Journal of Food Engineering 292 (2021) 110357

Contents lists available at ScienceDirect

Journal of Food Engineering

journal homepage: http://www.elsevier.com/locate/jfoodeng

Pine nut species recognition using NIR spectroscopy and image analysis

Roberto Moscetti [a,**], Daniel Hagos Berhe [a,b], Mariagrazia Agrimi [a], Ron P. Haff [c], Peishih Liang [c], Serena Ferri [d], Danilo Monarca [d], Riccardo Massantini [a,*]

[a] Department for Innovation in Biological, Agro-food and Forest System, University of Tuscia, Via S. Camillo de Lellis Snc, 01100, Viterbo, Italy
[b] Department of Natural Resources Management, Adigrat University, P. O. Box 50, Adigrat, Tigrai, Ethiopia
[c] United States Department of Agriculture, Agricultural Research Service, Western Regional Research Center, 800 Buchanan St, Albany, CA, 94710, United States
[d] Department of Agricultural and Forestry Sciences, University of Tuscia, Via S. Camillo de Lellis Snc, 01100, Viterbo, Italy

ARTICLE INFO

Keywords:
Computer vision
Colour
Shape and size
Chemometrics
Pinus pinea L.
Pinus sibirica Du Tour

ABSTRACT

NIR spectroscopy and physical properties derived from image analysis were evaluated as potential features for the classification of seed kernels from two pine nut species (*P. pinea* L. and *P. sibirica* Du Tour) using Partial Least Squares Discriminant Analysis (PLS-DA). Model performances were evaluated in terms of specificity, sensitivity and accuracy. Data pre-treatments were essential for achieving excellent performances (accuracy rate > 95%) in all tests. The interval PLS-DA highlighted that the most important features for (1) the NIR method were the absorption bands at 1640–1658, 1720–1738 and 1880–1998 nm, while for (2) the image analysis were kernel eccentricity, kernel major axis length, kernel lightness (L*) and kernel perimeter. The results demonstrate potential of both techniques for discriminating the two pine nut species.

1. Introduction

Pine nuts are edible seeds harvested from pinecones (family *Pinaceae*, genus *Pinus*) commonly used in cuisines worldwide. Although generally referred to as nuts, they are in fact classified as seeds with the edible part (the embryo) surrounded by a hard shell. While unshelled pine nuts have a long shelf life if kept dry, shelled nuts deteriorate rapidly and are susceptible to rancidity. Pine nuts have been harvested for human consumption since prehistoric times (Awan and Pettenella, 2017). Both naturally occurring stands and cultivated plantations of pine species are found in Asia, Europe, the Near East and North America (Nergiz and Dönmez, 2004). About 20 species are harvested in significant quantities, the most important for human consumption being *P. pinea* L. (Mediterranean stone pine), *P. koraiensis* Siebold & Zucc. (Korean pine), and *P. sibirica* Du Tour (Siberian pine). Nuts from the pine species *Pinus pinea* L. (commonly named Mediterranean stone pine or Italian stone pine) have long been an important component of the Mediterranean diet.

Pine nuts are a small (<1%) but rapidly growing segment of the global tree nut market. In fact, over the three-year period between 2015/2016 and 2018/2019 global production nearly doubled from around 18,600 metric tons to 34,000 metric tons (INC, 2017; Statista, 2019). Because the demand always exceeds the supply; prices vary widely by region (Sharashkin and Gold, 2004), and have been reported as high as 100 €/kg (retail) (Calama et al., 2016). Production costs are high in Europe due to multiple factors. Intensive manual labour is required for collection of cones and for separating the nuts from the shells, as no automated methods have yet been developed. Climate change has contributed to declines in pine nut production, which is also due to the spread of the western conifer seed bug (*Leptoglossus occidentalis* Heidemann) with the fungus *Diplodia sapinea* in some Mediterranean countries. Furthermore, local supplies are also insufficient to meet demand in North America (Awan and Pettenella, 2017; Parks, 2017; Vanhanen and Savage, 2013). Thus, most pine nuts are imported from China, Korea and Pakistan, with China being the main exporting country with 78% of total exports (INC, 2017).

The chronic global shortage of pine nuts from the traditional species, especially *P. pinea*, has led to introduction into the marketplace of nuts of other species, mostly but not exclusively *P. sibirica* and *P. koraiensis* from China (Ballin, 2012; Loewe et al., 2017). The latter nuts are sold in local markets at lower prices and are not easily distinguishable from *P. pinea* by consumers, even though they have different flavour, shape and size (Evaristo et al., 2010). In fact, while significant differences exist in price and quality of nuts from different species, this is not transparent

* Corresponding author. Tuscia University, Department for Innovation in Biological, Agro-food and Forest system, S. Camillo de Lellis snc, 01100, Viterbo, Italy.
** Corresponding author.
E-mail addresses: rmoscetti@unitus.it (R. Moscetti), massanti@unitus.it (R. Massantini).

https://doi.org/10.1016/j.jfoodeng.2020.110357
Received 16 May 2020; Received in revised form 27 August 2020; Accepted 20 September 2020
Available online 21 September 2020

to consumers since they are not reported on product labels (Awan and Pettenella, 2017; Mutke et al., 2012). Tracking information for nut production and distribution streams is not well maintained as processing centres lose track of points of origin (Awan and Pettenella, 2017; Mutke et al., 2012). Most nuts exported from China have their origins in other countries (Sharashkin and Gold, 2004). Shipping documents often lack product status details such as whether they are cones, unshelled pine nuts, or shelled kernels (Awan and Pettenella, 2017; Sharashkin and Gold, 2004), making the tracing of product origins difficult. Given the difficulty of product tracking coupled with visual similarities between species, the likelihood of product adulteration (undeclared mixing of lower value product with high value product) is high. For pine nuts, the issue of adulteration has been recently elevated to a health issue as well as an economic issue because of the occurrence dysgeusia, or Pine Nut Syndrome (PNS) in which consumers can experience a strong metallic or acid aftertaste that lasts up to 14 days in most cases but can last as much as 42 days in extreme cases (Awan and Pettenella, 2017). PNS has been linked to *P. armandii* Franch (Awan and Pettenella., 2017), its subspecies and other varieties, (Matthäus et al., 2018; Mikkelsen et al., 2014).

Given the economic and potential health impacts of undeclared mixing of pine nut species for consumer consumption it would be advantageous to develop methods to identify nut origins at all points of distribution (Loewe et al., 2017). Traceability has the potential to improve product quality and safety, provide geographic identification, and hinder black market trade. Finally, it would facilitate organic certification labelling (Mutke et al., 2012).

The use of NIR spectroscopy to detect adulteration in food products is widely reported and reviews of this research are available in the literature (Valand et al., 2020). NIR spectroscopy has also been reported as one potential method for tracing product origins, although available literature, especially for pine nuts, is limited. Tigabu et al. (2005) used NIR to trace the origins of seeds of Scots pine (*Pinus. Sylvestris* L.) for reforestation purposes. Loewe et al. (2017) used NIR spectroscopy for the discrimination of Mediterranean stone pine nuts from Chilean plantations. To the best of our knowledge, these are the only reported studies using NIR spectroscopy to discriminate the origin of pine nuts, while no literature on the image analysis of pine nuts is available.

The objective of this research is to assess the potential use of both NIR spectroscopy and image analysis for the discrimination of pine nuts species of different geographical ranges and cultivations.

2. Materials and methods

2.1. Sample preparation

Whole, dried and decorticated pine nuts kernels of species certified as *P. pinea* L. and *P. sibirica* Du Tour were obtained either from local markets (batches 1 to 5) or provided by the BIOSIC company (Viterbo, Central Italy) (batches 6 to 9), respectively (Table 1). Batches from different production areas were selected so that trained classification models would be insensitive to geographical origins, as well as to minimize the selection bias and secure chemical and physico-chemical representativeness to a large extent. One hundred samples (i.e. kernels) from each batch were randomly selected and assigned as either 'Class I' (*P. pinea* L.) or 'Class II' (*P. sibirica* Du Tour). Selection was performed manually by removing kernels with discoloration, infestation, infection or other visually apparent flaws. The 'Class I' and 'Class II' groups were thus composed of 500 and 400 kernels, respectively.

Table 1
List of the batches arranged into classes (i.e. pine nut species), which were used for the experimental activity.

Species	Class	Batch #	Packaging Country
Pinus pinea L.	1	1	Italy
		2	Spain
		3	Unknown
		4	Italy
		5	Italy
Pinus sibirica Du Tour	2	6	China
		7	Russia, Altai
		8	Russia, far east
		9	Russia, Buryatia

2.2. Spectral acquisition

A Luminar 5030 Acousto-Optic Tunable Filter-Near Infrared (AOTF-NIR) Miniature 'Hand-held' Analyzer coupled with the bundled 'SNAP! 2.04' software (Brimrose Corp., Baltimore, USA) was used to acquire NIR spectra. The analyzer was equipped with reflectance post-dispersive optics, a pre-aligned dual beam lamp assembly (5 W halogen lamp), and an indium gallium arsenide (InGaAs) array (range 1100–2300 nm, 2 nm resolution) with an integrating time of 60 ms. The instrument was allowed to warm up for at least an hour to reach a stable state before the use. The reference spectrum was automatically measured by the second detector of the instrument, which was a dual-beam spectrophotometer (Moscetti et al., 2016). Because each NIR measurement of the instrument covers only a single point, each sample was measured on opposite sides of the kernel in order to acquire more representative data. Diffuse reflectance spectra were transformed into absorbance ($A = \log_{10} R^{-1}$) and the average spectrum of each kernel was used for further computations.

2.3. Spectral pre-treatments

NIR spectra quite often suffer from problems of unwanted spectral variations and baseline shifts. In the present study the most common baseline correction methods (i.e. standard normal variation, SNV; multiplicative scatter correction, MSC; and Savitzky-Golay derivative) were tested to evaluate the presence of light scattering, which can be detrimental to quantitative/qualitative analysis and lead to inaccurate results. In a solid matrix, light scattering can be related to the variability in refractive index, morphology (e.g., surface roughness) and density of sample. SNV and MSC methods mathematically differ but are similar in their outcomes. The Savitzky-Golay first, second or third derivative (i.e. D1*f*, D2*f* and D3*f*, respectively) were used alone or in combination with SNV/MSC for its capability of improving the baseline correction as well as resolving overlapping bands, as resolution enhancement method. Each derivative pre-treatment employed a second or third order polynomial fitted over a window of five, seven, nine, or eleven features (Moscetti et al., 2017) with the aim of improving the signal to noise ratio of the spectra. The narrowest window size (i.e. filter length of 3 features) was excluded to circumvent noise inflating of the original spectrum due to derivative calculation. The highest window size was set to 11 to avoid rounding-off of the peaks and troughs, which happens when the filter length is too wide. Every possible combination of these spectral pre-treatments was tested and only the best model, in terms of performance metrices, was retained. Finally, spectral data were mean centered.

2.4. Image acquisition and image segmentation

Allowing for the exploratory nature of this research activity, a flatbed scanner was chosen as the imaging acquisition system as it is better suited for scanning still images than a digital camera, which is more suitable for high-speed applications (Sun, 2020). Thus, digital colour images of the kernels were acquired one batch at a time (i.e. 100 kernels per scan) using a CM2320nf (Hewlett-Packard, Palo Alto, USA) flatbed scanner with VueScan 9.2.11 Professional Edition software (Hamrick Software, Phoenix, AZ, USA). The scanner was profiled using the Colour Checker Passport target (X-Rite Ltd., U.K.). Each batch was scanned three times and averaged for the final image. The scan was performed only on one side of the sample because (1) all selected nuts were free from apparent visually flaws and discolouration and (2) to

avoid redundant data from spatial features. Each scan was conducted with the following parameters: resolution of 2466 × 3498 pixels (240 dpi); 48-bit colour intensity resolution (16 bits per RGB channel), and; Digital Negative (DNG) raw file format. Each DNG image was colour corrected by applying the scanner profile through the Camera Raw 6.0 software (Adobe Systems Inc., San Jose, CA) then saved in Tagged Image File format. Finally, image segmentation was performed by distinguishing kernels from the image background using a custom script written in Python 3.7.0 coupled with OpenCV 3.4.1, employing Otsu's binarization for clustering-based image thresholding (OpenCV, 2020).

2.4.1. CIELab colour measurements

The custom script also extracted the average colour data from each kernel. Specifically, the 'cvtColor ()' function of the OpenCV library was applied for RGB-to-CIELab colour space conversion. CIELab was chosen as a uniform colour space and then as a coordinate system in which perceived colour differences correspond to Euclidean distances. Thus, results were expressed in terms of lightness (L^*), red/green colour (a^*), yellow/blue colour (b^*), hue angle (h) and chroma (C^*) (Moscetti et al., 2013a). Moreover, the CIE 1976 colour difference ($\Delta E^* = [\Delta L^{*2} + \Delta a^{*2} + \Delta b^{*2}]^{1/2}$) was computed to evaluate how much the two species deviated from each other in terms of pine nut colour. Its value was expressed according to the evaluation scale used by Cecchini et al. (2011): imperceptible ($\Delta E^* < 1$), minimal ($1 \leq \Delta E^* < 2$), just perceptible ($2 \leq \Delta E^* < 3$), perceptible ($3 \leq \Delta E^* < 5$), strong difference ($5 \leq \Delta E^* < 12$) and different colour ($\Delta E^* \geq 12$).

2.4.2. Shape and size measurements

Kernel perimeter (mm), major and minor axis lengths (mm), surface area (mm^2) and eccentricity were extracted from each image using the 'regionprops ()' function (Matlab R2015a 'Image Processing' toolbox). The kernel perimeter was computed as the sum of the distances between adjacent pixels around the edge of the kernel. Major and minor axis lengths were computed on the ellipse with the same normalized second central moments (rotational inertia) of the kernel region. Area was acquired by calculating the number of pixels in the kernel, and eccentricity was computed as the ratio of the distance between the ellipse foci and the major axis length. Size and shape parameters were converted from pixel to metric unit (mm) based on the reference embedded in the Colour Checker Passport target (X-Rite Ltd., U.K.). In total, 10 features based on imaging data were extracted from each kernel image as features for classification models (Table 2).

3. Imaging features were scaled before model development

3.1. Classification models development

Spectral-based and imaging-based classification models were developed separately using Partial Least Squares Discrimination Analysis (PLS-DA) through the SIMPLS algorithm (de Jong, 1993). PLS-DA seeks linear combinations of the original independent variables with the ability to discriminate data classes while discarding irrelevant and unstable information and solving collinearity issues. In addition, Interval PLS-DA (iPLS-DA) was used to select subsets of features which could still achieve good prediction results (Xing and Guyer, 2008). The iPLS-DA algorithm used the stepwise forward mode to select a maximum of 10 intervals of 10 features each for spectral model and 10 intervals of 1 feature each for imaging model. The number of the interval which provides the lowest RMSECV was selected by the algorithm, allowing to circumvent under-/overfitting problems. Both selectivity ratio and β-coefficient were computed to assess the relative contribution of each feature or subset of features to the performance of each model (Rajalahti et al., 2009). The larger the selectivity ratio and/or the absolute value of β-coefficient, the more useful the given feature is in classification.

Table 2
Summary of descriptive statistics of the imaging features of pine nuts from *P. pinea* L. and *P. sibirica* Du Tour species. Mean values belonging to the same factor without common letters are statistically different according to HSD ($P \leq 0.05$).

Factor	Minimum	Q1[a]	Median	Mean		Q3[b]	Maximum	SE[c]
Lightness (L^*)								
P. pinea	74.09	81.19	82.48	82.23	a	83.59	86.78	0.08
P. sibirica	72.54	77.87	78.66	78.62	b	79.41	83.64	0.08
Red/green colour (a^*)								
P. pinea	−1.88	0.39	0.70	0.71	a	1.05	3.36	0.02
P. sibirica	−1.81	−0.35	0.18	0.20	b	0.73	3.15	0.05
Yellow/blue colour (b^*)								
P. pinea	11.28	15.13	16.32	16.95	b	17.94	30.91	0.12
P. sibirica	11.38	16.68	18.95	19.89	a	23.16	35.34	0.25
Hue angle (h)								
P. pinea	81.89	86.43	87.54	87.58	b	88.64	94.74	0.07
P. sibirica	80.44	87.68	89.46	89.15	a	90.95	94.11	0.14
Chroma (C^*)								
P. pinea	11.29	15.16	16.34	16.97	b	17.97	30.92	0.12
P. sibirica	11.39	16.72	18.97	19.91	a	23.17	35.36	0.25
Perimeter (mm)								
P. pinea	17.83	27.82	29.90	29.94	a	31.99	39.36	0.13
P. sibirica	16.99	20.68	22.31	23.30	b	25.81	34.57	0.20
Surface area (mm^2)								
P. pinea	21.82	45.83	52.54	52.98	a	60.10	82.56	0.42
P. sibirica	19.98	29.83	34.40	37.52	b	44.24	70.02	0.58
Eccentricity								
P. pinea	0.79	0.89	0.91	0.90	a	0.92	0.96	1.2E-03
P. sibirica	0.55	0.76	0.80	0.79	b	0.83	0.94	3.5E-03
Minor axis lenght (mm)								
P. pinea	3.73	4.90	5.30	5.32	ns	5.73	7.21	0.03
P. sibirica	3.71	4.85	5.21	5.34	ns	5.78	7.32	0.04
Major axis lenght (mm)								
P. pinea	7.07	11.69	12.72	12.75	a	13.79	18.02	0.07
P. sibirica	6.62	7.75	8.63	8.93	b	10.15	15.38	0.09

[a] First quartile.
[b] Third quartile.
[c] Standard error, ns = no significant difference.

In a first step, samples (spectra or images) for each class were randomly split between calibration set (75% or 675 kernels) and prediction set (25% or 225 kernels). As PLS regression performs a dimensionality reduction, it is essential to test the model and select the correct number of latent variables to find the optimal trade-off between underfitting and over-fitting. Thus, a model optimization was conducted using venetian blinds cross-validation with 10 data splits. Root Mean Square Error for calibration, cross-validation and prediction calculations (RMSEC, RMSECV and RMSEP, respectively) were used to evaluate each discriminant model with the purpose of selecting the optimal number of latent variables and, thus, to circumvent unrealistic results (Moscetti et al., 2015). In a second step, the prediction set was used to give an independent assessment of the accuracy, precision and robustness of the calibration model, following the criteria contained in the adopted validation guidelines (Broad et al., 2006). The Hotelling's t-squared and Q residuals were used in combination to identify potential outliers of each given model. Chemometrics was performed using Matlab software R2015a coupled with PLS_Toolbox software v8.1 (Eigenvector Research Inc., WA, USA).

The classification performance of selected models was tested in terms of sensitivity (Eq. (1)), selectivity (Eq. (2)) and accuracy (Eq. (3)) (Dejaegher et al., 2011). While accuracy rates are correlated with model predictivity, sensitivity and specificity rates are generally used to evaluate robustness (i.e. the capability of the model to resist to small changes in test conditions).

$$Sensitivity\ rate = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (1)$$

$$Selectivity\ rate = \frac{True\ Negatives}{False\ Positives + True\ Negatives} \quad (2)$$

$$Accuracy\ rate = \frac{True\ Positives + True\ Negatives}{Total\ Positives + Total\ Negatives} \quad (3)$$

3.2. Moisture content analysis

Moisture content was measured following the official oven-drying method 'Moisture in Dried Fruits' - AOAC 934.06. Results were obtained at batch level and expressed as a percentage by mass (grams per 100 g) (Horwitz, 2005). The analysis was performed with the aim of (i) assessing the potential difference in terms of moisture content between classes and (ii) evaluating the impact of moisture content on developed NIR models, as it is well known that NIR spectroscopy is sensitive to the water content of a food matrix.

3.3. Data handling and data analysis

One-way analysis of variance (ANOVA) was performed to evaluate statistical differences between classes in terms of moisture content, colour as well as shape and size. The Tukey's pairwise comparison method was performed, and the Honestly Significant Difference (HSD) was calculated for an appropriate level of interaction ($P \leq 0.05$). Results were reported as the mean and standard error of the mean. Data handling and ANOVA were both performed using R v3.3.3 software in combination with 'dplyr' v0.5.0 and 'agricolae' v1.2-4 R-packages (CRAN, 2017).

4. Results and discussion

4.1. Data overview

4.1.1. Spectral data

Although visual inspection of the spectra is not sufficient to distinguish kernels from different species, their graphical overview may help understanding spectral differences between classes. Fig. 1 thus illustrates the mean absorbance spectra of pine nuts from *P. pinea* and *P. sibirica* in the full NIR spectral range from 1100 to 2300 nm both without (a) and with (b) the best spectral pre-treatments. The best classification performance was obtained following spectral pre-treatments of SNV (standard normal variate) in combination with 1st order derivative of 2nd order polynomial over window 9, regardless of the classification algorithm used (i.e. PLS-DA or iPLS-DA). It means that the classification model was negatively affected by light scattering, while improved by noise filtering and unravelling overlapping bands. The PLS-DA model was based on the first two latent variables that were able to capture much of the variability between species. In the same Figure, the 10-feature intervals selected by the iPLS-DA algorithm are also showed. Each interval represents a window of 10 adjacent features which gave superior prediction over those obtained using all variables in the spectral data set.

4.1.2. Imaging data

Summary statistics and ANOVA results for image-based features are shown in Table 2. Lightness, hue angle, chroma, yellow/blue colour and red/green colour of kernels showed significant variation between species ($P \leq 0.05$) with higher values of lightness and red/green colour for *P. pinea* and higher values of hue angle, chroma and yellow/blue colour for *P. sibirica*. In other words, *P. sibirica* showed a yellower and vivid, but darker colour than *P. pinea*. This colour difference between the two species can be detectable by human eye as corroborated by the ΔE^*, which assumed an average value of 4.68 (i.e., $3 \leq \Delta E^* < 5$, perceptible difference of colour) and ranged from a minimum of 1.55 (i.e., $1 \leq \Delta E^* < 2$, minimal difference of colour) to a maximum of 26.37 (i.e., $\Delta E^* \geq 12$, different colours).

Kernel shape and size were significantly different ($P \leq 0.05$) in terms of perimeter, major axis, eccentricity and area, although the minor axis did not significantly differ between the two species. Thus, *P. pinea* was bigger and more elliptical (i.e. higher eccentricity) then *P. sibirica*.

While discrimination based on imaging techniques have been widely reported for many commodities including walnuts (Calama et al., 2017; Ercisli et al., 2012; Huang et al., 2016; Kuo et al., 2016; Liu et al., 2015; Lurstwut and Pornpanomchai, 2018; Menesatti et al., 2008; Pallottino et al., 2009; Rodríguez-Pulido et al., 2012; Wu et al., 2018; Zhang et al., 2016), none have been reported involving pine nuts making comparison to this study problematic.

4.2. Classification models

4.2.1. Prediction models based on NIR spectral data

Fig. 2a and b shows score plots derived from the PLS-DA and iPLS-DA models, respectively, for the prediction set. Good separation between classes was observed for both models (species). No outliers were detected for either classification model. Models affected by outliers were discarded due to low classification performances which were the result of the lack of proper selection of data pre-treatments and number of latent variables. Table 3 showed the performance metrices of both models. PLS-DA yielded 10 misclassifications versus 6 for iPLS-DA, corresponding to very good accuracy rates equal to 96% and 98%, respectively. For PLS-DA, 80% of misclassifications were for *P. pinea* and this increased to 100% for the iPLS-DA model. This suggests that derived models intending to evaluate the purity of *P. pinea* product might tend to have higher false negative (*P. pinea* misclassified; 1- sensitivity) than false positive results (*P. sibirica* misclassified; 1- specificity).

Misclassifications might be related to variances among features within each class or noise. Higher accuracy for iPLS-DA is not unexpected since it uses feature regions rather than discrete features (i.e. wavebands rather than wavelengths) thereby focusing on important spectral regions and removing interference from other regions (Xiaobo et al., 2010).

The PLS-DA model was characterized by 2 latent variables, while iPLS-DA only by one. The observed decrease in number of latent

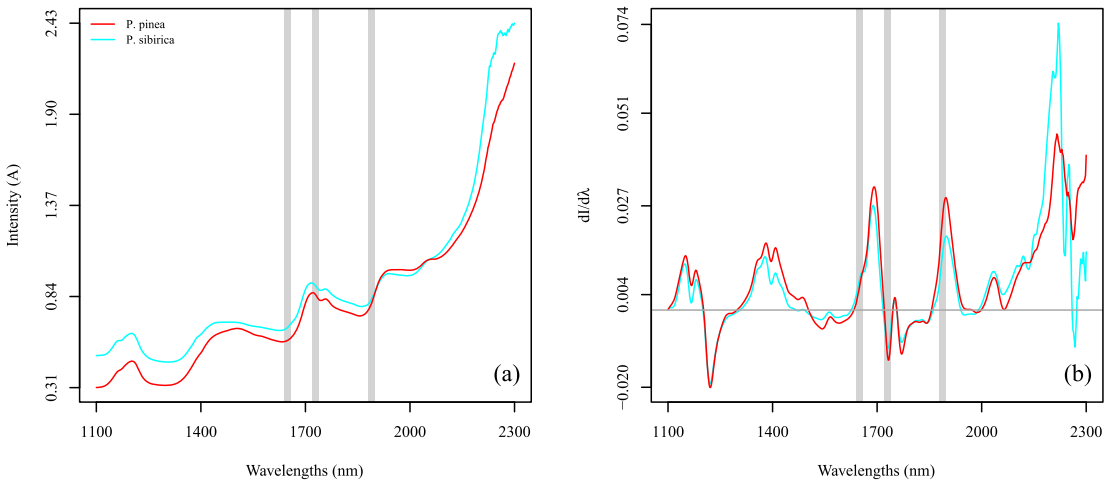

Fig. 1. Mean raw (a) and pre-treated (b) absorbance spectra for both *P. pinea* L. and *P. sibirica* Du Tour species. Spectral pre-treatment consisted of the Standard Normal Variate scatter correction followed by the 1st derivative Savitzky-Golay filter with 9-smoothing points. The vertical straight stripes represent the 10-features intervals at 1640–1658 nm, 1720–1738 nm and 1880–1898 nm selected by the iPLS-DA algorithm.

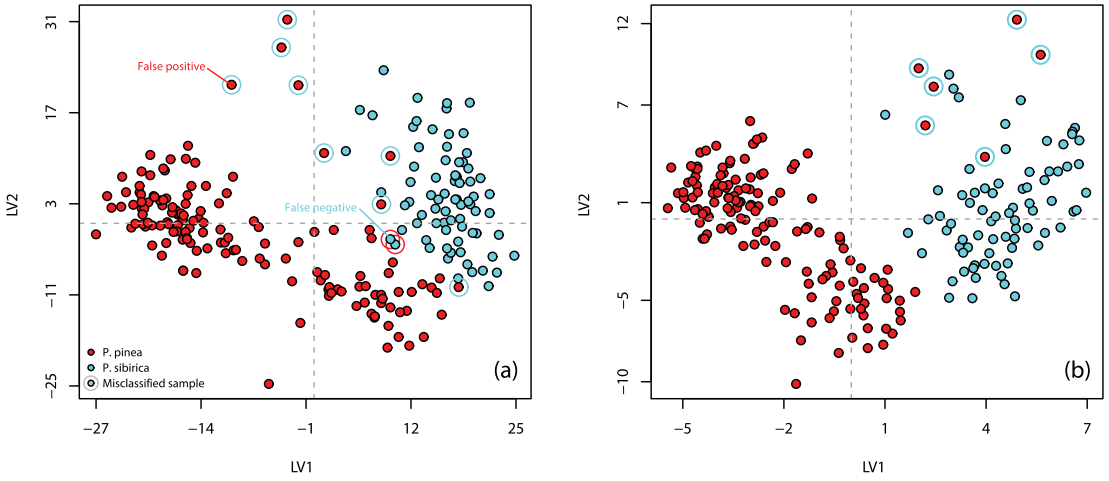

Fig. 2. Spatial distribution plots obtained from PLS-DA (a) and iPLS-DA (b) models developed using NIR spectral features of the prediction set. Red point with cyan outline corresponds to false positive error (i.e. *P. sibirica* sample erroneously classified as belonging to *P. pinea*). Cyan point with red outline corresponds to false negative error (i.e. *P. pinea* sample erroneously classified as belonging to *P. sibirica*). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

variables is quite common after variable selection, further reducing the complexity of models and making it easier to interpret. Therefore, it is also important to consider that a single latent variable may not be enough to allow a model to exploit all the available information, leading to the risk of underfitting. In our case, the choice for a single latent variable for the iPLS-DA model was driven by cross-validation, which proved that the additional latent variables did not significantly improve model performance.

To evaluate the contribution of each individual feature on the PLS-DA model two diagnostic methods were computed and compared: (i) the selectively ratio (Fig. 4a) and (ii) the ß coefficients (Fig. 4c). Conventionally, features with ß coefficients larger than ±2 times standard deviation of the regression vector were selected as important wavebands. Based on this, spectral regions around 1644, 1722 and 1858 nm were ranked as the most useful. However, this approach may lead to selecting or ignoring features with low or high contribution to the model, respectively (Rajalahti et al., 2009). Consequently, the selectivity ratio was also computed for its capability to highlight important features by combining predictive power (β-coefficients) with explanatory power (variance/covariance among features). Using this new diagnostic method, the cut-off ratio between the explained and the residual variance was computed and features above the cut-off were recognised as important for the model. In fact, the results indicated that the absorption bands in the regions around 1270, 1410–1450, 1550–1630, 1720 and 1870 nm had the highest selectivity ratio (up to 6.42), i.e. highest contribution to the model. However, below ~1250 nm, between 1724 and 1866 nm, as well as beyond ~1900 nm observed selectivity ratios were lower than the cut-off, suggesting a lower contribution to the model.

The iPLS-DA algorithm iteratively selected 10-feature intervals which provided the lowest model root-mean-square error of cross-validation (RMSECV). Accordingly, the absorption bands were: (i) 1640–1658 nm, (ii) 1720–1738 nm and (iii) 1880–1898 nm. These bands partially corresponded to the marker wavelengths selected through the selectivity ratio and, in general, are associated with the (i) lipids, (ii) proteins and (iii) carbohydrates, as well as moisture (Loewe et al., 2017; Tigabu et al., 2005; Workman and Weyer, 2008).

Absorption peaks have been reported in different NIR spectral regions of agricultural commodities and attributed to a variety of molecular processes. For instance, Loewe et al. (2017) reported that absorption peaks observed at 1200, 1500, 1720, 1760, 1940 and 2350 nm in spectra of Mediterranean pine nuts grown in Chile corresponded

Table 3
Summary of performance metrics for classification algorithm (i.e. PLS-DA and iPLS-DA) complexity which gave the best results for both analytical methods used in the experimentation (i.e. NIR spectroscopy and image analysis). The pre-treatments associated to each model were applied in combination.

Method	Algorithm	Features	Data pre-treatments			LVs[f]		Sensitivity			Specificity			Accuracy		
			SC[c]	Savitzky-Golay filter		n.	Variance (%)	C[g]	CV[h]	P[i]	C	CV	P	C	CV	P
				Derivative	Smoothing points											
NIR	PLS-DA[a]	Whole spectrum	SNV[d]	D1*f*	9	2	54.28	0.96	0.96	0.95	0.99	0.98	0.97	0.97	0.97	0.96
	iPLS-DA[b]	1) 1640–1658 nm 2) 1720–1738 nm 3) 1880–1898 nm	SNV	D1*f*	9	1	90.29	0.98	0.98	0.96	0.99	0.99	1.00	0.98	0.98	0.98
Imaging	PLS-DA	All imaging features				4	83.98	0.95	0.95	0.98	0.97	0.97	0.99	0.96	0.96	0.98
	iPLS-DA	1) Lightness (L*) 2) Red/green (a*) 3) Hue angle (h) 4) Eccentricity 5) Major Axis Length				1	52.89	0.95	0.95	0.97	0.97	0.97	0.97	0.96	0.96	0.97

[e]SP, Savitzky-Golay smoothing points.
[a] PLS-DA, Partial Least Squares Discriminant Analysis.
[b] iPLS-DA, Interval Partial Least Squares Discriminant Analysis.
[c] SC, Scatter Correction method.
[d] SNV, Standard Normal Variate.
[f] LVs, number of Latent Variables.
[g] C, calibration.
[h] CV, cross-validation.
[i] P, prediction.

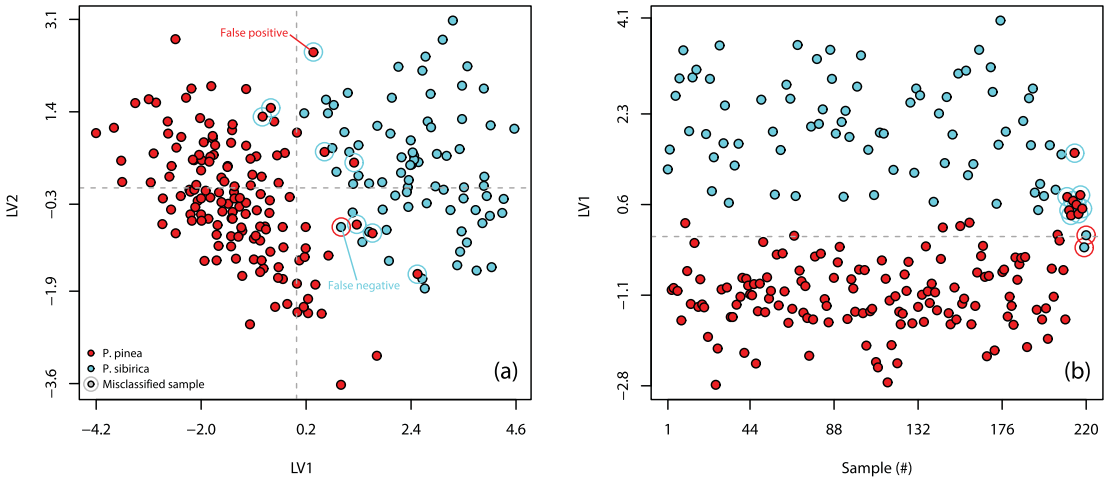

Fig. 3. Spatial distribution plots obtained from PLS-DA (a) and iPLS-DA (b) models developed using imaging features of the prediction set. Red point with cyan outline corresponds to false positive error (i.e. *P. sibirica* sample erroneously classified as belonging to *P. pinea*). Cyan point with red outline corresponds to false negative error (i.e. *P. pinea* sample erroneously classified as belonging to *P. sibirica*). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

to the most important wavebands for discrimination of product origin. The importance of these bands in derived classification models was attributed to known absorption peaks for proteins, amino acids and moisture (1200 and 1940 nm), lipids (1500; 1720; 1760 and 2350 nm), and polysaccharides and carbohydrates (2350 nm). Tigabu et al. (2005) identified strong NIR absorption peaks in the spectra of pine nut seeds from different sources at 1422 and 1930 nm along with less visible bands around 1500 and 2200 nm which were the most important features for discriminant models. Moreover, Moscetti et al. (2013b) reported complex bands in hazelnut spectra attributable to various molecular vibrations of functional groups, some linked with fatty acid content, unsaturated fatty acids with cis double bonds, protein acid and ester stretching vibrations, and others to polysaccharides, carbohydrates and lipids.

The moisture content of both kernel classes was measured to determine whether water bands had the potential to affect the NIR-based model development. The results showed a significantly different moisture content between *P. pinea* (5.36 ± 0.06%) and *P. sibirica* (3.22 ±

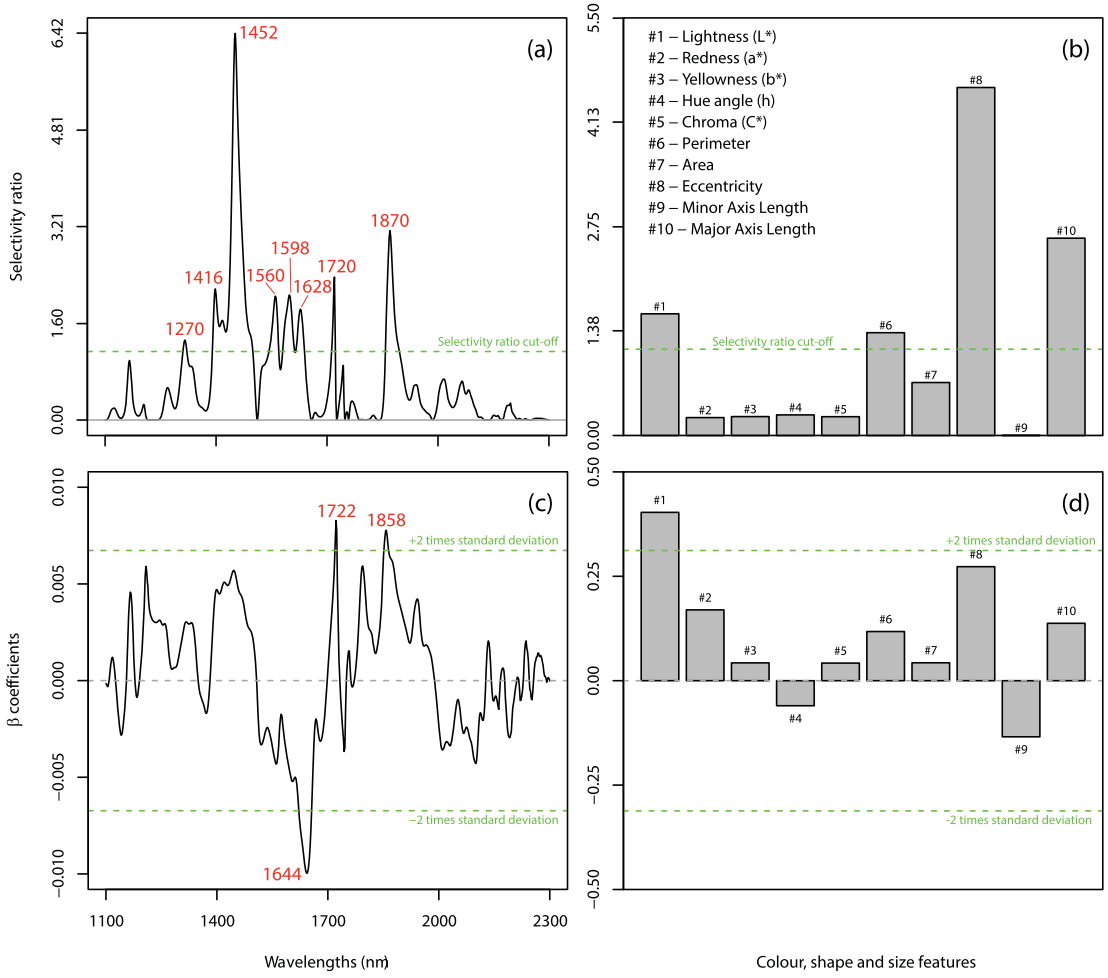

Fig. 4. Selectivity ratio plots for the PLS-DA models based on NIR spectral features (a) and imaging features (b). The horizontal green-dashed line corresponds to cut-off ratio between the explained and the residual variance. The larger the selectivity ratio, the more useful the given feature was for the classification task. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

0.09%). Results agreed with the literature, which showed that kernels from *P. pinea* had a higher moisture content (~5.10–5.60%) than kernels belonging to *P. sibirica* (<3.50%) (Evaristo et al., 2010; Nergiz and Dönmez, 2004). Therefore, both NIR-based models (i.e. PLS-DA and iPLS-DA) were probably affected by moisture content. In fact, the selectivity ratio showed that features close to the ~1480- and ~1900 nm water bands had high importance for both PLS-DA and iPLS-DA models. Considering that *"pine nut kernels should have a moisture content not exceeding 3.5* per *cent, except for Pinus pinea, which should not exceed 6.0* per *cent and Pinus gerardiana, which should not exceed 7.0* per *cent"* (UNECE, 2013), the moisture content was expected to significantly improve discriminant performances of models and, thus, the water bands were not excluded.

4.2.2. Prediction models based on imaging data

Fig. 3 reports the score plots of the PLS-DA and iPLS-DA of image analysis for the prediction set. Both models showed very good and similar performance metrices ranging from 95 to 98% (Table 3). Misclassifications were slightly higher for iPLS-DA (11 versus 9) with both heavily weighted towards misclassification of *P. pinea* as was also observed above for NIR-based models. Similar to the observation for the NIR-based models, the imaging-based iPLS-DA models was characterized by only one latent variable. In all cases, the sensitivity, specificity and accuracy rates showed similarity towards calibration, cross-validation and prediction, indicating that the models were robust. As observed for the NIR-based models, outliers were not detected for the selected imaging-based models.

Fig. 4b and d shows the selectivity ratio and the β-coefficients bar plots of the PLS-DA model, respectively. Regarding the β-coefficients, lightness (L*) was the only feature characterized by a value larger than ±2 standard deviation over the regression vector. However, similar to the observation for the NIR-based model, the selectivity ratio suggested a higher number of marker features. This suggests that β-coefficients may underestimate the feature's contribution in the model under the studied experimental conditions. In fact, the selectivity ratio indicated that among the 10 imaging features, lightness, perimeter, eccentricity and major axis length had the strongest contributions in the PLS-DA model, with eccentricity and major axis length scoring highest.

The features selected by the iPLS-DA algorithm were mostly the same as those observed as marker wavelengths for the PLS-DA model. In fact, they consisted of 3 colour attributes (i.e. lightness, red/green colour and hue angle) and 2 spatial properties (i.e. eccentricity and major axis length). Size features (i.e. area and perimeter) were discarded, suggesting that shape recognition played a major role in the classification of the two pine nut species considered in this study.

5. Conclusions

Pine nuts are widely consumed in domestic and foreign markets and, despite their importance, the industry faces market challenges related to adulteration and subsequent potential for PNS. This study addresses these challenges by demonstrating the potential use of NIR spectroscopy

and image analysis to distinguish pine nuts from different geographic origins.

NIR spectra of two species of pine nuts were subjected to different pre-treatments and presented as input features for PLS-DA and iPLS-DA discrimination. Absorption bands at 1640–1658 nm, 1720–1738 nm and 1880–1998 nm were found to be the most important for classification purposes.

Various features derived from image analysis, including CIELab colour space and measured physical properties, were also tested for their ability to distinguish classes. The dominant features in the resulting models were eccentricity, major axis length, and perimeter, based on calculations of selectivity ratios. PLS-DA model performance based on four latent variables was above 95% accuracy in classifying the pine nuts. The iPLS-DA model required only one latent variable to achieve greater than 95% accuracy for both calibration and prediction.

Based on the present study findings, it can be concluded that either NIR spectroscopy or image analysis coupled with chemometrics have potential for the classification of pine nuts species. Use of these techniques could improve the traceability of pine nuts, which is essential for controlling quality and the incidence of pine nut syndrome (PNS). However, prior to the implementation of this approach in industry further study is recommended to; (i) elucidate the major chemical constituents, i.e. moisture, fat, and protein content, and fatty acid profile, related to the spectral ranges on which classification models rely; and (ii) to validate each model with larger sample sizes and different regions, production years, agro-pedo-climatic conditions, and species. In addition, the effect of (i) fluctuations in the moisture content of fruit, (ii) excluding NIR water bands, and/or (iii) using spectral and spatial data in combination (i.e. multi-/hyperspectral imaging) should also be investigated in the future.

Conflict of interest and authorship conformation form

Please check the following as appropriate:

- ☑ All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- ☑ This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- ☑ The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

CRediT authorship contribution statement

Roberto Moscetti: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Daniel Hagos Berhe:** Writing - original draft, Formal analysis, Investigation, Resources, Formal analysis. **Mariagrazia Agrimi:** Conceptualization, Resources, Writing - original draft, Writing - review & editing, Project administration. **Ron P. Haff:** Resources, Writing - original draft, Writing - review & editing, Visualization. **Peishih Liang:** Resources, Writing - original draft, Writing - review & editing, Visualization. **Serena Ferri:** Supervision. **Danilo Monarca:** Conceptualization, Supervision, Funding acquisition. **Riccardo Massantini:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgments

The authors gratefully acknowledge (1) the 'Departments of excellence 2018' program (i.e. 'Dipartimenti di eccellenza') of the Italian Ministry of Education, University and Research for the financial support through the 'Landscape 4.0 food, wellbeing and environment' (DIBAF department of University of Tuscia, Italy); (2) the BIOSIC srl (Viterbo, Central Italy) for providing samples; and (3) MSc Gianpaolo Moscetti and Dr. Swathi Sirisha Nallan Chakravartula for the English language revision of the manuscript.

References

Awan, H.U.M., Pettenella, D., 2017. Pine nuts: a review of recent sanitary conditions and market development. Forests 8. https://doi.org/10.3390/f8100367.

Ballin, N.Z., 2012. Investigating cases of taste disturbance caused by pine nuts in Denmark. Case Studies in Food Safety and Authenticity. Elsevier, pp. 318–325. https://doi.org/10.1533/9780857096937.6.318.

Broad, N., Graham, P., Hailey, P., Hardy, A., Holland, S., Hughes, S., Lee, D., Prebble, K., Salton, N., Warren, P., Leiper, K., 2006. Guidelines for the development and validation of near-infrared spectroscopic methods in the pharmaceutical industry. Handb. Vib. Spectrosc. https://doi.org/10.1002/0470027320.s8303.

Calama, R., Fortin, M., Pardos, M., Manso, R., 2017. Modelling spatiotemporal dynamics of Pinus pinea cone infestation by Dioryctria mendacella. For. Ecol. Manage. 389, 136–148. https://doi.org/10.1016/j.foreco.2016.12.015.

Calama, R., Gordo, J., Madrigal, G., Mutke, S., Conde, M., Montero, G., Pardos, M., 2016. Enhanced tools for predicting annual stone pine (Pinus pinea L.) cone production at tree and forest scale in Inner Spain. For. Syst. 25 https://doi.org/10.5424/fs/2016253-09671.

Cecchini, M., Contini, M., Massantini, R., Monarca, D., Moscetti, R., 2011. Effects of controlled atmospheres and low temperature on storability of chestnuts manually and mechanically harvested. Postharvest Biol. Technol. 61, 131–136.

Cran, 2017. Comprehensive R archive network.

de Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. Chemometr. Intell. Lab. Syst. 18, 251–263.

Dejaegher, B., Dhooghe, L., Goodarzi, M., Apers, S., Pieters, L., Vander Heyden, Y., 2011. Classification models for neocryptolepine derivatives as inhibitors of the β-haematin formation. Anal. Chim. Acta 705, 98–110.

Ercisli, S., Sayinci, B., Kara, M., Yildiz, C., Ozturk, I., 2012. Determination of size and shape features of walnut (Juglans regia L.) cultivars using image processing. Sci. Hortic. 133, 47–55. https://doi.org/10.1016/j.scienta.2011.10.014.

Evaristo, I., Batista, D., Correia, I., Correia, P., Costa, R., 2010. Chemical profiling of Portuguese Pinus pinea L. nuts. J. Sci. Food Agric. 90, 1041–1049. https://doi.org/10.1002/jsfa.3914.

Horwitz, W., 2005. Official Methods of Analysis of AOAC International, eighteenth ed.

Huang, M., Tang, J., Yang, B., Zhu, Q., 2016. Classification of maize seeds of different years based on hyperspectral imaging and model updating. Comput. Electron. Agric. 122, 139–145. https://doi.org/10.1016/j.compag.2016.01.029.

INC, 2017. INTERNATIONAL NUT and DRIED FRUIT COUNCIL: Nuts & Amp. In: Dried Fruits Statistical Yearbook, vol. 76.

Kuo, T.Y., Chung, C.L., Chen, S.Y., Lin, H.A., Kuo, Y.F., 2016. Identifying rice grains using image analysis and sparse-representation-based classification. Comput. Electron. Agric. 127, 716–725. https://doi.org/10.1016/j.compag.2016.07.020.

Liu, D., Ning, X., Li, Z., Yang, D., Li, H., Gao, L., 2015. Discriminating and elimination of damaged soybean seeds based on image characteristics. J. Stored Prod. Res. 60, 67–74. https://doi.org/10.1016/j.jspr.2014.10.001.

Loewe, V., Navarro-Cerrillo, R.M., García-Olmo, J., Riccioli, C., Sánchez-Cuesta, R., 2017. Discriminant analysis of Mediterranean pine nuts (Pinus pinea L.) from Chilean plantations by near infrared spectroscopy (NIRS). Food Contr. 73, 634–643. https://doi.org/10.1016/j.foodcont.2016.09.012.

Lurstwut, B., Pornpanomchai, C., 2018. Image analysis based on color, shape and texture for rice seed (Oryza sativa L.) germination evaluation. Agric. Nat. Resour. 51, 383–389. https://doi.org/10.1016/j.anres.2017.12.002.

Matthäus, B., Li, P., Ma, F., Zhou, H., Jiang, J., Özcan, M.M., 2018. Is the profile of fatty acids, tocopherols, and amino acids suitable to differentiate Pinus armandii suspicious to Be responsible for the pine nut syndrome from other Pinus species? Chem. Biodivers. 15 https://doi.org/10.1002/cbdv.201700323.

Menesatti, P., Costa, C., Paglia, G., Pallottino, F., D'Andrea, S., Rimatori, V., Aguzzi, J., 2008. Shape-based methodology for multivariate discrimination among Italian hazelnut cultivars. Biosyst. Eng. 101, 417–424. https://doi.org/10.1016/j.biosystemseng.2008.09.013.

Mikkelsen, A.T., Jessen, F., Ballin, N.Z., 2014. Species determination of pine nuts in commercial samples causing pine nut syndrome. Food Contr. 40, 19–25. https://doi.org/10.1016/j.foodcont.2013.11.030.

Moscetti, R., Carletti, L., Monarca, D., Cecchini, M., Stella, E., Massantini, R., 2013a. Effect of alternative postharvest control treatments on the storability of "Golden Delicious" apples. J. Sci. Food Agric. 93, 2691–2697.

Moscetti, R., Haff, R.P., Aernouts, B., Saeys, W., Monarca, D., Cecchini, M., Massantini, R., 2013b. Feasibility of Vis/NIR spectroscopy for detection of flaws in hazelnut kernels. J. Food Eng. 118, 1–7.

Moscetti, R., Haff, R.P., Ferri, S., Raponi, F., Monarca, D., Liang, P., Massantini, R., 2017. Real-time monitoring of organic carrot (var. Romance) during hot-air drying using

near-infrared spectroscopy. Food Bioprocess Technol. 10 https://doi.org/10.1007/s11947-017-1975-3, 2046–2059.
Moscetti, R., Haff, R.P., Monarca, D., Cecchini, M., Massantini, R., 2016. Near-infrared spectroscopy for detection of hailstorm damage on olive fruit. Postharvest Biol. Technol. 120, 204–212. https://doi.org/10.1016/j.postharvbio.2016.06.011.
Moscetti, R., Saeys, W., Keresztes, J.C., Goodarzi, M., Cecchini, M., Danilo, M., Massantini, R., 2015. Hazelnut quality sorting using high dynamic range short-wave infrared hyperspectral imaging. Food Bioprocess Technol. 8, 1593–1604.
Mutke, S., Calama, R., González-Martínez, S.C., Montero, G., Gordo, F.J., Bono, D., Gil, L., 2012. Mediterranean stone pine: botany and horticulture. Hortic. Rev. 39, 153–201. https://doi.org/10.1002/9781118100592.ch4.
Nergiz, C., Dönmez, İ., 2004. Chemical composition and nutritive value of Pinus pinea L. seeds. Food Chem. 86, 365–368. https://doi.org/10.1016/j.foodchem.2003.09.009.
OpenCV, 2020. n.d. OpenCV: Image Thresholding [WWW Document]. URL. https://docs.opencv.org/master/d7/d4d/tutorial_py_thresholding.html. accessed 5.10.20.
Pallottino, F., Menesatti, P., Costa, C., Paglia, G., Salvador, F.R., Lolletti, D., 2009. Image analysis techniques for automated hazelnut peeling determination. Food Bioprocess Technol. 3, 155–159. https://doi.org/10.1007/s11947-009-0211-1.
Parks, S., 2017. n.d. Is the U.S. Pine Nut Industry on the Brink of Extinction? | Civil Eats [WWW Document]. URL. https://civileats.com/2017/06/01/is-the-u-s-pine-nut-industry-on-the-brink-of-extinction/. accessed 5.10.20.
Rajalahti, T., Arneberg, R., Berven, F.S., Myhr, K.-M., Ulvik, R.J., Kvalheim, O.M., 2009. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometr. Intell. Lab. Syst. 95, 35–48. https://doi.org/10.1016/j.chemolab.2008.08.004.
Rodríguez-Pulido, F.J., Gómez-Robledo, L., Melgosa, M., Gordillo, B., González-Miret, M. L., Heredia, F.J., 2012. Ripeness estimation of grape berries and seeds by image analysis. Comput. Electron. Agric. 82, 128–133. https://doi.org/10.1016/j.compag.2012.01.004.
Sharashkin, L., Gold, M., 2004. Pinenuts: species, products, markets, and potential for U. S. Production. In: Northern Nut Growers Association 95th Annual Report. Proceeding for the 95th Annual Meeting. Columbia, Missouri, August 16-19, 2004.
Statista, 2019. n.d. • Nuts: global production by type 2019 | Statista [WWW Document]. URL. https://www.statista.com/statistics/1030790/tree-nut-global-production-by-type/. accessed 5.10.20.
Sun, D.-W., 2020. Computer Vision Technology for Food Quality Evaluation, second ed.
Tigabu, M., Oden, P.C., Lindgren, D., 2005. Identification of Seed Sources and Parents of Pinus Sylvestris L . Using Visible – Near Infrared Reflectance Spectra and Multivariate Analysis. https://doi.org/10.1007/s00468-005-0408-5.
UNECE, 2013. UNECE Standard for Pine Nuts (DDP-12). Last access: 02 July 2020.
Valand, R., Tanna, S., Lawson, G., Bengtström, L., 2020. A review of Fourier Transform Infrared (FTIR) spectroscopy used in food adulteration and authenticity investigations. Food Addit. Contam. Part A Chem. Anal. Control. Expo. Risk Assess. https://doi.org/10.1080/19440049.2019.1675909.
Vanhanen, L., Savage, G., 2013. Mineral analysis of pine nuts (Pinus spp.) grown in New Zealand. Foods 2, 143–150. https://doi.org/10.3390/foods2020143.
Workman, J., Weyer, L., 2008. Practical Guide to Interpretive Near-Infrared Spectroscopy. CRC Press, London, UK, ISBN 978-1-57444-784-2.
Wu, Q., Xie, L., Xu, H., 2018. Determination of toxigenic fungi and aflatoxins in nuts and dried fruits using imaging and spectroscopic techniques. Food Chem. 252, 228–242. https://doi.org/10.1016/j.foodchem.2018.01.076.
Xiaobo, Z., Jiewen, Z., Povey, M.J.W., Holmes, M., Hanpin, M., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667, 14–32. https://doi.org/10.1016/j.aca.2010.03.048.
Xing, J., Guyer, D., 2008. Comparison of transmittance and reflectance to detect insect infestation in Montmorency tart cherry. Comput. Electron. Agric. 64, 194–201.
Zhang, C., Guo, C., Liu, F., Kong, W., He, Y., Lou, B., 2016. Hyperspectral imaging analysis for ripeness evaluation of strawberry with support vector machine. J. Food Eng. 179, 11–18. https://doi.org/10.1016/j.jfoodeng.2016.01.002.