well as defects detection issues [17, 18].

Despite the many advantages provided by this technique, a wider diffusion of HSI is hampered by the high amount of data that can be collected in very short times, considering that hyperspectral images with file sizes of 50 MB and more can be easily acquired in few seconds. Indeed this represents a crucial point since, in the main fields of use, applications requiring the simultaneous evaluation of a large number of images would be highly valuable. This issue, also referred to as *curse* of dimensionality, has been recently addressed by Burger and Gowen in [19], where multivariate analysis methods available for reducing the computational load involved in acquiring and managing HSI data are reviewed. Several approaches for dimensionality reduction have been recently discussed also by Gowen et al. in [20] about time series HSI data, where several hyperspectral images acquired on the same or similar samples at different times must be evaluated in order to gain information about the phenomena underlying dynamic processes and/or for prediction of the future behaviour of systems. In both cases, among the reported approaches, particular attention was paid to latent variables projection-based methods and to wavelet decomposition.

Among the latent variables projection-based methods, Principal Component Analysis (PCA) is the most frequently used technique in the frame of Multivariate Image Analysis (MIA) [21, 22]. In this case, data reduction is achieved by unfolding the hypercube, which means reorganising it into a two-dimensional data matrix with size $\{(x \times y), \lambda\}$, and then in projecting the high dimensional data, i.e., the pixels data in the λ spectral dimensions, into a new subspace defined by a limited number of uncorrelated variables (Principal Components, PCs), describing the major variability sources of the analysed data. The same concept is applied in Multivariate Curve Resolution (MCR)

[23], where the unfolded hypercube is decomposed into two matrices, taking advantage of the Lambert-Beer law; the first matrix contains the spectra recovered for the pure chemical components and the second one the corresponding concentration profiles for each pixel.

The Wavelet Transform (WT) [24, 25] allows to represent each analysed spectrum or image in an alternative domain, where the different frequencies are separated, but maintaining at the same time the localisation in the original domain. This is known as signal/image multiresolution. In this manner, in addition to the single intensity values, other useful aspects like, e.g., band widths and slopes of a spectrum, or discontinuities, noise and uniform areas of an image can be extracted from the data and compressed into a limited number of variables (called wavelet coefficients). Wavelet analysis can be applied to HSI both in the image space (two-dimensional WT) and in the spectral domain (one-dimensional WT). A number of WT-based approaches have been developed specifically for hyperspectral image analysis; as an example, the hyperspectral discrete wavelet transform proposed by Scholl and Dereniak [26], consists of a 2-D discrete wavelet transform (DWT) in the spatial dimension carried out independently of a 1-D DWT in the spectral dimension. Burger and Gowen [19] report a comparison between the use of PCA and WT for the compression of an image containing 318×256 pixels and 131 wavelength channels, which led to compression ratio values equal to 7.6 % and 1.2 %, respectively.

Notwithstanding the great potential of these techniques, they generally allow the simultaneous analysis of a relatively restricted number of hyperspectral images, since merging together more than few images of different samples is a computationally intensive task. However, when dealing with problems related to samples characterized by a large inter-sample variability such in the case of food industry, where several factors (e.g. harvest period or animal feeding) concur in defining the final quality of the product, it is necessary to consider an adequate number of samples in order to describe the real variability of the considered problem; to this aim, datasets composed by hundreds of hypercubes should be handled. Nowadays, this is usually achieved by analysing separately each image, in order to extract data, such as average spectra of a user-defined Region Of Interest (ROI), to be used for further analysis of the whole dataset. However, this procedure results to be quite laborious, time consuming and strictly depending on the problem at hand. Moreover, when averaging spectra, information about spatial (inter-pixel) variability is lost. Conversely, by investigating simultaneously hundreds of hypercubes, it could be possible to gain an overview of the acquired dataset, to identify specific patterns, as well as to properly verify the representativeness of training and test samples to be used for further classification, calibration or process monitoring purposes.

In this context, we propose a chemometric strategy that was developed to significantly reduce the dataset size, allowing to analyse at the same time from tens up to hundreds of hyperspectral images. This procedure is derived from the *colourgrams* approach, already developed by some of us for the elaboration of RGB images [27-29]. The proposed approach essentially consists in compressing each hyperspectral image into a signal, named *hyperspectrogram*, which is created by combining several quantities obtained by applying PCA to the unfolded hypercube data. Hyperspectrograms can then be used as a compact set of descriptors and subjected to further blind analysis techniques.

Briefly, hyperspectrograms are obtained by merging in sequence the frequency distribution curves of the score vectors obtained from a PCA model calculated separately on each HSI, and by adding also the frequency distribution curves of the Q residuals and of the Hotelling T^2 vectors, in order to preserve all the pixel-related variability of the hypercube. Moreover, in order to maintain the most relevant spectral features of the hypercube data, the PC loading vectors are also added at the end of the signal.

Using proper variable selection methods, hyperspectrograms can be further compressed to few significant descriptors, allowing to extract only the specific features that are useful to solve the problem at hand. Additionally, these features can be projected back into the image space, allowing to perform a visual evaluation of the choices made by the feature selection method, or into the spectral domain, in order to detect the spectral regions containing the information of interest.

The idea to use the whole images instead of the single pixels as objects in the context of hyperspectral imaging has been recently proposed by Kucheryavsky [30]. In this work, the frequency distribution curves of the score values of each principal component obtained from a PCA model calculated on the whole dataset of hyperspectral images were used to build a feature vector for each object. Conversely, in the *hyperspectrogram* approach PCA models are calculated separately for each hypercube, thus allowing to consider a much higher number of hyperspectral images at the same time.

The proposed approach was tested on two benchmark datasets of hyperspectral images of food samples, acquired by means of two different instruments working in the NIR and in the Vis-NIR ranges, and addressing a calibration and a defect detection issue, respectively.

2. Materials and methods

2.1. Dataset 1: wheat and rice kernels

In order to perform a preliminary test of the efficacy of the proposed approach, using an example where the sources of variation in the images are well known a priori, a first benchmark dataset

(which is available from the authors upon request) was created by acquiring hyperspectral images of binary mixtures of wheat and rice kernels. In particular, 15 samples containing percentages of rice ranging from 0 to 100% (Table 1) were imaged using a desktop NIR Spectral Scanner (DV Optic), embedding a reflectance imaging based spectrometer Specim N17E and operating in the 900 – 1700 nm spectral range (spectral resolution 5 nm). In particular, two repeated and three replicate images were acquired for each one of the 14 samples where the wheat and rice kernels were uniformly mixed (samples A–I and M–Q). Moreover, two repeated images of a sample containing 50% wheat kernels and 50% rice kernels grouped separately by kernel type (sample L) were also acquired. All the 86 images were acquired using as sample background a black silicon carbide (SiC) sandpaper sheet. An instrument calibration based on a high-reflectance standard reference and on dark current [31] was applied to convert the raw data into the corresponding reflectance values. The reflectance images were then cropped to obtain equal spatial dimensions of 231 × 229 pixels and furthermore, due to the low S/N ratio of the spectra extremes, only the 150 central wavelengths between 955 and 1700 nm were considered for further analysis.

2.2. Dataset 2: buns with surface defects

In order to evaluate the presence of a surface defect typical of industrial buns, namely *pale spots*, which is rather difficult to detect by means of classical RGB imaging techniques, hyperspectral images of 10 buns showing pale spots were compared with 4 control samples and 6 buns affected by another defect (*dark spots*), which is instead easily detectable by RGB imaging. In this context it has to be underlined that, although the assignment of the samples to the three classes was performed by expert assessors, this evaluation cannot be considered as free from a certain degree of uncertainty, due to the high variability of the extent and intensity of the two defects. This means that, for example, some samples assigned to the control class could actually be affected by pale spots, but with too limited extent and intensity to justify their assignation to the defective class.

Three repeated images were acquired for each sample and furthermore replicate images were acquired on 3 samples in different days. The 78 resulting hyperspectral images were acquired using a Specim ImSpector V10E Imaging VisNIR System operating in the 400-1000 nm range (spectral resolution 2.9 nm). Due to the low S/N ratio of the spectra extremes, only the 189 central wavelengths between 450 and 999 nm were considered for further analysis. Also in this case, a black sandpaper sheet was used as sample background and the instrument calibration previously described was applied to convert the raw intensity data into the corresponding reflectance values.

2.3. Preprocessing of hyperspectral images

Before converting images into hyperspectrograms, both the datasets were subjected to an image segmentation step [32] aimed at removing the pixels corresponding to the sandpaper sheet used as background, which was present in all the images of both datasets. To this purpose, according to a generally recognized procedure in the frame of hyperspectral image analysis, and thanks to the neat difference in the reflectance values between sample and background pixels, a fast thresholding procedure was employed. In particular, based on the preliminary evaluation of some sample images, the most discriminant wavelength was identified for each dataset by maximising the Fisher ratio, which led to the use of λ = 1090 nm for *Dataset 1* (threshold value: 0.2) and λ = 889 nm for *Dataset 2* (threshold value: 0.6).

2.4. Creation of the hyperspectrograms

As mentioned above, the proposed approach is based on the idea of codifying the potentially useful information contained in each hyperspectral image into a signal, named hyperspectrogram, which is obtained by merging together quantities derived by a PCA model calculated on the unfolded hypercube data. A schematic representation of the procedure followed to generate the hyperspectrogram is reported in Figure 1.

More in detail, starting from a dataset formed by a large number of hyperspectral images, the calculation the hyperspectrogram corresponding to each single image involves the following steps:

- the three-dimensional hypercube $\underline{\mathbf{H}}$ with size $\{x, y, \lambda\}$, where x and y are the number of pixel rows and columns, respectively, and λ is the number of wavelengths, is unfolded to a two-dimensional matrix $\mathbf{X}$ with size $\{(x \times y), \lambda\}$, containing as many rows as the number of pixels, and as many columns as the number of wavelengths, λ;
- a PCA model is calculated on meancentered spectra with a user-defined number of PCs , A, which is the same for all the analysed images, and considering only the r pixels retained after image segmentation, i.e. $r \leq (x \times y)$; the corresponding score vectors $\boldsymbol{t_a}$ and loading vectors $\boldsymbol{p_a}$ (with $1 \leq a \leq A$), Q-residuals vector, $\boldsymbol{q}$, and Hotelling T^2 vector, $\boldsymbol{h}$, are stored;
- in order to avoid problems due to the sign indeterminacy of PCA decomposition, starting from the second analyzed image, for each principal component a the sign of each loading vector $\boldsymbol{p_a}$ is defined in a way that the sum of the squared differences with respect to the corresponding loading vector calculated for the first image is minimum, and the sign of the corresponding score vector $\boldsymbol{t_a}$ is defined accordingly;

- the frequency distribution curves of each score vector and of the Q residuals and Hotelling T^2 vectors are calculated, considering a number of bins equal to the number of spectral variables, λ; each frequency distribution curve is then normalized by the number of pixels retained after segmentation of the corresponding image (*r*). The range considered for the calculation of the frequency distribution curves of the score vectors, (stored in the corresponding data vectors $\boldsymbol{Ft_a}$,) is defined separately for each principal component on the basis of the minimum and of the maximum score values calculated over all the images. Similarly, for the frequency distribution curve of Q residuals, $\boldsymbol{Fq}$, and of Hotelling T^2 , $\boldsymbol{Fh}$, the corresponding range is defined between 0 and the maximum value calculated over all the images. No outlier elimination is done at this step, since the pixels lying outside the 95% or 99.7% confidence limits could correspond to useful features, e.g. to sample defects when a defect detection issue is faced;
- the hyperspectrogram of each image is then created by joining in sequence the frequency distribution curves of the scores vectors, of the Q residual vector and of the Hotelling T^2 vector, and finally adding the loading vectors. For example, if the number of user-defined PCs, *A*, is set to 2, the hyperspectrogram is obtained by joining in sequence the vectors $\boldsymbol{Ft_1}$, $\boldsymbol{Ft_2}$, $\boldsymbol{Fq}$, $\boldsymbol{Fh}$, $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$, and the resulting length will be equal to $(2 \times A + 2) \times \lambda$, i.e., for 2 PCs, to 6 λ.

It must be noticed that any possible source of non-informative variability existing among the different images, due to factors such as e.g., instrumental instability, can be eliminated or minimized previous to hyperspectrograms calculation by means of a proper internal calibration step [8, 16]. As for the datasets considered in the present study, a preliminary explorative data analysis by PCA revealed that no internal calibration procedure was necessary.

A further remark concerns the choice of the most appropriate pretreatment to use for the calculation of the PCA models on the individual images. Indeed, although in the present work the original images of both datasets have been pretreated only by meancentering, it must be underlined that other pretreatments can be used, if a preliminary evaluation made on single images indicates that these allow to better point out the features of interest. As for the appropriate number of PCs to be retained in the PCA models used for the hyperspectrograms calculation, it has to be underlined that this does not really represent a crucial point. Indeed, hyperspectrograms can be further subjected to a variable selection step where the PCs accounting for variability sources which are not useful for solving the problem at hand will be discharged. However, also in this case, a preliminary evaluation

by PCA on a restricted number of representative images can be very useful in order to have an estimate of the number of PCs potentially bringing useful information.

Anyway, including in the hyperspectrograms also the Q residuals ensures that all the information which is potentially useful for the problem at hand is considered, independently of the number of retained PCs. Furthermore, the inclusion of the frequency distribution curve of Hotelling T^2, though being partially redundant, could be useful when a particular feature of interest is characterized by the simultaneous contribution of more PCs [22].

Concerning the time required to segment and to convert a set of hyperspectral images into the corresponding hyperspectrograms, as an example the calculation of the 86 hyperspectrograms of *Dataset 1* (overall size equal to 3.5 GB) using a personal computer running with Microsoft Windows 7–64 bit ® and equipped with an Intel Core ® i7-2600 CPU @ 3.40 GHz processor and 4.00 GB RAM required 227.84 seconds (2.65 s for image), including the time needed to load each image file from the hard disk, to segment the image and to save the resulting matrix of hyperspectrograms.

2.5. Explorative analysis of hyperspectrograms

As a first step, before calculating calibration/classification models, the matrices of hyperspectrograms of both the datasets have been subjected to explorative analysis by means of PCA, in order to obtain an overview of the whole structure of each dataset and to identify possible outlier samples. Moreover, PCA also helped us in understanding the effects of the different hyperspectrogram pretreatments on the resulting score plots. Indeed, considering that the pixel-related quantities are reported as frequency distribution curves, it should be taken into account that preprocessing by autoscaling leads to an enhancement of the contribution of those bins accounting for a small percentage of pixels. The effect of autoscaling could be therefore not very useful when dealing with classification or calibration issues based on general features of the samples, while on the other hand it could be helpful when a defect detection issue has to be faced. In this latter case, in fact, the classification is mainly based on few pixels that differ from the remainder ones, and which correspond to low peaks lying at extreme values of the frequency distribution curves of the hyperspectrograms.

In the light of these considerations, while in the case of *Dataset 1* the effects of all the main column pretreatments (none, meancentering and autoscaling) were investigated in order to examine their effects on the resulting models, the hyperspectrograms of *Dataset 2* were pretreated by autoscaling, since in this case the classification issue consisted in the detection of surface defects.

2.6. PLS on Dataset 1

In order to obtain a first estimate of the capability of hyperspectrograms to codify the useful information contained in the hyperspectral images, Partial Least Squares (PLS) regression models were developed to predict the mass fraction (% w/w) of rice kernels contained in each hyperspectral image of *Dataset 1* using hyperspectrograms. To this aim, samples were divided into a training of 48 signals and a test set of 38 signals corresponding to the samples composition reported in Table 1, and replicate measurements were included in the same set, in order to avoid overoptimistic results. Moreover, a customized cross-validation vector with 8 deletion groups was created in order to force the algorithm to keep the replicate measurements of each training set sample in the same deletion group. The effect of the different pretreatments described in the previous section was evaluated by comparing the respective calibration performances, and the best pretreatment was selected according to the lowest value of the Root Mean Square Error of Cross-Validation (RMSECV).

2.7. PLS-DA on Dataset 2

The pale spots defect of industrial bun samples of *Dataset 2* was detected by applying Partial Least Squares-Discriminant Analysis (PLS-DA) [33-35] to the autoscaled hyperspectrograms.
To this purpose, 2/3 of images were randomly assigned to the training set and the remaining 1/3 of images were kept in the test set, always including the replicate images of each sample in the same set. A customised cross-validation vector with 13 deletion groups was used, forcing the algorithm to keep the replicate measurements of each bun sample in the same group. The optimal number of Latent Variables (LVs) was chosen on the basis of the minimum value of the Root Mean Square Error in Cross-Validation (RMSECV). The classification results are reported, both in cross-validation and in prediction on the external test set, in terms of Efficiency %, which is the geometric mean of Sensitivity %, (the percentage of objects of the modelled class correctly accepted by the class model) and Specificity % (the percentage of objects of other classes correctly rejected by the class model).

2.8. Variable selection by means of interval PLS (iPLS) and interval PLS-DA (iPLS-DA)

As mentioned above, a further advantage of the proposed approach is represented by the possibility to apply variable selection methods to the hyperspectrograms, which may often allow to enhance

the performance and the robustness of the calibration/classification models by discharging non-informative or non-significant parts of the signals; moreover, variable selection also offers the possibility to obtain a better understanding of the problem at hand, by evaluating the selected signal regions. In the specific case of hyperspectrograms, it is possible to project back the selected portions of the loadings in the original spectral range, as well as to fold back the selected regions of the pixel-related parts of the hyperspectrogram (i.e., the frequency distribution curves) to visually evaluate the spatial features considered in the model.

Among the several existing methods for variable selection [4], in the present work the simple but effective iPLS and iPLS-DA algorithms were applied to *Dataset 1* and *Dataset 2*, respectively. As described in [36], iPLS works by dividing the whole signal in a user-defined number of intervals of equal width, and then by selecting the intervals most useful for calibration by an iterative procedure, which can follow either a *forward* or a *reverse* search strategy. More in detail, forward iPLS is conceived to calculate local PLS models on each subinterval, then to choose the best one on the basis of the lowest RMSECV value. In the second cycle, the first selected interval is used in all models but is combined with each of the remaining intervals one at a time, and the best combination of the two intervals is chosen again on the basis of the lowest RMSECV value. This iterative procedure is repeated until no further decrease of RMSECV is achieved. The reverse iPLS, on the contrary, works by initially including all the intervals in the model, then by discarding a single interval at a time. When discarding a certain interval produces the lowest RMSECV value, that interval is definitively excluded from the model. The same procedure is repeated by discarding the second "worst" interval and so on until no further decrease of the RMSECV values is obtained.

In the present work, the forward selection mode was used since it is the less conservative one, i.e., a smaller number of wavelengths are usually preserved in the final model when using forward selection with respect to the reverse mode. Concerning the interval size to be considered for variable selection, two approaches were applied to both datasets. In the first case, an interval size equal to the number of spectral variables, λ, was used in order to sequentially add a whole hyperspectrogram block. In the second case, a more refined selection was performed by considering narrower intervals so as to enable the selection of only the most informative portions within a block. In particular, iPLS with interval sizes of 150 and 10 variables and iPLS-DA with interval sizes of 189 and 18 variables were applied to *Dataset 1* and *Dataset 2*, respectively.

2.9. Image reconstruction using the selected spatial features

Despite the several advantages mentioned above, a main concern about the application of the proposed approach might be related to the loss of spatial (scene-related) information, due to the reduction of a hyperspectral image into a signal. To address this issue, a dedicated routine was developed to allow the representation in the original image domain of the hyperspectrogram features that have been selected (e.g. by iPLS), thus enabling a visual evaluation of the correctness of the choices made by the algorithm, in a similar way as it is done with colourgrams for RGB images [28, 29]. A schematic representation of the procedure followed to perform image reconstruction using the selected spatial features is reported in Figure 2.

More in detail, the image reconstruction procedure can be summarised in the following steps:

1. for each frequency distribution vector included in the hyperspectrogram, i.e., for each score frequency distribution vector $\boldsymbol{Ft_a}$ (with $1 \leq a \leq A$, were A is the number of PCs retained in the PCA models used for the hyperspectrograms calculation, *see* Section 2.4), for the Q residuals frequency distribution vector $\boldsymbol{Fq}$, and for the Hotelling T^2 frequency distribution vector $\boldsymbol{Fh}$, store the values related to the hyperspectrogram portions selected by iPLS or by iPLS-DA into the corresponding matrices of selected intervals $\boldsymbol{Int\_Ft_a}$, $\boldsymbol{Int\_Fq}$ and $\boldsymbol{Int\_Fh}$. Each matrix of selected intervals has as many columns as the number selected intervals, *j*, and each column contains the first and the last value of the selected interval in the first and in the second row, respectively. For example, if iPLS led to the selection of three intervals of the frequency distribution vector of PC2 scores, $\boldsymbol{Ft_2}$, in correspondence with the $\boldsymbol{t_2}$ values ranging from 10 to 20, from 25 to 35 and from 70 to 80, the resultant $\boldsymbol{Int\_Ft_2}$ matrix will be:

$$\boldsymbol{Int\_Ft_2} = \begin{bmatrix} 10 & 25 & 70 \\ 20 & 35 & 80 \end{bmatrix} \qquad (1)$$

2. from each matrix containing a number $j > 0$ of selected intervals, create the corresponding vector of selected pixel values $\boldsymbol{Sel\_t_a}$, $\boldsymbol{Sel\_q}$, $\boldsymbol{Sel\_h}$. For example, from the $\boldsymbol{Int\_Ft_2}$ matrix reported in equation (1), the corresponding vector of selected pixel values $\boldsymbol{Sel\_t_2}$ is given by:

$$\boldsymbol{Sel\_t_2} = \boldsymbol{t_2} \in \boldsymbol{Int\_Ft_2} \qquad (2)$$

   i.e., it contains only those elements of $\boldsymbol{t_2}$, whose values are included within the intervals specified in $\boldsymbol{Int\_Ft_2}$;

3. normalize each vector of selected pixel values between 0 and 1, considering the minimum and maximum values of the corresponding matrix of selected intervals. In the example reported above, the $\boldsymbol{Sel\_t_2}$ values are scaled considering the maximum and minimum values of $\boldsymbol{Int\_Ft_2}$, i.e., 10 and 80;

4. represent each vector of selected pixel values as a greyscale or pseudo-colour image with size $\{x, y\}$, i.e., with the same number of pixel rows x and pixel columns y as the original hyperspectral image **H** (as it has been defined in Section 2.4); all the pixels that have not been selected are set to NaN (Not a Number). Alternatively, up to three different vectors of selected pixel values can be represented altogether under the form of false-colour images, as reported in Figure 2.

The above described procedure is used when the feature selection has led to retain the pixel-related part of the hyperspectrogram (i.e. portions of the frequency distribution curves). Conversely, when the wavelength-related part of the hyperspectrograms (i.e., portions of the loading vectors) is selected, the easiest way to represent in the image domain the most relevant features of the problem at hand consists in calculating a PCA model of the hypercube data, where only the variables corresponding to the selected regions of the loading vectors are kept, and in representing the resultant score images.

3. Results and discussion

3.1. Dataset 1: wheat and rice kernels

Based on the preliminary indications obtained by PCA on some sample images, 3 PCs were used for the calculation of the hyperspectrograms. Each original hypercube consisting of more than 3×10^6 data points was therefore compressed into a 1200-points long hyperspectrogram (= 150 × 3 points for the frequency distribution curves of the 3 score vectors + 150 points for the frequency distribution curve of the Q residuals vector + 150 points for the frequency distribution curve of the Hotelling T^2 vector + 150 × 3 points for the 3 loading vectors). The average hyperspectrogram is reported in Figure 3, and the description of all the corresponding peaks is given in Table 2. On the whole, the initial dataset of 86 images with a size of 3.5 GB (average image size equal to about 40 MB) was compressed into a matrix of hyperspectrograms whose size was equal to 602 KB, which corresponds to a compression ratio of 1.66×10^{-2} %. This extremely low value of the compression ratio is essentially due to the fact that hyperspectrograms codify the useful information contained in the HSI data, but the localization of each single pixel in the image domain is lost. In fact, the main focus of the hyperspectrogram approach is to allow the evaluation of big datasets of hyperspectral images altogether, and not the reconstruction of the single images directly from the compressed data, as it can be done using other compression methods like, e.g., those based on Wavelet Transform. However, notwithstanding the transformation of HSI data into hyperspectrograms

implies the loss of spatial information, it is still possible to represent the selected features back into the original image domain.

The PCA model calculated on the meancentered dataset of hyperspectrograms was found to have an optimal dimensionality equal to 2 PCs, accounting for about 80% of the total variance; no outliers were identified considering the confidence limit of 99.7%. The PC1-PC2 score plot (Figure 4) showed the existence of a clear correlation between each image and the mass fraction of rice actually contained in the corresponding sample. Similar patterns were also observed when considering the PCA models calculated on raw and on autoscaled hyperspectrograms (data not shown).

The results of the calibration models calculated on raw, meancentered and autoscaled data are summarized in Table 3. Although similar calibration performances were obtained using the three pretreatments, the best performance in cross-validation was obtained using meancentering. The variable selection by means of iPLS considering 8 and 120 intervals was therefore applied to the meancentered signals. In both cases, the variable selection converged to the PC2 loadings region, and calibration performances equivalent to those obtained using the whole signal were obtained (Table 3). In particular, iPLS with window size 10 led to the selection of a unique interval corresponding to the PC2 loadings between 1405 and 1450 nm (highlighted in Figure 3), ascribable to the first overtone of the O-H stretching vibration and related to the starch content [37]. The 3 LVs calibration model calculated using the selected variables resulted in a R^2 in cross-validation equal to 0.9896 and in a R^2 in prediction of the external test set equal to 0.9718 (Figure 5). In order to visualize how the selected features retain the information related to the mass fraction of rice, a unique hyperspectral image was created by merging together a 100% wheat kernels image, a 100% rice kernels image and an image showing 50% wheat kernels and 50% rice kernels spatially separated. Two false-colour images were obtained by superimposing the PC1 and the PC2 score images resulting by the PCA models (meancentered spectra), calculated using both the whole spectral range (Figure 6a) and the ten selected variables only (Figure 6b). The comparison of the false-colour images points out an equivalent distinction of wheat and rice kernels, confirming that the information related to the selected variables was actually sufficient to discriminate wheat and rice kernels.

3.2. Dataset 2: buns with surface defects

The results of preliminary PCA models calculated on a restricted number of images of industrial buns showed that the number of significant PCs was always equal to 2, accounting for more than

99% of the total variance for all the analysed images. The hyperspectrograms were therefore created considering 2 PCs, to give a 1134 points-long signal for each image (= 189 × 2 points for the frequency distribution curves of the 2 score vectors + 189 points for the frequency distribution curve of the Q residuals vector + 189 points for the frequency distribution curve of the Hotelling T^2 vector + 189 × 2 points for the 2 loading vectors) as shown in Figure 7a. On the whole, the initial dataset having size of 7.32 GB was reduced to a matrix of hyperspectrograms whose size was equal to 477 KB, which corresponds to a compression ratio of 6.52×10^{-3} %. As it was mentioned above, during this step it was not performed any elimination of outlier pixels, since this could lead to the elimination of useful information related to sample defects. For example, Figure 8 reports the results of a PCA model calculated on a hyperspectral image of a sample showing pale spots, where the pixels lying outside the 95% confidence limits of the Hotelling T^2 values have been highlighted in magenta in the Q vs. T^2 plot (Figure 8a). The position of these outlying pixels corresponds to the defective portions of the sample surface, as one can see by comparing the Hotelling T^2 image (Figure 8b) with the RGB image of the same sample (Figure 8c).

The hyperspectrograms dataset was firstly analysed by PCA (model calculated on autoscaled variables), and no outliers were identified considering the 99.7% confidence limits for Q and Hotelling T^2. Then, in order to properly validate the PLS-DA classification models, the dataset of hyperspectrograms was split into a training set of 57 signals (corresponding to 13 bun samples) and a test set of 21 signals (corresponding to 7 bun samples). The PLS-DA model calculated on the autoscaled signals (3 LVs) led to classification efficiency values equal to 82.40% in cross-validation and to 100% in prediction of the test set. In order to check whether a simpler approach could lead to similar results, the performance of this PLS-DA model was compared with the performance of an analogous PLS-DA model calculated on mean spectra. To this aim, the mean spectrum of each segmented hypercube was computed obtaining a matrix with size {78, 189}. Then, this matrix was divided in a training and in a test set in the same way as for the hyperspectrogram matrix, and PLS-DA models were calculated considering mean centering as well as autoscaling as spectra pretreatments. In both cases, however, the classification models led to unsatisfactory results in terms of classification performances; in fact, the classification efficiency values in cross-validation and in prediction of the test set were equal to 66.69% and 52.67% using mean-centering and to 73.05% and 55.26% using autoscaling.

Concerning the use of variable selection on hyperspectrograms, when considering 150 variables-wide intervals the iPLS-DA algorithm led to discard the frequency distribution curve of PC1 scores as well as PC1 loadings, resulting in classification efficiency values of 90.28% and 94.29% in cross-validation and in prediction on the external test set, respectively (2 LVs). The use of 18

variable wide intervals led to the selection of the 180 variables highlighted in gray in Figure 7b, and the corresponding classification model (3 LVs) resulted in classification efficiency values equal to 100% both in cross-validation and in prediction of the external test set, as shown in Figure 9.

Among the hyperspectrogram portions selected by this latter iPLS-DA model, 18 variables were selected in the PC2 loading region corresponding to the spectral range between 529 and 579 nm, i.e., in the green colour region of the visible spectrum. As for the regions selected in the pixel-related part of the hyperspectrogram, their visual evaluation was made possible by building false-colour images showing the regions related to PC1 in the red channel, those related to PC2 in the green channel and those related to the Q residuals in the blue channel. As an example, the false-colour images obtained for a “dark spots”, a “control” and a “pale spots” samples are shown in Figure 10b, 10d and 10f, respectively. The comparison of these reconstructed images with the corresponding RBG ones, which are reported in Figure 10a, 10c and 10e, respectively, allows to interpret the choices made by the automated selection procedure. In fact, it can be noticed that the regions selected within the frequency distribution curve of PC1 account for the shape and the average colour of the samples, by selecting a ring of pixels (represented in red in the false-colour image) likely at equal sample height. The fact that an interval is selected on the frequency distribution curve of PC1 scores while no intervals are selected on the PC1 loadings is likely due to the fact that the useful features of PC1 are not related to particular chemical aspects, therefore localized in specific spectral regions, but to the average intensity of the whole spectrum. The hyperspectrogram regions selected within the frequency distribution curve of the Q-residuals correspond to the pixels characterized by the lowest Q values. By evaluating the spatial distribution of these (blue) pixels it can be noticed that they are mainly located in correspondence of the darkest regions of the sample surface, which are mostly due to shadow effects related to the shape of the sample itself. The most interesting features, namely the defective areas of the “pale spots” samples, are instead highlighted by the regions selected on the frequency distribution curve of PC2, showed as green pixels in Figure 10f, although large portions of the surfaces of samples not belonging to this class are also selected (Figures 10b and 10d). This can be explained by comparing the average frequency distribution curves of PC2 of the hyperspectrograms obtained for the “dark spots”, “pale spots” and the “control” samples reported in Figure 11a. In fact, a more in depth evaluation of the selected region which is most closely related to the “pale spots” detection issue (i.e., the portion of curve in Figure 11a within the red and green rectangles), reveals that it shows the largest difference of the “pale spots” curve with respect to the others, and at the same time the smallest difference between the “dark spots” and the “control” curves. Moreover, it can be observed that in the left part of this region (i.e. within the red rectangle) the number of pixels (corresponding to the area under

the curve) of the “control” and “dark spots” samples is much greater than the number of pixels for the “pale spots” samples. This can be explained comparing the lower colour homogeneity of the “pale spots” samples with respect to the control samples, but also with respect to the dark spots samples, whose defect is more localised and therefore affects a lower number of pixels. The lower colour homogeneity of the pale spots samples is reflected in turn into a broader shape of the PC2 frequency distribution curve, with lower values in the central part of the peak. In other words, the pixels falling within this region (i.e. the red pixels in Figures 11b, 11c and 11d) are those corresponding to a more homogeneous colour of the bun crust.

Conversely, when folding back in the original spatial domain the right part of the selected region (i.e. the region within the green rectangle in Figure 11a), it can be noticed that this one identifies the defective areas of the “pale spots” samples (green pixels in Figure 11d), characterized by higher values of PC2 scores. It can be noticed that a more limited amount of pixels corresponding to this part of the signal is also present within the reconstructed images of the “dark spots” and “control” samples reported in Figure 11b and 11c. Actually, also these samples present on their surface some regions characterized by a slightly pale aspect, which is emphasized in the green channel of the reconstructed images; however, in this case the intensity and extent of the defect is much more limited. Therefore this type of representation, by enhancing the presence of the sought defect, could be helpful for the quality control personnel, in addition to the output of the automated classification model.

4. Conclusions

In this paper, we have presented a novel chemometric strategy for efficient data compression of hyperspectral images. By compressing each hypercube into a signal of few hundreds of points, the proposed method enables the simultaneous evaluation of up to hundreds of hyperspectral images. The hyperspectrogram approach allows therefore the calculation of robust classification models, since it is possible to consider large datasets of samples. Moreover, a further improvement both in terms of data compression and of performance of the derived calibration/classification models can be achieved by applying a proper variable selection method to the dataset of hyperspectrograms.

A critical evaluation of the choices made by the feature selection algorithm is made possible by projecting back into the original image domain the pixel-related features of the hyperspectrograms retained during the variable selection step. Likewise, the interpretation of the chemical information underlying the wavelength-related part of the hyperspectrograms is enabled by projecting the corresponding selected features in the original spectral domain.

The use of hyperspectrograms to face two different issues concerning food samples of different nature confirmed the effectiveness of the proposed approach.

5. Acknowledgements

East Balt Italia is gratefully acknowledged for providing the bun samples and for furnishing useful information about buns defects.

References

[1] H. Grahn, P. Geladi, Techniques and Applications of Hyperspectral Image Analysis, John Wiley & Sons Ltd., Chichester, 2007.

[2] L. Cséfalvayová, M. Strlič, H. Karjalainen, Anal. Chem. 83 (2011) 5101-5106.

[3] D. Goltz, M. Attas, G. Young, E. Cloutis, M. Bedynski, J. Cult. Herit. 11 (2010) 19-26.

[4] R. Gosselin, D. Rodrigue, C. Duchesne, Chemom. Intell. Lab. Syst. 100 (2010) 12-21.

[5] G. Payne, C. Wallace, B. Reedy, C. Lennard, R. Schuler, D. Exline, C. Roux, Talanta 67 (2005) 334-344.

[6] G. Sciutto, P. Oliveri, S. Prati, M. Quaranta, S. Bersani, R. Mazzeo, Anal. Chim. Acta 752 (2012) 30-38.

[7] J.M. Amigo, Anal. Bioanal. Chem. 398 (2010) 93-109.

[8] C. Gendrin, Y. Roggo, C. Collet, J. Pharmaceut. Biomed. 48 (2008) 533–553.

[9] A.A. Gowen, C.P.O. Donnell, P.J. Cullen, G. Downey, J.M. Frias, Trends Food Sci. Tech. 18 (2007) 590-598.

[10] W. Wang, J. Paliwal, Sens. Instrum. Food Quality Saf. 1 (2007) 193-207.

[11] J. Burger, P. Geladi, The Analyst. 131 (2006) 1152-60.

[12] G. ElMasry, N. Wang, A. ElSayed, M. Ngadi, J. Food Eng. 81 (2007) 98-107.

[13] P. Rajkumar, N. Wang, G. EImasry, G.S.V. Raghavanb, Y. Gariepy, J. Food Eng. 108 (2012) 194-200.

[14] G. ElMasry, A. Iqbal, D.-W. Sun, P. Allen, P. Ward, J. Food Eng. 103 (2011) 333-344.

[15] P. Williams, P. Geladi, G. Fox, M. Manley, Anal. Chim. Acta 653 (2009) 121-130.

[16] A. Ulrici, S. Serranti, C. Ferrari, D. Cesare, G. Foca, G. Bonifazi, Chemom. Intell Lab. Syst., 122 (2013) 31-39.

[17] J. Li, X. Rao, Y. Ying, Comput. Electron. Agr. 78 (2011) 38-48.

[18] D.P. Ariana, R. Lu, Comput. Electron. Agr. 74 (2010) 137-144.

[19] J. Burger, A.A. Gowen, Chemom. Intell. Lab. Syst. 108 (2011) 13-22.

[20] A.A. Gowen, F. Marini, C. Esquerre, C. O'Donnell, G. Downey, J. Burger, Anal. Chim. Acta. 705 (2011) 272-82.

[21] P. Facco, A. Masiero, A. Beghi, J. Process Contr. 23 (2013) 89-98.

[22] J.M. Prats-Montalbán, A. de Juan, A. Ferrer, Chemom. Intell. Lab. Syst. 107 (2011) 1-23.

[23] A. de Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault, M. Maeder, TrAC – Trend Anal. Chem. 23 (2004) 70-79.

[24] B. Walczak, Wavelets in Chemistry, first ed., Elsevier, Amsterdam, 2000.

[25] M. Cocchi, R. Seeber, A. Ulrici, J. Chemom. 17 (2003) 512–527.

[26] J.F. Scholl, E.L. Dereniak, Proceedings of SPIE – Internat. Soc. Opt. Eng. 5208 (2004) 129-140.

[27] A. Antonelli, M. Cocchi, P. Fava, G. Foca, G.C. Franchini, D. Manzini, A. Ulrici, Anal. Chim. Acta 515 (2004) 3-13.

[28] G. Foca, F. Masino, A. Antonelli, A. Ulrici, Anal. Chim. Acta. 706 (2011) 238-245.

[29] A. Ulrici, G. Foca, M.C. Ielo, L.A. Volpelli, D.P. Lo Fiego, Innov. Food Sci. Emerg. Technol., 16 (2012) 417-426.

[30] S. Kucheryavsky, Chemom. Intell. Lab. Syst., 120 (2013) 126-135.

[31] J. Burger, P. Geladi, J. Chemom. 19 (2005) 355-363.

[32] M. Vidal, J.M. Amigo, Chemom. Intell. Lab. Syst. 117 (2012) 138-148.

[33] S. Chevallier, D. Bertrand, A. Kohler, P. Courcoux, J. Chemom. 20 (2006) 221-229.

[34] L. Pigani, A. Culetu, A. Ulrici, G. Foca, M. Vignali, R. Seeber, Food Chem. 129 (2011) 226-233.

[35] E. Ferrari, G. Foca, M. Vignali, L. Tassi, A. Ulrici, Anal. Chim. Acta. 701 (2011) 139-151.

[36] L. Nørgaard, A. Saudaland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413-419.

[37] J.S. Shenk, J.J. Workman, M.O. Westerhaus, in: D.A. Burns, E.W. Ciurczak (Eds.), Handbook of Near-Infrared Analysis, third ed., Marcel Dekker, New York, US, 2008, p. 356.

Tables

Sample name	Wheat amount (g)	Rice amount (g)	Rice mass fraction (% w/w)	Training / Test
A	10.04	0.00	0	Training
B	9.90	0.10	1	Training
C	9.82	0.23	2	Test
D	9.49	0.52	5	Training
E	9.03	1.00	10	Test
F	8.09	2.03	20	Training
G	7.02	3.03	30	Test
H	6.04	4.01	40	Training
I	5.00	5.00	50	Test
L	4.99	5.00	50	Test
M	4.04	6.00	60	Training
N	3.05	7.03	70	Test
O	2.01	8.00	80	Training
P	1.00	9.01	90	Test
Q	0.00	10.02	100	Training

Table 1: List of the samples included in *Dataset 1*, and their subdivision into training and test sets.

Peak number	Hyperspectrogram region	Definition
1	1-150	Frequency distribution curve of the 1st score vector
2	151-300	Frequency distribution curve of the 2nd score vector
3	301-450	Frequency distribution curve of the 3rd score vector
4	451-600	Frequency distribution curve of the Q-residual score vector
5	601-750	Frequency distribution curve of the Hotelling T^2 score vector
6	751-900	Normalised loading vector of the 1st PC
7	901-1050	Normalised loading vector of the 2nd PC
8	1051-1200	Normalised loading vector of the 3rd PC

Table 2: Description of the peaks present in the hyperspectrograms of Dataset 1, together with their relative positions (hyperspectrograms derived by a 3-PCs model calculated on an image with 150 spectral variables).

Pretreatment	# of variables	LVs	Calibration		Cross-validation		Prediction	
			R^2	RMSE	R^2	RMSE	R^2	RMSE
None	1200	4	0.9971	0.0192	0.9851	0.0508	0.9782	0.0480
Meancenter	1200	3	0.9971	0.0193	0.9855	0.0504	0.9774	0.0486
Autoscale	1200	3	0.9980	0.0161	0.9794	0.0563	0.9816	0.0445
Meancenter	150	3	0.9943	0.0271	0.9825	0.0495	0.9676	0.0587
Meancenter	10	3	0.9933	0.0294	0.9896	0.0368	0.9718	0.0524

Table 3: Results of the PLS models calculated on raw, meancentered and autoscaled data and of the iPLS models calculated on meancentered data.

Captions of figures

Figure 1. Procedure followed to generate the hyperspectrogram.

Figure 2. Image reconstruction using the hyperspectrogram selected spatial features.

Figure 3. *Dataset 1* average hyperspectrogram with the region selected by iPLS highlighted in gray. The numbers on the top of the figure indicate the hyperspectrogram regions described in Table 2.

Figure 4. PC1 vs. PC2 score plot obtanied from the PCA model on the mean centered hyperspectrograms of *Dataset 1*.

Figure 5. Actual mass fraction of rice (Y measured) vs. predicted mass fraction (Y predicted) resulting from the iPLS model calculated on *Dataset 1*.

Figure 6. False-colour image formed by the PC1 and PC2 score images resulted from the PCA model calculated using the whole range (a) and the selected variables only (b) on an image formed by merging together a 100% wheat image (left), a 100% rice image (middle) and a 50% wheat/50% rice image (right).

Figure 7. Hyperspectrograms obtained on *Dataset 2* (a) and variables selected by iPLS-DA highlighted in gray on the average signal (b).

Figure 8. Results of a 2 PCs model calculated on a hyperspectral image of a sample with pale spots: Q vs. T^2 plot (a) and Hotelling T^2 image (b). The pixels lying outside the 95% confidence limits of T^2 are highlighted in magenta. For comparison purposes, the RGB image of the same sample is also reported in (c).

Figure 9. Predicted values for the iPLS-DA model calculated on hyperspectrograms of Dataset 2. The vertical dashed line separates the cross-validation results for the training set samples (on the left) from the values predicted for the test set ones (on the right). The threshold value is indicated with the horizontal dash-dotted line.

Figure 10. Comparison between RGB images (left) and the corresponding false-colour reconstructions (right) of the hyperspectrograms selected features, where the red, green and blue channels account for the features selected for PC1, PC2 and Q, respectively; (a) and (b): “dark spots” sample; (c) and (d): “control” sample; (e) and (f): “pale spots” sample.

Figure 11. Average frequency distribution curves of the PC2 scores of the “dark spots”, “pale spots” and “control” samples (a), and false-colour images of a “dark spots” sample (b),

of a “pale spots” sample (c) and of a “control” sample (d). Images (b), (c) and (d) report in the red and green channels the pixels falling within the PC2 intervals of image (a) included in the red and in the green rectangles, respectively.

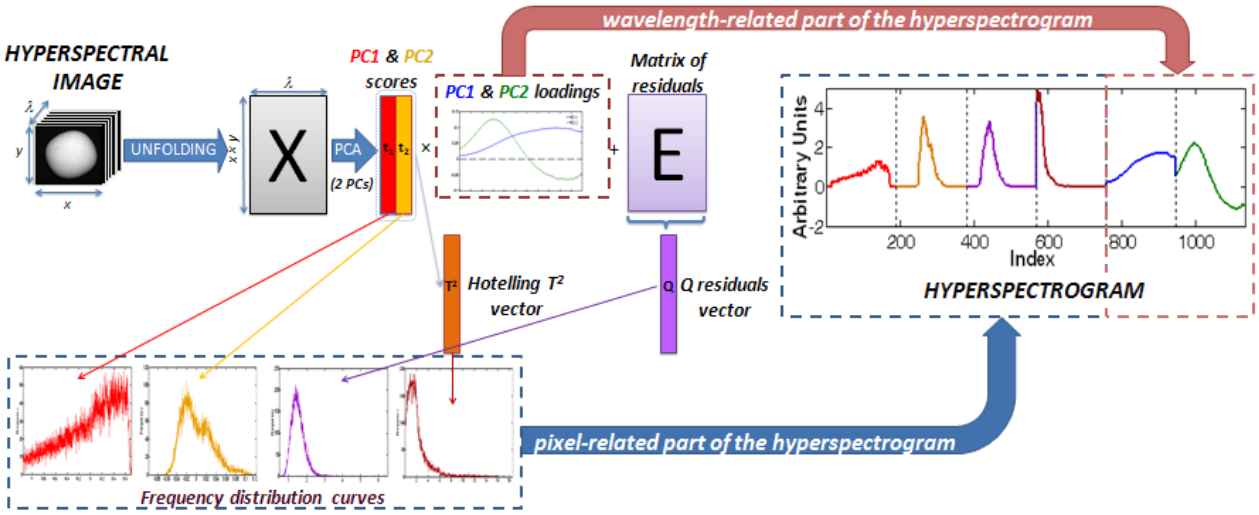
wavelength-related part of the hyperspectrogram
HYPERSPECTRAL IMAGE
UNFOLDING
X
PCA
(2 PCs)
PC1 & PC2 scores
PC1 & PC2 loadings
Matrix of residuals
E
Hotelling T² vector
Q residuals vector
Arbitrary Units
Index
HYPERSPECTROGRAM
pixel-related part of the hyperspectrogram
Frequency distribution curves

Figure 1

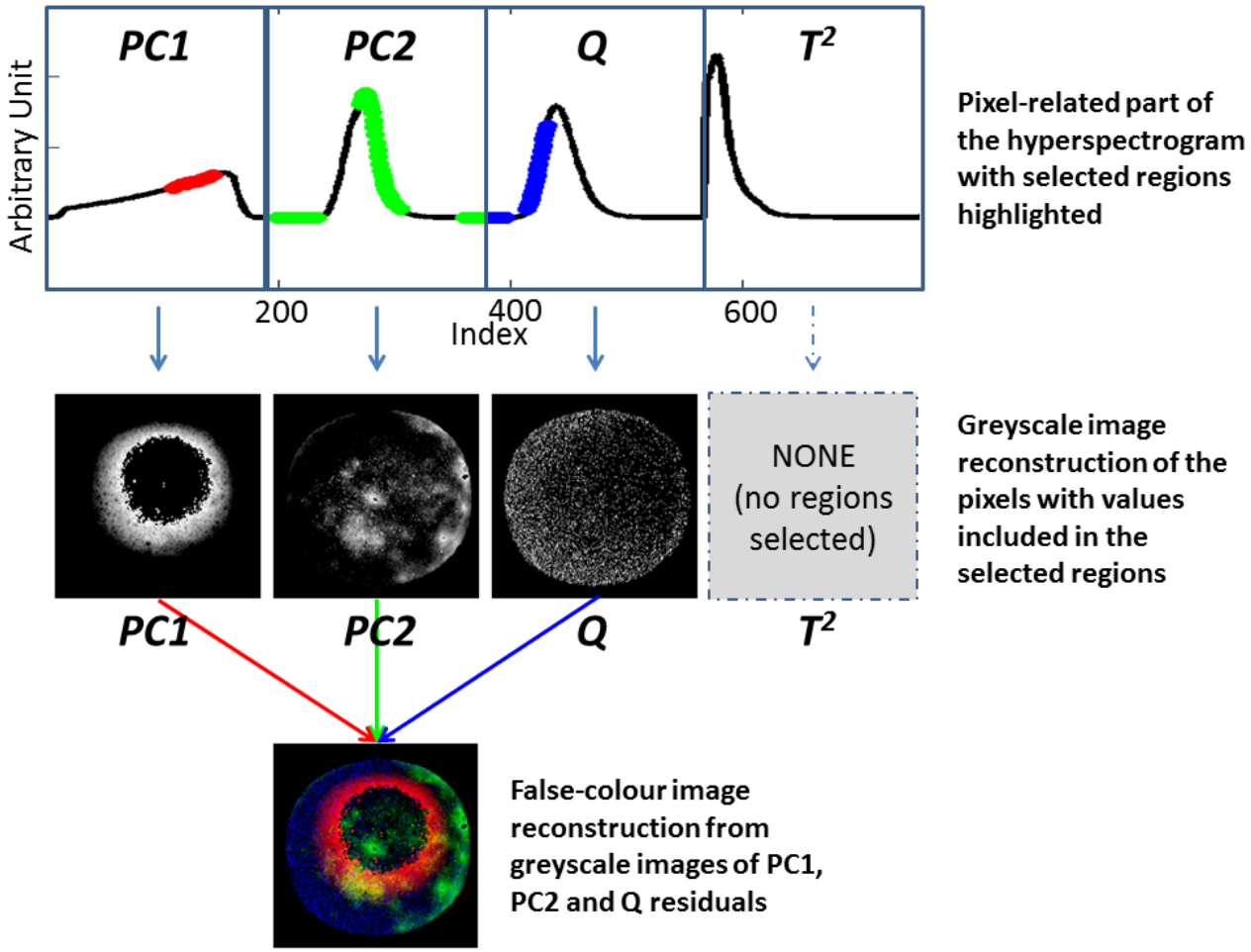
PC1
PC2
Q
T2
Arbitrary Unit
200
400
600
Index
Pixel-related part of the hyperspectrogram with selected regions highlighted
NONE
(no regions selected)
Greyscale image reconstruction of the pixels with values included in the selected regions
PC1
PC2
Q
T2
False-colour image reconstruction from greyscale images of PC1, PC2 and Q residuals

Figure 2

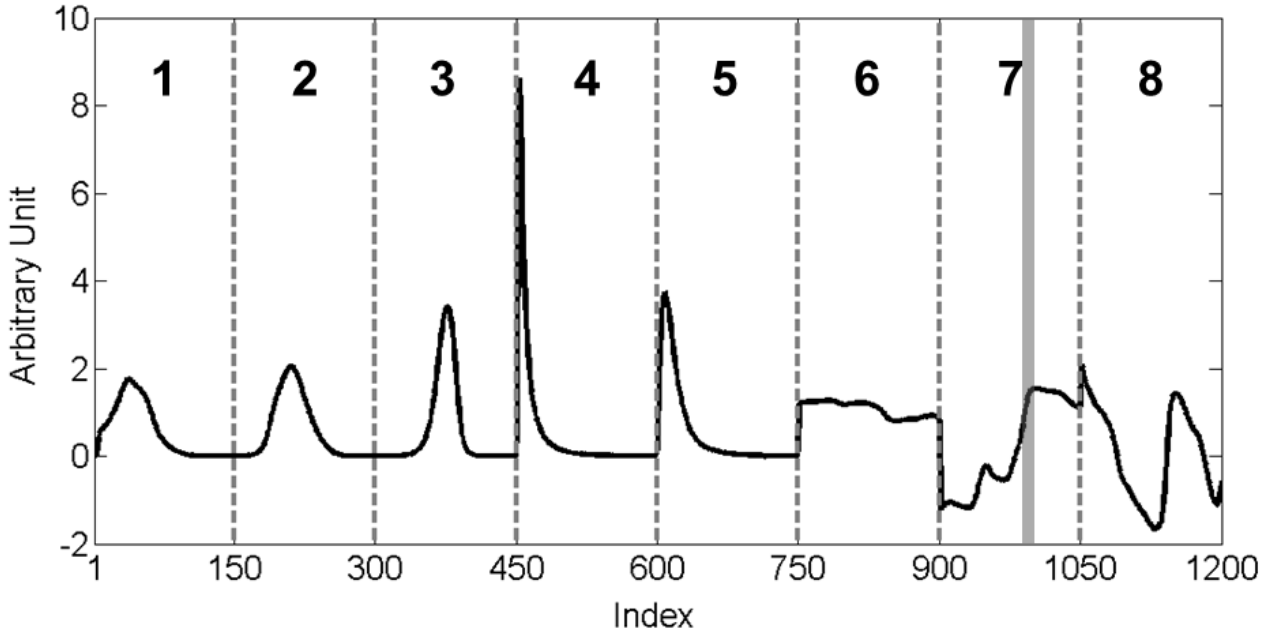
1
2
3
4
5
6
7
8
Arbitrary Unit
Index
10
8
6
4
2
0
-2
1
150
300
450
600
750
900
1050
1200

Figure 3

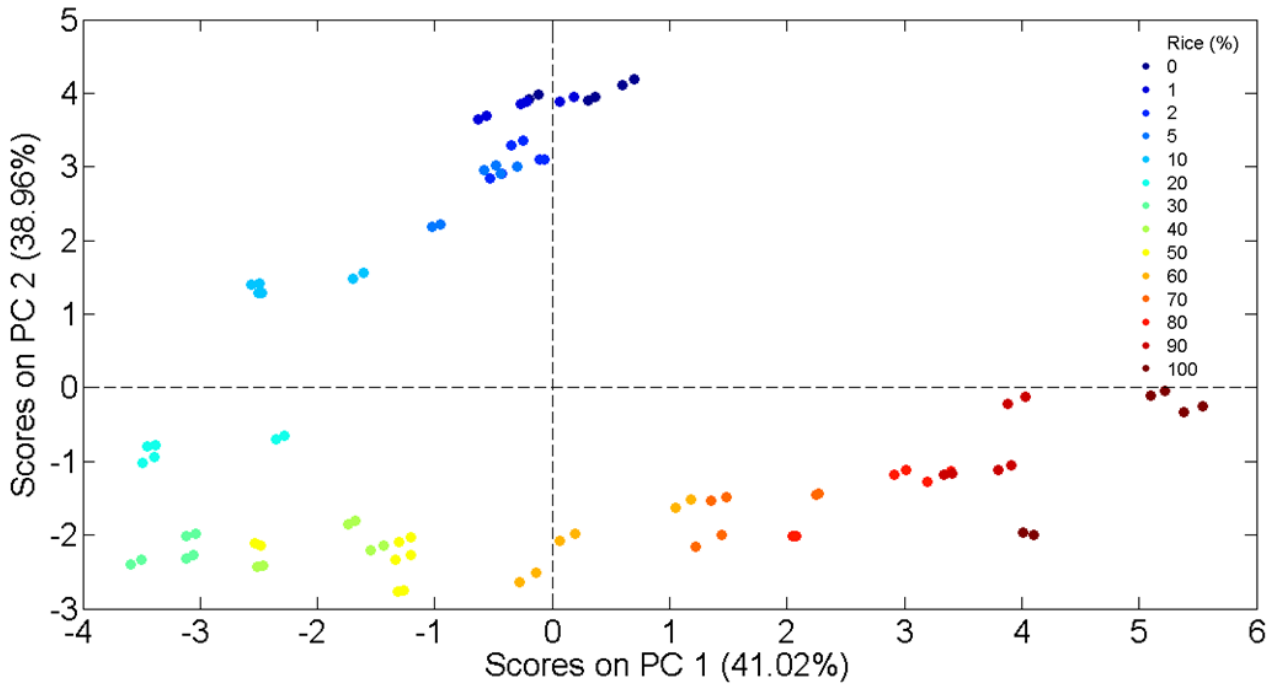
Rice (%)
0
1
2
5
10
20
30
40
50
60
70
80
90
100
Scores on PC 2 (38.96%)
Scores on PC 1 (41.02%)
5
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
5
6

Figure 4

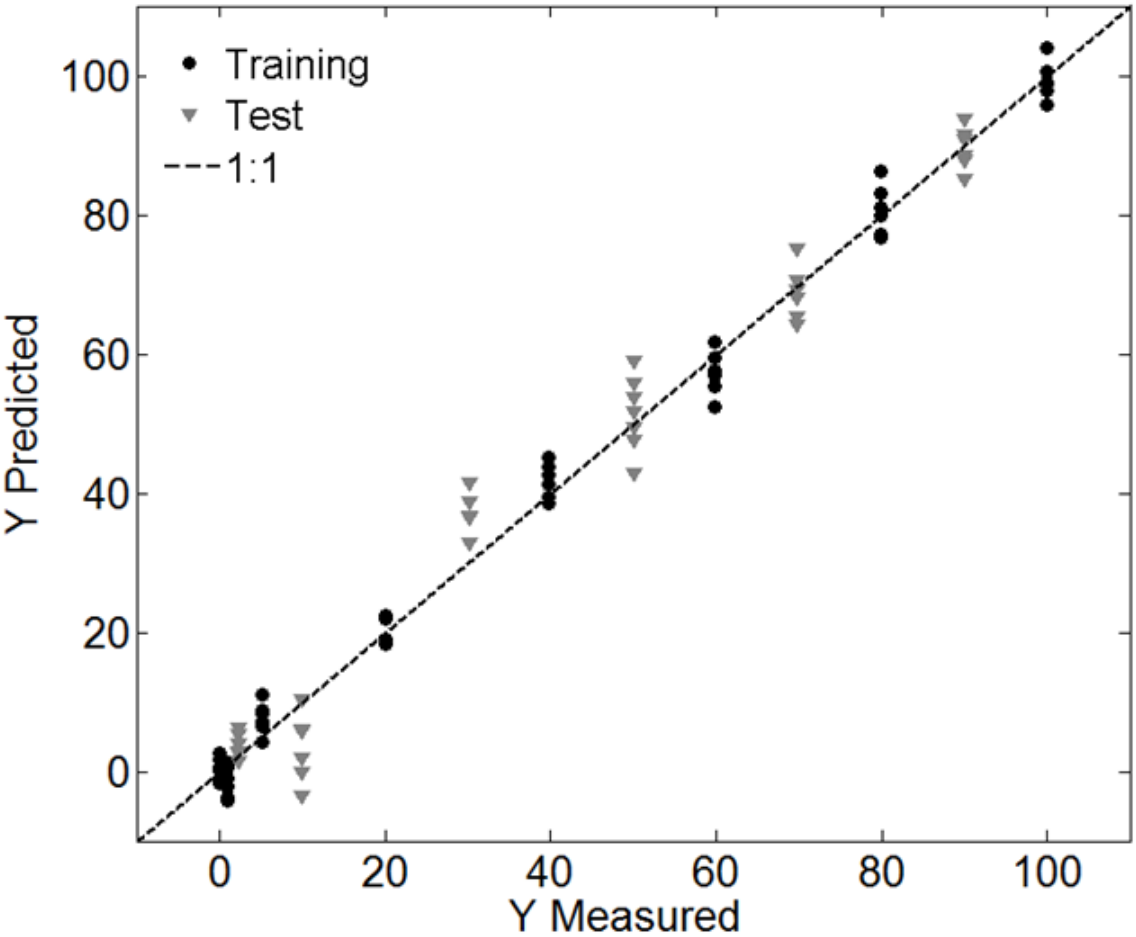
Training
Test
1:1
Y Predicted
Y Measured
100
80
60
40
20
0
0
20
40
60
80
100

Figure 5

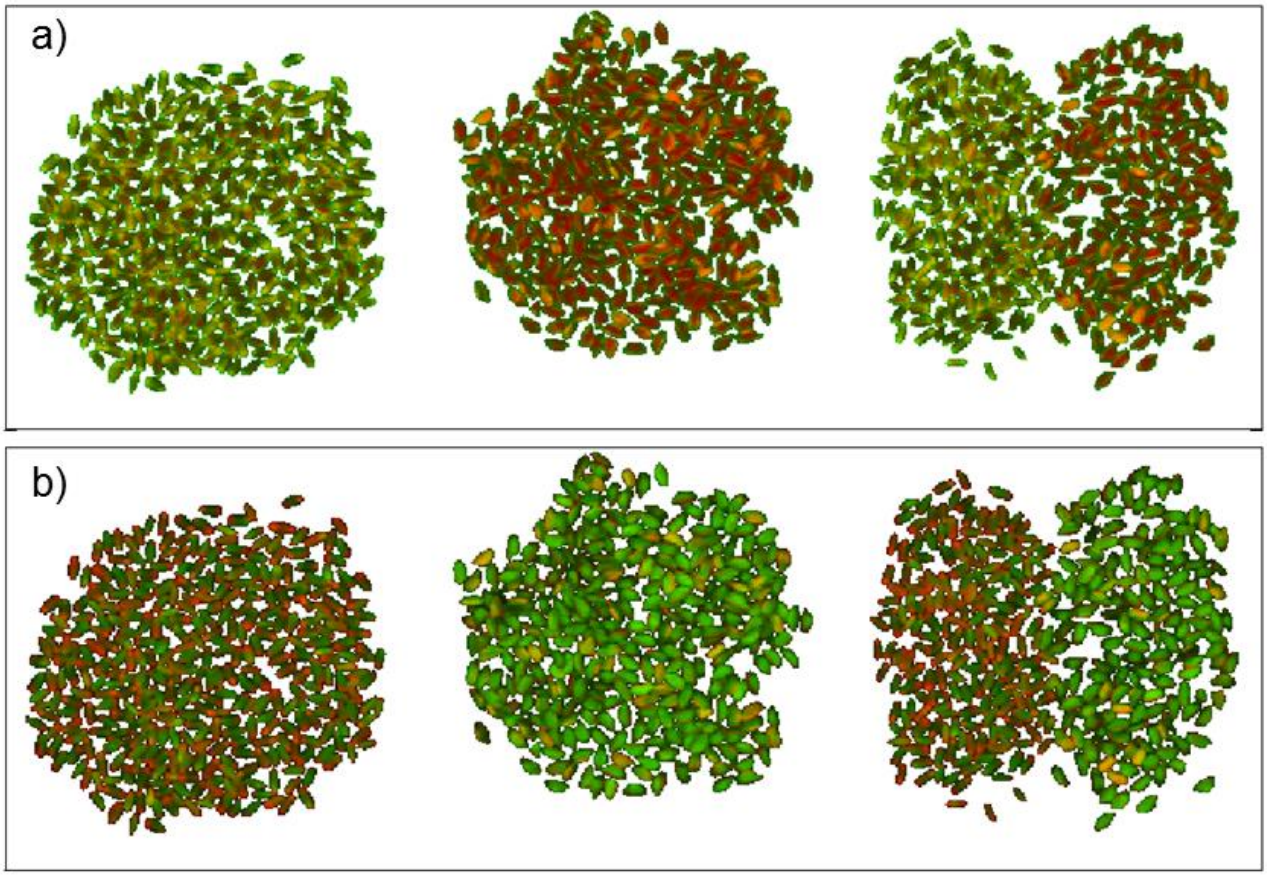

Figure 6

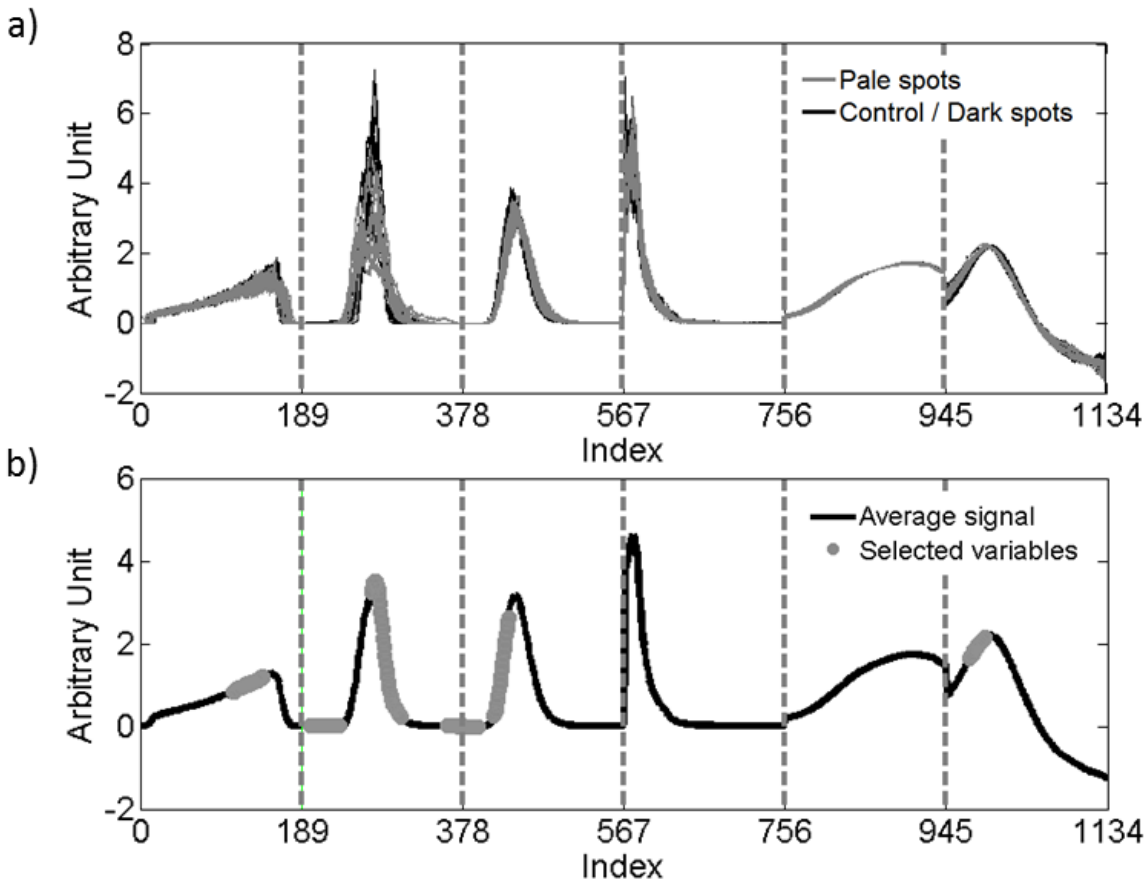
a)
Pale spots
Control / Dark spots
Arbitrary Unit
Index
b)
Average signal
Selected variables
Arbitrary Unit
Index

Figure 7

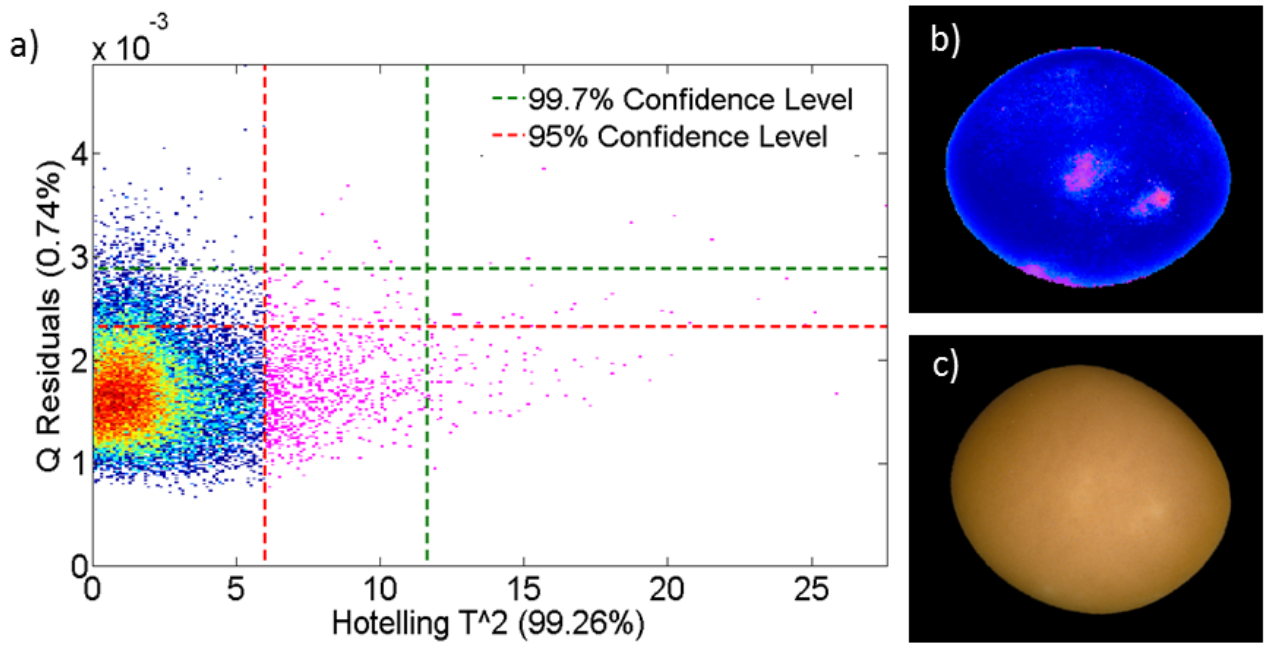
a)
x 10^-3
99.7% Confidence Level
95% Confidence Level
Q Residuals (0.74%)
Hotelling T^2 (99.26%)
0
1
2
3
4
0
5
10
15
20
25
b)
c)

Figure 8

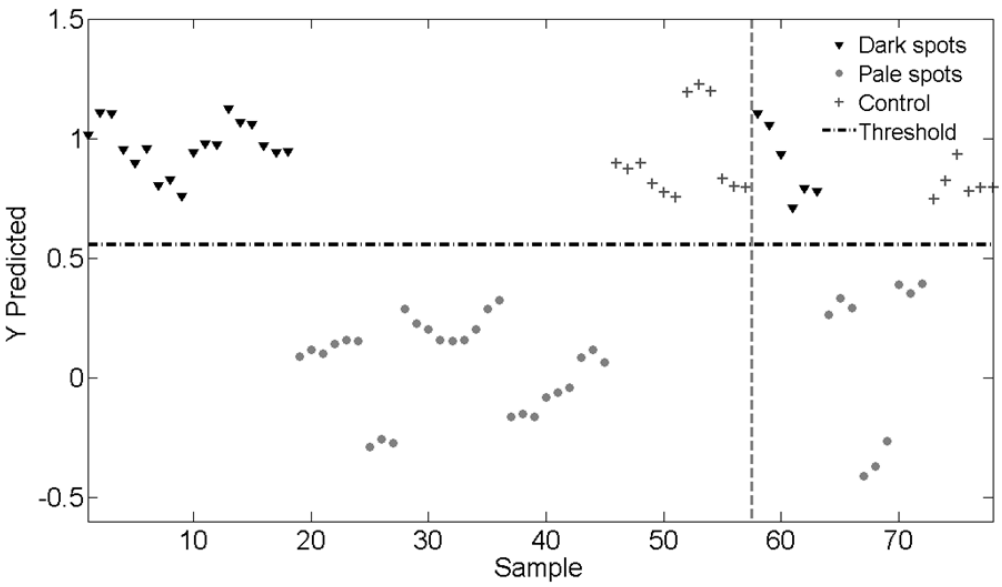
Dark spots
Pale spots
Control
Threshold
Y Predicted
Sample
1.5
1
0.5
0
-0.5
10
20
30
40
50
60
70

Figure 9

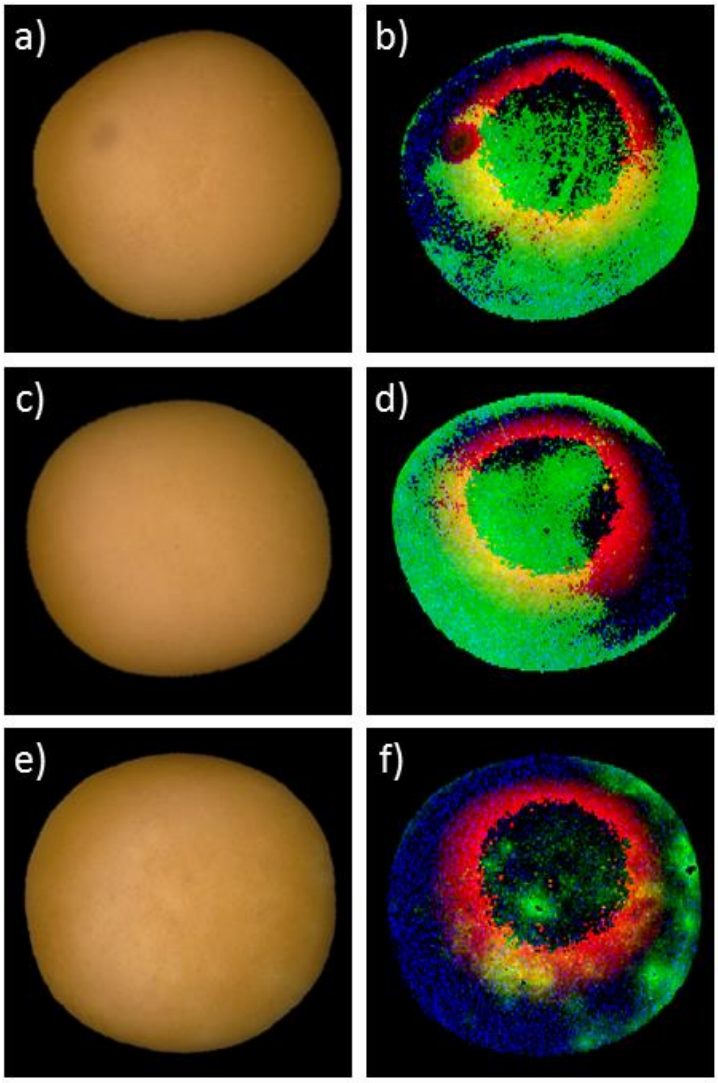

Figure 10

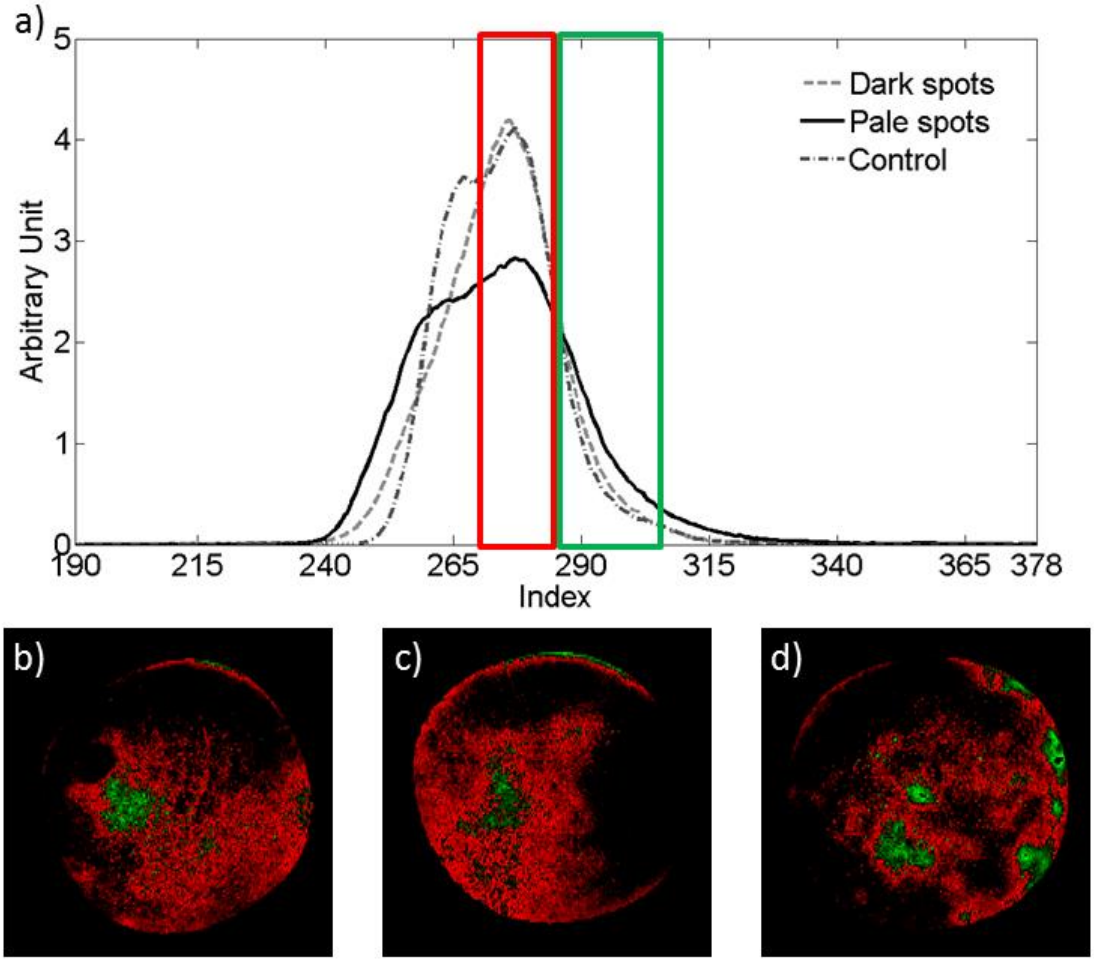
a)
Dark spots
Pale spots
Control
Arbitrary Unit
Index
0
1
2
3
4
5
190
215
240
265
290
315
340
365
378
b)
c)
d)

Figure 11