UNIVERSIDADE DE SÃO PAULO
CENTRO DE ENERGIA NUCLEAR NA AGRICULTURA

RACHEL FERRAZ DE CAMARGO

Development of a rapid and reliable X-ray fluorescence method for protein determination in soybean grains

Piracicaba
2023

RACHEL FERRAZ DE CAMARGO

Development of a rapid and reliable X-ray fluorescence method for protein determination in soybean grains

Dissertation presented to Center for Nuclear Energy in Agriculture of the University of São Paulo as a requisite to the Ms.Sc. Degree in Sciences

Concentration Area: Chemistry in Agriculture and Environment

Advisor: Prof. Dr. Hudson Wallace Pereira de Carvalho

Piracicaba
2023

Dados Internacionais de Catalogação na Publicação (CIP)

Seção Técnica de Biblioteca - CENA/USP

de Camargo, Rachel Ferraz

Desenvolvimento de um método rápido e confiável por fluorescência de raios X para a determinação da concentração de proteína bruta em grãos de soja / Development of a rapid and reliable X-ray fluorescence method for protein determination in soybean grains / Rachel Ferraz de Camargo; Hudson Wallace Pereira de Carvalho. - -Piracicaba, 2023.

58 p.

Dissertação (Mestrado – Programa de Pós-Graduação em Ciências. Área de concentração: Química na Agricultura e no Ambiente) – Centro de Energia Nuclear na Agricultura da Universidade de São Paulo, 2023.

1. Análise de alimentos 2. Espectrometria 3. Proteínas de plantas 4. Quimiometria 5. Raios X 6. Soja I. Título.

CDU 543.427.4 : 633.34

Elaborada por:
Marilia Ribeiro Garcia Henyei
CRB-8/3631
Resolução CFB Nº 184 de 29 de setembro de 2017

ACKNOWLEDGMENTS

I would like to thank God, for giving me the strength and courage to make this dream come true.

To my parents, Maria Rita and José Luiz, for guiding me, supporting me financially, and being my inspiration to always make the right decisions in my life and to become a better person.

To Prof. Dr. Hudson Wallace Pereira de Carvalho, for the opportunity, rich guidance, dedication, and high support during the research.

To Dr. Tiago Rodrigues Tavares, for always being there for me, and helping me out throughout the whole process.

To Dr. Eduardo de Almeida and MSc. Juliana Graciela Giovannini de Oliveira for the strong support in the laboratory and guidance.

To the staff of the Center for Nuclear Energy in Agriculture (CENA/USP) and to the students of the Laboratory of Nuclear Instrumentation (LIN) for their excellent support during the research.

To the National Nuclear Energy Commission (CNEN, Grant Nº 01341.001296/2021-11) and to the Coordination for the Improvement of Higher Education Personnel (CAPES, Grant Nº 88887.598066/2021-00) for the scholarships.

ABSTRACT

DE CAMARGO, R. F. **Development of a rapid and reliable X-ray fluorescence method for protein determination in soybean grains**. 2023. 58 p. Dissertação (Mestrado em Ciências) - Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, 2023.

X-ray fluorescence spectrometry (XRF) is a technique widely employed for elemental determination. However, the present study pointed to an unconventional direction evaluating the capability of XRF to determine the protein content of soybeans. This was motivated by the perception that soybean might be soon traded based on protein content rather than total grain weight. Additionally, XRF is simply operated, does not require gases or chemicals and the sample preparation and measurements are rapid. The study hypothesizes that sulfur concentration might proxy protein content in soybeans. The research was divided into two parts. Firstly, sample preparation and data acquisition methods were defined and optimized. Briefly, the proposed method consists in (1) coarsely grinding the grains with a household coffee grinder and then (2) scanning the samples for 90 s with the X-ray tube set at 40 kV and 30 µA. Employing 108 samples in the calibration set, a logistic regression model was developed to classify soybean into high- or low-protein groups. The model was validated using an independent set of 54 samples. At validation, the global accuracy and kappa index of the model were 0.81 and 0.61, respectively. The numbers indicate that the technique can be used for classifying soybean based on protein content. In the second part of the research, univariate linear regression, multiple linear regression, and partial least squares regression (PLS) models were established to evaluate the feasibility of quantifying the attribute. The models presented reasonable predictive performance (RPD > 1.57) and PLS presented the highest performance ($R^2 = 0.73$) at the validation, suggesting that the XRF technique can be used for rough screening applications. Additionally, samples prepared by mixing soybeans with soybean flours were added in the calibration (22 samples) and validation (10 samples) sets to widen the protein range. The protein content range was 33.8% - 43.9% and changed to 19.2% - 54% after including the mixtures. In this scenario, higher R^2 values were obtained (e.g., $R^2 = 0.89$ for PLS), confirming that protein can be predicted from XRF data. The hypothesis that sulfur proxies the protein content in soybeans was confirmed by the present study, since the sulfer emission line was the most important variable for prediction, regardless of the modeling strategy used.

Keywords: XRF. Food analysis. Chemometrics.

RESUMO

DE CAMARGO, R. F. **Desenvolvimento de um método rápido e confiável por fluorescência de raios X para a determinação da concentração de proteína bruta em grãos de soja**. 2023. 58 p. Dissertação (Mestrado em Ciências) - Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, 2023.

A espectrometria de fluorescência de raios X (XRF) é uma técnica amplamente utilizada para determinação elementar. Contudo, o presente estudo apontou uma direção não convencional, ao avaliar o desempenho da espectrometria de XRF para a determinação da concentração de proteína na soja. Este estudo foi motivado pela percepção de que a soja poderia em breve ser comercializada com base na concentração de proteína, em vez do peso total dos grãos. Além disso, o equipamento é de simples operação, não requer gases ou reagentes nocivos, e o preparo das amostras e as medições são rápidas. O estudo pressupõe que a concentração de enxofre pode ser utilizada para estimar a concentração de proteína na soja. A pesquisa foi dividida em duas partes. Na primeira, foram definidos e otimizados os métodos de preparo de amostra e de aquisição de dados. Resumidamente, o método proposto consiste (1) na moagem grosseira dos grãos com um moinho de café doméstico e (2) na realização de medidas de XRF de 90 s e com o tubo de raios X configurado em 40 kV e 30 µA. Com 108 amostras no conjunto de calibração, desenvolveu-se uma regressão logística para classificar a soja em grupos de alta ou baixa concentração de proteína. Testou-se o modelo com 54 amostras (conjunto de validação). Na validação, a acurácia global e o índice kappa do modelo foram 0,81 e 0,61, respectivamente. Os números indicam que a técnica pode ser utilizada para classificar a soja com base no teor de proteína. Na segunda parte da pesquisa, foram desenvolvidos modelos de regressão linear simples, regressão linear múltipla e regressão por quadrados mínimos parciais (PLS) para avaliar a potencialidade da espectrometria de XRF na quantificação do atributo. Os modelos apresentaram na validação desempenhos preditivos razoáveis (RPD > 1,57) e entre eles o PLS apresentou o melhor desempenho (R^2 = 0,73). Os resultados sugerem que o sensor pode ser utilizado para estimar a concentração de proteína na soja. Além disso, foram adicionadas 22 amostras no conjunto de calibração e 10 amostras no conjunto de validação, as quais foram preparadas misturando-se soja com farinhas de soja, com o objetivo de aumentar a faixa de concentração de proteína dos conjuntos. A faixa de concentração de proteína foi alterada de 33,8% - 43,9% para 19,2% - 54.0%, com essa inclusão. Neste cenário, foram observados valores mais elevados de R^2 (e.g., R^2 = 0,89 para PLS), confirmando que a concentração do atributo pode ser determinada com dados de XRF. A hipótese de que o enxofre pode ser

utilizado para estimar a concentração de proteína na soja foi confirmada no presente estudo, uma vez que a linha de emissão do enxofre foi a variável mais importante para prever proteína, independentemente da estratégia de modelagem utilizada.

Palavras-chave: XRF. Análise de alimentos. Quimiometria.

FIGURE LIST

Figure 2.1 - Sample preparation methods: (A) The measurements were performed directly on the soybeans; (B) on the soybeans peeled by hand; (C) on the soybeans ground for 60 s using a household coffee grinder (Cadence, model MDR302, Balneário Piçarras, SC, Brazil); (D) and on the soybeans ground using a cryogenic mill (Spex Sample Prep, Freezer/Mill 6870, Metuchen, NJ, USA), 5 grinding cycles of 2 min, each with 1 min of cooling between cycles and a pre-cooling time of 5 minutes 23

Figure 2.2 - (A) Spectral profile of a single sample acquired with a dwell time of 30 and 90 s; (B) CV of the Kα emission line intensities and (C) SNR for each equipment setting and dwell time. SNRs with different letters (a–b) are significantly different (p-value < 0.05) according to the variance analysis (ANOVA) 26

Figure 2.3 - Effect of different sample preparation methods on soybean spectrum (A) and intensities of the Kα emission lines and their respective coefficients of variation (labels displayed above the bars) (B). Results obtained by Tukey's test and means with different letters (a-d) are significantly different (p-value < 0.05) 27

Figure 2.4 - Binary logistic regression performance indicators for the training (A) and test (B) datasets and scatter plots of the probability of occurring the event (i.e., classification of a high-protein sample) versus the real protein content (reference analysis). Where true positive (TP) and true negative (TN) correspond to the number of samples that have respectively high and low protein content and were correctly classified as such; False negative (FN) and false positive (FP) respectively correspond to the samples incorrectly classified as high- and low-protein samples 30

Figure 3.1 - (A) Spectral profile of a soybean sample before (loose powder) and after pressed (2811 Pellet Press, Parr Instrument Company, Moline, IL, USA). (B) Kα peak intensities (labels inside the bars) and coefficients of variation (labels above the bars) of measurements performed on the loose powder and pressed powder sample. Means with different letters (a-b) are significantly different (p-value < 0.05) according to the variance analysis (ANOVA) 38

Figure 3.2 - (A) Summary of the modeling strategies applied to evaluate the predictive performance of XRF data for determining soybean protein content. Scenario A used only soybean samples and scenario B used soybean and blended samples 41

Figure 3.3 - Predictive performances achieved with scenarios A and B. The predicted and measured protein content is expressed in % or g/100g on a dry matter basis. ULR: univariate linear regression; MLR: multiple linear regression; stepwise MLR: multiple linear regression combined with backward stepwise variable selection; PLS: partial least squares regression; Cal and Val: refer to the calibration and validation samples, respectively; R^2: coefficient of determination; RMSE: root-mean-square error; RPD: ratio of performance to deviation. Rec 1: recovery of the model for the CRM LRI09091 reference material and Rec 2: recovery of the model for the NIST 3234 certified reference material 44

Figure 3.4 - Weighted standardized regression coefficients of the PLS models, calibrated with soybean samples (Scenario A) and soybean+blends (Scenario B), using in the models three and five latent variables, respectively 45

Figure 3.5 - Spectrum of soybean fiber flour, defatted soybean flour and soybean 47

TABLE LIST

Table 2.1 - Descriptive statistics of the soybean protein content and % of samples/class. 29

Table 2.2 - Ranges reported for protein content in soybeans 29

Table 3.1 - Proportion (% w/w) of soybean fiber flour (SFF), defatted soybean flour (DSF) and soybean[1] in the blended samples 36

Table 3.2 - Summary of the results: pre-processing techniques tested for PLS (scenario A). 40

Table 3.3 - Descriptive statistics for protein content of the calibration and validation datasets for both scenarios 41

Table 3.4 - Importance of Kα emission lines for protein content prediction on both datasets (soybeans and soybeans+blends). The values correspond to z-score standardized regression coefficients of the ULR and MLR models. 45

SUMMARY

1. INTRODUCTION 15
1.1. Hypothesis 16
1.2. Objectives 16
1.2.1. Specific objectives 17
1.3. Structure of the dissertation 17
2. SOYBEAN SORTING BASED ON PROTEIN CONTENT USING X-RAY FLUORESCENCE SPECTROMETRY 19
2.1. Introduction 20
2.2. Material and Methods 22
2.2.1. XRF device and dwell time evaluation 22
2.2.2. Definition of the sample preparation 23
2.2.3. Soybean samples, determination, and classification of its protein content 23
2.2.4. Data modeling with binary logistic regression 24
2.3. Results and Discussion 26
2.3.1. Effect of dwell time on the XRF data 26
2.3.2. Effect of sample preparation on XRF data 27
2.3.3. Reference analysis of protein content 28
2.3.4. Performance of the logistic regression model 29
2.4. Conclusion 32
3. QUANTIFYING SOYBEAN PROTEIN CONTENT BY X-RAY FLUORESCENCE SPECTROMETRY 33
3.1. Introduction 34
3.2. Material and Methods 36
3.2.1. Soybean samples 36
3.2.2. Blends 36
3.2.3. Total protein content 36
3.2.4. X-ray fluorescence spectrometry analysis 37
3.2.5. Data Modeling 38
3.2.5.1. Univariate and multiple linear regressions 39
3.2.5.2. Partial least squares regression (PLS) 39

3.3. Results and discussion 41
3.3.1. Descriptive statistics 41
3.3.2. Prediction performances 42
3.3.2.1. Models calibrated with soybean samples (scenario A) 42
3.3.2.2. Models calibrated with soybean and blended samples (Scenario B) 46
3.3.3. Challenges and future applications 47
3.4. Conclusion 48
4. FINAL REMARKS 49
REFERENCES 51
Apêndice A: Submitted publication 58

1. INTRODUCTION

Brazil is the world's largest soybean producer (USDA, 2023) and exporter (ATLAS, 2020). According to the latest report from the USDA, Brazilian soybean production (2020/21) was 139.5 million metric tonnes, representing up to 37.8% of the global production (USDA, 2023). The country is also the main exporter of this commodity. According to a survey carried out by the Growth Lab of Harvard University in 2020, Brazil was responsible for 44.4% of the world trade, which in monetary terms is equivalent to US$ 27.9 billion (ATLAS, 2020).

One problem that has affected the soybean business is the decline in protein content observed over the years. This has been a complaint in the animal feed market as the grains are used to produce soybean meal (LANDGRAF, 2015). This decline (ca. 2% per decade) has occurred because seed breeding companies, whose incentives are those pointed out by the market, have prioritized crop yield over grain quality, and grain yield holds an inverse relationship with protein content (UMBURANAS *et al*., 2022). Environmental (soil nutrient availability, temperature, and precipitation) and geographical (latitude and altitude) conditions are also known to affect the protein content of soybeans (PÍPOLO *et al*., 2015).

A country is well prepared for the future if it can predict the needs of the population and technological trends, to contribute to the economy. Since markets tend to become more efficient with time, it is reasonable to believe that soon soybean will be valued for its protein content, rather than weight. Today, soybeans are traded on the market per ton delivered, regardless of protein content. This trend is similar to what recently happened with sugar cane. Initially traded by weight, it is now valued by quality, e.g., based on the total recoverable sugar content (SACHS, 2007).

X-ray fluorescence spectrometry is an analytical technique employed to identify and determine the concentration of elements (e.g., S, P, K, Ca, Fe, Mn, and Zn) in several materials (MARGUI *et al*., 2022). It has been applied in many fields, such as food, agricultural, environmental sciences (FENG; ZHANG; YU, 2021). The main advantages of this technique are: (i) it requires minimal sample preparation, i.e., the samples do not need to be digested; (ii) it is environmentally friendly as it avoids hazardous reagents and consequent chemical waste disposal issues; and (iii) the equipment is simple to use and the measurements are rapid and accurate (MARGUI *et al*., 2022).

The technique can straightforwardly quantify chemical elements with atomic numbers above 12, as these elements have significant fluorescence yield (VAN GRIEKEN; MARKOWICZ, 2001. Although extracting information about light elements, such as hydrogen,

oxygen, nitrogen, and carbon is challenging, some studies explored the Rayleigh and Compton scattering peaks of the XRF spectrum to develop methods for determining organic compounds, e.g., sucrose (ALEXANDRE; GORAIEB; BUENO, 2010; MELQUIADES *et al.*, 2012), alcohol, fixed acidity (ALEXANDRE; GORAIEB; BUENO, 2010), fiber (MELQUIADES *et al.*, 2012), protein and carbohydrate (TERRA, 2009). This is possible because the scattering peaks are related to the average atomic number of the sample (BUENO *et al.*, 2005; VERBI; PEREIRA-FILHO; BUENO, 2005; ALEXANDRE; GORAIEB; BUENO, 2010; MELQUIADES *et al.*, 2012). These methods were developed utilizing multivariate calibration techniques, such as partial least squares regression, to associate a property of interest, obtained by a standard analytical method, with XRF spectral data.

1.1. Hypothesis

The hypothesis of this study were:

i) S Kα emission line and scattering peaks of soybean XRF spectra can be used as proxies for protein content;

ii) XRF spectrometry can be applied to classify soybeans into high- and low-protein categories;

iii) Soybean protein content can be determined using XRF spectrometry.

1.2. Objectives

The general objective of this research is to develop modeling approaches to explore the X-ray fluorescence spectrum (i.e., scattering peaks, as well as the sulfur emission line), to predict soybean protein content. The sulfur emission line is particularly important because the element is part of the methionine and cysteine amino acids structure of storage proteins. Hence, we seek to develop rapid and reliable methods for inferring soybean protein content, using XRF spectrometry.

1.2.1. Specific objectives

In chapter 2, the specific goals are:

(i) Establish sample preparation and data acquisition methods for the XRF measurements;

(ii) Develop a logistic regression model for classifying soybean as high- or low-protein, using the XRF spectra and protein contents of a large number of samples.

In chapter 3, the specific goals are:

(i) Evaluate the feasibility of using XRF data to quantify soybean protein content;

(ii) Compare predictive performances of different data modeling strategies (univariate linear regression, multiple linear regression, and partial least squares regression) for the proposed application of the XRF sensor.

1.3. Structure of the dissertation

Chapters two and three are manuscripts submitted to the *Food Chemistry* journal, entitled "Soybean sorting based on protein content using X-ray fluorescence spectrometry" and "Quantifying soybean protein content by X-ray fluorescence spectrometry", respectively. The first manuscript was recently published and is available at https://doi.org/10.1016/j.foodchem.2023.135548.

2. SOYBEAN SORTING BASED ON PROTEIN CONTENT USING X-RAY FLUORESCENCE SPECTROMETRY[1]

Abstract

The purpose of this research was to evaluate performance of an energy-dispersive X-ray fluorescence (XRF) sensor to classify soybean based on protein content. The hypothesis was that sulfur signals and other XRF spectral features can be used as proxies to infer soybean protein content. Sample preparation and equipment settings to optimize detection of S and other specific emission lines were tested for this application. A logistic regression model for classifying soybean as high- or low-protein was developed based on XRF spectra and protein contents. Additionally, the model was validated with an independent set of samples. Global accuracy of the method was 0.83 (training set) and 0.81 (test set) and the corresponding kappa indices were 0.66 and 0.61, respectively. These numbers indicated satisfactory performance of the sensor, suggesting that XRF spectral features can be applied for screening protein content in soybean.

Keywords: food analysis; XRF; Dumas; logistic regression; machine learning algorithms; chemometrics.

[1] Camargo, R.F. de; Tavares, T. R.; Silva, N.G. da C. da; Almeida, E. de; Carvalho, H.W.P. de. Soybean sorting based on protein content using X-ray fluorescence spectrometry. **Food Chemistry**, 2023. (submitted).

2.1. Introduction

Soybean is used in the manufacture of animal feed because of its high protein content (SUDARIĆ, 2020). Brazil is an important player in this scenario as the world's largest soybean producer, with about 138 million metric tonnes produced in 2020/21 (COMPANHIA NACIONAL DE ABASTECIMENTO, 2022). However, a reduction in Brazilian soybean protein content has been observed over the years (UMBURANAS *et al.*, 2022). The decline in protein level occurred as seed breeding companies prioritized crop yield, which is known to have an inverse relationship with protein concentration (UMBURANAS *et al.*, 2022).

It is predicted and reasonable to believe that soon soybean will be valued for its quality (UPDAW; BULLOCK; NICHOLS, 1976), i.e., it might be traded based on protein content rather than total weight. A similar move happened in recent years with sugarcane economic valuation. Instead of being paid by total weight, it is today valued for its quality, predominantly based on the total recoverable sugar content (MELO, 2015). This change has even raised a high demand for means of measuring the sugar quality in the field. Therefore, efforts have been made to incorporate sensor systems, including X-ray fluorescence (XRF) hardware, in sugar cane harvesters for measurements of quality attributes (CORRÊDO *et al.*, 2021).

XRF is a direct analytical technique used to evaluate a broad range of elements in samples, with little or no preparation (RODRIGUES *et al.*, 2018; MARGUÍ; QUERALT; ALMEIDA, 2022). Due to all these upsides, this technique has been used to evaluate the composition of soybean grains. When excited by a primary X-ray beam, atoms can emit photons at a characteristic energy, due to the renowned photoelectric effect (JENKINS, 1999). Every element with an atomic number above 12 can be identified by XRF analysis, as each has its characteristic X-ray photon emission profile (GRIEKEN; MARKOWICZ, 2001). Thus, the evaluation of XRF spectra allows qualitative and quantitative evaluation of the elements that compose the sample, the latter is possible because the intensity of emitted photons is proportional to the amount of emitting atoms (JENKINS, 1999).

It is worth mentioning two processes occurring and registered in the XRF spectrum when the X-ray interacts with matter: (i) absorption, which can lead to fluorescence emission as described above, and (ii) scattering, which generates the Thomson and Compton scattering peaks (GRIEKEN; MARKOWICZ, 2001). The intensity of scattering peaks being directly related to the average atomic number of the sample (JENKINS, 1999), can provide useful information about light elements (VERBI; PEREIRA-FILHO; BUENO, 2005), such as oxygen,

hydrogen, and carbon, whose fluorescence emission is not detectable under the instrumental conditions of portable XRF equipment (MELQUIADES *et al*., 2012).

Thus, our starting working hypothesis was that XRF data, specifically the sulfur emission line and scattering peaks, can be used as proxies for the classification of soybeans in terms of protein content. This assumption was driven by the fact that sulfur is a common element in the peptide chain of soybean proteins, more precisely methionine and cysteine amino acids. Additionally, the presence of other non-sulfur organic molecules could be inferred from the scattering peaks. To the best of our knowledge, the classification of soybean by XRF analysis has not yet been tested in the literature.

The nearest work addressing the aforementioned hypothesis is described in the doctoral thesis of Terra (2009), where X-ray fluorescence equipment was employed for protein quantification in soy-based products (flour, fiber, extract, and protein-enriched powder), reporting promising performance for this application (R^2 of 0.86 and RMSEP of 2.83%). The author obtained these results using 12 samples and applying partial least squares regression for modeling. The models were built using the intensity of the entire spectrum as explanatory variables, pointing out the scattering peaks as the most important regions of the model. Despite the good results found by this pioneering study, it is well known that multivariate calibrations require larger databases (e.g., n>50) to be robust (TANG *et al*., 2017), especially for indirect predictions such as protein content. In addition, Terra (2009) used a variety of soy-based products that have differences in composition and granulometric aspects (powder, paste, and granule), which can lead to distinct matrix effect, impacting the fluorescence emission (MARUYAMA *et al*., 2008; SAPKOTA *et al*., 2019). Therefore, it is recommended to standardize the sample preparation.

In light of this previous work, the present study was carried out using a larger database and standardized sample preparation. Hence, the present study aimed to answer the following questions: (i) what is the best sample preparation and instrumental condition to obtain accurate XRF data on soybeans? (ii) Is it possible to classify/sort soybeans into high and low protein content categories using X-ray fluorescence spectrometry?

2.2. Material and Methods

2.2.1. XRF device and dwell time evaluation

The measurements were performed with a portable XRF spectrometer (Tracer III–SD, Bruker AXS, Madison, WI, USA), furnished with Rh anode X-ray tube (4 W power) and X-Flash® Peltier-cooled Silicon Drift Detector (Bruker AXS, Madison, WI, USA). The X-ray tube was set at 40 kV and 30 μA, a condition that maintained the detector deadtime around 30%. The above settings were selected to favor the Rh Kα Compton and Thomson scattering peaks in the XRF spectra. All measurements were carried out under a vacuum condition and without a primary filter.

Two scenarios of dwell time were evaluated: for 30 s and 90 s. The following criteria were evaluated to select the optimal dwell time for the subsequent analyses: (i) higher signal-to-noise ratio (SNR) and (ii) lower coefficient of variation (CV) for the net intensities. The evaluation of the dwell time scenarios was performed with soybeans prepared using the coarse grinding method, as described in Section 2.2.2. For the measurements, 3 g of the sample was loaded into a 31 mm diameter XRF cup (n. 1530, Chemplex Industries Inc., Palm City, FL, USA). The XRF cup was sealed in the bottom with a 6 mm thick polypropylene film (VHG Labs, Manchester, NH, USA). This procedure was replicated five times and scanned using both dwell times, totalizing 10 measurements.

The spectra were obtained with the Bruker S1PXRF® software (Bruker AXS, Madison, WI, EUA). The net and background emission line intensities were determined using the Bayesian deconvolution process as per the Artax® software (Bruker AXS, Madison, WI, EUA). The net and background intensities were normalized by the detector live time, being reported in counts of photons per second (cps). The noise (N) was determined as the square root of the background intensity (BG, cps) divided by the detector live time, t(s), given by the (Eq. 1).

$$N = (BG\ /t\ )^{1/2} \quad (1)$$

Subsequently, the SNR (unitless) was calculated by the ratio between the net intensity of the elemental emission line (cps) and its noise (cps).

2.2.2. Definition of the sample preparation

Five sample preparation strategies were tested (detailed in Figure 2.1) to evaluate the trade-off between the effort invested in sample preparation and XRF data quality. The goal was to find the simplest preparation method, that allows obtaining CV values of the emission lines of interest (e.g., P Kα, S Kα, K Kα, Ca Kα, Mn Kα, Fe Kα, and Zn Kα) lower than 10%, when performing the measurements in replicates (5 scans per method).

After defining the optimal dwell time and sample preparation procedure, the XRF measurements were performed accordingly using the samples described in Section 2.2.3. Each sample was measured in biological triplicate.

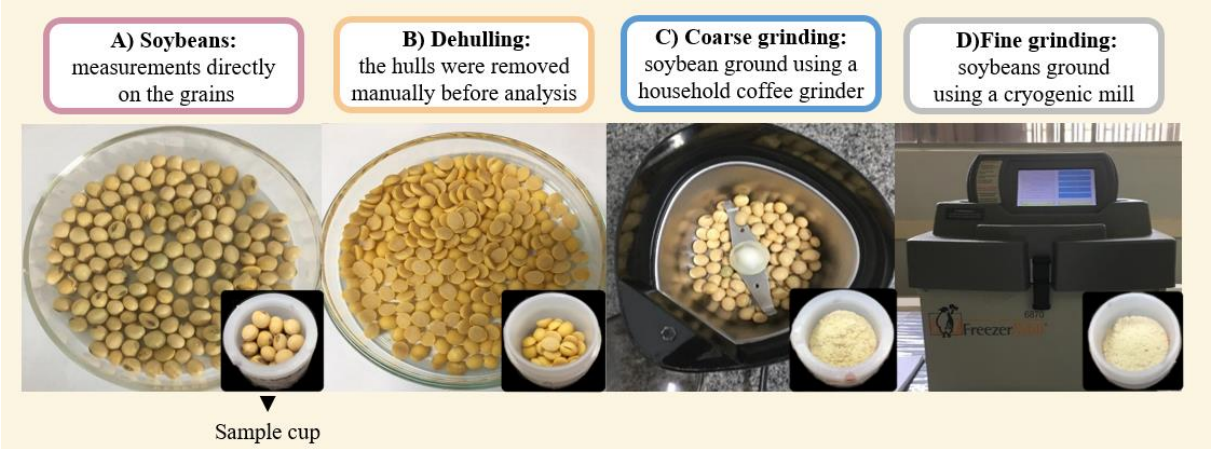


Figure 2.1 - Sample preparation methods: (A) The measurements were performed directly on the soybeans; (B) on the soybeans peeled by hand; (C) on the soybeans ground for 60 s using a household coffee grinder (Cadence, model MDR302, Balneário Piçarras, SC, Brazil); (D) and on the soybeans ground using a cryogenic mill (Spex Sample Prep, Freezer/Mill 6870, Metuchen, NJ, USA), 5 grinding cycles of 2 min, each with 1 min of cooling between cycles and a pre-cooling time of 5 minutes.

2.2.3. Soybean samples, determination, and classification of its protein content

One hundred and sixty-two soybean samples of different cultivars from four locations in Brazil (Iracemápolis – SP, Sacramento – MG, Uberaba – MG, and Perdizes – MG) were used in this study.

The protein content was determined by the combustion method (AOAC, 1997), using the FP–528 protein/nitrogen analyzer (LECO Corp, St Joseph, MI, USA). This method indirectly quantifies the amount of protein by the total nitrogen content. To obtain the crude protein content, the nitrogen content on a dry basis was multiplied by the factor 6.25. The

accuracy and precision of the reference method were checked using a certified reference material (NIST SRM 3234 – Soy flour, National Institute of Standards and Technology, Gaithersburg, MD, USA). The soybean dry matter was determined by oven drying at 105 °C for 24h (BRASIL, 2009). All analyses were carried out using two biological replicates.

The average number between the maximum and minimum value observed in this data set was the threshold to distinguish high from low protein samples: samples presenting values equal to or below 38.8% were considered low protein content and those above this threshold as high content.

2.2.4. Data modeling with binary logistic regression

The entire dataset was divided into training and test datasets. 70% of the samples (n = 108) were employed to build a logistic regression model associated with a stepwise procedure for variable selection. The remaining samples (30% of the total, n = 54) were used for the model validation. The selection of the training and test datasets was done using the Kennard-Stone algorithm (KENNARD; STONE, 1969), to ensure similarity between the datasets in terms of protein content range and variation.

Binary logistic regression is a supervised machine-learning method used for classification. This model uses explanatory variables (i.e., XRF data) to calculate the probability of occurrence of an event, in this case, high protein content sample. The P Kα, S Kα, K Kα, Ca Kα, Mn Kα, Fe Kα, Zn Kα, Rh Kα Compton, and Rh Kα Thomson net intensities were the independent variables, whereas the protein content classification was the dependent variable.

The probability (P, Eq. 2) of the event (high protein content sample) was estimated using the logit function (Z, Eq. 3), a linear combination of the independent variables. The coefficients $\beta_0, \beta_1, \ldots \beta_K$ were determined by maximizing the log-likelihood (LL, Eq. 4) function (FÁVERO; BELFIORE, 2017).

$$P_i = 1/(1 + e^{-Z_i}) \tag{2}$$

$$Z_i = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} \ldots + \beta_k . X_{ki} \tag{3}$$

$$LL = \sum_{i=1}^{n} \left\{ \left[(Y_i) . ln(\frac{e^{z_i}}{1 + e^{z_i}}) \right] + \left[(1 - Y_i) . ln\ (\frac{1}{1 + e^{z_i}}) \ \right] \right\} \tag{4}$$

As far as the probability was known, the cutoff value could be defined. When the probability is higher than the cutoff, the observation is categorized as an event (FÁVERO; BELFIORE, 2017). Therefore, the probability of a soybean sample belonging to the high protein content category was predicted, accordingly, using the R statistical software (version 4.1.2, R Core Team, Vienna, Austria). The cutoff value was chosen to maximize the sensitivity and specificity (FÁVERO; BELFIORE, 2017).

The performance of the model was assessed by the global accuracy, sensitivity, specificity, and Cohen's kappa statistic. The global accuracy (Eq. 5) corresponds to the proportion of observations correctly classified.

$$Global\ accuracy = N_1/N_0 \tag{5}$$

where N_1 corresponds to the number of observations correctly categorized and N_0 is the total number of observations. Sensitivity (Eq. 6) is the proportion of observations classified as positive that are positive. Likewise, specificity (Eq. 7) is the proportion of negative observations correctly classified (FÁVERO; BELFIORE, 2017).

$$Sensitivity = TP/(TP + FN) \tag{6}$$

$$Specificity = TN/(TN + FP) \tag{7}$$

where TP and TN correspond, respectively, to true positive and true negative, i.e., the number of high and low protein samples correctly categorized. FN and FP represent, respectively, false negative and false positive, i.e., the number of high and low protein samples mistakenly classified. The kappa statistic indicator is used to access the level of agreement between the true and predicted rating, and its values were interpreted according to the classes suggested by Landis and Kock (1997): poor (kappa statistic < 0.00), slight (0.00 – 0.20), fair (0.21 – 0.40), moderate (0.41 – 0.60), substantial (0.61 – 0.80) and almost perfect (0.81 – 1.00).

2.3. Results and Discussion

2.3.1. Effect of dwell time on the XRF data

Figure 2.2 shows the soybean XRF spectra, CV, and SNR of the emission lines acquired using a dwell time of 30 s and 90 s. Both conditions showed CV lower than 10% when evaluating the replicates. Besides that, increasing the dwell time from 30 s to 90 s, the SNRs increased between 70.9% and 75.6%, and the CV reduced between 0.2% and 6.6% for most of the emission lines (P Kα, S Kα, K Kα, Mn Kα, Zn Kα, and Rh Kα Thomson). Specifically, the SNR of the S Kα emission line increased by 75.6%, and the corresponding CV reduced from 1.6 to 0.9%, compared to 30 s. Hence, to warrant precise measurements, with higher SNR and lower CV, the adopted dwell time was 90 s

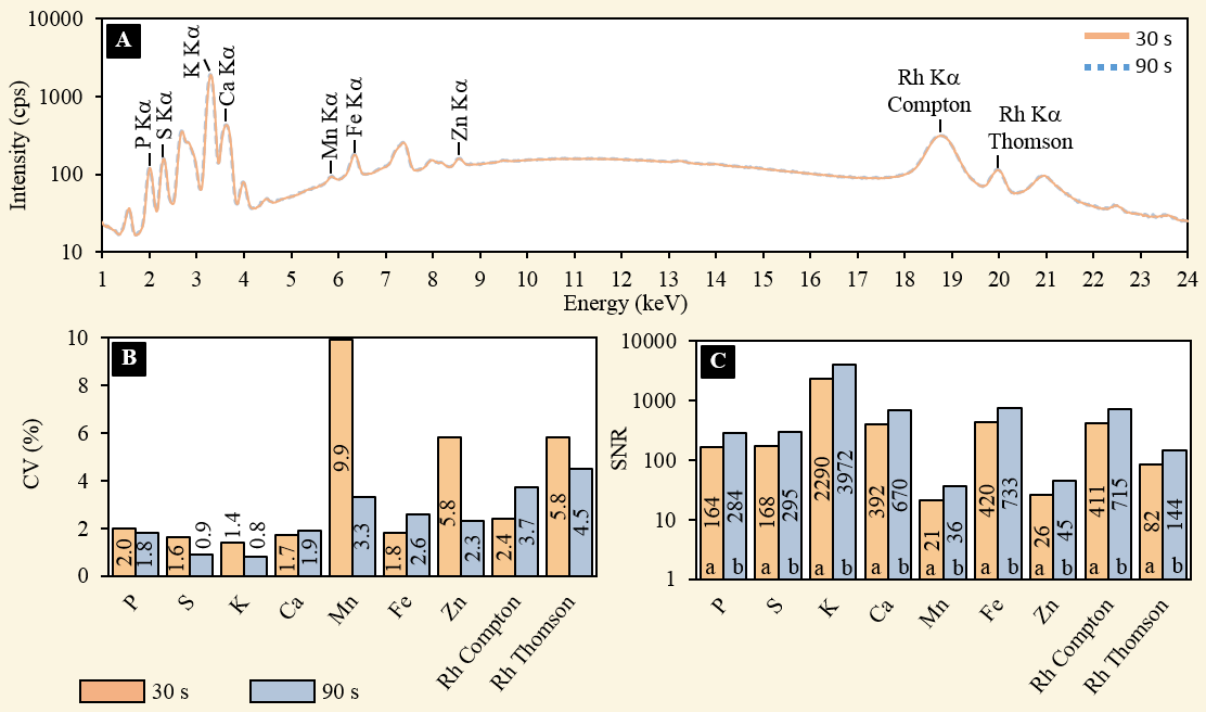


Figure 2.2 - (A) Spectral profile of a single sample acquired with a dwell time of 30 and 90 s; (B) CV of the Kα emission line intensities and (C) SNR for each equipment setting and dwell time. SNRs with different letters (a–b) are significantly different (p-value < 0.05) according to the variance analysis (ANOVA).

2.3.2. Effect of sample preparation on XRF data

Figure 2.3 shows the effect of different sample preparation tested on the XRF spectrum and coefficient of variation of P Kα, S Kα, K Kα, Ca Kα, Mn Kα, Fe Kα, Zn Kα, Rh Kα Compton, and Rh Kα Thomson intensities. In general, the measurements obtained with the whole and dehulled soybean presented lower net intensity and higher CV. This result was expected because these preparation methods do not guarantee sample homogenization. On the other hand, the results indicate that fine and coarse grinding reached more accurate measurements, as in both cases the CV was below 10% (the limit considered acceptable for direct measurements in XRF spectroscopy). Besides, the P Kα, S Kα, and K Kα emission line intensities were higher when compared with the measurements obtained with the whole and dehulled soybean.

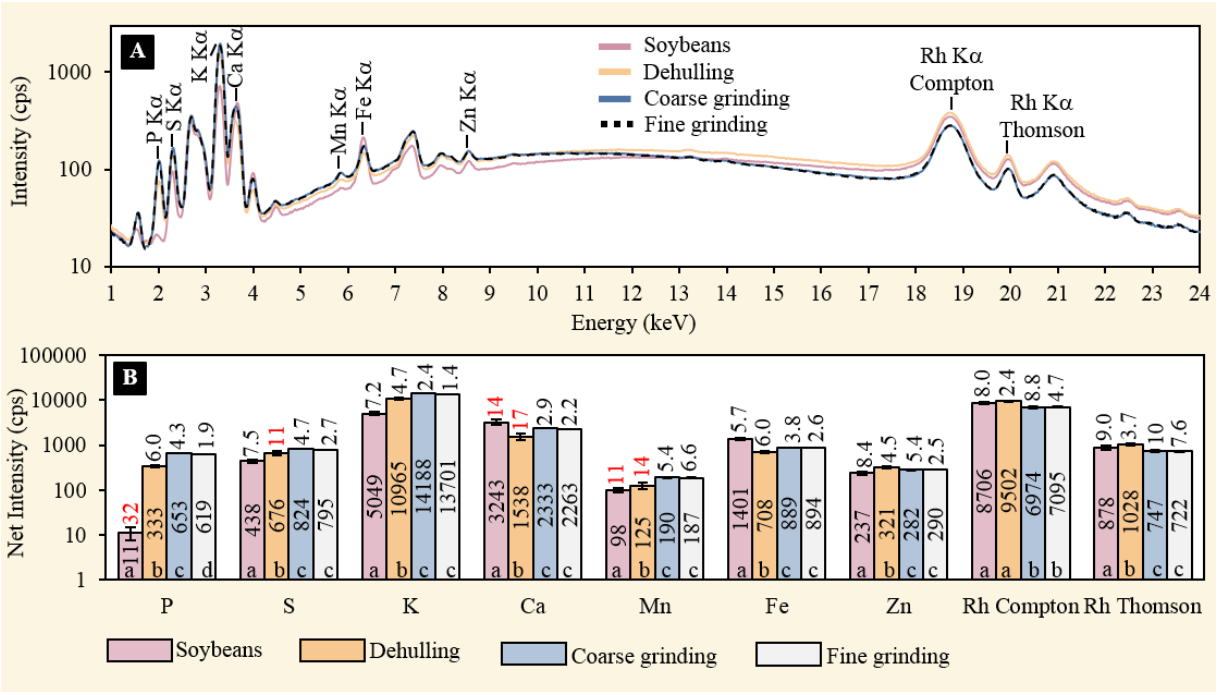


Figure 2.3 - Effect of different sample preparation methods on soybean spectrum (A) and intensities of the Kα emission lines and their respective coefficients of variation (labels displayed above the bars) (B). Results obtained by Tukey's test and means with different letters (a-d) are significantly different (p-value < 0.05).

The intensities of the Ca Kα and Fe Kα reached higher values with no sample preparation, while the Zn Kα intensity was higher in dehulled grains, which may be explained by the heterogeneous distribution of elements in the grain. It is well known that the particle size reduction promotes the homogenization of the different elements in the sample since there are

regions in soybeans where these elements are agglomerated (nuggets). Our results corroborate it, as the grinding of the samples promoted a considerable increase in precision for all XRF peaks (The CV reduced at a maximum of 30.1% for P Kα, 8.3% for S Kα, 5.8% for K Kα, 14.8% for Ca Kα, 7.4% for Mn Kα, 3.4% for Fe Kα, 5.9% for Zn Kα), except for the scattering peaks.

Comparing coarse and fine grinding, the effectiveness of cryogenic grinding in providing homogeneous samples was evidenced. This procedure contributed to the lowest CV for all emission lines and therefore is highly recommended for XRF spectroscopy. However, there are some drawbacks to using it, such as the time required for sample preparation (about 240 s), the relatively high cost of the equipment (approximately US$ 12,000), and the requirement for liquid nitrogen for its operation. In contrast, a household coffee grinder is cheap (about US$ 35) and easy to use, comparing the two grinding procedures, the results are quite similar. Tukey's test showed that all XRF peaks, except for P, had similar net intensities (p-value < 0.05) when comparing the data obtained with samples prepared with household coffee and cryogenic grinders. In light of the abovementioned results, coarse grinding was chosen as the sample preparation method for further analyses of this study.

2.3.3. Reference analysis of protein content

The protein content of the samples, determined by the reference method, ranged from 33.8% to 43.9% on a dry basis, with mean and median of 39.1% and 39.4%, respectively (Table 2.1). The CV and accuracy of the reference method were 0.63% and 101%, respectively, by measuring the NIST SRM 3234. The average between the maximum and the minimum values (38.8%) was the threshold to split the observations into the classes: high and low protein contents. It is important to mention that the range of protein content found in this study is comparable to the ones reported in the literature (Table 2.2). For example, Lee, Kim and Hwang (2021) and Assefa *et al.* (2019) observed ranges of 28.7% - 44.6% and 27.3% - 45.4%, respectively. Overall, the samples of our study presented similar protein content variability to soybean samples worldwide. Regarding data modeling, it is important that the training and test sets have similar descriptive statistics since different ranges and standard deviations between them can lead to biased interpretations of model performance (STENBERG *et al.*, 2010). As shown in Table 2.1 these characteristics were preserved for the training and test subsets in the present study.

Table 2.1 - Descriptive statistics of the soybean protein content and % of samples/class

		Protein content (%)						% of samples/class	
Dataset	**No.**	**Min.**	**Max.**	**Mean**	**Median**	**SD**	**CV**	**High**	**Low**
Full	162	33.8	43.9	39.1	39.4	2.1	5.4	56.8	43.2
Training	108	33.8	43.9	39.1	39.4	2.1	5.4	56.5	43.5
Test	54	33.9	43.6	39.1	39.4	2.1	5.4	57.4	42.6

No.: number of samples, Min.: minimum, Max.: maximum, SD: standard deviation, CV: coefficient of variation, Full dataset: all samples (training and test samples).

Table 2.2 - Ranges reported for protein content in soybeans

Reference	**Min.**	**Max.**	**Mean**	**No**	**Description**
(Lee; Kim; Hwang, 2021)	28.7	44.6	39.1	300	Samples from different countries (Korea, China, Japan, USA, Russia, and North Korea)
(Uikey *et al*., 2022)	36.1	41.2	38.0	154	Samples of different genotypes (154)
(Grieshop *et al*., 2001)	39.4	44.5	41.5	133	Samples from different countries (Brazil, China, and USA)
(Jiang, 2020)	33.4	47.7	41.8	20	Samples of different genotypes (16)
(Ferreira *et al*., 2014)	32.9	42.2	38.3	40	Samples of different varieties (20) and cities (2) in Brazil
(Assefa *et al*. 2019)	27.3	45.4	35.7	13574	Samples from several locations across the USA
(Dong; Qu, 2012)	34.5	48.7	40.9	114	Samples from supermarkets in China
(Singh *et al*., 2018)	26.0	37.9	31.2	31	Samples of different cultivars (31)
(Armstrong, 2006)	29.0	55.0	40.9	300	Samples of different varieties (3)
(Wei *et al*., 2021)	32.0	47.0	–	75	Samples collected in China
(Zhu *et al*., 2018)	37.0	43.2	40.4	360	Samples of different varieties (50) and areas (2) of China
(Choung *et al*., 2001)	36.0	51.8	43.2	300	Samples of different varieties (300)

Min.: minimum, Max.: maximum and No.: number of samples

2.3.4. Performance of the logistic regression model

The performance of the classification model is shown in Figure 2.4. The logistic regression model classified more than 80% of the observations correctly, with a global accuracy of 0.83 and 0.81 for the training and test data sets, respectively. The training and test datasets presented kappa statistic values of 0.66 and 0.61, respectively. This indicates a good level of agreement between the true and classified ratings. Indeed, according to Landis and Koch (1997), kappa statistics presenting values between 0.61 and 0.80 indicate substantial strength of agreement. In this research, the cutoff was chosen to maximize the specificity and sensitivity of the training datasets. Adopting a threshold of 0.6, yielded a sensitivity and specificity of 0.84 and 0.83, respectively. For the test dataset, the sensitivity and specificity were

0.94 and 0.65, respectively. Consequently, the model had better performance in identifying high-protein samples than low ones in the test data set.

Furthermore, in the stepwise procedure, that selects the most important variables for the model calibration, only the S Kα and Mn Kα emission lines were statistically significant (p-value < 0.01 for Mn Kα and p-value < 0.001 for S Kα). Thus, they were included in the logistic regression model. Soybean proteins have sulfur-containing amino acids, more precisely methionine and cysteine (MA *et al.*, 2019), therefore the contribution of the S Kα emission line was expected. As for the Mn Kα emission line, its contribution to the logistic regression model might be related to the physiological processes involved in soybean protein synthesis. In addition, it is common to apply fertilizers containing S and Mn during the development of soybean crops, since both nutrients can improve the yield and protein content of soybean production (NAZAROVNA *et al.*, 2020). Differently from Terra (2009), here a logistic regression model was designed with a much larger sample set and the scattering region of the spectra was not statistically significant (p-value > 0.05).

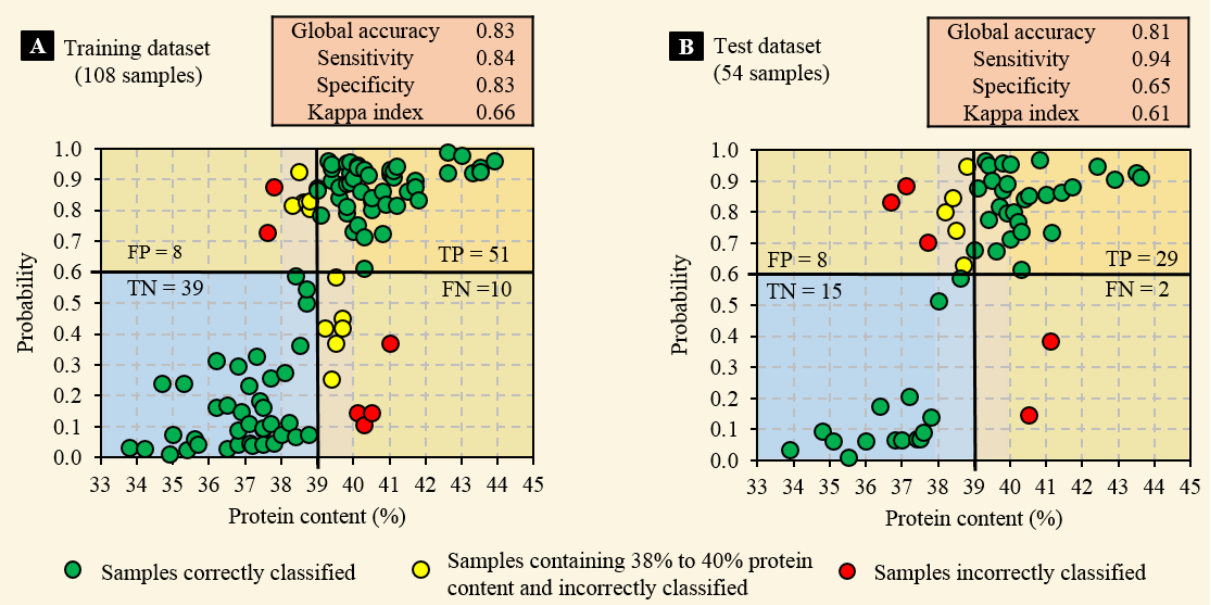


Figure 2.4 - Binary logistic regression performance indicators for the training (A) and test (B) datasets and scatter plots of the probability of occurring the event (i.e., classification of a high-protein sample) versus the real protein content (reference analysis). Where true positive (TP) and true negative (TN) correspond to the number of samples that have respectively high and low protein content and were correctly classified as such; False negative (FN) and false positive (FP) respectively correspond to the samples incorrectly classified as high- and low-protein samples.

Figure 2.4 shows that the model misclassified 18 out of 108 samples of the training set. Among them, 12 samples contain 38% to 40% of protein, i.e., around 38.8% (approximately 39 ± 1%), which corresponds to the value adopted to split the samples into the high and low

categories. Likewise, 10 out of 54 samples from the test dataset were misclassified, with 5 of them containing 38% to 40% of protein content. Thus, the main error of the model is the classification of samples containing a protein content of around 39%. One factor that may have contributed to the model misclassification was the error of the reference method, which is near 0.6%. The reference method is always critical since its systematic and random errors are inadvertently added to the statistical model. Thus, the accuracy of the reference method is fundamental for the evaluation of predictive models and should be considered in future studies.

The development of a rapid, environmentally friendly, and easy-to-use technique for protein analysis in soybean will enable the expansion of the diagnosis of this quality-related attribute. There are methods for quantifying crude protein content in soybean, such as near-infrared spectroscopy (NIR). Although NIR is a well-established and non-destructive method, it measures overtones and combination tones of molecular vibrations and not the concentration of macro and micronutrients (such as P, K, S, Fe, Mn, and Zn,). The XRF technique is also able to quantify those elements (OTAKA; HOKURA; NAKAI, 2014; LAI *et al*., 2020), which are essential for animal and human nutrition. In addition, as previously indicated, this information can be used to infer protein content.

The XRF analysis features (rapid and easy operation) make the method attractive for direct analysis in the field, e.g., with equipment embedded in agricultural machinery. Recent works have suggested the use of XRF for the evaluation of sugarcane quality indicators (CORRÊDO *et al*., 2021), pointing out that the use of this sensor in harvesters is a promising alternative for spatial variability mapping of sugarcane quality in the field, enabling its management through Precision Farming approaches. In turn, the XRF technique can directly benefit farmers, by allowing them to plan each season's agricultural management based on spatial variability of quality indicators of previous and current harvests. In this scenario, it is well known that the accuracy of field analysis is lower than that of laboratory analysis (KUANG *et al*., 2012; GREDILLA *et al*., 2016). On the other hand, this is compensated by the high spatial resolution of the data collection, i.e., the predictions are compared with the predictions of their neighbors and inconsistent values can be filtered using spatial algorithms (MALDANER; MOLIN; SPEKKEN, 2021). Reliable soil P mapping using a NIR sensor on an agricultural platform was reported by Mouazen and Kuang (2016), who used a predictive model with reasonable performance (e.g., R^2 of 0.60). The authors used a fine resolution of 1,000 data points/ha for the spatial characterization of this variable in the field.

The results of our study showed that XRF sensors can be used as a practical tool for identifying soybean with high and low protein content. Nevertheless, further studies are needed

to strengthen the hypothesis that protein determination can be done via quantification of elemental S, as well as, to optimize strategies to improve the predictive/classificatory performance of XRF sensors. In this sense, some topics can be suggested for future research, such as: (i) increasing the number of samples to cover the full protein range (26 to 55 %) reported in the literature; (ii) testing other modeling approaches (e.g., computational models); (iii) considering X-ray tube configurations with lower voltages (e.g., < 10 kV) to enhance the S fluorescence emission; (iv) combining XRF data with other direct analytical techniques (e.g., vis-NIR sensors) to explore the synergy between sensors. In addition, recent portable equipment (e.g., Tracer 5, Bruker AXS, Madison, WI, EUA) have incorporated technologies that improve the detection of light elements, such as thinner beryllium window and optimized geometry between the X-ray tube and detector (MIGLIORI *et al.*, 2011). The use of this type of equipment could also be considered in future research.

2.4. Conclusion

In the present study, a portable energy-dispersive XRF sensor was successfully applied for classifying soybean based on protein content, using a simple sample preparation method (grinding the samples with a household coffee grinder). The performance of the logistic regression model, used for classifying the samples into high and low protein content, was appropriate. The global accuracy and Kappa Index of the test data set were 0.81 and 0.61, respectively, showing that the strength of agreement between the true and predicted ratings is good. In addition, the sensitivity of the test dataset was 0.94, indicating the high performance of the model in identifying high-protein soybean samples. The variables that were important for the model calibration were the S Kα and Mn Kα emission lines (p-value < 0.01 for Mn Kα and p-value < 0.001 for S Kα). Thus, the hypothesis that sulfur could be used as a proxy for the classification of soybeans in terms of protein content was evidenced in this research. The scattering peaks did not present a significant contribution to the model in our study, contradicting hypothesis already raised in the literature. Finally, the findings of this research open doors for further investigations, such as sorting soybean by protein content in the field, e.g., with the equipment embedded in agricultural machinery.

3. QUANTIFYING SOYBEAN PROTEIN CONTENT BY X-RAY FLUORESCENCE SPECTROMETRY

Abstract

X-ray fluorescence (XRF) is a technique for elemental analysis. This study evaluated the feasibility of using XRF to predict soybean protein content. Univariate linear regression, multiple linear regression, and partial least squares regression (PLS) were compared as modeling strategies. Two dataset scenarios were considered: (A) 108 soybean samples for calibration and 54 for validation; (B): added 22 and 10 samples, prepared by mixing soybean grain and concentrates, to the calibration and validation datasets, respectively, of scenario A. The aim of scenario B was to widen the protein range of the datasets. Among the modeling strategies, PLS showed the best performance in the validation, with R^2 of 0.73 and 0.89 in scenarios A and B, respectively. Results show that the XRF sensor is suitable for screening applications, creating the possibility of incorporating the technique in approaches that require high throughput analysis, such as agricultural robots and mobile laboratories.

Keywords: XRF; chemometrics; multiple linear regression; partial least squares regression; food analysis; multivariate calibration.

3.1. Introduction

Soybeans contain up to 45 wt.% of protein (ASSEFA *et al*., 2019), for this reason, it is widely used in animal and human nutrition (SUDARIĆ, 2020). However, as grain yields have increased, soybean protein content has been declining, at a rate of *ca*. 2% per decade (UMBURANAS *et al*., 2022). In this context, there is a global market trend to purchase soybean based on protein content rather than grain weight. China, the main importer of the commodity, is already pushing suppliers in that direction (WILLIAM; DAHL; HERTSGAARD, 2020).

Hence, it is crucial to develop agile methods for quantifying protein content. Well-established methods for determining protein content are based on Kjeldahl, Dumas, and near-infrared spectrometry approaches (CHANG; ZHANG, 2017). Kjeldahl and Dumas indirectly quantify protein content by applying a conversion factor to the total nitrogen content of the sample. Kjeldahl is the most popular method for protein determination, however, it requires hazardous and costly chemical reagents, not fully complying with green chemistry practices. In addition, the Kjeldahl method is laborious and time-consuming if the process is not automated. The Dumas method is faster than Kjeldahl, the nitrogen analyzer is automatic and avoids the use of corrosive and hazardous chemicals (JUNG *et al*., 2003; CHANG; ZHANG, 2017). Another popular method for protein determination employs near-infrared spectroscopy (NIR). This method relies on the use of a multivariate calibration model, which converts the spectra into a protein content prediction result (CHANG; ZHANG, 2017). NIR technique detects O-H, C-H, N-H, and S-H bonds in the sample, which are related to the protein molecule (INGLE *et al*., 2016). It has the same advantages as the Dumas method, plus it does not require ground samples for determination (SHI *et al*., 2022).

X-ray fluorescence spectroscopy (XRF) is a rapid and reagent-free analytical technique that requires little sample preparation for accurate measurements, being an alternative to the traditional wet chemistry methods (TERRA, 2009). XRF is traditionally used to identify and quantify several chemical elements in solid and liquid samples (RODRIGUES *et al*., 2018; MARGUÍ; QUERALT; ALMEIDA, 2022).

In recent years, studies reported the use of XRF for the quantification of organic molecules or correlated properties. They argue that the predictions of these attributes are associated with the X-ray scattering spectral region, which depends on the average atomic number of the sample. Melquiades *et al*. (2012), Alexandre, Goraieb and Bueno (2010), and Terra *et al*. (2010) explored the above-mentioned relationship. The first one developed a method for simultaneous determination of sugar cane quality parameters (i.e., sucrose and fiber

content); The second study determined the sucrose content of cashew juice; and fixed acidity, alcohol, and sucrose content of an alcoholic beverage; and the third mentioned research developed a method to determine the energy value of vegetable-based dried products of different origins (e.g., corn, soy, wheat, cassava, milk).

XRF was previously proposed to estimate the protein content of soybeans. In 2009, Terra combined XRF with chemometrics to quantify the protein content of soybean-based products (i.e., flour, fiber, extract, and protein-enriched powder), achieving high predictive performances (R^2 of 0.86). A partial least squares regression (PLS) model was developed with the spectra of 12 samples, revealing the scattering peaks (i.e., Compton and Thomson scattering regions) as the most important variables for prediction. Despite the promising results of that study, the reliability of multivariate calibration can be enhanced if more samples (e.g., $n > 50$) are used (TANG *et al*. 2017). So, it is important to assess whether this relationship will persist across larger databases.

Another potential relationship between XRF and protein data may be related to the S Kα emission line. S is present in the methionine and cysteine amino acids, of soybean storage proteins (MA *et al*., 2019). If the proportion of such proteins in soybeans is constant and considering that the primary structures of proteins have specific amino acid profiles, then protein content might be estimated from the S signal count rate of the XRF spectrum. The relationship between the S Kα and Mn Kα emission lines intensities of the XRF spectra and protein content was recently reported (CAMARGO *et al*., 2023). These emission lines were employed to develop a logistic regression for classifying soybeans into high and low protein content categories. It is important to mention that similar reasoning is used in the traditional Kjeldahl and Dumas methods, which estimate the concentration of protein based on the total N content.

Thus, the hypothesis that motivated the present study is that protein content may be estimated via XRF measurements, by (i) SKα emission line and (ii) scattering peaks. To advance the findings recently reported by Camargo *et al*. (2023), the aim of this study was to evaluate the feasibility of quantifying soybean protein content using XRF, reproducing the experiment of Terra (2009) with a larger sample set (n = 194). Additionally, different data modeling approaches were evaluated, comparing PLS, multiple linear regression (MLR), and univariate linear regression (ULR) models to establish an optimal predictive strategy for rapid and environmentally friendly soybean protein analysis via XRF.

3.2. Material and Methods

3.2.1. Soybean samples

One hundred and sixty-two soybean samples were used in this study. They were acquired from four places in Brazil: (i) Iracemápolis, SP, (ii) Sacramento, MG (iii) Uberaba, MG and (iv) Perdizes, MG. The grains were 60 s ground into flour using a coffee grinder (Cadence, model MDR302, Balneário Piçarras, SC, Brazil).

3.2.2. Blends

Thirty-two blended samples (denominated as blends) were prepared by homogeneously mixing soybean samples (detailed in Section 3.2.1.) in different proportions (Table 3.1) with two commercial soybean flours: soybean fiber flour (Inabel Alimentos, Jumirim, SP, Brazil) and defatted soybean flour (Inabel Alimentos, Jumirim, SP, Brazil), which contained 19% and 54% of protein content, respectively. The goal was to widen the protein range of the calibration models by using samples with protein levels below 33% and above 44%.

Table 3.1 - Proportion (% w/w) of soybean fiber flour (SFF), defatted soybean flour (DSF), and soybean[1] in the blended samples.

Sample	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Soybean[1]	0	12	18	24	29	35	39	41	46	51	54	57	63	68	74	79
SFF	100	88	82	76	71	65	61	59	54	49	46	43	37	32	26	21

Sample	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
Soybean[1]	87	88	81	73	65	61	53	50	49	45	38	30	23	22	14	0
DSF	13	12	19	27	35	39	47	50	51	55	62	70	77	78	86	100

[1]Each blended sample was prepared with a different soybean sample.

3.2.3. Total protein content

The total nitrogen content was determined by the Dumas method (AOAC, 1997), using a FP-528 protein/nitrogen analyzer (LECO Corp, St Joseph, MI, USA). This method uses a conversion factor of 6.25 to convert the total nitrogen (%, in a dry matter basis) into protein content. Dry matter was determined by the oven-drying method (105°C for 24h). The analyses were performed with two biological replicates.

A certified reference material from the National Institute of Standards and Technology (NIST 3234 - soy flour) and another reference material obtained from the Brazilian Agricultural Research Corporation (soybean flour CRM LRI09091, Embrapa) were utilized to validate the reference method (Dumas) and the developed method (XRF). The recoveries of the reference method using NIST 3234 and CRM LRI09091 were 101% and 98%, respectively.

3.2.4. X-ray fluorescence spectrometry analysis

Prior to the XRF measurements, a preliminary test was performed to evaluate the accuracy of the measurements performed on loose and pressed powder samples, to decide which sample preparation to use in the present study. For this, a soybean sample was measured in triplicates, before and after pressing it with a pellet press (Parr Instrument Company, Moline, IL, USA). The precision of the measurements was evaluated by the coefficient of variation (CV) of the P, S, K, Zn, and scattering peaks fluorescence emission (Figure 3.1). As pressed powder samples showed more accurate results and the time required for the procedure is short, approximately 10 s per sample, this sample preparation was then used for the subsequent analyses.

Three grams of each sample were transferred into a 31 mm diameter XRF cup (n. 1530, Chemplex Industries Inc., Palm City, FL, USA) sealed on the bottom with a 6 mm thick polypropylene film (VHG Labs, Manchester, NH, USA). The samples were pressed directly in the XRF cup using the pallet press. This procedure resulted in 7.3 mm thick pellets.

The samples were scanned with a portable X-ray fluorescence spectrometer (Tracer III–SD, Bruker AXS, Madison, WI, USA), equipped with a Rh Anode X-ray tube (4 W power) and a X-Flash® Peltier-cooled Silicon Drift Detector (Bruker AXS, Madison, WI, USA). The X-ray tube was set to 40 kV and 30 μA and the measurements were done under a vacuum atmosphere, without a primary filter, during 90s (dwell time). Biological triplicate readings were taken for each sample.

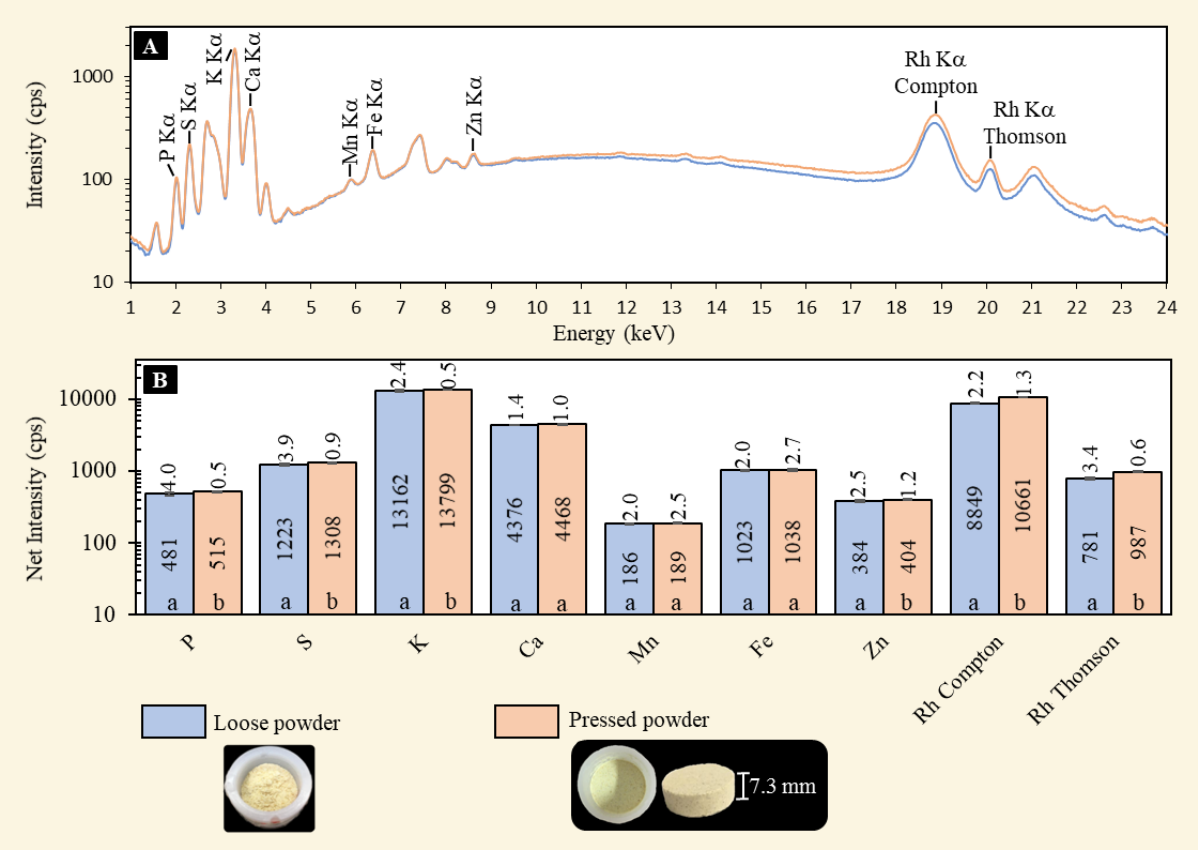


Figure 3.1 - (A) Spectral profile of a soybean sample before (loose powder) and after pressed (2811 Pellet Press, Parr Instrument Company, Moline, IL, USA). (B) Kα peak intensities (labels inside the bars) and coefficients of variation (labels above the bars) of measurements performed on the loose powder and pressed powder sample. Means with different letters (a-b) are significantly different (p-value < 0.05) according to the variance analysis (ANOVA).

3.2.5. Data Modeling

Two dataset scenarios were considered: (A) 108 soybean samples for calibration and 54 for validation; (B): added 22 and 10 blended samples (detailed in Section 3.2.2) to the calibration and validation datasets, respectively, of scenario A. Thus, scenario B came up to 130 and 64 samples for calibration and validation, respectively. The aim of scenario B was to widen the protein range of the datasets. The selection of calibration and validation datasets was made using the Kennard-Stone algorithm, which ensures similarity between the datasets in terms of protein content range and variation.

Using the spectral data as explanatory variables (independent variables) and the reference data as dependent variables, three predictive modeling strategies (described in detail in the following sections) were considered for both scenarios. The quality of the models was evaluated by the coefficient of determination (R_C^2 and R_P^2 for the calibration and validation sets, respectively), root-mean-square error (RMSEC and RMSEP, for the calibration and validation

sets, respectively), and residual prediction deviation (RPD). Regarding the RPD values, the models were categorized as poor (RPD < 1.40), reasonable (1.40 ≤ RPD < 2.00), good (2.0 ≤ RPD < 3.00), and excellent (RPD ≥ 3.00) performance (CHANG *et al.*, 2001). The recovery of the models was accessed using CRM LRI09091 and NIST 3234 reference materials (described in section 3.2.3).

3.2.5.1. Univariate and multiple linear regressions

Univariate linear regressions (ULR) were developed with the net intensity of the S Kα emission line after being normalized by the Rh Kα Compton, as suggested by Tavares *et al.* (2020). Multiple linear regressions (MLR) were developed with the net intensities of the following emission lines: P Kα, S Kα, K Kα, Ca Kα, Fe Kα, Mn Kα, and Zn Kα (all of them normalized by the Rh Kα Compton net intensity), as well as the scattering peaks (Rh Kα Compton and Rh Kα Thomson). Net intensities were determined with the Bayesian deconvolution process of the Artax® software (Bruker AXS, Madison, EUA) and normalized by the detector live time, being reported in counts of photons per second (cps). MLR was also associated with a stepwise variable selection method, i.e., backward elimination (CHAN *et al.*, 2022), here referred to as stepwise MLR.

3.2.5.2. Partial least squares regression (PLS)

The entire spectrum, between 1 and 25 keV, was used to develop the PLS models. As pre-treatment, firstly the spectra were aligned using the correlation optimized warping method (NIELSEN; CARSTENSEN; SMEDSGAARD, 1998), performed with a step and slack of 80 and 8, respectively. Then, a test was performed to select an optimal combination of pre-processing techniques, being selected the one with the lowest root-mean-square error of prediction (RMSEP) (Table 3.2). The following sequence was chosen: multiplicative signal correction + interquartile range scaling + mean centering. For PLS calibrations, the optimal number of latent variables (LVs) was selected by the minimum value of the root-mean-square error of cross-validation (RMSECV). Figure 3.2 shows a summarized description of the modeling strategies used in this research.

Table 3.2 - Summary of the results: pre-processing techniques tested for PLS (scenario A).

Pre-preprocessing	Calibration			Validation		
	LV	R^2	RMSEC (%)	R^2	RMSEP (%)	Recovery (%)
MC	4	0.62	1.29	0.63	1.26	100.2
SD scaling + MC	3	0.70	1.14	0.66	1.22	101.6
IQR scaling + MC	3	0.68	1.18	0.68	1.18	101.6
Range scaling + MC	3	0.72	1.11	0.67	1.19	101.1
SG + MC	4	0.62	1.29	0.63	1.27	101.0
SG + SD scaling + MC	3	0.67	1.21	0.68	1.19	100.1
SG + IQR scaling + MC	3	0.64	1.26	0.63	1.26	102.0
SG + Range scaling + MC	3	0.67	1.21	0.69	1.17	101.5
1st der + MC	4	0.59	1.34	0.57	1.37	95.7
1º der + SD scaling + MC	1	0.61	1.31	0.56	1.37	97.3
1º der + IQR scaling + MC	1	0.60	1.32	0.56	1.38	97.2
1º der + Range scaling + MC	1	0.60	1.33	0.57	1.36	97.6
2nd der + MC	4	0.57	1.37	0.56	1.38	95.2
2º der + SD scaling + MC	1	0.63	1.28	0.56	1.38	96.8
2º der + IQR scaling + MC	1	0.62	1.30	0.56	1.39	96.6
2º der + Range scaling + MC	1	0.62	1.30	0.57	1.37	97.4
SNV + MC	3	0.61	1.30	0.63	1.27	100.5
SNV + SD scaling + MC	3	0.72	1.11	0.71	1.12	103.8
SNV + IQR scaling + MC	3	0.73	1.08	0.72	1.10	104.9
SNV + Range scaling + MC	3	0.70	1.14	0.70	1.15	101.9
MSC + MC	3	0.61	1.30	0.62	1.27	100.5
MSC + SD scaling + MC	3	0.72	1.10	0.72	1.11	104.0
MSC + IQR scaling + MC	3	0.74	1.07	0.73	1.09	105.2
MSC + Range scaling + MC	3	0.71	1.14	0.70	1.14	102.0

LV: number of latent variables; R^2: coefficient of determination; RMSEC and RMSEP: root-mean-square error of calibration and validation, respectively; Recovery of the reference material (LRI090921); MC: mean centering; SD scaling: standard deviation scaling; IQR: interquartile range scaling; SG: Savitzky-Golay smoothing; 1st der: Savitzky-Golay first derivative; 2nd der: Savitzky-Golay second derivative; SNV: standard normal variate; MSC: multiplicative signal correction. A third-order polynomial fit with 11 smoothing points was employed in the SG, 1st dev and 2nd der (SANTOS *et al.*, 2021)

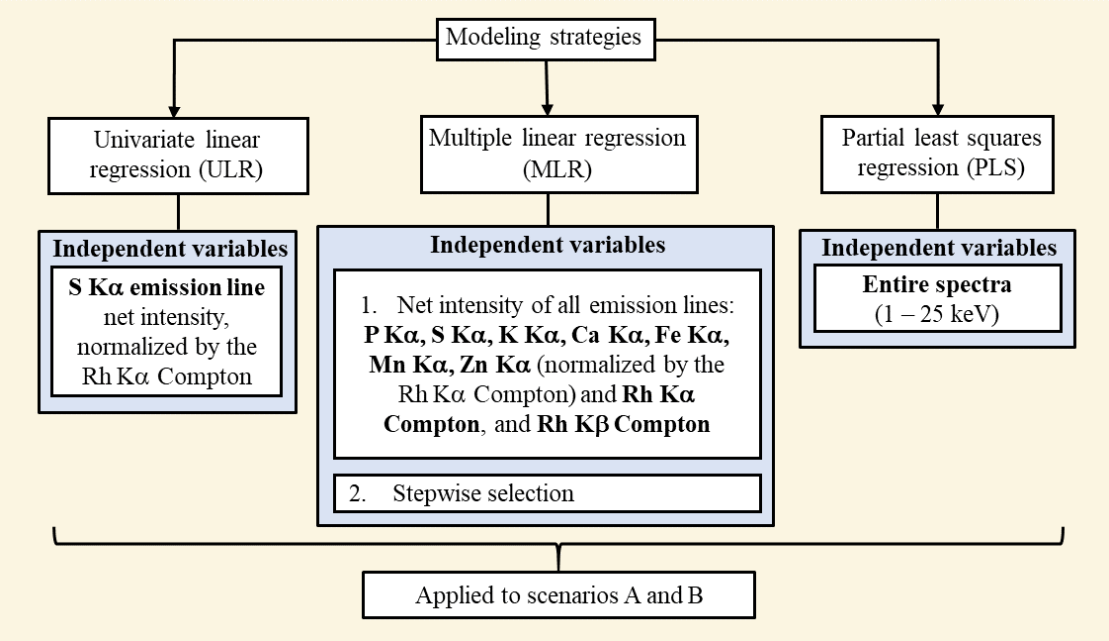


Figure 3.2 - Summary of the modeling strategies applied to evaluate the predictive performance of XRF data for determining soybean protein content. Scenario A used only soybean samples and scenario B used soybean and blended samples.

3.3. Results and discussion

3.3.1. Descriptive statistics

Table 3.3 shows the descriptive statistics of protein content. Note that the calibration and validation datasets have a similar range, mean, median, standard deviation (SD), and coefficient of variation (CV). Special attention was given to achieve these aspects, as different ranges and standard deviations between the calibration and validation datasets can affect the model's performance and bias (STENBERG *et al*., 2010).

Table 3.3 - Descriptive statistics for protein content of the calibration and validation datasets for both scenarios.

Dataset	Soybean (scenario A)		Soybean + blends (scenario B)	
	Calibration	Validation	Calibration	Validation
No.	108	54	130	64
Min.	33.8	34.2	19.2	19.7
Max.	43.9	43.6	54.0	53.4
Mean	39.0	39.1	38.7	39.0
Median	39.4	39.4	39.4	39.5
SD	2.1	2.1	5.1	5.0
CV	5.4	5.3	13.1	12.8

No.: number of samples, Min.: minimum value of protein content (expressed % or g/100g in dry matter basis), max.: maximum value of protein content (%), SD: standard deviation (%), CV: coefficient of variation (%).

3.3.2. Prediction performances

3.3.2.1. Models calibrated with soybean samples (scenario A)

The predictive performances of the ULR, MLR, stepwise MLR, and PLS models calibrated with the scenario A samples are shown in Figure 3.3 These models exhibited reasonable predictive performance ($1.57 \leq RPD \leq 1.92$). Comparing the modeling strategies, PLS exhibited the highest predictive performance (R_P^2 = 0.73), followed by MLR (R_P^2 = 0.66), stepwise MLR (R_P^2 = 0.62), and ULR (R_P^2 = 0.60).

The R_P^2 of the PLS model (0.73) is lower than the one reported by Terra (2009) (0.86). The difference between the results might be explained by the narrower protein range of the samples of the current study. In the pioneering study (TERRA, 2009), a PLS model was calibrated with soybean flours, which protein contents ranged between 21% and 45%, while in the current study the model was calibrated with soybean, with protein contents between 34% and 44%. It is known that the narrower the range of variable Y, the lower the performance of models calibrated with sensed data (ADAMCHUK *et al.*, 2004).

The S Kα emission line was the main variable responsible for the prediction, regardless of the modeling strategy used. For example, the emission lines selected in the stepwise MLR were S Kα, Ca Kα, Fe Kα, and Mn Kα (p-value < 0.05). Among these, S Kα exhibited the highest standardized regression coefficient (Table 4). The addition of other emission lines to S Kα one, i.e., MLR *versus* URL, promoted little improvement in the prediction, the R_P^2 increases only slightly from 0.60 (i.e. ULR) to 0.62 (i.e. stepwise MLR). This indicates low importance of these emission lines compared to S Kα.

The weighted regression coefficients (Bw) of the PLS models provide information on the importance of each variable to predict the targeted attribute. Those with high Bw values play an important role in the PLS model (Figure 3.4). S Kα presented the highest Bw value (at 2.3 keV) and therefore was the most important variable in the model. This is consistent with what was observed in the other models (i.e., MLR and ULR). Also, Zn Kα (8.7 keV), Mn Kα (5.9 keV), Rb Kα (13.4 keV), Sr Kα (14.2 keV) Fe Kα (6.4 keV), Rh Kα Compton (18.9 keV), and Rh Kα Thomson (20.2 keV) contributed to the model.

The S Kα being the main variable responsible for prediction, regardless of the modeling strategy utilized, confirms the hypothesis of this research. As mentioned earlier (Section 3.1), the relationship between protein content and the S Kα emission line should be related to the presence of the element in the structure of cysteine and methionine amino acids (MA *et al.*,

2019). Terra (2009) reported that protein content could be estimated from the XRF scattering region, thus the contribution of the Rh Kα Compton and Thomson peaks to the models was expected. In the current study, the contribution of the scattering peaks was only observed in the PLS model. Peaks of other elements contributed to the models, e.g., Zn, Mn, Fe, Rb, and Sr in the PLS model. Zn, Mn, and Fe are known to be involved in protein synthesis in soybeans (BUTTROSE, 1978; ECKERMANN; EICHEL; SCHRÖDER, 2000). This may explain their relationship with protein content. However, to the best of our knowledge, there is no explanation for why Rb Kα and Sr Kα contributed to the model. Thus, future studies can be conducted, e.g., with other samples, to evaluate the relationship between protein and these elements (Zn, Mn, Fe, Rb, and Sr).

In the PLS model, the full spectrum was used for calibration, which might explain its superior performance (R_P^2 = 0.73), compared to the other approaches. For example, PLS captured background information and even emission lines of very low signals (e.g., Rb and Sr). Just like the scattering peaks, the background intensity of XRF spectra depends on the mean atomic number of the sample, i.e., the lighter the matrix, the higher is the background intensity (ALLEGRETTA *et al*., 2020). This may explain the observed negative association between background and protein content. PLS is the most popular multivariate calibration method (FERREIRA, 2015), its algorithm uses predictor variables (e.g., spectral data) and response variable (e.g., protein content) to calculate a new set of variables (i.e., latent variables). In this calculation, spectral information, highly correlated with the response variable, receive extra weight in the regression (MILLER; MILLER, 2010). The PLS method is recognized as a robust calibration technique and is highly recommended for complex systems, e.g., in cases of collinearity between variables and in the presence of interferences (FERREIRA, 2015; CHAN *et al*., 2022). In fact, Wang, Zhao and Kowalski (1990) suggested PLS as a general modeling method for EDXRF analysis, because of its superior performance in the prediction of elements compared to a conventional method.

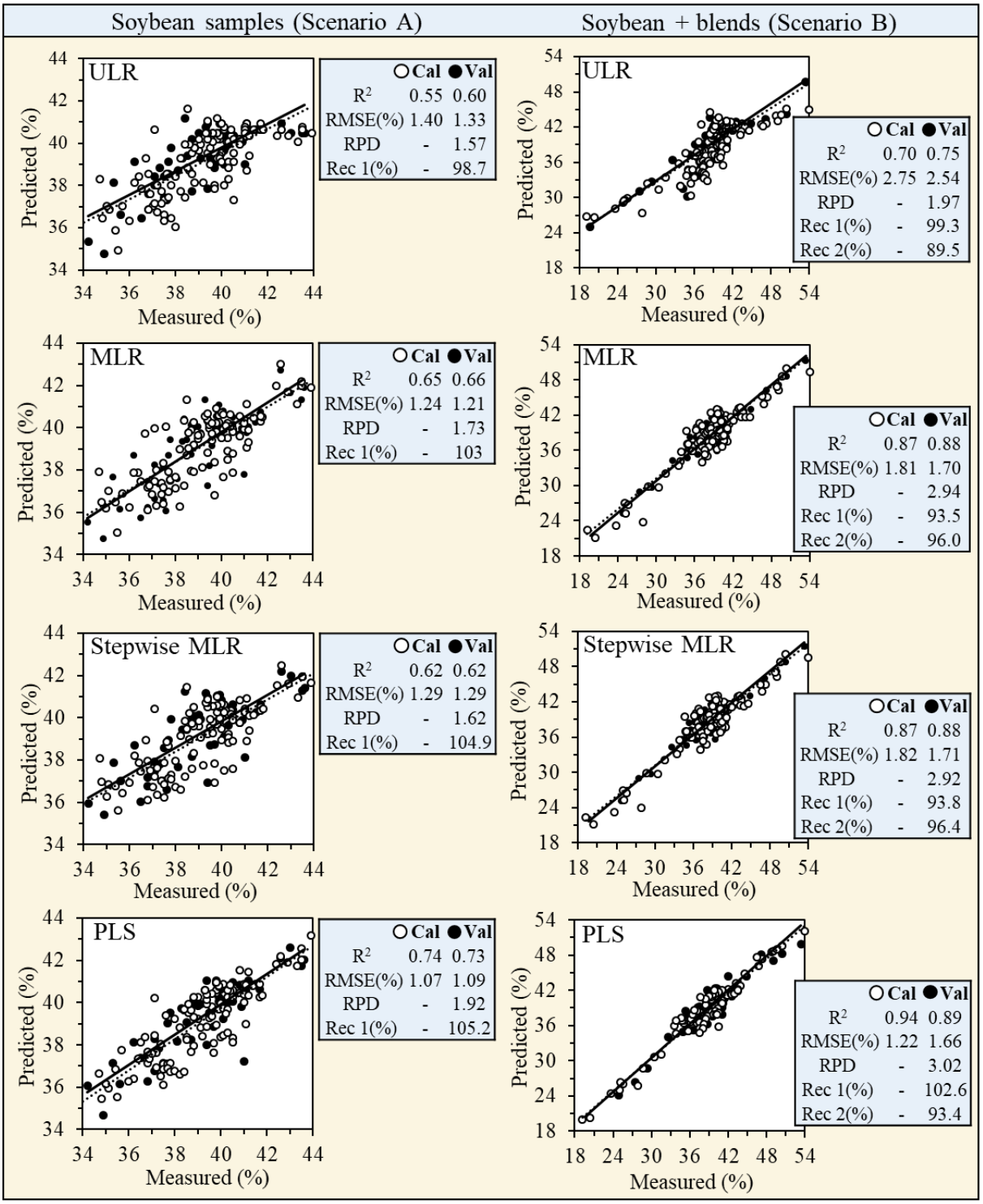


Figure 3.3 - Predictive performances achieved with scenarios A and B. The predicted and measured protein contents are expressed in % or g/100g on a dry matter basis. ULR: univariate linear regression; MLR: multiple linear regression; stepwise MLR: multiple linear regression combined with a backward stepwise variable selection; PLS: partial least squares regression; Cal and Val: refer to the calibration and validation samples, respectively; R^2: coefficient of determination; RMSE: root-mean-square error; RPD: ratio of performance to deviation. Rec 1: recovery of the model for the CRM LRI09091 reference material and Rec 2: recovery of the model for the NIST 3234 certified reference material.

Table 3.4 - Importance of Kα emission lines for protein content prediction on both datasets (soybeans and soybeans+blends). The values correspond to the z-score standardized regression coefficients of the ULR and MLR models.

	Soybean (scenario A)			Soybean + blends (scenario B)		
	MLR	Stepwise MLR	ULR[2]	MLR	Stepwise MLR	ULR[2]
P Kα[1]	-0.07			**-0.34**	**-0.34**	
S Kα[1]	**0.52**	**0.60**	**0.74**	**0.38**	**0.42**	**0.84**
K Kα[1]	-0.05			**0.32**	**0.31**	
Ca Kα[1]	**0.22**	**0.13**		0.01		
Mn Kα[1]	0.17	**0.18**		0.04		
Fe Kα[1]	**0.16**	**0.16**		**-0.76**	**-0.71**	
Zn Kα[1]	**0.16**			0.03		
Rh Compton Kα	0.13			**-0.40**	**-0.42**	
Rh Thomson Kα	-0.14			**0.21**	**0.20**	

[1]Emission line net intensity normalized by Rh Compton Kα; [2]Corresponds to the correlation coefficient of ULR; Significant regression coefficients at the probability level of 0.05 are in bold.

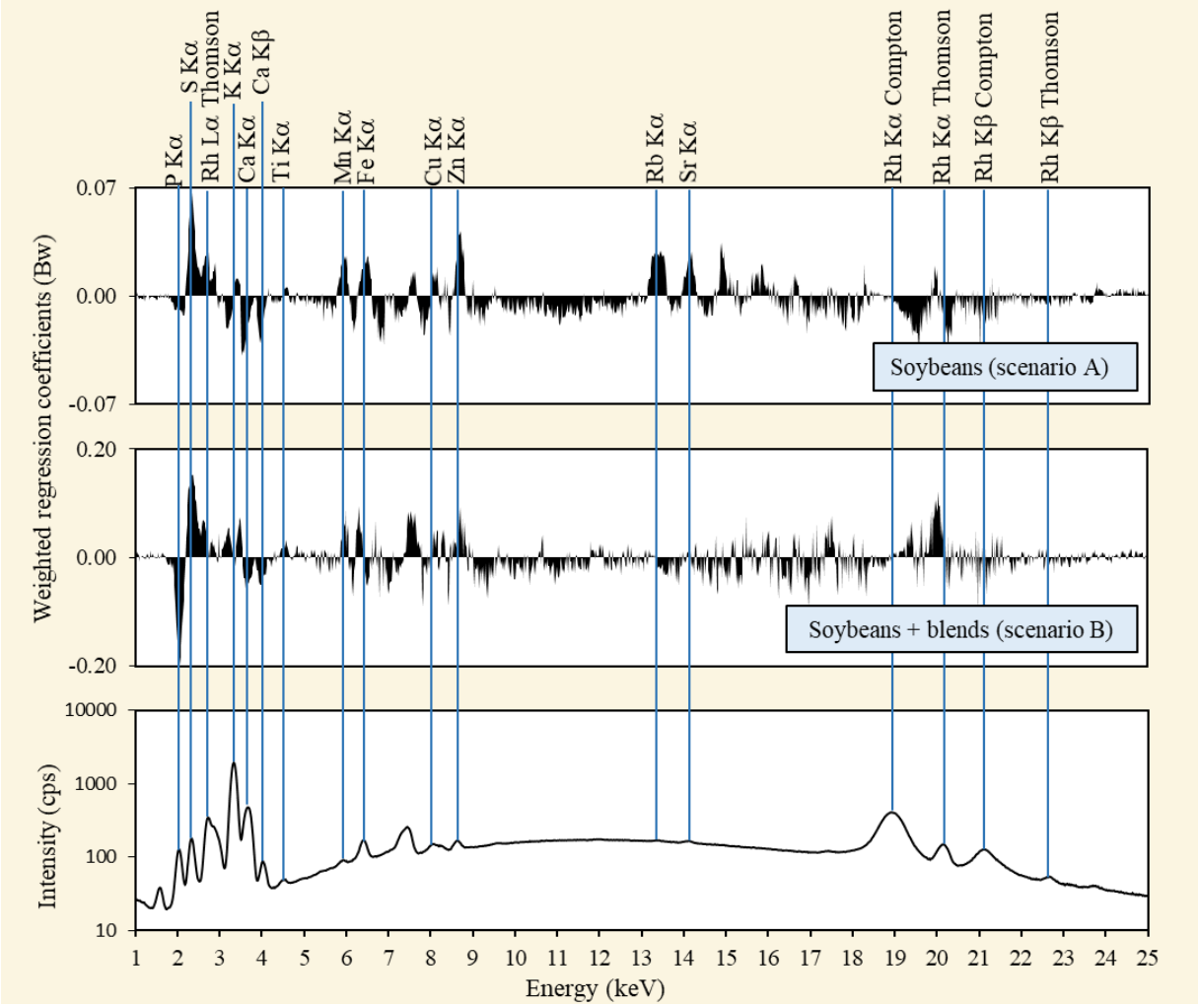


Figure 3.4 - Weighted standardized regression coefficients of the PLS models, calibrated with soybean samples (Scenario A) and soybean+blends (Scenario B), using in the models three and five latent variables, respectively.

3.3.2.2. Models calibrated with soybean and blended samples (Scenario B)

The purpose of including blends in the calibration models was to increase the protein range, going from 34% - 44% to 19% - 54%. The R_P^2 values increased after including the blends in the models (Figure 3.3), going from 0.60 to 0.75 (for ULR), from 0.66 to 0.88 (for stepwise MLR), from 0.62 to 0.88 (for MLR), and from 0.73 to 0.89 (for PLS). Thus, the performance of the models increased when using scenario B. The ULR reached a reasonable performance (RPD = 1.97), while MLR and PLS had good performances ($2.94 \leq \text{RPD} \leq 3.02$).

The R_P^2 = 0.89 obtained with PLS was slightly higher than the R_P^2 = 0.86 reported by Terra (2009), confirming that the performance of the model is influenced by the protein range. The protein range of the soybean+blends dataset (19.2% and 54%) is wider than of Terra (2009) (21 - 45%).

Again, S Kα was one of the most important emission lines for protein prediction. For example, in the PLS model, S Kα presented the highest standardized regression coefficient (2.3 keV) and in the stepwise MLR it was statically significant (p-value < 0.05). However, the importance of other variables in the models changed after including the blends. The emission lines selected in the stepwise MLR were P Kα, S Kα, K Kα, Fe Kα, Rh Kα Thomson and Rh Kα Compton (Table 3.4). Thus, K Kα, Rh Kα Compton, and Rh Kα Thomson became statistically significant (p-value < 0.05), while Mn Kα and Ca Kα became insignificant. Fe Kα presented a higher effect in the model, by means of standardized regression coefficient. Furthermore, in the PLS model, P Kα gained importance, while Rb Kα, Sr Kα and scattering peaks lost (Figure 3.4). The importance of spectral regions changed probably due to the elemental profile of the components used to produce the blends (soybean fiber flour and defatted soybean flour). For example, defatted soybean fiber presents lower P Kα, S Kα, K Kα net intensities (i.e., lower concentration of the elements) and higher Ca Kα and Fe Kα net intensities, compared to the whole grain (Figure 3.5). Most likely, these differences in composition led to a higher matrix effect on the XRF data and consequently affected the models.

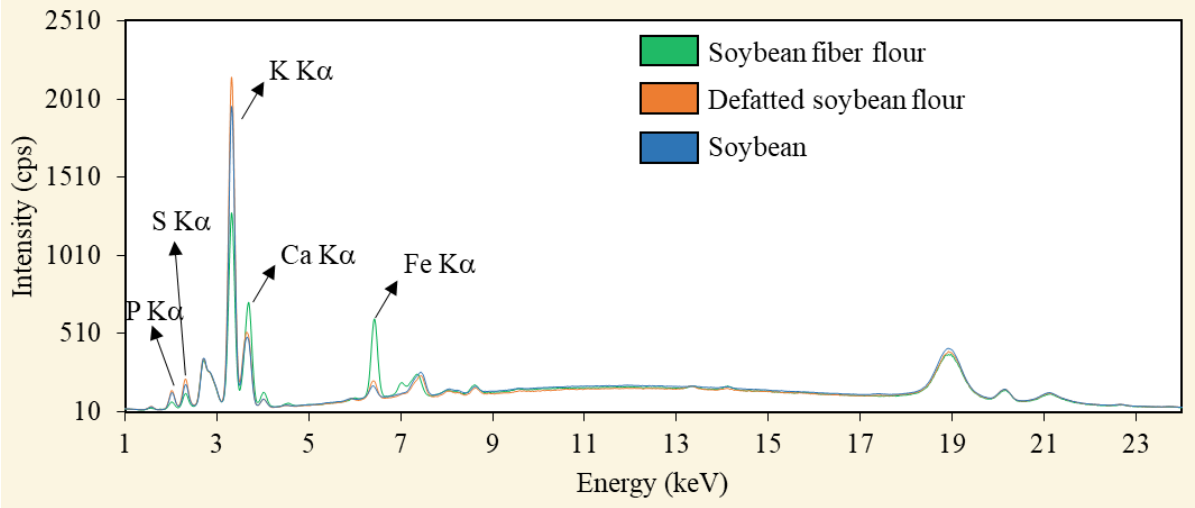


Figure 3.5 - Spectrum of soybean fiber flour, defatted soybean flour and soybean.

3.3.3. Challenges and future applications

The present study indicates that XRF can be used to predict soybean protein content, especially with calibration samples that have a wide protein range (e.g., 19.2% – 54%). As the models calibrated with soybean+blends (scenario B) were superior to those calibrated with only soybean (scenario A), we recommend future studies to use a larger sample size and include many soybean varieties to increase the protein range, and consequently, the model performance. Anyway, our results showed that it is possible to obtain soybean protein content predictions with errors lower than 2% (PLS obtained RMSEP of 1.09% in scenario A and 1.66% in scenario B). Predictions with this accuracy open opportunities for applications that require practical and rapid analysis, as highlighted below.

Sensor-based analyses, which can be conducted in the field, *in situ* or mobile laboratory, are not meant to beat traditional chemistry laboratory. They intend to offer users accessible measurements, faster results, and higher sampling density. XRF sensors offer all these features, being a technique compatible with *in situ* applications. A potential application of this sensor is soybean quality mapping. For instance, XRF sensors have been employed to directly analyze sugar cane quality parameters (MELQUIADES *et al*., 2012). Additionally, the use of sensors in sugar cane harvesters has been suggested as a promising alternative to improve data collection and quality maps (CORRÊDO *et al*., 2021). Embedding the technology into agricultural machinery allows fine-resolution surveying, which is a much sought-after form of monitoring the precision farming approaches, due to the increased mapping accuracy they bring (CORRÊDO *et al*., 2021). Sensors with reasonable performances can be applied in this case,

because the accuracy loss can be traded off by the high spatial resolution of the collected data, i.e., inconsistent values can be identified and corrected with spatial filters (MALDANER; MOLIN; SPEKKEN, 2021). For instance, Mouazen and Kuang (2016) developed reliable maps of P in agricultural soils using a sensing approach, that showed a reasonable relationship with the target variable (R^2 of 0.60).

Hence, the XRF sensor is a promising tool for the evaluation of soybean quality. It can quantify the elemental content of a wide range of elements and, as revealed in the present study, it also screens protein content. Further studies might be able to refine the method, improving its analytical quality.

3.4. Conclusion

Univariate linear regression (ULR), multivariate linear regression (MLR), and partial least squares regression (PLS) models for the prediction of soybean protein content were established, using data of the X-ray fluorescence (XRF) spectra and protein content reference values. These models presented reasonable performances ($1.57 \leq RPD \leq 1.92$), indicating that the XRF technique can be employed to predict protein content. Among the modeling strategies, PLS resulted in the highest predictive performance ($R_P^2 = 0.73$ and $RPD = 1.92$). Its superior performance was attributed to the larger number of variables used for calibration (the entire spectrum). Among the variables, the sulfur signal was the main one for protein prediction, confirming the hypothesis of the study. It happens because soybean's storage proteins contain methionine and cysteine amino acids.

Furthermore, ULR, MLR, and PLS models were also developed using soybean and samples prepared with two soybean flours (soybean fiber flour and defatted soybean flour). Encouraging R_P^2 values were obtained with these models ($0.75 \leq R_P^2 \leq 0.89$). Finally, XRF was effective in predicting soybean protein content, especially with calibration samples with a wide protein content range (e.g., 19.2% – 54%), offering a rapid, practical, and environmentally-friendly alternative method for the evaluation of soybean quality.

4. FINAL REMARKS

This research explored the potential use of X-ray fluorescence spectrometry for screening protein content in soybeans. Two calibration modeling approaches were considered: (i) to classify soybean into high- or low-protein groups and (ii) to quantify soybean protein content.

The results of the first approach indicated satisfactory performance of the sensor for that application (global accuracy and kappa index of 0.81 and 0.61, respectively), especially in the identification of high-protein soybean samples (sensitivity of 0.94). Additionally, the hypothesis that sulfur could be used as a proxy for the classification of soybeans in terms of protein content was confirmed.

Considering the results of the second approach, we concluded that XRF can reasonably quantify soybean protein content ($0.60 \leq R^2 \leq 0.73$). Models developed using soybean and samples prepared with two soybean concentrates showed encouraging R_P^2 values ($0.75 \leq R_P^2 \leq 0.89$). The performance of the models increased because the samples had a wider protein range (e.g., 19.2% – 54%). In future studies, we recommend the employment of a larger sample size and to include more soybean varieties to increase the protein range, and consequently, the model performance. The results also showed that it is possible to obtain soybean protein content predictions with errors lower than 2%. Predictions with this accuracy open opportunities for applications that require practical and rapid protein determination.

We strongly believe that soybeans will soon be traded based on its protein content, rather than solely grain weight. Thus, quick and user-friendly methods will be necessary to value the harvested grains. Although the present study revealed the potential of the XRF technique for such purpose, some advances are still necessary to make it of practical use. As a perspective, we can point out that:

i) The method here presented can be improved, we suggest to work the mathematical and sampling strategies, as well as finding a solution to bypass sample grinding;

ii) instrumentation should be adapted in order to produce cheaper instruments, which would make the technique more competitive compared to near-infrared spectrometry and nitrogen analyzers;

iii) additional value can be extracted from the technique since the information regarding mineral content, straightforwardly determined by XRF, can guide farmers on how to proceed soil fertilization during the next crop season;

iv) XRF analyzers bearing an improved version of the method proposed by this dissertation can also carry codes for analyzing leaves, soils, fertilizers and other commodities making the technique a dry chemical lab close to farmers.

REFERENCES

ADAMCHUK, V.I.; HUMMEL, J.W.; MORGAN, M.T.; AND UPADHYAYA, S.K. On-the-go soil sensors for precision agriculture. **Computers and Electronics in Agriculture**, v. 44, n. 1, p. 71-91, 2004.

ALEXANDRE, T. L.; GORAIEB, K.; BUENO, M. I. M. S. Quality control of beverages using XRS allied to chemometrics: determination of fixed acidity, alcohol and sucrose contents in Brazilian cachaça and cashew juice. **X-Ray Spectrometry**, v. 39, n. 4, p. 285-290, 2010.

ALLEGRETTA, I.; MARANGONI, B.; MANZARI, P.; PORFIDO, C.; TERZANO, R.; PASCALE, O.; SENESI, G. S. Macro-classification of meteorites by portable energy dispersive X-ray fluorescence spectroscopy (pED-XRF), principal component analysis (PCA) and machine learning algorithms. **Talanta**, v. 212, art. 120785, p. 1-9, 2020.

AOAC. **Official Methods of Analysis**. 6. ed. Arlington: AOAC International, 1997.

ARMSTRONG, P. R. Rapid single-kernel NIR measurement of grain and oil-seed attributes. **Applied Engineering in Agriculture**, v. 22, n. 5, p. 767-772, 2006.

ASSEFA, Y.; PURCELL, L. C.; SALMERON, M.; NAEVE, S.; CASTEEL, S. N.; KOVÁCS, P.; ARCHONTOULIS, S.; LICHT, M.; BELOW, F.; KANDEL, H.; LINDSEY, L. E.; GASKA, J.; CONLEY, S.; SHAPIRO, C.; ORLOWSKI, J. M.; GOLDEN, B. R.; KAUR, G.; SINGH, M.; THELEN, K.; LAURENZ, R.; DAVIDSON, D.; CIAMPITTI, I. A. Assessing variation in US soybean seed composition (protein and oil). **Frontiers in Plant Science**, v. 10, art. 298, 2019.

ATLAS OF ECONOMIC COMPLEXITY. What did Brazil export in 2020? Available at: <https://atlas.cid.harvard.edu/explore?country=undefined&queryLevel=location&product=753&year=2020&productClass=HS&target=Product&partner=undefined&startYear=undefined Accesse+d%20at%3A%2013%20Nov.%20202>. Accessed at: 2 Jan. 2023.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Regras para análise de sementes**. Brasília, DF: Ministério da Agricultura, Pecuária e Abastecimento. Secretaria de Defesa Agropecuária, 2009. Available at: <https://www.gov.br/agricultura/pt-br/assuntos/insumos-agropecuarios/arquivos-publicacoes-insumos/2946_regras_analise__sementes.pdf>>. Accessed at: 2 Jan. 2023.

BUENO, M. I. M. S.; CASTRO, M. T. P. O.; de SOUZA, A. M.; de OLIVEIRA, E. B. S.; TEIXEIRA, A. P. X-ray scattering processes and chemometrics for differentiating complex samples using conventional EDXRF equipment. **Chemometrics and Intelligent Laboratory Systems**, v. 78, p. 96-102, 2005.

BUTTROSE, M. S. Manganese and iron in globoid crystals of protein bodies from Avena and Casuarina. **Functional Plant Biology**, v. 5, n. 5, p. 631-639, 1978.

CAMARGO, R. F.; TAVARES, T. R.; DA SILVA, N. G. D. C.; DE ALMEIDA, E.; DE CARVALHO, H. W. P. Soybean sorting based on protein content using X-ray fluorescence spectrometry. **Food Chemistry**, v. 412, art. 135548, 2023.

CHAN, J. Y. L.; LEOW, S. M. H.; BEA, K. T.; CHENG, W. K.; PHOONG, S. W.; HONG, Z. W.; CHEN, Y. L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. **Mathematics**, v. 10, n. 8, art. 1283, 2022.

CHANG, C. W.; LAIRD, D. A.; MAUSBACH, M. J.; HURBURGH, C. R. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. **Soil Science Society of America Journal**, v. 65, n. 2, p. 480-490, 2001.

CHANG, S. K. C.; ZHANG, Y. Protein analysis. In: NIELSEN, S. S (Ed.). **Food analysis**. 5. ed. West. Lafayette: Springer, 2017. chap. 18, p. 315-331.

CHOUNG, M. G.; BAEK, I. Y.; KANG, S. T.; HAN, W. Y.; SHIN, D. S.; MOON, H. P.; KANG, H. K. Determination of protein and oil contents in soybean seed by near infrared reflectance spectroscopy. **Korean Journal of Crop Science**, v. 46, n. 2, p. 106-111, 2001.

COMPANHIA NACIONAL DE ABASTECIMENTO. Boletim da Safra de Grãos: 10º Levantamento Safra 2021/22. Brasília, DF, 2022. Available at: https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos/item/18435-10-levantamento-safra-2021-22 Accessed at: 6 Jan. 2020.

CORRÊDO, L. D. P.; CANATA, T. F.; MALDANER, L. F.; LIMA, J. D. J. A.; MOLIN, J. P. Sugarcane Harvester for In-field Data Collection: State of the Art, Its Applicability and Future Perspectives. **Sugar Tech**, v. 23, n. 1, p. 1-14, 2021.

DONG, Y.; QU, S. Y. Nondestructive method for analysis of the soybean quality. **International Journal of Food Engineering**, v. 8, n. 4, art. 103536, 2012.

DOS SANTOS, F. R.; DE OLIVEIRA, J. F.; BONA, E.; BARBOSA, G. M.; MELQUIADES, F. L. Evaluation of pre-processing and variable selection on energy dispersive X-ray fluorescence spectral data with partial least square regression: A case of study for soil organic carbon prediction. **Spectrochimica Acta Part B: Atomic Spectroscopy**, v. 175, art. 106016, 2021.

ECKERMANN, C.; EICHEL, J.; SCHRÖDER, J. Plant Methionine Synthase: New Insights into Properties and Expression. **Biological Chemistry**, v. 381, n. 8, p. 695-703, 2000.

FÁVERO, L. P.; BELFIORE, P. **Manual de Análise de Dados***:* Estatística e Modelagem com Excel, SPSS e Stata. 1st ed. Elsevier, 2017.

FENG, X.; ZHANG, H.; YU, P. X-ray fluorescence application in food, feed, and agricultural science: a critical review. **Critical reviews in food science and nutrition**, v. 61, n. 14, p. 2340-2350, 2021.

FERREIRA, D. S.; GALÃO, O. F.; PALLONE, J. A. L.; POPPI, R. J. Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for determination of quality parameters in soybean samples. **Food Control**, v. 35, n. 1, p. 227-232, 2014.

FERREIRA, M. M. C. **Quimiometria**: conceitos, métodos e aplicações. Campinas, SP: Editora da Unicamp, 2015.

GREDILLA, A.; DE VALLEJUELO, S. F. O.; ELEJOSTE, N.; DE DIEGO, A.; MADARIAGA, J. M. Non-destructive Spectroscopy combined with chemometrics as a tool for Green Chemical Analysis of environmental samples: A review. **TrAC Trends in Analytical Chemistry**, v. 76, p. 30-39, 2016.

GRIESHOP, C. M.; FAHEY JUNIOR, G. C. Comparison of Quality Characteristics of Soybeans from Brazil, China, and the United States. **Journal of Agricultural and Food Chemistry**, v. 49, n. 5, p. 2669-2673, 2001.

INGLE, P. D.; CHRISTIAN, R.; PUROHIT, P.; ZARRAGA, V.; HANDLEY, E.; FREEL, K.; ABDO, S. Determination of protein content by NIR spectroscopy in protein powder mix products. **Journal of AOAC International**, v. 99, n. 2, p. 360-363, 2016.

JENKINS, R. **X-ray Fluorescence Spectrometry**. 2. ed. Hoboken: John Wiley & Sons, 1999.

JIANG, G.-L. Comparison and Application of Non-Destructive NIR Evaluations of Seed Protein and Oil Content in Soybean Breeding. **Agronomy**, v. 10, n. 1, art. 77, 2020.

JUNG, S.; RICKERT, D. A.; DEAK, N. A.; ALDIN, E. D.; RECKNOR, J.; JOHNSON, L. A.; MURPHY, P.A. Comparison of Kjeldahl and Dumas Methods for Determining Protein Contents of Soybean Products. **Journal of the AOCS**, v. 80, n. 12, p. 1169-2013, 2003.

KENNARD, R. W.; STONE, L. A. Computer aided design of experiments. **Technometrics**, v. 11, n. 1, p. 137-148, 1969.

KUANG, B.; MAHMOOD, H. S.; QURAISHI, M. Z.; HOOGMOED, W. B.; MOUAZEN, A. M.; VAN HENTEN, E. J. Sensing soil properties in the laboratory, in situ, and on-line: a review. **Advances in Agronomy**, v. 114, p. 155-223, 2012.

LAI, H.; XI, J.; SUN, J.; HE, W.; WANG, Z.; ZHENG, C.; MAO, X. Multi-elemental analysis by energy dispersion X-ray fluorescence spectrometry and its application on the traceability of soybean origin. **Atomic Spectroscopy**, v. 41, n. 1, p. 20-28, 2020.

LANDGRAF, L. Soja sofre redução no teor de proteína ao longo do tempo. Embrapa Soja, 2015. Available at: https://www.embrapa.br/busca-de-noticias/-/noticia/7693893/soja-sofre-reducao-no-teor-de-proteina-ao-longo-do-tempo. Accessed at: 2 Jan. 2023.

LANDIS, J. R.; KOCH, G. G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. **Biometrics**, v. 33, n. 2, p. 363-374, 1977.

LEE, J.-S.; KIM, H.-S.; HWANG, T.-Y. Variation in Protein and Isoflavone Contents of Collected Domestic and Foreign Soybean (Glycine max (L.) Merrill) Germplasms in Korea. **Agriculture**, v. 11, n. 8, art. 735, 2021.

MA, Y.; MA, W.; HU, D.; ZHANG, X.; YUAN, W.; HE, X.; YU, D. QTL mapping for protein and sulfur-containing amino acid contents using a high-density bin-map in soybean (Glycine max L. Merr.). **Journal of Agricultural and Food Chemistry**, v. 67, n. 44, p. 12313-12321, 2019.

MALDANER, L. F.; MOLIN, J. P.; SPEKKEN, M. Methodology to filter out outliers in high spatial density data to improve maps reliability. **Scientia Agricola**, v. 79, n. 1, e20200178, 2021.

MARGUÍ, E.; QUERALT, I.; ALMEIDA, E. X-ray fluorescence spectrometry for environmental analysis: Basic principles, instrumentation, applications and recent trends. **Chemosphere**, v. 303, n.1, art. 135006, 2022.

MARUYAMA, Y.; OGAWA, K.; OKADA, T.; KATO, M. Laboratory experiments of particle size effect in X-ray fluorescence and implications to remote X-ray spectrometry, of lunar regolith surface. **Earth Planets Space**, v. 60, p. 293-297, 2008.

de MELO, F. D. A. Remuneration System of Sugarcane. In: SANTOS, F.; BORÉM, A.; CALDAS, C. (eds.). **Sugarcane**: agricultural production, bioenergy and ethanol. New York: Academic Press, 2015. chap. 19, p. 407-422.

MELQUIADES, F. L.; BORTOLETO, G. G.; MARCHIORI, L.F.S.; BUENO, M. I. M. S. Direct determination of sugar cane quality parameters by X-ray spectrometry and multivariate analysis. **Journal of Agricultural and Food Chemistry**, v. 60, n. 43, p. 10755-10761, 2012.

MIGLIORI, A.; BONANNI, P.; CARRARESI, L.; GRASSI, N.; MANDO, P.A. A novel portable XRF spectrometer with range of detection extended to low-Z elements. **X-Ray Spectrometry**, v. 40, n. 2, p. 107-112, 2011.

MILLER, J.; MILLER, J. C. **Statistics and chemometrics for analytical chemistry**. 6. ed. London: Pearson Education, 2010.

MOUAZEN, A. M.; KUANG, B. On-line visible and near infrared spectroscopy for in-field phosphorous management. **Soil and Tillage Research**, v. 155, p. 471-477, 2016.

NAZAROVNA, A. K.; BAKHROMOVICH, N. F.; ALAVKHONOVICH, K. A.; UGLI, K. S. S. Effects of Sulfur and Manganese Micronutrients on the Yield of Soybean Varieties. **Agricultural Sciences**, v. 11, n. 11, p. 1048-1059, 2020.

NIELSEN, N.P.V.; CARSTENSEN, J. M.; SMEDSGAARD, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. **Journal of Chromatography A**, v. 805, n. 1-2, p. 17-35, 1998.

OTAKA, A.; HOKURA, A.; NAKAI, I. Determination of trace elements in soybean by X-ray fluorescence analysis and its application to identification of their production areas. **Food Chemistry**, v. 147, p. 318-326, 2014.

PÍPOLO, A. E.; HUNGRIA, M.; FRANCHINI, J.C.; JUNIOR, A. A. B.; DEBIASI, H.; MANDARINO, J. M. G. **Teores de óleo e proteína em soja**: fatores envolvidos e qualidade para a indústria. Londrina, PR: Embrapa Soja, 2015. 14 p. (Comunicado Técnico, 86). Available at: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/130450/1/comunicadotecnico-86OL.pdf. Accessed at: 6 Jan. 2020.

RODRIGUES, E. S.; GOMES, M. H. F.; DURAN N. M.; CASSANJI, J. G. B.; DA CRUZ, T. N. M.; SANT'ANNA N. A.; SAVASSA, S. M.; DE ALMEIDA, E.; DE CARVALHO, H. W. P. Laboratory Microprobe X-Ray Fluorescence in Plant Science: Emerging Applications and Case Studies. **Frontiers in Plant Science**, v. 9, art. 1588, p. 1-15, 2018.

SACHS, R. C. C. Remuneração da tonelada de cana-de-açúcar no estado de São Paulo. **Informações Econômicas**, v. 37, n. 2, 2007. Available at: http://www.iea.sp.gov.br/ftpiea/ie/2007/pag%2055-66.pdf. Accessed at: 6 Jan. 2020.

SAPKOTA, Y.; MCDONALD, L. M.; GRIGGS, T. C.; BASDEN, T. J.; DRAKE, B. L. Portable X-ray Fluorescence Spectroscopy for Rapid and Cost–Effective Determination of Elemental Composition of Ground Forage. **Frontiers in Plant Science**, v. 10, art. 317, p. 1–9, 2019.

SHI, D.; HANG, J.; NEUFELD, J.; ZHAO, S.; HOUSE, J. D. Estimation of crude protein and amino acid contents in whole, ground and defatted ground soybeans by different types of near-infrared (NIR) reflectance spectroscopy. **Journal of Food Composition and Analysis**, v. 111, art. 104601, 2022.

SINGH, S.; PATEL, S.; LITORIA, N.; GANDHI, K.; FALDU, P.; PATEL, K.G. Comparative Efficiency of Conventional and NIR Based Technique for Proximate Composition of Pigeon Pea, Soybean and Rice Cultivars. **International Journal of Current Microbiology and Applied Sciences**, v. 7, n. 1, p. 773-782, 2018.

STENBERG, B.; ROSSEL, R. A. V.; MOUAZEN, A. M.; WETTERLIND, J. Visible and near infrared spectroscopy in soil science. **Advances in Agronomy**: v. 107, p. 163–215, 2010.

SUDARIĆ, A. **Soybean for human consumption and animal feed**. Rijeka, Croatia: IntechOpen, 2020.

TANGE, R. I.; RASMUSSEN, M. A.; TAIRA, E.; BRO, R. Benchmarking support vector regression against partial least squares regression and artificial neural network: Effect of sample size on model performance. **Journal of Near Infrared Spectroscopy**, v. 25, n. 6, p. 381–390, 2017.

TAVARES, T. R.; MOUAZEN, A. M.; ALVES, E. E. N.; DOS SANTOS, F. R.; MELQUIADES, F. L.; DE CARVALHO, H. W. P.; MOLIN, J. P. Assessing soil key fertility attributes using a portable X-ray fluorescence: A simple method to overcome matrix effect. **Agronomy**, v. 10, n. 6, art. 787, 2020.

TERRA, J. **Potencialidade da aliança da espectroscopia de raios X e quimiometria na determinação de valor energético e teores de alguns macronutrientes em amostras de farinhas para consumo humano**. 2009. 101 p. Tese (Doutorado) - Instituto de Química, Universidade Estadual de Campinas, Campinas, 2009.

TERRA, J.; ANTUNES, A. M.; PRADO, M. A.; BUENO, M. I. M. S. Energy value determinations of industrialized foods: the potential of using X-ray spectroscopy and partial least squares. **X-ray Spectrometry**, v. 39, n. 3, p. 167-175, 2010.

UIKEY, S.; SHARMA, S.; AMRATE, P. K.; SHRIVASTAVA, M. K. Identification of Rich Oil–Protein and Disease Resistance Genotypes in Soybean [Glycine max (L.) Merrill]. **International Journal of Bio-Resource and Stress Management**, v. 13 n. 5, p. 497-506, 2022.

UMBURANAS, R. C.; KAWAKAMI, J.; AINSWORTH, E. A.; FAVARIN, J. L.; ANDERLE, L. Z.; DOURADO–NETO, D.; REICHARDT, K. Changes in soybean cultivars released over the past 50 years in southern Brazil. **Scientific Reports**, v. 12, n. 1, p. 1–14, 2022.

USDA. **World Agricultural Production**. Table 04 Corn Area, Yield, and Production. Washington, DC, 2023. p. 27, Available at: https://apps.fas.usda.gov/psdonline/circulars/production.pdf. Accessed at: 4 Jan. 2020.

UPDAW, N. J.; BULLOCK, J. B.; NICHOLS, T. E. Pricing soybeans on the basis of oil and protein content. **Journal of Agricultural and Applied Economics**, v. 8, n. 2, p. 129-132, 1976.

VAN GRIEKEN, R. E.; MARKOWICZ, A. A. **Handbook of X-ray spectrometry**: Methods and techniques. 2. ed. Boca Raton: CRC Press, 2001.

VERBI, F. M.; PEREIRA-FILHO, E. R.; BUENO, M. I. M. S. Use of X-Ray Scattering for Studies with Organic Compounds: a Case Study Using Paints. **Microchimica Acta**, v. 150, p. 131-136, 2005.

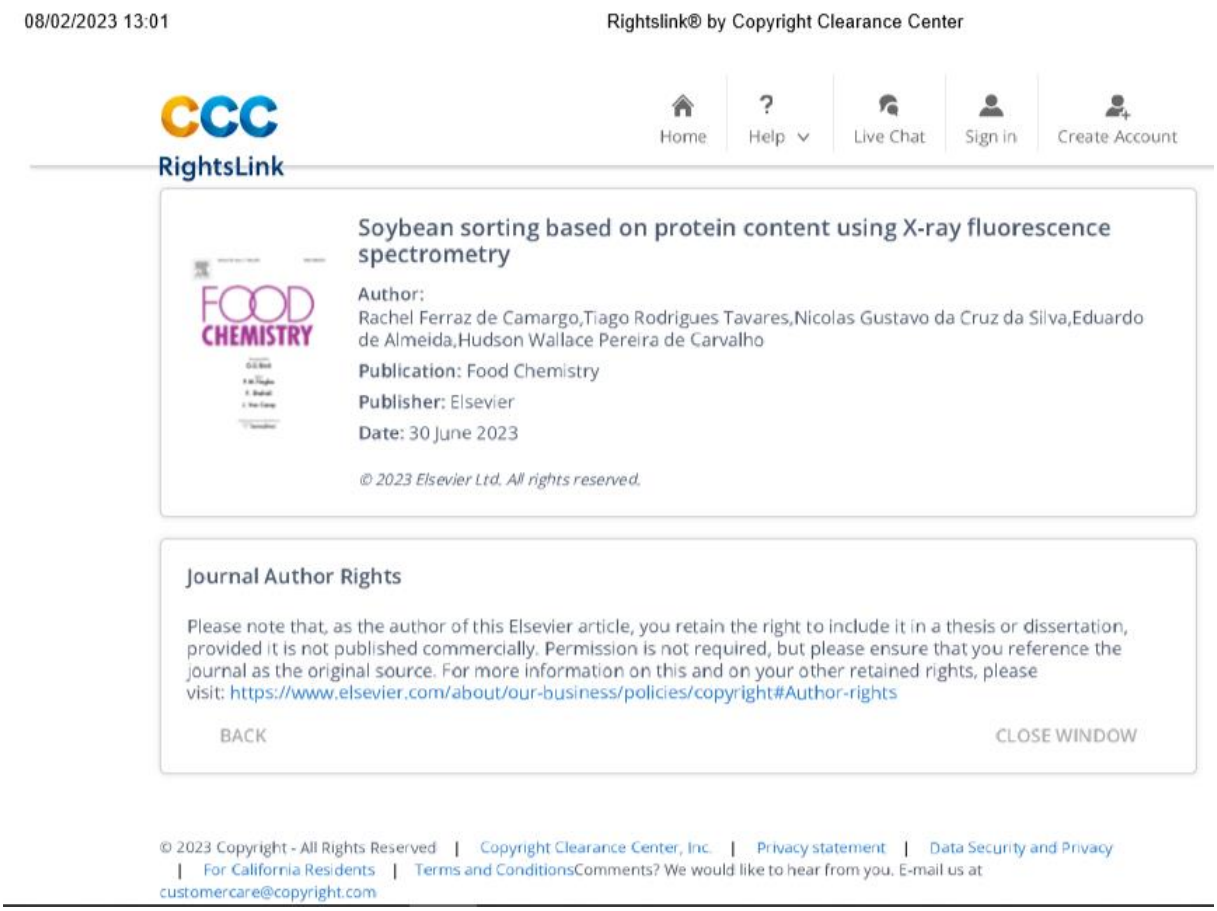
WANG, Y; ZHAO, X.; KOWALSKI, B. R. X-ray fluorescence calibration with partial least-squares. **Applied spectroscopy**, v. 44, n. 6, p. 998-1002, 1990.

WEI, X.; LI, S.; ZHU, S.; ZHENG, W.; ZHOU, S.; WU, W.; XIE, Z. Quantitative analysis of soybean protein content by terahertz spectroscopy and chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 208, art. 104199, 2021.

WILLIAM, W.; DAHL, B.; HERTSGAARD, D. Soybean quality differentials, blending, testing and spatial arbitrage. **Journal of Commodity Markets**, v. 18, art. 100095, 2020.

ZHU, Z.; CHEN, S.; WU, X.; XING, C.; YUAN, J. Determination of soybean routine quality parameters using near-infrared spectroscopy. **Food Science & Nutrition**, v. 6, n. 4, p. 1109-1118, 2018.

Apêndice A: Submitted publication



08/02/2023 13:01 Rightslink® by Copyright Clearance Center

CCC RightsLink

Home | Help | Live Chat | Sign in | Create Account

Soybean sorting based on protein content using X-ray fluorescence spectrometry

Author:
Rachel Ferraz de Camargo,Tiago Rodrigues Tavares,Nicolas Gustavo da Cruz da Silva,Eduardo de Almeida,Hudson Wallace Pereira de Carvalho

Publication: Food Chemistry

Publisher: Elsevier

Date: 30 June 2023

© 2023 Elsevier Ltd. All rights reserved.

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW

© 2023 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com