Classification of wheat varieties by PLS-DA and LDA models and investigation of the spatial distribution of protein content using NIR spectroscopy

Behnam Foroozani[1], Hossein Bagherpour[1*], Nicola Caporaso[2], Khalil Zaboli[3]

(1. Department of Biosystems Engineering, Faculty of Agriculture, Bu-Ali Sina University, 65178-38695, Hamedan, Iran;
2. Department of Agricultural and Food Sciences, University of Naples Federico II, Via Universit à 100, 80055 Portici (NA), Italy
3. Department of Animal Science, Faculty of Agriculture, Bu-Ali Sina University, 65178-38695, Hamedan, Iran)

Abstract: Near-Infrared (NIR) spectroscopy gives information about the chemical properties of objects, and it is of particular interest for agricultural and food science applications given its rapid and non-destructive nature. This paper presents a study on the prediction of protein content in whole wheat kernels by NIR, comparing two statistical discriminant analysis techniques, namely PLS-DA and LDA, to classify wheat varieties and protein levels at the farm's site. Data were collected from three varieties collected from nine farms located in the same region, with a total of 54 samples analyzed. The NIR spectrometer used had a range of 950-1650 nm and it was used to classify different wheat samples according to their varieties and protein level. The optimal spectral pre-processing for the current application was Savitzky-Golay followed by Multiplicative Scatter Correction (SG+MSC), which resulted in R^2 of 0.82 and 0.79 and RMSE of 0.73 and 0.79 for the calibration and validation datasets, respectively. Among the three varieties investigated, only Gaskojhen (Gas) variety had a classification rate above 75%, while the two varieties Mih (Mihan) and Pish (Pishgam) were regarded as one class. Comparing PCA-LDA and PLS-DA, the latter showed better potential in varietal identification compared to PCA-LDA. Investigating the protein changes at different points of the farm revealed that sampling location had a significant effect on the protein content. The ability of NIR to classify different varieties indicates that NIR can be useful in assessing wheat quality, and can give helpful information in varietal identification.

Keywords: NIR spectroscopy, wheat protein, wheat variety identification, protein variability.

Citation: Foroozani, B., H. Bagherpour, N. Caporaso, and K. Zaboli. 2022. Classification of wheat varieties by PLS-DA and LDA models and investigation of the spatial distribution of protein content using NIR spectroscopy. Agricultural Engineering International: CIGR Journal, 24(2): 184-193.

1 Introduction

Wheat is an important source of carbohydrates, and it is the leading source of vegetal protein for human nutrition. It has a protein content of about 12%, which is relatively high compared to other major cereals (Scherf et al., 2016). Wheat is considered as the preferable crop for bread and other flour products because of its baking performance compared to other cereal like barley (Dewettinck et al., 2008). Protein content of wheat grain is one of the most important factors for wheat quality evaluation. It affects the

Received date: 2020-11-09 **Accepted date:** 2022-01-23

***Corresponding author: Hossein Bagherpour,** Assistant Professor, Department of Biosystems Engineering, Faculty of Agriculture, Bu-Ali Sina University, Hamedan, Iran. Tel: +988134425401. Fax: +988134425402. Email: h.bagherpour@basu.ac.r.

technological efficiency in baked products and has a significant impact on wheat price, thus protein has significant economic outcomes for wheat producers and related industries (Caporaso et al., 2018).

In the milling industry, the wheat delivery stage is one of the critical points in obtaining homogeneous flours; a wrong classification of varieties can lead to low homogeneity in the batches produced. Generally, different varieties of wheat have different properties, and within the same variety, if the parameters such as gluten and protein are different, they can be classified into other silos so that the final composition is more uniform, with respect to a specific compound or property. Usually, varietal classification is carried out by visual evaluation, requiring a long, dedicated training process (Miralbés, 2008).

Traditional methods for measuring wheat protein content are usually complex, expensive and relatively tedious and slow. Therefore, simple and economical techniques are needed to speed up the measurement of a large number of samples. Near-infrared reflectance spectroscopy (NIRS) is a nondestructive analytical technique that can be effectively used for classification purposes or for quantification of properties of materials (Abu-Khalaf et al., 2004). This technique is widely used as quality assessment tool in the food, agricultural and pharmaceutical sector, as well as other industries.

NIR spectroscopy measurement of wheat protein content is traditionally done on batches of wheat flours (Osborne, 1984). The possibility of using near infrared (NIR) measurement on whole kernels was demonstrated by Williams & Norris in 1987. Because of nondestructive scanning of whole grain and its speed of analysis, NIR applications are attractive to breeders, engineers and food scientists, who have the possibility of on-line data acquisition (Williams and Sobering, 1993).

Long et al. (2008) evaluated an improved version of an NIR diffuse reflectance instrument that works in the range from 600 to 1100 nm, and the authors applied it for measuring wheat grain protein content during harvest. In this study the spectrometer was evaluated in both laboratory bench and field conditions. In the laboratory bench experiment, the results on white winter wheat samples showed a high level of protein variability (R^2 = 0.91, standard errors of prediction (SEP) =3.1 g kg^{-1}). In the field, the predicted protein content correlated well with the measured reference protein (R^2 = 0.94, SEP = 3.1 g kg^{-1}). Other studies have been conducted to predict other parameters such as dry gluten, wet gluten, biochemical properties, and dough-handling properties of hard red winter wheat by reflectance NIR spectrometry (Delwiche et al., 1998; Miralbés, 2008).

So far, several studies have been conducted to investigate the feasibility of NIR for wheat protein prediction and classification of classes (Miralbés, 2008; Delwiche and Norris, 1993; Delwiche et al., 1995; Mao et al., 2014; Long et al., 2008). These studies generally reported good results. However, for the in-line protein measurement, there are some questions that have been raised about the spatial variability of protein content in the fields, especially in those countries where most fields are smaller than three hectares and there are several varieties cultivated in neighboring regions or when the pedological greatly vary (Sarrafian, 2009). Therefore, the objectives of this study were to: i) develop a classification model to discriminate wheat varieties, ii) investigate the ability of NIR spectroscopy to predict the protein content of wheat samples, and iii) investigate the protein changes at the different points of farms that depending on the location.

2 Materials and methods

2.1 Samples

The study area was situated in Hamedan County (latitude of 35°2104″ and longitude of 49°1929″) in the west of Iran. Wheat is the dominant crop in this region, and the alteration between cultivation and fallow is done mostly in this region. Tests were conducted during the crop season 2017-2018, by using three common varieties namely Mihman (Mih), Gaskojhen (Gas) and Pishgam (Pish). In the 2017 season, at the beginning of the autumn, nine wheat fields were selected. They are located within

approximately 20 meter distance from each other, having a size of at least one hectare and the same soil type (loam sandy). Three experimental fields were randomly assigned to each cultivar.

To measure the available nitrogen in the soil, three soil samples were collected from each field at 6 different locations at 0-15 cm depth of soil profile. After collection, samples were crushed into fine particles using Pestle and Mortar. Then, samples were air-dried, ground, and sieved for calculating physical and chemical properties of soils. The nitrogen content was analyzed by alkaline permanganate and phosphorous and potassium were estimated using flame photometer method (Barrios et al., 1997).

To analyse the spatial variability of protein over the fields, six samples were collected in July 2018 from different locations across each farm such that a total of 54 samples were collected for nine fields. Wheat samples were threshed manually and then each sample (900 g grain) was separated into three fractions to run analytical replicates.

From each sample, about 100 g was ground and the total Nitrogen content was determined using the Kjeldahl method. To remove the effect of moisture on the protein content, all samples were dried by oven at 105℃ until no weights change was measured (AACC, 2000).

2.2 NIR Reflectance measurements

A near infrared spectrophotometer DA7200 (Perten Instruments, Sweden) was used to collect reflectance spectra of whole wheat grain samples. The samples were scanned over the range of 950–1650 nm at 5 nm intervals. Spectra were obtained by averaging three scans and the absorbance spectra were calculated as log (1/R), where R is the relative reflectance, after normalization with a white reference and with the dark current signal. The sample set used in this study was divided into a calibration (36 samples) and a validation sample set (16 samples).

2.3 Pre-processing of the spectral data and statistical analysis

In order to remove some of the noise in the spectra, such as high frequency noise and dispersion, and some of the changes unrelated to chemical composition such as light scattering effects, spectra were pre-processed by using some common pre-processing techniques, including Savitzky-Golay (SG) smoothing (He et al., 2007). The optimal band-width of smoothing windows was considered as 7 nm. The multiplicative scatter correction (MSC) was used to modify the spectra by taking into account the additive and multiplicative effects and to remove the constant off-sets or linear baselines from the spectra. To the same aim, derivatives, namely the first or second derivatives were used. Optimal values of the parameters for the Savitzky-Golay algorithm, the window size, polynomial order, and the derivative order (first or second) were optimized on the basis of standard error of cross-validation of Partial Least Squares (PLS) calibrations.

Calibration and prediction datasets were chosen such that the calibration set contained 70% of the data and the prediction set contained 30% of the data. Cross validation was then applied on the calibration dataset to build the calibration model based on the minimum value of root mean square of error of calibration (RMSEC). This calibration model was subsequently applied to the separate prediction set to evaluate the model performance. The best prediction models have been selected based on the values of root mean square of error of prediction (RMSEP), atio of prediction to deviation (RPD), and the number of the latent variables (LVs). In partial least square - discrimination analysis (PLS-DA), a dummy variable was used as a dependent variable, with values of -1 for Mih, 0 for Pish and +1 for Gas variety. By using PLS as a prediction model, predicted values below -0.5, between –0.5 and 0.5 and the values above 0.5 were assigned to the Mih, Pish and Gas groups, respectively.

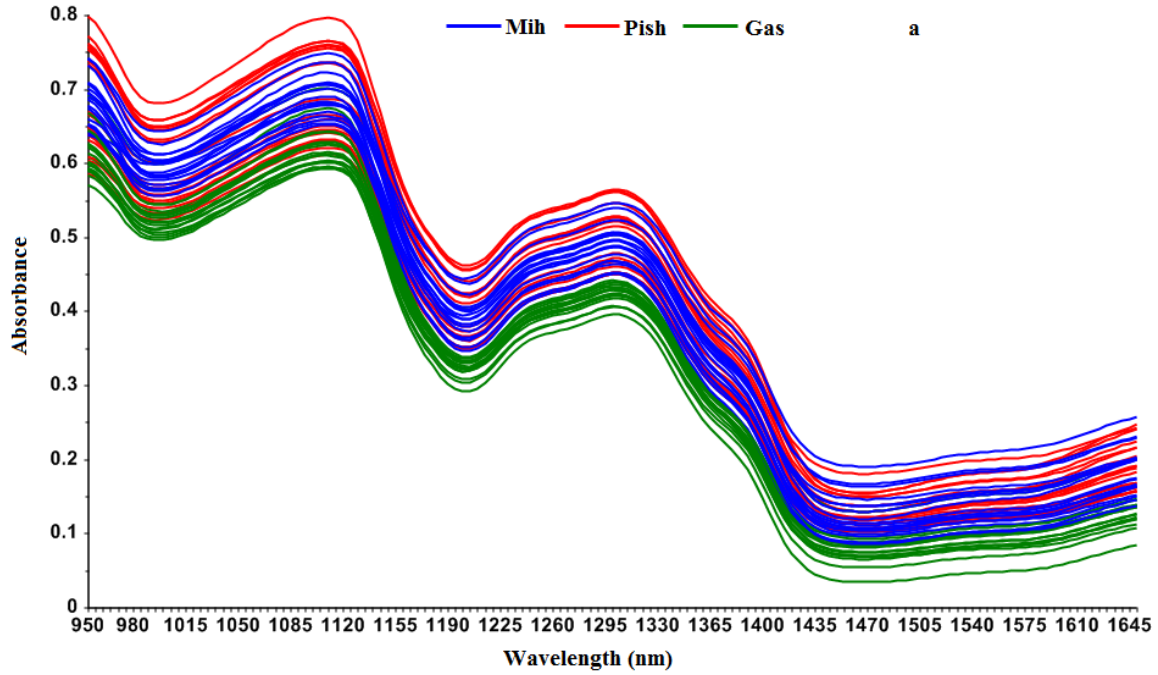
Linear discriminant analysis (LDA) is used for classifying samples into groups based on features that can be used to describe the objects. This could include developing classifications models for a library of products, good vs. bad quality product, or healthy vs. cancerous cells. In this study to prevent over fitting in the LDA model,

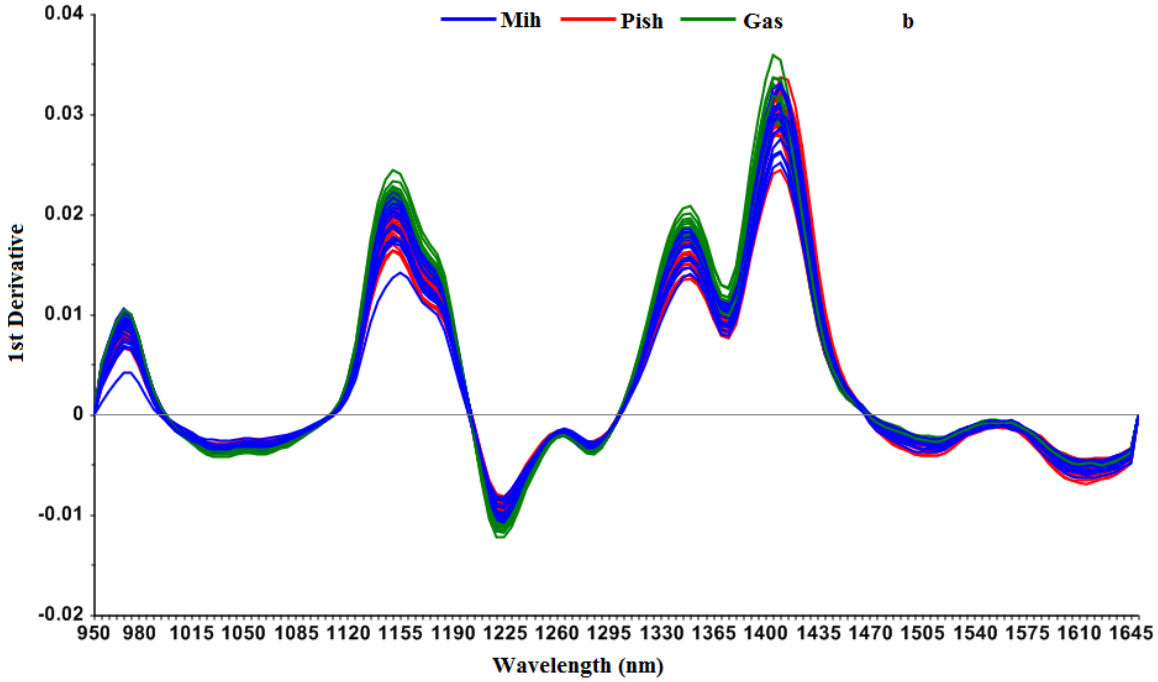
Principal Components Analysis (PCA) scores are used as input variables in LDA algorithm. Confusion matrix was used to assess the accuracy of the proposed model.

3 Results and discussion

3.1 Protein content prediction

The mean NIR reflectance spectrum for each of the three wheat varieties analyzed in the present study is shown in Figure 1(a). This Figure shows that the changes in the spectra are very similar for all varieties, but the mean reflectance spectrum of the Gas variety was higher than the other ones. However, reflectance spectra *per se* is often not a useful indication of possible chemical differences, and absorbance spectra is often used, together with pre-processing methods to remove scattering effects, e.g. derivatives and standard normal variate (SNV). These spectra are shown in Figure 1(b and c).





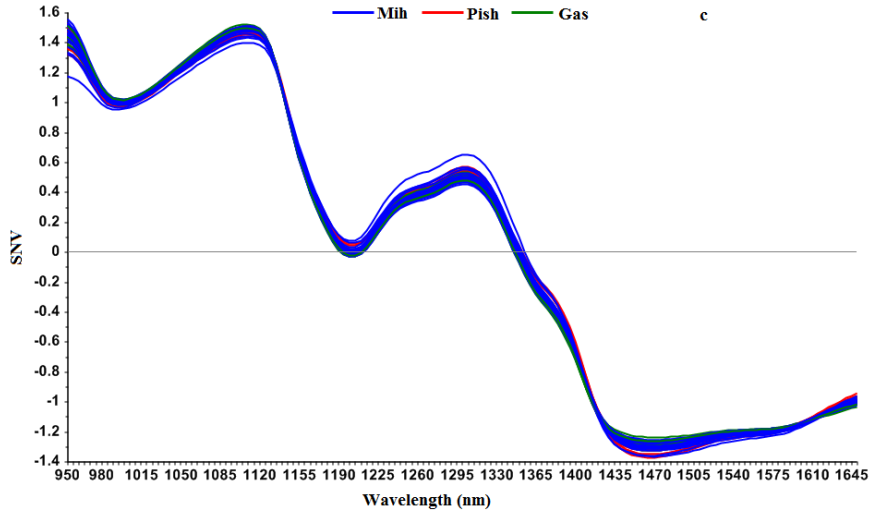


Figure 1 Mean NIR spectra for all wheat samples a) absorbance spectra, b) First derivative and c) Standard normal variate (SNV)

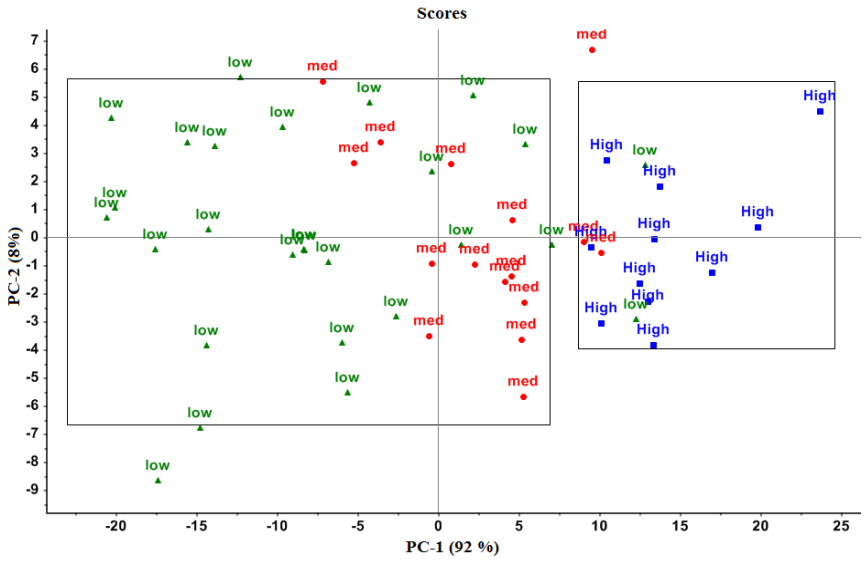


Figure 2 Score plot of NIR reflectance spectra in which the samples have been grouped into the three categories according to their protein levels

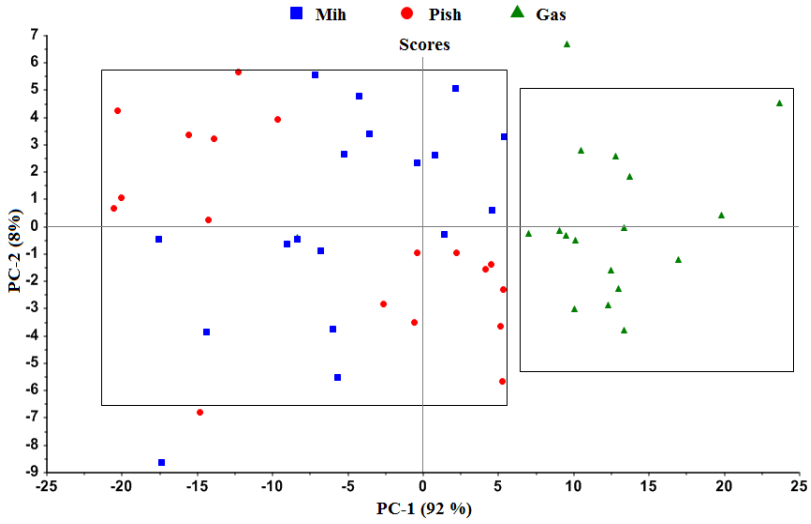


Figure 3 Score plot of the first two principal components (PCs) for spectral data included three wheat varieties

The score plots from the PCA showed similarities among the samples when plotting the first two principal components. Conversely, samples far away from each other are different from each other. Therefore, to investigate the potential of NIR to predict protein content and classify the wheat kernels for their variety, PCA was firstly applied as the unsupervised exploratory tool. The maximum and minimum protein content values from the reference measurement were 16% and 10%, respectively. Thus, kernels were grouped according to their protein content into a group comprising those below 11%, those comprised between 11% and 14%, and those with protein content higher than 14%. It was noticed that samples having high protein content could be potentially detected in the score plot of the Savitzky-Golay and Multiplicative Scatter Correction (MSC) pre-processed NIR reflectance. Most of samples with high protein value were located on the right side of the score plot (Figure 2). In this Figure, although most of the low protein samples were on the left side of score plot, there was no clear distinction between medium and low protein content samples. As shown in Figure 3, it appears that NIR-based prediction has potential to classify wheat varieties, as all Gas variety samples were grouped closely and other were located in the right side of the score plot. This indicates that samples with high protein content are relatively far from other varieties and it is possible to distinguish this variety in terms of their high protein content. Other possible factors influencing the reflectance NIR spectra were reduced or removed, e.g. the difference in moisture content of the samples, which were dried to the same moisture level.

The results of comparing the mean value of total protein content of the three varieties (Figure 4) showed that there was a significant difference between Gas and the other two varieties, but there was no significant difference between Pish and Mih varieties. The results of this test confirms the results of score plot shown in Figure 3.

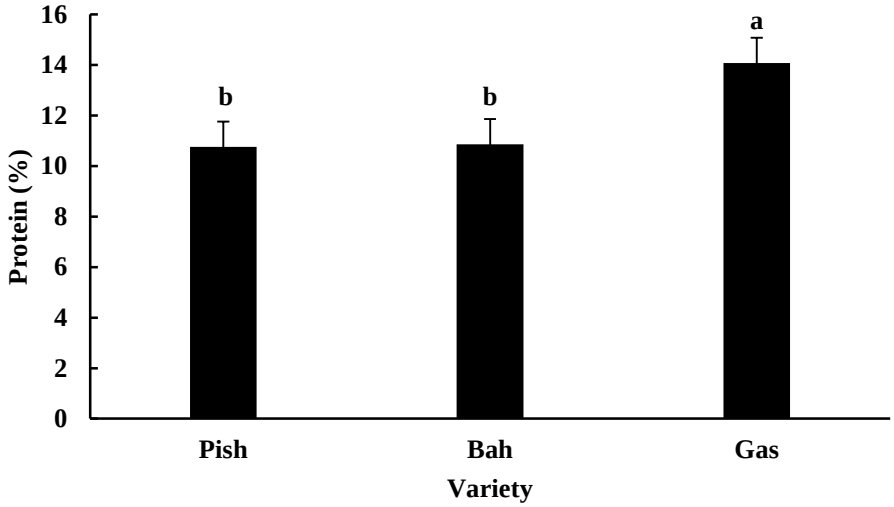


Figure 4 Mean value of protein for different varieties. Different letters indicate a statistical significance difference at $p \leq 0.05$.

Using the NIR spectral data and PLS-DA as the chemometric approach, the whole wheat kernels could be differentiated based on their variety and the level of protein content. Using two dimensional plots of the score vectors for PC1 and PC2, the levels of wheat protein content were separated by these first PCs (Figure 2). PC1 explained about 92% of the variation between samples. Among the several pre-processing techniques tested, the best performance was achieved for SG+MSC, resulting in R^2 values of 0.86 and 0.77 and a RMSE of 0.73 and 0.97, for the calibration and prediction datasets, respectively. The results of some pre-processing and the selected regression model are shown in Table 1, while Figure 5 reports the reference vs. predicted plot of the best prediction model.

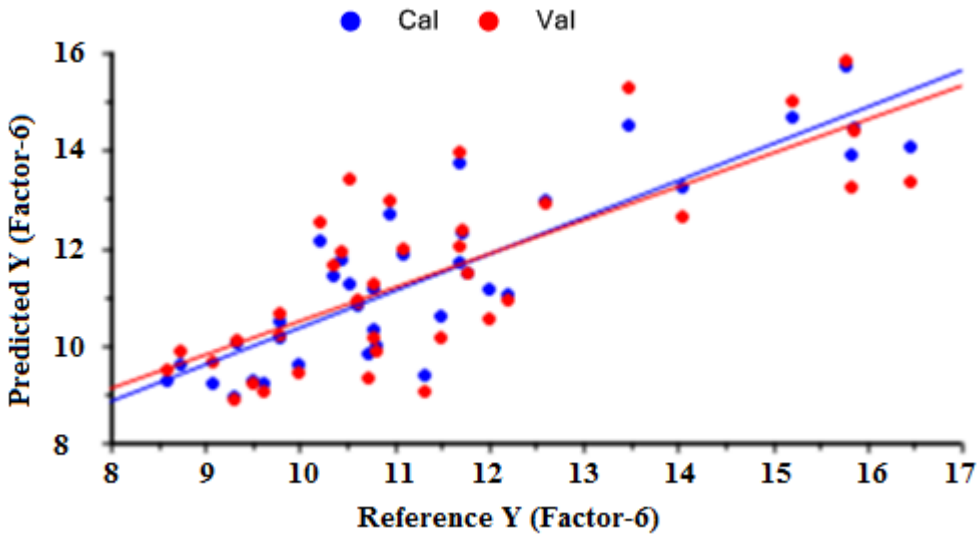


Figure 5 PLS regression model for the calibration dataset in predicting total protein content.

These results are similar to a previous similar study that performed for the quantitative prediction of single wheat kernel protein content using hyperspectral imaging in the spectral region 980–2500 nm (Caporaso et al., 2018). In this case, the best performance was obtained by using standard normal variate (SNV) treatment followed by first derivative with Savitzky-Golay smoothing. The performance for single kernels compares slightly better than the results of this study that was done for bulk samples. Even if the spectrometer used in the present study had a lower spectral range, reaching 1650 nm, while the one used in the previous study could read 2500 nm, the acquisition conditions were significantly different, as Caporaso et al. (2018) acquired single kernels data while they were moving on a moving stage, and acquisition was done by a hyperspectral camera rather than a conventional NIR spectrometer. Also, averaging spectra from the same sample can dramatically improve the resulting spectra by removing some noise and interference, thus ideally improving NIR-based prediction models.

Table1 Performance of PLS regression models for predicting protein content (% DW) of whole wheat using reflectance NIR spectroscopy, depending on different spectra pre-processing

Preprocessing	LV	R_C^2	RMSEC	R_P^2	RMSEP	SDR
SG+2nd Der.+SNV	8	0.783	0.936	0.719	1.382	1.59
MA+1st Der.+SNV	8	0.851	0.786	0.768	0.988	2.02
MA+1st Der.+BL	9	0.854	0.754	0.762	0.966	1.91
SG +MSC	6	0.861	0.736	0.774	0.978	2.18
SG+SNV	8	0.792	0.908	0.727	1.267	1.68
MA+1st Der.	9	0.804	0.844	0.745	1.121	1.77

Note: SG: Savitzky-Golay smoothing treatment; SNV: Standard normal variate; Der.: Derivative; BL: Baseline; MSC: Multiplicative scattering correction; MA: Moving average. LV: Latent variable chosen for the model. RMSEC/RMSEP: Root mean square error of calibration/prediction. SDR: Standard deviation ratio.

3.2 Discriminant analysis for variety classification

Results of discriminant analysis indicated that both PCA-LDA and PLS-DA demonstrated good performance for discrimination of wheat variety. As previously mentioned, Mih and Pish varieties have a very similar mean protein content, which could be a possible reason for which the prediction of 'Mih' and 'Pish' varieties was not successful, whereas 'Gas' variety was correctly classified. The main conclusion to be drawn from this result is that the protein content by itself cannot be used as a reliable factor for discrimination of Mih and Pish varieties. Other features such as texture, or visual characteristics such as color and shape might be better features to be investigate for classifying these varieties (Miralbés, 2008). As reported in Table 2, just Gas variety can be identified well with the classification rate of >75% and two varieties of Mih and

Pish can be assumed as one class.

Table2 Classification rate of PCA-LDA and PLS-DA models for discrimination of three varieties.

Variety	PCA-LDA		PLS-DA	
	Training	Test	Training	Test
Gaskojhen	81%	76%	90%	87%
Mihan	54%	39%	53%	44%
Pishgman	58%	38%	55%	42%

By comparing PCA-LDA and PLS-DA models, it can be concluded that the PLS-DA had good potential in identifying 'Gas' variety compared to PCA-LDA. Confusion and prediction matrices of LDA model showed the classification rates of 81% and 76% for training and test data, respectively. In comparison, PLS-DA model showed a classification accuracy of 90% and 87%, respectively. In examining the capability of models for discriminating the protein levels, the results shown in Table 2 demonstrate better classification accuracy using PLS-DA compared to LDA. These results indicated that both PLS-DA and PCA-LDA models are capable to detect high protein content samples (>14%) with a satisfactory performance for a potential industrial application or screening purposes. However, according to the classification rates, these models have low performance in the classifying of low (<11%) and medium (between 11% to 14%) protein samples. The results of reported in the Table 3 indicate that samples with high protein content, which are located on the right side of the PCA score plots (Figures 2 and 3), are more closely correlated to 'Gas' variety, while the other samples can be related to Mih or Pish varieties.

Table3 Classification accuracy of wheat kernels according to three protein content levels, using PCA-LDA and PLS-DA models

Protein level	PCA-LDA		PLS-DA	
	Training	Test	Training	Test
High	78%	67%	80%	74%
Medium	55%	43%	54%	42%
Low	59%	41%	60%	45%

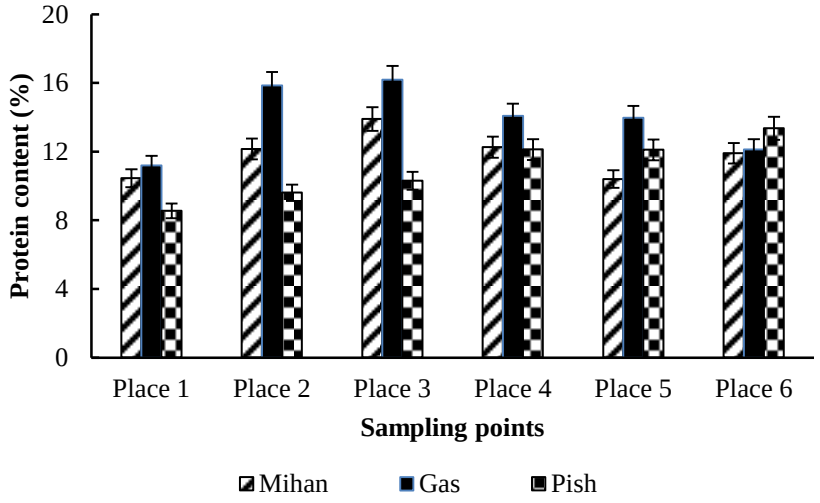


Figure 6 Mean protein content at different points of three selected farms for three varieties

3.3 Spatial variability of whet protein content

Figure 6 illustrates the changes in the protein content at the different points across one farm. Investigation of protein changes at the different points of farm indicated that the sampling location has a significant impact on the protein content. However, comparing the average protein of farms for each variety revealed no significant differences among them. Protein variability within a farm may be explained by temporal or permanent variation in the physical or chemical composition of soils at different

points. These changes are very important for those who want to manage the quality of their products. These changes across the farm are sufficient to justify the use of precision farming technology and NIR spectroscopy for producing yield quantity and quality maps. Long et al. (2008) studied the feasibility of using in-line NIR reflectance spectroscopy to measure wheat protein in a moving grain stream. Although this study evaluated the potential of NIR spectroscopy to measure protein, it had not reported any results about the impact of sampling points on the protein content. The result of this study is also compatible with the study conducted by Wu et al. (2009) who reported that there are significant differences among soil nitrogen content at different places of fields.

4 Conclusion

The present study reports a method for classifying wheat varieties and identifying wheat protein content by NIR spectroscopy. Results indicated that while this method was an appropriate approach for predicting the total protein content, it did not perform satisfactorily in identifying the three studied wheat varieties. Results of classification models indicated that only one variety, namely Gas, could be identified using reflectance NIR spectroscopy. Other varieties could not be separated from each other in terms of their protein content. For discriminating these varieties, other physical features such as color, shape, or size could be potentially be helpful, and further studies need to confirm whether visual differences exist among these varieties. Comparing PLS-DA and LDA, the results of this study showed a better performance for the former than the latter, in terms of classification accuracy with respect to the variety and protein level. Comparing protein differences at different points of the growing fields, the present results indicated that sampling location had a significant effect on the protein content but there was no significant difference between the mean protein content of growing locations for the varieties studied. It should also be noted that a larger number of samples would be needed for a more robust varietal classification model, and this could be done in future studies by also including other varieties of which the classification is of scientific of industrial interest.

References

AACC. 2000. Approved Methods of American Association of Cereal Chemists. 10th ed. St. Paul, Minnesota: AACC.

Abu-Khalaf, N., B. S. Bennedsen, and G. K. Bjørn. 2004. Distinguishing carrot's characteristics by near infrared (NIR) reflectance and multivariate data analysis. CIGR Journal, 6: FP 03012.

Barrios, E., R. J. Buresh, F. Kwesiga, and J. I. Sprent. 1997. Light fraction soil organic matter and available nitrogen following trees and maize. Soil Science Society of America Journal, 61(3): 826-831.

Caporaso, N., M. B. Whitworth, and I. D. Fisk. 2018. Protein content prediction in single wheat kernels using hyperspectral imaging. Food Chemistry, 240: 32-42.

Delwiche, S. R., and K. H. Norris. 1993. Classification of hard red wheat by near-infrared diffuse reflectance spectroscopy. Cereal Chemistry, 70: 29-35.

Delwiche, S. R., Y. R. Chen, and W. R. Hruschka. 1995. Differentiation of hard red wheat by near-infrared analysis of bulk samples. Cereal Chemistry, 72: 243-247.

Delwiche, S. R., R. A. Graybosch, and C. J. Peterson. 1998. Predicting protein composition, biochemical properties, and dough-handling properties of hard red winter wheat by near-infrared reflectance. Cereal Chemistry, 75: 412-416.

Dewettinck, K., F. Van Bockstaele, B. Kühne, D. Van de Walle, T. M. Courtens, and X. Gellynck. 2008. Nutritional value of bread: Influence of processing, food interaction and consumer perception. Journal of Cereal Science, 48(2): 243-257.

He, Y., X. Li, and Y. Shao. 2007. Fast discrimination of apple varieties using Vis/NIR spectroscopy. International Journal of Food Properties, 10(1): 9-18.

Long, D. S., R. E. Engel, and M. C. Siemens. 2008. Measuring grain protein concentration with in-line near infrared reflectance spectroscopy. Agronomy Journal, 100(2): 247-252.

Miralbés, C. 2008. Discrimination of European wheat varieties using near infrared reflectance spectroscopy. Food Chemistry, 106(1): 386-389.

Mao, X., L. Sun, G. Hui, and L. Xu. 2014. Modeling research on wheat protein content measurement using near-infrared reflectance spectroscopy and optimized radial basis function neural network. Journal of Food and Drug Analysis, 22(2): 230-235.

Osborne, B. G. 1984. Investigations into the use of near infrared

reflectance spectroscopy for the quality assessment of wheat with respect to its potential for bread baking. Journal of the Science of Food and Agriculture, 35(1): 106-110.

Sarrafian, M. R. 2009. Country Report of Iran Submitted to Fourth Session of the Technical Committee of APCAEM 10-12 February, Chiang Rai, Thailand. Available at: http://www.unapcaem .org/ Activities Files/ A0902/ ir-p.pdf.

Scherf, K. A., P. Koehler, and H. Wieser. 2016. Gluten and wheat sensitivities–an overview. Journal of Cereal Science, 67: 2-11.

Williams, P. C., and D. Sobering. 1993. Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. Journal of Near Infrared Spectroscopy, 1(1): 25–32.

Wu, J., Y. Liu, D. Chen, J. Wang, and X. Chai. 2009. Quantitative mapping of soil nitrogen content using field spectrometer and hyperspectral remote sensing. In 2009 International Conference on Environmental Science and Information Application Technology, 379-382. Wuhan, China, 4-5 July.