UNTARGETED CLASSIFICATION FOR PAPRIKA POWDER AUTHENTICATION USING VISIBLE – NEAR INFRARED SPECTROSCOPY (VIS-NIRS)

Olga Monago-Maraña[a*], Carl Emil Eskildsen[a], Teresa Galeano-Díaz[b,c], Arsenio Muñoz de la Peña[b,c], Jens Petter Wold[a]

[a]Nofima AS – Norwegian Institute of Food, Fisheries and Aquaculture Research, PB 210, N-1431, Ås, Norway

[b]Department of Analytical Chemistry, University of Extremadura, Badajoz 06006, Spain

[c]Research Institute on Water, Climate Change and Sustainability (IACYS), University of Extremadura, Badajoz 06006, Spain

*corresponding author. E-mail: olgamonago@gmail.com

Abstract

This paper describes a non-destructive screening method for authentication of paprika belonging to the Spanish Protected Designation of Origin (PDO) *"Pimentón de La Vera"*. Different multivariate classification models were developed in order to differentiate PDO and non-PDO samples, using visible-near infrared spectra as fingerprint for each paprika sample. Sample treatment was not required. Principal component analysis (PCA) was applied in different spectral ranges: 400 - 2500, 400 - 800 and 800 - 2500 nm. In all spectral ranges, PCA was largely able to differentiate PDO from non-PDO samples. Partial least-squares - discriminant analysis (PLS-DA), PCA-linear discriminant analysis (LDA) and PCA-quadratic discriminant analysis (QDA) were used as classification methods in the different spectral ranges. All methods were able to differentiate PDO from non-PDO samples, with error rates (ER) lower than 0.15. The best models were those obtained with PLS-DA in the NIR range (800 - 2500 nm), showing ERs lower than 0.07 and error indexes (I_{ERROR}) (false positives) lower than 0.05.

Keywords: Protected Designation of Origin (PDO); paprika; authentication; Visible-Near Infrared Spectroscopy (Vis-NIRS); multivariate analysis

1. Introduction

Paprika powder is used as a spice in many countries. In Spain, there are three traded types of paprika, which differ in their drying process (air, sun and smoke drying). Air-dried paprika, using heated air, is produced mainly in the south-east and central-east of Spain (Murcia), where the high temperature conditions allow peppers to undergo rapid dehydration. Sun-dried paprika are imported from South America and South Africa. Smoked paprika originates from La Vera region, Extremadura in the south-west of Spain. Here, a traditional drying process is used, where oak logs are burnt to heat the paprika to 40 ºC and give it a smoked flavor (Martín et al., 2017).

Smoked paprika is recognized under the quality seal Protected Designation of Origin (PDO) *"Pimentón de La Vera"* by the European Union since 2006 (Unión Europea, 2006). This product is considered a high-quality product obtained by drying the fruit of autochthonous varieties of peppers *(Capsicum annum L.)*. Moreover, the traditional drying process confers the paprika its aroma, flavor, and color (Martín et al., 2017). Adulteration of smoked paprika *"Pimentón de La Vera"* with foreign paprika of lower quality, primarily to increase profit margins, has been a concern for many years to the smoked paprika industry (Hernández, Martín, Aranda, Bartolomé, & Córdoba, 2007). Therefore, inexpensive and high throughput screening tools to differentiate paprika based on origin is interesting for the industry.

Recent reviews show how spectroscopic techniques, including near-infrared spectroscopy (NIRS), can be used for detection of adulteration in herbs and spices (Kucharska-Ambrożej & Karpinska, 2020; Marciano M. Oliveira, Cruz-Tirado, & Barbin, 2019). However, not many studies about paprika powder adulteration were found. In the case of paprika or related products, NIRS has been mainly used for quantification. For example, to quantify ASTA color, moisture (Bae, Han, & Hong, 1998), capsaicinoids (Lim, Kim, Mo, & Kim, 2015; Park et al., 2008), arsenic and lead (Moros et al., 2008), soluble solids content (SSC), firmness of peppers (Penchaiya, Bobelyn, Verlinden, Nicolaï, & Saeys, 2009) and mycotoxins (Hernández-Hierro, García-Villanova, & González-Martín, 2008). In addition, Vis-NIRS combined with multivariate analysis has been used to determine total carotenoids, chlorophylls, as well as maturity stage of intact peppers (Timea Ignat et al., 2013) and ascorbic acid (T. Ignat, Schmilovitch, Fefoldi, Steiner, & Alkalai-Tuvia, 2012). Few works about the adulteration and/or authentication of paprika powder using NIRS as analytical technique have been found in the literature. A recent work about this topic was based on the detection of adulterants such us potato starch, annatto and acacia gum in paprika powder samples from Spain (n = 3) and Brazil (n = 2) (M. M. Oliveira, Cruz-Tirado, Roque, Teófilo, & Barbin, 2020). Detection and quantification of adulterants was done using a portable NIR instrument in combination with partial least squares (PLS) regression and PLS-Discriminant Analysis (PLS-DA). The results were promising with a specificity greater than 90% and error rate lower than 2 % for the PLS-DA models.

In another study, paprika samples were clustered based on origin using NIRS and Principal Component Analysis (PCA) (Molnár et al., 2018). However, only six paprika samples from Spain were included in the analysis, and PDO specifications were not taken into account.

Only few studies have investigated the possibility of differencing between paprika samples belonging to the PDO *"Pimentón de La Vera"* and samples not belonging to the PDO. Discrimination has been based on color measurements with visible spectrophotometry, being samples, belonging to the PDO *"Pimentón de La Vera*" or not, correctly grouped in two groups with PCA (Monago Maraña, Bartolomé García, & Galeano Díaz, 2016). Then, samples were classified as different PDOs *("Pimentón de La Vera"* or *"Pimentón de Murcia")* with classification efficiencies ranging from 92 to 95 % when visible spectra and multilayer perceptrons artificial neural networks (MLP-ANN) were used (A. Palacios-Morillo, Jurado, Alcázar, & Pablos, 2016).

Regarding to destructive methods, liquid chromatography has been widely used for the paprika authentication. Classification and authentication have been done with different Spanish PDOs, *"Pimentón de La Vera"*, *"Pimentón de Murcia"*, and Czech Republic paprika samples without PDO. Employing ultra-high-performance liquid chromatography coupled with high-resolution mass spectrometry (UHPLC-HRMS), samples were discriminated on a non-target way (Barbosa, Saurina, Puignou, & Núñez, 2020) and based on the polyphenolic and capsaicinoid profiling

(Barbosa, Saurina, & Oscar, 2020) with classification results of 100%. On the other hand, HPLC-UV was used to obtain the phenolic profile of paprika for their authentication, confirming that was enough to discriminate between PDOs (Cetó, Sánchez, Serrano, Díaz-Cruz, & Núñez, 2020).

Also, the presence or absence of sub-products from the smoking process (Polycyclic Aromatic Hydrocarbons, PAHs) (Monago-Maraña, Galeano-Díaz, & Muñoz de la Peña, 2017), hydrophobic proteins (Hernández et al., 2007) or metallic content (Ana Palacios-Morillo, Jurado, Alcázar, & De Pablos, 2014) have allowed differentiation of paprika at different conditions. Although being very selective, discriminating on these compounds requires sample extraction steps, which normally is time consuming. For this reason, high throughput screening methods are interesting for practical use in the paprika industries.

In this study, Vis-NIR measurements will be used, which are cost effective, high throughput and non-destructive, to discriminate paprika powder samples belonging to the PDO *"Pimentón de La Vera"* from paprika powder samples not belonging to the PDO. To achieve this goal, we use multivariate qualitative analytical methods for authenticating the PDO *"Pimentón de La Vera"* paprika powder samples. Different methods for classification of multivariate data were compared and ranked.

2. Material and methods

2.1. Samples

A total of 49 paprika powder samples under the PDO *"Pimentón de La Vera"* were included in the study. These samples were from five different producers and were made over a period of ten years (2010 – 2020). Samples from 2010 to 2017 were obtained in 2017 (n = 35) from producers and measured in that year. Samples from 2017 – 2020 (n = 14) were acquired in Spanish markets in 2020 and measured that year. The samples were made under smoked conditions, following the traditional process from La Vera, in Extremadura, Spain. Among these samples, there were sweet, sweet/hot and hot paprika samples.

A total of 50 samples not belonging to any PDO were acquired from different markets in Spain and Norway. Samples acquired in Norway (n = 9) were bought and measured in 2017, but samples acquired in Spanish markets (n = 23) were acquired in 2017 and 2020 (n = 18), and measured the corresponding year of acquisition. The production processes of these samples are unknown as well as the peppers used for their production due to the fact that it is not mandatory to include that information in labels of paprika samples. Among these samples, there were sweet and hot paprika samples.

2.2. Spectroscopic acquisition

The VIS-NIRS measurements were obtained in reflectance mode using a FOSS NIRS Systems XDS Rapid Content^{TM} Analyzer (FOSS Analytical A/S, Hillerød, Denmark). In order to obey *Beer's law*, the NIR spectra were transformed from reflectance (R) units into absorbance-like units (log(1/R)). An internal ceramic standard was used as reference. Spectra were obtained from 400 to 2500 nm, with a resolution of 0.5 nm. Paprika powder samples were measured in circular sample cups of approximately 79 cm^2 (FOSS Analytical A/S, Hillerød, Denmark). Spectra from each sample were acquired in triplicate, mixing the powder for obtaining different surfaces each time to obtain a representative sample spectrum. The average spectrum was used for further analysis.

2.3. Data processing and multivariate analysis

2.3.1. Principal component analysis

Principal component analysis (PCA) was applied to explore the main variation over samples. During PCA all samples were included. Prior to PCA the spectral measurements were preprocessed by extended multiplicative signal corrected (EMSC) (Martens & Stark, 1991) and mean centered variable-wise.

The objective of PCA is to compress the data, reducing it from the high dimensional variable space into a lower dimensional principal component space. Each new principal component (PC) is a linear combination of the original variables. The loadings describe the direction of each principal component in the original X-space and the scores are the projections of the original data onto the loading vectors (Wold, Esbensen, & Geladi, 1987).

PCAs was performed separately for the entire spectral range, the visible range (from 400 to 800 nm) and the NIR (800 - 2500 nm) range.

2.3.2. Classification analysis

For the classificatory analysis, samples were divided in two sets (training and test). Approximately 60 % of the samples were used for training and the remaining 40 % of the samples were used for validation. Hence, the training set was composed by 59 samples (29 PDO and 30 non-PDO) and the test set was formed by 40 samples (20 PDO and 20 non-PDO). The split of samples was based on the recently published EuroLab Guide (TR No 01/2015, 2015), which recommends a minimum of 20 samples for each class in the test sets. The training and test samples were randomly chosen. Hence, this division was performed three times, and three different training and test sets were obtained and used for building different calibration models. As a result, the average results of three training and test sets were given with the corresponding standard deviation.

The following classification algorithms were tested for discrimination of the sample spectra: discriminant partial least-squares (PLS-DA) (Barker & Rayens, 2003), linear discriminant analysis based on the PC scores of the spectra (PCA-LDA) (Mohanty, John, Manmatha, & Rath, 2013), and quadratic discriminant analysis based on the PC scores of the spectra (PCA-QDA) (Tharwat, 2016).

PLS-DA involves performing a multivariate regression model to establish class limits and placing a numeric value to each object/sample first, and then classifying them into a specific class. As in PLS regression, the relation between instrumental response in X (spectra) and y (class coding) is established, and the optimal number of latent variables is chosen based on the error range by cross-validation.

To apply LDA or QDA, it is necessary to reduce the dimensionality of the spectral data. For that PCA is used. After PCA, LDA is used when the decision line between the two groups can be represented by a linear function. However, if a curved line is needed to separate the groups, then QDA is more effective.

Prior to classification the spectral training data were preprocessed by EMSC and variable-wise mean centered. Classification models were fitted on the training set using full-cross validation to determine the optimal models. Then the models were tested with the external test set (pre-processed with the EMSC model obtained for training previously). Data analysis was done using a graphical interface (Ballabio & Consonni, 2013) in Matlab (R2016b, The MathWorks, Inc., Natick, MA, USA).

2.3.3. Evaluation of the methodology

In order to evaluate the screening methodology, the confusion matrices were obtained and the performance parameters such as precision (PREC), sensitivity (SENS), error rate (ER), accuracy (ACCU) and specificity (SPEC) were calculated.

The PREC is defined as the number of samples correctly assigned as belonging to the PDO (i.e. true positives (TP)) over the total number of samples assigned as belonging to the PDO (i.e. the total number of true positives and false positives (FP)) (Eq. 1). The SENS is the number of true positives over the total number of samples belonging to the PDO (i.e. the total number of true positives and false negatives (FN)) (Eq. 2). The ER is the number of samples incorrectly classified by the model (i.e. the total number of false positives and false negatives) over the total number of samples (Eq. 3). The ACCU is the number of samples correctly classified by the model (i.e. the total number of true positives and true negatives (TN)) over the total number of samples (Eq. 4). The SPEC is the number of samples correctly assigned as not belonging to the PDO (i.e. true negatives) over the total number of samples not belonging to the PDO (Eq. 5).

$$PREC = \frac{TP}{TP+FP} \qquad (1)$$

$$SENS = \frac{TP}{TP+FN} \qquad (2)$$

$$ER = \frac{FN + FP}{TP+TN+FP+FN} \quad (3)$$

$$ACCU = \frac{TP + TN}{TP+TN+FP+FN} \quad (4)$$

$$SPEC = \frac{TN}{TN + FP} \quad (5)$$

Where TP and TN are the number true positive and number of true negative, respectively, and FN and FP are the number of false negative and number of false positive, respectively.

Furthermore, two recently proposed indexes, error index (I_{ERROR}) and loss index (I_{LOSS}), for assigning a specification-based quality grade for a PDO label are calculated (Cuadros-Rodríguez, Valverde-Som, Jiménez-Carvelo, & Delgado-Aguilar, 2020).

I_{ERROR} is the probability of a sample being incorrectly assigned to the PDO class (Eq. 6). I_{LOSS} is the probability of obtaining false negatives and thus the risk of economic loss due to assignment error.

$$I_{ERROR} = \frac{FP}{TP+TN+FP+FN} \quad (6)$$

$$I_{LOSS} = \frac{FN}{TP+TN+FP+FN} \quad (7)$$

3. Results and discussion

3.1. VIS-NIRS spectral profiling

Figure 1A shows the mean of the absorption spectra for both classes (PDO and non-PDO). The mean spectrum of non-PDO shows higher intensity over the whole spectral range as compared with the mean spectrum for PDO. More subtle differences can be seen after pre-processing by EMSC (Figure 1B). The main difference in the visible range was observed at 670 nm, and in the NIR range at 1450, 1940, 2305, 2346 and 2490 nm. The visible range was previously reported to be useful for the quantification of total carotenoids and chlorophylls in intact bell pepper (Timea Ignat et al., 2013). In the case of NIR bands, some of them might be due to water peaks (1450 and 1940 nm) and the other three main peaks (2305, 2350 and 2490 nm) do most likely originate from fat (Núñez-Sánchez et al., 2016).

3.2. Exploratory analysis

In order to study the most important spectral variation for discriminating PDO and non-PDO samples, detect potential outliers and systematic artifacts in the samples, PCA was performed on the EMSC pre-processed spectra. All 99 samples were included in the analysis. As described above, PCAs were performed on different spectral ranges.

When including the whole spectral range, the first three principal components (PCs) explain 84 % of the total variation in the data set. The first principal component (PC1) explains 50 % of the variation, and the corresponding loading plot (not shown) reveals the most important peaks at approximately 480 and 600 nm in the visible range and at 1450 and 1940 nm in the NIR range (water peaks). However, this component does not differentiate PDO from non-PDO paprika samples.

The best discrimination is observed for scores of PC3 and PC5, explaining 12 and 4 % of the total variation, respectively (Figure 2A). Clearly, two groups are established according to PDO and non-PDO samples. However, the two groups are slightly overlapping. PC3 provides the clearest discrimination of the groups. The clear unsupervised clustering is a good basis for supervised classification.

The loadings for PC3 and PC5 are presented in Figure 2B. The main variables affecting the separation of the groups were 540 and 670 nm in the visible range and water peaks in the NIR range (Figure 2A). Score values for PC3 are generally high for the PDO samples, which means that positive loadings, representing certain chemical components, are positively related to PDO samples. The negative loadings observed at 1720 and 1760 nm are related with first overtone C-H stretching vibration of methyl (-CH_3), methylene (-CH_2) and ethenyl (-CH=CH-) groups. The loadings close to 1725 nm has been related to oleic acid and the band close to 1760 nm to saturated components. The bands at 2305 and 2350 nm have previously been assigned to combination of C-H stretches and deformations (Núñez-Sánchez et al., 2016; Pérez-Juan et al., 2010). Also, the small band at 1207 nm is related with fat. All bands related to fat are negative loadings, suggesting a relatively low concentration of fat in PDO samples.

Scores for PCA in the visible spectral range are presented in Figure 2C. PC4, explaining 6 % of the variance, discriminates quite well between the two groups. Note that the overlap of the groups is stronger when using only the visible range, compared to using the whole range. The main variables affecting the clustering are those mentioned before (570 and 670 nm) as seen in the loading for PC4 (Figure 2D).

Finally, for the NIR range, a quite good grouping of the samples is obtained in PC2 (Figure 2E) due to variables corresponding to water and fat peaks. Interestingly, some peaks are more pronounced in the loadings in this case. These peaks can be attributed to proteins bands: 2056 nm (N-H stretching vibrations) and 2478 nm (-C-N-C stretching first overtone).

3.2. Classificatory analysis

As detailed in the section 2.4.2, samples were divided into training and test sets. This step was performed three times and the classification model was obtained for each case. Average results for confusion matrices from different sets and the corresponding validation parameters are shown in Tables 1 and 2, respectively. The numbers in parentheses correspond with the standard deviations from the three sets assayed.

For PLS-DA, the best classification results were obtained for the NIR range in both training and test samples. The ERs obtained for this range were overall lower than for other ranges. Interestingly, from a quality-point of view, the I_{ERROR} was lower for the NIR spectral range as compared with the other spectral ranges, for both the training set and test set. This is important for avoiding non-PDO samples being classified as PDO samples. The visible range gave slightly less correct classifications than the whole range, but all models provided acceptable results, with ERs lower than 0.11 and I_{ERROR} lower than 0.10. According to (Cuadros-Rodríguez et al., 2020), a good screening method should offer an I_{ERROR} equal to or lower than 0.1 in order to minimize the false-compliance error. Hence, the best choice with PLS-DA would be with the NIR range, although in some cases that means that some samples would be false-negative and refused categorized as PDO (PDO samples categorized as non-PDO samples).

Regarding the other performance parameters, SENS and SPEC present similar values (Table 2), mainly in the NIR range. This means that the error is balanced, and there is not a clear trend in the models for false positives, or vice versa. PREC values were higher for the NIR range, which means that false positives were lower in these models, as observed in the I_{ERROR} values as well.

The regression coefficients for each spectral range (Figure 3) were evaluated in order to elucidate the main variables contributing to the classification. For the visible range, the main variables were 570 and 670 nm with negative values, and 540 nm with positive value. It might be expected that the variation in the visible range would be related to total carotenoids, ASTA values (extractable color), as other authors reported (A. Palacios-Morillo et al., 2016). In these samples, the ASTA value was not so relevant since some PDO samples were old and therefore had low ASTA values (between 25 - 70). Therefore, it was expected that some samples were incorrectly classified when using the visible range. However, acceptable results for classification were obtained due to other variables, not related to total carotenoids. The VIP scores (not shown) were also investigated. Similar information was retrieved from the VIP scores and the regression vectors (Figure 3).

The absorption around 670 nm has previously been related with chlorophylls (Timea Ignat et al., 2013) and could be also related with pheophytins formed from chlorophylls during ripening or drying process (Bonaccorsi et al., 2016) . This peak has negative regression coefficients (Figure 3A and 3B), which suggests that non-PDO samples have lower content of chlorophyll compared to PDO samples. This is also observed in Figure 1B.

Regarding the NIR range, the regression coefficient positive wavelength bands associated with fat, such as, 1725, 2305, 2350 and 2490 nm, again suggesting a relatively high fat content in non-PDO samples. A higher fat content can have different reasons. Different types of peppers used for paprika production vary in the fatty acid composition depending on genotype and environmental factors. Kim et al., 2019 recently reported this for some varieties of peppers and this could be extended to other kind of peppers (Kim et al., 2019). Another reason may be related with the addition of sunflower vegetal oil to give stronger brightness of the powder. In the case of PDO *"Pimentón de La Vera"* the amount of oil is limited to 3 % (w/w) (Unión Europea, 2006).

However, there are not specifications reported about other kind of paprika samples, which are not under the PDO. This could mean that other paprika samples contain a higher percentage of sunflower oil to give more brightness. A third reason could be related to the addition of seeds from peppers used in the paprika production, which would influence in the fatty acid composition. This kind of addition is not allowed in PDO samples (Unión Europea, 2006).

PCA-LDA and PCA-QDA gave results in accordance with PLS-DA; better results were obtained when the NIR range or whole range were used to classify samples, giving ERs lower than 0.15 and I_{ERROR} lower than 0.11. Another important result was that PCA-QDA offered better results than PCA-LDA in all cases. In the case of PCA-LDA and PCA-QDA, PREC, SENS and SPEC values were slightly better for the NIR range. As in previous case, SENS and SPEC values were similar, which proved that errors did not follow a clear trend.

Finally, it must be highlighted that these good results were obtained for three training/test sets, which proved the robustness of the methods. To our knowledge, this is the first work where non-destructive classification of PDO *“Pimentón de La Vera”* has been performed. The method is easy and quick to use and could with some more development contribute to effective control in the paprika industries.

4. Conclusions

Vis-NIR spectroscopy with different multivariate classification techniques have been proven to discriminate between paprika samples belonging to the PDO *“Pimentón de La Vera”* and other paprika samples. The variability of samples and the random choice of samples for training and test, indicate that the models are quite robust. The visible range offered the good classification due to chlorophylls or pheophytin compounds and NIR range showed slightly better classification based on differences in absorbance of fat. PLS-DA offered somewhat better results than other classification methods. It can be highlighted that all methods offered acceptable ERs and I_{ERROR}, always lower than 0.15 and 0.11, respectively. This method is easy, rapid and non-destructive, being an advantage in order to implement the method for industrial purposes.

Acknowledgements

Olga Monago Maraña thanks to the Fundación Ramón Areces for a postdoctoral fellowship for studies abroad in the field of Life and Matter Sciences (XXXI edition of grants, 2019/2020) to support her postdoctoral studies at Nofima, Ås, Norway.

Financial support was provided by the Junta de Extremadura (Ayuda GR18041-Research Group-FQM003 and Project IB16058) and Ministerio de Ciencia, Innovación y Universidades of Spain (Project CTQ2017-82496-P), both co-financed by the Fondo Social Europeo funds. Funding was also given by Norwegian Agricultural Food Research Foundation through the project FoodSMaCK – Spectroscopy, Modelling & Consumer Knowledge, No. 262308 /F40.

References

Bae, M.-J., Han, E.-S., & Hong, S.-H. (1998). Use of near infrared spectroscopy in quality control of red pepper powder. *Journal of Near Infrared Spectroscopy*, *6*, A333–A337.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*, 3790–3978.

Barbosa, S., Saurina, J., & Oscar, N. (2020). Capsaicinoid profiling for the chemometric characterization and classification of Paprika with Protected Designation of Origin (PDO) attributes. *Molecules, 25,* 1–16.

Barbosa, S., Saurina, J., Puignou, L., & Núñez, O. (2020). Classification and authentication of paprika by UHPLC-HRMS fingerprinting and multivariate calibration methods (PCA and PLS-DA). *Foods, 9,* 1–10.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*, 166–173.

Bonaccorsi, I., Cacciola, F., Utczas, M., Inferrera, V., Giuffrida, D., Donato, P., … Mondello, L. (2016). Characterization of the pigment fraction in sweet bell peppers (Capsicum annuum L.) harvested at green and overripe yellow and red stages by offline multidimensional convergence chromatography/liquid chromatography–mass spectrometry. *Journal of Separation Science*, *39*(17), 3281–3291. https://doi.org/10.1002/jssc.201600220

Cetó, X., Sánchez, C., Serrano, N., Díaz-Cruz, J. M., & Núñez, O. (2020). Authentication of paprika using HPLC-UV fingerprints. *LWT - Food Science and Technology, 124,* 109153.

Cuadros-Rodríguez, L., Valverde-Som, L., Jiménez-Carvelo, A. M., & Delgado-Aguilar, M. (2020). Validation requirements of screening analytical methods based on scenario-specified applicability indicators. *TrAC - Trends in Analytical Chemistry*, *122*.

Hernández-Hierro, J. M., García-Villanova, R. J., & González-Martín, I. (2008). Potential of near infrared spectroscopy for the analysis of mycotoxins applied to naturally contaminated red paprika. *Analytica Chimica Acta*, *2*, 189–194.

Hernández, A., Martín, A., Aranda, E., Bartolomé, T., & Córdoba, M. de G. (2007). Application of temperature-induced phase partition of proteins for the detection of smoked paprika adulteration by free zone capillary electrophoresis (FZCE). *Food*

Chemistry, *105*, 1219–1227.

Ignat, T., Schmilovitch, Z., Fefoldi, J., Steiner, B., & Alkalai-Tuvia, S. (2012). Non-destructive measurement of ascorbic acid content in bell peppers by VIS-NIR and SWIR spectrometry. *Postharvest Biology and Technology*, *74*, 91–99.

Ignat, Timea, Schmilovitch, Z., Feföldi, J., Bernstein, N., Steiner, B., Egozi, H., & Hoffman, A. (2013). Nonlinear methods for estimation of maturity stage, total chlorophyll, and carotenoid content in intact bell peppers. *Biosystems Engineering*, *114*, 414–425.

Kim, E. H., Lee, S. Y., Baek, D. Y., Park, S. Y., Lee, S. G., Ryu, T. H., … Oh, S. W. (2019). A comparison of the nutrient composition and statistical profile in red pepper fruits (Capsicums annuum L.) based on genetic and environmental factors. *Applied Biological Chemistry*, (1), 62–48.

Kucharska-Ambrożej, K., & Karpinska, J. (2020). The application of spectroscopic techniques in combination with chemometrics for detection adulteration of some herbs and spices. *Microchemical Journal*, *153*, 104278.

Lim, J., Kim, G., Mo, C., & Kim, M. (2015). Design and fabrication of a real-time measurement system for the capsaicinoid content of Korean red pepper (Capsicum annuum L.) powder by visible and Near-Infrared Spectroscopy. *Journal of Biosystems Engineering*, *15*, 47–60.

Martens, H., & Stark, E. (1991). Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, *9*(8), 625–635.

Martín, A., Hernández, A., Aranda, E., Casquete, R., Velázquez, R., Bartolomé, T., & Córdoba, M. G. (2017). Impact of volatile composition on the sensorial attributes of dried paprikas. *Food Research International*, *100*, 691–697.

Mohanty, N., John, A. L. S., Manmatha, R., & Rath, T. M. (2013). Shape-based image classification and retrieval. *Handbook of Statistics*, *31*, 249–267. https://doi.org/10.1016/B978-0-444-53859-8.00010-2

Molnár, H., Kónya, É., Zalán, Z., Bata-Vidács, I., Tömösközi-Farkas, R., Székács, A., & Adányi, N. (2018). Chemical characteristics of spice paprika of different origins. *Food Control*, *83*, 54–60. https://doi.org/10.1016/j.foodcont.2017.04.028

Monago-Maraña, O., Galeano-Díaz, T., & Muñoz de la Peña, A. (2017). Chemometric Discrimination Between Smoked and Non-Smoked Paprika Samples. Quantification of PAHs in Smoked Paprika by Fluorescence-U-PLS/RBL. *Food Analytical*

Methods, *10*, 1128–1137.

Monago Maraña, O., Bartolomé García, T. de J., & Galeano Díaz, T. (2016). Characterization of Spanish Paprika by Multivariate Analysis of Absorption and Fluorescence Spectra. *Analytical Letters*, *49*, 1184–1197.

Moros, J., Llorca, I., Cervera, M. L., Pastor, A., Garrigues, S., & de la Guardia, M. (2008). Chemometric determination of arsenic and lead in untreated powdered red paprika by diffuse reflectance near-infrared spectroscopy. *Analytica Chimica Acta*, *613*, 196–206.

Núñez-Sánchez, N., Martínez-Marín, A. L., Polvillo, O., Fernández-Cabanás, V. M., Carrizosa, J., Urrutia, B., & Serradilla, J. M. (2016). Near Infrared Spectroscopy (NIRS) for the determination of the milk fat fatty acid profile of goats. *Food Chemistry*, *190*, 244–252.

Oliveira, M. M., Cruz-Tirado, J. P., Roque, J. V., Teófilo, R. F., & Barbin, D. F. (2020). Portable near-infrared spectroscopy for rapid authentication of adulterated paprika powder. *Journal of Food Composition and Analysis*, *87*, 103403.

Oliveira, Marciano M., Cruz-Tirado, J. P., & Barbin, D. F. (2019). Nontargeted analytical methods as a powerful tool for the authentication of spices and herbs: a review. *Comprehensive Reviews in Food Science and Food Safety*, *18*, 670–689.

Palacios-Morillo, A., Jurado, J. M., Alcázar, A., & Pablos, F. (2016). Differentiation of Spanish paprika from Protected Designation of Origin based on color measurements and pattern recognition. *Food Control*, *62*, 243–249.

Palacios-Morillo, Ana, Jurado, J. M., Alcázar, Á., & De Pablos, F. (2014). Geographical characterization of Spanish PDO paprika by multivariate analysis of multielemental content. *Talanta*, *128*, 15–22.

Park, T. S., Candidate, P. D., Bae, Y. M., Researcher, S., Sim, M. J., & Student, G. (2008). Analysis of Capsaicinoids from Hot Red Pepper Powder by Near-Infrared Spectroscopy. *ASABE Annual International Meeting*, (January 2008), 1–7. https://doi.org/10.13031/2013.25077

Penchaiya, P., Bobelyn, E., Verlinden, B. E., Nicolaï, B. M., & Saeys, W. (2009). Non-destructive measurement of firmness and soluble solids content in bell pepper using NIR spectroscopy. *Journal of Food Engineering*, *94*, 267–273.

Pérez-Juan, M., Afseth, N. K., González, J., Díaz, I., Gispert, M., Furnols, M. F. i., … Realini, C. E. (2010). Prediction of fatty acid composition using a NIRS fibre optics probe at two different locations of ham subcutaneous fat. *Food Research*

International, *43*(5), 1416–1422.

Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, *3*(2), 145. https://doi.org/10.1504/ijapr.2016.079050

TR No 01/2015. (2015). *Guide to NMR Method Development and Validation – Part II : Multivariate data analysis*.

Unión Europea, U. (2006). *Reglamento (CE) Nº 510/2006 del Consejo. S. Diario Oficial de la Unión Europea* (Vol. C 287/2).

Wold, S., Esbensen, K. I. M., & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*, 37–52.

Figure captions

Figure 1. (A) Average of absorption spectra (B) Average of EMSC pre-processed spectra. Black lines correspond to the PDO samples and red lines correspond to the non-PDO samples.

Figure 2. Loadings (B, D, F) and scores values (A, C, E) obtained from PCA of the spectra in wavelength ranges: 400 - 2500 nm, 400 - 800 nm and 800 - 2500 nm.

Figure 3. Regression coefficients for non-PDO samples obtained for the PLS-DA models for the different spectral ranges studied.

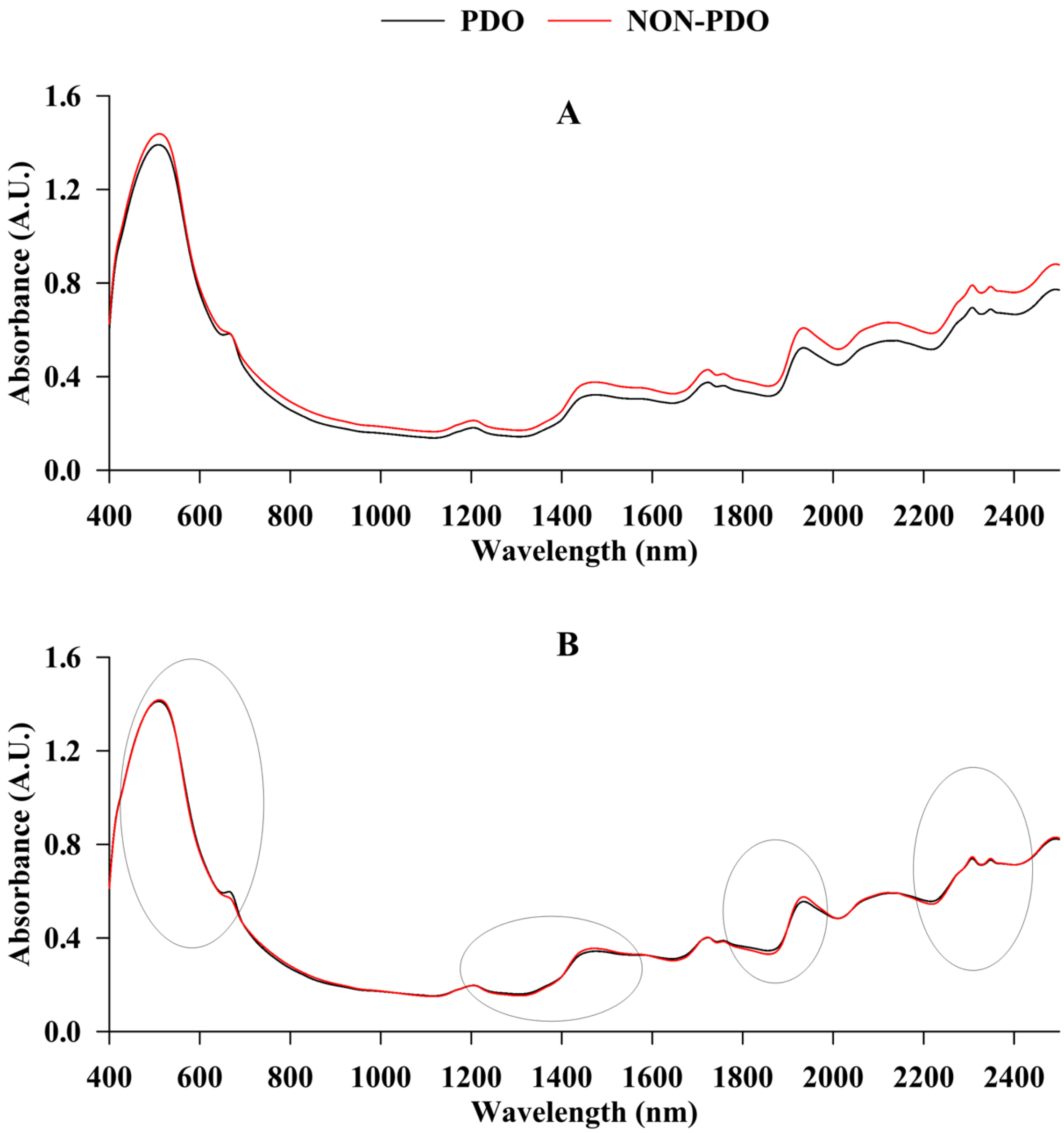
PDO
NON-PDO
A
B
Absorbance (A.U.)
0.0
0.4
0.8
1.2
1.6
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
Wavelength (nm)

Figure 1

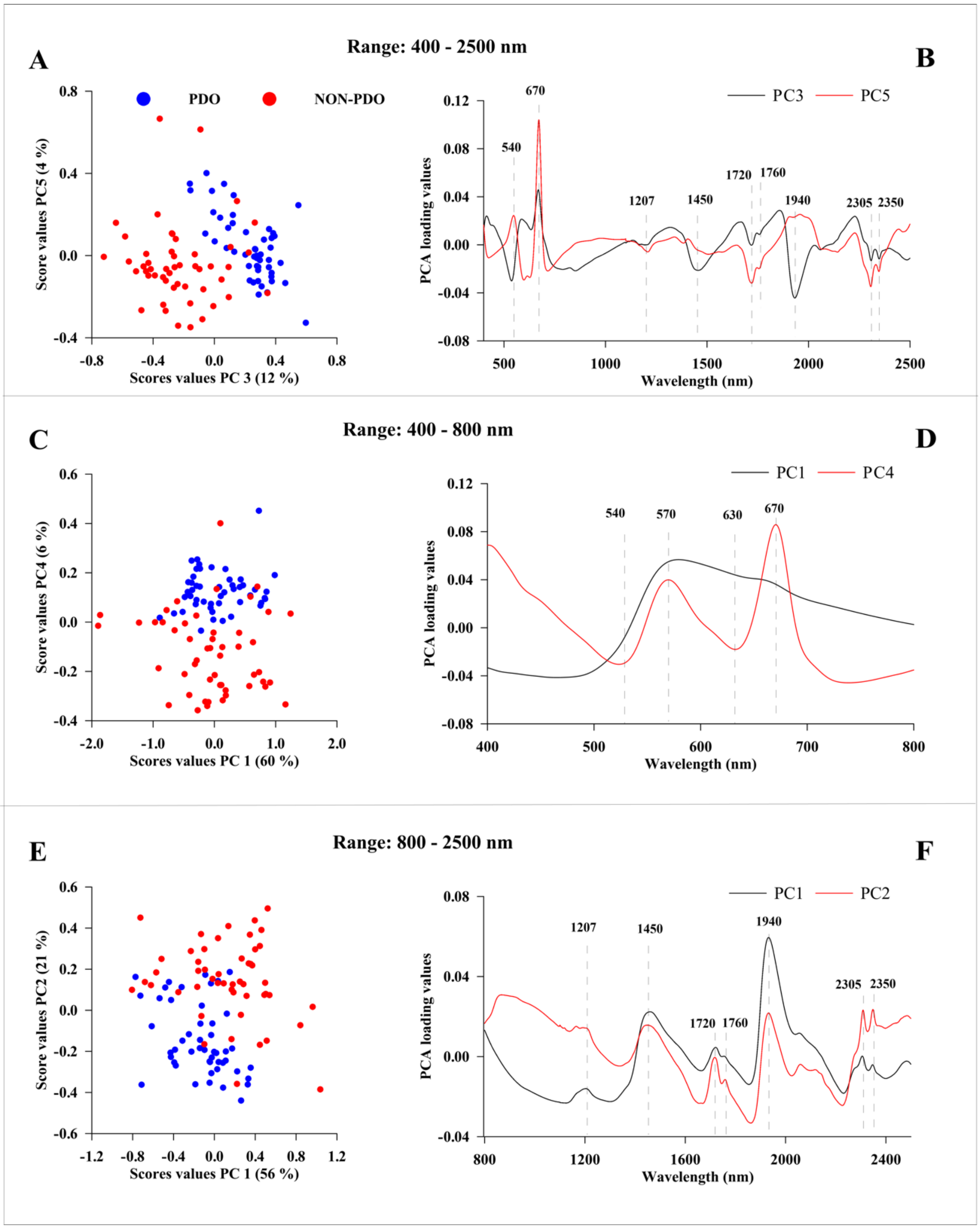
A
Range: 400 - 2500 nm
B
PDO
NON-PDO
Score values PC5 (4 %)
Scores values PC 3 (12 %)
0.8
0.4
0.0
-0.4
-0.8
-0.4
0.0
0.4
0.8
PC3
PC5
PCA loading values
Wavelength (nm)
0.12
0.08
0.04
0.00
-0.04
-0.08
500
1000
1500
2000
2500
540
670
1207
1450
1720
1760
1940
2305
2350
C
Range: 400 - 800 nm
D
Score values PC4 (6 %)
Scores values PC 1 (60 %)
0.6
0.4
0.2
0.0
-0.2
-0.4
-2.0
-1.0
0.0
1.0
2.0
PC1
PC4
PCA loading values
Wavelength (nm)
0.12
0.08
0.04
0.00
-0.04
-0.08
400
500
600
700
800
540
570
630
670
E
Range: 800 - 2500 nm
F
Score values PC2 (21 %)
Scores values PC 1 (56 %)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-1.2
-0.8
-0.4
0.0
0.4
0.8
1.2
PC1
PC2
PCA loading values
Wavelength (nm)
0.08
0.04
0.00
-0.04
800
1200
1600
2000
2400
1207
1450
1720
1760
1940
2305
2350

Figure 2

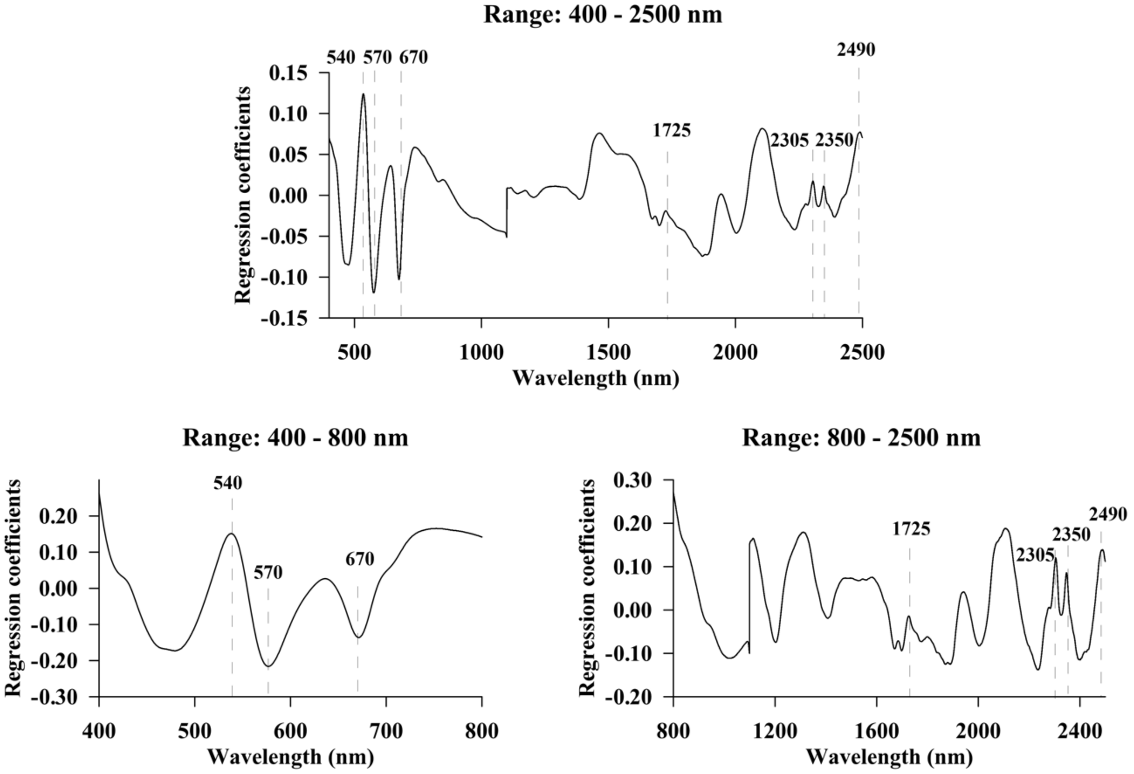
Range: 400 - 2500 nm
540
570
670
1725
2305
2350
2490
Regression coefficients
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
500
1000
1500
2000
2500
Wavelength (nm)
Range: 400 - 800 nm
540
570
670
Regression coefficients
0.20
0.10
0.00
-0.10
-0.20
-0.30
400
500
600
700
800
Wavelength (nm)
Range: 800 - 2500 nm
1725
2305
2350
2490
Regression coefficients
0.30
0.20
0.10
0.00
-0.10
-0.20
800
1200
1600
2000
2400
Wavelength (nm)

Figure 3

Table 1. Confusion matrices for the different algorithms and ranges studied in the training and test sets.

					Training set		Test set	
Algorithm	Range (nm)	Nº comp	%EV (X)		PDO (CV)	NON- PDO (CV)	PDO (val)	NON-PDO (val)
PLS-DA	400 - 2500	6	96 (1)	PDO	28 (1)	1 (1)	19 (1)	1 (1)
				NON-PDO	5 (3)	25 (3)	2 (2)	18 (2)
	400 - 800	5	99 (0)	PDO	28 (1)	1 (1)	19 (1)	1 (1)
				NON-PDO	6 (2)	24 (2)	3 (1)	17 (1)
	800 - 2500	6	98 (1)	PDO	28 (0)	1 (0)	19 (1)	1 (1)
				NON-PDO	3 (2)	27 (2)	1 (2)	19 (2)
PCA-LDA	400 - 2500	5	96 (0)	PDO	27 (1)	2 (1)	19 (1)	1 (1)
				NON-PDO	7 (2)	23 (2)	2 (2)	18 (2)
	400 - 800	5	99 (0)	PDO	28 (1)	1 (1)	19 (1)	1 (1)
				NON-PDO	6 (2)	24 (2)	3 (1)	17 (1)
	800 - 2500	5	98 (1)	PDO	25 (2)	4 (2)	17 (2)	3 (2)
				NON-PDO	5 (2)	25 (2)	2 (2)	18 (2)
PCA-QDA	400 - 2500	5	97 (1)	PDO	27 (1)	2 (1)	17 (2)	2 (2)
				NON-PDO	3 (1)	27 (1)	2 (2)	18 (2)
	400 - 800	5	99 (0)	PDO	26 (2)	3 (2)	18 (1)	2 (1)
				NON-PDO	3 (1)	27 (1)	2 (2)	18 (2)
	800 - 2500	5	96 (1)	PDO	26 (1)	3 (1)	16 (3)	4 (3)
				NON-PDO	2 (1)	28 (1)	2 (2)	18 (2)

*CV: cross-validation; numbers in parentheses correspond to the standard deviation of three sets assayed.

Table 2. Validation parameters calculated for the target class (PDO class) in the different classification methods.

		Training set							Test set						
Algorithm	**Range (nm)**	**SPEC**	**SENS**	**PREC**	**ER**	**ACCUR**	**I_{ERROR}**	**I_{LOSS}**	**SPEC**	**SENS**	**PREC**	**ER**	**ACCUR**	**I_{ERROR}**	**I_{LOSS}**
PLS-DA	**400 - 2500**	0.85 (0.11)	0.98 (0.02)	0.86 (0.09)	0.09 (0.06)	0.91 (0.06)	0.08 (0.05)	0.01 (0.05)	0.92 (0.10)	0.97 (0.03)	0.93 (0.09)	0.06 (0.06)	0.94 (0.05)	0.04 (0.05)	0.02 (0.01)
	400 - 800	0.81 (0.08)	0.97 (0.04)	0.83 (0.07)	0.11 (0.05)	0.89 (0.06)	0.10 (0.04)	0.02 (0.02)	0.87 (0.06)	0.97 (0.06)	0.88 (0.05)	0.08 (0.06)	0.92 (0.06)	0.07 (0.03)	0.02 (0.03)
	800 - 2500	0.90 (0.07)	0.97 (0.0)	0.91 (0.06)	0.07 (0.04)	0.93 (0.04)	0.05 (0.03)	0.02 (0.00)	0.93 (0.08)	0.97 (0.03)	0.94 (0.07)	0.05 (0.04)	0.95 (0.04)	0.03 (0.04)	0.02 (0.01)
PCA-LDA	**400 - 2500**	0.78 (0.07)	0.94 (0.02)	0.80 (0.05)	0.14 (0.04)	0.86 (0.04)	0.11 (0.03)	0.03 (0.00)	0.88 (0.08)	0.97 (0.06)	0.89 (0.06)	0.08 (0.04)	0.92 (0.04)	0.06 (0.04)	0.02 (0.03)
	400 - 800	0.80 (0.06)	0.97 (0.04)	0.82 (0.05)	0.12 (0.03)	0.88 (0.03)	0.10 (0.04)	0.02 (0.01)	0.87 (0.06)	0.97 (0.06)	0.88 (0.05)	0.08 (0.06)	0.92 (0.06)	0.07 (0.03)	0.02 (0.03)
	800 - 2500	0.82 (0.05)	0.87 (0.05)	0.82 (0.05)	0.15 (0.04)	0.85 (0.04)	0.09 (0.03)	0.04 (0.02)	0.90 (0.09)	0.87 (0.10)	0.90 (0.07)	0.12 (0.06)	0.88 (0.06)	0.05 (0.04)	0.07 (0.05)
PCA-QDA	**400 - 2500**	0.91 (0.02)	0.92 (0.02)	0.91 (0.02)	0.08 (0.01)	0.92 (0.00)	0.04 (0.01)	0.04 (0.01)	0.92 (0.08)	0.87 (0.10)	0.92 (0.07)	0.11 (0.02)	0.89 (0.01)	0.04 (0.04)	0.07 (0.05)
	400 - 800	0.90 (0.03)	0.89 (0.05)	0.90 (0.04)	0.11 (0.04)	0.89 (0.04)	0.04 (0.02)	0.05 (0.03)	0.92 (0.08)	0.90 (0.05)	0.92 (0.07)	0.09 (0.02)	0.91 (0.01)	0.05 (0.05)	0.06 (0.01)
	800 - 2500	0.92 (0.02)	0.90 (0.04)	0.92 (0.02)	0.09 (0.03)	0.91 (0.03)	0.04 (0.01)	0.05 (0.02)	0.92 (0.10)	0.78 (0.16)	0.92 (0.10)	0.15 (0.07)	0.85 (0.07)	0.04 (0.05)	0.11 (0.08)

Numbers in parentheses correspond to the standard deviation of three sets assayed.