Folch Fortuny, A.; Prats-Montalbán, JM.; Cubero-García, S.; Blasco Ivars, J.; Ferrer, A. (2016). VIS/NIR hyperspectral imaging and N-way PLS-DA models for detection of decay lesions in citrus fruits. Chemometrics and Intelligent Laboratory Systems. 156:241-248. doi:10.1016/j.chemolab.2016.05.005.

VIS/NIR hyperspectral imaging and N-way PLS-DA models for detection of decay lesions in citrus fruits

A. Folch-Fortuny[a,*], J.M. Prats-Montalbán[a], S. Cubero[b], J. Blasco[b], A. Ferrer[a]

[a]*Multivariate Statistical Engineering (GIEM). Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universidad Politècnica de Valencia, Camino de Vera s/n, Edificio 7A, 46022 Valencia, Spain*

[b]*Centro de Agroingeniería, Instituto Valenciano de Investigaciones Agrarias (IVIA), Cra. Moncada-Náquera, km 5, 46113 Moncada, Valencia, Spain*

Abstract

In this work an N-way partial least squares regression discriminant analysis (NPLS-DA) methodology is developed to detect symptoms of disease caused by *Penicillium digitatum* in citrus fruits (green mould) using visible/near infrared (VIS/NIR) hyperspectral images. To build the discriminant model a set of oranges and mandarins was infected by the fungus and another set was infiltrated just with water for control purposes. A double cross-validation strategy is used to validate the discriminant models. Finally, permutation testing is used to select a few bands offering the best correct classification rates in the validation set. The discriminant models developed here can be potentially implemented in a fruit packinghouse to detect infected citrus fruits at their arrival from the field with affordable multispectral (3-5 channels) cameras installed in the packinglines.

Keywords: hyperspectral imaging, NIR, NPLS-DA, variable selection, permutation test

1. Introduction

Citrus production exceeded 115 million tons in 2011 [1]. They are cultivated in over one hundred countries world wide, being Spain one of the most important producer countries and the world leader in fresh citrus exports [1]. Citrus are, indeed, the most widely produced fruits for human consumption, especially oranges (62%) and mandarins (23%). To ensure product quality and reduce production losses, it is mandatory to enhance postharvest handling in citrus packinghouses. Many issues arise in this process due to pathological diseases in fruits. This problem can be potentially harmful, since a small set of rotten and sporulated fruits can contaminate the whole batch, especially during storage or transport. *Penicillium digitatum* (the cause of green mould) and *Penicillium italicum* (the cause of blue mould) are two examples of the most deleterious fungi

causing fruit decay, and they affect several cultivars over the world [2,3].

Green mould lesions at early stages cannot be detected with the naked eye because the appearance of the damage is very similar to the appearance of sound fruit. The first symptoms of these disease appear as a slightly discolored soft, water soaked around a point of injury. The spot expands rapidly to a 30-40 mm diameter. As the infection advances, a white fungal growth appears on the surface of the rot [4]. Before the sporulation, the appearance of the lesions is very similar to the sound skin being difficult for the workers to detect damaged fruit, especially when they work on an inspection table, examining fruit traveling at high speed. Therefore, the application of visual inspection or computer vision systems based on colour images is limited. Nowadays, novel machine vision technologies are being incorporated in the citrus postharvest to detect this dangerous disease, mostly based on ultraviolet (UV) induced fluorescence. Ogawa *et al*. [5] presented a system to detect decay lesions in citrus using fluorescence images, and Blanc *et al*. [6] patented an automatic machine for in-line decay detection and fruit sorting using UV illumination. However, Momin *et al*. [7] demonstrated that different cultivars of citrus fruits have different excitation wavelengths to produce UV induced fluorescence in the infected areas, which makes it difficult to create a system valid for all cases based only on this technology. Also, this kind of automatic detection can be potentially jeopardized by fluorescence measurements from other non-related defects [8]. Alternatively, this disease can often be observed using other techniques like image backscattering [9,10] or hyperspectral imaging (HSI) [11]. In this sense, different hyperspectral sensors are being investigated to detect non-visible fruit damage [12] like decay lesions in citrus fruits [13].

Using spectral devices, a set of images is obtained at different wavelengths, capturing a huge amount of chemical information. Some works have been focused on reducing the redundant information in this procedure, compressing the high-dimensional original variable space into a low-dimensional one that preserves the main properties of the data. Gomez-Sanchis *et al*. [14] and Lorente *et al*. [15] used the features from spectral images of infected fruit as inputs for classification algorithms, in order to improve the discrimination between sound and symptomatic skin. In addition, HSI systems have also been developed to detect other dangerous diseases. Qin *et al*. [16] used a portable imaging spectrograph to acquire hyperspectral images of red grapefruits affected by canker and other defects. In that work, the spectral images of the different defects were analysed using principal component analysis (PCA) [17] and spectral information divergence as classification method, detecting 97.6% of infected fruits. In Qin *et al*. [18], the authors exploited the bands selected using PCA and correlation analysis to obtain a system capable of detecting the canker using ratios of two bands. Afterwards, a system to detect canker lesions in-line was

developed by Qin *et al*. [19]. Also, PCA and band ratios were used by Li *et al*. [20,21] to select relevant bands for the detection of this disease among other common defects.

Multivariate Image Analysis (MIA) uses a wide number of models and approaches to deal with hyperspectral images [22,23]. PCA is probably the most used method within MIA (some examples are shown in the previous paragraph), but other two-way methods are commonly used, as partial least squares regression (PLS) [24] or multivariate curve resolution (MCR) [25]. In some cases, it is convenient or interesting to use three-way models such as N-way PLS (NPLS) [26] or Tucker [27].

The aim of this work is to develop multivariate models based on hyperspectral images able of discriminating between infected and sound citrus fruits while at the same time reducing as much as possible the number of wavelengths used. For this, we will use NPLS discriminant analysis (NPLS-DA) [28,29] to build a latent variable-based regression model using specific features extracted from a pool of images of different orange and mandarin cultivars collected at the Instituto Valenciano de Investigaciones Agrarias (IVIA) (Valencia, Spain). This kind of models has been succesfully applied in many research works within fruit industry, *e.g*. for tomato [30], coffee [31], loquats [32] and apple [33] discrimination. In our case, the present study represents an attempt to implementing automatic classification procedures in fruit packinghouses to prevent the storage of infected citrus fruits, which may ultimately rot and sporulate causing contamination of packinghouse facilities and spread of the disease to healthy stored fruit.

The structure of the paper is as follows. Section 2 gives specific details on the data and the image acquisition. In Section 3 the data preprocessing, feature extraction and latent variable modelling are described. Section 4 shows the results of the multivariate discriminant. Finally, some conclusions are drawn on Section 5.

2. Materials

Eight different orange and mandarin varietes are analysed in this paper: Clementine, Navel Lane Late, Mioro, Nadorcott, Nova, Salustiana, Blood orange, and Washington Navel. In each variety, 150 fruits were harvested from the field collection of the Citrus Germplasm Bank at the IVIA (Spain) [34]. After two days of storage with controlled temperature and humidity, 100 fruits of each variety were inoculated with a concentration of 10^6 spores/ml of *P. Digitatum* [35]. These citrus fruits represent the fungus group. The remaining 50 fruits were inoculated with water, and they

represent the control group to know if the innoculation process influences the results. Both inoculations were produced around 2 days after the fruit collection.

Between 1 and 4 days after inoculation, when the fruit started to show slight external symptons of decay, a camera coupled with a visible/near infrared (VIS/NIR) liquid crystal tunable filter (LCTF) was used at IVIA to obtain a RGB and hyperspectral images from each fruit of each variety. Figure 1 shows the RGB images of a control and an infected mandarin, in order to illustrate how difficult is to discriminate between both classes with visual inspection. 44 wavelentghs were registered from 650 to 1080 nm with a resolution of 10 nm. Each image has 1040 times 1392 pixels per wavelentgh. Therefore, the hyperspectral images can be represented as $1040 \times 1392 \times 44$ datacubes.

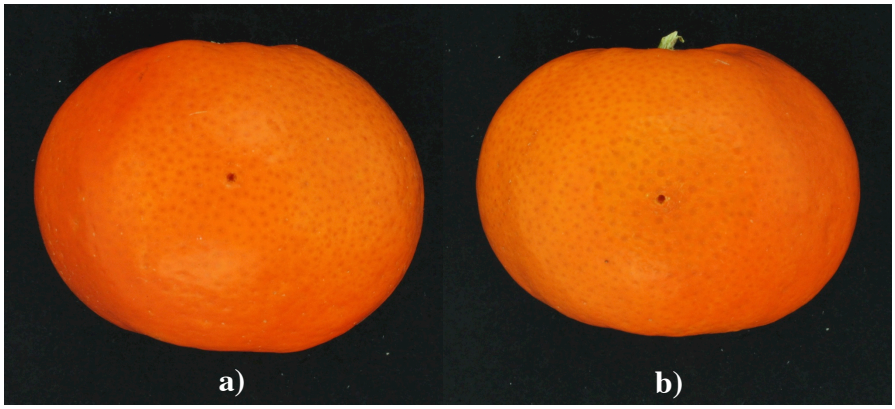


Figure 1: RGB images of a control (a) and an infected (b) mandarin.

3. Methods

3.1. Data preprocessing

The citrus fruits appear centered in the images (see Figure 2, first block of images). The spherical shape of the fruits causes some undesirable effects in the fruit images, one of the most important being that the pixels in the borders (pale blue areas around the fruit in Figure 2) appear darker than those in the centre of the fruit due to the reflexion laws of the light. Therefore, it is convenient to remove the pixels near the border from the analysis, which was done in this experiment by applying a mask. After defining an intensity threshold, pixels exceeding this limit were selected, representing the inner area of the fruit (see Figure 2, second block of images). The pixel selection is performed at each wavelength of the image. Then, the joint area across all wavelengths is defined as the mask for

the whole image. This way, if a pixel is above the threshold for, at least, one wavelength, it is guaranteed that it is included in the fruit mask. This procedure is repeated for all fruits in each variety.

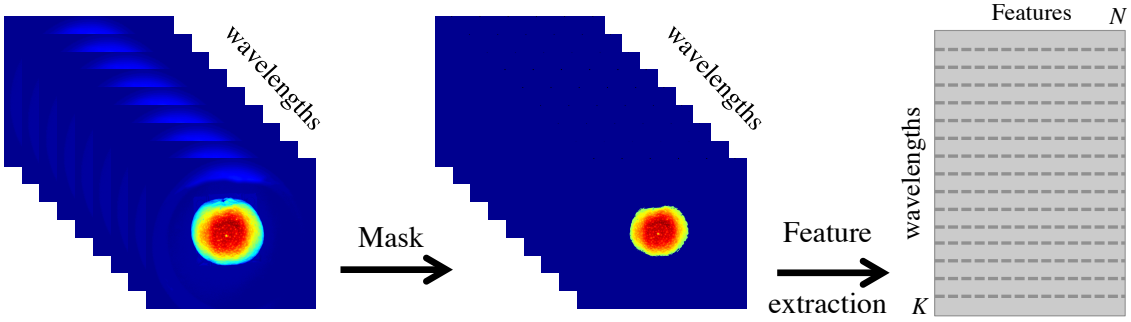


Figure 2: Hyperspectral image preprocessing. First, a mask was calculated using all wavelengths for a particular citrus fruit, and then the area inside the limits was used to perform the feature extraction. A vector with first order statistics was extracted for each wavelength image.

Five different data preprocessings were applied in this study. The first one consisted of analysing the images using the original intensities (no preprocessing), i, measured with the VIS/NIR-LCTF system. The second one consisted of transforming the intensities into reflectance values, r, using black (b) and white (w) references taken with the HSI system:

$$r = 100 \times \frac{i - i_b}{i_w - i_b} \tag{1}$$

The third preprocessing consisted of obtaining the absorbance values, a, from the reflectance. That is:

$$a = log_{10}(\frac{r}{100}) \tag{2}$$

The fourth and fifth preprocessings consisted of applying multiplicative scatter correction (MSC) and standard normal variate (SNV) methods [36] to the absorbance values, respectively. The complete study presented in Section 4 was reproduced using the five different preprocessings. A table with the main results is shown in the Appendix.

3.2. Feature extraction

Once the mask was applied, each wavelength image was converted into a one-dimensional numerical array using an image-based approach [23] (see Figure 2). In each vector a set of first order statistics were included as features describing the corresponding wavelength image. Specifically, the mean, standard deviation, and third to fifth order moments were used. After feature extraction, the data were arranged in a 3-way data cube, containing the whole set of fruits in each variety by rows, the features by columns, and the 44 wavelengths as third mode (see Figure 3).

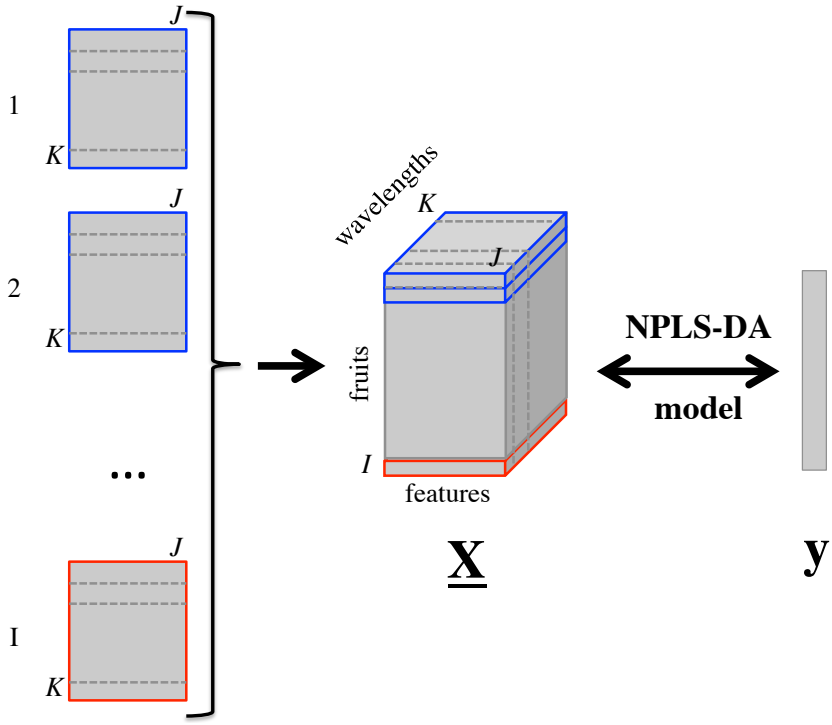


Figure 3: NPLS-DA modelling. The feature matrices extracted from each citrus fruit are arranged as row slices of the three-way array **X**. Then, the datacube is used jointly with the dummy variable **y**, representing the fungus/control group, in the NPLS-DA model.

3.3. N-way partial least squares discriminant analysis (NPLS-DA)

In this work we were interested in selecting a few wavelengths (ideally three to five) to discriminate between control and fungus oranges. For this, a variable selection performed within 2-way PLS could be applied. However, PLS allows for interactions between features and wavelengths, so a solution using a combination of different features at several wavelengths can be obtained. Since we were interested in using all features of a subset of wavelengths, and our data set is a three-way structure, we decided to apply N-way PLS. Therefore, N-way PLS [37] was proposed for studying

the three-way data structure with discriminant purposes between the control and fungus citrus fruits (*i.e.* NPLS-DA). NPLS is the natural extension of PLS to N-way structures, which tries to maximize the covariance between the $\underline{\mathbf{X}}$ and $\mathbf{Y}$ data arrays.

In the present study, the $\underline{\mathbf{X}}$ $(I \times J \times K)$ data matrix is the datacube represented in Figure 3. It is worth noting that here we are not considering the datacube of each image (described in Section 2). For the NPLS-DA analysis we build a new three-way array using the $(K \times J)$ feature array (see Figure 2) of each of the I oranges. Each row slice of $\underline{\mathbf{X}}$ represent the set of features of citrus fruit i (therefore $K = 44$ wavelengths and $J = 5$ features). Considering $\mathbf{X}$ $(I \times JK)$ the unfolded version of the datacube $\underline{\mathbf{X}}$, NPLS tries to find latent spaces $\mathbf{W}^{\mathrm{K}}$ and $\mathbf{W}^{\mathrm{J}}$ that maximise the covariance between $\mathbf{X}$ and, in this case, the dummy vector $\mathbf{y}$ (containing 1s for the fungus citrus fruits and 0s for the control ones), so it can be expressed as:

$$\mathbf{X} = \mathbf{T}(\mathbf{W}^{\mathrm{K}} \otimes \mathbf{W}^{\mathrm{J}})^{\mathrm{T}} + \mathbf{R} \tag{3}$$

afterwards decomposing $\underline{\mathbf{X}}$ from $\mathbf{X}$ using the improved NPLS version expression [38], in order to obtain residuals with better statistical properties:

$$\mathbf{X} = \mathbf{TG}(\mathbf{W}^{\mathrm{K}} \otimes \mathbf{W}^{\mathrm{J}})^{\mathrm{T}} + \mathbf{R}' \tag{4}$$

here, $\otimes$ is the Kronecker product, $\mathbf{W}^{\mathrm{J}}$ and $\mathbf{W}^{\mathrm{K}}$ refer to the weights of the wavelengths mode (3rd) and of the features mode (2nd), respectively; whereas $\mathbf{T}$ matrix gathers the scores of the fruits at each component extracted, in the 1st mode, and $\mathbf{G}$ is the core array of a Tucker3 decomposition when using $\mathbf{T}$, $\mathbf{W}^{\mathrm{J}}$ and $\mathbf{W}^{\mathrm{K}}$ as loadings, in order to obtain a better (or at least not worse) approximation of the $\underline{\mathbf{X}}$ 3-way array [39]. Finally, $\mathbf{R}'$ incorporates the residuals. More details on the matrices can be found in [38,39].

3.4. Validation procedure and wavelength selection

Proper validation of discriminant models is a subtle issue in chemometrics. Here, a double cross validation strategy [40] was applied. Using this procedure, the data from each variety and treatment (fungus/control) were split in three groups with the same number of observations in each group (16 fruits) (see Figure 4, using the compact 3-way array $\underline{\mathbf{X}}$). The first group is the calibration set, used to build the NPLS-DA model. The second group is the test set, used for selecting the number of components. And the third group is the validation set, used to evaluate the predictive power of the NPLS-DA.

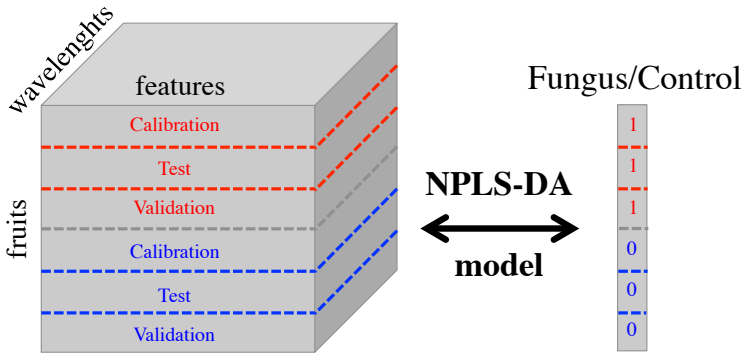


Figure 4: Data partition for the double cross validation procedure using images of a particular variety. The model was obtained using the calibration samples, the number of components was selected projecting the test samples, and the correct classification rate was obtained using the validation samples.

The ultimate goal of the present study is the creation of an affordable automatic procedure to discriminate between sound and infected fruit in packinghouses. The main drawback of using the VIS/NIR-LCTF system to obtain the spectral information is the relative high price of the equipment. On the other hand, HSI-based systems in general capture a huge amount of data that is sometimes redundant and needs large time to be acquired. Hence, there is a need to reduce the dimensionality of the data by selecting only those important wavelengths that still retain most of the information. Current state of technology allows the development of multispectral cameras capable of working in production lines with three to five charge-coupled device (CCD) sensors that can be customized to capture specific wavelengths. Hence, the goal is to perform a variable selection on the third mode of the data, the spectral bands, to assess whether a few wavelengths (three to five) have enough discriminant power to classify each fruit correctly. Permutation testing was used since it is one of the most used techniques to perform variable selection in PLS-DA [40-43].

The double-cross validation and the variable selection were performed as follows:

1. 500 different calibration, test and validation sets were built, including 32 random samples in each group: 16 fungus and 16 control citrus fruits.

2. For each of the 500 group selections:

2.1. The 32 calibration fruits were used to build NPLS-DA models, with number of components ranging from 1 to 25.

2.2. The test set was projected onto each of the 25 models to decide the number of components. The interest was in maximising first the F-score and then the parsimony of the final model. The F-score was calculated using precision and recall of the decay prediction of the NPLS-DA model. The precision, p, was computed as:

$$p = \frac{TP}{TP + FP} \tag{5}$$

where TP are the true positives, *i.e.* fungus citrus fruits correctly classified in the model, and FP the false positives, *i.e.* control fruits classified as fungus. The recall, r, was defined as:

$$r = \frac{TP}{TP + FN} \tag{6}$$

where FN refers to the false negatives, *i.e.* fungus citrus fruits classified as control. So, the F-score was computed as:

$$F = \frac{2pr}{p + r} \tag{7}$$

Therefore the F-score was maximum when all the samples were classified correctly, both control and fungus (first criterion). If this was achieved selecting different number of components, the lowest number was selected following the principle of parsimony (second criterion). See Figure 5 for an example of this selection using Nadorcott variety. The NPLS-DA model built with the calibration fruits and the components selected using the *test* set was the called the "real model".

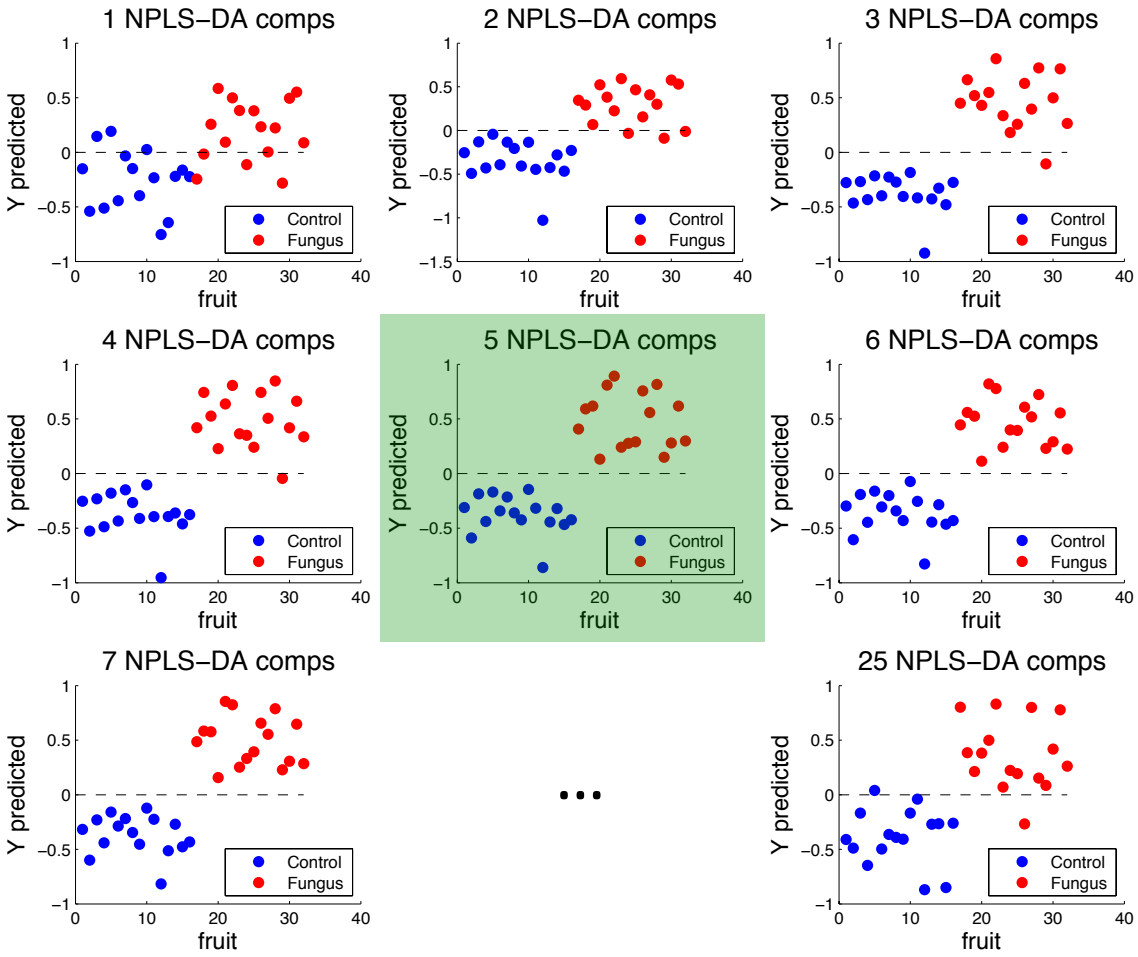


Figure 5: Predicted class for the test samples using different number of components in the NPLS-DA model fitted with the calibration data. 5 components were selected, since it was the model with highest F-score and parsimony.

2.3. The VIP values (variable importance in projection) [44,45] of the real model were collected. The VIP value of the variable (wavelength) k was computed as:

$$VIP_k^2 = \frac{K \sum_{a=1}^{A} [(w_{k,a}^K)^2 (RSSY_{a-1} - RSSY_a)]}{RSSY_0 - RSSY_A} \quad (8)$$

where $w_{k,a}^K$ is the loading value of the kth variable at the ath component, A is the number of latent variables in the NPLS-DA model and $RSSY_a$ is the residual $\mathbf{Y}$-sum of squares of the model with a components ($a = 0$ to A).

2.4. Steps 2.1-2.3 were then repeated destroying the relationships between **X** and **Y**, thus creating a random model. This was done by permuting the rows of **Y** before applying step 2.1. Finally, the VIP values using the random model were collected after i) fitting the NPLS-DA model with the *calibration* set and ii) deciding the number of components using the *test* set.

2.5. The remaining validation samples were projected onto the real model to obtain the correct classification rates. Figure 6 exemplifies the projection of the validation set in the Nadorcott variety using the model selected in step 2.2 (see Figure 5).

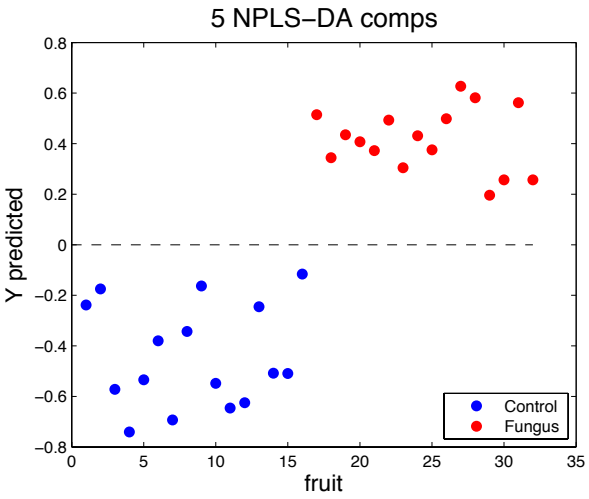


Figure 6: Validation samples projected onto the NPLS-DA model with 5 components built with calibration samples.

2.6. Steps 2.1-2.5 were repeated three times, moving the samples from group to group, that is: calibration-test-validation (first model, as in Figure 4), test-validation-calibration (second model), and validation-calibration-test (third model).

2.7. The VIP values of both the real and the random models were averaged among the three models.

2.8. The results of the external validation using the three real models were integrated.

3. Once step 2 was performed for all group selections, the statistical significance between the real and random models was assessed. The distribution of the random VIP values represents the null distribution, so the real VIP values, or the their mean, m, can be compared with the previous distribution to compute the statistical p-value, that is the probability of obtaining at random a value

equal or higher than m. Figure 7 shows an example using the VIP values of a particular wavelength in the Nadorcott variety.

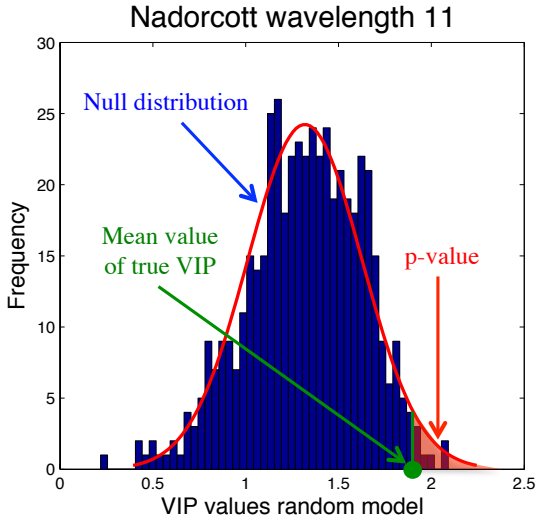


Figure 7: VIP values of a particular wavelength. The red line denotes the null distribution from the random models. The green dot represents the mean VIP value of the real models. The red area is the p-value associated to the green dot in the red null distribution.

4. The mean p-values of each wavelength across all varietes were averaged to obtain the mean p-values. Then, after sorting the p-values, the wavelengths with lower mean p-values were classified as the most discriminant variables in all fruit varieties.

5. The mean correct classification rates were obtained using the results of the validation set in the 500 models.

6. Steps 2.1, 2.2, 2.5, 2.6, 2.8 and 5 (double-cross validation procedure) were repeated using the 3, 4, 5, 10, 15, 20, 25, 30, 35, and 40 most discriminant wavelengths determined in step 4. This way, the degradation of the missclassifications was evaluated in all varieties in terms of the number of wavelengths considered in the NPLS-DA model.

3.5. Software

NPLS-DA models were fitted using N-way toolbox for MATLAB [46]. The VIP values and permutation testing were obtained using own code.

4. Results

Figure 8 shows the p-values computed using the random and real VIPs. It is clear that different distributions of p-values in the wavelengths were observed among varieties. For example, wavelengths 4-9 have the highest discriminant power (lowest p-values) in Clementine, while the best ones in Mioro are wavelengths 37-43. Despite these differences in the best bands per variety, it seems the initial 15 wavelengths, corresponding to 650-790 nm, tend to have low p-values (high discriminant power).

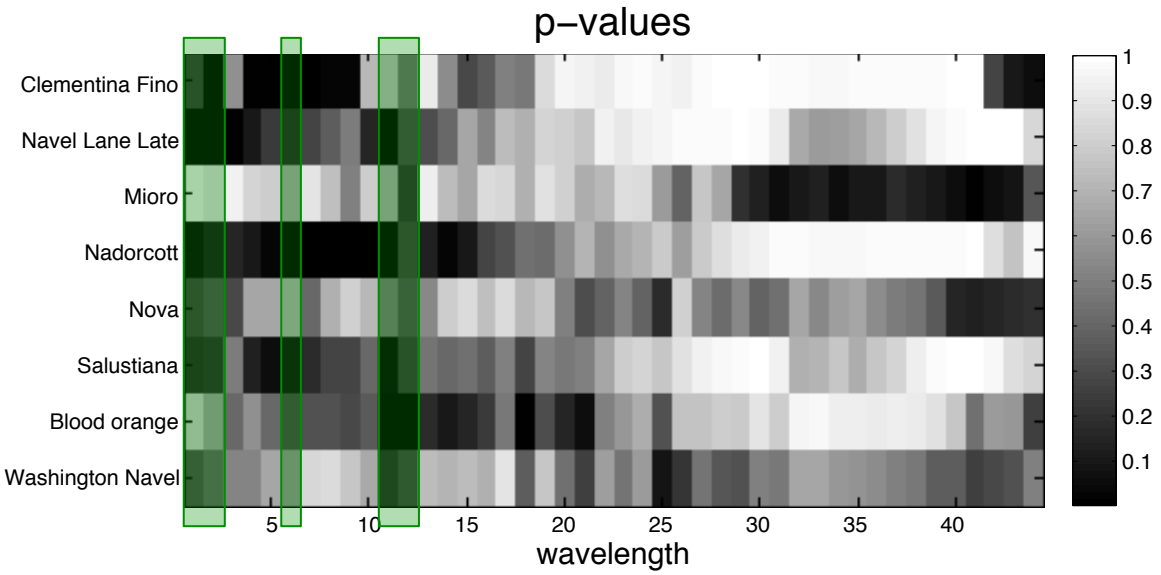
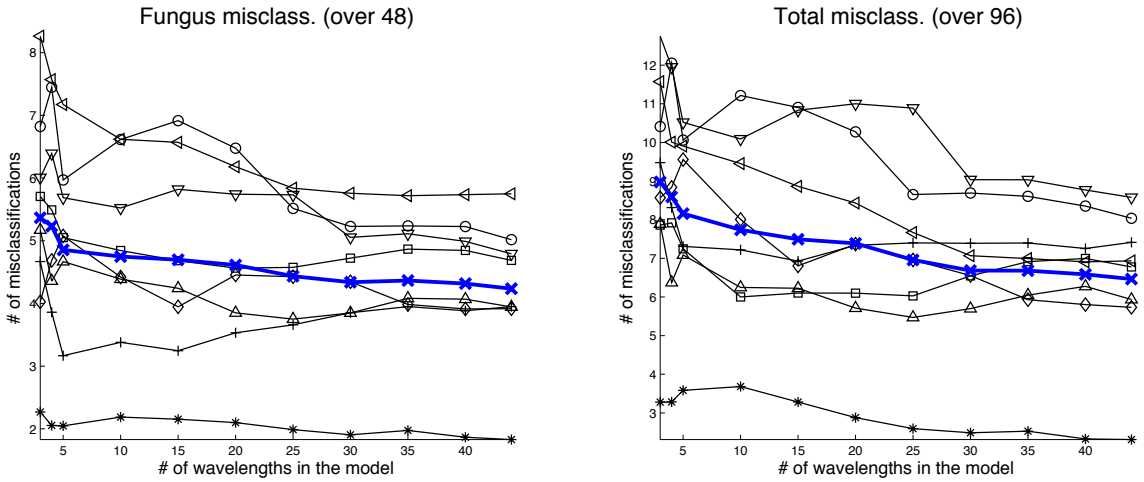


Figure 8: P-values computed using the random and the real VIPs. The darker the square, the lower is the corresponding p-value. Green areas mark the 5 best wavelengths attending to the highest mean across varieties: 1, 2, 6, 11 and 12.

From a theoretical point of view, the best choice would be to fit different NPLS-DA models (step 6.) in each variety including the wavelengths with the smallest p-values in that particular variety. From a practical point of view, this would imply to build different digital cameras incorporating different wavelengths depending on the variety.

Here, a compromise approach was applied, and we decided to select the wavelengths according to the list of ordered mean p-values obtained in step 4. The results in terms of fungus and total missclassifications can be visualized in Figure 9. The average number of missclassifications in the

fungus class decreases notably from 3 to 5 wavelengths, and then the values decrease slowly from 5 to 44 wavelengths. A similar behaviour is shown in the average total number of misclassifications.



a) b)

Figure 9: Fungus (a) and total (b) number of missclassifications when varying the number of wavelengths included in the NPLS-DA model. The black lines show the mean over the 500 models of Clementine (upper triangles), Lanelate ('+' symbols), Mioro (circles), Nadorcott (asterisks), Nova (squares), Salustiana (left triangles), Blood orange (diamonds), and Washington Navel (lower triangles). Also, the mean values over all varieties are shown in bold blue lines with crosses.

The best combination of 5 wavelengths was 1, 2, 6, 11 and 12 (see Figure 8), corresponding to 650, 660, 700, 750 and 760 nm. Table 1 shows the number of correct classifications and the corresponding percentages using all wavelenghts and only the five most discriminant ones. The variety best discriminated in both cases was Nadorcott, having 99% and 96.8% of correct classification in control and 96.2% and 95.7% in fungus fruit, respectively. The second-best classified, in terms of disease detection, were Navel Lane Late and Clementine, with near 92% of correct classification rate using all wavelengths, and around 93% and 90% using five wavelengths, respectively. Salustiana attained the lowest classification rate. The Appendix shows the results of applying different preprocessings on the original images using all wavelengths, showing that for this variety it was better to use the absorbance values or the absorbance with SNV. For the rest of varieties, it was statistically better to use the intensity values (no preprocessing). The average correct classifications rates using all wavelengths were 95.6% and 91.2% for control and fungus oranges, respectively. When using five wavelenghts the percentages decreased to 93.1% and 90.0%, respetively.

Table 1: Correct classification results in all orange and mandarin varieties using all wavelengths and only the five most discriminant ones. * denotes a statistically better dicrimination power in the corresponding class (control or fungus) between using all wavelengths or only the five most discriminant ones.

Variety	All wavelengths		Five most discriminant wavelengths	
	Control	Fungus	Control	Fungus
	Correct class./%	Correct class./%	Correct class./%	Correct class./%
Clementine	46.0* / 95.9%	44.1* / 91.8%	45.6 / 95.0%	43.3 / 90.3%
Navel Lane Late	44.5* / 92.8%	44.1 / 91.8%	43.9 / 91.4%	44.8* / 93.4%
Mioro	45.0* / 93.7&	43.0* / 89.6%	43.9 / 91.5%	42.0 / 87.6%
Nadorcott	47.5* / 99.0%	46.2* / 96.2%	46.5 / 96.8%	46.0 / 95.7%
Nova	45.9 / 95.6%	43.3* / 90.2%	45.8 / 95.4%	43.0 / 89.5%
Salustiana	46.8* / 97.5%	42.3* / 88.0%	45.3 / 94.3%	40.8 / 85.1%
Blood Orange	46.2* / 96.2%	44.1* / 91.8%	43.5 / 90.7%	42.9 / 89.4%
Washington Navel	44.2* / 92.1%	43.2* / 90.0%	43.2 / 89.9%	42.3 / 88.2%
AVERAGE	45.8 / 95.6%	43.8 / 91.2%	44.7 / 93.1%	43.1 / 90.0%
MINIMUM	44.2 / 92.1%	42.3 / 90.2%	43.2 / 89.9%	40.8 / 85.1%
MAXIMUM	47.5 / 99.0%	46.2 / 96.2%	46.5 / 96.8%	46.0 / 95.7%

To assess the statistical differences between using 5 or 44 wavelengths, a paired t-test was applied on each variety. The results are presented also in Table 1. In general, the results using the selected 5 wavelengths are statistically worse than using all wavelengths. Due to the high sample size (500)

used in the paired t-test, small differences in the number of correct classifications become statistically significant. However, comparing the results of both models, the mean loss in correct classification using five wavelengths instead of 44 is 0.8 and 1.1 fruits out of 48 fruits in fungus and control cases, respectively. Anyway, the dramatic reduction in the price of a 5-channel camera clearly compensates for the small reduction of correct classification.

5. Conclusions

N-way PLS-DA applied on features extracted from a set of hyperspectral images reveals as a powerful tool for discrimination between infected and sound citrus fruits. The methodology applied on several orange and mandarin varieties shows that, on average, 91% of fruit with decay lesions caused by *P. digitatum* can be detected at early stages when the damage is barely visible or even invisible and therefore cannot be detected in postharvest by manual inspection. The predictive models were properly validated using a double cross validation procedure, computing up to 500 models with different fruit groupings.

Permutation testing on VIP values was used here to select a few spectral channels with the most discriminant power in all citrus fruit varieties. Despite the number of correct classifications becomes stable from five selected wavelengths onwards, there exist statistically significant differences between using five and all wavelengths captured by the VIS/NIR-LCTF system, being the latter significantly better.

Nevertheless, there is a strong cost reduction by selecting a few wavelengths, since a digital camera can be customised to capture up to five filters to reproduce the VIS/NIR-LCTF hyperspectral system. Therefore, from a practical point of view, the NPLS-DA models including information from the best five wavelengths are sufficient to reduce the losses in fruit warehouses due to storage of infected fruits.

The knowledge obtained in this work is a key step towards the achievement of a potential automatic fruit sorting system using these modified cameras in which fruits are photographed and instantly classified using the predictions from the NPLS-DA model. This way, suspicious fruits can be expelled from the commercial chain prior to affecting sound fruits.

Acknowledgements

This research was partially funded by the Spanish Ministry of Science and Innovation through grants DPI2011-28112-C04-02 and DPI2014-55276-C05-1R, and by INIA through grant RTA2012-00062-C04-01. In all cases with the support of European FEDER funds. Authors thank Lluis Palou from the Centro de Tecnología Postcosecha at the IVIA for the help and supervision in the innoculation process of the fruits.

Appendix

The results of the five different preprocessings (intensity, reflectance, absorbance, absorbance + MSC, absorbance + SNV) obtained using 100 models from the 500 used in Section 4, are depicted in Table 2. Based on the results of a paired t-test applied between the intensity values and the rest of preprocessings, it is sensible to use the intensity values to fit the models. Only in the case of Salustiana, the results of absorbance and absorbance + SNV were statistically better than the intensity values.

Table 2: Correct classification rates using different preprocessing (all wavelengths). The +/- superindices mark the statistical superiority/inferiority of the results in the preprocessing compared to the raw intensity values.

Fungus oranges and mandarins										
Variety	**Intensity**		**Reflectance**		**Absorbance**		**Abs. + MSC**		**Abs. + SNV**	
	Corr.	**%**	**Corr.**	**%**	**Corr.**	**%**	**Corr.**	**%**	**Corr.**	**%**
Clementine	44.3	92.3%	43.1^{-}	89.7%	41.7^{-}	86.9%	41.4^{-}	86.2%	42.2^{-}	88.0%
Navel Lane Late	43.8	91.3%	42.5^{-}	88.4%	40.9^{-}	85.1%	34.1^{-}	71.0%	34.5^{-}	71.8%
Mioro	43.0	89.7%	43.0	89.6%	40.1^{-}	83.6%	38.0	79.1%	42.7^{-}	88.9%
Nadorcott	46.1	96.0%	46.0	95.8%	45.1^{-}	94.0%	41.5^{-}	86.5%	41.7^{-}	86.8%
Nova	43.5	90.7%	42.0^{-}	87.5%	40.3^{-}	84.0%	39.0^{-}	81.3%	41.3^{-}	86.1%
Salustiana	42.3	88.2%	43.0	89.6%	43.4^{+}	90.4%	43.8^{+}	91.2%	42.1	87.6%
Blood Orange	44.2	92.0%	42.8^{-}	89.2%	42.0^{-}	87.5%	38.2^{-}	79.5%	36.9^{-}	76.8%

Washington Navel	43.2	90.0%	43.0	89.6%	41.8[-]	87.1%	35.0[-]	72.9%	38.2[-]	79.5%
AVERAGE	43.8	91.3%	43.2	89.9%	41.9	87.3%	38.9	81.0%	40.0	83.2%
MINIMUM	42.3	88.2%	42	87.5%	40.1	83.6%	34.1	71.0%	34.5	71.8%
MAXIMUM	46.1	96.0%	46	95.8%	45.1	94.0%	43.8	91.2%	42.7	88.9%

References

[1] Citrus Fruit, Fresh and Processed, Annual Statistics 2012, Tech. rep., Food and Agriculture Organization of the United Nations (FAO) (2012). URL http://www.fao.org/

[2] L. Palou, *Penicillium digitatum*, *Penicillium italicum* (Green Mold, Blue Mold), in: Postharvest Decay, Elsevier, 2014, pp. 45–102.

[3] G.J. Holmes, J.W. Eckert, Sensitivity of *Penicillium digitatum* and *P. italicum* to Postharvest Citrus Fungicides in California, Phytopathology 89(9) (1999) 716–721.

[4] R.A. Fullerton, J.L. Tyson, P.R. Sale, Citrus diseases. In: Growing Citrus in New Zealand. A practical guide. (Pauline Mooney ed). New Zealand Citrus Growers Inc, Wellintong, New Zealand (2001).

[5] Y. Ogawa, M. M. Abdul, M. Kuramoto, Y. Kohno, T. Shiigi, K. Yamamoto, N. Kondo, Detection of Rotten Citrus Fruit Using Fluorescent Images, The Review of Laser Engineering 39(4) (2011) 255–261.

[6] P. G. R. Blanc, I. J. Blasco, G. E. Moltó, S. J. Gómez, G. S. Cubero, System for the automatic selective separation of rotten citrus fruits, international Classification B07C5/342; Cooperative Classification G01N2021/845, G01N21/85, G01N2021/8466, B07C5/363, B07C5/3422 (Dec. 2013).

[7] M.A. Momin, N. Kondo, M. Kuramoto, Y. Ogawa, K. Yamamoto, T. Shiigi, Investigation of Excitation Wavelength for Fluorescence Emission of Citrus Peels based on UV-VIS Spectra, Engineering in Agriculture, Environment and Food 5(4) (2012) 126–132.

[8] D. Obenland, D. Margosan, J. L. Smilanick, B. Mackey, Ultraviolet Fluorescence to Identify Navel Oranges with Poor Peel Quality and Decay, Hort Technology 20(6) (2010) 991–995.

[9] D. Lorente, M. Zude, C. Regen, L. Palou, J. Gómez-Sanchis, J. Blasco, Early decay detection in

citrus fruit using laser-light backscattering imaging, Postharvest Biology and Technology 86 (2013) 424–430.

[10] D. Lorente, M. Zude, C. Idler, J. Gómez-Sanchis, J. Blasco, Laser-light backscattering imaging for early decay detection in citrus fruit using both a statistical and a physical model, Journal of Food Engineering 154 (2015) 76–85.

[11] J. Gómez, J. Blasco, E. Moltó, G. Camps-Valls, Hyperspectral detection of citrus damage with Mahalanobis kernel classifier, Electronics Letters 43(20) (2007) 1082–1084.

[12] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O.L. García-Navarrete, J. Blasco, Recent Advances and Applications of Hyperspectral Imaging for Fruit and Vegetable Quality Assessment, Food and Bioprocess Technology 5(4) (2011) 1121–1142.

[13] J. Gómez-Sanchis, D. Lorente, E. Soria-Olivas, N. Aleixos, S. Cubero, J. Blasco, Development of a Hyperspectral Computer Vision System Based on Two Liquid Crystal Tuneable Filters for Fruit Inspection. Application to Detect Citrus Fruits Decay, Food and Bioprocess Technology 7(4) (2014) 1047–1056.

[14] J.Gómez-Sanchis, J. Blasco, E. Soria-Olivas, D. Lorente, P. Escandell-Montero, J.M. Martínez-Martínez, M. Martínez-Sober, N. Aleixos, Hyperspectral LCTF-based system for classification of decay in mandarins caused by *Penicillium digitatum* and *Penicillium italicum* using the most relevant bands and non-linear classifiers, Postharvest Biology and Technology 82 (2013) 76–86.

[15] D. Lorente, J. Blasco, A.J. Serrano, E. Soria-Olivas, N. Aleixos, J.Gómez-Sanchis, Comparison of ROC Feature Selection Method for the Detection of Decay in Citrus Fruit Using Hyperspectral Images, Food and Bioprocess Technology 6(12) (2012) 3613–3619.

[16] J. Qin, T.F. Burks, M.A. Ritenour, W.G. Bonn, Detection of citrus canker using hyperspectral reflectance imaging with spectral information divergence, Journal of Food Engineering 93(2) (2009) 183–191.

[17] J.E. Jackson, A User's Guide to Principal Components, Wiley Series in Probability and Statistics, 2004.

[18] J. Qin, T.F. Burks, X. Zhao, N. Niphadkar, M.A. Ritenour, Multispectral Detection of Citrus Canker Using Hyperspectral Band Selection, Transactions of the ASABE 54(6) (2011) 2331–2341.

[19] J. Qin, T.F. Burks, X. Zhao, N. Niphadkar, M.A. Ritenour, Development of a two-band spectral imaging system for real-time citrus canker detection, Journal of Food Engineering 108(1) (2012) 87–93.

[20] J. Li, X. Rao, Y. Ying, D. Wang, Detection of navel oranges canker based on hyperspectral imaging technology, Nongye Gongcheng Xuebao/Transactions of the Chinese Society of

Agricultural Engineering 26 (2010) 222–228.

[21] J. Li, X. Rao, Y. Ying, Development of algorithms for detecting citrus canker based on hyperspectral reflectance imaging, Journal of the Science of Food and Agriculture 92(1) (2012) 125–134.

[22] P. Geladi, H. Grahn, Multivariate Image Analysis, 1st Edition, Wiley, Chichester; New York, 1997.

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

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

[25] R. Tauler, Multivariate curve resolution applied to second order data, Chemometrics and Intelligent Laboratory Systems 30(1) (1995) 133–146.

[26] J. M. Prats-Montalbán, A. Ferrer, Integration of colour and textural information in multivariate image analysis: defect detection and classification issues, Journal of Chemometrics 21(1-2) (2007) 10–23. 12

[27] J. M. Prats-Montalbán, M. Cocchi, A. Ferrer, N-way modeling for wavelet filter determination in multivariate image analysis, Journal of Chemometrics 29(6) (2015) 379–388.

[28] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.

[29] B. Matthew, R. William, Partial least squares for discrimination, Journal of Chemometrics 17 (2003) 166–173.

[30] S. Shrestha, L. C. Deleuran, M. H. Olesen, R. Gislum, Use of Multispectral Imaging in Varietal Identification of Tomato, Sensors (Basel, Switzerland) 15(2) (2015) 4496–4512.

[31] R. Calvini, A. Ulrici, J. M. Amigo, Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging, Chemometrics and Intelligent Laboratory Systems 146 (2015) 503–511.

[32] K.-Q. Yu, Y.-R. Zhao, Z.-Y. Liu, X.-L. Li, F. Liu, Y. He, Application of Visible and Near-Infrared Hyperspectral Imaging for Detection of Defective Features in Loquat, Food and Bioprocess Technology 7(11) (2014) 3077–3087.

[33] J.-F. Gao, H.-L. Zhang, W.-W. Kong, Y. He, Nondestructive discrimination of waxed apples based on hyperspectral imaging technology, Spectroscopy and Spectral Analysis 33(7) (2013) 1922–1926.

[34] L. Navarro, J.A. Pina, J. Juarez, J.F. Ballester-Olmos, J.M. Arregui, C. Ortega, A. Navarro, N. Duran-Vila, J. Guerri, P. Moreno, M. Cambra, S. Zaragoza, The citrus variety improvement program in Spain in the period 1975-2001. In: Proceedings of the 15th Conference of the International Organization of Citrus Virologists. IOCV, Riverside, 2002, pp. 306-316.

[35] L. Palou, J. Smilanick, J. Usall, I. Viñas, Control postharvest blue and green molds of oranges by hot water, sodium carbonate, and sodium bicarbonate, Plant Disease 85 (2001) 371-376.

[36] T. Fearn, C. Riccioli, A. Garrido-Varo, J. E. Guerrero-Ginel, On the geometry of SNV and MSC, Chemometrics and Intelligent Laboratory Systems 96(1) (2009) 22–26.

[37] R.Bro, Multiway calibration. Multilinear PLS, Journal of Chemometrics 10(1) (1998) 47–61.

[38] R. Bro, A. K. Smilde, S. de Jong, On the difference between low-rank and subspace approximation: improved model for multi-linear PLS regression, Chemometrics and Intelligent Laboratory Systems 58(1) (2001) 3–13.

[39] A.Smilde, R.Bro, P.Geladi, Multi-way Analysis: Applications in the Chemical Sciences, John Wiley & Sons, 2004.

[40] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J.v. Velzen, J.P.M.v. Duijnhoven, F.A.v. Dorsten, Assessment of PLSDA cross validation, Metabolomics 4(1) (2008) 81–89.

[41] G. Quintás, N. Portillo, J. C. García-Cañaveras, J.V. Castell, A. Ferrer, A. Lahoz, Chemometric approaches to improve PLSDA model outcome for predicting human non-alcoholic fatty liver disease using UPLC-MS as a metabolic profiling tool, Metabolomics 8(1) (2012) 86–98.

[42] E. Szymanska, E. Saccenti, A. K. Smilde, J. A. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 8(1) (2012) 3–16.

[43] S. Bijlsma, I. Bobeldijk, E. R. Verheij, R. Ramaker, S. Kochhar, I. A. Macdonald, B. van Ommen, A. K. Smilde, Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation, Analytical Chemistry 78(2) (2006) 567–574.

[44] A. Conesa, J.M. Prats Montalbán, S. Tarazona, M.J. Nueda, A. Ferrer, A multiway approach to data integration in systems biology based on Tucker3 and N-PLS, Chemometrics and Intelligent Laboratory Systems 104(1) (2010) 101–111.

[45] S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems 129 (2013) 76-86.

[46] C.A. Andersson, R. Bro, The N-way Toolbox for MATLAB, Chemometrics and Intelligent Laboratory Systems 52(1) (2000) 1–4.