Infrared Physics & Technology 116 (2021) 103757

Contents lists available at ScienceDirect

Infrared Physics and Technology

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

Identifying the best rice physical form for non-destructive prediction of protein content utilising near-infrared spectroscopy to support digital phenotyping

Puneet Mishra [a,*], Mariagiovanna Angileri [b], Ernst Woltering [a,c]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA Wageningen, the Netherlands
[b] CHIBIOFARM, Univesity of Messina, Piazza Pugliatti, 1, 98122 Messina, ME, Italy
[c] Horticulture and Product Physiology Group, Wageningen University, Droevendaalsesteeg 1, P.O. Box 630, 6700AP Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Chemometrics
Food quality
High-throughput
Phenotyping

ABSTRACT

Digital rice phenotyping requires rapid assessment of protein content of rice to kernels to support high-throughput crop phenotyping experiments. A fast and non-destructive approach can allow rapid decision making to breed and select relevant rice varieties. Hence, this study compares the predictive potential of near-infrared (NIR) spectroscopy for three physical forms of rice i.e., rice kernel (with glume), whole grain brown rice and powdered rice. The aim is to identify the best physical form to be adapted in future use for high-throughput protein content prediction in rice samples. The models were optimized by selecting key wavelengths most correlated to the protein content in rice. For variable selection, a total of 8 recently developed chemometric variable selection techniques were used. As a baseline comparison to variable selection techniques, partial-least square (PLS) regression analysis was used. The results showed that for all forms of rice samples, variable selection improved the predictive performance compared to the PLS regression modelling. The best accuracies were obtained for the brown rice samples with a prediction error of 0.349%. Further, this was achieved with only 12 wavelengths compared to the 304 wavelengths available in the original data set. Based on the results, this study indicates that there is no need to grind the rice samples into powder for using NIR spectroscopy. Hence, NIR spectroscopy can directly be used on brown rice samples and can support the rapid assessment of protein content in rice to support digital phenotyping.

1. Introduction

Digital crop phenotyping involves rapid assessment of crop physical and chemical traits to support crop breeders constantly involved in improving the crop varieties [1,2]. To rice, a major interest of breeders is to attain high protein content with limited growing resources [3]. This is because rice yield and quality are closely linked to the rice protein content [4,5]. Digital crop phenotyping aims to understand and identify key genetic characteristics that bring the improvement in protein content in rice [3]. Since the plant breeding experiments are carried out on a large scale, it is often time-consuming and costly to assess protein content of rice samples with destructive wet chemistry analysis [6–8].

In recent years, a wide interest in the use of near-infrared (NIR) spectroscopy for digital crop phenotyping is emerging [2,9]. The NIR spectroscopy is a vibrational spectroscopy technique that deals with the interaction of the infrared spectrum with the material [10]. The main benefit of NIR spectroscopy is that, in a non-destructive way, it captures the overtones of chemical bonds such as OH, CH and NH [10]. The overtones are captured as peaks in the NIR spectra and are modelled with the reference property of interest to develop a calibration model [11]. Once the calibration model is developed then it can be used to replace the wet chemistry analysis. Several applications of NIR spectroscopy can be found related to protein prediction in rice samples [12,13]. However, until this study, no previous study identified the best physical form of rice to measure with NIR spectroscopy. An identification of the best physical form can save extra sample preparation efforts, and hence, the associated cost. For example, if the NIR spectroscopy can be directly used in the intact brown kernels then it can reduce the time needed to grind the rice samples and avoid the complex handling and measurement of rice flour.

* Corresponding author.
E-mail address: puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.infrared.2021.103757
Received 1 January 2021; Received in revised form 18 April 2021; Accepted 20 April 2021
Available online 24 April 2021

Table 1
A summary of reference protein content in calibration and test set.

Chemical component	Calibration (%)	Test (%)
Protein	9.17 + 1.39	8.72 + 1.75

NIR spectra are highly correlated and are made up of overlapping peaks corresponding to overtones of chemical bonds such as OH, CH and NH [14]. To reduce the high-collinearity, latent space modelling approaches that facilitates spectral decorrelation are used. Partial least-square regression[15] analysis is one of popular method to extract the decorrelated variable from the NIR spectra and is commonly used for processing the NIR data. However, PLS regression analysis is based on utilising the complete spectra range (wavelengths), where now it is widely developed fact that pre-selecting key NIR wavelengths before the regression analysis can improve the robustness of NIR models [16,17]. Hence, this study utilizes several variable selection approaches to identify key wavelengths that improve the prediction of protein in rice samples.

The objective of the study is to compare the predictive potential of near-infrared (NIR) spectroscopy for three physical forms of rice i.e., rice kernel (with glume), brown rice and powdered rice. The aim is to identify the best physical form to be adapted in future use for high-throughput protein content prediction in rice samples. The models were optimized by selecting key wavelengths most correlated to the protein content in different physical forms of rice. For variable selection, a total of 8 recently developed chemometric variable selection techniques were used. The eight techniques were bootstrapping soft shrinkage analysis (BOSS) [18], variable combination population analysis (VCPA) [19], variable combination population analysis combined with iteratively retaining informative wavelengths (VCPA-IRIV) [20], competitive adaptive reweighted sampling (CARS) [21], Monte-Carlo uninformative wavelength elimination (MC-UVE) [22], covariates selection (CovSel) [23], variable iteration space shrinkage approach (VISSA) [24] and interval variable iteration space shrinkage approach (iVISSA) [18]. As a baseline comparison to variable selection techniques, partial-least square (PLS) regression analysis was used.

2. Materials and method

2.1. Data set

The data set includes the NIR transmission spectra and reference protein content of 201 rice samples. Out of 201 samples, 143 samples were available in the calibration set and 52 remaining were in the external validation set. The calibration samples were from 25 rice mutant varieties with different protein content. The external validation samples were selected from the 2018 national crop variety regional test materials, with a total of 21 rice varieties. More specific details on the sample varieties can be accessed in [12]. The NIR transmission spectra were measured with MPA Fourier transform NIR spectrometer (Bruker, Germany) in the full spectral range (800–1726 nm), with a spectral resolution of 1 nm. Before NIR measurements, the samples were stabilized in an incubator for 24 h at 25 °C and relative humidity (RH) of 50%. For sample placement, a circular aluminum sheet with a 2 mm diameter hole in the center was fixed in the detection window of the spectrometer [12]. The rice kernel with glume was measured first, then the brown rice (after manually removing glume) followed by the rice flour of the same kernel. The rice flour was obtained by manual grinding using a small mortar and pestle. After the NIR measurements, the rice flour was analyzed for protein content determination using the Dumas combustion method [12]. Each rice flour sample was weighed to 4.0 ± 0.2 mg using the electronic scale (Mettler Toledo, Switzerland) and wrapped in foil into a ball shape, and approximately 5 mg of pure benzene sulfonic acid was weighed and wrapped in foil, as the standard. All samples were analyzed in an elemental analyzer (Elementar,

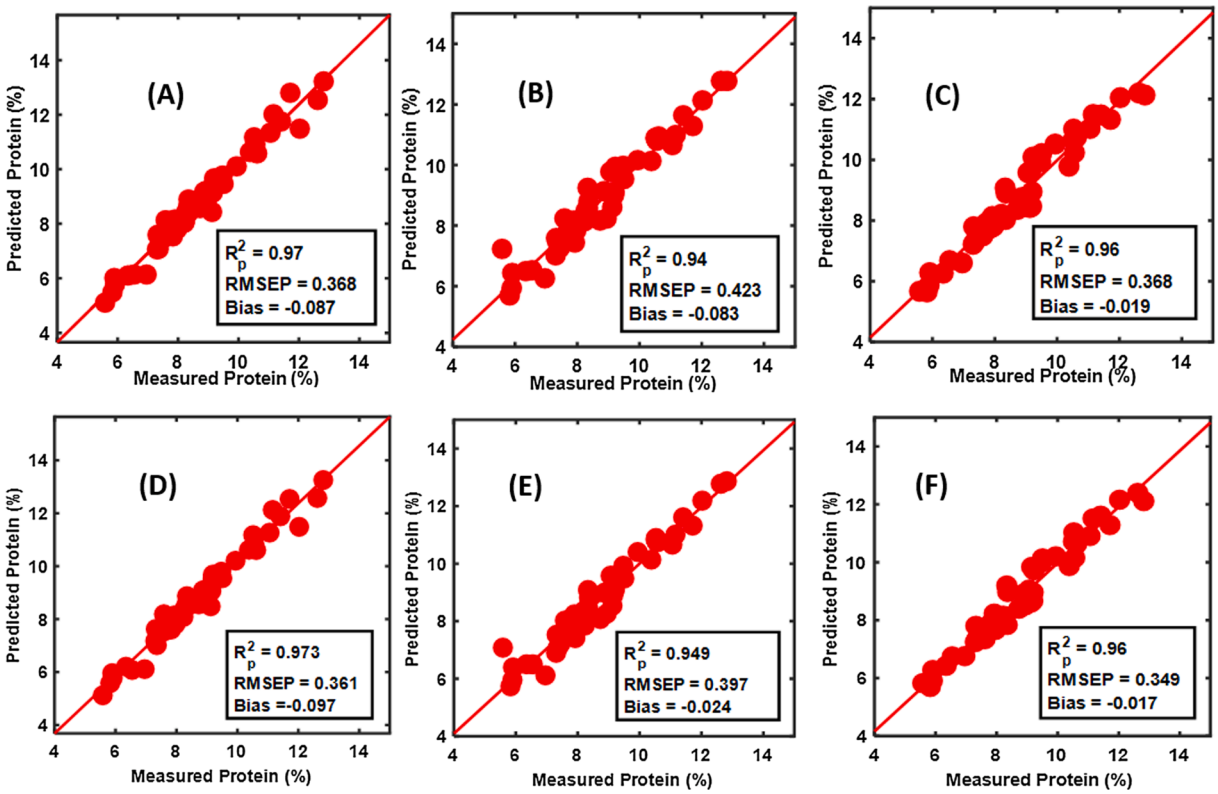

Fig. 1. Summary of protein prediction with partial least-square (PLS) regression and variable selection techniques. PLS test for (A) rice flour, (B) rice kernel (with glume), and (C) brown rice. Variable selection: (D) rice flour with variable combination population analysis-Iteratively retains informative variables, (E) rice kernel (with glume) with covariate selection and (F) brown rice variable combination population analysis.

Table 2
A summary of model performances for protein prediction in rice flour, rice kernel with glume and brown rice. The best performing variable selection technique is highlighted in red and the results from partial least-square regression (PLSR) are highlighted in green. Bootstrapping soft shrinkage analysis (BOSS) [18] (Deng et al., 2016), variable combination population analysis (VCPA), variable combination population analysis combined with iteratively retaining informative wavelengths (VCPA-IRIV), competitive adaptive reweighted sampling (CARS), Monte-Carlo uninformative wavelength elimination (MC-UVE), covariates selection (CovSel), variable iteration space shrinkage approach (VISSA) and interval variable iteration space shrinkage approach (iVISSA).

Technique	Variables			RMSEP (%)		
	Flour	Kernel with Glume	Brown rice	Flour	Kernel with Glume	Brown rice
PLSR	304	304	304	0.368	0.423	-0.019
BOSS	18	45	34	0.374	0.567	0.016
CovSel	10	10	10	0.38	0.397	0.045
CARS	45	58	38	0.417	0.933	-0.041
MCUVE	40	80	109	0.44	0.858	-0.153
VISSA	60	60	109	0.362	0.425	-0.119
iVISSA	127	147	188	0.377	0.796	-0.138
VCPA	11	304	12	0.42	0.452	-0.017
VCPA-IRIV	40	45	31	0.361	0.488	0.066

Germany). The protein content was calculated according to the instrument output of nitrogen (N% × 5.95) [12]. A summary of reference protein content in calibration and test set is described in Table 1. The data set used in this study was the same as used in the study demonstrating the calibration transfer between multiple physical forms of rice [12]. The original data set used in this study is/are accessible at: https://data.mendeley.com/datasets/zvgy65m2rc/1.

2.2. Data analysis

At first, the data were modelled with the PLS regression analysis [15] where the optimal latent variables were optimized with a 5-fold cross-validation procedure. Later, 8 wavelength selection techniques were used for optimizing the models and to select a small subset of wavelengths.

2.2.1. Brief description of wavelength selection techniques

BOSS techniques assigns weights to the wavelengths by generating sub-models with a combination of bootstrap and model population analysis (MPA). Later, key wavelengths are filtered out using a soft-shrinkage [18]. In VCPA [19], a population of sub-models is developed using exponentially decreasing function (EDF) and binary matrix sampling (BMS). Later, subsets of wavelengths carrying lowest error is identified by the MPA [19]. VCPA-IRIV [20] advances VCPA with an iterative selection step to retain informative wavelengths. CARS [21] implements a Monte-Carlo sampling of the regression vector of PLS regression to select a subset of wavelengths. Later, key wavelengths are selected with EDF and adaptive reweighted sampling. MC-UVE [22] evaluates with the stability of the regression coefficients of several PLS models built using random samples and then eliminates the wavelengths with poor stability. CovSel selects the wavelength having maximum covariance with the response(s) and later orthogonalized the data with respect to the selected wavelengths [23]. These two steps are repeated until a pre-defined criterion is met such as the cross-validation error [23]. VISSA generates sub-models using the weighted binary matrix sampling and later the wavelength set is selected based on the variable shrinkage [24]. An extension of VISSA to select wavelength intervals called iVISSA [25] carries out global and local search alternately to iteratively and intelligently optimize the locations, widths and combinations of the spectral intervals.

All the data analysis was performed in MATLAB (2018b, Natick, MA, USA).

3. Results and discussion

A summary of model accuracies for different rice forms are shown in Fig. 1. The RMSEP obtained with PLS regression model for predicting protein content in rice flour (Fig. 1A) and brown rice (Fig. 1C) samples were similar i.e., 0.368% and better than the rice kernel (with glume) (Fig. 1B). A poor performance rice kernel samples could be due to the presence of glume on the kernel which may have affected the NIR spectra. This glume was not present in the brown rice and flour spectra as it was manually removed [12]. The prediction bias for brown rice was lower compared to the rice flour samples, indicating that the protein content in brown rice can be predicted more precisely than the rice flour. A better prediction of protein content in brown rice could be due to less scattering effects in the brown rice samples compared to the rice flour samples. The interaction of NIR light with the powdered samples lead to scattering effects, masking the real absorbance profiles [26]. The results obtained with the best variable selection algorithm for the three different rice forms is also depicted in Fig. 1D, E and F. The variable selection reduced the RMSEP for all the rice forms i.e., flour (Fig. 1D), kernel (Fig. 1E) and brown rice (Fig. 1F). After, variable selection, the RMSEP for the brown rice (0.349%) was the lowest followed by rice flour (0.361%) and rice kernel (0.397%).

Table 3
Summary of selected wavelengths for protein prediction in different rice forms.

Sample type for rice	Total	Wavelengths (nm)
Flour	10	1185.7, 1137.2, 1161.0, 1206.5, 950.2, 975.9, 1219.4, 1249.1, 1043.2, 962.0
Kernel	10	1250.0, 1142.2, 950.2, 1158.0, 1118.4, 1181.8, 1201.5, 1218.4, 985.8, 1003.6
Brown kernel	12	960.1, 978.9, 986.8, 1001.6, 1006.6, 1051.1, 1086.7, 1155.0, 1194.6, 1202.5, 1218.4, 1235.2

A summary of models based on different variable selection approaches for rice flour, rice kernel (with glume) and brown rice are shown in Table 2, respectively. Almost all the variable selection techniques selected a low number of variables and attained either better performance or similar (sometimes slightly higher RMSEP) to that of the PLS regression. In the case of rice flour, the lowest RMSEP was obtained with VCPA-IRIV selected wavelengths. In the case of the rice kernel (with glume), the CovSel selected wavelengths attained the lowest RMSEP and in the case of brown rice, the VCPA selected wavelengths attained the lowest RMSEP. It can be noted that not a single variable selection technique attained the lowest RMSEP for different rice forms, but, different techniques performed differently for each rice forms. However, for all the three rice forms, two techniques i.e., CovSel and VCPA selected the lowest number of variables compared to other techniques.

The wavelengths selected for different rice forms are shown in Table 3. For different rice forms, not the same wavelengths were selected. However, different rice forms have some neighbour wavelengths in common. For example, in the case of the rice kernel (with

glume), wavelength 985.8 nm was selected, and in the case of brown rice, wavelength 986.8 nm was selected. Similarly, in the case of rice flour, wavelength 962 nm was selected, and in the case of brown rice, wavelength 960.1 nm was selected. Such a minute difference in selected wavelengths could be related to the local variabilities in each physical form such as scattering effects. However, in general, the selected wavelengths for different rice forms were related to the second overtones of the protein and peptide groups [12] responsible to predict protein content in rice samples.

4. Conclusions

The study aimed to identify the best rice form to be used for calibration with NIR spectroscopy to predict protein content in rice. The results showed that the brown rice samples attained the lowest RMSEP for protein prediction in rice. The results were confirmed by both the PLS regression modelling as well as wavelength selection modelling. Further, the RMSEP was much lower with the models based on selected wavelengths for all the rice forms. The lowest RMSEP attained for the brown rice samples was 0.349%. Moreover, such a low RMSEP model was obtained with only 12 discrete wavelengths compared to the 304 wavelengths present in the original data. The study finally concludes two messages, the first is that NIR measurements on brown rice samples are sufficient to predict protein content in rice, therefore, grinding the rice and measuring it in flour form is not required, thus, saving time and resources. The second is that identifying key NIR wavelengths can improve the predictive performance of models. The 12 selected wavelengths were 960.1, 978.9, 986.8, 1001.6, 1006.6, 1051.1, 1086.7, 1155.0, 1194.6, 1202.5, 1218.4 and 1235.2 nm.

Declaration of Competing Interest

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.

References

[1] C. Zhao, Y. Zhang, J. Du, X. Guo, W. Wen, S. Gu, J. Wang, J. Fan, Crop Phenomics: current status and perspectives, Front. Plant Sci. 10 (2019) 714.
[2] P. Mishra, S. Lohumi, H. Ahmad Khan, A. Nordon, Close-range hyperspectral imaging of whole plants for digital phenotyping: Recent applications and illumination correction approaches, Comput. Electron. Agric. 178 (2020) 105780.
[3] K. Chattopadhyay, L. Behera, T.B. Bagchi, S.S. Sardar, N. Moharana, N.R. Patra, M. Chakraborti, A. Das, B.C. Marndi, A. Sarkar, U. Ngangkham, K. Chakraborty, L. K. Bose, S. Sarkar, S. Ray, S. Sharma, Detection of stable QTLs for grain protein content in rice (Oryza sativa L.) employing high throughput phenotyping and genotyping platforms, Sci. Rep. 9 (2019) 3196.
[4] M. Martin, M.A. Fitzgerald, Proteins in rice grains influence cooking properties!, J. Cereal Sci. 36 (2002) 285–294.
[5] C.F. Jenner, T.D. Ugalde, D. Aspinall, The physiology of starch and protein deposition in the endosperm of wheat, Funct. Plant Biol. 18 (1991) 211–226.
[6] R. Pieruschka, U. Schurr, Plant Phenotyping: Past, Present, and Future, Plant Phenomics 2019 (2019) 6.
[7] C. Costa, U. Schurr, F. Loreto, P. Menesatti, S. Carpentier, Plant phenotyping research trends, a science mapping approach, Front. Plant Sci. 9 (2019) 1933.
[8] R.T. Furbank, M. Tester, Phenomics – technologies to relieve the phenotyping bottleneck, Trends Plant Sci. 16 (2011) 635–644.
[9] P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, Close range hyperspectral imaging of plants: A review, Biosyst. Eng. 164 (2017) 49–67.
[10] C. Pasquini, Near infrared spectroscopy: A mature analytical technique with new perspectives – A review, Analytica Chim. Acta 1026 (2018) 8–36.
[11] W. Saeys, N.N. Do Trong, R. Van Beers, B.M. Nicolai, Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: A review, Postharvest Biol. Technol. (2019) 158.
[12] Z. Xu, S. Fan, J. Liu, B. Liu, L. Tao, J. Wu, S. Hu, L. Zhao, Q. Wang, Y. Wu, A calibration transfer optimized single kernel near-infrared spectroscopic method, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 220 (2019) 117098.
[13] D.S. Himmelsbach, F.E. Barton Ii, A.M. McClung, E.T. Champagne, Protein and apparent amylose contents of milled rice by NIR-FT/Raman spectroscopy, Cereal Chem. 78 (2001) 488–492.
[14] B.G. Osborne, Near-infrared spectroscopy in food analysis, encyclopedia of, Anal. Chem. (2006).
[15] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[16] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemometrics Intell. Lab. Syst. 118 (2012) 62–69.
[17] T. Mehmood, S. Sæbø, K.H. Liland, Comparison of variable selection methods in partial least squares regression, J. Chemomet. n/a (2020) e3226.
[18] B.-C. Deng, Y.-H. Yun, D.-S. Cao, Y.-L. Yin, W.-T. Wang, H.-M. Lu, Q.-Y. Luo, Y.-Z. Liang, A bootstrapping soft shrinkage approach for variable selection in chemical modeling, Anal. Chim. Acta 908 (2016) 63–74.
[19] Y.-H. Yun, W.-T. Wang, B.-C. Deng, G.-B. Lai, X.-B. Liu, D.-B. Ren, Y.-Z. Liang, W. Fan, Q.-S. Xu, Using variable combination population analysis for variable selection in multivariate calibration, Anal. Chim. Acta 862 (2015) 14–23.
[20] Y.-H. Yun, J. Bin, D.-L. Liu, L. Xu, T.-L. Yan, D.-S. Cao, Q.-S. Xu, A hybrid variable selection strategy based on continuous shrinkage of variable space in multivariate calibration, Anal. Chim. Acta 1058 (2019) 58–69.
[21] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (2009) 77–84.
[22] W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemomet. Intell. Lab. Syst. 90 (2008) 188–194.
[23] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, CovSel: Variable selection for highly multivariate and multi-response calibration: Application to IR spectroscopy, Chemomet. Intell. Lab. Syst. 106 (2011) 216–223.
[24] B.-C. Deng, Y.-H. Yun, Y.-Z. Liang, L.-Z. Yi, A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling, Analyst 139 (2014) 4836–4845.
[25] B.-C. Deng, Y.-H. Yun, P. Ma, C.-C. Lin, D.-B. Ren, Y.-Z. Liang, A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals, Analyst 140 (2015) 1876–1885.
[26] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, TrAC Trends Anal. Chem. (2020) 116045.