Journal of Food Composition and Analysis 77 (2019) 84–90

Contents lists available at ScienceDirect

Journal of Food Composition and Analysis

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

Original research article

Predicting the physicochemical properties and geographical ORIGIN of lentils using near infrared spectroscopy

I. Revilla[a,*], C. Lastras[a], M.I. González-Martín[b], A.M. Vivar-Quintana[a], R. Morales-Corts[c], M.A. Gómez-Sánchez[c], R. Pérez-Sánchez[c]

[a] Area of Food Technology, University of Salamanca, EPS de Zamora, 49022, Zamora, Spain
[b] Department of Analytical Chemistry, Nutrition and Bromatology, University of Salamanca, Chemistry Faculty, 37008, Salamanca, Spain
[c] Area of Vegetable Production, University of Salamanca, Faculty of Agricultural and Environmental Sciences, 37007, Salamanca, Spain

ARTICLE INFO

Keywords:
Lentil
Physicochemical properties
Calibration
Protected geographical indication
NIRS
Discriminant analysis
Food analysis
Food composition

ABSTRACT

Calibration statistical descriptors for both whole and ground lentils using Near Infrared Spectroscopy (NIRS), combined with fiber-optic probe, are presented and discussed. The models were developed for estimating the weight, size, total raw protein, moisture, total fat, total fiber, and ash. Standard methods were used to determine compositional parameters of 42 samples of different varieties of lentils. The calibration curves show a wide range of validity for all parameters. The results showed excellent predictability for the determination of weight, fiber, and ash in whole lentils. However, size, moisture, and total fat were predicted satisfactorily in ground lentils. The total protein content could be predicted for both whole and ground lentils. Moreover, NIRS and Direct Partial Least Squares (DPLS) were used to determine whether a sample of lentils belonged to the Protected Geographical Indication (PGI) "Lenteja de La Armuña" or not. The results showed that 95% of the samples were correctly classified as belonging to a PGI. This result demonstrates that this technique allows the differentiation of samples from nearby regions.

1. Introduction

Pulses are currently arousing more interest in the field of the development of healthy and functional foods (Faris et al., 2012). Among them lentils make a significant contribution to the diet of people from many countries and are included as a staple food in the Mediterranean diet (Schneider, 2002). Lentils are low in fat and rich in proteins, complex hydrocarbons, and minerals (Campos-Vega et al., 2010; Gujral et al., 2011). Indeed, they have a lower glycaemic index than other starchy foods. Epidemiological and intervention studies have shown that legume consumption is inversely associated with the risk of coronary heart disease, type II diabetes mellitus, and obesity. Their consumption is associated with lower LDL cholesterol and higher HDL cholesterol (Rizkalla et al., 2002; Bazzano et al., 2011; Villegas et al., 2008; Pistollato and Battino, 2014). Lentils have been classified as soft seed-coated pulses that require a shorter cooking time. Because of this they have smaller nutrient losses than those with a hard seed coat (Satya et al., 2010). For these reasons, the incorporation of lentils into western diets has been highly recommended (Aguilera et al., 2010; Barbana and Boye, 2013).

The new trend in pulse consumption and its potential health benefits require a closer look at the composition of lentils. Food quality programs usually involve extensive evaluations of the quality components of interest. Owing to this, large numbers of screenings are usually performed by the standard analytical methods of seed lines. Although the standard analytical techniques usually provide a high level of accuracy and precision, they also have some disadvantages such as high costs, high labor input, and a delay in reporting. On the other hand, Near Infrared Spectroscopy (NIRS) is a non-destructive and rapid technique that does not require chemical reagents. It also saves a considerable amount of time by testing all parameters simultaneously. Due to the high penetration and scatter efficiency of NIRS, samples could be tested directly irrespective of the physical status of samples (Manley, 2014).

Abbreviations: Dt, de trend; DPLS, direct partial least squares; MPLS, modified partial squares; MSC, multiplicative scatter correction; NIRS, near infrared spectrometry; PCA, principal component analysis; PCR, principal component regression; PGI, protected geographical indication; PLS, partial least squares; RMSEC, root mean square error of calibration; RPD, ratio performance deviation; RSQ, multiple correlation coefficient SECV square error of cross-validation; SEP, square error of prediction; SEP (C), square error of prediciton corrected by BIAS SNV, standard normal variate; UPOV, union for the protection of new varieties of plants

* Corresponding author.
E-mail address: irevilla@usal.es (I. Revilla).

https://doi.org/10.1016/j.jfca.2019.01.012
Received 10 April 2018; Received in revised form 5 December 2018; Accepted 18 January 2019
Available online 23 January 2019
0889-1575/

Although NIRS has been used for calibration purposes in a wide range of food products, legumes have attracted less attention in general (Lee and Choung, 2009; Pojić et al., 2010; Szigedi et al., 2013; Kamboj et al., 2017). Regarding lentils, the NIRS technique has been used only to determine protein (Lazzeri et al., 1990) and color (Black and Panozzo, 2004). Indeed, NIRS in combination with chemometrics has been successfully used to distinguish samples depending on the geographical origin of diverse food products such as virgin olive oil (Casale et al., 2010), cheese (Karoui et al., 2005), wine (Liu et al., 2006), honey (Woodcock et al., 2009), and wheat (Zhao et al., 2013). In the case of lentils, Diffuse Reflectance Fourier Transform Infrared Spectroscopy has been applied to determine whether they are produced in Greece or imported from other countries (Kouvoutsakis et al., 2014).

European Council Regulation (EC) 2081/92 establishes that foodstuffs which are produced, processed and prepared in a given geographical area with a specific quality, reputation, or other characteristics attributable to that geographical origin are distinguished by a Protected Geographical Indication (PGI). The "Lenteja de La Armuña" PGI originates from the province of Salamanca in northwest Spain and includes the variety generally known as Rubia de La Armuña. “La Armuña” lentils are normally planted in autumn, preferably in October, and are harvested in late June or early July. “La Armuña” lentils are light green, sometimes with dark marbling. The seeds are notably flat and large and may attain a diameter of 9 mm in diameter; they are irregular in shape. Lentils from” La Armuña” are known for their nutritional value and have a high protein, fiber, calcium and iron content and a very characteristic sensory profile (Dueñas et al., 2016). Once baked the grains remain whole, the tegument is fine, the texture of the albumen is soft and consistent, and the lentils have an intense nutty flavor.

These PGI products need to be clearly authenticated and differentiated from very similar products that are often obtained from areas very close to “La Armuña”. As far as we know, no studies have been carried out on Protected Geographical Indication discrimination using this technique.

Taking into account the above, the aim of this study was to develop a rapid method to quantify the main physical properties (size and weight) and chemical characteristics of lentils. Moreover, this study aims to investigate whether NIRS combined with chemometrics can distinguish lentils of the “Lenteja de La Armuña” PGI from samples of different geographical origin.

2. Materials and methods

2.1. Samples

For calibration purposes a total of 42 dried lentil seed samples were used. 40 samples were of the macrosperma type, of which 39 samples were of the Guareña cultivar (Rubia de La Armuña ecotype) and 1 sample of the Rubia Castellana cultivar, and 2 samples were of the microsperma type (Pardina cultivar). All of them were harvested the same year.

For discriminant purposes a total of 37 dried lentil seed samples were used, 20 from the “Lenteja de La Armuña” Protected Geographical Indication (PGI) (EU 1151/2012) and 17 samples not belonging to the PGI from other Spanish regions. The cultivars and geographical origin of these samples are summarized in Table 1.

2.2. Physicochemical analysis

Seed weight was calculated as the average weight of 100 seeds and size was determined as the average diameter of three seeds chosen at random (UPOV). Moisture content was determined by oven-drying in accordance with the ISO standard (ISO, 1973), nitrogen content was determined by the Kjeldahl analysis, and a factor of 6.25 was used to estimate the protein content (AOAC, 1990). Total fat was analyzed using the Soxhlet method according to standard AOAC (1990) procedures, the total fiber content was determined according to the method proposed by Goering and Van Soest (1970) using ANKOM apparatus (Ankom Technology, Macedone, NY), and ash was analyzed using the gravimetric method according to ISO methodology (ISO, 2002). All the determinations were performed in triplicate.

2.3. NIR spectroscopy

A Foss NIRSystem 5000 (Foss A/S, Hillerød, Denmark) was used with a standard 1.5 m 210/210 bundle fiber-optic probe (ref. nº R6539-A). The spectral range was set at 1100–2000 nm since above this value (2000 nm) significant attenuation of the signal occurred due to strong absorption of the OH groups present in the fiber optic. The probe used a remote reflectance system and a ceramic plate as a reference. The window was of quartz with a 5 cm × 5 cm surface area. The NIR spectrum was obtained for each of the samples by applying the remote reflectance fiber-optic probe directly to the whole lentil sample or to the sample ground up in a hammer mill. The spectra were recorded at 2 nm intervals, and 32 scans were taken for both the reference and the samples. All samples were analyzed in triplicate in order to minimize sampling error.

2.4. NIR-chemometric methods

The models of calibration were developed by using the data obtained from the analytical determinations and the spectral data obtained from NIR spectra of the 42 samples assessed. Initially a Principal Component Analysis (PCA) was used. Once the number of principal components had been determined, the spectral anomalies were analyzed using the Mahalanobis distance (H statistic) method, establishing a value of H = 3 as the upper limit. Samples with an H value of above 3.0 were accordingly considered different from the spectral population and were removed from the set. Another criterion to be considered is the T criterion which is based on chemical parameters (the differences between the laboratory value and the NIRS-predicted value). Samples with a T value exceeding 2.5 were removed from the set because they were different from the population. The quantification of the different analytical parameters was performed using the modified partial squares (MPLS) regression method. Partial least squares (PLS) regression is similar to principal component regression (PCR) but uses both reference data (chemical, physical, etc.) and spectral information to form the factors useful for suitable purposes (Martens and Naes, 2001). MPLS is often more stable and accurate than the standard PLS algorithm. In MPLS, the NIR residuals obtained after each factor and at each wavelength were calculated and standardized (dividing them by the standard deviations of the residuals at each wavelength) before the next factor was calculated. Taking into account both H and T criteria, the calibration process was performed with the spectra of the resulting samples and their chemical data.

Scattering effects were removed using multiplicative scatter correction (MSC), standard normal variate (SNV), DeTrend (DT), or SNV–DT. Moreover, the mathematical treatments were tested in the development of the NIRS calibrations by using a nomenclature of 2,4,4,1 in which the first digit is the number of the derivative, the second is the gap over which the derivative is calculated, the third is the number of data points in a running average or smoothing, and the fourth is the second smoothing. When developing the MPLS equations, cross-validation is recommended in order to select the optimal number of factors and to avoid overfitting (Shenk and Westerhaus, 1995) and the calibration set is divided into several groups for cross-validation. Each group is then validated using calibration based on the other samples. Finally, any validation errors generated are combined into a root mean square error of cross-validation SECV (Davies and Williams, 1996). The prediction capacity of the model was assessed using the ratio performance deviation (RPD) parameter, which can be defined as

Table 1
Lentils samples used for discriminant analysis.

	Cultivar	Geographical origin	Samples
PGI	Guareña (Rubia de la Armuña ecotype)	Monterrubio de la Armuña, Salamanca (Castilla-León)	1
		Tardáguila Salamanca (Castilla-León)	2
		Villaverde de Guareña, Salamanca (Castilla-León)	2
		Villares de la Reina, Salamanca (Castilla-León)	2
		Palencia de Negrilla, Salamanca (Castilla-León)	2
		Pedrosillo el Ralo, Salamanca (Castilla-León)	1
		La Vellés, Salamanca (Castilla-León)	2
		Castellanos de Moriscos, Salamanca (Castilla-León)	2
		Pajares de la Laguna, Salamanca (Castilla-León)	2
		Negrilla de Palencia, Salamanca (Castilla-León)	1
		Cabrerizos, Salamanca (Castilla-León)	1
		Villaverde de Guareña, Salamanca (Castilla-León)	1
		Cabezabellosa de la Calzada, Salamanca (Castilla-León)	1
Not PGI	Rubia Castellana	Cuenca (Castilla-La Mancha)	2
	Rubia Castellana	Albacete (Castilla-La Mancha)	2
	Rubia Castellana	Salamanca (Castilla y León)	1
	Pardina	Valladolid (Castilla y León)	3
	Pardina	Cuenca (Castilla La Mancha)	2
	Guareña	Albacete (Castilla la Mancha)	1
	Guareña	Guadalajara (Castilla la Mancha)	1
	Guareña	Valladolid (Castilla y León)	1
	Guareña	Burgos (Castilla y León)	1
	Crimson Red	Albacete (Castilla la Mancha)	1
	Du-Puy	Albacete (Castilla la Mancha)	1
	Eston	Guadalajara (Castilla la Mancha)	1

PGI. Protected Geographical Indication.

Table 2
Data of physicochemical parameters of lentils samples determined by reference methods.

	Calibration set (42 samples)			
	Minimum	Maximum	Mean	SD
Weight (gr)	3.23	8.24	6.61	0.97
Size (mm)	4.30	7.63	6.65	0.67
Protein (%)	19.9	26.8	24.1	1.8
Moisture (%)	4.98	9.07	6.39	0.91
Total fat (%)	0.52	1.06	0.67	0.11
Total fiber (%)	5.63	7.81	6.52	0.48
Ash (%)	1.98	2.94	2.31	0.21

SD: Standard Deviation.

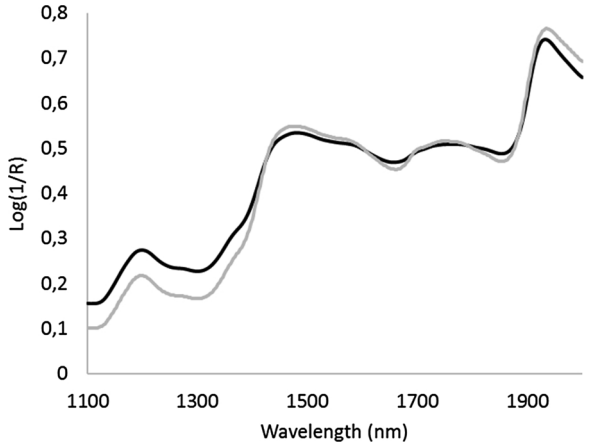

Fig. 1. Mean spectra of whole (black) and ground (grey) lentils (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

the relationship between the standard deviation of the chemical method (SD ref.) and that of the root mean square error of prediction (SEP) in the NIR model (Williams and Sobering, 1996). It is desirable for this ratio to be higher than 2 for satisfactory calibration (Monazen et al., 2005). An RPD ratio of less than 1.5 indicates poor predictions and the model cannot therefore be used to make further predictions. The statistics used to select the best equations were R^2, the determination coefficient, and the RMSEC (the root mean square error of calibration).

A discriminant analysis of the geographical origin of samples belonging to the “La Armuña” PGI and samples not belonging to the PGI was carried out. The calibration method applied to this procedure was D-PLS (Ståhle and Wold, 1987). Essentially the PLS method aims to concentrate the relevant information contained in the variables measured in a lower number of variables with no loss of relevant information. Regression is carried out with these new variables, simplifying the calibration model and the interpretation of the results. PLS 1 or PLS 2 should be used depending on whether the model is applied to a single variable or to several dependent variables (Brereton, 2003; Vandeginste et al., 1998). As during the development of the model the group to which the samples belong is known, files are prepared to contain the spectra of all samples belonging to the same category; there is one file for each category. This leads to the automatic generation of a temporal matrix with samples from all categories and as many new dummy variables as categories. In each variable there is an indication of whether the sample belongs to a given group or not by means of a binary code of ones or zeros respectively. The PLS 2 regression is performed with the matrix of these new dummy variables and the spectral data of the samples. Cross-validation is used to establish the number of PLS factors and assess the model. Once the latter has been obtained, a prediction is made as to the value of each dummy variable for each sample; these predicted values are then changed by the addition of one unit. The number of samples correctly classified in the categories to which they belong indicates the acceptability of the model developed. Some studies have used NIRS with this technique (Xie et al., 2007; Fernandez-Ibañez et al., 2009). The models obtained in each case were validated. The software used for the chemometric treatments was WinISI version 1.50 programme (Foss A/S, Hillerød, Denmark).

Table 3
Statistical descriptors of calibration by NIR of the phisycochemical characteristics of whole and crushed lentils.

	Recording mode	Math treatment	N	Mean	SD	Est. Min	Est. Max	SEC	R^2	SECV	RPD
Weight (gr)	Whole	Detrend only 2,4,4,1	39	6.79	0.66	4.80	8.78	0.20	0.91	0.35	5.4
Size (mm)	Whole	Detrend only 0,0,1,1	40	6.66	0.64	4.73	8.58	0.38	0.64	0.55	1.9
Protein (%)	Whole	Detrend only 1,4,4,1	38	24.0	1.78	18.7	29.4	0.36	0.96	0.33	5.4
Moisture (%)	Whole	Standard MSC 2,8,6,1	37	6.31	0.81	3.89	8.73	0.26	0.89	0.31	3.1
Total fat (%)	Whole	Standard MSC 0,0,1,1	39	0.66	0.09	0.39	0.92	0.06	0.48	0.09	1.8
Total fiber (%)	Whole	Standard MSC 1,4,4,1	38	6.57	0.46	5.19	7.95	0.25	0.61	0.29	2.0
Ash %)	Whole	Standard MSC 2,4,4,1	39	2.29	0.19	1.72	2.85	0.06	0.91	0.18	4.2
Weight (gr)	Ground	None 2,8,6,1	37	6.72	0.66	4.73	8.70	0.49	0.44	0.56	1.9
Size (mm)	Ground	Standard MSC 0,0,1,1	36	6.80	0.44	5.47	8.13	0.23	0.72	0.32	2.8
Protein (%)	Ground	SNV only 1,4,4,1	39	24.1	1.77	18.8	29.4	0.33	0.97	0.30	5.9
Moisture (%)	Ground	Standard MSC 2,10,10,1	38	6.42	0.93	3.64	9.20	0.22	0.94	0.24	4.3
Total fat (%)	Ground	SNV only 2,4,4,1	39	0.65	0.09	0.39	0.92	0.02	0.93	0.08	3.7
Total fiber (%)	Ground	Standard MSC 2,8,6,1	38	6.49	0.38	5.36	7.62	0.21	0.69	0.39	1.6
Ash %)	Ground	None 2,4,4,1	41	2.31	0.22	1.67	2.96	0.14	0.58	0.21	1.6

N, number of samples; SNV, standard normal variate; MSC, multiplicative scatter correction; SD, standard deviation; R^2, determination coefficient; SEC, Standard Error of Calibration SECV, Standard Error of Cross-Validation, Est. Min- Est Max: Minimum and Maximum value estimated by the model developed.

3. Results and discussion

3.1. Chemical analyses

Table 2 shows the minimum and maximum, the mean concentration, and the standard deviations of the physicochemical parameters determined by the reference methods in the 42 samples of lentils. The reference values for lentils of "Lenteja de la Armuña PGI" (Ministry of Agriculture, 1992) are 16,6% for moisture, 55% for total carbohydrates, 26.3% for protein, 0.8% for fat and 4.7% for total fiber.

The weight and size values reflect a wide range of variation between the minimum and the maximum. This is because the samples included macrosperma varieties such as "Rubia de La Armuña" with a weight and size varying between 5–8 grams and 3–6 mm respectively, and microsperma varieties such as Pardina, the lentils of which weigh less than 3 g.

The protein content interval found for the samples analyzed (19.9–26.8 %) was within the interval previously reported in the bibliography, which varies from the 20.6 to 31.4 (Lazzeri et al., 1990; Urbano, 2007). The mean content was slightly lower than those found by these authors although it coincided with the mean values found in other studies (Faris et al., 2012; Moldovan et al., 2015; Dueñas et al., 2016). Large differences in protein content were noted among samples of the same variety. This suggested that raw protein content could be used within a variety as an indicator of a general "environmental" effect (Wang and Daun, 2006). In other agriproducts such as nuts, protein content has been reported to be also affected by ecological conditions (Kodad et al., 2016; Rabadán et al., 2018). This is because protein content seems to be particularly sensitive to environmental stress such as rainfall, light intensity, the length of the growing season, daylight hours, and temperature in addition to agronomic factors such as plant density, weeds, and soil fertility (Singh et al., 1972; McLean et al., 1974; Delouche, 1980).

As far as moisture is concerned, the maximum value fell within the interval previously reported for lentils (6.53–10.6 %) (Faris et al., 2012; Moldovan et al., 2015; Moreiras et al., 2013) but the minimum value was slightly lower. Lentils have a relatively low fat content as can be seen from Table 2. Several researchers have found that lentil seeds had a total fat content of approximately 0.7–2.4 g/100 g (Ryan et al., 2007; Moldovan et al., 2015; Moreiras et al., 2013). The results of this work are within the lower limit of the reported range as previously observed for Rubia de la Armuña variety (Dueñas et al., 2016). The average levels of total fiber (6.52%) coincide with those found in green and red varieties of lentils (Moldovan et al., 2015) and the maximum and minimum levels are within the range described in other studies (Bhatty, 1988; Moreiras et al., 2013). The total ash content of the samples analyzed (1.98–2.94 %) was consistent with the results obtained by other researchers (Wang et al., 2009; Faris et al., 2012) but slightly lower than that found by Moldovan et al., (2015) for green lentils.

The reason why data are not always comparable is due to differences in genotypes, environments, methods of analysis, and in many cases the analysis of a single sample (Bhatty, 1988).

3.2. Spectral characteristics and NIRS calibration development

NIRS calibration is carried out on samples of ground and whole lentils in order to confirm on the one hand the possibility of predicting rapidly the parameters of interest of this pulse and on the other hand to find out for which physical status (ground or whole) the calibration models are most effective.

To predict the chemical composition of lentil seed samples using NIR technology calibrations with chemometric techniques are required. In order to do so it is necessary to have a large set of representative samples, to register the spectra, and to analyze the samples using a reference method. The mean spectral curves of the registered samples according to the physical status (whole or ground) are shown in Fig. 1. At a glance, the spectra show noticeable differences in the wavelengths at which they absorb water molecules (1490 and 1920 nm). This depends on the physical status as the absorbance values of these wavelengths are higher in ground samples.

A principal component analysis was carried out with the samples randomly selected in the calibration. The spectral variability explained exceeded 98%; between 6 and 10 principal components were required for each component studied when the calibrations were carried out on whole lentils and between 4 and 7 principal components were necessary in the case of ground lentils. Calibrations were performed by modified partial least squares regression (MPLS). The statistical parameters (R^2, SEC, SECV and RPD) of the calibration generated by the MPLS models show the quantification of the majority of the physicochemical constituents analyzed in the lentil (Table 3). In this table the calibration of ground and whole lentils is given separately; N indicates the number of samples that have been used in the development of the model after the elimination of samples owing to spectral criteria (H) or chemical criteria (T). Owing to the H criterion therefore two samples were eliminated for size, moisture, protein, and ash, one sample for fiber, and none for weight. In the case of ground samples one sample was eliminated for all calibrated parameters. This sample was not the same in all parameters. For their part, in accordance with criterion T three samples were eliminated for weight, moisture, and fiber, two samples for protein, one for fat, and none for size in the case of whole samples, while for ground lentils four samples were eliminated for weight, five for size, three for moisture and fiber, two for fat, one for

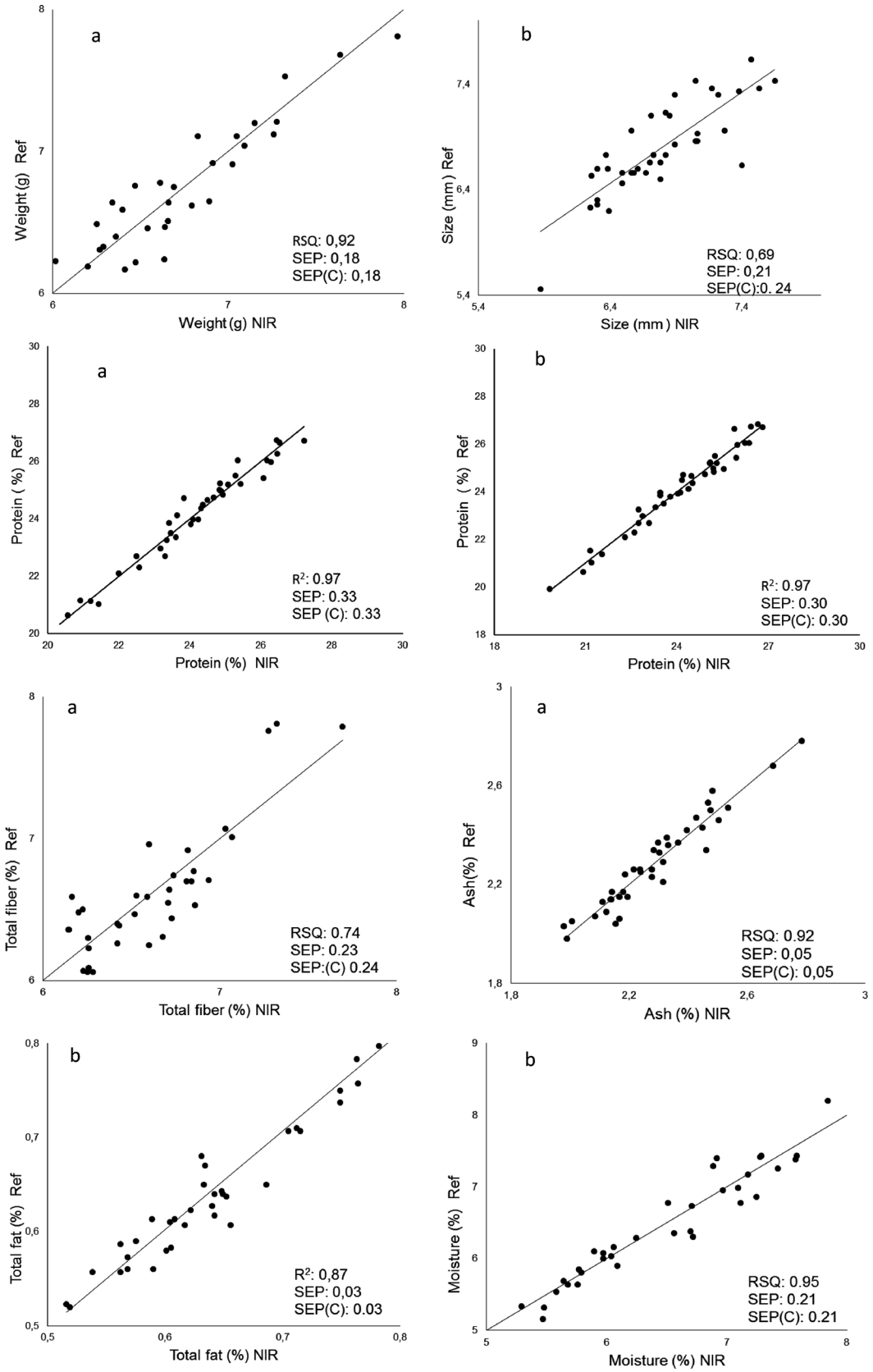

Fig. 2. Comparison of reference values with predicted values by NIRS model for each physicochemical parameter analyzed according to the recording mode of the samples: whole lentils (a) and ground lentils (b). (RSQ, multiple correlation coefficient, SEP, square error of prediction, SEP(C), standard error of prediction corrected by BIAS).

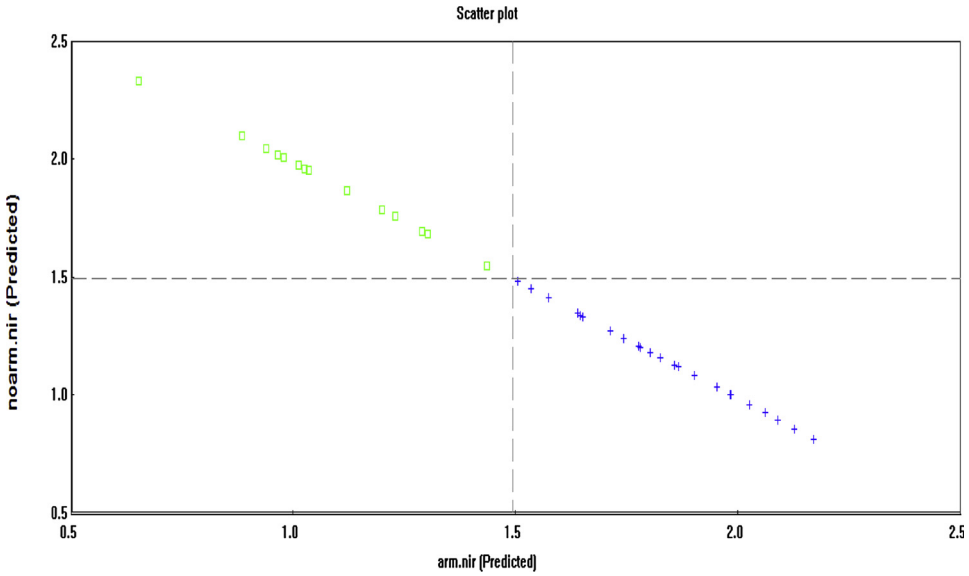

Fig. 3. Discriminant analysis (D-PLS) of NIR spectra of PGI de la Armuña lentils (□) and not PGI Armuña lentils (+).

protein, and none for ash.

It can be observed that with similar calibration margins the calibration models obtained in the case of whole lentils present better statistical descriptors for weight, total fiber, and ash. With ground lentils however better results are obtained in the calibration of the size, moisture, and total fat, which we believe to be a consequence of the correlations with the water molecules (moisture) and the C═O (fat) bonds of these parameters. Prediction of physical parameters (weight and size) are relevant because lentils from "La Aramuña" PGI are characterised in particular by these parameters. The regression model used can be justified by means of the correlation between the value of the parameter determined and the various wavelengths, given by the values of the β coefficients. These β coefficients are obtained by calculating the parameters of the equation $y = \beta_o + \beta_1 X_{\lambda_1} + \beta_2 X_{\lambda_2} + \beta_3 X_{\lambda_3} + \ldots \beta_n X_{\lambda_n}$.In which ß$_0$, ß$_1$, ß$_2$ … are the coefficients and X_{λ_1}, X_{λ_2}, X_{λ_3} are the wavelengths in which the correlation with the value of the variables (in this case for the weight and size) shows maximum reflectance. The β coefficients were very high for size and weight at the most significant wavelengths, for example β = 25128 at 1530 nm for weight and β = 78588 at 1735 nm for size.

Proteins deserve a special mention as they can be calibrated in both whole and ground lentils as previously observed in peas (Lee and Choung, 2009). This is highly significant as proteins are among the most valuable components of these pulses, which makes them an important source of protein. Good calibrations are reflected in the high prediction capacity values (RPD) and in the R^2 values. Despite the poor R^2 results obtained for fat in whole samples (0.48) and for weight in ground samples (0.44), the data of these calibration models have been maintained in Table 3 merely so as to compare the two forms of presentation, as these values are unacceptable in an NIR model.

3.3. Validation. Prediction capacity of the models

Cross-validation was used to evaluate the models obtained. In this method the set of calibration samples was divided into a series of subsets for them to be calibrated. Six series were established of which 5 were taken for the calibration set and one was taken for the prediction set. The process was repeated as many times as there were sets so that all pass through the calibration set and the prediction set. The models used were thus validated and their prediction capacities were confirmed. Fig. 2 shows the correlation of the values obtained in the laboratory (Ref.) with those values predicted by NIR with a remote reflectance fiber-optic probe for the physicochemical components studied in the lentils, comparing the prediction graphs for whole lentils (a) and for ground lentils (b). The results showed excellent predictability for the determination of the weight, the total fiber, and the ash in whole lentils and the size, moisture, and total fat in ground lentils. The total protein content could be predicted for either whole or ground lentils. This confirms that the models established by the NIR technology for these physicochemical constituents can be applied to unknown samples.

3.4. Discriminant analysis

In order to differentiate between samples from the "Lenteja de La Armuña" PGI (lentils from the region of La Armuña in northwest Spain) and those collected from other Spanish regions a discriminant analysis was performed. NIR spectral data of whole lentils of 20 samples from the "La Armuña" PGI and 17 samples not belonging to this PGI were used with no type of mathematical treatment. The DPLS models were then constructed with 37 samples and 4 principal components.

The DPLS model correctly classified 95% of the PGI samples and 71% of the non PGI samples, with RSQ values of 0.69 and SECV of 0.40. The results of modeling the classes are shown in Fig. 3, in which the "dummy variables" are represented on the axes. Spectral differences may be due to the clayey nature of the soil of La Armuña and its low-acidity pH (Gallardo et al., 1984), which is reflected in the NIR spectra of the lentils produced there.

Previous studies have shown that discriminant analysis together with DRIFTS (Diffuse Reflectance Fourier Transform Infrared Spectroscopy) allows the differentiation of lentil samples according to geographical origin, i,e. whether they are Greek or imported (Kouvoutsakis et al., 2014). However, given the results it may be concluded that the NIR spectral information and the DPLS discriminant analysis allows the differentiation of lentils not only according to the country of origin but also to production areas within the same country.

4. Conclusions

Calibration statistical descriptors of the models developed for estimating the weight, size, total raw protein, moisture, total fat, total fiber, and ash using Near Infrared Spectroscopy (NIRS) together with a fiber-optic probe showed excellent predictability for determining the weight, fiber, and ash of whole lentils and the size, moisture and total fat of ground lentils. It is noteworthy that the total protein content

could be predicted for both whole and ground lentils. Moreover, the calibration curves show a wide range of validity for all parameters. As far as geographical discrimination is concerned, the analysis developed using NIRS and DPLS correctly classified 95% of the PGI samples, which shows that this technique allows the differentiation of samples from regions very near to one another.

Funding

This work was financed by University of Salamanca (Project 18KBCN/463AC01).

Acknowledgment

The authors would like to give special thanks to the Lenteja de La Armuña PGI for its cooperation.

References

Aguilera, Y., Dueñas, M., Estrella, I., Hernández, T., Benitez, V., Esteban, R.M., Martín-Cabrejas, M.A., 2010. Evaluation of phenolic profile and antioxidant properties of Pardina lentil as affected by industrial dehydration. J. Agric. Food Chem. 58, 10101–10108.

Barbana, C., Boye, J.I., 2013. *In vitro* protein digestibility and physico-chemical properties of flours and protein concentrates from two varieties of lentil (*Lens culinaris*). Food Funct. 4 (2), 310–321.

Bazzano, L.A., Thompson, A.M., Tees, M.T., Nguyen, C.H., Winham, D.M., 2011. Non-soy legume consumption lowers cholesterol levels: a meta-analysis of randomized trials. Nutr. Metab. Cardiovasc. Dis. 21, 94–103.

Bhatty, R.S., 1988. Composition and quality of lentil (*Lens culinaris* Medik): a review. Can. Inst. Food Sci. Technol. J. 21, 144–160.

Black, C.K., Panozzo, J.F., 2004. Accurate technique for measuring color values of grain and grain products using a Visible-NIR instrument. Cereal Chem. 81, 469–474.

Brereton, R.G., 2003. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. J. Wiley, Chichester, West Sussex, England.

Campos-Vega, R., Loarca-Piña, G., Oomah, B.D., 2010. Minor components of pulses and their potential impact on human health. Food Res. Int. 43, 461–482.

Casale, M., Casolino, C., Oliveri, P., Forina, M., 2010. The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil. Food Chem. 118, 163–170.

Davies, A.M.C., Williams, P.E., 1996. Near Infrared Spectroscopy: the Future Waves Proceeding of the Seventh International Conference on Near Infrared Spectroscopy. NIR Publication, Chichester, West Sussex.

Delouche, J.C., 1980. Environmental effects on seed development and seed quality. Hortic. Sci. 15 (6), 13–18.

Dueñas, M., Sarmento, T., Aguilera, Y., Benitez, V., Mollá, E., Esteban, R.M., Martín-Cabrejas, M.A., 2016.). Impact of cooking and germination on phenolic composition and dietary fibre fractions in dark beans (*Phaseolus vulgaris* L.) and lentils (*Lens culinaris* L.). LWT-Food Sci. Technol. 66, 72–78.

Faris, M.A.I.E., Takruri, H.R., Issa, A.Y., 2012. Role of lentils (*Lens culinaris* L.) in human health and nutrition: a review. Med. J. Nutr. Metab. 6, 3–16.

Fernandez-Ibañez, V., Soldado, A., Martinez-Fernandez, A., de la Roza-Delgado, B., 2009. Application of near infrared spectroscopy for rapid detection of aflatoxin B1 in maize and barley as analytical quality assessment. Food Chem. 113 (2), 629–634.

Gallardo, J.F., Moyano, A., Prat, L., Escudero, F., 1984. La fertilidad de los suelos de la Armuña, Tierra de Alba y Campo de Peñaranda. Consejo Superior de Investigaciones Científicas, Salamanca, Spain.

Goering, H.K., Van Soest, P.J., 1970. Forage Fiber Analyses (apparatus, Reagents, Procedures, and Some Application). USDA: Agricultural Research Service.

Gujral, H.S., Angurala, M., Sharma, P., Singh, J., 2011. Phenolic content and antioxidant activity of germinated and cooked pulses. Int. J. Food Prop. 14, 1366–1374.

Kamboj, U., Guha, P., Mishra, S., 2017. Characterization of chickpea flour by near Infrared Spectroscopy and chemometrics. Anal. Lett. 50, 1754–1766.

Karoui, R., Dufour, E., Pillonel, L., Schaller, E., Picque, D., Cattenoz, T., Bosset, J.O., 2005. The potential of combined infrared and fluorescence spectroscopies as a method of determination of the geographic origin of Emmental cheeses. Int. Dairy J. 15, 287–298.

Kodad, O., Estopañán, G., Juan, T., Sociasi-Company, R., Sindic, M., 2016. Genotype and year variability of the chemical composition of walnut oil of Moroccan seedlings from the high Atlas Mountains. Grasas Aceites 67 (1) e116.

Kouvoutsakis, G., Mitsi, C., Tarantilis, P.A., Polissiou, M.G., Pappas, C.S., 2014. Geographical differentiation of dried lentil seed (*Lens culinaris*) samples using diffuse reflectance fourier transform infrared spectroscopy (DRIFTS) and discriminant analysis. Food Chem. 145, 1011–1014.

Lazzeri, L., Leoni, O., Palmieri, S., 1990. Near infrared reflectance analysis for determining the total protein content of chick-pea, lentil and pea seeds. Agric. Mediterranea 120, 103–109.

Lee, J.H., Choung, M., 2009. Determination of protein content in pea by near infrared spectroscopy. Food Sci. Biotechnol. 18, 60–65.

Liu, L., Cozzolino, D., Cynkar, W.U., Gishen, M., Colby, B., 2006. Geographic classification of Spanish and Australian Tempranillo red wines by visible and near-infrared spectroscopy combined with multivariate analysis. J. Agric. Food Chem. 54, 6754–6759.

Manley, M., 2014. Near-Infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. Chem. Soc. Rev. 43, 8200–8214.

Martens, H., Naes, T., 2001. Multivariate Calibration. John Wiley and Sons, Ltd, Chichester, New York, USA.

McLean, L.A., Sosulski, F.W., Youngs, C.G., 1974. Effects of nitrogen and moisture on yield and protein in field peas. Can. J. Plant Sci. 54, 301–305.

Moldovan, O., Păucean, A., Vlaic, R., Bors, M.D., Mute, S., 2015. Preliminary assessment of the nutritional quality of two types of lentils (*Lens culinaris*) by near Infrared Reflectance Spectroscopy Technology (Nirs). Bull. UASVM Food Sci. Technol. 72, 127–132.

Monazen, M., De Baedemaeker, J., Ramon, H., 2005. Towards development of on-line soil moisture content sensor using a fiber-type NIR spectrophotometer. Soil Tillage Res. 80, 171–183.

Moreiras, O., Carbajal, A., Cabrera, L., Cuadrado, C., 2013. Tablas de composición de alimentos. Guía de prácticas (16º ed). Pirámide, Madrid, Spain.

Pistollato, F., Battino, M., 2014. Role of plant-based diets in the prevention and regression of metabolic syndrome and neurodegenerative diseases. Trends Food Sci. Technol. 97, 1093–1101.

Pojić, M., Mastilović, J., Palić, D., Petorić, M., 2010. The development of near-infrared spectroscopy (NIRS) calibration for prediction of ash content in legumes on the basis of two different reference methods. Food Chem. 123, 800–805.

Rabadán, A., Pardo, J.E., Pardo-Giménez, A., Álvarez-Ortí, M., 2018. Effect of genotype and crop year on the nutritional value of walnut virgin oil and defatted flour. Sci. Total Environ. 634, 1092–1099.

Rizkalla, S.W., Bellisle, F., Slama, G., 2002. Health benefits of low glycaemic index foods, such as pulses, in diabetic patients and healthy individuals. Br. J. Nutr. 88, 255–262.

Ryan, E., Galvin, K., O'Connor, T.P., Maguire, A.R., O'Brien, N.M., 2007. Phytosterol, squalene, tocopherol content and fatty acid profile of selected seeds, grains, and legumes. Plant Foods Hum. Nutr. 62, 85–91.

Satya, S., Kaushik, G., Naik, S.N., 2010. Processing of food legumes: a boon to human nutrition. Med. J. Nutrition Metab. 3, 183–195.

Schneider, A.V.C., 2002. Overview of the market and consumption of pulses in Europe. Br. J. Nutr. 88, 243–250.

Shenk, J.S., Westerhaus, M.O., 1995. Routine Operation, Calibration, Development and Network System Management Manual. Silver Spring, Maryland, USA NIR Systems.

Singh, B., Campbell, W.F., Salunkhe, D.K., 1972. Eff ;ects of s- triazines on protein and fine structure of cotyledons of bush beans. Am. J. Bot. 59, 568–572.

Ståhle, L., Wold, S., 1987. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. J. Chemom. 1 (3), 185–196.

Szigedi, T., Fodor, M., Pérez-Marin, D., Garrido-Varo, A., 2013. Fourier Transfrom Near-Infrared Spectroscopy to predict the gross energy content of food grade legumes. Food Anal. Methods 6, 1205–1211.

Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., 1998. Handbook of Chemometrics and Qualimetrics. Elsevier Science Inc., New York, NY, USA.

Villegas, R., Gao, Y.T., Yang, G., Li, H.-L., Elasy, T.A., Zheng, W., Shu, X.O., 2008. Legume and soy food intake and the incidence of type 2 diabetes in the Shanghai Women's Health Study. Am. J. Clin. Nutr. 87, 162–167.

Wang, N., Daun, J.K., 2006. Eff ;ects of variety and crude protein content on nutrients and anti-nutrients in lentils (*Lens culinaris*). Food Chem. 95 (3), 493–502.

Wang, N., Hatcher, D.W., Toews, R., Gawalko, E.J., 2009. Influence of cooking and dehuling on nutritional composition of several varieties of lentils (*Lens culinaris*). Lwt - Food Sci. Technol. 42, 842–848.

Williams, P.C., Sobering, D.C., 1996. How do we do it: a brief summary of the methods we use in developing near infrared calibrations. In: Davies, A.M.C., Williams, P.C. (Eds.), Near Infrared Spectroscopy: The Future Waves. NIR Publications., Chichester, NY, USA, pp. 185–188.

Woodcock, T., Downey, G., O'Donnell, C.P., 2009. Near infrared spectral fingerprinting for confirmation of claimed PDO provenance of honey. Food Chem. 114, 742–746.

Xie, L., Ying, Y., Ying, T., 2007. Combination and comparison of chemometrics methods for identification of transgenic tomatoes using visible and near infrared diffuse transmittance technique. J. Food Eng. 82 (3), 395–401.

Zhao, H., Guo, B., Wei, Y., Zhang, B., 2013. Near infrared reflectance spectroscopy for determination of the geographical origin of wheat. Food Chem. 138, 1902–1907.