Subpixel detection of peanut in wheat flour using a matched subspace detector algorithm and near-infrared hyperspectral imaging

Antoine Laborde, Benoît Jaillais, Jean-Michel Roger, Maxime Metz, Delphine Jouan-Rimbaud Bouveresse, Luc L. Eveleigh, Christophe Cordella

Subpixel detection of peanut in wheat flour using a matched subspace detector algorithm and near-infrared hyperspectral imaging

Authors/Affiliations

Antoine Laborde
AgroParisTech, UMR914 PNCA, INRAE/AgroParisTech/Université Paris-Saclay, Paris, France.
Benoît Jaillais
Unité de Statistiques, Sensométrie, Chimiométrie, INRAE/ONIRIS, Nantes, France.
Jean-Michel Roger
INRAE, UMR ITAP, Montpellier University, Montpellier, France.
Maxime Metz
INRAE, UMR ITAP, Montpellier University, Montpellier, France.
Delphine Jouan-Rimbaud Bouveresse
AgroParisTech, UMR914 PNCA, INRAE/AgroParisTech/Université Paris-Saclay, Paris, France.
Luc Eveleigh
AgroParisTech, UMR1145 Ingénierie Procédés Aliments, INRAE/AgroParisTech/Université Paris-Saclay, Paris, France.
Christophe Cordella
AgroParisTech, UMR914 PNCA, INRAE/AgroParisTech/Université Paris-Saclay, Paris, France.

Abstract

The detection of adulterations in food powder product represents a high interest especially when it concerns the health of the consumers. The food industry is concerned by peanut adulteration since it is a major food allergen and it is often used in transformed food products. Near-infrared hyperspectral imaging is an emerging technology for food inspection. It was used in this work to detect peanut flour adulteration in wheat flour. The detection of peanut particles was challenging for two main reasons: the particle size is smaller than the pixel size leading to impure spectral profiles; peanut and wheat flour exhibit similar spectral signature and variability. A Matched Subspace Detector (MSD) was designed to take into account those difficulties and detect peanut adulteration at the pixel scale using

the associated spectrum. Defatted peanut flour and wheat flour were mixed in eight different proportions (from 0.02% to 20%) to test the performances of the hyperspectral measurement technique with the detection algorithm. Additionally, a set of simulated data was generated to overcome the lack of reference value at the pixel scale. The application of MSD on simulated data showed to be relevant with experimental observations and helped to design the detector. The most performant designs for the detector were compared using the simulation by estimating the sensitivity as well as the experimental data by comparing the detection maps. Finally, the number and the position of detections on experimental data were investigated to show the relevancy of the results. This proved that the use of hyperspectral imaging and a fine-tuned MSD enables to detect a global adulteration of 0.2% of peanut in wheat flour using a single hyperspectral image acquisition.

Keywords
Hyperspectral imaging, near-infrared spectroscopy, detection algorithm, spectral simulation, matched subspace detector.

Abbreviations
HSI: Hyperspectral Imaging
NIR: Near-Infrared
LMM: Linear Mixing Model
MSD: Matched Subspace Detector
PCA: Principal Component Analysis

1. Introduction

Hyperspectral Imaging (HSI) is a technique combining spectroscopy with spatial imaging to obtain both spatial and spectral information from a sample. When associated with the near-infrared (NIR) spectral range, HSI is a powerful technique to provide a fast, non-destructive

and cost-effective control method. Since many samples may be chemically heterogeneous, the spatial information provides a high added value for many application fields. Indeed, HSI has been investigated for several decades in remote sensing for earth observation [1], food industry [2], agriculture [3] or medical uses [4]. Particularly, HSI is an emerging technology for food inspection since it provides non-destructive analysis of heterogeneous samples [5] and hyperspectral images allow the visualization of chemical maps of the samples. As an example, Elmasry et al. use this technology to provide water, fat and protein distributions on beef samples [6]. The study of the estimation of food nutrients in samples lead to the assessment of the quality of such products [7].

Hyperspectral imaging technique enables to obtain a spectral measurement for each individual detector of the sensor. As a consequence, each spectrum is representative of a small surface of the sample. Spatial imaging makes the technology more sensitive to minor components since they have more influence on the field of view of one pixel than on the entire sample. This offers the opportunity to detect adulterants in food by characterizing each pixel of the image [8], [9], [10]. In practice, it consists of classifying each pixel spectrum of the hyperspectral image as a target or a background spectral signature. As NIR spectroscopy has shown to be a powerful technique for characterizing organic matter [11], [12], HSI appears to be a promising tool for adulteration detection in the food industry. The literature shows plenty of such applications: Vermeulen et al. studied the detection of ergot bodies in cereal flour [13], Fernández Pierna et al. investigated the detection of melamine in milk powder [14] [15], Verdú et al. proposed to study the adulteration of wheat products [16].

Some adulteration cases are linked to serious public health problems like allergic reactions. The case of peanut allergy is interesting as a typical example. Whereas this allergy is a major issue worldwide [17] [18], peanuts are still widely used in the food industry [19] and maybe also linked to adulteration cases. HSI seems to be a relevant technology to tackle the problem of adulteration in the food industry as a detection tool. Such detection has already been made using crushed peanut in wheat flour [20] [21]. Defatted peanut flour contains smaller peanut particles without fatty acid components. Detecting such a product in wheat flour may be of high interest since powder samples are more frequently used in the food industry.

However, the use of HSI for detection may be a challenge for several reasons. Indeed, the purpose is to identify different materials based on their spectral signature to label each pixel as a target or background pixel. However, there is not a unique spectral signature for each material for two main reasons [22]: the reflectance value at each wavelength of a given material is not deterministic but is a random variable. Its variability is linked to the lighting conditions, the material surface, the sample heterogeneity, and many other factors. Moreover, two different materials may have very similar spectra. Especially for food industry products, the shape of NIR spectra is often similar since they are the result of the mix of the main nutrients. This is particularly the case for wheat and peanut as they are similar products once transformed into flour. More specifically, as peanuts are defatted, the fatty acid NIR fingerprint does not appear in the spectrum. Thus, the ambiguity of spectral information and the spectral variability issue are two main challenges for detection purposes in the food industry.

Another difficulty arises when dealing with powder samples. Indeed, the particle size may be smaller than the pixel size. For a hyperspectral camera, a pixel integrates the radiance signal from all materials in its field of view. Thus, if this field of view contains particles with different spectral signatures, the pixel is considered as mixed and the resulting spectrum does not correspond to any pure chemical defined as target or background. This problem is known as the subpixel detection [22].

The detection problem can be seen as a binary hypothesis test for which each pixel is assigned to the target or the background class. In this condition, no spatial pattern is taken into account and each pixel is considered as an independent situation. For the null hypothesis, the pixel is assumed to contain only background particles whereas, for the alternative hypothesis, the pixel contains some target particles. Detection algorithms have been designed to take into account both variability and subpixel issues using spectral modeling [22].

The mixed pixel issue is tackled using the Linear Mixing Model (LMM) [23] which assumes the radiance measured by a pixel is the sum of the chemical radiances weighted by their surface contribution in the pixel field of view. The spectral variability issue can be addressed using subspace modeling that is integrated into the LMM formulation. Finally, a detector can be designed using the likelihood ratio and by comparing its value to a threshold. Particularly, the Matched Subspace Detector (MSD) [22] [24] is derived using the assumption that both target and background variabilities are modeled.

Such a detector requires to be fine-designed to give high performances. In the case of MSD, the way the subspaces are designed as well as their dimensions highly influence its

performances. Such detectors are often evaluated on reference hyperspectral images that are already labeled at the pixel scale (example of Hydice data set for remote sensing applications [22]). Thus, detector performances can be calculated on reference images so that the design can be optimized. For the inspection of food samples, the field of view of a pixel is much smaller (1 mm × 1 mm) [25] than for remote sensing applications. In this situation, having chemical reference information for each pixel would require performing reference analysis for each pixel field of view surface. This process is not realistic when dealing with food samples. As a consequence, no reference data can be used at the pixel scale for designing the detector.

Spectral data simulation consists of generating new spectral data from a known statistical distribution. This technique can be used to compare different modeling techniques [26]. Such a simulation can take the spectral variability of a pure sample into account by using Principal Component Analysis (PCA) [20]**.** Additionally, the LMM may be used to perform mixed spectra resulting from an interaction between several materials.

The use of a near-infrared hyperspectral imaging system for the detection of adulterant in food has been studied in plenty of applications. However, the issues regarding the variability of the samples, the spectral ambiguity between species, the mixed pixels and the lack of reference data for the detector design make the detection difficult. To our knowledge, no study has been proposed to tackle detection for such samples with particle size smaller than the pixel. The purpose of this study is to evaluate how the MSD approach using the LMM and the modeling of spectral variability can provide performant detection for such a detection problem. As no reference values for the detector design are available, a spectral

simulation method is proposed. After studying the performances of the detector on simulated data, we propose to study the detection using real hyperspectral measurements of flour mix at graduated concentrations of peanut.

2. Materials and methods

2.1 Samples

White wheat flour (Grands Moulins de Paris, Francine batch number **ER510 – FTU104**, France) was used as the background sample. Samples were taken from two different packs. Defatted peanut flour (KoRo Handels GmbH, batch number **C170151**, Germany) was used for the target sample. Flour samples were mixed together to obtain 8 different mass concentrations of peanut flour: 20 %, 10 %, 5 %, 2 %, 1 %, 0.5 %, 0.2 % and 0.02 % for a total mass of 13.75 g. Mass measurements were performed using a precision balance (**Sartorius Entris**, 0.01 g precision). Additionally, pure peanut flour and pure wheat flour were prepared. Each sample was prepared in triplicate. Mixed samples were put in a container to be shaken and mixed with a spatula. Samples were put in a rectangular sample holder (30 mm width x 70 mm length) made of a 7 mm depth cavity. The top of the sample holder was skimmed to remove excess powder without affecting the packing density.

2.2 Hyperspectral imaging system

A line-scan pushbroom Specim SWIR camera (SPECIM, Oulu, Finland) was used for the image acquisition. The hyperspectral camera acquired 288 spectral bands from 900 nm to 2500 nm with a 5.6 nm step. The camera acquired 392 pixels per line and the pixel size was 250 µm × 250 µm. Six halogen lamps were used for the measurement and heated up for 30 minutes

before the acquisition. A white reference measurement was performed before each acquisition using a white diffuse reflectance standard (Spectralon®, SRS-99-010, Labsphere). Additionally, the dark reference image was acquired after closing the shutter of the camera. Each sample was measured independently leading to 30 data cubes.

2.3 Data processing

Each image was cropped to focus on the sample in the central cavity of the sample holder leading to data cubes of size 200 × 320 × 188. The white reference image was averaged along the perpendicular direction of the sensor array to obtain one spectrum for every pixel of the sensor line (I_0).The reflectance image is calculated using (Eq. 1):

$$R = \frac{I - I_B}{I_0 - I_B} \qquad \text{(Eq. 1)}$$

where I_B corresponds to the black measurement for each pixel and I is the raw intensity measurement of the sample. First (under 1200 nm) and last (over 2200 nm) wavelengths were removed as spectra were too noisy. Spectra were processed using a Savitsky-Golay filter to reduce the noise for the remaining wavelengths (2^{nd} order polynomial, 7-points window, and no derivative). A Standard Normal Variate (SNV) transformation was applied to compensate for scattering effects.

2.4 Spectral simulation using Principal Component Analysis

A spectrum can be considered as a vector lying in a p-dimensional space. Each dimension is defined by one variable, namely a wavelength of the spectral range. The variability of

spectral data is the variance of the reflectance for each wavelength in this p-dimensional space. NIR spectral data exhibits a high correlation between the variables. As a consequence, defining the spectral variability independently for each variable is not efficient.

PCA is a method for dimensionality reduction that decomposes the data matrix $X \in \mathbb{R}^{n \times p}$ according to orthogonal sources of the highest possible variance. The data matrix can be decomposed as follow:

$$X = TP^T + R \quad \text{(Eq. 2)}$$

where $T \in \mathbb{R}^{n \times p}$ is the score matrix, $P \in \mathbb{R}^{p \times p}$ is the loading matrix, $R \in \mathbb{R}^{n \times p}$ is the matrix of residuals, and the upper script symbol ^{T} refers to the transposed matrix.

Under this representation, the distribution of scores for each component can be considered as Gaussian with mean μ - which is, in practice, null - and variance σ^2. Each component of the matrix P provides a unit vector spanning the subspaces of the sample variability. As a consequence, a new spectrum can be simulated by generating its coordinates in the principal components space. In this representation, the p coordinates are generated using Gaussian distributions.

The Linear Mixing Model (LMM) describes the linear interaction between spectral signatures of pure materials in a mixture context [23]. In this model, a spectrum x is described by (Eq. 3).

$$x = \sum_{i=1}^{A} c_i s_i + w \quad \text{(Eq. 3)}$$

where s_i is the pure spectrum of the i-th material, $c_i \in [0\,,1]$ is the associated concentration, w is the residual vector and $A \in \mathbb{N}$ is the number of species in the model.

The PCA and LMM are combined considering that s_i may represent a source of variability as well as a pure spectral signature. The process of data simulation used in this study is described by the following procedure (Figure 1):

1) PCA is performed on the data matrix of pure wheat data X_w and pure peanut data X_p distinctly. The average spectra of both matrices $\overline{X}_w$ and $\overline{X}_w$ are calculated and considered as the pure spectral signatures of the materials.

2) For every dimension $i \in [1\,, p]$, the distribution of scores T^i is assumed to be Gaussian of mean μ^i (which is equal to 0) and standard deviation σ^i. These parameters are estimated on the pure sample image for wheat and peanut distinctly.

3) For a given peanut percentage c, the average spectrum is simulated using the LMM:

$$\widetilde{X}_0 = c\overline{X}_p + (c-1)\overline{X}_w \qquad \text{(Eq. 4)}$$

c being a scalar varying between 0 and 1.

4) The simulation is completed by adding the random variability to the average simulated spectrum $\widetilde{X}_0$. The variability is obtained by multiplying the simulated scores with the principal component loadings. The simulated scores $\widetilde{T}^i$ are generated from Gaussian distributions with the parameters calculated in step 2) for peanut and wheat separately. The total variability attributed to $\widetilde{X}$ is a balance between peanut and wheat controlled by the proportion c.

$$\widetilde{\mathrm{X}} = \widetilde{\mathrm{X}}_0 + \mathrm{c}\widetilde{\mathrm{T}}_\mathrm{p}\mathrm{P}_\mathrm{p}^\mathrm{T} + (\mathrm{c}-1)\widetilde{\mathrm{T}}_\mathrm{w}\mathrm{P}_\mathrm{w}^\mathrm{T} \qquad \text{(Eq. 5)}$$

where the tilde symbol designates the simulated matrices.

[Insert Figure 1]

Figure 1: Procedure for spectral simulation. For the first two steps (1-2), the procedure is only described for the wheat flour sample. The same procedure is applied on the peanut flour sample.

The simulation procedure is applied to obtain 100 spectra for each peanut concentration 5%, 10%, 15% and 20%.

2.5 Matched subspace detector design

According to the LMM, a given pixel has a spectrum x described by:

$$x = \sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{a_i s_i} + \mathrm{w} \qquad \text{(Eq. 6)}$$

where the vectors $\mathrm{s_i}$ define the variability space of x . These vectors can be obtained from statistical techniques like PCA or Non-Negative Matrix Factorization as shown in the section 2.5. The coefficients $\mathrm{a_i}$ are the abundance coefficients associated to the $\mathrm{s_i}$. w is the noise vector coming from the sensor or the measurement conditions. Despite the fact Eq. 6 is mathematically equivalent to Eq. 3, it does not hold the same physical sense in the context of this work. Eq. 3 is the model used for representing the interaction between peanut and wheat flour in a simulation context. Eq. 6 is the model used to represent the mixture context of a pixel for a detection purpose.

In the context of a subpixel detection problem, two competing hypotheses are tested. For the null hypothesis associated with the background class, the pixel is assumed to contain only the background sample. Consequently, the LMM decomposes the pixel spectrum

according to the sources of variability associated to the background sample s_i^b. For the alternative hypothesis associated with the target class, the pixel contains background particles as well as target particles. Thus, x is modeled using the LMM with the sources of variability associated with the background and the target s_i^p. The MSD is thus based on the following statistical test:

$$H_0: x = \sum_{i=1}^{k_b} a_i s_i^b$$

$$H_1: \sum_{i=1}^{k_b} a_i^b s_i^b + \sum_{i=1}^{k_s} a_i^p s_i^p$$

where k_b and k_s define the dimensionality associated to the variability for the background and the target respectively. Two matrices are defined corresponding to these hypotheses: B contains the vectors s_i^b in columns, and S contains the vectors s_i^b and s_i^p in columns as shown below.

$$B = \left(s_1^b,\ s_2^b, \dots, s_{k_b}^b\right) \text{ and } S = \left(s_1^b, s_2^b, \dots, s_{k_b}^b,\ s_1^p, s_2^p, \dots, s_{k_b}^p\right)$$

The generalized likelihood ratio approach gives the detection statistic for the MSD [22]:

$$T_{MSD}(x) = \frac{x^T(P_B^\perp - P_S^\perp)x}{x^T P_S^\perp x} \qquad \text{(Eq. 7)}$$

where $P_B^\perp$ and $P_S^\perp$ are the projection matrices on the orthogonal subspace of B and S respectively. These projectors are obtained from the B and S matrices by the following formula:

$$P_X^\perp = I - X(X^T X)^{-1} X^T \qquad \text{(Eq. 8)}$$

After calculating the detection statistic for each pixel spectrum of a sample using (Eq. 7), a threshold must be applied to classify between both classes: target (the pixel contains target and background) and background. This threshold is chosen using the Neyman-Pearson approach that consists of maximizing the detection rate by keeping the false alarm rate

under a given limit. For this purpose, the threshold is fixed as the maximum value of the detection statistic obtained on one replicate of the pure wheat sample:

$$T_{MSD}(x) \gtrless \eta_{NP} \qquad \text{(Eq. 9)}$$

$$\text{with } \eta_{NP} = \max\left(T_{MSD}(X_W)\right).$$

The two other pure wheat replicates were used in order to assess the robustness of the thresholding method.

For both data matrix X_W and X_P, a PCA on non-centered data was performed and the first loadings were extracted to obtained vectors s_i^b on wheat flour and s_i^p on peanut flour. k_b and k_s correspond to the number of extracted loadings for wheat and peanut flours respectively.

As Eq. 7 and 8 show, the design of the MSD highly depends on the design of B and S. The way of constructing these matrices depends on the choice of the spectral profiles s_i^b and s_i^p. In the method presented in this work, those spectral profiles are obtained by selecting consecutive components from PCA performed on the pure samples. However, the number of components to choose to obtain B and S : k_b and k_s are tunable parameters that change the calculation of T_{MSD}. The design of the MSD consists of finding optimal values fork_b and k_s. In the following, the performance of the detector is qualified using its sensitivity. In this paper, the sensitivity refers to the minimum local concentration (at the pixel scale) for which the detection rate is over 99%.

2.6 Software

The data processing is performed using Python 3.7. For data simulation, the PCA is performed using the Scikit-Learn 0.18.1 implementation consisting of a Singular Value

Decomposition (SVD). For the design of MSD, the PCA on non-centered data is performed using the eigenvalue decomposition of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ performed with Numpy 1.16.4.

3. Results and discussions

3.1 Data simulation for the detector design

Figure 2 shows the factorial plan of PCA performed on pure wheat and pure peanut samples. Two different sets of scores are plotted: empty-square markers are from pure peanut and wheat samples used for the PCA calculation; filled markers are the simulated data obtained by projecting the spectra onto the two first PCA loadings. This factorial plan shows that simulated data are ordered according to the first principal component: lower peanut concentrations are closer to pure wheat flour on the left and higher concentrations are on the right side. The focus shows the variability of simulated data is similar to the one of real measurements of wheat flour. Peanut flour shows a greater variance on the second principal component because the surface of the sample exhibits more heterogeneity. The Figure 2 shows that the simulated data exhibit less variability on PC 2 compared to pure peanut and, even less so, to pure wheat. As Eq. 5 shows, the amount of variability added to the model depends on the parameter c which represents the simulated target concentration. Consequently, when the simulated concentration is low, the amount of variability is more similar to the one of wheat flour. Moreover, the difference observed on

the variability level of pure wheat flour and simulated data is explained by the Gaussian simplification of the real distribution of the data. Indeed, the real distribution of data points in the PCA may not fit a Gaussian perfectly. More particularly, extreme points on PC 2 seem to be more present on real data. These results show that simulated data can be relevantly used for the estimation of the sensibility of detectors.

[Insert Figure 2]

Figure 2: Simulated data are projected on the score plot of the PCA performed on real measurements of pure samples. Real measurements are shown with empty square markers. Projected simulated data are shown for low concentrations between 5% and 20% of peanut. Only 400 representative data points among 130 000 are plotted for peanut and wheat flour.

Table 1: Design parameters of three detectors of interest. shows the details of the design (the values for k_b and k_s) for three detectors. These detector designs are selected because they show the best and most interesting results among all those which have been tested. The next section shows the results for other design and focuses on the choice of the parameters.

Table 1: Design parameters of three detectors of interest.

Name of the detector	**Detector 1**	**Detector 2**	**Detector 3**
k_b	1	2	2
k_s	1	1	2

The design of the detector requires to evaluate its sensitivity to optimize the choice of the parameter values (k_b and k_s). Figure 3 shows the detection rate of the three different detector designs described in Table 1: Design parameters of three detectors of interest.. The detection rate indicates the fraction of detected targets for a given peanut concentration of the simulated data. For zero peanut concentration, spectra from real wheat flour images were used. The graph shows that all detectors do not have any false alarm on real wheat measurements. This means that the thresholding method is robust for all three detector designs. Detectors 2 and 3 reach a detection rate of 100% for a simulated peanut concentration of 20% and they both have similar detection rates for smaller peanut

concentrations. Detector 1 exhibits a smaller detection rate for every concentration, and it does not reach a 100% detection rate for 20% of peanut adulteration. This shows that, according to the simulated data, the detectors 2 and 3 have a similar sensitivity which is higher than the one of the detector 1.

[Insert Figure 3]

Figure 3: The detection rate according to the peanut concentration in simulated data for three detector design. For zero peanut concentration, real wheat flour measurement data are used to calculate the detection rate.

Figure 4 shows the comparison of a detection map for the three detectors presented in Figure 3. For this purpose, a focus is made on the hyperspectral image measured on a sample containing 2% of peanut flour. The map represents an area of 104 × 62 pixels (2.6 × 1.5 cm) and each color corresponds to an output of the comparison of 2 detectors: detectors 1 and 2 for the top map and detectors 2 and 3 for the bottom map. The top map shows several groups of blue pixels (top left-hand corner) meaning that many pixels are only detected by detector 2 and not by detector 1. Since real measurements do not contain any reference value at the pixel scale, the real position of the targets is unknown. However, the fact that neighbor pixels are simultaneously detected strengthens the probability that there is effectively peanut in these pixels. In other words, the detection of a neighborhood of pixels is more credible than the detection of an isolated pixel. This argument is developed in the section Detection position. On the other hand, detector 1 only detects one pixel exclusively. The comparison of the detection maps tends to show that detector 1 is less sensitive than detector 2.

The maps in Figure 4 show only a few colored pixels which means that detectors have approximately the same performances. The detector 3 detects 13 more pixels than detector 2 and only one (in the middle of the map) can be considered suspicious since it is isolated.

On the other hand, the detector 2 shows only one exclusive detection (in the middle of the map) which is isolated.

These observations made on real data are relevant to the results obtained with the simulation method (Figure 3): the detectors 2 and 3 have similar performance and are more sensitive than detector 1. This shows that simulated data can be relevantly used for designing and analyzing matched subspace detectors. Additionally, detection map comparison shows that detector 3 seems to be more sensitive than detector 2 despite the fact that they exhibit almost equal performances on the simulated data.

[Insert Figure 4]

Figure 4: Detection map comparison on real data (focus on a sample with 2% of peanut – replicate A). The map above shows the comparison between the detectors 1 and 2, below is the comparison between the detectors 2 and 3. Each pixel is colored according to the output of both detectors (see the legend).

3.2 Matched subspace detector

3.2.1 Design of the detector

Figure 5 shows the sensitivity of several detectors calculated on the simulated data. The first graph on the left shows the effect of varying $\mathrm{k_b}$ with $\mathrm{k_s} = 2$. This means the background dimensionality is varying whereas the dimensionality of the target model is fixed. $\mathrm{k_b} = 1$ gives poor performances since no spectra are detected even for a peanut concentration of 20%. The best performances are obtained for $\mathrm{k_b} = 2$. Then, increasing $\mathrm{k_b}$ leads to lower detection rates. The graph on the right shows the evolution of the sensitivity when fixing $\mathrm{k_b} = 2$. In this condition, the performances of the detectors are identical for $\mathrm{k_s} = 1$ and $\mathrm{k_s} = 2$. Then, choosing a higher value for $\mathrm{k_s}$ decreases the detection rate. We also show the design with $\mathrm{k_s} = 1$ and $\mathrm{k_b} = 1$ (see Figure 3) provides a lower sensitivity than

the detector 3. Consequently, the results show that there is an optimal design $k_b = 2$, $k_s = 2$ for the MSD regarding the performances on the simulated data.

[Insert Figure 5]

Figure 5 : Detector sensitivity for varying k_s and k_b. On the left, k_s is fixed and k_b goes from 1 to 4. On the right, k_s goes from 1 to 4 and k_b is fixed. Detection rate are calculated on simulated data for concentration from 5% to 20% and on real wheat measurement data for 0%.

Figure 6 depicts the geometrical interpretation of the matched subspace detector. For simplicity, we assume the spectrum vector x belongs to a 3-dimensional subspace defined by three virtual wavelength bands λ_1, λ_2 and λ_3. The target and the background subspaces are assumed to be 1-dimensional and are represented by vectors s^b and s^p. The common subspace is a 2-dimensional subspace represented by the plane S. The MSD compares the residuals of the decomposition of x under hypothesis H_0 and H_1. Geometrically, the squared residuals under H_0 corresponds to AW^2 which is the norm of the projection of x onto the orthogonal subspace of s^b. Similarly, under the alternative hypothesis H_1, the residuals AP^2 are calculated as the norm of the projection of x onto the orthogonal subspace of S. As a consequence, the quantities involved in the matched subspace detector formulation can be translated in terms of vector norms on the graph of the Figure 6: $x^T P_b^\perp x = AW^2$ and $x^T P_S^\perp x = AP^2$. Then, the matched subspace detector metric can be rewritten as follow:

$$T_{MSD}(x) = \frac{x^T(P_b^\perp - P_S^\perp)x}{x^T P_S^\perp x} = \frac{AW^2 - AP^2}{AP^2} = \frac{WP^2}{AP^2} \quad \text{(Eq. 10)}$$

The last equality of (Eq. 10) can be deduced from the fact that P is the orthogonal projection of A on the subspace S leading to the relationship: $\mathrm{AW}^2 = \mathrm{WP}^2 + \mathrm{PA}^2$.

[Insert Figure 6]

Figure 6: Geometrical interpretation of the matched subspace detector. A spectrum is represented by a vector x in a 3-dimensional subspace ($\lambda_1, \lambda_2, \lambda_3$). The target and background subspaces are represented as 1-dimensional and the common subspace is the red plane defined by the union. For more clarity, the triangle OWA is rectangle in W and the triangle OAP is rectangle in P.

The geometrical interpretation is useful to visualize that the matched subspace detector compares two models with two different dimensionalities. If the pixel spectrum only contains wheat, the modeling of x on S^{b} is sufficient. Consequently, WP^2 is small because adding vectors from the target subspace S^{p} should not significantly improve the model fitting. However, if the pixel spectrum contains some peanut, adding vectors from the target variability to the model should significantly improve it. The role of the detector threshold is to define the limit for which the distance between both models is large enough to consider that a peanut particle is present or not.

From this interpretation, the results from different detector designs can be explained. When $\mathrm{k_b} < \mathrm{k_s}$, the models for H_0 and H_1 become highly unbalanced. For instance, for the design where $\mathrm{k_b} = 2$, and $\mathrm{k_s} = 3$, x is modeled using a 2-dimensional subspace under H_0 compared to a 5-dimensional subspace under H_1. This unbalanced situation always holds whatever the design of the MSD. For this reason, the matched subspace detector metric does not compare directly the residuals under both hypotheses but a ratio. The role of the thresholding phase is to take into account this unbalanced problem. One method consists of tuning the threshold on the detector statistic obtained on a pure background sample. This method works when the histograms of the detector statistic between the pure wheat flour and the adulterated pixels can be separated. However, when the design is too highly

unbalanced, the model fitting for H_0 and H_1 becomes very competitive because additional vectors from the peanut subspace can be used to improve the fitting on x even if it only contains wheat. This is a consequence of the fact the spectral signatures of peanut and wheat flours are similar. In these conditions, it becomes difficult to find a threshold value that meets the needs for a high detection rate.

When k_b and k_s are too high, each model takes into account a large variability. However, this is not a good strategy since peanut and wheat spectral signatures are similar and their variabilities are high. Thus, increasing the parameters k_b and k_s leads to more ambiguity for the detector and a high-dimensional model will be preferred. Finally, when k_b and k_s are too small ($k_b = 1$ and $k_s = 1$), the model cannot consider the background variability which is detrimental for this application.

Finally, the interpretation of the MSD metric is useful to understand why the set of appropriate values for the parameter k_b and k_s is relatively small. The results on the simulated data are relevant with this interpretation and enable to find the best parameters for the design of the detector.

3.2.2 Number of detections

Figure 8 shows the detection rate of three selected detectors for all the measured samples. The scatter plot shows the detection rate increases when the sample concentration increases so the application of the MSD detector on real measurement is relevant. The results also show a high variance in the detection rates for the same detector applied to the three replicates of the same concentration. This can be explained by the experimental conditions. Indeed, hyperspectral measurements are representative of the material through a depth of some millimeters. However, the peanut concentration is a global characteristic of

the sample volume. As it is not possible to make sure that the sample is homogenous, the global concentration and the apparent concentration at the surface may not be equal. Consequently, sample replicates may exhibit real concentration variance at the surface.

Additionally, the results show a nonlinear behavior in the evolution of the detection rate according to the peanut concentration. This is visible on the left-graph of Figure 8. There are three phenomena to take into account to explain this behavior.

1) Let us assume a sample is perfectly homogeneous at the pixel-scale with a concentration of 20% and a detector with a sensitivity of 10% is applied on a hyperspectral data cube. Then, each pixel has a contribution of 20% of peanut and is detectable. As a consequence, every pixel is detected and the detection rate is 1. Conversely, if the concentration is 5%, no pixel is detected and the detection rate is 0. Figure 7 shows the explained behavior and the resulting detection rate curve. This phenomenon may explain part of the nonlinearity of the results. This highly depends on the scale of scrutiny for which the sample is declared to be homogeneous: that is to say, the sampling size for which the homogeneity is guaranteed.

[Insert Figure 7]

Figure 7 : A virtual Matched Subspace Detector is considered to have a pixel-wise sensitivity of 10%. The sample is assumed to be perfectly homogeneous so that the situation of every pixel is identical: there is one target particle in the pixel field of view for the sample with 5 % of peanut, and there is four target particles in the pixel field of view for the sample with 20% of peanut.

2) More realistically, the fact a sample has a global concentration of 20% does not mean each pixel surface has the same concentration. In other words, it is reasonable to assume the sample is heterogeneous at the pixel-scale. Assuming there is no spatial relationship between pixels, an image of 100 000 pixels can be statistically considered as 100 000 independent experiments. Each one can be seen as a series of Bernoulli processes for which there are as many trials as the number of particles in

the pixel. The probability of selecting a peanut particle corresponds to the global concentration of peanut. The pixel-wise concentration is thus given by a binomial distribution. Such a simulation provides a detection rate curve shown on the left graph of Figure 8. The result is obtained by simulating a sensitivity of 25%. This is relevant to the estimated sensitivity using the simulated data.

3) In practice, pixels are not independent, so several neighbor pixels are likely detected as targets. This phenomenon may happen for two main reasons. Firstly, because flour samples do contain several particle sizes. Some may be higher than 150 µm. With such a size, some particles may overlap several pixels and make all of them detectable. Secondly, because the particle size study of flours shows that particles tend to agglomerate with each other. Despite the fact the median particle size is approximately 50 µm in wheat and peanut flours, some agglomerations may have a size of several millimeters which leads to the coverage of several pixels. The agglomeration occurs particularly often for peanut flour because of the remaining fatty acids.

[Insert Figure 8]

Figure 8: The detection rate for the three selected detectors for all samples (from 0.02% to 20% of peanut concentration). The graph on the right shows a focus in the concentration range from 0.02% to 2.5%. The dot line is the result of the statistical simulation performed with a virtual detector of sensitivity 25%.

These arguments show all detector designs provide relevant results regarding the real samples with different concentrations. However, the relationship between the number of detections and the peanut concentration of samples is complex. This is related to the scale for which the sample homogeneity is assumed as well as the sensitivity of the detector and the particle size.

3.2.3 Detection positions

Previous results show that detector 3 seems to provide the most sensitive results. Figure 9 shows the detection maps of three different concentrations and three replicates. These maps show that the number of detections is repeatable among the replicates as Figure 8 showed. They also show the detection locations are credible: most of them are made on neighbor pixels so that peanut agglomeration can be seen. Furthermore, the location of these agglomerations is randomly distributed across the sample. These results show the MSD can be used to detect difficult targets as peanut flour in wheat and give their position efficiently.

[Insert Figure 9]

Figure 9: Detection maps obtained after applying Detector 3 on real samples for three concentration: 20%, 5% and 0.2%. A black pixel means no detection, a white pixel means target is detected.

CONCLUSIONS

The purpose of this study was to tackle a difficult detection problem dealing with similar materials with high spectral variability and particle size involving subpixel detections. The development of a Matched Subspace Detector was proposed to overcome these difficulties. The spectral variability was tackled using subspace modeling through PCA whereas the Linear Mixing Model was used to consider the subpixel detection issue. Moreover, data simulation of different peanut concentrations was proposed to provide an estimation of the performances of the detectors. This technique was used to choose the most appropriate design.

As a result, the data simulation method provides realistic data regarding the measurement variability. Detector designs giving a high detector rate for the low concentrated samples were conserved and applied on real measurements. Results show that the MSD and the data simulation were relevant to overcome the detection issue. Despite the lack of local reference values, the number and the position of the detections show that MSD provides reliable results.

Additional work could be provided for further improvements for this kind of detection situation. Firstly, the data simulation process could be improved by selecting a subset of loadings to simulate the data. This may provide more reliability on the simulation. Indeed, the expected sensitivity obtained on simulated data (20%) does not seems to be reached in practice. Then, even if no spatial a priori hypothesis can be done regarding the particle size, the detection results on real data show that most of the detections are made on neighbor pixels. This is because of particle agglomeration which is a phenomenon that applies on very small flour particles. Such an effect could be taken into account to improve the detection by adding some spatial dependence in the detector algorithm. Finally, the statistical simulation for the number of detected pixels according to the peanut concentration may be improved. For example, the hypothesis that each pixel is independent of the other could be changed to get a more accurate approach.

REFERENCES

[1] Y. Xie, Z. Sha, M. Yu, Remote sensing imagery in vegetation mapping : a review, J. Plant Ecol. 1 (2008) 9–23. https://doi.org/10.1093/jpe/rtm005.

[2] G. Elmasry, M. Kamruzzaman, D. Sun, P. Allen, Principles and Applications of Hyperspectral Imaging in Quality Evaluation of Agro-Food Products : A Review, Crit. Rev. Food Sci. Nutr. 52 (2012) 999–1023. https://doi.org/10.1080/10408398.2010.543495.
[3] L.M. Dale, A. Thewis, C. Boudry, I. Rotar, P. Dardenne, V. Baeten, J.A. Fernández Pierna, Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review, Appl. Spectrosc. Rev. 48 (2013) 142–159. https://doi.org/10.1080/05704928.2012.705800.
[4] G. Lu, B. Fei, Medical hyperspectral imaging : a review, J. Biomed. Opt. 19 (2014). https://doi.org/10.1117/1.JBO.19.1.010901.
[5] Y. Feng, D. Sun, Application of Hyperspectral Imaging in Food Safety Inspection and Control : A Review, Crit. Rev. Food Sci. Nutr. 52 (2012) 1039–1058. https://doi.org/10.1080/10408398.2011.651542.
[6] G. Elmasry, D. Sun, P. Allen, Chemical-free assessment and mapping of major constituents in beef using hyperspectral imaging, J. Food Eng. 117 (2013) 235–246. https://doi.org/10.1016/j.jfoodeng.2013.02.016.
[7] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, G. Downey, J.M. Frias, Hyperspectral imaging - an emerging process analytical tool for food quality and safety control, Trends Food Sci. Technol. 18 (2007) 590–598. https://doi.org/10.1016/j.tifs.2007.06.001.
[8] M. Kamruzzaman, Y. Makino, Assessment of Visible Near-Infrared Hyperspectral Imaging as a Tool for Detection of Horsemeat Adulteration in Minced Beef, Food Bioprocess Technol. 8 (2015) 1054–1062. https://doi.org/10.1007/s11947-015-1470-7.
[9] M. Kamruzzaman, Y. Makino, S. Oshita, Rapid and non-destructive detection of chicken adulteration in minced beef using visible near-infrared hyperspectral imaging and machine learning, J. Food Eng. 170 (2016) 8–15. https://doi.org/10.1016/j.jfoodeng.2015.08.023.
[10] W.H. Su, D.W. Sun, Fourier Transform Infrared and Raman and Hyperspectral Imaging Techniques for Quality Determinations of Powdery Foods: A Review, Compr. Rev. Food Sci. Food Saf. 17 (2018) 104–122. https://doi.org/10.1111/1541-4337.12314.
[11] D.I. Ellis, V.L. Brewster, W.B. Dunn, J.W. Allwood, A.P. Golovanov, D.I. Ellis, V.L. Brewster, W.B. Dunn, J.W. Allwood, A.P. Golovanov, R. Goodacre, Fingerprinting food: current technologies for the detection of food adulteration and contamination, Chem. Soc. Rev. 41 (2012) 5706–5727. https://doi.org/10.1039/c2cs35138b.
[12] L. Manning, J. Soon, Developing systems to control food adulteration, J. Food Policy. 49 (2014) 23–32. https://doi.org/10.1016/j.foodpol.2014.06.005.
[13] P. Vermeulen, M.B. Ebene, B. Orlando, J.A. Fernández Pierna, V. Baeten, Online detection and quantification of particles of ergot bodies in cereal flour using near-infrared hyperspectral imaging, Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess. 34 (2017) 1312–1319. https://doi.org/10.1080/19440049.2017.1336798.
[14] J.A. Fernández Pierna, D. Vincke, V. Baeten, C. Grelet, F. Dehareng, P. Dardenne, Use of a multivariate moving window PCA for the untargeted detection of contaminants in agro-food products, as exemplified by the detection of melamine levels in milk using vibrational spectroscopy, Chemom. Intell. Lab. Syst. 152 (2016) 157–162. https://doi.org/10.1016/j.chemolab.2015.10.016.
[15] J.A. Fernández Pierna, D. Vincke, P. Dardenne, Z. Yang, L. Han, V. Baeten, Line scan

hyperspectral imaging spectroscopyfor the early detection of melamineand cyanuric acid in feed, J. Near Infrared Spectrosc. 22 (2014) 103–112. https://doi.org/10.1255/jnirs.1109.
[16] S. Verdú, F. Vásquez, R. Grau, E. Ivorra, A.J. Sánchez, J.M. Barat, Detection of adulterations with different grains in wheat products based on the hyperspectral image technique: The specific cases of flour and bread, Food Control. 62 (2016) 373–380. https://doi.org/10.1016/j.foodcont.2015.11.002.
[17] A.W. Burks, Peanut allergy, Lancet. 371 (2008) 1538–1546. https://doi.org/10.1016/S0140-6736(08)60659-5.
[18] S.L. Hefle, S.L. Taylor, Food allergy and the food industry, Curr. Allergy Asthma Rep. 4 (2004) 55–59. https://doi.org/10.1007/s11882-004-0044-y.
[19] X. Zhao, J. Chen, F. Du, Potential use of peanut by-products in food processing: A review, J. Food Sci. Technol. 49 (2012) 521–529. https://doi.org/10.1007/s13197-011-0449-2.
[20] P. Mishra, A. Herrero-Langreo, P. Barreiro, J.M. Roger, Detection and quantification of peanut traces in wheat flour by near infrared hyperspectral imaging spectroscopy using principal-component analysis, J. Near Infrared Spectrosc. 23 (2015) 15–22. https://doi.org/10.1255/jnirs.1142.
[21] P. Mishra, C.B.Y. Cordella, D.N. Rutledge, P. Barreiro, J.M. Roger, B. Diezma, Application of independent components analysis with the JADE algorithm and NIR hyperspectral imaging for revealing food adulteration, J. Food Eng. 168 (2016) 7–15. https://doi.org/10.1016/j.jfoodeng.2015.07.008.
[22] D.G. Manolakis, G. Shaw, Detection algorithms for hyperspectral imaging applications, IEEE Signal Process. Mag. 19 (2002) 29–43. https://doi.org/10.1109/79.974724.
[23] J.M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, J. Chanussot, Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5 (2012) 354–379. https://doi.org/10.1109/JSTARS.2012.2194696.
[24] L.L. Scharf, B. Friedlander, Matched Subspace Detectors, IEEE Trans. Signal Process. 42 (1994) 2146–2157.
[25] P. Geladi, J. Burger, T. Lestander, Hyperspectral imaging : calibration problems and solutions, Chemom. Intell. Lab. Syst. 72 (2004) 209–217. https://doi.org/10.1016/j.chemolab.2004.01.023.
[26] K. Jörgensen, V. Segtnan, K. Thyholt, N. Tormod, A comparison of methods for analysing regression models with both spectral and designed variables, J. Chemom. 18 (2004) 451–464. https://doi.org/10.1002/cem.890.