applied sciences

MDPI

Article

E-Eye-Based Approach for Traceability and Annuality Compliance of Lentils

Martina Foschi [1], Valerio Di Maria [2], Angelo Antonio D'Archivio [1], Federico Marini [2,*] and Alessandra Biancolillo [1]

[1] Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio, 67100 Coppito, Italy
[2] Department of Chemistry, University of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185 Rome, Italy
* Correspondence: federico.marini@uniroma1.it

Abstract: In recent years, thanks to their numerous nutritional benefits, legumes have been rediscovered and have attracted interest from many consumers. However, these products, the most valuable ones traditionally produced in smaller communities in particular, can be objects of fraud; this is the case of Italian lentils, which, being a dry product, have a fairly long shelf life, but, due to the minimal visual changes that can affect them, it is possible that expired lentils may be sold alongside edible ones. The present work aims at creating a non-destructive method for classifying Italian lentils according to their harvest year and origin, and for discriminating between expired and edible ones. In order to achieve this goal, Red-Green-Blue (RGB) imaging, which could be considered as a sort of e-eye and represents a cutting-edge, rapid, and effective analytical method, was used in combination with a discriminant classifier (Sequential Preprocessing through ORThogonalization-Linear Discriminant Analysis, SPORT-LDA) to create novel testing models. The SPORT-LDA models built to discriminate the different geographical origins provided an average correct classification rate on the test set of about 88%, whereas an overall 90% accuracy was obtained (on the test samples) by the SPORT-LDA model built to recognize whether a sample was still within its expiry date or not.

Keywords: e-eye; lentils; ANOVA simultaneous component analysis (ASCA); sequential preprocessing through orthogonalization-linear discriminant analysis (SPORT-LDA); classification; traceability; harvesting year; image analysis

Citation: Foschi, M.; Di Maria, V.; D'Archivio, A.A.; Marini, F.; Biancolillo, A. E-Eye-Based Approach for Traceability and Annuality Compliance of Lentils. *Appl. Sci.* **2023**, *13*, 1433. https://doi.org/10.3390/app13031433

Academic Editor: Claudio Medana

Received: 26 December 2022
Revised: 17 January 2023
Accepted: 18 January 2023
Published: 21 January 2023

1. Introduction

Lentils (*Lens culinaris* Medik.) are a traditional and ancient staple food of the Mediterranean basin, representing an inexpensive and important source of plant-based protein [1–3]. The annual production is around 4.5 million tons, and they are majorly produced in Canada, the United States, Turkey, Australia, and India [3]. This pulse is a source of biologically active proteins such as lectins and protease inhibitors [3]. Additionally, lentils present a remarkable protein nitrogen content, despite their main components being carbohydrates. Interestingly, Johnson et al. have demonstrated that this legume is a relevant source of prebiotic carbohydrates [4]. Furthermore, lentils contain fibers, oligosaccharides, and mineral ingredients [3]. The high and valuable nutritional properties of this legume have been [2], and are still being, widely declaimed and promoted by the Food and Agriculture Organization of the United Nations (FAO) [5] and the scientific community. These qualities make lentils a sustainable and resilient crop that could also help address diet-related diseases, from malnutrition to obesity [6]. The great interest in this legume is demonstrated by the large amount of literature on it [4,7–11] and is resulting in an increase in consumption and production worldwide.

In Italy, numerous different cultivars of pulses are commonly cultivated. Zaccardelli et al. investigated several types of lentils grown in various Italian areas and demonstrated that these legumes present a large genetic variation [12]. This leads to the fact the different ecotypes (grown relatively close to one another) have very dissimilar chemical (and,

consequently, organoleptic) characteristics. Despite this richness in lentil cultivars, and although production has almost doubled in the last ten years, their cultivation has suffered a considerable decline and is now based on a multitude of traditional ecotypes well adapted to the severe conditions of marginal mountainous rural areas [12]. This circumstance makes Italian lentil production inefficient in meeting the internal demand, turning Italian landraces into niche products, valuable in terms of cost, quality, and uniqueness. In addition to this scenario, there are the effects that climate change could have on these typical varieties; although this research area is still unexplored, two different studies have confirmed the adverse effects that climate change (increased temperatures and reduced water availability) can have on the quantitative yield (size and number of seeds) and nutritional quality of different lentil varieties [13,14].

Therefore, this context, combined with reduced yields, could dramatically affect the economy of small producers and the entire production of high-value lentils, leading producers to commit illegal practices such as adulteration and counterfeiting. Accordingly, several studies in the literature have aimed at the control and authentication of Italian lentils against imported varieties by employing chemometrics and different analytical techniques such as Isotope Ratio Mass Spectrometry (IRMS) [15], Proton nuclear magnetic resonance (^{1}H-NMR) [16], Inductively Coupled Plasma-Optical Emissions Spectrometry (ICP-OES) [17], or Infrared Spectroscopy [18].

On the other hand, numerous studies have focused on the genetic wealth of Italian varieties by developing a method helpful in safeguarding the on-farming-survival landraces that are easily exposed to genetic erosion [19–21]. For example, Biancolillo et al. developed a non-destructive fingerprint approach, coupled with chemometrics, to verify autochthonous materials and prevent genetic contamination [22]. Torricelli and colleagues extensively characterized several Italian landraces, paying particular attention to the Abruzzo varieties traditionally selected by growers around Santo Stefano di Sessanio (a locality in the province of L'Aquila) [23]. These studies confirmed the existence of a genetic metapopulation of the Abruzzo landraces, which will be named in this study as the L'Aquila class, that is differentiated from the most known Italian lentil variety, i.e., the PGI lentil of Castelluccio di Norcia. These studies are of enormous importance in monitoring the diversity or similarity of autochthonous varieties, and, over time, they can provide information about crop resilience against climate change. Focusing on this last reported study, it is interesting to note the agreement of the image analysis performed by the authors with the experimental evidence of the morphological and genetic characterization, demonstrating the power of this tool as a preliminary approach for the authentication and characterization of local varieties belonging to the same species. To date, image analysis has been widely used to study and characterize lentils [24]. This has been accomplished using different methods, depending on the end purpose. Shahin et al. employed image analysis to obtain lentil seed size and shape, demonstrating that seed diameter, thickness, plumpness, and degree of edge roundness, coupled with multivariate linear regression can predict dehulling efficiency [25]. A flatbed scanner and an image processing program were used to obtain seed morphological characteristics. The method was also combined with the mean color information and a Linear Discriminant Analysis to identify five Sicilian landraces and three common Canadian accessions [26]. However, a mere colorimetric analysis has been shown to be efficient for quality control [27], allowing for distinguishing between the Abruzzo and Castelluccio di Norcia ecotypes [23], recognizing deteriorated lentils [28], and identifying adulterated lentil flour [29].

Thus, based on the evidence reported in the literature, Multivariate Image Analysis (MIA), which is a rapid non-destructive, and objective method, was applied as a preliminary stage of landraces authentication. In detail, we have applied MIA to distinguish the typical genetic traits of the studied ecotypes, namely L'Aquila, Castelluccio di Norcia, and Colfiorito. Due to the way the dataset was constructed (more details will be given in Section 2.1), the possibility of distinguishing expired lentils from those that are edible was also considered. The most advanced chemometric methods were applied to the colorgrams

obtained from MIA. In addition, ANOVA-Simultaneous Component Analysis (ASCA) was employed as an exploratory analysis to evaluate the significant factors that actually characterize the dataset. Sequential Preprocessing through ORThogonalization (SPORT) was the supervised pattern recognition method employed to discriminate between the Italian landraces and among expired and edible samples. Therefore, although several image analysis methods have been applied to lentil characterization, no examples are reported in the literature regarding the use of MIA coupled with such chemometric approaches.

2. Materials and Methods

2.1. Samples

Different lots of lentils were collected by local retailers or growers farming in three different areas of Central Italy: L'Aquila (Abruzzo), Colfiorito (Umbria), and Castelluccio di Norcia (Umbria). Samples were harvested at different time points, in particular, every year in July, from 2016 to 2021. Samples were stored in sealed plastic bags at room temperature in dark and dry conditions. Details on harvesting years and retailers are reported in Table 1.

Table 1. Origin, harvesting year, and the number of retailers of the investigated lentil samples.

	Harvesting Year and Number of Retailers					
Origin	**2021**	**2020**	**2019**	**2018**	**2017**	**2016**
L'Aquila (AQ)	0	0	0	2	3	0
Colfiorito (COL)	0	0	0	2	3	0
Castelluccio di Norcia (CDN)	2	6	3	4	4	1

2.2. E-Eye Analysis

Images were collected by arranging the lentils on a black plate, ensuring that their quantity allowed a homogeneous covering of the support. Pictures were taken by means of the RS Pro Wi-Fi USB Microscope (1280 × 1024 pixel resolution, magnification power from 10× to 160×, 36 mm diameter, 142 mm length, and light-emitting diode lighting) kept at the same constant distance from the sample plate. A total of 560 images, organized as described in Table 2, were collected.

Table 2. The number of investigated samples divided according to the origin and the harvesting year.

	Harvesting Year and Number of Samples					
Origin	**2021**	**2020**	**2019**	**2018**	**2017**	**2016**
L'Aquila (AQ)	0	0	0	42	84	0
Colfiorito (COL)	0	0	0	40	60	0
Castelluccio di Norcia (CDN)	40	85	55	55	79	20

2.3. ASCA

ANOVA-simultaneous component analysis (ASCA) was employed as an exploratory method to perform an ANalysis Of Variance (ANOVA), which is generally used to confirm the significance of factors on a single variable in a complex multivariate system [30]. The ASCA can be schematically explained by two steps [31]:

1. The mean-centered experimental data matrix $\boldsymbol{X}$ (N × V), where N is the number of analyzed samples and V is that of measured variables, is decomposed, following the ANOVA approach, into individual matrices accounting for the effects of each design term. For instance, if two factors are controlled in the design, here labeled as α and β, the ANOVA decomposition can be expressed as:

$$\boldsymbol{X} = \boldsymbol{X}_{\alpha} + \boldsymbol{X}_{\beta} + \boldsymbol{X}_{\alpha\beta} + \boldsymbol{X}_{E} \quad (1)$$

where $\boldsymbol{X}_\alpha$ and $\boldsymbol{X}_\beta$ are the arrays accounting for the main effect of the factors, while $\boldsymbol{X}_{\alpha\beta}$ describes the binary interaction among them, and $\boldsymbol{X}_E$ contains the residual variability, i.e., the portion of the experimental variance not approximated by the ANOVA model. Although for balanced designs this decomposition can be obtained from the simple "differences of means" procedure, a more general approach relies on expressing the ANOVA model as a linear regression problem, according to the following relation:

$$\boldsymbol{X} = \boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{X}_E \tag{2}$$

where $\boldsymbol{D}$ (N × P) is the design matrix that encodes the levels of the factors and their interaction (s) and consists of P columns, while $\boldsymbol{\theta}$ (P × V) is the matrix of regression coefficients. The peculiarity of both matrices $\boldsymbol{D}$ and $\boldsymbol{\theta}$ is that they are block-partitioned so that an individual linear equation can be written for each design term as follows [32]:

$$\boldsymbol{X}_i = \boldsymbol{D}_i\,\boldsymbol{\theta}_i = \boldsymbol{D}_i\left(\boldsymbol{D}_i^T\boldsymbol{D}_i\right)^{-1}\boldsymbol{D}_i^T\boldsymbol{X} \text{ with } i = \alpha, \beta, \alpha\beta \tag{3}$$

where it is apparent that each effect matrix is obtained by projecting the experimental data matrix onto the suitably coded design sub-matrix corresponding to the specific term.

(2) The second step consists of the Principal Component Analysis, performed separately on each effect matrix, thus:

$$\boldsymbol{X}_i = \boldsymbol{T}_i\,\boldsymbol{P}_i^T \text{ with } i = \alpha, \beta, \alpha\beta \tag{4}$$

where $\boldsymbol{T}_i$ and $\boldsymbol{P}_i$ are the scores and loadings matrices for each factor. The significance of the effects and related loadings was performed by using a permutation test and bootstrap procedure, respectively.

2.4. Sequential Preprocessing through ORThogonalization (SPORT)

Sequential Preprocessing through ORThogonalization (SPORT) is a regression approach conceived for ensemble preprocessing [33]. This method originates from another multi-block regression approach, called Sequential and Orthogonalized Partial Least Squares (SO-PLS) [34]. SPORT exploits the SO-PLS algorithm to extract information from the blocks subjected to the various pre-treatments, but, at the same time, avoids redundancies among them [35].

In the present case, due to the nature of the analyzed data, two pre-treatments were tested: mean-centering and autoscaling. This corresponds to building a SPORT model handling two different data blocks, one given by the colorgram matrix preprocessed by mean-centering ($\boldsymbol{X}_1$) and one obtained by autoscaling the original data matrix ($\boldsymbol{X}_2$). Given these circumstances, the SPORT algorithm can be summarized by the following steps:

(1) Y is fitted to $\boldsymbol{X}_1$ by PLS.
(2) $\boldsymbol{X}_2$ is orthogonalized with respect to the X_1-scores estimated at step (1), resulting in $\boldsymbol{X}_{2,Orth}$.
(3) A second PLS regression model is calculated between $\boldsymbol{X}_{2,orth}$ and the Y-residuals from step 1.
(4) The predicted $\boldsymbol{Y}$ ($\hat{\boldsymbol{Y}}$) is given by the combination of regressions at steps (1) and (3):

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}_1\boldsymbol{B} + \boldsymbol{X}_2\boldsymbol{C} \tag{5}$$

where $\boldsymbol{B}$ and $\boldsymbol{C}$ are the regression coefficients matrices.

In the case where SPORT is used as a classification strategy (as in the present work), the response $\boldsymbol{Y}$ referred to in step (1) is the so-called dummy Y [36], which binarily encodes the class-membership of the samples. However, $\hat{\boldsymbol{Y}}$ is not binary-coded as its target values as its elements are all real-valued [37]. Classification can then be achieved by different strategies; in the present study, it was accomplished by applying linear discriminant analysis on

the predicted responses $\hat{\boldsymbol{Y}}$ [38]. Once the model is calibrated on training individuals and the regression coefficient matrices are estimated, new objects can be classified by solving Equation (6):

$$\hat{\boldsymbol{Y}}_{new} = \boldsymbol{X}_{1,new}\boldsymbol{B} + \boldsymbol{X}_{2new}\boldsymbol{C} \quad (6)$$

where $\boldsymbol{X}_{1,new}$ and $\boldsymbol{X}_{2new}$ are the data matrices associated with new observations.

3. Results and Discussion

3.1. ASCA for Exploring the Significance of the Harvesting Year and the Origin

Once all the images were collected, colorgrams were obtained using in-house MATLAB (R2015b; The Mathworks, Natick, MA, USA) functions; examples of the collected images and colorgrams are reported in Appendix A (Figures A1 and A2, respectively).

The ASCA was used to inspect the significance of two factors: the origin and harvesting year. To perform the ASCA, we decided to work on the fraction of the dataset that could provide a crossed-factor design with all the cells populated; therefore, only the samples of the three varieties produced in the years 2017 and 2018 were considered. Moreover, in order to have a balanced design, 40 samples per each combination of year and origin were selected out of the available number reported in Table 2, using the Kennard–Stone algorithm [39] on the colorgram data. This decision stems from the fact that the ASCA does not allow the independent interpretation of effect matrices for unbalanced experimental designs, as in our case [40]. The significance has been evaluated by the permutation test (104 permutations), and both effects and their interaction appeared to be significant.

The inspection of samples projected onto the space spanned by the first SCs associated with the model of the original effect (Figure 1) reveals a clear grouping trend of individuals according to their harvesting area (L'Aquila-AQ, Colfiorito-COL, Castelluccio di Norcia-CDN). In fact, CDN samples (blue diamonds) fall at negative values of SC1 whereas COL (green squares) and AQ objects (red dots) present positive values of this component. Samples harvested in AQ seem divided into two subgroups: one falling at negative SC2 scores and one at slightly negative values of this component, overlapping with CDN samples. This similarity can certainly be attributed to the proximity among the diverse areas and to similar pedoclimatic conditions. However, it must be stressed that the class AQ consists of different ecotypes cultivated in the vast province of L'Aquila. This condition results in a class with a higher variance than the others considered.

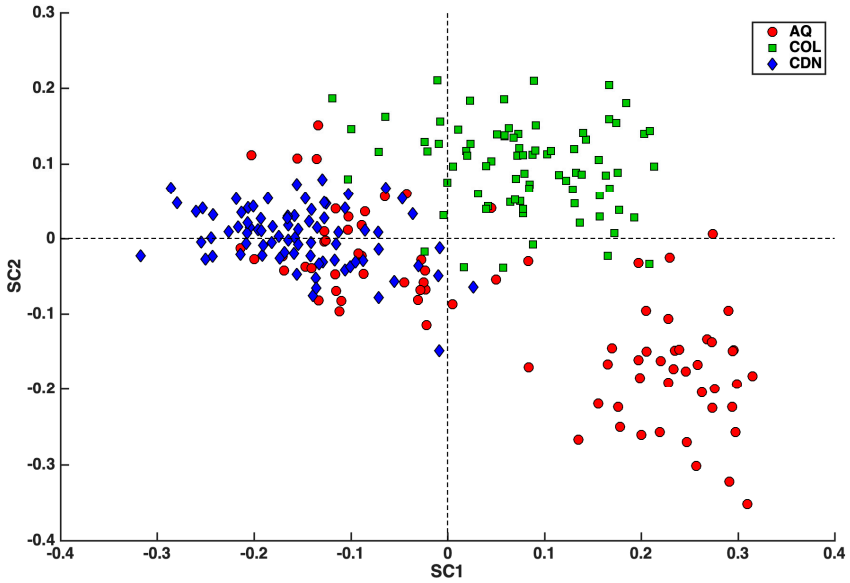

Figure 1. ASCAanalysis: SCA model of the effect of the *geographical origin* factor. Sample scores are plotted onto the two SCs of the model after the projection of the residual matrix to show within-level variability. Legend: Red dots—AQ; Green Squares—COL; Blue diamonds—CDN.

Inspection of the loadings for the SC model, reported in Appendix B (Figure A3), indicated that the CDN samples are characterized, in general, by a higher amount of pixels with higher intensity of the red channel and a lower intensity of the green and blue channels, and lightness. At the same time, the CDN samples have a higher amount of pixels with higher hue, saturation, and intensity values than the lentils from the other two origins.

On the other hand, by looking at the loadings on SC2, along which COL and AQ are differentiated, it is possible to affirm that the images from COL samples are mostly characterized by a lower intensity of the red channel and intermediate intensity of the blue one, the lightness also being lower. With respect to AQ, they are also characterized by slightly higher hue, intermediate saturation, and lower intensity.

The inspection of the SC1 associated with the year effect factor unveils a clear trend related to annuality (Figure 2). Indeed, the 2018 production year is characterized by predominantly positive scores, whereas 2017 is mostly described by negative values. The corresponding loadings (Appendix B, Figure A4), which were found to be statistically different from zero only for a few variables, indicate that the 2018 samples are characterized by a higher amount of pixels with an intermediate intensity of the red, green, and, to a lower extent, blue channels.

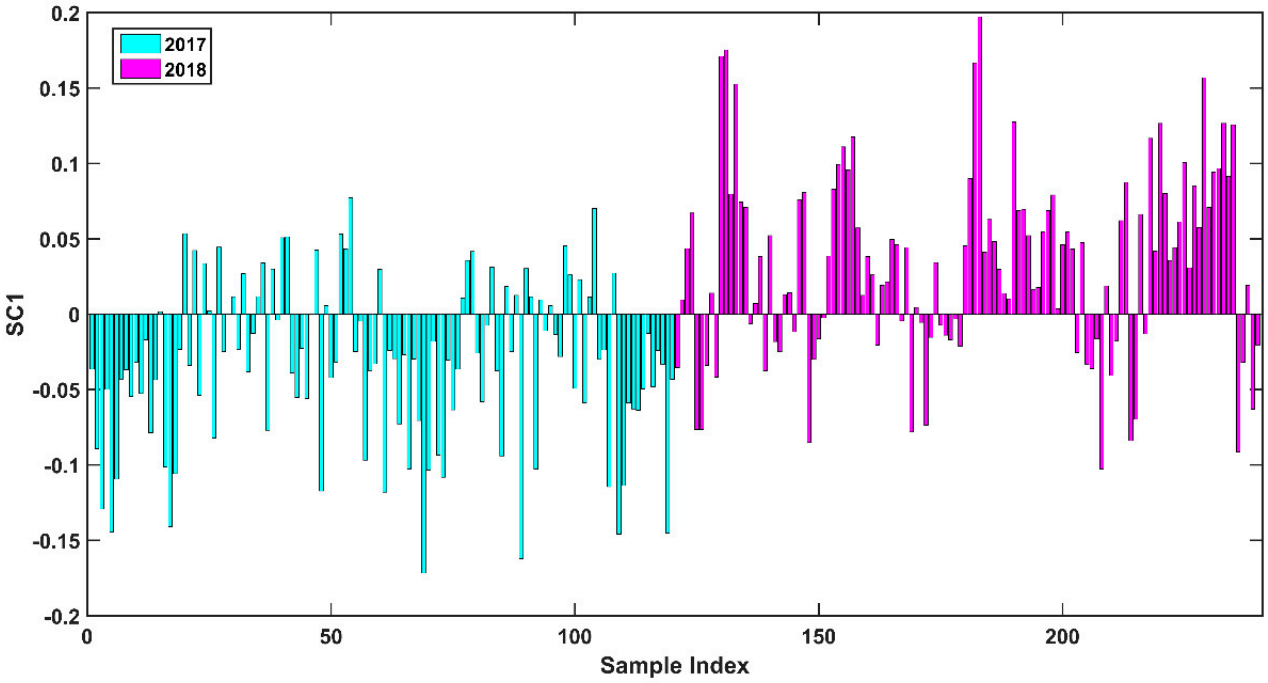

Figure 2. ASCA analysis: SCA model of the effect of the *harvesting year* factor. Sample scores were plotted onto the only SC of the model after the projection of the residual matrix to show within-level variability. Legend: Cyan Bars—2017; Magenta Bars—2018.

Accordingly, in order to assess the trends due to the production year in more detail, a principal component analysis considering the samples of Castelluccio di Norcia, the only class having six years available, was performed (Table 1).

From the scores plot in Figure 3, it can be appreciated that the three most aged groups of samples (2016, 2017, and, to a lesser extent, 2018) appear superimposed. On the other hand, lentils harvested in 2019 and 2020 overlap (samples not yet expired). It is obvious that the most recent samples (2021) present a behavior different from all the other individuals. In fact, they show a quite narrow distribution. This trend may be associated with the dry climatic conditions of this year. This high drought very likely influenced the average color of lentil samples and their distribution. Inspection of the loadings (Appendix C, Figure A5) shows how, with aging, samples are characterized by a decrease in the intensity of the green and blue channels and a corresponding increase in the red channel. At the same time,

the lightness also decreases, i.e., aged samples are characterized by a higher saturation and lower hue.

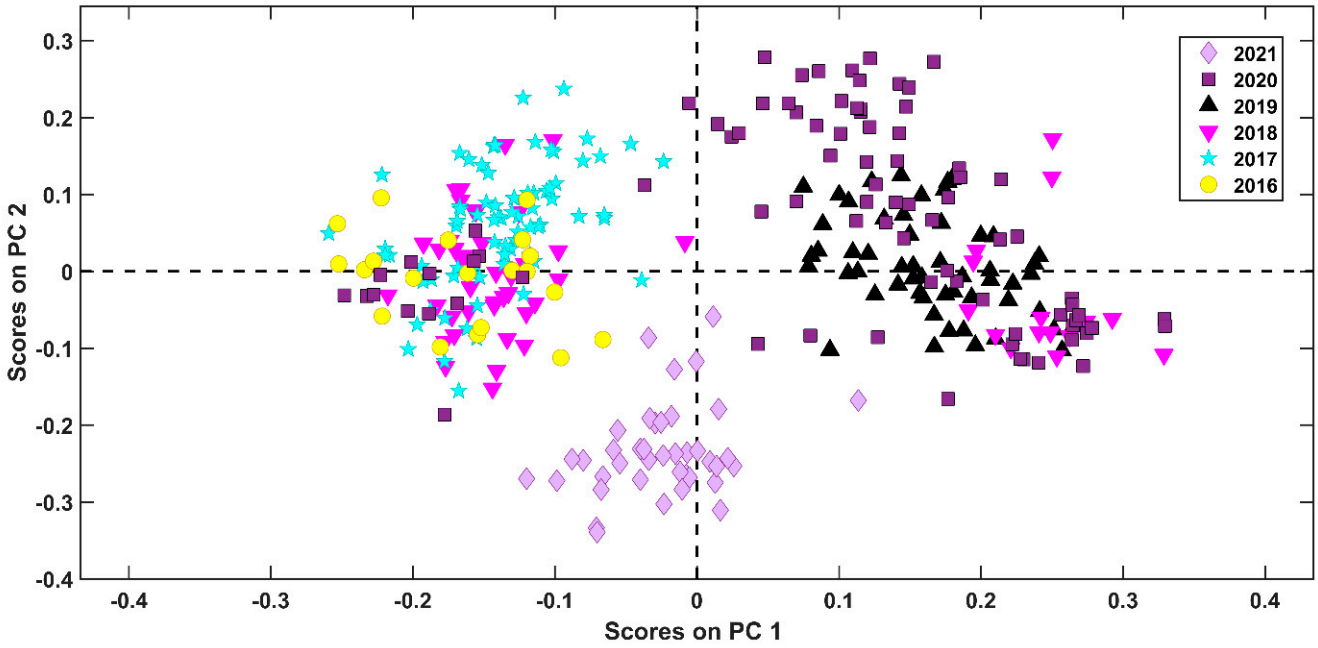

Figure 3. PCA of samples from Castelluccio di Norcia. Legend: Lilac diamonds: 2021; Purple squares: 2020; Black triangles: 2019; Magenta down-ward triangles: 2018; Cyan stars: 2017; and Yellow circles: 2016.

However, it should be noted that, since the analyses were conducted in 2022, the aging process has certainly produced a variation in the color of the samples, introducing a variability that may affect the one related to the year of production. Accordingly, seed deterioration is related to progressive seed coat browning due to the production of brown polymeric compounds that result from lipidic peroxidation and the formation of free radicals. The effect occurred even though the samples were carefully stored in dry and dark conditions.

Therefore, in conclusion, the inspection of the *year effect* factor essentially interconnects with the samples' aging, not allowing a clear and independent interpretation. Nevertheless, this extensive exploratory analysis allowed us to direct the subsequent classification phase.

3.2. SPORT Classification According to the Harvesting Year, the Origin, and the Edibility

Since the purposes of the present study were manifold, three different classification models were built and validated, one aimed at discriminating samples according to their origin (*model I*), one designed for the classification of lentils according to their harvesting year *(model II)*, and one tailored to discern edible (i.e., photographed prior to their expiration date) and expired (i.e., photographed after to their expiration date) samples. All models were built testing two different data pre-treatments: mean-centering and autoscaling; from now on, the mean-centered block is referred to as $\mathbf{X}_1$, whereas the autoscaled one will be called $\mathbf{X}_2$. Accordingly, all three models were built using the SPORT approach, where the differently preprocessed data matrices were simultaneously analyzed by SO-PLS-LDA, as described in Section 2.4. In all cases, the optimal model complexity, i.e., the number of latent variables to be extracted from each block, was chosen as the one leading to the lowest mean classification error (the average of the classification error for the different categories) in a seven-fold cross-validation procedure. The choice of the mean classification error instead of the total classification error is recommended when the number of samples in the different categories is unbalanced, as in the present case. In the remainder of the section, the different models will be individually discussed.

3.2.1. Classification of the Samples According to Their Geographical Origin

At first, a model to discriminate among the three geographical origins of the lentil samples was built and validated. To this purpose, the 560 samples were split into a training and a test set by means of the Duplex algorithm [41] applied category-wise with a 70:30 splitting ratio. Accordingly, 392 samples (88 AQ, 70 COL, and 234 CDN) were included in the training set, while the remaining 168 (38 AQ, 30 COL, and 100 CDN) were left out to constitute the external test set. The training samples were then used to build the SPORT classification model, whose optimal complexity, estimated in cross-validation, resulted to be 11 and 3, for the mean-centered and autoscaled colorgram, respectively. The results are summarized in Table 3.

Table 3. Results of the SPORT-LDA modeling for the discrimination of samples according to their geographical origin (Model I).

	LVs				Sensitivity (%)			Specificity (%)		
	MC	AS	Accuracy%	Mean CCR%	AQ	COL	CDN	AQ	COL	CDN
Calibration	11	3	96.7	96.3	94.3	97.1	97.4	99.3	97.5	98.1
CV			90.1	89.5	84.1	92.9	91.5	98.0	92.9	93.7
Prediction			87.5	88.0	86.8	90.0	87.0	93.8	94.2	92.6

Legend: MC—Mean-centered data; AS—Autoscaled data; CCR—Correct classification rate; CV—Cross-validation.

The results reported in Table 3 show how the SPORT classification model led to very good results not only in the calibration stage but also when it was applied to the external validation samples. Indeed, an average correct classification rate of 88.0% was achieved, corresponding to the correct prediction of 87% of the AQ samples, 90% of the COL images, and 87% of the CDN lentils. The model results were also very specific, with values always higher than 92%. When looking at the directions of the misclassifications, it could be observed how the wrongly predicted COL samples were all assigned to the CDN category, whereas for the other two classes, misclassifications were equally distributed among the other categories.

It might be useful to emphasize that, in contrast to previous works [23,26] aimed at discriminating among Italian landraces by image analysis, the present study demonstrates the method's versatility by correctly distinguishing between typical varieties from relatively close areas that also span several production years.

3.2.2. Classification of the Samples According to Their Harvesting Year

A model to discriminate the samples according to their harvesting year was built and validated. In this case, the 560 samples were split into a training and a test set by means of the Duplex algorithm applied category-wise with a 70:30 splitting ratio. Accordingly, 383 samples (13 from 2016, 153 from 2017, 97 from 2018, 35 from 2019, 60 from 2020, and 25 from 2021) were included in the training set, while the remaining 177 (7 from 2016, 70 from 2017, 40 from 2018, 20 from 2019, 25 from 2020, and 15 from 2021) were left out to constitute the external test set. The training samples were then used to build the SPORT classification model, whose optimal complexity, estimated in cross-validation, resulted to be 11 and 3, for the mean-centered and autoscaled colorgram, respectively. The results are summarized in Table 4.

While the results in the calibration phase (i.e., when the model is applied to the same samples used for calculating its parameters) were very good, with an overall accuracy higher than 96%, those in cross-validation and prediction (i.e., on the external test set) were significantly lower, though comparable with one another. Indeed, the overall accuracy on the test set was about 69% and the mean correct classification rate was slightly less than 71%. This was mainly due to the inaccuracy associated with samples harvested in 2018, 2020, and, to a lesser extent, 2016, whose correct classification rates were 40.0%, 48.0%, and 57.1%, respectively. On the other hand, the odd harvesting years were accurately predicted; in fact, the model properly classified 100% of samples grown in 2021 and 2019, and 78%

of those from 2017. Inspection of the confusion matrix for the test set evidenced how the wrongly predicted 2016 samples were mostly classified as being from 2018 and only to a lesser extent as from 2017. On the other hand, the 2017 and 2018 samples were mostly confused with one another: 17.1% of the 2017 samples were predicted as 2018, while 42.5% of the 2018 samples were classified as 2017. Finally, the misclassified sample from 2020 was mostly predicted as 2018 and, to a lesser extent, as 2017. When trying to rationalize these results, it was found that, taking into account the climatological variables of the area, no direct and straightforward interpretation was possible for the area. Indeed, climatic information can only suggest that there was no actual heat stress. On the other hand, except for the year 2016, which was the wettest year, the crops were affected by water stress for all the sampled production years, with a peak in 2021 [42]; additionally, this phenomenon is known to affect plant composition [43]. One could then think that the less than completely satisfactory results observed on the test set may be related to the uneven distribution of samples among the geographical origins and years (with lentils of different origins being available only for two years, 2017 and 2018), or to the natural aging of the specimens, which could play a significant role in the data variability.

Table 4. Results of the SPORT-LDA modeling for the discrimination of samples according to their harvesting year (Model II).

		Calibration	CV	Prediction
LVs	MC		14	
	AS		11	
Accuracy (%)		96.9	64.2	68.9
Mean CCR (%)		96.2	68.4	70.6
Sensitivity (%)	2016	92.3	53.9	57.1
	2017	96.7	69.3	78.6
	2018	99.0	48.5	40.0
	2019	100.0	94.3	100.0
	2020	93.3	48.3	48.0
	2021	96.0	96.0	100.0
Specificity (%)	2016	98.1	95.1	97.7
	2017	99.1	77.0	80.4
	2018	100.0	85.0	84.7
	2019	99.7	99.1	97.5
	2020	99.4	94.4	96.7
	2021	100.0	99.4	100.0

Legend: MC—Mean-centered data; AS—Autoscaled data; CCR—Correct classification rate; CV—Cross-validation.

3.2.3. Classification of the Samples According to Their Edibility

Finally, a model to discriminate the samples according to whether they could still be considered edible at the time of analysis or not was built and validated. The Duplex algorithm was applied category-wise with a 70:30 splitting used to split the 560 samples into a training set of 383 samples (85 edible and 295 expired) and a test set of 177 samples (40 edible and 137 expired). The optimal model was found to be the one including 15 and 5 latent variables for the mean-centered and autoscaled colorgram, respectively. The results are summarized in Table 5.

The results summarized in Table 5 indicate that the model possesses a very good predictive ability both in the calibration and the validation phase. In particular, when looking at the results of the test set, an overall accuracy of about 90% is obtained. This value corresponds to the correct prediction of 75% of the edible lentils and 94% of the expired samples. The results can also be graphically visualized in Figure 4, where the values of the predicted response for the training and the test samples are displayed together with the

classification threshold (which, as described in Section 2.4, is calculated by applying LDA to the predicted Y values of the training set).

Table 5. Results of the SPORT-LDA modeling for the discrimination of samples according to their edibility (Model III).

	LVs				Sensitivity (%)		Specificity (%)	
	MC	AS	Accuracy%	Mean CCR%	Edible	Expired	Edible	Expired
Calibration	15	5	100.0	100.0	100.0	100.0	100.0	100.0
CV			87.0	78.1	62.1	94.1	94.1	62.1
Prediction			89.8	84.6	75.0	94.6	94.6	75.0

Legend: MC—Mean-centered data; AS—Autoscaled data; CCR—Correct classification rate; CV—Cross-validation.

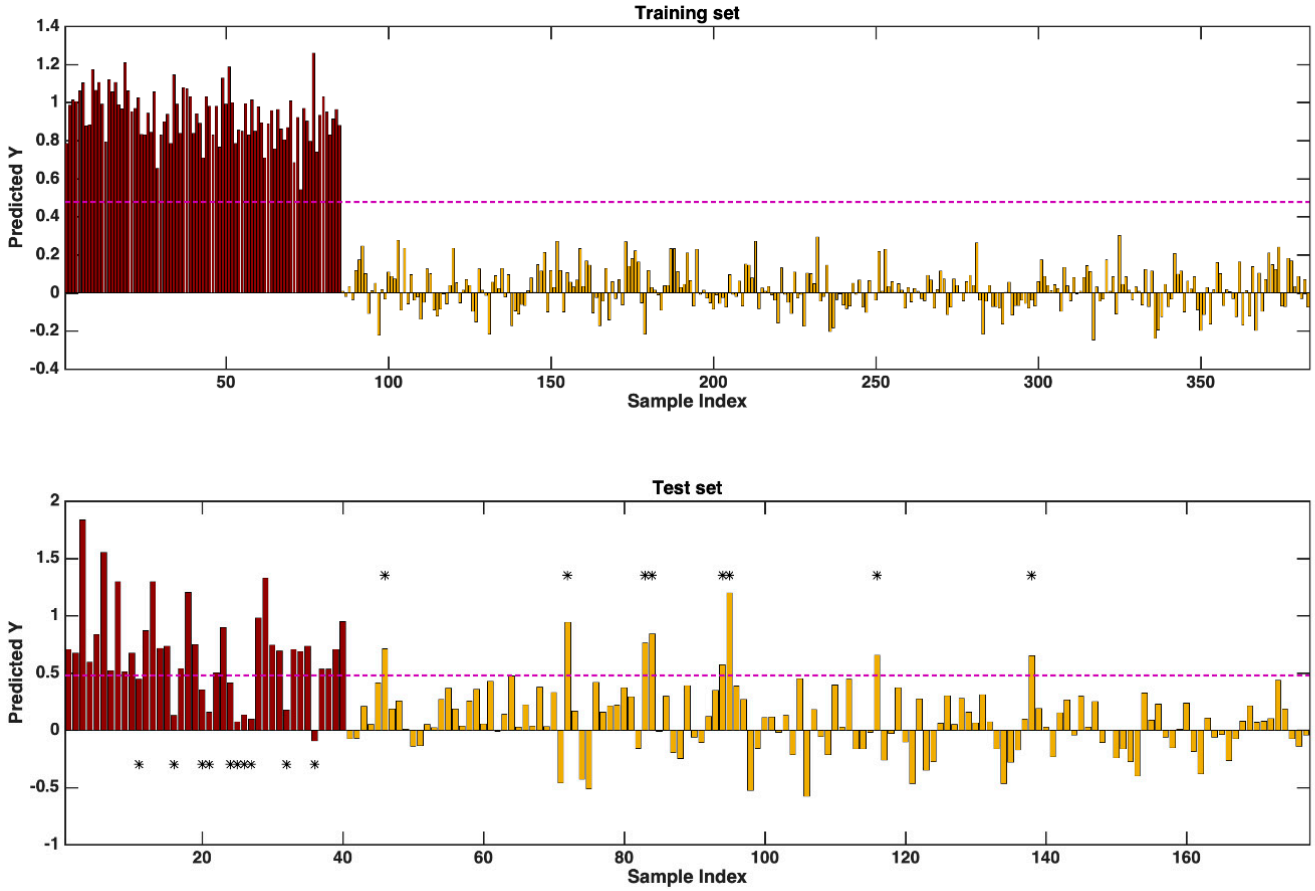

Figure 4. SPORT model for the discrimination of edible vs. expired lentils. Bar plot of the predicted response for the training (upper panel) and the test (lower panel) samples. The purple line represents the classification threshold, while the star symbols (*) indicate the misclassified samples. Legend: Bordeaux—Edible; Mustard yellow—Expired.

In the figure, all the samples with a predicted response value above the threshold are predicted as being edible, while those below the threshold are classified as expired. The misclassified samples are highlighted in the plot by a black star. It is then evident that all the training samples are perfectly classified since the predicted values of the response all fall on the right side of the threshold. On the other hand, when looking at the test set, ten compliant samples (Bordeaux bars) have predicted Y values lower than the threshold and are therefore predicted by the model as expired; analogously, eight images of expired lentils (mustard yellow bars) present response values higher than the classification limit and so are wrongly predicted as edible.

Nevertheless, it has to be pointed out that the predictions obtained by this model are particularly satisfactory. In fact, although not 100% accurate, the model misclassifies a lower percentage of expired samples, which is preferred, as it ensures that only a few expired samples are wrongly recognized as still edible.

It should be noted that Dell'Aquila [28] proved that a medium RGB index could be used to predict seed deterioration; nevertheless, through MIA, the classification of expired and edible lentil seeds was generalized, testing the applicability of this method on samples of different origin and production years.

4. Conclusions

In this study, the feasibility of using RGB imaging, which can be considered as a sort of e-eye, together with a chemometric discriminant classifier for the characterization of Italian lentils was demonstrated. In particular, after recording the images, MIA was applied to obtain the characteristic colorgrams of the samples, which were then processed by different state-of-the-art multivariate statistical tools. In particular, an initial ANOVA-simultaneous component analysis applied to a subset of the original data suggested that both the geographical origin and the harvesting year significantly affect the recorded experimental profiles. Successively, SPORT-LDA was applied to the dataset to build models able to discriminate the samples according to different categorizations (geographical origin, harvesting year, and compliance with respect to the expiry date). The approach showed good classification efficiency despite the complexity of the problem and the multiple sources of variability. Indeed, the model built to discriminate the different geographical origins resulted in an average correct classification rate on the test set of about 88%, with comparable sensitivity for all the three investigated categories. A lower classification accuracy (close to 70% on the test set) was instead obtained for the model discriminating the samples according to their harvesting year. Lastly, an overall 90% accuracy was obtained on the test samples by the model built to recognize whether a sample was still within its expiry date or not.

Thus, through this work, it was verified that MIA, coupled with the latest pattern recognition methods, is a suitable, fast, inexpensive, and non-destructive approach for the quality control of typical Italian varieties of lentils, useful for monitoring and authenticating the local populations over the years as well as ensuring their edibility.

Author Contributions: Conceptualization, M.F. and A.B.; methodology, A.B.; software, A.B. and F.M.; validation, M.F., A.B. and F.M.; formal analysis, A.B., M.F. and V.D.M.; investigation, A.B., M.F. and V.D.M.; resources, A.A.D.; data curation, M.F., A.B. and A.A.D.; writing—original draft preparation, M.F., F.M. and A.B.; writing—review and editing, A.A.D. and F.M.; visualization, M.F. and A.B.; supervision, A.A.D. and F.M.; project administration, A.A.D.; funding acquisition, A.A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Examples of the collected images are shown below (Figure A1); in detail, images of lentil samples produced, in the same year in the three considered production areas are reported, as well as those collected from the Castelluccio di Norcia grown over several years (from 2016 to 2018).

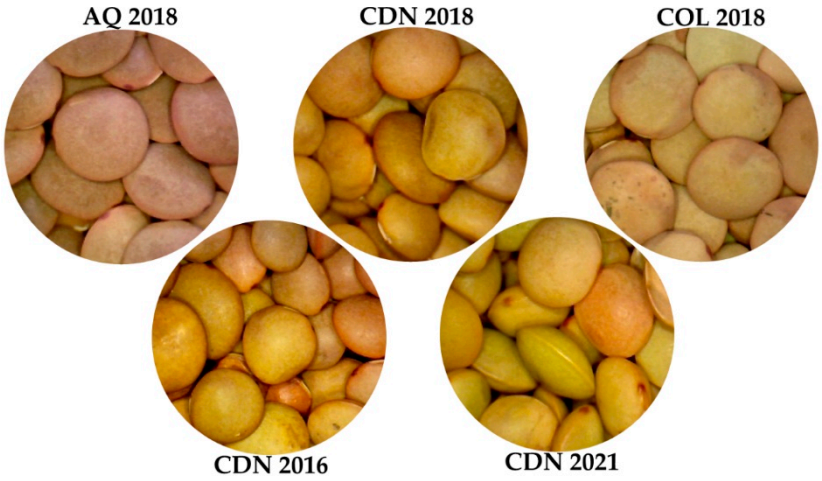

Figure A1. Some photos of lentil samples taken according to the indications reported in Section 2.2 and divided based on the area (AQ 2018, CDN 2018, and COL 2018) and the year of production (CDN 2016, CND 2018, and CDN 2021).

Figure A2 shows the average colorgrams for the classes involved in the main categorizations considered for the present study. In particular, in the left panel, the average profiles for the three geographical origins (AQ, COL, and CDN) are shown, whereas, in the right panels, the mean colorgrams calculated from samples from the different harvesting years (from 2016 to 2021) are displayed.

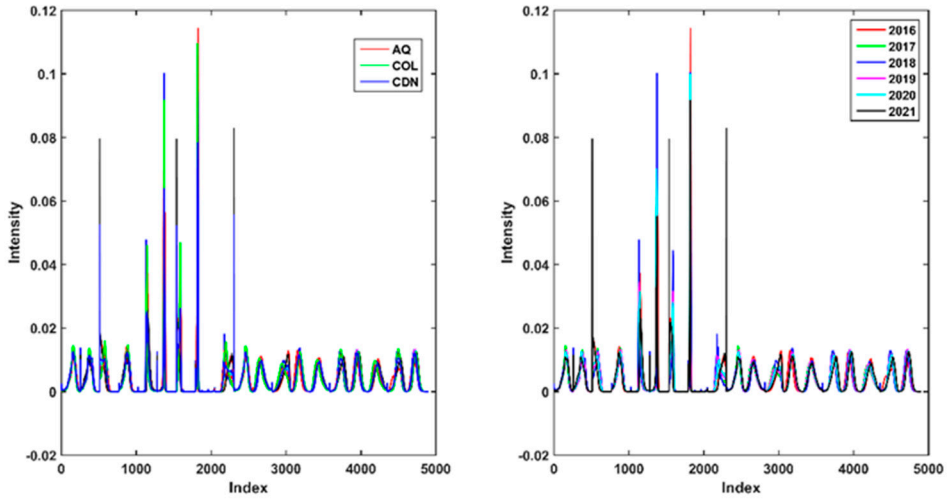

Figure A2. Mean colorgrams calculated for the different categories considered in the study. Left panel: average colorgrams corresponding to the three geographical origins; right panel: average colorgrams corresponding to the six harvesting years.

The colorgrams were obtained by globally considering the pixels of the acquired images and by concatenating 19 frequency distribution vectors of 256 elements related to the following ordered parameters: red channel (1–256 variable index), green channel (257–512), blue channel (513–768), lightness (769–1024), relative red (1025–1280), relative green (1281–1536), relative blue (1537–1792), hue (1793–2048), saturation (2049–2304), intensity (2305–2560), distribution curve of the first, second, and third score vectors from the PCA on the raw unfolded RGB matrix (2561-3328), distribution curve of the first, second, and third score vectors from the PCA on the mean-centered unfolded RGB matrix (3329–4096), distribution curve of the first, second, and third score vectors from the PCA on the autoscaled unfolded RGB matrix (4097–4864), and normalized loading vectors and eigenvalues of the three PCA models (4865–4900) [44].

Appendix B

Appendix B reports the loading plots obtained by the ASCA and differentiated into significant and not significant according to the bootstrap procedure.

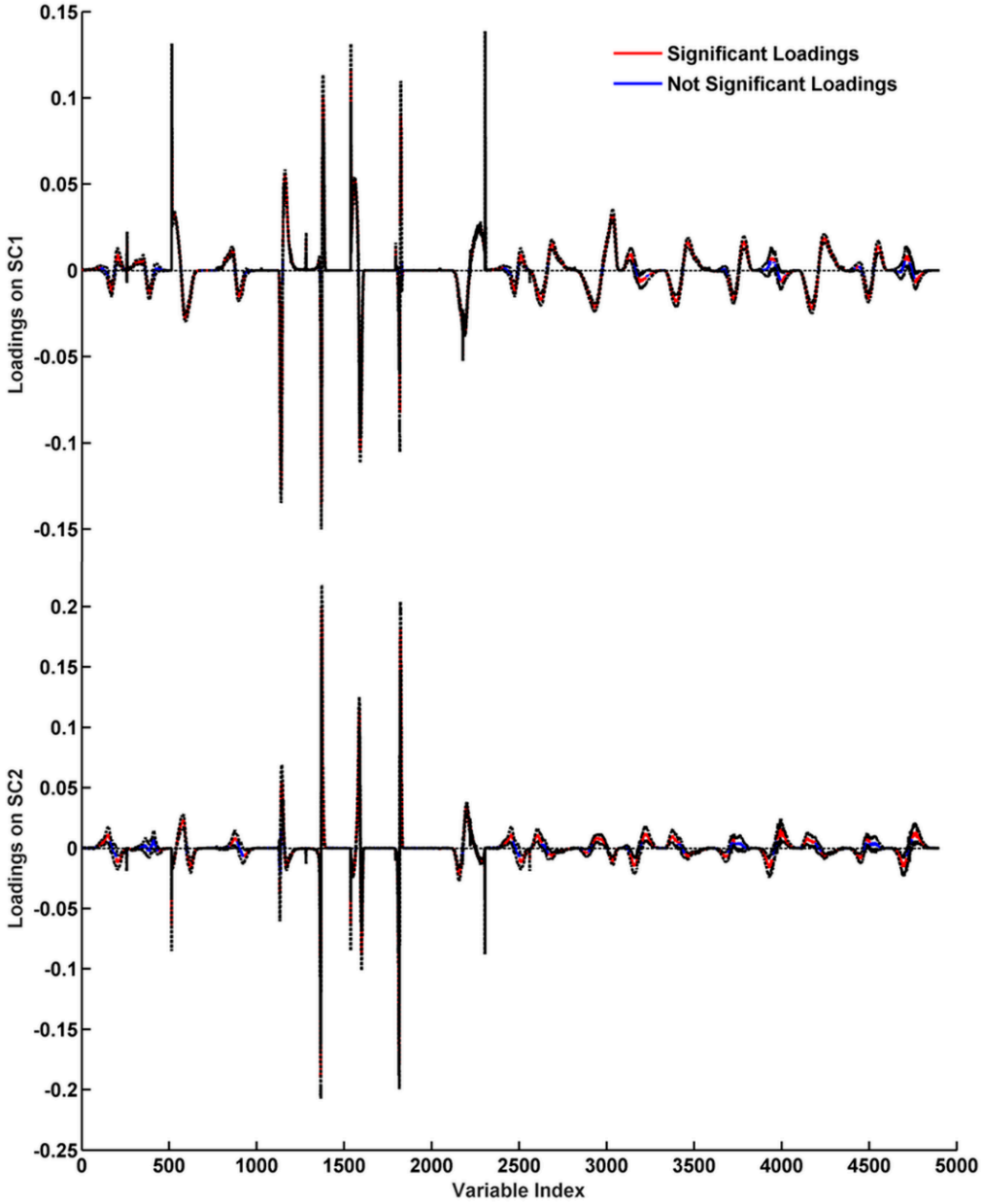

Figure A3. Loadings on SC1 and SC2 for the geographical origin factor together with their 95% confidence interval estimated by bootstrapping. The variables results highlighted in red were significantly different than zero.

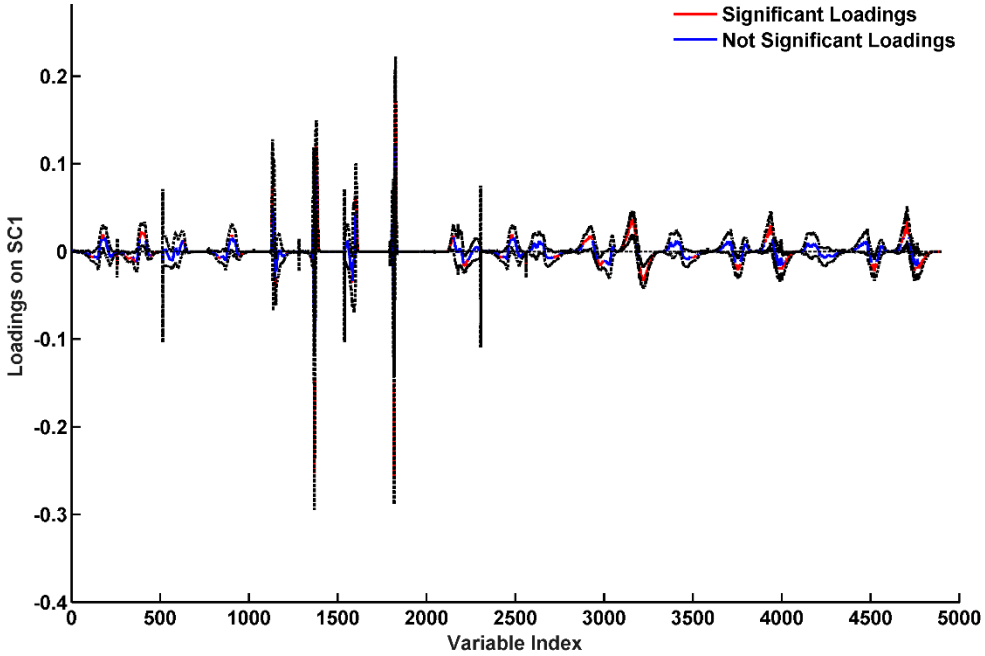

Figure A4. Loadings on SC1 for the *Harvesting year* factor together with their 95% confidence interval estimated by bootstrapping. The variable results highlighted in red were significantly different than zero.

Appendix C

Appendix C reports the loading plots obtained by PCA for the analysis of Castelluccio di Norcia samples.

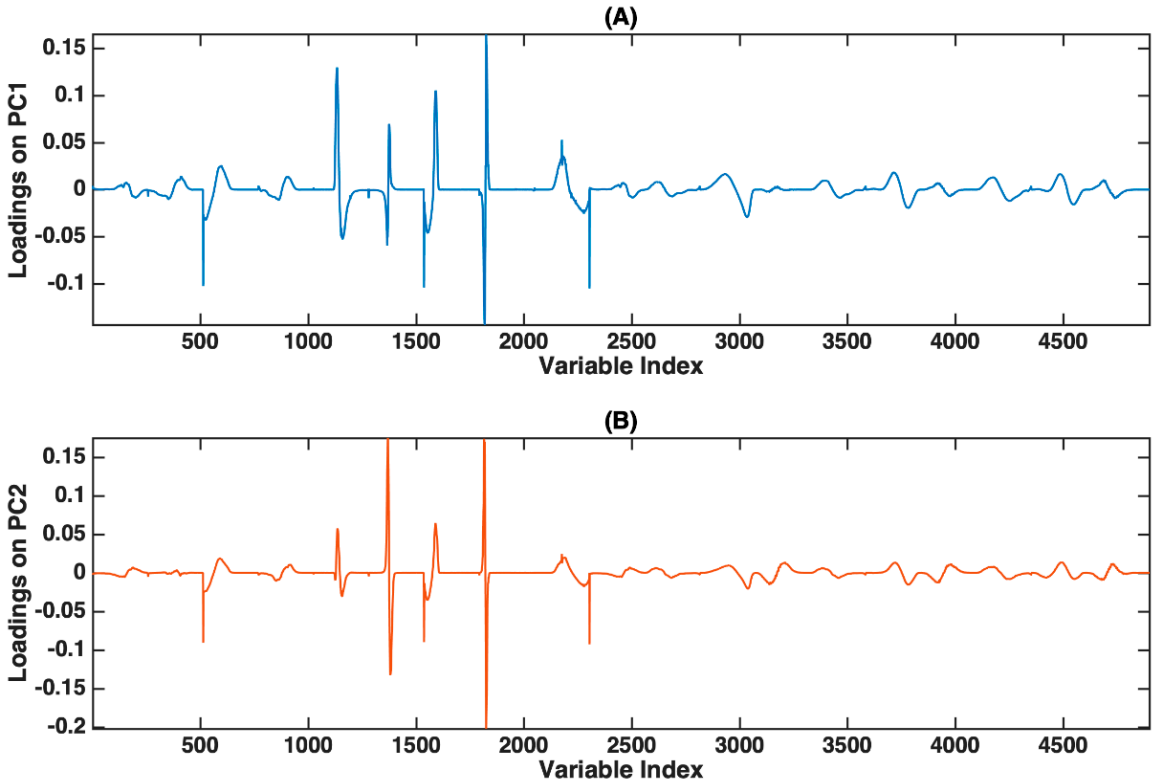

Figure A5. PCA of samples from Castelluccio di Norcia: Loadings on PC1 (**A**) and PC2 (**B**).

References

1. Sonnante, G.; Pignone, D. The major Italian landraces of lentil (*Lens culinaris* Medik.): Their molecular diversity and possible origin. *Genet. Resour. Crop Evol.* **2007**, *54*, 1023–1031. [CrossRef]
2. Kaale, L.D.; Siddiq, M.; Hooper, S. Lentil (*Lens culinaris* Medik) as nutrient-rich and versatile food legume: A review. *Legum. Sci.* **2022**, e169. [CrossRef]
3. Chelladurai, V.; Erkinbaev, C. Lentils. In *Pulses*; Manickavasagan, A., Thirunathan, P., Eds.; Springer International Publishing: Cham, Swizterland, 2020; pp. 129–143. ISBN 978-3-030-41376-7.
4. Johnson, N.; Johnson, C.R.; Thavarajah, P.; Kumar, S.; Thavarajah, D. The roles and potential of lentil prebiotic carbohydrates in human and plant health. *Plants People Planet* **2020**, *2*, 310–319. [CrossRef]
5. Food and Agriculture Organization of the United Nations (FAO). Everything You Need to Know About Lentils. Available online: https://www.fao.org/pulses-2016/blog/everything-you-need-to-know-about-lentils/en/ (accessed on 17 January 2023).
6. Faris, M.A.-I.E.; Takruri, H.R.; Issa, A.Y. Role of lentils (*Lens culinaris* L.) in human health and nutrition: A review. *Med. J. Nutr. Metab.* **2013**, *6*, 3–16. [CrossRef]
7. Duranti, M. Grain legume proteins and nutraceutical properties. *Fitoterapia* **2006**, *77*, 67–82. [CrossRef]
8. Lombardi, M.; Materne, M.; Cogan, N.O.I.; Rodda, M.; Daetwyler, H.D.; Slater, A.T.; Forster, J.W.; Kaur, S. Assessment of genetic variation within a global collection of lentil (*Lens culinaris* Medik.) cultivars and landraces using SNP markers. *BMC Genet.* **2014**, *15*, 150. [CrossRef]
9. Sen Gupta, D.; Thavarajah, D.; Knutson, P.; Thavarajah, P.; McGee, R.J.; Coyne, C.J.; Kumar, S. Lentils (*Lens culinaris* L.), a rich source of folates. *J. Agric. Food Chem.* **2013**, *61*, 7794–7799. [CrossRef] [PubMed]
10. Khazaei, H.; Caron, C.T.; Fedoruk, M.; Diapari, M.; Vandenberg, A.; Coyne, C.J.; McGee, R.; Bett, K.E. Genetic diversity of cultivated lentil (*Lens culinaris* Medik.) and its relation to the world's agro-ecological zones. *Front. Plant Sci.* **2016**, *7*, 1093. [CrossRef]
11. Laskar, R.A.; Khan, S.; Deb, C.R.; Tomlekova, N.; Wani, M.R.; Raina, A.; Amin, R. Lentil (*Lens culinaris* Medik.) Diversity, cytogenetics and breeding. In *Advances in Plant Breeding Strategies: Legumes: Volume 7*; Al-Khayri, J.M., Jain, S.M., Johnson, D.V., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 319–369. ISBN 978-3-030-23400-3.
12. Zaccardelli, M.; Lupo, F.; Piergiovanni, A.R.; Laghetti, G.; Sonnante, G.; Daminati, M.G.; Sparvoli, F.; Lioi, L. Characterization of Italian lentil (*Lens culinaris* Medik.) germplasm by agronomic traits, biochemical and molecular markers. *Genet. Resour. Crop Evol.* **2012**, *59*, 727–738. [CrossRef]
13. Sehgal, A.; Sita, K.; Bhandari, K.; Kumar, S.; Kumar, J.; Vara Prasad, P.V.; Siddique, K.H.M.; Nayyar, H. Influence of drought and heat stress, applied independently or in combination during seed development, on qualitative and quantitative aspects of seeds of lentil (*Lens culinaris* Medikus) genotypes, differing in drought sensitivity. *Plant. Cell Environ.* **2019**, *42*, 198–211. [CrossRef]
14. Choukri, H.; Hejjaoui, K.; El-Baouchi, A.; El haddad, N.; Smouni, A.; Maalouf, F.; Thavarajah, D.; Kumar, S. Heat and Drought Stress Impact on Phenology, Grain Yield, and Nutritional Quality of Lentil (*Lens culinaris* Medikus). *Front. Nutr.* **2020**, *7*, 596307. [CrossRef]
15. Longobardi, F.; Casiello, G.; Cortese, M.; Perini, M.; Camin, F.; Catucci, L.; Agostiano, A. Discrimination of geographical origin of lentils (*Lens culinaris* Medik.) using isotope ratio mass spectrometry combined with chemometrics. *Food Chem.* **2015**, *188*, 343–349. [CrossRef] [PubMed]
16. Longobardi, F.; Innamorato, V.; Di Gioia, A.; Ventrella, A.; Lippolis, V.; Logrieco, A.F.; Catucci, L.; Agostiano, A. Geographical origin discrimination of lentils (*Lens culinaris* Medik.) using1H NMR fingerprinting and multivariate statistical analyses. *Food Chem.* **2017**, *237*, 443–448. [CrossRef] [PubMed]
17. Foschi, M.; Archivio, A.A.D.; Rossi, L. Geographical discrimination and authentication of lentils (*Lens culinaris* Medik.) by ICP-OES elemental analysis and chemometrics. *Food Control* **2020**, *118*, 107438. [CrossRef]
18. Innamorato, V.; Longobardi, F.; Lippolis, V.; Cortese, M.; Logrieco, A.F.; Catucci, L.; Agostiano, A.; De Girolamo, A. Tracing the Geographical Origin of Lentils (*Lens culinaris* Medik.) by Infrared Spectroscopy and Chemometrics. *Food Anal. Methods* **2019**, *12*, 773–779. [CrossRef]
19. Ceccobelli, S.; Ciancaleoni, S.; Lancioni, H.; Veronesi, F.; Albertini, E.; Rosellini, D. Genetic distinctiveness of a Protected Geographic Indication lentil landrace from the Umbria region, Italy, over 20 years. *Genet. Resour. Crop Evol.* **2019**, *66*, 1483–1493. [CrossRef]
20. Scippa, G.S.; Rocco, M.; Ialicicco, M.; Trupiano, D.; Viscosi, V.; Di Michele, M.; Arena, S.; Chiatante, D.; Scaloni, A. The proteome of lentil (*Lens culinaris* Medik.) seeds: Discriminating between landraces. *Electrophoresis* **2010**, *31*, 497–506. [CrossRef]
21. Caprioli, G.; Cristalli, G.; Ragazzi, E.; Molin, L.; Ricciutelli, M.; Sagratini, G.; Seraglia, R.; Zuo, Y.; Vittori, S. A preliminary matrix-assisted laser desorption/ionization time-of-flight approach for the characterization of Italian lentil varieties. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 2843–2848. [CrossRef]
22. Biancolillo, A.; Foschi, M.; Di Micco, M.; Di Donato, F.; D'Archivio, A.A. ATR-FTIR-based rapid solution for the discrimination of lentils from different origins, with a special focus on PGI and Slow Food typical varieties. *Microchem. J.* **2022**, *178*, 107327. [CrossRef]
23. Torricelli, R.; Silveri, D.D.; Ferradini, N.; Venora, G.; Veronesi, F.; Russi, L. Characterization of the lentil landrace Santo Stefano di Sessanio from Abruzzo, Italy. *Genet. Resour. Crop Evol.* **2012**, *59*, 261–276. [CrossRef]

24. Mahajan, S.; Das, A.; Sardana, H.K. Image acquisition techniques for assessment of legume quality. *Trends Food Sci. Technol.* **2015**, *42*, 116–133. [CrossRef]
25. Shahin, M.A.; Symons, S.J.; Wang, N. Predicting dehulling efficiency of lentils based on seed size and shape characteristics measured with image analysis. *Qual. Assur. Saf. Crop. Foods* **2012**, *4*, 9–16. [CrossRef]
26. Venora, G.; Grillo, O.; Shahin, M.A.; Symons, S.J. Identification of Sicilian landraces and Canadian cultivars of lentil using an image analysis system. *Food Res. Int.* **2007**, *40*, 161–166. [CrossRef]
27. Shahin, M.A.; Symons, S.J. A machine vision system for grading lentils. *Can. Biosyst. Eng. Genie Biosyst. Canada* **2001**, *43*, 77–714.
28. Dell'Aquila, A. Red-Green-Blue (RGB) colour density as a non-destructive marker in sorting deteriorated lentil (*Lens culinaris* Medik.) seeds. *Seed Sci. Technol.* **2006**, *34*, 609–619. [CrossRef]
29. Pradana-López, S.; Pérez-Calabuig, A.M.; Otero, L.; Cancilla, J.C.; Torrecilla, J.S. Is my food safe?—AI-based classification of lentil flour samples with trace levels of gluten or nuts. *Food Chem.* **2022**, *386*, 132832. [CrossRef] [PubMed]
30. Jansen, J.J.; Hoefsloot, H.C.J.; van der Greef, J.; Timmerman, M.E.; Westerhuis, J.A.; Smilde, A.K. ASCA: Analysis of multivariate data obtained from an experimental design. *J. Chemom.* **2005**, *19*, 469–481. [CrossRef]
31. Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.J.; Lamers, R.-J.A.N.; van der Greef, J.; Timmerman, M.E. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043–3048. [CrossRef]
32. Bertinetto, C.; Engel, J.; Jansen, J. ANOVA simultaneous component analysis: A tutorial review. *Anal. Chim. Acta X* **2020**, *6*, 100061. [CrossRef]
33. Roger, J.-M.; Biancolillo, A.; Marini, F. Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. *Chemom. Intell. Lab. Syst.* **2020**, *199*, 103975. [CrossRef]
34. Næs, T.; Tomic, O.; Mevik, B.-H.; Martens, H. Path modelling by sequential PLS regression. *J. Chemom.* **2011**, *25*, 28–40. [CrossRef]
35. Biancolillo, A.; Næs, T. The Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions. *Data Handl. Sci. Technol.* **2019**, *31*, 157–177. [CrossRef]
36. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173. [CrossRef]
37. Nocairi, H.; Qannari, E.M.; Vigneau, E.; Bertrand, D. Discrimination on latent components with respect to patterns. Application to multicollinear data. *Comput. Stat. Data Anal.* **2005**, *48*, 139–147. [CrossRef]
38. Indahl, U.G.; Martens, H.; Næs, T. From dummy regression to prior probabilities in PLS-DA. *J. Chemom.* **2007**, *21*, 529–536. [CrossRef]
39. Kennard, R.W.; Stone, L.A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148. [CrossRef]
40. Thiel, M.; Féraud, B.; Govaerts, B. ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J. Chemom.* **2017**, *31*, e2895. [CrossRef]
41. Snee, R.D. Validation of Regression Models: Methods and Examples. *Technometrics* **1977**, *19*, 415–428. [CrossRef]
42. Associazione Linea Meteo Rete Stazioni Meteo Linea Meteo. Available online: http://www.lineameteo.it/retemeteo.php (accessed on 15 August 2022).
43. Borràs, D.; Plazas, M.; Moglia, A.; Lanteri, S. The influence of acute water stresses on the biochemical composition of bell pepper (*Capsicum annuum* L.) berries. *J. Sci. Food Agric.* **2021**, *101*, 4724–4734. [CrossRef]
44. Antonelli, A.; Cocchi, M.; Fava, P.; Foca, G.; Franchini, G.C.; Manzini, D.; Ulrici, A. Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm. *Anal. Chim. Acta* **2004**, *515*, 3–13. [CrossRef]