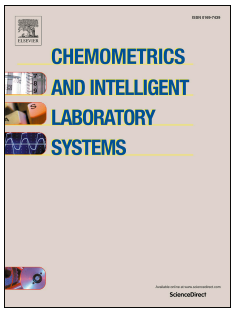
Journal Pre-proof



Spectroscopic Fingerprinting and Chemometrics for the Discrimination of Italian Emmer Landraces

Martina Foschi, Alessandra Biancolillo, Simona Vellozzi, Federico Marini, Angelo Antonio D'Archivio, Ricard Boqué

PII: S0169-7439(21)00116-7

DOI: https://doi.org/10.1016/j.chemolab.2021.104348

Reference: CHEMOM 104348

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 10 February 2021

Revised Date: 15 April 2021

Accepted Date: 19 May 2021

Please cite this article as: M. Foschi, A. Biancolillo, S. Vellozzi, F. Marini, A.A. D'Archivio, R. Boqué, Spectroscopic Fingerprinting and Chemometrics for the Discrimination of Italian Emmer Landraces, *Chemometrics and Intelligent Laboratory Systems*, https://doi.org/10.1016/j.chemolab.2021.104348.

Author contribution

Martina Foschi: investigation, methodology, formal analysis and writing - original draft

Alessandra Biancolillo: methodology, formal analysis and writing - original draft

Simona Vellozzi: investigation, methodology

Federico Marini: methodology, writing – review & editing

Angelo Antonio D'Archivio: resources, conceptualization and writing - original draft

Ricard Boqué: methodology, conceptualization and writing - review & editing

Spectroscopic Fingerprinting and Chemometrics for the Discrimination of Italian Emmer Landraces

Martina Foschi [a*], Alessandra Biancolillo [a], Simona Vellozzi [b], Federico Marini [b], Angelo Antonio D'Archivio [a], Ricard Boqué [c]

[a]Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio 67100, Coppito, L'Aquila, Italy; martina.foschi@graduate.univaq.it ; alessandra.biancolillo@univaq.it ; angeloantonio.darchivio@univaq.it

[b]Department of Chemistry, Sapienza University of Rome, P.le Aldo Moro 5, Rome, 00185, Italy; federico.marini@uniroma1.it ; vellozzi.1482655@studenti.uniroma1.it

[c] Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira i Virgili, Campus Sescelades, Edifici N4, C/Marcel·lí Domingo s/n, Tarragona, 43007, Spain; ricard.boque@urv.cat

*****Correspondence: martina.foschi@graduate.univaq.it

Abstract

Emmer is a traditional Italian wheat species attracting growing attention for the high-nutritive and dietary value. The growth of emmer consumption and the recent spreading even in areas where production was not traditional pose a risk to biodiversity and to the geographical identities. Thus, the present work aims to develop a non-destructive and routine-compatible method to discriminate three Italian landraces and lay the basis for a possible authentication method. One-hundred and forty-seven emmer samples, harvested in 2019 in three traditional production areas (Garfagnana, Monteleone di Spoleto, Gran Sasso and Monti della Laga National Park), were investigated by Mid-Infrared (MIR) and Near-Infrared (NIR) spectroscopy. Two different approaches of multiclass

Partial Least Squares-Discriminant Analysis (PLS-DA) were applied on the collected fingerprinting profiles. Eventually, Data-Fusion strategies have been employed to combine the different information sources and classify the samples according to the geographical origin. The most accurate predictions were provided by the Sequential and Orthogonalized-Partial Least Squares-Discriminant Analysis (SO-PLS-DA) model, which misclassified only one test sample over 44 (in external validation). Finally, a chemical interpretation of the most discriminant variables was performed.

Keywords: Emmer; Infrared; Classification; Multi-block; Data Fusion; SO-PLS

1 Introduction

Triticum dicoccum (Shubler), commonly known as emmer, is a tetraploid hulled wheat species, whose cultivation is supposed to date back to the First Agricultural Revolution during the Neolithic period [1]. The tightly bond glume, covering the "hulled wheats", helps the grains to retain nutrients and to be tolerant against abiotic (heat and soil conditions) and biotic (fungal deceases like Fusarium head blight, powdery mildew, rusts) stresses [2]. At the same time, the thick emmer husk has to be removed making the grains more laborious to process than modern wheat; regardless, the very high stem, which characterizes the old wheat, makes it unsuitable for intensive agricultural practice. The progress in grain production, which has encouraged higher-yielding and free-thrashing wheat cultivation, has made the old cereals actually a marginal crop [3]. Nowadays, the key concepts of food sovereignty and agroecology, as an alternative solution to food industry globalization, are gaining popularity both among consumers and producers [4]. As a result, alternative foodstuffs with high added-value, as emmer is, have attracted attention meeting the demand of consumers for social, economic and environmental reasons. Indeed, emmer is considered a high-nutritive cereal (rich in minerals fiber and antioxidants) showing also some dietary values

such as high protein digestibility and low glycaemic index [5,6]. Furthermore, emmer is a low-input plant and, due to the ability in growing even in unfavorable soils and climatic conditions, it is suitable for organic farming and offers an alternative to common wheat for the requalification of marginal rural areas. In Italy, the cultivation of emmer represents a long-lasting tradition, proved by the large use of this ancient wheat by the Etruscan and Roman populations and by the numerous landraces well-adapted to the typical local production areas [7,8]. Traditional landraces of *dicoccum* wheat are still grown in the upland areas located in Tuscany, Abruzzo, Umbria, Lazio and in the southern regions of Italy. The value of those native populations is recognized and protected by European and national seals that are linked to biodiversity and geographical identities, such as "protected geographical indication" (PGI), Traditional Agri-food Products (PAT) and Protected Designation of Origin (PDO). The growth of emmer consumption and the resulting recent spreading of the crop, even in not typical and more fruitful areas, are increasing the risk related to the genetic contamination and to the loss of competitiveness of traditional farmers, a scenario that calls for increased control and monitoring due to the difficulty to ensure traceability of the production process and to the higher probability of frauds.

Therefore, analytical methods aimed at food traceability could play an important role to safeguard consumers, honest producers and the wealth of the genetic diversity that has been preserved by farmers over generations. Several analytical techniques combined with chemometrics have proved to be suitable to distinguish the origin and the cultivar of different wheat species. Head-space solid-phase microextraction coupled with gas chromatography-mass spectrometry (HS - SPME/GC - MS) and chemometrics, such as Multivariate ANalysis Of Variance (MANOVA), Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), was used to assess the influence of the combined effects of regional provenience and cultivar on the volatile composition of Chinese winter wheat [9]. Non-targeted Liquid Chromatography coupled to High-Resolution Mass Spectrometry (LC-HRMS) and chemometrics (PCA and Orthogonal-Partial Least Squares- Discriminant

Analysis) was used to detect chemical markers able to correctly discriminate Italian, UE and non-UE Durum wheat samples [10]. Several studies of geographical discrimination of wheat and wheat products have employed Nuclear Magnetic Resonance (NMR), Isotope Ratio Mass Spectrometry (IRMS), high resolution-inductively coupled plasma-mass spectrometry (HR-ICP-MS) and X-ray fluorescence (XRF) techniques coupled with multivariate statistical analysis [11–15].

Among the analytical techniques aimed at food authenticity and traceability, Fourier Transform infrared (FT-IR) spectroscopy is gaining increasing popularity mainly due to the possibility of better exploiting the data through chemometric tools [16]. IR spectroscopy-based methods are suitable for routine quality analysis of agri-food products since they are non-destructive, fast and do not require complex sample pre-treatments [17]. In this regard, FT-NIR and ATR-FT-MIR were used and compared to discriminate naturally contaminated wheat samples according to the high or low Ochratoxin A levels [18], as well as to detect durum wheat pasta adulteration with common wheat [19] or to determine crude protein and intestinal protein digestibility [20]. NIR spectroscopy coupled with multivariate calibration techniques has been applied to classify wheat flour samples according to quality category [21], to predict milling and baking parameters [22] and has been also widely used as a geographical fingerprinting of very different food matrices[23–26]. In this context, classification of wheat samples according to the variety [27] or geographic origin [28,29] was carried out based on NIR spectra coupled to discriminant pattern-recognition methods. To our knowledge, studies that involve emmer wheat are mainly based on chemical characterization aimed at valorising and evaluating the nutritional and nutraceutical values of different ancient wheat species [30] and assessing the possibility to differentiate between *Triticum* species (*monococcum, dicoccum, spelta* and *turgidum*) [31–34]. Thus, geographical discrimination of emmer is still undescribed.

In the present work, geographical discrimination of three valuable Italian *Triticum diccoccum* landraces, coming from Garfagnana (PGI), Monteleone di Spoleto (PDO) and from the Gran Sasso e monti della Laga National Park (PAT), was attempted through the chemometric elaboration of their IR spectroscopic data (Fig.1A). In detail, 147 whole kernels were analysed by ATR-FT-MIR and FT-NIR spectroscopies. Each sample was cut in two exploiting the fragility near the grain cavity and the derived two cross-sections were further analysed by the above-mentioned techniques. The spectral information (coming from the outer part and from the endosperm (Fig.1B)) was used to build PLS-DA models in order to discriminate samples according to the origin.

----------Insert Figure 1 approx. here----------

Two different multiclass approaches, multiclass-probabilistic PLS-DA and "*one versus one"* PLS-DA, were applied and compared for each dataset [35,36]. Furthermore, an improvement of the classification ability was attempted by applying low level, parallel and sequential mid-level data fusion strategies which have already proven to provide better classification performances in similar agri-food discrimination problems [37–39]. Therefore, the aim of the work is to find the best chemometric elaboration and combination of the data matrices that better discriminate Italian emmer produced in relatively close areas, in order to lay the groundwork for a possible fast methodology to assess compliance of this specific value-added food matrix with the production specifications.

2 Experimental

2.1 Emmer samples

Emmer samples were purchased directly from local producers and supermarkets, in which the traceability and authenticity of the product were ensured. The study focused on three different landraces harvested in the traditional production areas during 2019. The PGI Emmer "Farro della

Garfagnana", hereinafter indicated with the acronym GA, is an autumn variety whose production has to strictly comply with the regulations. In this regard, the polishing of the husked grain ("brillatura") has to be mechanical and, in some cases, is still carried out with stone mills [40]. The PDO "Farro di Monteleone di Spoleto", which will be named MS, is a spring variety perfectly adapted to the pedoclimatic conditions of the area located over 700 meters above the sea level. To study an analogous marketed product, it was decided to purchase the one commercialized as semi-pearled since it is, as required by the product specification, "slightly scratched with a milling machine to remove the husk" [40]. The last selected landrace was the PAT product known as "Farro d'Abruzzo", an autumn variety produced in the Gran Sasso and Monti della Laga National Park that will be indicated with GS. Although there are no detailed production rules for the "Farro d'Abruzzo", the Italian quality mark ensures certain uniformity of the traditional production process throughout the production area [41]. Also in this case, the selected samples were purchased as "semi-pearled product" like in the other classes. A total number of 147 whole kernels were sampled from the collected lots and, of these, 53 belong to the class GA, 53 to the class MS and 41 to GS. The entire dataset was also properly divided (see section 3.1) in a training set of 103 samples, used to build the models, and a test set of 44 samples employed for the external validation.

2.2 IR spectroscopic analysis

The samples were analysed using an FT-NIR Nicolet 6700 (Thermo Scientific Inc., Madison, WI) equipped with a halogen-tungsten lamp. Spectra were acquired in reflectance mode with an indium gallium arsenide (InGaAs) detector and an integrating sphere working in the range of 4000-10,000 cm^{-1} and with a nominal resolution of 4 cm^{-1} (82 scans). Four NIR analyses were performed on each grain: two replicates on the outer part, which is the one mechanically processed, and two obtained by exposing the internal part of each of the two grain sections, which were derived from the whole kernel fracture. The same sections were further analysed, following the same procedure, with a PerkinElmer Spectrum Two™ (PerkinElmer, Waltham, MA, USA) FT-IR spectrometer equipped with a PerkinElmer Universal Attenuated Total Reflectance (uATR) sampling accessory

and a deuterated triglycine sulphate (DTGS) detector. After air background, the spectra acquisition was performed from 4000 cm^{-1} to 400 cm^{-1} (instrumental resolution of 4 cm^{-1}) and was carried out at room temperature. A controlled pressure was applied to the sample to make it adhere to the crystal and increase the signal-to-noise ratio by using an integrated loading monitoring system. All the spectra (NIR and MIR spectra from the outer and the inner part of each grain) were exported in MATLAB 2012b (The Mathworks, Natick, MA) and were converted into pseudo-absorbance (log(1/R)). The two replicates per side were averaged obtaining two MIR (147x1800) data matrices and two NIR (147x3112) matrices that were subsequently processed and analysed through chemometrics.

2.3 Partial least squares - discriminant analysis

Partial least squares-discriminant analysis (PLS-DA) is a linear classification method commonly applied to spectroscopic data since, exploiting the PLS regression algorithm, it is suitable to treat ill-conditioned matrices [42]. It is actually possible to apply the regression method in classification problems when the **X** data matrix is related to a **Y** dependent binary matrix (NxG) that encodes the class membership of each of the N objects in the G classes. PLS-DA allows dealing with highly-correlated variables by searching for the orthogonal directions (Latent Variables) of maximum covariance with the **Y**-block and while ensuring an explanation of the relevant sources of data variability [43]. The regression model can be expressed in terms of original **X** and **Y** blocks by the relation: $\mathbf{Y} = \mathbf{X}\boldsymbol{B}_{PLS.} + \mathbf{E}$, where **E** is the residual matrix and $\boldsymbol{B}_{PLS}$ is the matrix of the regression coefficients, estimated for the calibration model, that can be used together with $\mathbf{X}_{new}$ to predict $\mathbf{Y}_{new}$ [16]. Since the computed or predicted responses are not binary but continuous real values, a class-assignment criterion must be adopted. In this work, a classification rule based on the Bayesian theorem was chosen [44]. In this phase, according to the binary multiclass approach, two Gaussian distributions (one for the in-class samples and the other for all the not-in-class ones) were built from the continuous Y-values and for each column of the predicted **Y**-block (NxG). This allows calculating the probability to observe a Y-value for a given sample that belongs to the G-th class or

that does not belong to it. A probabilistic threshold for each class can be pinpointed at the intersection of the curves that, due to the normalization step, corresponds to an *a posteriori* probability of 0.5 [35]. In discrimination problems, where a single class assignment is required, the probabilistic criterion may result in a univocal assignation when the objects are attributed to the class with the higher *a posteriori* probability. The results obtained with the above-mentioned approach were compared with those derived from the "*one versus one*" strategy proposed by Pérez et al. [36]. According to this method, G(G-1)/2 binary models were built, thus, as explained above, the two normal class distributions were computed for each model involving only two classes. The strategy permits overcoming limitations that could arise when different classes, which may be incompatible, are grouped together also producing unbalanced categories. Based on the Y-values computed by the optimal binary models, G-1 Probability Density Functions (PDFs) were estimated for each class. Therefore, to classify an unknown object, a combination of the predicted probabilities was carried out according to the procedure described in ref. [36]. Eventually, the class assignment was performed according to the highest combined probability.

2.4 Data fusion strategy

Data Fusion (DF) represents a typical and interdisciplinary tool useful for handling data of different nature (acquired by different modalities) but that share a common mode [45]. The goal of the integration of datasets from multiple-sources is to build a more informative and better-quality model that is achieved by assessing the common or complemental information and data-block interactions [39]. Depending on the level at which the fusion takes place, it is possible to categorize the strategy as a low level, mid-level and, finally, high-level data fusion, which was not performed in the present work. In the low-level DF, the original datasets are concatenated according to the shared mode, which is commonly the sample mode [46]. In the present work, raw data as well as the block-scaled ones were concatenated sample-wise and processed according to the two multiclass pattern recognition methods described above. Mid-level DF consists of extracting the relevant features from each data block and of concatenating them into a single matrix, which is further

206 elaborated. In this case, the feature extraction may be performed independently for each block
207 (Multi-Block mid-level DF), thus the blocks are exchangeable. Otherwise, it can be performed
208 sequentially; in the latter case, the model will depend on the specific order of the data blocks [47].
209 Attempting the sequential methodology, Sequential and Orthogonalized-Partial Least Squares-
210 Discriminant Analysis (SO-PLS-DA) algorithm, described in refs. [48,49], was used.
211
212 ----------Insert Fig. 2 approx. here----------
213
214 The SO-PLS-DA's algorithm for two data blocks (schematized in Fig. 2) proceeds by regressing the
215 responses on the first data block (**X**) by PLS algorithm (1st step) and by searching for improvement
216 of predictions using additional and orthogonal information provided by the second matrix (**Z**orth).
217 The 2nd regression consists of fitting the **Y**-residuals of the first PLS regression to the **Z** matrix,
218 which has been orthogonalized with respect to the **X**-score matrix of the 1st step. **Y** is computed
219 summing up the predictions of the two individual regressions as follows: $\mathbf{Y} = \mathbf{X}\boldsymbol{B} + \mathbf{Z}\boldsymbol{C} + \mathbf{E}$, where
220 **X** (N x J) and **Z** (N x L) are predictor blocks, **Y** (N x G) is the response matrix, $\boldsymbol{B}$ (J x G) and $\boldsymbol{C}$ (L
221 x G) are the regression coefficients and **E** (N x G) the residual matrix. The real-valued **Y**, which
222 was computed or predicted by the method, was used to perform the classification based on the
223 probabilistic approaches described above.
224
225 **3 Results and discussion**
226 *3.1 Exploratory analysis*
227 As mentioned, NIR and MIR spectra, collected analysing the outer processed part and the
228 endosperm of each grain, were used to evaluate the ability of each technique and their combination
229 to differentiate three traditional emmer landraces.
230

----------Insert Fig. 3 approx. here----------

Fig.3 shows the four spectra (NIRin NIRout MIRin MIRout) obtained by averaging the signals for each given class (GA MS and GS). The exploratory analysis was conducted on raw spectra; the portion of the spectrum, which was worked on, was reduced to the range from 850 to 4000 cm^{-1} for MIR spectra and to the interval from 4000 to 9000 cm^{-1} for NIR spectra since the removed regions were noisy and strongly influenced by the light scattering compared to the information provided by the signals. Since it was decided to apply a multi-block strategy, some preliminary decisions had to be taken. Firstly, the complementarity of the information sources was verified; in detail, the available datasets were divided according to the sample membership and PCA models were built, for each class, concatenating variable-wise the matrices resulting from the same instrumental technique. The outcomes of the MIR-PCA models reveal a systematic differentiation between the samples analysed from the outer and from the inner part, supporting the decision to consider MIRin and MIRout as two different blocks providing complemental information. The same procedure applied to NIRin and NIRout data did instead not show a substantial difference between the two analyses, which could be related to a greater penetration depth of the more energetic wavelengths. Based on this evidence NIRin and NIRout were averaged and treated as a single information source, thereafter indicated as NIRin/out. To externally validate the classification models, a data splitting procedure that would guarantee a good representation of both the three categories under study and of the variability in all the considered data-blocks was proposed by Biancolillo et al. [50]. The procedure consists of applying the Kennard-Stone duplex algorithm [51], separately for each class, on the row-augmented matrices obtained by concatenating the significant principal components extracted from all the three data-blocks (MIRin MIRout NIRin/out). Eventually, all the single-class subsets were collected resulting in a training set of 103 samples and in a test set of 44 samples. After the splitting procedure, single-block PLS-DA models were built to determine the best data

pre-treatment and to assess the best technique that could be able to discriminate the samples according to their geographical origin.

3.2 Single-block PLS-DA

First and second derivatives, calculated using the Savitzky–Golay approach (19 points window and third-order interpolating polynomial)[52], Standard Normal Variate (SNV) and their combinations were tested to find the most suitable data pre-treatment; in all cases, mean-centering was further applied before building the classifier. A 10-fold cross-validation procedure was applied to establish the optimal data pre-processing and complexity (number of latent variables) of each model, which was subsequently validated on the test set. Table 1 shows the chosen data pre-treatment, the classification rate in cross-validation and the related prediction ability. It can be deduced, by the correct classification rate both in internal and external validation, that MIRout and NIRin/out provided promising results with a correct classification rate (CCR) of 86.8% (in cross-validation) and 92.3% (on the test set) and a CCR of 89.3% (internal cross-validation) and 86.8% (in external validation), respectively. On the basis of the outcomes of the MIRin model (81.6% in cross-validation, 86.8% on the external set), which was produced analysing the fractured samples, it could be deduced that the benefits when considering this data-block are not enough compared to the drawback of having a slower and destructive method. Nevertheless, the possibility of improving the classification ability by a DF strategy using all the three data-blocks was attempted (method described in the dedicated section 3.4). Fig.4A and B graphically display the probabilistic criterion through which the class attribution took place, in the representative case of the PLS-DA model on MIRout data. In particular, Fig.4A shows the target class distribution intersecting the Probability Density Function (PDF) which results from all the samples not belonging to the class. The reported PDFs were estimated from the computed $\hat{Y}$ of the training set and were for each pair (class/not-class) normalized. Fig.4B, in which the confused samples are highlighted, shows the normalized probabilities and the posterior probability values attributed to the samples by interpolation of the $\hat{Y}$ with the PDFs.

----------Insert Fig. 4 approx. here----------

3.3 "one against one" approach

The "one against one" binarization method was employed to evaluate the possible problems that could have arisen by using a criterion that is binary by construction in a multi-class setting. The approach was tested for all the single blocks and even for the fused matrices (the "*1vs1*" mid-level DF outcome, which does not differ from the one obtained with the Mid-Level Multi-Block-PLS-DA approach, is reported in Tab.1). Even if the binarization method has led to a partial redistribution of the classification errors, whose outcomes in external validation are reported in Tab.1 (cross-validation results are not reported to keep the table readable), the total classification rates remain almost unchanged compared to the multi-class probabilistic PLS-DA models. It can therefore be assessed that no real problems related to unbalanced or incompatible groups have occurred when, for the estimation of the class probability, a super-class (which is the grouping of all the not-in-class samples) is considered.

----------Insert Tab.1 approx. here----------

3.4 Data-Fusion approach

Low-Level DF was applied by concatenating sample-wise the matrices that were pre-processed according to the optimal pre-treatment identified during the optimization of the single-block PLS-DA models. Before building the Low-Level PLS-DA model, the fused matrix was block-scaled and mean-centered. The Frobenius' norm (total sum of squares of each matrix) was used as a block-scaling factor to make the blocks comparable and to avoid the model being driven only by the blocks with the greater number of variables. The best Low-Level DF classification ability was achieved by involving all the three blocks (MIRout, NIRin/out and MIRin) and 12 Latent Variables but satisfactory results, compared to the ones obtained by single-block models, were not reached (see Tab.1). Eventually, to test the Mid-Level DF strategy, the scores of the LVs separately

extracted in the optimal single-block PLS-DA models were concatenated row-wise and a multi-block PLS-DA model was developed using the fused matrix (103x21). Excellent results were obtained using only the MIRout and NIRin/out matrices with an optimal complexity of 3 Latent Variables. Although in the Mid-Level DF an additional optimization step is required, a more balanced representation of the information carried by each block is ensured, which is a relevant factor especially when blocks containing a large number of variables are fused.

Besides, a sequential Mid-Level DF was applied and compared to the above-described method. SO-PLS-DA has several advantages compared to the Multi-Block PLS-DA method; indeed, although it requires a greater computational effort, it potentially involves only one optimization step. As described in section 2.4, SO-PLS-DA sequentially extracts the information from the blocks defining an optimal complexity for each dataset; thus, it permits the independent evaluation of the information sources and the removal of redundancy among them. Furthermore, a direct identification of significant variables and recognition of complemental information may be performed through the VIP (Variable Importance in Projection) analysis leading, in this case, to a direct chemical interpretation [53]. Fig.5A shows a 3D plot of the $\hat{Y}$ in calibration (full symbols) and in validation (empty symbols) resulting from the SO-PLS-DA model. The three class-distributions are well distinguished as well as the incorrect attribution that the model makes in the external validation (one GS sample confused with GA). Moreover, the external samples proved to be representative of the class they belong to and they fall relatively close to the training samples. MS is the most sensible and specific class, correctly recognizing all the compliant samples and rejecting all the samples that do not belong to the class. $\hat{Y}_{GA}$ and $\hat{Y}_{GS}$ appear more dispersed than $\hat{Y}_{MS}$ but, despite this, the SO-PLS-DA model achieved sensitivities of 100% and 91.7% with a complexity of 9 LVs for the MIRout block and 6 LVs for the NIRin/out. GS is the class showing the highest confusion, both for GA and MS. This could be explained considering the concurrence of different factors such as the analogy in the variety, which is autumnal in the case of GA and GS

categories, but also the actual proximity of the production areas of GS and MS (Gran Sasso e Monti della Laga National Park and Monteleone di Spoleto respectively) (Fig.1A). In the present work, High-Level data fusion was not tested since, taking place at the decision level, the evaluation of the complementarity of the blocks results troublesome. Moreover, the SO-PLS-DA model produced satisfactory results without compromising the chemical interpretability in the discrimination process.

----------Insert Fig. 5 approx. here----------

3.5 VIP analysis and chemical interpretation

Fig. 1(B) shows the structure of the cereal caryopsis and the rough chemical composition concerning the different seed tissue. Proteins are mainly concentrated in the cells of the aleurone layer, which is rich in albumins and globulins and characterized by a higher content of essential amino acids and lysine [54]. Galterio et al. [55] quantified the protein percentage in three Italian landraces (Garfagnana, Leonessa and Trivento) and found values not greater than 10%. However, the high variability of this value was confirmed in the literature demonstrating the strong influence of the growing conditions and the genetic background on the protein content that also has proved to be higher in spring emmer than in the autumn or winter varieties [56]. The amide I (~1650 cm^{-1}) and amide II (~1550 cm^{-1}) bands, which arise from specific stretching and bending vibrations of the protein backbone [57], are clearly distinguishable in the MIRout spectrum demonstrating a significant contribution of the aleurone layer to the signals. The pericarp as well as the seed coat and, to a lesser extent, the aleurone layer, are also characterized by structural carbohydrates, such as cellulose or hemicellulose, and by lignin, whose characteristic signals can be found in the MIRout spectrum (~1240 for the cellulosic material and the shoulders at ~1600 and ~1515, not present in the MIRin spectrum, indicative of the aromatic character of the lignin). Starch is the main storage carbohydrate in emmer kernels, accounting for 61 – 68% of the grain; MIRin shows a greater influence of starch typical bands, such as the α-1,4 Glycosidic bonds skeleton vibration at ~ 930 cm^{-}

^{1}, demonstrating the higher starch amount in the endosperm [58]. Finally, the lipids, which are the minor constituents (2% of the emmer kernel), are mainly concentrated in the germ and pericarp. Lipid signals are found in the region 2800–3000 cm^{-1}. The signals are predominantly asymmetric and symmetric stretching vibrations (~2922 and ~2852 cm^{-1}) of the CH_2 groups of the acyl chains. In addition, the band at ~1736 cm^{-1}, which arises from a stretching vibration of the carboxyl group C=O in the lipid ester linkage, is higher in the MIRout spectrum [59]. The highest intensity could demonstrate the seed coat and the pericarp highest lipids amount. The VIP analysis was performed for the SO-PLS-DA model in order to identify the variables giving the highest contribution in discriminating emmer landraces. VIP coefficients express the importance of each variable in defining the LVs subspace; a VIP index equal to 1, which is the average of the square values, is assumed as a cut-off to define which spectral variables are the most significant. Fig.5B reports a graphical representation of VIP analysis: the red points correspond to the spectral variables having VIP scores higher than one, while the black line is the sample mean spectrum. The CH_2 symmetric and asymmetric stretching (~2922 $cm^{-1}$and ~2856 cm^{-1}), as well as the CH bending at 1469 cm-1, are selected. Variables related to the amide I and II signals and to the C=O stretching vibration show VIP scores higher than 1. Finally, the region of the spectrum from 1180 cm^{-1} to 950 cm^{-1} related to the CO stretching in structural and not structural carbohydrates is highlighted as significant, as well as the starch characteristic band at ~930 cm^{-1} and the signal detected at ~1235 cm^{-1} that is due to coupled OH bending and CO stretching vibrations. The selected NIRin/out features are shown in Fig. 5B and were selected considering the information already explained by the first block (MIRout). In particular, the variables in the range of ~5000 and ~5500 cm^{-1} are highlighted as relevant predictors and associated to the combination bands of the NH, OH and C=O bonds [21]. VIP indexes higher than 1 are shown by variables in the spectral range ~7000 to ~7200 cm^{-1}, an area associable to the presence of carbohydrates, or to the absorption of non-bonded O-H groups in fatty acids [60]. Finally, a limited number of variables is selected between ~8000 and ~8900 cm^{-1} and 4100- 4600cm^{-1} probably due to the correlated information already took into

account in the MIRout block and linked to the second overtone and combination bands of C-H bond [29].

4 Conclusion

Promising results were obtained in the first part of the study by applying the multiclass PLS-DA approaches to the individual data blocks. The Data-Fusion strategy was applied to handle the multi-block dataset leading to comparable or better results than single-block models. The best outcome was obtained with MIRout and NIRin/out matrices elaborated through the SO-PLS-DA method (97.2% of total correct classification rate). The satisfactory results confirm that MIR and NIR spectroscopic techniques could be used to analyse the whole emmer grain and to provide information that could be combined to assess the geographical origin of Italian emmer. The proposed methodology could also encourage the development of a similar approach aimed at emmer authentication.

References

[1] L. Hlisnikovský, M. Hejcman, E. Kunzová, L. Menšík, The effect of soil-climate conditions on yielding parameters, chemical composition and baking quality of ancient wheat species Triticum monococcum L., Triticum dicoccum Schrank and Triticum spelt L. in comparison with modern Triticum aestivum L., Arch. Agron. Soil Sci. 65 (2019) 152–163. https://doi.org/10.1080/03650340.2018.1491033.

[2] M. Zaharieva, N.G. Ayana, A. Al Hakimi, S.C. Misra, P. Monneveux, Cultivated emmer wheat (Triticum dicoccon Schrank), an old crop with promising future: a review, Genet. Resour. Crop Evol. 57 (2010) 937–962. https://doi.org/10.1007/s10722-010-9572-6.

[3] S. Padulosi, K. Hammer, J. Heller, Hulled wheats, Promoting the conservation and used of underutilized and neglected crops 4, IPGRI, Rome, 1996.

[4] A. Wezel, S. Bellon, T. Doré, C. Francis, D. Vallod, C. David, Agroecology as a science, a movement and a practice. A review, Agron. Sustain. Dev. 29 (2009) 503–515. https://doi.org/10.1051/agro/2009004.

[5] S. Marino, R. Tognetti, A. Alvino, Effects of varying nitrogen fertilization on crop yield and grain quality of emmer grown in a typical Mediterranean environment in central Italy, Eur. J. Agron. 34 (2011) 172-180. https://doi.org/10.1016/j.eja.2010.10.006.

[6] A. Arzani, M. Ashraf, Cultivated Ancient Wheats (Triticum spp.): A Potential Source of Health-Beneficial Food Products, Compr. Rev. Food Sci. Food Saf. 16 (2017) 477–488. https://doi.org/10.1111/1541-4337.12262.

[7] G. Barcaccia, L. Molinari, O. Porfiri, F. Veronesi, Molecular characterization of emmer (Triticum dicoccon Schrank) Italian landraces, Genet. Resour. Crop Evol. 49 (2002) 417–428. https://doi.org/10.1023/A:1020650804532.

[8] A. Troccoli, P. Codianni, Appropriate seeding rate for einkorn, emmer, and spelt grown under rainfed condition in southern Italy, Eur. J. Agron. 22 (2005) 293–300. https://doi.org/10.1016/j.eja.2004.04.003.

[9] S.A. Wadood, G. Boli, Z. Xiaowen, A. Raza, W. Yimin, Geographical discrimination of Chinese winter wheat using volatile compound analysis by HS-SPME/GC-MS coupled with multivariate statistical analysis, J. Mass Spectrom. 55 (2020) e4453. https://doi.org/10.1002/jms.4453.

[10] D. Cavanna, C. Loffi, C. Dall'Asta, M. Suman, A non-targeted high-resolution mass spectrometry approach for the assessment of the geographical origin of durum wheat, Food Chem. 317 (2020) 126366. https://doi.org/10.1016/j.foodchem.2020.126366.

[11] R. Lamanna, L. Cattivelli, M.L. Miglietta, A. Troccoli, Geographical origin of durum wheat studied by 1H-NMR profiling, Magn. Reson. Chem. 49 (2011) 1–5. https://doi.org/10.1002/mrc.2695.

[12] F. Longobardi, D. Sacco, G. Casiello, A. Ventrella, A. Sacco, Characterization of the Geographical and Varietal Origin of Wheat and Bread by Means of Nuclear Magnetic Resonance (NMR), Isotope Ratio Mass Spectrometry (IRMS) Methods and Chemometrics: A Review, Agric. Sci. 06 (2015) 126–136. https://doi.org/10.4236/as.2015.61010.

[13] H. Liu, Y. Wei, H. Lu, S. Wei, T. Jiang, Y. Zhang, B. Guo, Combination of the $^{87}Sr/^{86}Sr$ ratio and light stable isotopic values ($\delta^{13}C$, $\delta^{15}N$ and δD) for identifying the geographical origin of winter wheat in China, Food Chem. 212 (2016) 367–373. https://doi.org/10.1016/j.foodchem.2016.06.002.

[14] R. Consonni, L.R. Cagliani, Chapter 4-Nuclear Magnetic Resonance and Chemometrics to Assess Geographical Origin and Quality of Traditional Food Products, in: S. L. Taylor (Ed.), Adv. Food Nutr. Res. 59, Academic Press, Cambridge, 2010, pp. 87–165. https://doi.org/10.1016/S1043-4526(10)59004-1.

[15] H. Zhao, B. Guo, Y. Wei, B. Zhang, Multi-element composition of wheat grain and provenance soil and their potentialities as fingerprints of geographical origin, J. Cereal Sci. 57 (2013) 391–397. https://doi.org/10.1016/j.jcs.2013.01.008.

[16] A. Biancolillo, F. Marini, Chapter 4-Chemometrics Applied to Plant Spectral Analysis, in: J. Lopes, C. Sousa (Eds.), Vibrational Spectroscopy for Plant Varieties and Cultivars Characterization, Compr. Anal. Chem. 80, Elsevier, Amsterdam, 2018, pp. 69–104. https://doi.org/10.1016/bs.coac.2018.03.003.

[17] S. Lohumi, S. Lee, H. Lee, B.-K. Cho, A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration, Trends Food Sci. Technol. 46 (2015) 85–98. https://doi.org/10.1016/j.tifs.2015.08.003.

[18] A. De Girolamo, C. von Holst, M. Cortese, S. Cervellieri, M. Pascale, F. Longobardi, L. Catucci, A.C.R. Porricelli, V. Lippolis, Rapid screening of ochratoxin A in wheat by infrared

458 spectroscopy, Food Chem. 282 (2019) 95–100.
459 https://doi.org/10.1016/j.foodchem.2019.01.008.

460 [19] A. De Girolamo, M.C. Arroyo, S. Cervellieri, M. Cortese, M. Pascale, A.F. Logrieco, V.
461 Lippolis, Detection of durum wheat pasta adulteration with common wheat by infrared
462 spectroscopy and chemometrics: A case study, LWT. 127 (2020) 109368.
463 https://doi.org/10.1016/j.lwt.2020.109368.

464 [20] H. Shi, Y. Lei, L. Louzada Prates, P. Yu, Evaluation of near-infrared (NIR) and Fourier
465 transform mid-infrared (ATR-FT/MIR) spectroscopy techniques combined with
466 chemometrics for the determination of crude protein and intestinal protein digestibility of
467 wheat, Food Chem. 272 (2019) 507–513. https://doi.org/10.1016/j.foodchem.2018.08.075.

468 [21] M. Cocchi, M. Corbellini, G. Foca, M. Lucisano, M.A. Pagani, L. Tassi, A. Ulrici,
469 Classification of bread wheat flours in different quality categories by a wavelet-based feature
470 selection/classification algorithm on NIR spectra, Anal. Chim. Acta. 544 (2005) 100–107.
471 https://doi.org/10.1016/j.aca.2005.02.075.

472 [22] O. Jirsa, M. Hrušková, I. Švec, Near-infrared prediction of milling and baking parameters of
473 wheat varieties, J. Food Eng. 87 (2008) 21–25.
474 https://doi.org/10.1016/j.jfoodeng.2007.09.008.

475 [23] P. Firmani, S. De Luca, R. Bucci, F. Marini, A. Biancolillo, Near infrared (NIR)
476 spectroscopy-based classification for the authentication of Darjeeling black tea, Food
477 Control. 100 (2019) 292–299. https://doi.org/10.1016/j.foodcont.2019.02.006.

478 [24] X. Li, L. Zhang, Y. Zhang, D. Wang, X. Wang, L. Yu, W. Zhang, P. Li, Review of NIR
479 spectroscopy methods for nondestructive quality analysis of oilseeds and edible oils, Trends
480 Food Sci. Technol. 101 (2020) 172–181. https://doi.org/10.1016/j.tifs.2020.05.002.

481 [25] P. Firmani, R. Bucci, F. Marini, A. Biancolillo, Authentication of “Avola almonds” by near
482 infrared (NIR) spectroscopy and chemometrics, J. Food Compos. Anal. 82 (2019) 103235.
483 https://doi.org/10.1016/j.jfca.2019.103235.

484 [26] S. Ghanavati Nasab, M. Javaheran Yazd, F. Marini, R. Nescatelli, A. Biancolillo,
485 Classification of honey applying high performance liquid chromatography, near-infrared
486 spectroscopy and chemometrics, Chemom. Intell. Lab. Syst. 202 (2020) 104037.
487 https://doi.org/10.1016/j.chemolab.2020.104037.

[27] C. Miralbés, Discrimination of European wheat varieties using near infrared reflectance spectroscopy, Food Chem. 106 (2008) 386–389. https://doi.org/10.1016/j.foodchem.2007.05.090.

[28] S.A. Wadood, B. Guo, X. Zhang, Y. Wei, Geographical origin discrimination of wheat kernel and white flour using near-infrared reflectance spectroscopy fingerprinting coupled with chemometrics, Int. J. Food Sci. Technol. 54 (2019) 2045-2054. https://doi.org/10.1111/ijfs.14105.

[29] H. Zhao, B. Guo, Y. Wei, B. Zhang, Near infrared reflectance spectroscopy for determination of the geographical origin of wheat, Food Chem. 138 (2013) 1902–1907. https://doi.org/10.1016/j.foodchem.2012.11.037.

[30] S. Dhanavath, U.J.S. Prasada Rao, Nutritional and Nutraceutical Properties of Triticum dicoccum Wheat and Its Health Benefits: An Overview, J. Food Sci. 82 (2017) 2243–2250. https://doi.org/10.1111/1750-3841.13844.

[31] E. Giambanelli, F. Ferioli, F.L. D'Antuono, Alkylresorcinols and fatty acids in primitive wheat populations of Italian and Black sea region countries origin, J. Food Compos. Anal., 69 (2018) 62-70. https://doi.org/10.1016/j.jfca.2018.02.009.

[32] E. Suchowilska, M. Wiwart, Z. Borejszo, D. Packa, W. Kandler, R. Krska, Discriminant analysis of selected yield components and fatty acid composition of chosen Triticum monococcum, Triticum dicoccum and Triticum spelta accessions, J. Cereal Sci. 49 (2009) 310–315. https://doi.org/10.1016/j.jcs.2008.12.003.

[33] J.U. Ziegler, M. Leitenberger, C.F.H. Longin, T. Würschum, R. Carle, R.M. Schweiggert, Near-infrared reflectance spectroscopy for the rapid discrimination of kernels and flours of different wheat species, J. Food Compos. Anal. 51 (2016) 30–36. https://doi.org/10.1016/j.jfca.2016.06.005.

[34] G.A. Toole, G. Le Gall, I.J. Colquhoun, S. Drea, M. Opanowicz, Z. Bedő, P.R. Shewry, E.N.C. Mills, Spectroscopic analysis of diversity in the spatial distribution of arabinoxylan structures in endosperm cell walls of cereal species in the HEALTHGRAIN diversity collection, J. Cereal Sci. 56 (2012) 134–141. https://doi.org/10.1016/j.jcs.2012.02.016.

[35] N.F. Pérez, J. Ferré, R. Boqué, Calculation of the reliability of classification in discriminant partial least-squares binary classification, Chemom. Intell. Lab. Syst. 95 (2009) 122–128.

https://doi.org/10.1016/j.chemolab.2008.09.005.

[36] N.F. Pérez, J. Ferré, R. Boqué, Multi-class classification with probabilistic discriminant partial least squares (p-DPLS), Anal. Chim. Acta. 664 (2010) 27–33. https://doi.org/10.1016/j.aca.2010.01.059.

[37] P. Firmani, A. Nardecchia, F. Nocente, L. Gazza, F. Marini, A. Biancolillo, Multi-block classification of Italian semolina based on Near Infrared Spectroscopy (NIR) analysis and alveographic indices, Food Chem. 309 (2020) 125677. https://doi.org/10.1016/j.foodchem.2019.125677.

[38] A. Biancolillo, M. Foschi, A.A. D'Archivio, Geographical Classification of Italian Saffron (Crocus sativus L.) by Multi-Block Treatments of UV-Vis and IR Spectroscopic Data, Molecules. 25 (2020) 2332. https://doi.org/10.3390/molecules25102332.

[39] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal. Chim. Acta. 891 (2015) 1–14. https://doi.org/10.1016/j.aca.2015.04.042.

[40] eAmbrosia – the EU geographical indications register, European Commission. https://ec.europa.eu/info/food-farming-fisheries/food-safety-and-quality/certification/quality-labels/geographical-indications-register/#, 2020 (accessed 20 September 2020).

[41] Parco Nazionale Gran Sasso e Monti Della Laga, PAT - Prodotti Agroalimentari Tradizionali. http://www.gransassolagapark.it/dettaglio_prodotto.php?id_prodotti=2512, 2020 (accessed 20 September 2020).

[42] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[43] H. Nocairi, E. Mostafa Qannari, E. Vigneau, D. Bertrand, Discrimination on latent components with respect to patterns. Application to multicollinear data, Comput. Stat. Data Anal. 48 (2005) 139–147. https://doi.org/10.1016/j.csda.2003.09.008.

[44] D. Ballabio, R. Todeschini, Chapter 4-Multivariate Classification for Qualitative Analysis, in: D.-W. Sun (Ed.), Infrared Spectrosc. Food Qual. Anal. Control, Academic Press, Cambridge, 2009, pp. 83–104. https://doi.org/10.1016/B978-0-12-374136-3.00004-3.

[45] D. Lahat, T. Adali, C. Jutten, Multimodal Data Fusion: An Overview of Methods,

Challenges, and Prospects, Proc. IEEE. 103 (2015) 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697.

[46] A. Biancolillo, R. Boqué, M. Cocchi, F. Marini, Chapter 10-Data Fusion Strategies in Food Analysis, in: M. Cocchi (Ed.), Data Fusion Methodology and Applications, Data Handl. Sci. Technol. 31, Elsevier, Amsterdam, 2019, pp. 271–310. https://doi.org/10.1016/B978-0-444-63984-4.00010-7.

[47] T. Næs, R. Romano, O. Tomic, I. Måge, A. Smilde, K.H. Liland, Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects, J. Chemom. (2020). https://doi.org/10.1002/cem.3243.

[48] A. Biancolillo, T. Næs, Chapter 6-The Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions, in: M. Cocchi (Ed.), Data Fusion Methodology and Applications, Data Handl. Sci. Technol. 31, Elsevier, Amsterdam, 2019, pp. 157–177. https://doi.org/10.1016/B978-0-444-63984-4.00006-5.

[49] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, Chemom. Intell. Lab. Syst. 141 (2015) 58–67. https://doi.org/10.1016/j.chemolab.2014.12.001.

[50] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication, Anal. Chim. Acta. 820 (2014) 23–31. https://doi.org/10.1016/j.aca.2014.02.024.

[51] R.D. Snee, Validation of Regression Models: Methods and Examples, Technometrics. 19 (1977) 415–428. https://doi.org/10.1080/00401706.1977.10489581.

[52] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures., Anal. Chem. 36 (1964) 1627–1639. https://doi.org/10.1021/ac60214a047.

[53] S. Wold, E. Johansson, M. Cocchi, PLS: Partial Least Squares Projections to Latent Structures, in: H. Kubinyi (Ed.), 3D-QSAR in drug design, theory, methods, and applications, ESCOM Sci., Leiden, (1993), pp. 523-550.

[54] V. Čurná, M. Lacko-Bartošová, Chemical composition and nutritional value of emmer wheat (Triticum dicoccon schrank): A review, J. Cent. Eur. Agric. 18 (2017) 117–134. https://doi.org/10.5513/JCEA01/18.1.1871.

[55] G. Galterio, P. Codianni, A.M. Giusti, B. Pezzarossa, C. Cannella, Assessment of the agronomic and technological characteristics of Triticum turgidum ssp. dicoccum Schrank and T. spelta L., Nahrung/Food. 47 (2003) 54–59. https://doi.org/10.1002/food.200390012.

[56] V. Giacintucci, L. Guardeño, A. Puig, I. Hernando, G. Sacchetti, P. Pittia, Composition, protein contents, and microstructural characterisation of grains and flours of emmer wheats (Triticum turgidum ssp. dicoccum) of the central Italy type, Czech J. Food Sci. 32 (2014) 115–121. https://doi.org/10.17221/512/2012-CJFS.

[57] A. Barth, Infrared spectroscopy of proteins, Biochim. Biophys. Acta - Bioenerg. 1767 (2007) 1073–1101. https://doi.org/10.1016/j.bbabio.2007.06.004.

[58] F. Huang, H. Song, L. Guo, P. Guang, X. Yang, L. Li, H. Zhao, M. Yang, Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 235 (2020) 118297. https://doi.org/10.1016/j.saa.2020.118297.

[59] P. Yu, H. Block, Z. Niu, K. Doiron, Rapid characterization of molecular chemistry, nutrient make-up and microlocation of internal seed tissue, J. Synchrotron Radiat. 14 (2007) 382-390. https://doi.org/10.1107/S0909049507014264.

[60] L. Amendola, P. Firmani, R. Bucci, F. Marini, A. Biancolillo, Authentication of Sorrento Walnuts by NIR Spectroscopy Coupled with Different Chemometric Classification Strategies, Appl. Sci. 10 (2020) 4003. https://doi.org/10.3390/app10114003.

596 **Figure captions**

597 Fig.1: (A) Geographical origin of the Italian emmer samples. (B) Structure of emmer caryopsis and
598 chemical composition of the different seed tissues

599 Fig.2: Scheme of the SO-PLS-DA algorithm for two predictor blocks

600 Fig.3: NIR and MIR mean spectra of emmer samples averaged according to the class membership
601 and collected analysing both the external (MIRout, NIRout) and the internal part **(**MIRin, NIRin)

602 Fig.4: Probability Density Functions (PDFs) of the classes estimated from the $\hat{Y}$ continuous values
603 in calibration (A). Normalized Probability and posterior probability values of the training (full
604 symbols) and the external (empty symbols) samples (B)

605 Fig.5: (A) 3D plot of the $\hat{Y}$ of each class of the training set (full symbols) and the test set (empty
606 symbols). (B) VIP analysis of MIRout and NIRin/out spectra for the SO-PLS-DA model (9LVs for
607 MIRout and 6LVs for NIRin/out): the black line is the average spectrum; thick red trait points the
608 variables with VIP > 1.

609 **Table**

	Cross-validation					
	Pre-processing	**LVs**	**GA**	**MS**	**GS**	**Total CCR**
MIR_{OUT}	SNV-1st derivative	11	91.9	89.2	79.3	86.8
MIR_{IN}	SNV-1st derivative	13	91.9	83.8	69.0	81.6
$NIR_{IN/OUT}$	1st derivative	10	94.6	100.0	73.3	89.3
L.L. (MIR_{OUT} $NIR_{IN/OUT}$ MIR_{IN})	Optimal pre-proscessing	12	94.6	86.5	79.3	86.8
M.L. (MIR_{OUT} $NIR_{IN/OUT}$)	autoscaling	2	100.0	100.0	100.0	100.0
SO-PLS-DA (MIR_{OUT} $NIR_{IN/OUT}$)	Optimal pre-proscessing	9-6	100.0	100.0	100.0	100.0

	External valiation			
	GA	**MS**	**GS**	**Total CCR**
***1 VS1* MIR_{OUT}**	93.7	93.7	91.7	93.0
MIR_{OUT}	93.7	100.0	83.3	92.3
***1VS 1* MIR_{IN}**	81.2	93.7	83.3	86.1
MIR_{IN}	81.2	87.5	91.7	86.8
***1 VS 1* $NIR_{IN/OUT}$**	100.0	81.2	75.0	85.4
$NIR_{IN/OUT}$	100.0	93.7	66.7	86.8
L.L. (MIR_{OUT} $NIR_{IN/OUT}$ MIR_{IN})	100.0	87.5	75.0	87.5
***1 VS 1* M.L (MIR_{OUT} $NIR_{IN/OUT}$)**	100.0	100.0	83.3	94.4
M.L. (MIR_{OUT} $NIR_{IN/OUT}$)	100.0	100.0	83.3	94.4
SO-PLS-DA (MIR_{OUT} $NIR_{IN/OUT}$)	100.0	100.0	91.7	97.2

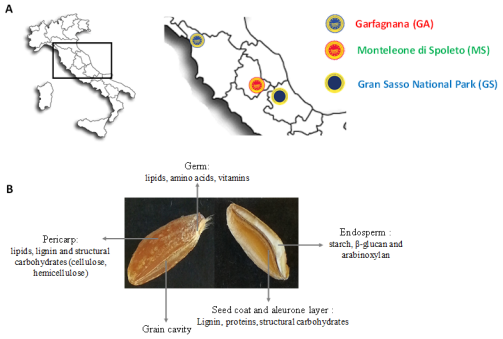
610

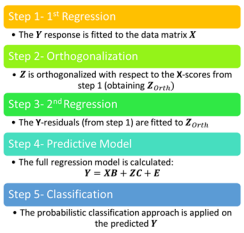
611 Tab.1: Correct Classification Rate (CCR) in cross- and external validation, related complexity
612 (LVs) and optimal pre-processing for the single-block PLS-DA models, for the related "one against
613 one" approach (1*VS*1) and for the multi-block PLS-DA models: Low- Level Data Fusion (L.L.),
614 Mid-Level Data Fusion (M.L.) and Sequential and Orthogonalized-Partial Least Squares-
615 Discriminant Analysis (SO-PLS-DA)

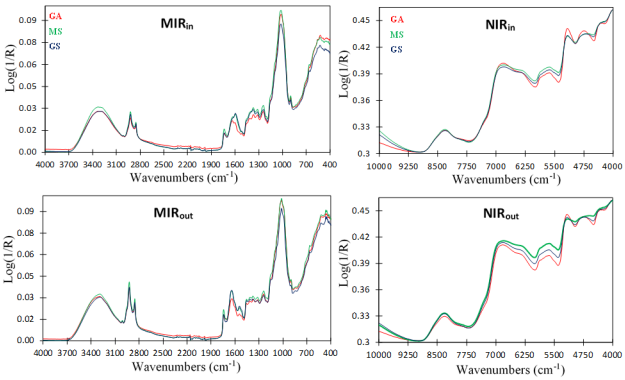
	Cross-validation					
	Pre-processing	**LVs**	**GA**	**MS**	**GS**	**Total CCR**
MIR_{OUT}	SNV-1st derivative	11	91.9	89.2	79.3	86.8
MIR_{IN}	SNV-1st derivative	13	91.9	83.8	69.0	81.6
$NIR_{IN/OUT}$	1st derivative	10	94.6	100.0	73.3	89.3
L.L. (MIR_{OUT} $NIR_{IN/OUT}$ MIR_{IN})	Optimal pre-proscessing	12	94.6	86.5	79.3	86.8
M.L. (MIR_{OUT} $NIR_{IN/OUT}$)	autoscaling	2	100.0	100.0	100.0	100.0
SO-PLS-DA (MIR_{OUT} $NIR_{IN/OUT}$)	Optimal pre-proscessing	9-6	100.0	100.0	100.0	100.0

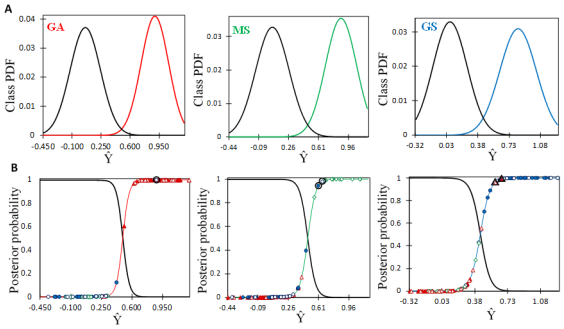
	External valiation			
	GA	**MS**	**GS**	**Total CCR**
1 VS1 **MIR_{OUT}**	93.7	93.7	91.7	93.0
MIR_{OUT}	93.7	100.0	83.3	92.3
1VS 1 **MIR_{IN}**	81.2	93.7	83.3	86.1
MIR_{IN}	81.2	87.5	91.7	86.8
1 VS 1 **$NIR_{IN/OUT}$**	100.0	81.2	75.0	85.4
$NIR_{IN/OUT}$	100.0	93.7	66.7	86.8
L.L. (MIR_{OUT} $NIR_{IN/OUT}$ MIR_{IN})	100.0	87.5	75.0	87.5
1 VS 1 **M.L (MIR_{OUT} $NIR_{IN/OUT}$)**	100.0	100.0	83.3	94.4
M.L. (MIR_{OUT} $NIR_{IN/OUT}$)	100.0	100.0	83.3	94.4
SO-PLS-DA (MIR_{OUT} $NIR_{IN/OUT}$)	100.0	100.0	91.7	97.2

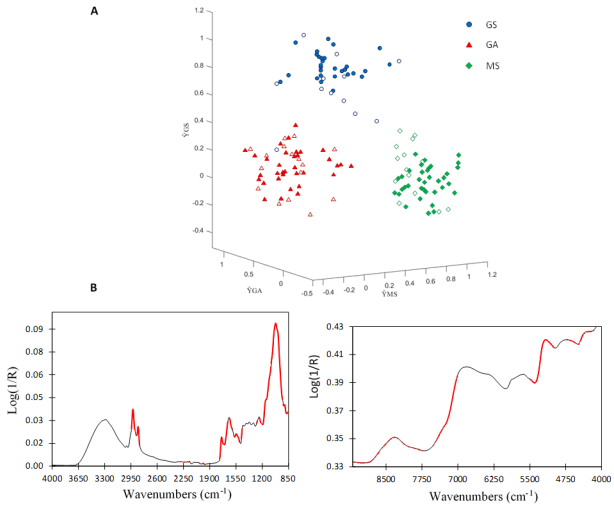
Tab.1: Correct Classification Rate (CCR) in cross- and external validation, related complexity (LVs) and optimal pre-processing for the single-block PLS-DA models, for the related "one against one" approach (1*VS*1) and for the multi-block PLS-DA models: Low- Level Data Fusion (L.L.), Mid-Level Data Fusion (M.L.) and Sequential and Orthogonalized-Partial Least Squares-Discriminant Analysis (SO-PLS-DA)











Highlights

- Geographical discrimination study for the traceability of traditional Italian emmer
- Characterization of emmer by means of Middle and Near Infrared spectroscopy
- Single- and multi-block chemometric elaboration of the spectroscopic profiles
- Development of non-distructive method to asses geographical origin of Italian emmer
- Mid-level data fusion approach achieved an external prediction ability of 97.2%

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: