Contents lists available at ScienceDirect

LWT

journal homepage: www.elsevier.com/locate/lwt

Raman spectroscopy coupled with chemometric modeling approaches for authentication of different paprika varieties at physiological maturity

Stefan Kolašinac [a,*], Ilinka Pećinar [a], Dario Danojević [b], Zora Dajić Stevanović [a]

[a] Faculty of Agriculture, Department of Agricultural Botany, Nemanjina 6, 11000, Belgrade-Zemun, Serbia
[b] Institute of Field and Vegetable Crops, Maksima Gorkog 30, 21000, Novi Sad, Serbia

ARTICLE INFO

Keywords:
Capsicum annuum
Carotenoids
Raman spectroscopy
Multivariate classification

ABSTRACT

Five Balkan paprika varieties at physiological maturity were investigated by means of Raman spectroscopy in order to discriminate the differences which stemmed from their genetic variability since the plants were grown under the same experimental conditions. The spectra were obtained using the 532 nm wavelength. In an effort to find the best classification power, several pre-processing methods were applied: 1) baseline correction, unit vector normalization; 2) baseline correction, unit vector normalization and first Savitzky-Golay derivative; 3) baseline correction, unit vector normalization and second Savitzky-Golay derivative; 4) baseline correction, unit vector normalization and third Savitzky-Golay derivative. All of the pre-processing methods were followed by making PCA-LDA (Principal Component Analysis-Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), and PLS-DA (Partial Least Square - Discriminant Analysis) classification models. QDA showed the best discrimination power (83.87–100% and 89.47–100% for the training and the test data, respectively), then PCA-LDA (0.00–100 and 0.00–100% for the training and the test data, respectively) and PLS-DA (19.35–100% and 0.00–100.00% for the training and the test data, respectively). The results pointed out the applicability of chemometric modeling associated with Raman spectroscopy in the assessment of nutritionally similar samples, such as the studied red paprika varieties.

1. Introduction

Paprika (*Capsicum* spp.) is among the most cultivated vegetable crops. According to the recent FAO reports, the global production of fresh and dried green chilies and peppers in 2019 was estimated at 38,024,154 and 4,255,050 tonnes, respectively (www.fao.org). Paprika is characterized by a high number of varieties differing in shape and color, starting from green, yellow or white (unripe fruits), and turning to red, dark red or brown. Significant quantities of red paprika are produced in Hungary, Serbia, Croatia and North Macedonia as it is one of the favorite and most frequently used vegetables in traditional cuisine (Vinković et al., 2018). Because of its high nutritive value, and especially due to the presence of different bioactive compounds, mainly fibers, L-ascorbic acid (vitamin C), carotenoids and polyphenols, sweet paprika is recommended for daily consumption in human diet (Greco, Riccio, Bergero, Del Re, & Trevisan; Kolašinac, Dajić-Stevanović, Kilibarda, & Kostić, 2021). Different carotenoids of paprika are responsible for different fruit coloration, as well as for the significant health-related beneficial effects, which are mainly attributed to capsanthin, capsorubin, beta carotene and lutein (Deli, Molnár, Matus, & Tóth, 2001). During the ripening period, the color of red paprika transforms from green to deep red due to the accumulation and transformation of carotenoids, which mostly refers to the conversion of lutein and beta carotene into zeaxanthin and capsanthin (Deli et al., 2001).

The level of bioactive compounds in paprika is influenced by the ripening stage, the genotype and cultivation practice (Hallmann & Rembiałkowska, 2012). The content of polyphenols, carotenoids and ascorbic acid increases during the fruit maturation reflecting increased antioxidant activity (Kim, Ahn, Ha, Rhee, & Kim, 2011).

Different analytical tools are used for food quality assessment, as well as for the confirmation of the geographic origin of target products and their possible adulteration. Regarding the quality assessment of different paprika varieties and their products, a standard technique and some advanced techniques are used, including NMR (Nuclear magnetic resonance) (Ramírez-Meraz et al., 2020), HPLC (High-performance liquid chromatography) (Cetó et al., 2018), VIS-NIRS (Monago-Maraña, Eskildsen, Galeano-Díaz, Muñoz de la Peña, & Wold, 2021), HPLC-FLD (High-performance liquid chromatography with fluorescence

* Corresponding author.
E-mail address: stefan.kolasinac@agrif.bg.ac.rs (S. Kolašinac).

https://doi.org/10.1016/j.lwt.2022.113402
Received 30 September 2021; Received in revised form 3 March 2022; Accepted 27 March 2022
Available online 5 April 2022

detection) (Campmajó, Rodríguez-Javier, Saurina, & Núñez, 2021), UHPLC (Mudrić et al., 2017), FT-MIR (Fourier transform mid-infrared) (Horn, Esslinger, Pfister, Fauhl-Hassek, & Riedl, 2018), UHPLC–APCI–HRMS (ultra-high-performance liquid chromatography coupled to high-resolution mass spectrometry using atmospheric pressure chemical ionization) (Arrizabalaga-Larrañaga et al., 2021), and some others.

Today, there is a great need for the application of non-destructive and rapid analytical methods in food quality control, as well as food product authentication and adulteration, since the standard instrumental methods are high-priced, time-consuming and require special sample preparation. In general, all of these methods target the detection of a narrow group of chemical compounds (such as proteins, carbohydrates, lipids, polyphenols, etc.), as well as some specific (target) compounds or even the isotopic forms of an element (such as ^{18}O, ^{13}C) (Danezis, Tsagkaris, Camin Brusic, & Georgiou, 2016). Moreover, high chemical complexity of plant and food samples makes quality control analysis difficult. The compound-based approach in the chemical characterization of a complex sample is focused on the precise targeting of one or a small number of desired compounds (Cetó, Sánches, Serrano, Díaz-Cruz, & Núñez, 2020). Consequently, a relatively new approach, which shifted from a component-based to a pattern-based approach, should be applied (Esteki, Shahsavari, & Simal-Gandara, 2019). This approach allows simultaneous recording of a number of compounds and together with chemometrics (pattern recognition), it could provide respectable information about the biological sample.

Raman spectroscopy is a molecular vibrational spectroscopic technique based on the interaction between monochromatic light (from VIS and UV region) and a sample. The observation of the Raman scattering signal depends on the change of its polarizability during a particular mode of vibration (Larkin, 2011). During this collision, the vibrational (rotational) energy of the molecule is changed, and the scattered radiation is shifted to a different wavelength, a Raman shift (Yang & Ying, 2011). Obtained Raman spectra contain information about the chemical bonds in the sample (Yang & Ying, 2011). Raman microspectroscopy combined with multivariate chemometric analysis has already been used as a promising tool for the discrimination and authentication of different carotenoid-rich food samples. In such approaches, a number of classification methods are used, including unsupervised (Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA)) and supervised classification models (Partial least square discrimination analysis (PLS-DA), Linear discriminant analysis (LDA), k-nearest neighbors (KNN), and Soft independent modeling of class analogy (SIMCA) (Akpolat et al., 2020; Ivleva, Niessner, & Panne, 2005; Kolašinac, Pećinar, Danojević, Aćić, & Dajić-Stevanović, 2021b; Monago-Maraña et al., 2019). Carotenoids are especially detectable by Raman (micro)spectroscopy due to their polyene molecular structure (Schulz, Baranska, & Baranski, 2005). In addition, carotenoids are among the key nutrients in red paprika, contributing to both sensory, i.e. visual attractiveness (as they are responsible for fruit and product coloration) and nutritive quality of the fruit (Pugliese et al., 2014).

There are several reports addressing the use of Raman spectroscopy in the characterization of carotenoids in paprika, including the studies on the carotenoid content during ripening (e.g. Sharma, Sarika Bharti, Singh, & Uttam, 2019), the distribution of carotenoids along the fruit pericarp (fruit wall) (e.g. Baranski, Baranska, & Schulz, 2005) and the determination of particular carotenoid compounds in chili peppers (Sharma et al., 2019). However, to the best of our knowledge, no reports addressed the application of Raman spectroscopy as a tool in the discrimination of paprika varieties based on the differences between the components appearing within the target spectral range, mostly the total carotenoids and some individual carotenoid compounds.

The main objective of this paper is to establish the best multivariate classification model for the discrimination and therefore, the authentication of different paprika varieties at their final maturity stage, considering the needs for defining the best harvest practices, the quality assessment and the control of final food products.

2. Material and methods

2.1. Plant material

The samples of the traditional Balkan red paprika varieties (Amfora, Una, Kurtovska Kapija and Vrtka), as well as one inbreed line (derived from a cross between cultivar Amfora and Una) were used in the experiment (Fig. 1). The plants were grown under the same agro-ecological conditions, at the experimental field plots of the Institute of Field and Vegetable Crops, Novi Sad, the Republic of Serbia. The samples were harvested in the last week of September depending on the full maturity phase of each variety, i.e. as soon as the deep red color appeared across the entire fruit's surface. The fruits were previously marked to ensure the same harvest conditions for all of the samples. Five fruits were collected from each variety. After washing the fruits with deionized water, several fragments of about 2 cm^2 were taken from each sample for Raman spectroscopy recording. Spectra were recorded from the epidermal layer, i.e the fruits' surface was studied.

2.2. Raman instrumentation and spectra recording

Horiba Raman spectrometer system (Horiba Jobin Yvon, France) equipped with the Olympus BX 41 microscope was used in this study. This system possesses 532-nm and 785-nm laser sources. During the spectral recording, the 532-nm laser was focused onto the sample on the microscope stage through a 50 LWD (long working distance) objective (Olympus, Tokyo, Japan). The spectrometer is equipped with 1200 lines/mm grating. Raman scattering signals were detected by a charge-coupled-device (CCD) detector, the detection range from 900 to 1800 cm^{-1} in the extended mode. The measurement was conducted with a 5s integration time, with 10 spectral accumulations, and 20–25 mW maximum output laser power. The spectral data were collected with LabSpec 6 (Horiba, France). The spectral resolution was about 3 cm^{-1} and the calibration was checked by a 520.47 cm^{-1} line of silicon. In order to consider a possible sample inhomogeneity, the experiments were performed in five replicates of each paprika variety. Ten spectra were recorded per each replicate making a total of 50 spectra per variety, i.e. 250 spectra in total. The assignment of the major bands was carried out using the literature data.

2.3. Chemometric analysis

Obtained raw spectral data were arranged in a matrix with 250 rows (objects-each variety 50 repetitions) and 334 columns (variables-wavenumbers). Raw spectra were pre-processed by applying the following procedures: 1) baseline correction, unit vector normalization; 2) baseline correction, unit vector normalization and first Savitzky-Golay derivative; 3) baseline correction, unit vector normalization and second Savitzky-Golay derivative; 4) baseline correction, unit vector normalization and third Savitzky-Golay derivative. After derivation, the spectra were smoothed with 13 smoothing points (Fig. 2).

The final stage requires the introduction of the Principal Component Analysis (PCA), prior to the application of different classification methods. PCA is required since a large number of independent variables usually cause misclassification problems (Abdul Rashid, Siti Esah Che Hussain, Razak Ahmad, & Norazami Abdullah, 2019). PCA is a method for reducing a large number of data sets of possibly correlated variables to a smaller number of uncorrelated variables (property of orthogonality) called principal components (PCs). Acquired PCs serve as the input for tested classification methods, including: PCA-LDA (Principal Component Analysis-Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), and PLS-DA (Partial Least Square - Discriminant Analysis).

LDA method serves as a model for separation into two classes

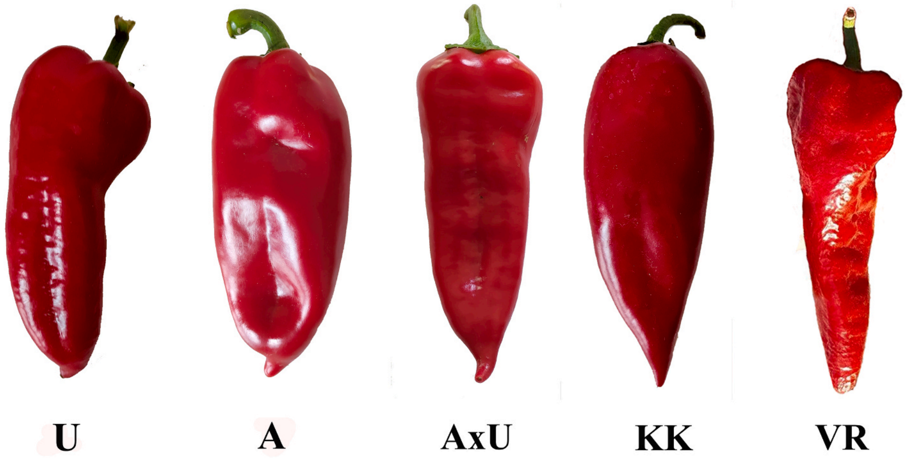

Fig. 1. Varieties used in the experiment: U-Una, A-Amfora, AxU-Amfora x Una, KK-Kurtovska kapija, VR-Vrtka.

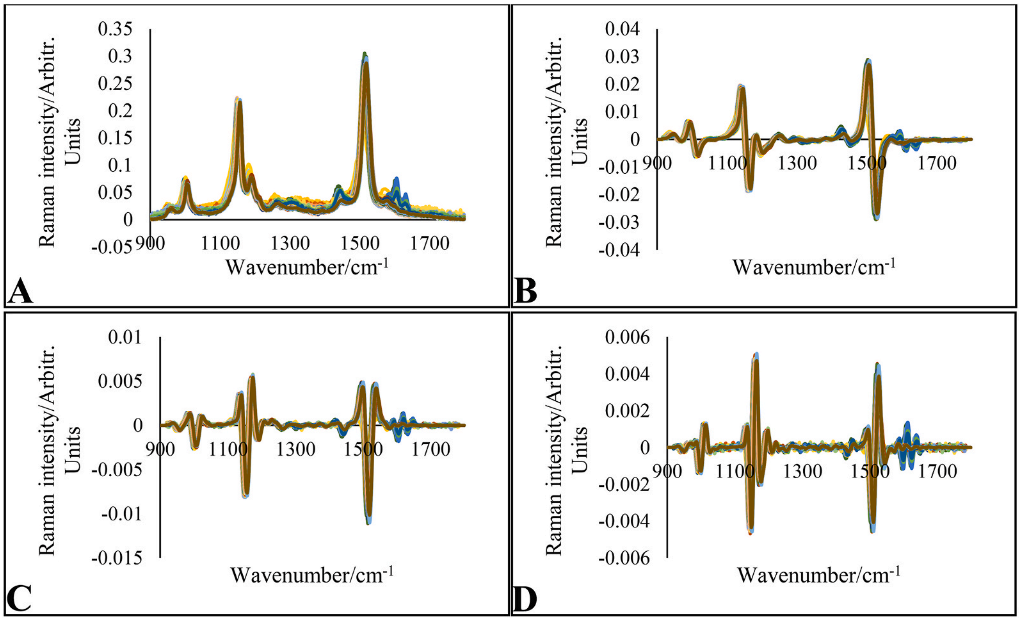

Fig. 2. Pre-processed spectra: A) baseline correction, unit vector normalization; B) baseline correction, unit vector normalization and the first Savitzky-Golay derivative; C) baseline correction, unit vector normalization and second Savitzky-Golay derivative; D) baseline correction, unit vector normalization and the third Savitzky-Golay derivative.

(Pomerantsev, 2014). The aim is to find the ideal projections and carry out the discrimination on the projected subspace. Distances between classes are maximized through projection while keeping minimum distance between the objects in the same class (Chen & Jiang, 2018). LDA creates linear boundaries by dividing the variable space into regions with a straight line or hyperplane (Dixon & Brereton, 2009).

QDA is a multiclass method, which can be used for the discrimination of several classes simultaneously. In contrast to LDA, QDA obtains quadratic boundaries, where a quadratic curve divides the variable space into regions (Dixon & Brereton, 2009). PLS-DA classification method is based on the search for the components or latent variables which serve to discriminate two or more different groups. Discrimination is implemented according to their maximum covariance with a target class (Uarrota et al., 2014).

To conduct multivariate classification methods, all data were divided into two sets: the training set (75% of the samples) and the test set (25% of the samples). The training set was used to establish the classification rules while the test set was used to validate them. Goodness of models was evaluated based on the highest value of sensitivity (SE) specificity (SP) and precission (P) in both training and test sample sets. SE refers to the percentage of samples of a given class that the model correctly recognises as belonging to that class:

$$SE = TP/(TP + FN)$$

SP refers to samples that do not belong to a given class and are correctly rejected by the model

$$SP = TN/(TN + FP)$$

Precision tells what fraction of predictions as a positive class were actually positive. To calculate precision, use the following formula:

$$P = TP/(TP + FP)$$

Computation of sensitivity, specificity and precission for PCA-LDA model is based on 2 × 2 confusion matrix (Table 1).

Unlike binary classification, PCA-QDA and PLS-DA were performed as multiclass discrimination models. Hence, metrics needed to calculate SE, SP and P for each variety were computed according more complex

Table 1
Confusion matrix. The table compares the prediction against the correct group assignment.

		Actual values	
		positive	negative
Predicted values	positive	TP	FP
	negative	FN	TN

*TP- true positive; TN- true negative; FP-false positive; FN-false negative.

confusion matrix (Table 2) and following equations:

3. Results and discussion

3.1. Raman spectra

Investigated paprika fruit samples were very similar in shape and color at the final maturity stage (Fig. 1). According to the obtained Raman spectra, it was clear that some subtle differences between the samples existed (Fig. 3). Fig. 2 shows the results of different pre-processing methods. Accordingly, 2B, 2C and 2D show that the changes in the Raman intensity and the bands' position (especially in the low wavenumber region) are more clearly visible using the 1st, 2nd and 3rd order derivatives. However, since there was a lot of noise coming from fluorescence, at that point, useful information was not obtained.

The literature data indicate that the red paprika varieties can slightly differ in their chemical composition at the final (deep red) physiological state, which is mainly related to some differences in the concentration of total polyphenols (Mudrić et al., 2017), flavonoids (Tundis et al., 2011) and carotenoids (Collera-Zúñiga, García Jiménez, & Meléndez Gordillo, 2005). Discrimination analysis in spectroscopy is a model of classification of observed objects (samples) according to their spectra. Generally, classification methods are divided into two groups: supervised, where the data are portioned according to their similarity into pre-defined groups; and unsupervised, where there is no prior information about the groups (Pomerantsev, 2014). So far, Raman spectroscopy coupled with chemometric classification methods has been used as a tool for discrimination of several fennel chemotypes (Gudi, Krähmer, Krüger, Hennig, & Schulz, 2014) and coffee varieties (Keidel, von Stetten, Rodrigues, Máguas, & Hildebrandt, 2010; Luna, da Silva, da Silva, Lima, & de Gois, 2019. In these studies, the Hierarchical Cluster Analysis (HCA), Soft Independent Modeling of Class Analogy (SIMCA), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were applied.

Significant bands associated with the pericarp essential constituents were observed at ∼1511–1519 cm^{-1} (very strong), ∼1149–1151 cm^{-1} (medium) and ∼998–1006 cm^{-1} (low intensity) (Fig. 3B), representing the characteristic bands of carotenoids. Although the carotenoids were present in the sample as minor components (in a quantity less than 1 mg kg^{-1}), they were very sensitive in the Raman visible region due to the signal enhancement caused by the pre-resonance effect of the analyte (Baranski et al., 2005). It is thought that the observed bands were assigned to the stretching vibration of the $C{=}C$, $C{-}C$ bonds and the $C{-}CH_3$ in-plane group rocking vibrations, respectively (de Oliveira, Castro, Edwards, & de Oliveira, 2009; Schulz et al., 2005). These bands suggest the presence of carotenoids as major secondary metabolites in a wide range of plant species including the paprika fruit (Sharma et al., 2019). According to Schulz et al. (2005), the main bands of the red paprika fruit are observed at 1517, 1158 and 1004 cm^{-1} which is assigned to the capsanthin, the main carotenoid in red paprika at the physiological maturity stage. de Oliveira et al. (2009) reported that the bands at 1527, 1156 and 1006 cm^{-1} can be allocated to beta carotene.

In addition to the bands corresponding to carotenoids, the weak bands were identified below 1000, at 1572–1575, 1440–1442 and ∼1260 cm^{-1}, and in the region of 1600–1625 cm^{-1} (Fig. 3B). As the parenchyma cells of paprika pericarp are very rich in various carbohydrates, the bands observed below 1000 cm^{-1} probably refer to glycosidic linked stretches (Synytsya, Čopí;ková, Matějka, & Machovič, 2003). The bands observed at ∼1445 cm^{-1} (related to $\delta(CH_2)$ vibrational mode) and at ∼1256 cm^{-1} are most probably linked to the polygalacturonic acid (Chylińska, Szymańska-Chargot, & Zdunek, 2014) (Fig. 3B). The weak band at 1575 cm^{-1} is associated with the presence of benzene ring (Trebolazabala, Maguregui, Morillas, de Diego, & Madariaga, 2017) which is the main constituent of phenolic compounds. In addition to this one, the weak band at ∼1600 cm^{-1} can also be attributed to phenolic compounds (Prats Mateu, Hauser, Heredia, & Gierlinger, 2016). It is known that the red pepper is rich in flavonoid glycosides and phenolic acids such as quercetin-3-glycoside (Lekala et al., 2019).

Table 2
Confuson matrix of multiclass classification problem.

		Actual values				
Predicted values		U	A	AU	KK	VR
	U	a	b	c	d	e
	A	f	g	h	i	j
	AU	k	l	m	n	o
	KK	p	q	r	s	t
	VR	u	v	w	x	y

3.2. Multivariate classification analysis

It has already been discussed that the target compound classification analysis could sometimes be useless because of the fact that there is rarely a specific compound present only in one single type of a sample, which could consequently be used to separate (classify) samples according to such fine chemical differences. The non-targeted analysis also has some disadvantages, mainly because it requires many descriptors, some of which might be irrelevant, thus interfering with the data processing stage and possibly causing degradation of model performance. In our experiment, all classification rules are built upon the spectral region between 900 and 1800 cm^{-1}, which is mostly associated with the bands corresponding to the carotenoids.

Since the classification models are unpredictable due to unknown discrimination power, it is necessary to perform several models based on accepted pre-processing methods (Devos, Downey, & Duponchel, 2014). In all tested classification models, five principal components were used, explaining 99% of the entire variability.

The precision of PCA-LDA, PLS-DA and PCA-QDA in the training data was 0–100, 19.35–100 and 83.87–100%, respectively, while in the test data it was 0.00–100.00 0.00–100.00 and 89.47–100.00%, respectively. The results varied depending on the performed pre-processing method. In general, the applied pre-processing methods did not change the classification precision but PCA-LDA and PLS-DA showed uncertainty in the case where the number of corrected classified samples was zero. PCA-QDA showed the best classification power and consistency in all of the pre-processing methods used (for the graphical representation of the classification results see Supplementary material).

The main problem of the PLS-DA model was to classify the Amfora variety and Amfora x Una inbreed line. The effect could be connected with the similar genetic structure of these two samples, since they share a high number of common genes, including those possibly determining the metabolic pathways of synthesis of carotenoids, polyphenols, sugars and other metabolites. PCA-LDA showed better performance in the classification of these two samples, while PCA-QDA did not show weakness in the classification of genetically related samples. PLS-DA and PCA-LDA are based on Mahalanobis distance, but LDA presumes a single variance-covariance matrix over all classes. However, QDA presumes different variance-covariance matrices for each class separately, creating a more powerful classification rule (Dixon & Brereton, 2009) which can be the reason for the best discrimination power. In the event of incorrect classification, the varieties were randomly assigned to a

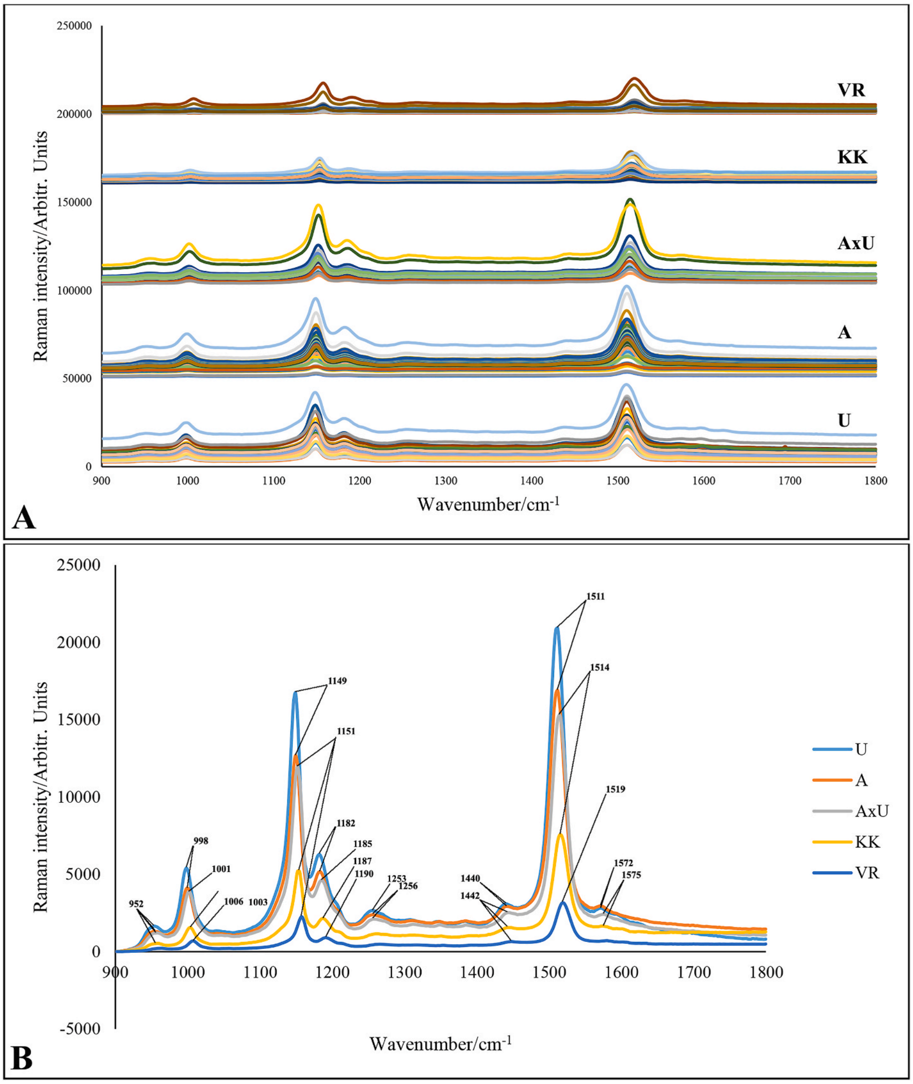

Fig. 3. Raw (A) and averaged (B) Raman spectra of paprika varieties: U-Una; A-Amfora; AxU-Amrofa x Una; KK-Kurtovska kapija; VR-Vrtka.

certain class.

The summary of the discrimination of each model and each pre-processing model combination is shown in Tables 3 and 4.

QDA is a very simple algorithm, and as opposed to LDA, it computes the variance structures for each class separately, creating a more powerful discrimination rule for classes with different covariance matrices, such as for biological spectra sets in which the variability within the class is the key issue.

Table 3
Classification results of Training sets of PCA-LDA, PLSDA and QDA models.

Method	Variety	Training set												Overall
		BC + N			BC + N+ 1st D			BC + N+ 2nd D			BC + N+ 3rd D			
		P	SE	SP	P	SE	SP	P	SE	SP	P	SE	SP	
	U	100.00	96.77	100.00	100.00	87.10	79.03	100.00	87.10	97.58	93.10	87.10	98.39	93.85
	A	50.94	100.00	100.00	50.94	100.00	100.00	90.00	100.00	100.00	100.00	96.77	100.00	90.72
	AU	100.00	100.00	100.00	100.00	83.87	100.00	0.00	0.00	100.00	100.00	9.68	100.00	74.46
	KK	0.00	100.00	100.00	0.00	0.00	100.00	0.00	0.00	100.00	50.00	100.00	0.00	45.83
PCA-LDA	VR	0.00	100.00	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00	100.00	41.67
	U	54.54	83.87	80.64	60.00	87.10	79.84	75.00	88.57	84.68	72.73	90.32	81.45	78.23
	A	52.00	19.35	95.97	51.92	19.35	96.77	58.70	34.29	96.77	54.90	25.81	97.58	58.62
	AU	96.87	100.00	99.19	62.22	90.32	86.29	100.00	83.87	100.00	100.00	83.87	100.00	91.89
	KK	100.00	100.00	100.00	82.35	51.61	97.58	86.11	100.00	95.97	86.11	100.00	95.97	91.31
PLS-DA	VR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	U	93.94	93.55	100.00	85.71	83.87	99.19	88.57	87.10	100.00	91.18	90.32	100.00	92.79
	A	100.00	100.00	98.39	96.30	96.77	95.97	100.00	100.00	96.77	100.00	100.00	97.58	98.48
	AU	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	KK	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
PCA-QDA	VR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 4
Classification results of Test sets of PCA-LDA, PLSDA and QDA models.

Method	Variety	Test set												Overall
		BC + N			BC + N+ 1st D			BC + N+ 2nd D			BC + N+ 3rd D			
		P	SE	SP	P	SE	SP	P	SE	SP	P	SE	SP	
PCA-LDA	U	0.00	0.00	50.00	100.00	100.00	86.84	100.00	100.00	92.10	100.00	100.00	94.74	76.97
	A	0.00	0.00	33.33	65.51	84.21	100.00	76.00	94.74	100.00	82.61	94.74	100.00	69.26
	AU	33.33	100.00	0.00	86.36	100.00	95.16	100.00	100.00	100.00	100.00	100.00	100.00	84.57
	KK	50.00	100.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	87.50
	VR	50.00	100.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	87.50
PLS-DA	U	0.00	0.00	75.00	61.29	100.00	84.21	61.29	100.00	84.21	59.37	100.00	82.90	67.36
	A	20.00	0.00	80.26	100.00	26.32	100.00	100.00	26.32	100.00	84.04	21.05	100.00	63.17
	AU	0.00	5.26	94.74	70.83	89.47	90.79	77.27	89.47	93.42	94.74	89.47	96.05	74.29
	KK	0.00	0.00	76.32	87.50	73.68	97.37	83.33	78.95	96.05	92.63	84.21	94.74	72.07
	VR	0.00	0.00	50.00	100.00	94.74	100.00	100.00	94.74	100.00	98.95	94.74	100.00	77.76
PCA-QDA	U	100.00	100.00	100.00	90.00	89.47	100.00	100.00	89.47	100.00	94.44	89.47	100.00	96.07
	A	100.00	94.74	100.00	100.00	94.74	97.37	90.47	100.00	97.37	90.00	94.74	97.37	96.40
	AU	95.00	100.00	98.68	95.00	100.00	98.68	100.00	100.00	100.00	100.00	100.00	100.00	98.95
	KK	100.00	100.00	100.00	100.00	100.00	100.00	100.00	88.00	100.00	89.47	89.47	100.00	97.25
	VR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

BC + N - baseline correction, unit vector normalization; BC + N+1st D - baseline correction, unit vector normalization and first Savitzky-Golay derivative; BC + N+2nd D - baseline correction, unit vector normalization and second Savitzky-Golay derivative; BC + N+3rd D - baseline correction, unit vector normalization and third Savitzky-Golay derivative; P-precission; SE-sensitivity; SP-specificity.

Kolašinac, Pećinar, Danojević, Aćić, and Stevanović (2021) investigated different chemometric models in the classification of red pepper cultivars at different maturity stages. According to the training data in this research, PCA-LDA and PLS-DA indicated precision between 95% and 100%, respectively, while in the test data they showed 90–100% and 100% precision, respectively. Luna et al. (2019) used PCA-LDA, QDA and PLS-DA in combination with different pre-processing methods coupled with Raman spectroscopy for the classification of different coffee varieties. The results showed that precision was equal to 100% for all tested models except for QDA (97.8%), using MSC (multiplicative scatter correction) as a pre-processing method. On the other hand, when MC (mean centering) was used, the correct classification of the samples was 62.7% for PCA-LDA, 62.7% for QDA and 61.3% for PLS-DA.

The importance of this research is reflected in the potential use of Raman spectroscopy in the discrimination of very similar plant samples, such as some crop varieties grown at the same conditions and harvested at the same maturity phase. The paprika samples investigated in this paper are mostly used as raw material to obtain various food products that are available at the market. The most preferable and the most consumed product is the traditional Balkan dish called "ajvar", in addition to roasted and ground sweet red paprika. "Ajvar" is acknowledged Balkan food, prepared from just a few paprika varieties, known as "ajvar" or "ajvarusa" paprika types in Serbian, Bosnian and North-Macedonian languages. Such varieties as "Kurtovska kapija", "Amfora", "Crvena roga" and some others, have typical fleshy and taste-rich mesocarp (paprika "flesh"), shiny deep red color and elongated, cylindrical and somewhat flattened regular shape, suitable for efficient roasting. Best quality ajvar is made from just roasted paprika, vegetable oil and salt, where only one variety is used depending on the region, availability on the market and the price. In addition to homemade ajvar, there is significant industrial production offered by several food processing companies in the Balkans (e.g. "Bakina tajna", "Premija", "Podravka", "Dijamant", etc.), due to high demand and consumer preferences. Alongside ajvar, fleshy deep red paprika varieties are also used fresh, roasted in different products, as well as in the form of dried paprika and sweet ground paprika product. Therefore, future research should be directed towards the quality assessment of paprika products and the differences which could be accurately estimated by a non-destructive and rapid method such as Raman spectroscopy associated with proper chemometric models. Moreover, Raman imaging could be recommended for studies of alterations in nutrient content and composition in different parts of the fruit allowing a deeper understanding of carotenoid transformation during the ripening period and therefore performing the best harvest practices for different capsicum species and varieties.

4. Conclusion

Results obtained in this study show that the Raman spectroscopy coupled with multivariate classification analysis can be used as a tool for the discrimination of different paprika varieties. Overall, this paper aims to demonstrate the advantages derived from the use of chemometric methods as an alternative to a component-based approach. The pattern-based approach has proved to be precise, fast, low-cost and time-saving. The best classification results were acquired using QDA in all pre-treatment procedures, especially with baseline-correction and normalization. Finally, our study confirmed the possible applicability of the chemometrics-based Raman spectroscopy in food quality analyses, as well as in the authentication of food products, food adulteration and the assessment of the product's geographic origin, since the method allows determination of subtle alterations in physico-chemical traits of different products and at different processing phases.

CRediT authorship contribution statement

Stefan Kolašinac: Conceptualization, Data curation, Methodology, Investigation, Writing – original draft. **Ilinka Pećinar:** Formal analysis, Investigation. **Dario Danojević:** Resources. **Zora Dajić Stevanović:** Conceptualization, Methodology, Formal analysis, Funding acquisition, Project administration, Writing – review & editing, Supervision.

Acknowledgments

This work was supported by Ministry of Education, Science and Technological Development of Republic of Serbia and Science Fund of the Republic of Serbia, #GRANT No 7750160, R-SPECT.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lwt.2022.113402.

BC + N - baseline correction, unit vector normalization; BC + N+1st D - baseline correction, unit vector normalization and first Savitzky-Golay derivative; BC + N+2nd D - baseline correction, unit vector normalization and second Savitzky-Golay derivative; BC + N+3rd D -

baseline correction, unit vector normalization and third Savitzky-Golay derivative; P-precission; SE-sensitivity; SP-specificity.

References

Abdul Rashid, N., Siti Esah Che Hussain, W., Razak Ahmad, A., & Norazami Abdullah, F. (2019). Performance of classification analysis: A comparative study between PLS-DA and integrating PCA+LDA. *Mathematics and Statistics, 7*(4A), 24–28. https://doi.org/10.13189/ms.2019.070704

Akpolat, H., Barineau, M., Jackson, K. A., Akpolat, M. Z., Francis, D. M., Chen, Y.-J., et al. (2020). High-throughput phenotyping approach for screening major carotenoids of tomato by handheld Raman spectroscopy using chemometric methods. *Sensors, 20*(13), 3723. https://doi.org/10.3390/s20133723

Arrizabalaga-Larrañaga, A., Campmajó, G., Saurina, J., Núñez, O., Santos, F. J., & Moyano, E. (2021). Determination of capsaicinoids and carotenoids for the characterization and geographical origin authentication of paprika by UHPLC–APCI–HRMS. *Lebensmittel-Wissenschaft und -Technologie- Food Science and Technology, 139*, Article 110533. https://doi.org/10.1016/j.lwt.2020.110533

Baranski, R., Baranska, M., & Schulz, H. (2005). Changes in carotenoid content and distribution in living plant tissue can be observed and mapped in situ using NIR-FT-Raman spectroscopy. *Planta, 222*(3), 448–457. https://doi.org/10.1007/s00425-005-1566-9

Campmajó, G., Rodríguez-Javier, L. R., Saurina, J., & Núñez, O. (2021). Assessment of paprika geographical origin fraud by high-performance liquid chromatography with fluorescence detection (HPLC-FLD) fingerprinting. *Food Chemistry, 352*, Article 129397. https://doi.org/10.1016/j.foodchem.2021.129397

Cetó, X., Sánches, C., Serrano, N., Díaz-Cruz, J., & Núñez, O. (2020). Authentication of paprika using HPLC-UV fingerprints. *Lebensmittel-Wissenschaft und -Technologie- Food Science and Technology, 124*, Article 109153. https://doi.org/10.1016/j.lwt.2020.109153

Cetó, X., Serrano, N., Aragó, M., Gámez, A., Esteban, M., Díaz-Cruz, J., et al. (2018). Determination of HPLC-UV fingerprints of Spanish paprika (capsicum annuum L.) for its classification by linear discriminant analysis. *Sensors, 18*(12), 4479. https://doi.org/10.3390/s18124479

Chen, L.-H., & Jiang, C.-R. (2018). Sensible functional linear discriminant analysis. *Computational Statistics & Data Analysis, 126*, 39–52. https://doi.org/10.1016/j.csda.2018.04.005

Chylińska, M., Szymańska-Chargot, M., & Zdunek, A. (2014). Imaging of polysaccharides in the tomato cell wall with Raman microspectroscopy. *Plant Methods, 10*(1), 14. https://doi.org/10.1186/1746-4811-10-14

Collera-Zúñiga, O., García Jiménez, F., & Meléndez Gordillo, R. (2005). Comparative study of carotenoid composition in three mexican varieties of Capsicum annuum L. *Food Chemistry, 90*(1–2), 109–114. https://doi.org/10.1016/j.foodchem.2004.03.032

Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. *TRAC Trends in Analytical Chemistry, 85*, 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Deli, J., Molnár, P., Matus, Z., & Tóth, G. (2001). Carotenoid composition in the fruits of red paprika (capsicum annuum var. lycopersiciforme rubrum) during ripening; biosynthesis of carotenoids in red paprika. *Journal of Agricultural and Food Chemistry, 49*(3), 1517–1523. https://doi.org/10.1021/jf000958d

Devos, O., Downey, G., & Duponchel, L. (2014). Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. *Food Chemistry, 148*, 124–130. https://doi.org/10.1016/j.foodchem.2013.10.020

Dixon, S. J., & Brereton, R. G. (2009). Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on. *Chemometrics and Intelligent Laboratory Systems, 95*(1), 1–17. https://doi.org/10.1016/j.chemolab.2008.07.010

Esteki, M., Shahsavari, Z., & Simal-Gandara, J. (2019). Food identification by high performance liquid chromatography fingerprinting and mathematical processing. *Food Research International, 122*, 303–317. https://doi.org/10.1016/j.foodres.2019.04.025

Greco, L., Riccio, R., Bergero, S., Del Re, A. A. M., & Trevisan, M. (2007). Total reducing capacity of fresh sweet peppers and five different Italian pepper recipes. *Food Chemistry, 103*(4), 1127–1133. https://doi.org/10.1016/j.foodchem.2006.10.013

Gudi, G., Krähmer, A., Krüger, H., Hennig, L., & Schulz, H. (2014). Discrimination of fennel chemotypes applying IR and Raman spectroscopy: Discovery of a new γ-asarone chemotype. *Journal of Agricultural and Food Chemistry, 62*(16), 3537–3547. https://doi.org/10.1021/jf405752x

Hallmann, E., & Rembiałkowska, E. (2012). Characterisation of antioxidant compounds in sweet bell pepper (Capsicum annuum L.) under organic and conventional growing systems. *Journal of the Science of Food and Agriculture, 92*(12), 2409–2415. https://doi.org/10.1002/jsfa.5624

Horn, B., Esslinger, S., Pfister, M., Fauhl-Hassek, C., & Riedl, J. (2018). Non-targeted detection of paprika adulteration using mid-infrared spectroscopy and one-class classification – is it data preprocessing that makes the performance? *Food Chemistry, 257*, 112–119. https://doi.org/10.1016/j.foodchem.2018.03.007

Ivleva, N. P., Niessner, R., & Panne, U. (2005). Characterization and discrimination of pollen by Raman microscopy. *Analytical and Bioanalytical Chemistry, 381*(1), 261–267. https://doi.org/10.1007/s00216-004-2942-1

Keidel, A., von Stetten, D., Rodrigues, C., Máguas, C., & Hildebrandt, P. (2010). Discrimination of green Arabica and Robusta coffee beans by Raman spectroscopy. *Journal of Agricultural and Food Chemistry, 58*(21), 11187–11192. https://doi.org/10.1021/jf101999c

Kim, J.-S., Ahn, J.-Y., Ha, T.-Y., Rhee, H.-C., & Kim, S.-A. (2011). Comparison of phytochemical and antioxidant activities in different color stages and varieties of paprika harvested in Korea. *Korean Journal of Food Science and Technology, 43*(5), 564–569. https://doi.org/10.9721/KJFST.2011.43.5.564

Kolašinac, M. S., Dajić Stevanović, P., Kilibarda, Z. N., & Kostić, A. (2021a). Carotenoids: New applications of "Old" pigments. *Phyton, 90*(4), 1041–1062. https://doi.org/10.32604/phyton.2021.015996

Kolašinac, S., Pećinar, I., Danojević, D., Aćić, S., & Stevanović, Z. D. (2021b). Raman spectroscopic-based chemometric modeling in assessment of red pepper ripening phases and carotenoids accumulation. *Journal of Raman Spectroscopy*. https://doi.org/10.1002/jrs.6197. jrs.6197.

Larkin, P. (2011). *Infrared and Raman spectroscopy - principles and spectral interpretation*. Amsterdam: Elsevier Ltd. (Chapter 2).

Lekala, C. S., Madani, K. S. H., Phan, A. D. T., Maboko, M. M., Fotouo, H., Soundy, P., et al. (2019). Cultivar-specific responses in red sweet peppers grown under shade nets and controlled-temperature plastic tunnel environment on antioxidant constituents at harvest. *Food Chemistry, 275*, 85–94. https://doi.org/10.1016/j.foodchem.2018.09.097

Luna, A. S., da Silva, A. P., da Silva, C. S., Lima, I. C. A., & de Gois, J. S. (2019). Chemometric methods for classification of clonal varieties of green coffee using Raman spectroscopy and direct sample analysis. *Journal of Food Composition and Analysis, 76*, 44–50. https://doi.org/10.1016/j.jfca.2018.12.001

Monago-Maraña, O., Eskildsen, C. E., Afseth, N. K., Galeano-Díaz, T., Muñoz de la Peña, A., & Wold, J. P. (2019). Non-destructive Raman spectroscopy as a tool for measuring ASTA color values and Sudan I content in paprika powder. *Food Chemistry, 274*, 187–193. https://doi.org/10.1016/j.foodchem.2018.08.129

Monago-Maraña, O., Eskildsen, C. E., Galeano-Díaz, T., Muñoz de la Peña, A., & Wold, J. P. (2021). Untargeted classification for paprika powder authentication using visible – near infrared spectroscopy (VIS-NIRS). *Food Control, 121*, Article 107564. https://doi.org/10.1016/j.foodcont.2020.107564

Mudrić, S.Ž., Gašić, U. M., Dramićanin, A. M., Ćirić, I.Ž., Milojković-Opsenica, D. M., Popović-Đorđević, J. B., et al. (2017). The polyphenolics and carbohydrates as indicators of botanical and geographical origin of Serbian autochthonous clones of red spice paprika. *Food Chemistry, 217*, 705–715. https://doi.org/10.1016/j.foodchem.2016.09.038

de Oliveira, V. E., Castro, H. V., Edwards, H. G. M., & de Oliveira, L. F. C. (2009). Carotenes and carotenoids in natural biological samples: A Raman spectroscopic analysis. *Journal of Raman Spectroscopy, 41*(6), 642–650. https://doi.org/10.1002/jrs.2493

Pomerantsev, A. L. (2014). *Chemometrics in excel*. John Wiley & Sons, Inc.

Prats Mateu, B., Hauser, M. T., Heredia, A., & Gierlinger, N. (2016). Waterproofing in Arabidopsis: Following phenolics and lipids in situ by confocal Raman microscopy. *Frontiers of Chemistry, 4*. https://doi.org/10.3389/fchem.2016.00010

Pugliese, A., O'Callaghan, Y., Tundis, R., Galvin, K., Menichini, F., O'Brien, N., et al. (2014). In vitro investigation of the bioaccessibility of carotenoids from raw, frozen and boiled red chili peppers (Capsicum annuum). *European Journal of Nutrition, 53*(2), 501–510. https://doi.org/10.1007/s00394-013-0555-1

Ramírez-Meraz, M., Méndez-Aguilar, R., Hidalgo-Martínez, D., Villa-Ruano, N., Zepeda-Vallejo, L. G., Vallejo-Contreras, F., et al. (2020). Experimental races of Capsicum annuum cv. jalapeño: Chemical characterization and classification by 1H NMR/machine learning. *Food Research International, 138*, Article 109763. https://doi.org/10.1016/j.foodres.2020.109763

Schulz, H., Baranska, M., & Baranski, R. (2005). Potential of NIR-FT-Raman spectroscopy in natural carotenoid analysis. *Biopolymers, 77*(4), 212–221. https://doi.org/10.1002/bip.20215

Sharma, S., Sarika Bharti, A., Singh, R., & Uttam, K. N. (2019). Non-destructive phenotyping of chili pepper ripening using spectroscopic probes: A potential approach for shelf-life measurement. *Analytical Letters, 52*(10), 1590–1613. https://doi.org/10.1080/00032719.2018.1558231

Synytsya, A., Čopíková, J., Matějka, P., & Machovič, V. (2003). Fourier transform Raman and infrared spectroscopy of pectins. *Carbohydrate Polymers, 54*(1), 97–106. https://doi.org/10.1016/S0144-8617(03)00158-9

Trebolazabala, J., Maguregui, M., Morillas, H., de Diego, A., & Madariaga, J. M. (2017). Portable Raman spectroscopy for an in-situ monitoring the ripening of tomato (Solanum lycopersicum) fruits. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 180*, 138–143. https://doi.org/10.1016/j.saa.2017.03.024

Tundis, R., Loizzo, M. R., Menichini, F., Bonesi, M., Conforti, F., Statti, G., et al. (2011). Comparative study on the chemical composition, antioxidant properties and hypoglycaemic activities of two capsicum annuum L. Cultivars (Acuminatum small and Cerasiferum). *Plant Foods for Human Nutrition, 66*(3), 261–269. https://doi.org/10.1007/s11130-011-0248-y

Uarrota, V. G., Moresco, R., Coelho, B., Nunes, E. da C., Peruch, L. A. M., Neubert, E. de O., et al. (2014). Metabolomics combined with chemometric tools (PCA, HCA, PLS-DA and SVM) for screening cassava (Manihot esculenta Crantz) roots during postharvest physiological deterioration. *Food Chemistry, 161*, 67–78. https://doi.org/10.1016/j.foodchem.2014.03.110

Vinković, T., Gluščić, V., Mendaš, G., Vinković Vrček, I., Parađiković, N., Tkalec, M., et al. (2018). Phytochemical composition of ground paprika from the eastern Danube region. *Poljoprivreda, 24*(2), 3–12. https://doi.org/10.18047/poljo.24.2.1

Yang, D., & Ying, Y. (2011). Applications of Raman spectroscopy in agricultural products and food analysis: A review. *Applied Spectroscopy Reviews, 46*(7), 539–560. https://doi.org/10.1080/05704928.2011.593216