UNIVERSIDADE FEDERAL DE VIÇOSA

Índices multivariados na seleção de cultivares de soja e espectroscopia NIR para a predição do teor de proteína em grãos

Letícia Maria Sartori Carneiro
Magister Scientiae

VIÇOSA - MINAS GERAIS
2025

LETÍCIA MARIA SARTORI CARNEIRO

Índices multivariados na seleção de cultivares de soja e espectroscopia NIR para a predição do teor de proteína em grãos

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Fitotecnia, para obtenção do título de *Magister Scientiae*.

Orientador: Felipe Lopes da Silva

VIÇOSA - MINAS GERAIS
2025

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

C289i
2025

Carneiro, Letícia Maria Sartori, 1999-

Índices multivariados na seleção de cultivares de soja e espectroscopia NIR para predição do teor de proteína em grãos / Letícia Maria Sartori Carneiro. – Viçosa, MG, 2025.

1 dissertação eletrônica (69 f.): il. (algumas color.).

Orientador: Felipe Lopes da Silva.

Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Agronomia, 2025.

Inclui bibliografia.

DOI: https://doi.org/10.47328/ufvbbt.2025.681

Modo de acesso: World Wide Web.

1. Soja - Seleção. 2. Espectroscopia de infravermelho próximo. 3. Soja - Teor proteico. 4. Óleo de soja. I. Silva, Felipe Lopes da, 1981-. II. Universidade Federal de Viçosa. Departamento de Agronomia. Programa de Pós-Graduação em Fitotecnia. III. Título.

CDD 22. ed. 633.34

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

LETÍCIA MARIA SARTORI CARNEIRO

Índices multivariados na seleção de cultivares de soja e espectroscopia NIR para a predição do teor de proteína em grãos

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Fitotecnia, para obtenção do título de *Magister Scientiae*.

APROVADA: 31 de julho de 2025.

Assentimento:

_____

Letícia Maria Sartori Carneiro
Autora

_____

Felipe Lopes da Silva
Orientador

Dedico este trabalho ao meu pai, Edgar (in memoriam), que me ensinou o verdadeiro significado de força e coragem, até o seu último dia de vida. Embora não tenha tido a oportunidade de presenciar, em vida, sua finalização, sei que segue comigo em cada conquista, me acompanhando de onde estiver.

AGRADECIMENTOS

A Deus que foi minha força e alicerce durante toda a minha vida.
Aos meus pais Ana Aparecida Sartori e Edgar Elias Carneiro (in memoriam) que nunca mediram esforços para que eu pudesse chegar até aqui. Seus conselhos e amor incondicional foram fundamentais, este trabalho é um reflexo dos valores e educação que vocês me deram. Obrigada por acreditarem em mim!
A minha avó Geralda Borges de Almeida (in memoriam) que sempre me apoiou a buscar voos maiores e a nunca desistir.
Aos meus irmãos obrigada pela cumplicidade, carinho e companheirismo.
A minha companheira Emilly, por nunca soltar minha mão e ser, em todos os momentos, uma fonte constante de força, apoio.
Ao meu orientador Felipe Lopes da Silva, pela orientação e amizade.
A técnica Edna Mayer do Programa soja UFV, por todo auxilio e conselhos durante o Mestrado.
Aos estagiários do Programa Soja UFV, cuja dedicação foi essencial tanto para a condução deste trabalho quanto para a manutenção da excelência do programa de melhoramento genético.
Aos colegas da Pós-Graduação por toda ajuda e conhecimento compartilhado.
À Universidade Federal de Viçosa e ao Departamento de Agronomia (DAA), pela minha formação acadêmica e incentivo à pesquisa durante o mestrado.
Este trabalho foi realizado com o apoio das seguintes agências de pesquisa brasileiras: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).
A todos que de alguma forma contribuiu para a realização deste trabalho.
Muito Obrigada!

RESUMO

CARNEIRO, Letícia Maria Sartori, M.Sc., Universidade Federal de Viçosa, julho de 2025. **Índices multivariados na seleção de cultivares de soja e espectroscopia NIR para a predição do teor de proteína em grãos**. Orientador: Felipe Lopes da Silva.

A soja se destaca entre os cerais e leguminosas pelo elevado teor de óleo e proteína nos grãos. No entanto, a correlação negativa entre essas características representa um desafio para o desenvolvimento de cultivares com alta produtividade e qualidade nutricional. Diante disso, este trabalho teve como objetivos: i) desenvolver modelos preditivos para estimar o teor de proteína em sementes de soja por meio de espectroscopia no infravermelho próximo (NIR); e ii) avaliar a eficiência de métodos de seleção univariados e multivariados na identificação de cultivares superiores quanto a características agronômicas e de qualidade de grãos. Foram avaliados 110 cultivares de soja, analisados na forma de grãos inteiros e moídos, com valores de proteína determinados pelo método de Kjeldahl. Os modelos foram calibrados com regressão PLS, e os melhores desempenhos foram observados para grãos moídos (R^2 = 0,87; RPD = 5,75), em comparação aos grãos inteiros (R^2 = 0,82; RPD = 4,17). Na análise de seleção, foram avaliadas cinco características (TO, TP, PROD, AP e DAM) por métodos univariados e pelos índices MGIDI e FAI-BLUP. Os métodos multivariados promoveram ganhos mais equilibrados entre as características, sendo mais eficazes para seleção simultânea de cultivares com alta produtividade e qualidade nutricional. Conclui-se que a espectroscopia NIR é uma ferramenta promissora para análises não destrutivas do teor de proteína, e que os índices multivariados são mais indicados em programas de melhoramento com múltiplos objetivos.

Palavras-chave: óleo; proteína; infravermelho

ABSTRACT

CARNEIRO, Letícia Maria Sartori, M.Sc., Universidade Federal de Viçosa, July, 2025. **Multivariate indexes in the selection of soybean cultivars and NIR spectroscopy in the prediction of protein content in grains**. Adviser: Felipe Lopes da Silva.

Soybean stands out among cereals and legumes for its high oil and protein content. However, the negative correlation between these traits poses a challenge for developing cultivars with high productivity and nutritional quality. Thus, this work aimed to: i) develop predictive models to estimate protein content in soybean seeds using near-infrared (NIR) spectroscopy; and ii) evaluate the efficiency of univariate and multivariate selection methods in identifying superior cultivars for agronomic and grain quality traits. One hundred and ten soybean cultivar were evaluated, analyzed as whole and ground grains, with protein values determined by the Kjeldahl method. The models were calibrated with PLS regression, and the best performance was observed for ground grains (R^2 = 0.87; RPD = 5.75) compared to whole grains (R^2 = 0.82; RPD = 4.17). In the selection analysis, five traits (TO, TP, PROD, AP, and DAM) were evaluated using univariate methods and the MGIDI and FAI-BLUP indices. Multivariate methods promoted more balanced gains among traits, being more effective for simultaneously selecting cultivars with high productivity and nutritional quality. It is concluded that NIR spectroscopy is a promising tool for non-destructive analysis of protein content, and that multivariate indices are more suitable for breeding programs with multiple objectives.

Keywords: oil; protein; infrared

SUMÁRIO

1 INTRODUÇÃO GERAL

A soja (Glycine max (L.) Merril) é uma das culturas mais cultivadas no mundo. Na safra de 2023/24 a produção mundial do grão alcançou cerca de 396 milhões de toneladas (USDA/PSD, 2024). O Brasil é o maior produtor e exportador de soja do mundo, com produção de mais de 140 milhões de toneladas (CONAB, 2025). Esse destaque se deve ao fato da cultura ser amplamente usada na alimentação humana e animal, além de ganhar destaque no cenário de biocombustível (SILVA, et al. 2022).

Devido ao seu alto teor de proteína e óleo, a cultura da soja assume grande impacto social e econômico no cenário mundial (RANGEL, 2004). Essas características são determinantes para o valor comercial do grão de soja (BOERMA, 2004). Os cultivares de soja encontrados hoje no mercado possuem cerca de 21% e 38% de óleo e proteína, respectivamente (ROESSING e GUEDES, 1993). A busca por alimentos práticos e saudáveis têm impulsionado o processamento da soja e difundido seus derivados em diferentes segmentos industriais e alimentícios (VIEIRA et al., 1999; PENHA et al., 2007).

A alta quantidade de proteica confere a soja múltiplas aplicações na alimentação humana e segundo a Organização Mundial de Saúde (OMS) a proteína de soja contém aminoácidos que o corpo humano não pode produzir por conta própria, como lisina e leucina. Além disso, a proteína presente no grão possui boa digestibilidade e comparada a proteína animal seu valor de produção e aquisição pelo consumidor é menor (CARCIOFI et al., 2006).

O óleo de soja é o segundo óleo vegetal mais consumido no mundo (CONAB, 2019). Utilizado tanto para o consumo humano quanto como matéria-prima industrial, sua composição é influenciada tanto pelo genótipo quanto pelas condições ambientais de cultivo. Sendo sua composição representada por ácido palmítico (13%), ácido esteárico (4%), ácido oleico (20%), ácido linoleico (55%) e ácido linolênico (8%) (GOETTEL et al. 2014). No Brasil, a soja possui relevância ainda mais expressiva, sendo responsável por mais de 80% do mercado de óleos vegetais destinados à indústria alimentícia e atendendo a mais de 80% da demanda nacional para a produção de biodiesel (APROSOJA, 2019).

A demanda por soja de alta qualidade aumentará nas próximas décadas devido quantidade de produtos que surgirão para atender o crescimento populacional e à relação positiva que existe entre aumento da renda e a ingestão de proteína, sendo ela vegetal ou animal

(BHEEMANAHALLI, 2022). Desta forma, é notório que a produção de proteína e óleo precisa aumentar para atender o mercado. Existem empresas de processamento de soja no Brasil que já oferecem bonificação para fornecedores que entregam soja com níveis de proteína acima de 35% (WILLIAM et al., 2019). Da mesma maneira, existe requisitos mínimos a serem cumpridos para teor de óleo e proteína para exportação (HERTSGAARD et al., 2019).

O teor de óleo e proteína pode ser influenciado por fatores genéticos (G), ambientais (A), pela interação GxA e pelo manejo agronômico da cultura (GRASSINI et al., 2019). Por se tratar de características quantitativas, variações ambientais, como estresse térmico, hídrico e luminosidade, podem influenciar a quantidade desses componentes no grão. (PÍPOLO et al., 2002; RANGEL et al., 2004).

Os subprodutos do grão de soja são utilizados tanto na alimentação humana quanto animal, além disso, são fontes de matéria prima para diversos produtos industriais. Diante disto, existe a necessidade de realizar análises para verificar a qualidade bioquímica dos grãos de soja, bem como os teores de proteína, óleo, etc. (VIANA, 2022).

2 REFERÊNCIAS BIBLIOGRÁFICAS

APROSOJA. 2019. Aprosoja Brasil. Disponível em: < https://aprosojabrasil.com.br/estatisticas-da-soja/levantamento-de-safra/> Acesso em: 24 de junho. 2025.

BHEEMANAHALLI, R. et al. Fenotipagem de cultivares de soja do sul dos Estados Unidos para composições de peso potencial de sementes e qualidade de sementes. **Agronomy.** v.12, 839, 2022.

BOERMA, H. Roger; SPECHT, James Eugene. **Soybeans: improvement, production and uses.** 2004.

CARCIOFI, A.C.; PONTIERI, R.; FERREIRA, C.F.; PRADA, F. Avaliação de dietas com diferentes fontes protéicas para cães adultos. **Revista Brasileira de Zootecnia**, v. 35. n. 3, p.754-760, 2006.

CONAB, 2025. Companhia nacional de abastecimento. Brasília http://www.conab.gov. br/, (Acessado em 22 de maio de 2025).

CONAB. 2019. Companhia Nacional de Abastecimento. Acompanhamento da Safra Brasileira de Grãos. Décimo levantamento. Safra 2018/2019. v. 6, n.10.

GESTEIRA, G. S. et al. Selection of Early Soybean Inbred Lines Using Multiple Indices. **Crop Science**, v. 58, n. 6, p. 2494-2502, 2018.

GOETTEL, W.; XIA. E.; UPCHURCH, R.; WANG, M.L.; Chen, P.; Identification and characterization of transcript polymorphisms in soybean lines varying in oil composition and content. **BMC Genomics**. v. 15, n. 1, p. 299, 2014.

GRASSINI, P., et al.. **Soybean**. In: Sadras, V.O., Calderini, D.F. (Eds.), Crop Physiology. Case Histories for Major Crops. Elsevier, p. 283–308, 2021.

HERTSGAARD, D.J. et al.. Custos e riscos de testes e misturas para aminoácidos essenciais em soja. **Agribusiness** 35, 265–280, 2019.

PÍPOLO, A.E. **Influência da temperatura sobre as concentrações de proteína e óleo em sementes de soja (Glycine max (L.) Merrill).** Piracicaba: Escola Superior de Agricultura "Luiz de Queiroz", 2002. 128p. Tese Doutorado.

RANGEL, M. A. S. et al. **Efeito do genótipo e do ambiente sobre os teores de óleo e proteína nos grãos de soja, em quatro ambientes da região sul de Mato Grosso do Sul, safra 2002/2003**. 2004.

ROESSING,A.C.;GUEDES,L.C.A. **Cultura da Soja no Cerrado.** Piracicaba: Associação Brasileira para a Pesquisa da Potassa e do Fosfato, 1993.

SILVA, F. L. DA et al. **Soja: do plantio à colheita**. 2. ed. São Paulo, SP: **Oficina de Textos**, 2022. 312p.

USDA- United States Department of Agriculture. Foreign Agricultural Service online disponível em: Acesso em: 20/10/2024.

VIANA, Valdomiro Teixeira. **Comparativo entre os métodos nitrogênio de Kjeldahl e NIRS para análise de proteína bruta em farelo de soja**. 2022.

VIEIRA, C.R.; CABRAL, L.C.; PAULA, A.C.O. de. Proximate composition and amino acid, and fatty acid and mineral contents of six soybean cultivars for human consumption. **Pesquisa Agropecuária Brasileira**, v.34, p.1277-1283, 1999.

WILLIAM, W., Dahl, B., Hertsgaard, D.J. **Diferenciais de qualidade da soja, mistura, testes e arbitragem espacial.** J. Commod. Mark. v. 18, 100095, 2019.

Capítulo I. Eficiência de métodos de seleção em características agronômicas e de qualidade de grãos de soja

RESUMO

A soja é uma cultura grande importância econômica, destacando-se pelo alto teor de óleo e proteína nos grãos. No entanto, a correlação negativa entre essas características representa um desafio à obtenção de cultivares superiores simultaneamente produtivos e com qualidade nutricional elevada. Este trabalho teve como objetivo avaliar 110 cultivares de soja quanto a características agronômicas e de qualidade dos grãos, estimando os ganhos esperados por diferentes métodos de seleção, com ênfase na comparação entre abordagens univariadas e multivariadas. O experimento foi conduzido em delineamento de blocos aumentados (DBA), com avaliação de cinco características: teor de óleo (TO), teor de proteína (TP), produtividade (PROD), altura da planta na maturação (AP) e dias até a maturação fisiológica (DAM). Foram aplicados critérios de seleção direta e os índices multivariados MGIDI e FAI-BLUP, considerando como ideótipo plantas produtivas, com alto teor de óleo e proteína, porte adequado e ciclo precoce. Os métodos univariados resultaram em maiores ganhos para os caracteres-alvo, porém com penalizações severas em outras características. Já os índices multivariados promoveram ganhos mais equilibrados entre os atributos avaliados, sendo eficazes na identificação de cultivares com melhor desempenho geral. Os resultados demonstram que a escolha do método de seleção deve ser pautada pelos objetivos do estudo, sendo os índices multivariados recomendados quando se busca equilíbrio entre produtividade e qualidade dos grãos.

Palavras-chave: índice MGIDI; índice FAI-BLUP; seleção fenotípica; qualidade de grãos.

Chapter I. Efficiency of selection methods in agronomic and quality traits of soybean Grains

ABSTRACT

Soybean is a crop of great economic importance, standing out for its high oil and protein content in its grains. However, the negative correlation between these traits represents a challenge to obtain superior cultivars that are simultaneously productive and have high nutritional quality. This study aimed to evaluate 110 soybean cultivars regarding agronomic and grain quality traits, estimating the expected gains from different selection methods, with emphasis on the comparison between univariate and multivariate approaches. The experiment was conducted in an augmented block design, with evaluation of six traits: oil content (OC), protein content (P), productivity (PROD), plant height at maturity (HM) and days to physiological maturity (DAM). Direct selection criteria and the multivariate indices MGIDI and FAI-BLUP were applied, considering as ideotype productive plants, with high oil and protein content, adequate size and early cycle. Univariate methods resulted in greater gains for the target traits, but with severe penalties for other traits. Multivariate indices, on the other hand, promoted more balanced gains among the attributes evaluated, being effective in identifying cultivars with better overall performance. The results demonstrate that the choice of selection method should be guided by the objectives of the study, with multivariate indices being recommended when seeking a balance between productivity and grain quality.

Keywords: MGIDI index; FAI-BLUP index; phenotypic selection; grain quality.

1 INTRODUÇÃO:

A soja é uma das mais importantes culturas agrícolas, seu valor comercial depende diretamente do teor de óleo e proteína (WILSON, 2004), visto que principal forma de utilização do grão, é como matéria prima para indústria produtora de óleo e farelo (PIPOLO, 2015). Dentre os objetivos dos programas de pesquisas pode-se citar o desenvolvimento de cultivares cada vez mais promissoras quanto ao teor de óleo (TO) e proteína (TP), visando atender às exigências do mercado.

O óleo de soja é um dos óleos vegetais mais consumidos e importantes do mundo (SILVA et al., 2022). No Brasil, corresponde a mais de 80% do mercado de óleos vegetais na indústria alimentícia (Aprosoja, 2019). Além disso, é fonte de gorduras saudáveis, vitaminas, matéria prima para indústria de farmacêuticos, cosméticos, produtos de limpeza, tintas, entre outros (DEMIRBAS, 2008). O óleo presente no grão além de fonte de energia, possui ácidos graxos essenciais, dentre eles os ácidos palmítico, esteárico, oleico, linoleico e linolênico (GRAEF, 2009). O óleo de soja também é utilizado para produção de biodiesel. O biodiesel é obtido a partir de fontes renováveis, que contribuem com redução de impactos ambientais e menor prejuízo a saúde humana (ANP, 2020). Entre os meses de janeiro a dezembro de 2024, mais de 74% do biodiesel produzido no Brasil foi proveniente do óleo de soja (ABIOVE, 2024).

A soja, possui alta quantidade de proteína, cerca de 40% de proteína, quando comparada a outros grãos e leguminosas (CARCIOFI et al., 2006). O farelo de soja é a forma mais comum de utilização da soja como fonte de proteína na alimentação animal (ROSTAGNO et al., 2005). O crescimento no consumo de proteína animal tem impulsionado a demanda por farelo de soja, na formulação de rações (CARRÃO-PANIZZI, 2021).

A estimativa das correlações entre os caracteres de interesse representa uma informação valiosa, pois permite compreender como a seleção direcionada a uma característica pode impactar a expressão de outras (MIRANDA, 2006). Essas correlações são fundamentais na definição de estratégias de seleção mais eficientes e integradas. É importante compreender que a composição do grão de soja depende de muitos fatores como genótipo, ambiente que a cultura é cultivada, práticas agronômicas adotadas, e, a interação destes fatores. O genótipo é responsável por aproximadamente 50% do rendimento final da cultura, portanto é essencial que sua escolha seja criteriosa e alinhada ao objetivo de uso do cultivar (NUNES, 2015). Apesar de

serem características fortemente influenciadas pelo ambiente, a busca por cultivares com elevado potencial para produtividade, teor de proteína e óleo é essencial para atender às exigências do mercado e garantir eficiência na produção (FERREIRA, 2020). Contudo a correlação negativa e de alta magnitude entre os teores de proteína e óleo nos grãos vem dificultando a busca de cultivares com altos teores para estas características.

Para contornar a forte influência ambiental sobre as características, recomenda-se a seleção indireta por meio de traços com alta herdabilidade e boa correlação com a característica-alvo, pois permite ganhos genéticos mais consistentes (FERREIRA, 2020). No entanto, como a escolha de genótipos envolve múltiplas características, torna-se necessário o uso de índices de seleção. Os índices de seleção consistem em ferramentas multivariadas que permitem a associação de informações de vários caracteres de interesse agronômico (GRANATE et al., 2002), como os propostos por Smith (1936) e Hazel (1943). Ainda assim, a presença de colinearidade entre as características pode comprometer a confiabilidade dos índices (CARVALHO, 1995). Os índices de seleção propostos por Rocha et al. (2017) e Olivoto e Nardino (2020), demonstra capacidade de contornar tanto os problemas de multicolinearidade quanto a subjetividade na atribuição de pesos econômicos, frequentemente presente nos índices de seleção tradicionais.

Diante desse contexto, torna-se fundamental compreender as correlações entre os caracteres agronômicos e de qualidade dos grãos, de modo a estabelecer estratégias mais eficientes de seleção de cultivares com dupla aptidão. Assim, o presente trabalho teve como objetivo avaliar o desempenho de cultivares de soja quanto a características agronômicas e qualidade de grãos, estimando os ganhos com a seleção proporcionados por diferentes métodos, a fim de comparar a eficiência entre abordagens univariadas e multivariadas no processo seletivo.

2 MATERIAL E MÉTODOS

Este trabalho foi conduzido durante o período compreendido entre os meses de junho a dezembro, em 2024, na Unidade de Ensino Pesquisa e Extensão Horta Nova, (20º45'14" S e 42º52'55" O, altitude de 648 m) pertencente ao Departamento de Agronomia da Universidade Federal de Viçosa, em Viçosa, Minas Gerais. Para a sua realização foram utilizados 110 cultivares de soja (Tabela 1), representando as principais macrorregiões produtoras do Brasil, com grupos de maturidade relativa variando de 5.3 a 8.6.

O plantio foi realizado no dia 12/06/2024, seguindo as práticas culturais recomendadas para soja conforme estabelecido por Silva et al. (2022), em sistema de plantio direto. O delineamento experimental utilizado foi o de Blocos Aumentados de Federer (Federer, 1955), considerando dois blocos, com dois cultivares comuns e 108 não comuns. Cada parcela experimental foi composta de uma linha de um metro de comprimento, com espaçamento de 0,5 metros entre linhas e densidade de semeadura de 16 sementes por metro.

Foram avaliados os seguintes caracteres:

a) Produtividade (PROD) - A produtividade foi obtida na maturidade final, após a colheita, por meio da debulha dos legumes e pesagem dos grãos provenientes da área útil de cada parcela. Os valores obtidos foram expressos em gramas por parcela (g m^{-1}), com o peso corrigido para umidade padrão de 13%.

b) Teor de óleo (TO) e Teor de proteína (TP) - Os teores de óleo e proteína nos grãos de soja foram determinados por espectroscopia no infravermelho próximo (NIR), utilizando o equipamento FT-NIR (Thermo Scientific, modelo Antaris II), operando na faixa espectral de 1.000 a 2.500 nm, em modo de reflectância (R). Para a análise, foram selecionadas amostras aleatórias dos grãos para cada um dos 110 cultivares, as quais foram previamente moídas, para realizar a leitura no equipamento. Cada amostra foi analisada em triplicata, totalizando três espectros por amostra. Os espectros obtidos foram posteriormente tratados, e os teores de óleo e proteína foram estimados por meio de modelos previamente calibrados. Os resultados foram expressos em porcentagem (%).

c) Altura da planta na maturação (AP) - A altura da planta na maturação foi mensurada, em centímetros, da base da planta em contato com o solo até a extremidade da haste principal.

As medições foram realizadas após a maturação fisiológica, em plantas localizadas na área útil de cada parcela.

d) Dias até a maturação fisiológica (DAM): foi determinado por meio da contagem dos dias transcorridos entre a emergência das plântulas e o momento em que mais de 50% das plantas da parcela atingiram o estádio fenológico R8, conforme a escala proposta por Fehr e Caviness (1977).

Tabela 1 - Características dos 110 cultivares de soja avaliados quanto ao grupo de maturação relativa (GMR) e ao tipo de crescimento (TC) (continua)

CULTIVAR	GMR	TC	CULTIVAR	GMR	TC	CULTIVAR	GMR	TC
H05310_IPRO	5.3	IND	ST700_I2X	7	IND	79I81RSF_IPRO	7.9	IND
CZ15B40_IPRO	5.4	IND	TMG7067_IPRO	7	SD	790_IPRO	7.9	IND
FPS1755_IPRO	5.5	IND	71_E	7	IND	BRS267	7.9	IND
57K58RSF_CE	5.7	IND	VLP	7.1	SD	NS_7901_RR	7.9	IND
56I59RSF_IPRO	5.9	IND	74I77_RSF_IPRO	7.2	IND	80I84RSF_IPRO	8	IND
B5595_CE	5.9	IND	720_I2X	7.2	IND	80K80RSF_CE	8	DET
B5802_CE	5.9	IND	CD2728_IPRO	7.2	IND	81I84RSF_IPRO	8	IND
BRS1074_IPRO	6	IND	730_RR	7.3	IND	81I85RSF_IPRO	8	IND
NEO610_IPRO	6.1	IND	CZ37B39_I2X	7.3	IND	80I85RSF_IPRO	8	IND
TMG2375_IPRO	6.1	IND	73I70RSF_IPRO	7.3	IND	80IX81RSF_I2X	8	IND
PEKING	6.2	IND	BRS_7482_RR	7.4	IND	NS8400_IPRO	8	IND
6260RSF_IPRO	6.2	IND	TMG_123	7.4	DET	FTR_4280_IPRO	8	IND
63I64RSF_IPRO	6.3	IND	NEO740_IPRO	7.4	IND	80I79RSF_IPRO	8	IND
FPS2063_IPRO	6.3	IND	74H0112_TP_IPRO	7.4	DET	81IX82RSF_I2X	8.1	IND
NS6990_IPRO	6.3	IND	CZ37B43_IPRO	7.4	IND	81H0110_IPRO	8.1	IND
630_IPRO	6.3	IND	NEO750_IPRO	7.5	IND	M8210_IPRO	8.2	DET
64H0114_IPRO	6.4	IND	8473RSF	7.5	IND	82I86_RSF_IPRO	8.2	IND
ADV4672_IPRO	6.4	IND	IAC_100	7.5	IND	80E87RSF_E	8.2	IND

(conclusão)

64IX60RSF_I2X	6.4	IND	770_I2X	7.5	IND	82I78RSF_IPRO	8.2	IND
64H0133_IPRO	6.4	DET	7869RSF	7.6	IND	8121_IPRO	8.2	IND
64I63RSF_IPRO	6.4	IND	75I76RSF_IPRO	7.6	IND	CZ58B28IPRO	8.2	IND
NS6433_I2X	6.4	IND	CZ_47B90_IPRO	7.6	IND	M8210_IPRO	8.2	DET
65I65RSF_IPRO	6.5	IND	CZ_48B32_IPRO	7.6	IND	8321_CE3	8.3	IND
BRS_316RR	6.5	DET	76IX77RSF_I2X	7.6	IND	83H0113_TP_IPRO	8.3	IND
65K67RSF_CE	6.5	IND	76IX78RSF_I2X	7.6	IND	TMG_115	8.3	SD
TMG2165_IPRO	6.5	IND	74I78RSF_IPRO	7.7	IND	820_IPRO	8.3	IND
BRAGG	6.5	IND	BRS_133	7.7	DET	DM5.9i	8.3	IND
65IX67RSF_I2X	6.5	IND	CZ47B74_I2X	7.7	IND	AGN8019_IPRO	8.3	IND
NEO660_IPRO	6.6	DET	77E78RSF_E	7.7	IND	M8001	8.3	IND
NEO 661_I2X	6.6	IND	77H0111_I2X	7.7	IND	84I86_RSF_IPRO	8.4	IND
CD_202	6.6	DET	ST777_IPRO	7.7	IND	84I86_RSF_IPRO	8.4	IND
7166RSF_IPRO	6.6	IND	77I79RSF_IPRO	7.7	IND	840_IPRO	8.4	IND
CONQUISTA	6.8	IND	75I77RSF_IPRO	7.7	IND	84I86RSF_IPRO	8.4	IND
68I70RSF_IPRO	6.8	SD	NS7780_IPRO	7.8	DET	8576RSF	8.5	IND
68_XTD	6.8	IND	78IX80RSF_I2X	7.8	IND	C2860_E	8.6	IND
69IX69RSF_I2X	6.9	IND	TMG2378_IPRO	7.8	SD	9086RSF_IPRO	8.6	IND
NS6906_IPRO	6.9	IND	7921_IPRO	7.9	IND			

IND: indeterminada; DET: determinada; SD: semi determinada.

2.1 Análise de correlação

A análise de correlação entre as características avaliadas, foi realizada com o objetivo de avaliar a existência de colinearidade e compreender as relações entre os caracteres avaliados. A matriz de correlação foi construída utilizando a função cor do pacote corrplot software R 4.2.2 (R CORE TEAM, 2020).

2.2 Índices de Seleção

Para identificar os cultivares mais promissores quanto as características agronômicas e de qualidade, foram utilizados diferentes índices de seleção. O ideótipo foi definido com base em critérios agronômicos e de qualidade de grãos, considerando como desejáveis: alto teor de óleo e proteína, PROD e valores próximos a 90 cm para altura e menores valores de DAM.

Para cada característica, os valores genotípicos foram ajustados por meio de modelos lineares mistos (REML/BLUP), os componentes de variância foram estimados pelo método da máxima verossimilhança restrita (REML) e os valores genotípicos foram preditos pela melhor predição linear não viesada (BLUP). A predição dos efeitos genéticos foi realizada via BLUP (do inglês, *Best Linear Unbiased Prediction*), utilizando o pacote lme4 por meio do software R (R CORE TEAM, 2020).

Considerando-se o delineamento experimental em blocos aumentados de Federer, os dados foram analizados conforme o modelo estatístico:

$$y = Xf + Zg + Wb + e$$

Em que: y é o vetor de dados; f é o vetor dos efeitos assumidos como fixos (média geral); g é o vetor dos efeitos genotípicos (assumidos como aleatórios); b é o vetor dos efeitos ambientais de blocos (assumidos como aleatórios); e o vetor de erros ou resíduos (aleatórios); X, Z e W representam as matrizes de incidência para os referidos efeitos (f, g, e e, respectivamente).

A significância dos efeitos aleatórios foi avaliada por meio da análise de deviance, utilizando o teste da razão de verossimilhança (LRT). Para isso, foram comparados os modelos com e sem o efeito aleatório de interesse, sendo os testes implementados pela função ranova() do pacote lmerTest no software R. A decisão quanto à significância foi baseada na estatística do LRT, que segue uma distribuição Qui-quadrado com um grau de liberdade, conforme descrito por Resende (2007).

O coeficiente de variação genética (CVg) foi estimado utilizando a seguinte equação:

$$CVg\ (\%) = \left(\frac{\sqrt{\sigma_g^2}}{\mu}\right) \times 100$$

Em que:

σ_g^2 e μ são a variância genética e a média geral da característica avaliada, respectivamente

O coeficiente de variação residual (CVe) foi estimado através da equação:

$$CVe\ (\%) = \left(\frac{\sqrt{\sigma_e^2}}{\mu}\right)$$

Em que,

σ_g^2 é a variância residual.

A predição dos efeitos genéticos foi realizada via BLUP (do inglês, *Best Linear Unbiased Prediction*), utilizando o pacote lme4 no software R.

Os índices de seleção multivariados MGIDI (*Multi-trait Genotype-Ideotype Distance Index*), proposto por Rocha et al. (2017) e FAI-BLUP (*Factor analysis and ideotype-design*) que consideram a distância dos cultivares em relação ao um ideotipo ideal, conforme proposto por Olivoto e Nardino (2020), foram utilizados para a seleção de cultivares que se assemelham ao ideótipo proposto.

O índice FAI-BLUP baseia-se inicialmente na análise fatorial exploratória, seguida pela definição de um ideótipo, por meio da combinação dos atributos desejados e indesejados, conforme estabelecido anteriormente.

A estimação do índice FAI-BLUP é dada por:

$$P_{ij} = \frac{\frac{1}{d_{ij}}}{\sum_{i=1;j=1}^{i=n;j=m} \frac{1}{d_{ij}}}$$

Em que:

P_{ij}: probabilidade do i-ésimo cultivar (i = 1, 2, ..., n) ser semelhante ao j-ésimo ideótipo (j = 1, 2, ..., m);

d_{ij}: distância cultivar-ideótipo do i-ésimo cultivar ao j-ésimo ideótipo com base na distância euclidiana média padronizada.

O índice MGIDI, baseia-se na distância dos cultivares ao ideótipo por meio de métodos multivariados agregando todas as características de interesse do estudo. Para tanto, as características em estudo devem ser codificadas em um intervalo de 0 a 100. Assim, por meio de análises multifatorial é possível identificar estruturas de correlação. Então o ideótipo é estimado com base nas características de interesse do pesquisador e, por fim, a distância cultivar/ideótipo é estimada, segundo a equação:

$$MGIDI_I = \sqrt{\sum_{j=1}^{f} (F_{ij} - F_i)^2} \qquad (2)$$

Em que:

$MGIDI$: índice da distância cultivar/ideótipo multicaracterísticas;

F_{ij}: pontuação do i-ésimo cultivar no j-ésimo fator (i = 1, 2, ..., g; j = 1, 2, ..., f), sendo g e f o número de cultivares e fatores, respectivamente;

F_i = o j-ésimo pontuação do ideótipo.

Para cada índice de seleção aplicado, foram calculados o ganho de seleção relativo percentual (GS, %).

$$GS\ (\%) = \left(\frac{\bar{X}selecionados - \bar{X}original}{\bar{X}\ original}\right) \times 100$$

Em que, GS (%) é o ganho percentual;

$\bar{X}selecionados$ refere-se à média de 20 cultivares selecionadas e,

$\bar{X}original$ é a média original do ensaio.

Os ganhos obtidos para cada índice foram comparados com os ganhos oriundos da seleção truncada para cada característica, considerando a seleção de 20 cultivares para cada característica avaliada.

As análises para obtenção dos índices e dos respectivos ganhos foram realizadas utilizando o pacote metan no software R, versão 4.2.2.

3 RESULTADOS E DISCUSSÃO

Os resultados da Análise de Deviance para as características avaliadas estão apresentados na Tabela 2. A análise de deviance permitiu identificar as características que apresentaram variância genética significativamente diferente de zero, indicando potencial para resposta à seleção. A análise de deviance indicou significância estatística para a maioria das características avaliadas.

Embora o teor de óleo (TO) e produtividade (PROD) não foram estatisticamente significativas, foram mantidas nas etapas seguintes em razão da expressiva relevância agronômica e econômica, uma vez que constituem características-chave nos objetivos de seleção de programas de melhoramento voltados à obtenção de cultivares com elevado desempenho produtivo e qualidade de grãos.

Para a característica dias até a maturação (DAM), o valor de herdabilidade (h^2) foi igual a 1, com variância genética estimada em 28,43 e erro residual nulo. Consequentemente, os coeficientes de variação residual e a razão foram iguais a zero. Apesar disso, o LRT foi altamente significativo, indicando que a inclusão do efeito dos cultivares no modelo foi estatisticamente relevante. Em situações como essa, é importante considerar que a ausência de variância residual pode decorrer de diferentes fatores, como número limitado de repetições, uniformidade extrema nas unidades experimentais ou baixa sensibilidade do delineamento em captar variações não genéticas (CARVALHEIRO, et al. 2002). A herdabilidade igual a 1 também deve ser interpretada com cautela, pois embora teoricamente possível, é rara em condições experimentais reais e pode refletir limitações no ajuste do modelo (DE VILLEMEREUIL, et al. 2017).

As características teor de proteína (TP) e altura de planta na maturação (AP) apresentaram efeito genotípico significativo a 5% e 10% de probabilidade, com valores de Qui-quadrado (LRT) de 5,41 e 3,98, respectivamente. Sendo assim, para as características TP, AP e DAM, os componentes de variância genética diferiram significativamente de zero, o que assegura maior confiabilidade na seleção de genótipos com base nesses atributos (PIMENTEL, 2014).

Tabela 2 – Estimativas dos componentes de variância e parâmetros genéticos para teor de proteína (TP), teor de óleo (TO), produtividade de grãos (PROD), altura de planta na maturação (AP), dias até a maturidade fisiológica (DAM)

Características	μ	h^2	r_{yy}	CV_g/CV_e (%)	σ_g^2	σ_e^2
TP	34,80	0,99	0,99	33,6	4,94**	0,0045
TO	20,42	0,85	0,92	2,35	$1,45^{ns}$	0,26
PROD	259,72	0,54	0,73	1,08	$8156,80^{ns}$	7002,36
AP	83,89	0,90	0,94	12,00	223,02*	23,8
DAM	171,40	1	1	0	28,43***	0

, **, ** significativo estatisticamente (χ^2, g.l. = 1) a 10%, 5% e 1% de probabilidade, respectivamente; ns (não significativo), média (μ), herdabilidade (h^2), acurácia seletiva (r_{yy}), razão entre o coeficiente de variação genético e coeficiente de variação ambiental (CVg/CVe), variância genética entre os cultivares (σ_g^2), variância residual (σ_e^2).

A eficiência do processo de seleção está diretamente ligada à herdabilidade dos caracteres. Características com baixa herdabilidade comprometem a eficiência da seleção, pois a maior influência do ambiente sobre os genótipos reduz a proporção da variância atribuída ao controle genético e, consequentemente, o ganho esperado com a seleção (MIRANDA, 2006). Uma vez que se conhece a importância desses parâmetros é possível selecionar de forma mais estratégica e eficaz de genótipos superiores e adequar a intensidade de seleção para tal característica. De acordo com Reis et al. (2004), a herdabilidade no sentido amplo tem sido utilizada como uma medida de precisão na seleção de características agronômicas em soja.

Para a característica DAM, o modelo linear misto não convergiu durante o processo de estimação, impossibilitando o cálculo dos componentes de variância. Como consequência, não foi possível estimar a herdabilidade nem aplicar métodos de seleção para essa variável.

As herdabilidades variaram de 0,54 a 0,99. A característica com menor herdabilidade foi a PROD e maior herdabilidade foi o TP. Resultados semelhantes foi obtido por Cavallin (2020), que ao analisar a herdabilidade para diferentes características em soja, observou que a produtividade foi o caráter com menor estimativa de herdabilidade. Segundo Allard et al. (1971), alta variabilidade genética permite a aplicação de métodos seletivos mais simples e eficazes. A acurácia seletiva (r_{yy}) apresentou valores de 0,99 a 0,73. A razão entre o coeficiente de variação genético e ambiental (CVg/Cve), variou de 1,08 a 33,6%. Os maiores valores foram observados para o teor de proteína e altura de planta indicando que, para essas características,

a variabilidade genética é substancial em relação à variação experimental, o que favorece a seleção de cultivares superiores.

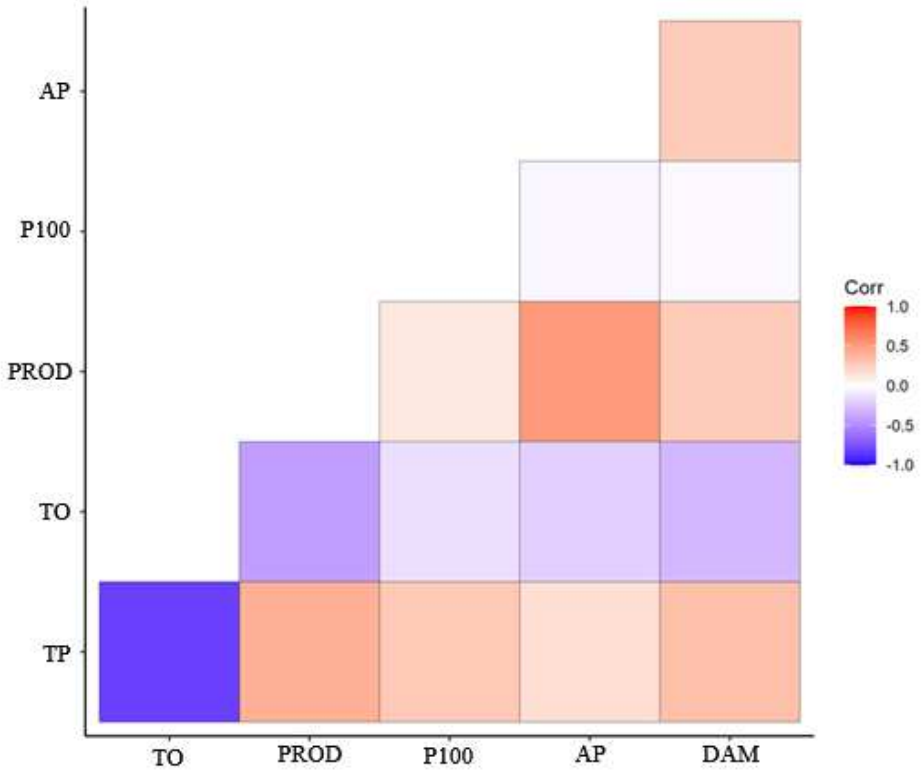
As correlações genéticas entre as características foram estimadas com base nos valores genotípicos preditos. As correlações obtidas entre os caracteres avaliados estão representadas na Figura 1. Observa-se correlação negativa entre teor de proteína e teor de óleo (-0,82), evidenciando o antagonismo clássico existente entre essas duas variáveis, que é amplamente relatado na literatura para a cultura da soja (COBER e VOLDENG, 2000; PIPOLO, 2002; MIRANDA, 2006; NAOE, 2015; VIOTTO DEL CONTE, 2020). O teor de proteína correlacionou-se positivamente com a produtividade e com a massa de cem sementes assumindo valores de 0,41 e 0,28, respectivamente, sugerindo que cultivares com maior teor de proteína apresentou maior produtividade. Embora a literatura registre predominantemente uma correlação negativa entre teor de proteína e produtividade de grãos (PIPOLO, 2002; MIRANDA, 2006, VIOTTO DEL CONTE, 2020), esse comportamento pode ser atribuído à especificidade do conjunto genético avaliado e a época de cultivo, visto que o ensaio foi conduzido entre os meses de junho a dezembro. Temperaturas mais amenas tendem a aumentar o conteúdo de proteína nos grãos, pois favorecem o metabolismo de nitrogênio, permitindo maior deposição de compostos nitrogenados (ROTUNDO e WESTGATE, 2009).

Correlação positiva de 0,52 e 0,27 também foram observadas entre peso total e altura de planta e entre peso total e dias até a maturação, respectivamente. O que sugere que cultivares mais altas e com ciclo mais longo tendem a ser mais produtivos (SOUZA, 2013).

O TO e PROD, apresentou correlação negativa de -0,417, indicando que à medida que o teor de óleo aumentou, houve uma tendência de redução na produtividade dos cultivares. Embora a literatura frequentemente aponte uma correlação positiva entre o teor de óleo e a produtividade, esse padrão pode ser fortemente influenciado pelas condições ambientais durante o ciclo da cultura (FERREIRA, 2020). No presente estudo, as temperaturas registradas durante o período experimental variaram, com máxima de 34 °C ao longo do ciclo e, durante o enchimento de grãos, temperaturas mais amenas foram observadas, com máxima de 21 °C e mínima de 11 °C. Essa variação térmica, especialmente nas fases finais do desenvolvimento, pode ter afetado negativamente o acúmulo de óleo nos grãos, o que contribui para explicar a correlação negativa observada entre essas variáveis. Temperaturas mais baixas nessa fase crítica reduzem a atividade de enzimas envolvidas na biossíntese de lipídios, resultando em menor acúmulo de óleo (ROTUNDO e WESTGATE, 2009).

Os resultados apresentados via análise das correlações evidenciam a complexidade das inter-relações entre os caracteres agronômicos e qualidade de grãos, reforçando a importância da aplicação de índices de seleção que considerem múltiplas características de forma simultânea.

Figura 1 - Mapa de correlação de Pearson para os caracteres altura de planta (AP,cm), massa de cem sementes (P100, g), produtividade (PROD,g m^{-1}), teor de óleo (TO,%), teor de proteína (TP, %), dias até a maturação fisiológica (DAM, dias)



Corr: correlação

3.1 Índices de Seleção

Os ganhos genéticos estimados com a seleção, via índices multivariados adotados e seleção truncada para as características teor de óleo (TO), teor de proteína (TP), produtividade (PROD) e altura da planta (AP) estão apresentados na Tabela 3.

Evidencia-se que, quando utilizado a seleção truncada, para todas as características analisadas houveram ganhos favoráveis após a seleção. Este resultado reforça a existência de variabilidade genética existente na população de cultivares o que favoreceu a seleção. O maior

ganho foi observado para a característica PROD, ganho percentual de 55,37%. Em seguida, destaca-se a altura da planta (AP), com ganho de 26,86%. Os teores de proteína e óleo apresentaram ganhos de 8,53% e 8,16%, respectivamente.

Esses resultados demonstram a eficácia da seleção truncada, em promover ganhos genéticos para a característica individuais de qualidade e rendimento, ainda que em proporções distintas. Contudo, em programas de melhoramento busca-se genótipos superiores para características agronômicas e de qualidade de grãos desejadas.

Neste sentido, adotou-se no presente trabalho os índices multivariados MGIDI E FAI-BLUP. Além disso, adotou-se como ideótipo cultivares com valores elevados de TO, TP e PROD, com valores próximos de 85 cm para AP e menores valores para DAM. Os resultados obtidos via seleção das cultivares pelos índices MGIDI e FAI-BLUP se encontram na Tabela 3.

Tabela 3 - Médias originais, médias após seleção e ganhos genéticos percentuais (GS,%) obtidos para os caracteres teor de proteína (TP,%), teor de óleo (TO,%), produtividade (PROD, g m^{-1}), altura da planta na maturação (AP, cm) e dias até a maturação fisiológica (DAM, dias), com base nos métodos de seleção truncada, MGIDI e FAI-BLUP

Característica	Média original	Seleção truncada		Seleção MGIDI		Seleção via FAi-BLUP	
		Média dos selecionados	GS (%)	Média dos selecionados	GS (%)	Média dos selecionados	GS (%)
TO	20.42	22.09	8.16	20.46	0.16	19.6	-4.03
TP	34.78	37.74	8.53	34.87	0.26	36.71	5.57
PROD	259.72	403.52	55.37	255.94	-1.45	336.93	29.73
AP	83.89	86,28	2.77	79.99	-4.65	92.93	10.77

Os ganhos de seleção obtidos pelo índice FAI-BLUP para as características teor de proteína e produtividade foram superiores àqueles alcançados com o uso do índice MGIDI, no entanto, para teor de óleo e altura de planta, o MGIDI apresentou desempenho superior, evidenciando variações na eficiência dos métodos conforme a característica considerada.

O uso do índice MGIDI resultou em ganhos genéticos discretos para TO e TP, de 0,16% e 0,26%, respectivamente (Tabela 3). A natureza multivariada e balanceada do índice favoreceu a manutenção da qualidade bioquímica do grão. A seleção por meio do MGIDI favoreceu a seleção de cultivares de porte reduzido, características essas desejadas no ideótipo. O ganho genético de PROD foi de -1,45. Embora tenha sido observada uma correlação genética positiva entre teor de proteína (TP) e produtividade (PROD), os ganhos de seleção indicaram aumento em TP acompanhado de redução em PROD. A correlação positiva, embora favoreça uma tendência geral de ocorrência simultânea entre os caracteres, não é suficiente para garantir ganhos simultâneos, especialmente quando a correlação não é elevada ou quando há escassez de genótipos que combinem alto TP e alta produtividade (HELMS e ORF, 1998). Aliado a isso a obtenção de cultivares superiores quanto ao TO é dificultada devido a correlação negativa com outros atributos. O acréscimo no teor de óleo (TO) foi acompanhado por uma redução na produtividade (PROD), comportamento semelhante ao relatado por Bandillo (2015).

Tais resultados evidenciam a eficácia do índice MGIDI em realizar uma seleção balanceada, favorecendo cultivares com desempenho geral compatível com o ideótipo desejado, ainda que alguns ganhos genéticos foram desfavoráveis à seleção, como o apresentado para produtividade. Além disso, reforça o desafio à seleção simultânea de características com correlação antagônica, como TO, TP e PROD, o que representa um dos principais entraves enfrentados por programas de melhoramento voltados à obtenção de cultivares altamente produtivas e com elevada qualidade de grãos (PIPOLO et al., 2015). Contudo, o índice MGIDI foi favorável à seleção de cultivares com valores de TO e TP acima da média da população original.

Os resultados obtidos no índice FAI-BLUP, demonstraram que a seleção foi eficiente em promover ganhos relevantes para as características TP, PROD e AP alinhadas ao ideótipo previamente estabelecido.

A PROD apresentou 29,73% de ganho, sendo o maior incremento observado dentre todas as variáveis. Para as características TP e AP os ganhos foram de 5,57% e 10,77%, respectivamente. A seleção via FAI-BLUP favoreceu obtenção de plantas mais altas, porém

plantas mais altas podem apresentar maior área foliar e maior potencial produtivo, desde que não haja acamamento (SANGOI, 2011). Isso se deve ao fato de que, à medida que a planta se desenvolve por mais tempo, há maior formação de nós produtivos, o que potencializa a capacidade de acúmulo de biomassa e produção de grãos (MIRANDA, 2006). O TO apresentou queda de -4,03%. Estes resultados evidenciam as correlações entre as variáveis PROD, TP e TO encontradas neste trabalho, em que, as correlações entre TO e as características PROD e TP foram negativas. Isto favoreceu a obtenção de ganhos genéticos positivos para as características PROD e TP e negativo para TO, frente ao ideótipo proposto para o índice FAI-BLUP. Resultados semelhantes foram observados por Naoe (2015) e Hayati et al. (1996), os quais destacaram as dificuldades em se obter ganhos simultâneos em características que apresentam correlação negativa expressiva.

Nas últimas décadas programas de pesquisa têm investido no desenvolvimento de cultivares produtivas com maior TO e TP (Li et al. 2018). Contudo, ambas características são antagônicas. Essa relação é um dos maiores empecilhos para o avanço simultâneo dessas variáveis, dificultando assim a obtenção de cultivares com dupla aptidão. O antagonismo entre óleo e proteína é explicado na literatura pelo fato deles competirem pelos mesmos percursores metabólicos, átomos de carbono (C) e nitrogênio (N), principalmente (PIPOLO et al., 2015). O óleo é sintetizado a partir de ácidos graxos, derivado de carbono. E a proteína é composta por aminoácidos, cuja síntese depende fortemente de C e N. Quando há maior alocação de carbono para produção de óleo, uma menor quantidade fica disponível para síntese de aminoácidos, portanto menos proteína será formada e vice-versa (HAYATI et al., 1996).

Neste sentido, o uso de índices de seleção multivariados permite a identificação de cultivares superiores de forma mais equilibrada, minimizando perdas em características agronomicamente relevantes (GRANATE et al., 2002). Nesse contexto, torna-se fundamental avaliar o desempenho desses índices em comparação à seleção univariada, uma vez que sua aplicação tende a oferecer vantagens superiores ao considerar simultaneamente múltiplas variáveis de interesse.

Métodos univariados são mais eficientes para obtenção de ganhos direcionados. Contudo, frequentemente estes métodos levam geralmente a perdas em características importantes que possuem correlações baixas com aquelas alvo da seleção. Em contrapartida, os índices multivariados, possibilita a seleção de cultivares de forma balanceada, alinhando aos múltiplos critérios de interesse agronômico, ainda que com ganhos individuais ligeiramente inferiores (DALAROSA et al., 2021). Dessa forma, a adoção de índices de seleção

multivariados representa uma estratégia robusta para programas de pesquisa voltados à obtenção de genótipos cultivares em produtividade e qualidade de grãos.

No presente trabalho foi possível verificar que a utilização do índice multivariado MGIDI, evidenciou sua principal vantagem, que é a capacidade de promover uma seleção mais balanceada, priorizando cultivares com perfil próximo do ideótipo (DALAROSA et al., 2021). Embora os ganhos tenham sido moderados em comparação à seleção univariada, o índice reduziu as perdas em características que, devido à baixa correlação com as demais, tendem a ser desfavorecidas nos métodos tradicionais. No presente trabalho o índice MGIDI foi favorável para seleção de cultivares com valores de TO e TP acima da média geral da população

Ao utilizar o índice FAI-BLUP, os resultados demonstraram excelente desempenho na maximização do ganho genético geral, especialmente em contextos onde a produtividade é o principal alvo de seleção como é o caso dos programas de melhoramento (PIPOLO et al., 2015).

4 CONCLUSÃO

A utilização de índices de seleção univariados e multivariados permitiu identificar cultivares com potencial para conciliar produtividade e qualidade de grãos.

Métodos multivariados, como MGIDI e FAI-BLUP, proporciona ganhos equilibrados entre produtividade e qualidade dos grãos, demonstrando maior eficiência em comparação às abordagens univariadas, que por sua vez se mostram eficientes para seleção direta da característica.

5 REFERÊNCIAS BIBLIOGRÁFICAS

ABIOVE. Estatísticas. Disponível em: https://biodiesel.abiove.org.br/estatisticas/. Acesso em: 20/07/2025.

ALLARD, R.W. **Principles of plant breeding**. New York: John Wiley, 485p. 1971.

APROSOJA. 2019. **Aprosoja Brasil**. Disponível em: <http://aprosojabrasil.com.br/>. Acesso em: 17/03/2025.

BANDILLO, N. et al. A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. **The Plant Genome**. v. 8, n. 3, p.1-13, 2015.

BENZAIN, B.; LANE, P.W. Protein concentration of grains in relation to some weather and soil factors during 17 years of English winter-wheat experiments. **Journal of Science of Food and Agriculture, Barking**, v. 37, p. 435-444, 1986.

BRASIL. **Agência Nacional do Petróleo, Gás Natural e Biocombustíveis**. Biodiesel. 2020. Disponível em: https://www.gov.br/anp/pt-br/assuntos/producao-e-fornecimento-de-biocombustiveis/biodiesel #:~:text=A%20obrigatoriedade%20veio%20no%20artigo,em%20todo%20o%20território%20nacional. Acesso em: 18/03/2025.

CARCIOFI, A.C.; PONTIERI, R.; FERREIRA, C.F.; PRADA, F. Avaliação de dietas com diferentes fontes protéicas para cães adultos. **Revista Brasileira de Zootecnia**, v. 35. n. 3, p.754-760, 2006.

CARRÃO-PANIZZI, Mercedes Concórdia et al. Teores de óleo e proteína em genótipos de soja em diferentes situações de manejo. **Passo Fundo**, 2021.

CARVALHO, Samuel Pereira de. **Métodos alternativos de estimação de coeficientes de trilha e índices de seleção, sob multicolinearidade.** Orientador: Cosme Damião Cruz. 1995. 163f. Tese (Doutorado Genética e Melhoramento) 1995.

CARVALHEIRO, R. et al. Efeitos da heterogeneidade de variância residual entre grupos de contemporâneos na avaliação genética de bovinos de corte. **Revista Brasileira de Zootecnia**, v. 31, p. 1680-1688, 2002.

CAVALLIN, Isabella Cristina. **Predição genômica no melhoramento de soja visando tolerância ao deficit hídrico: tamanho populacional e seleção de cultivares**. 2020.

COBER, E.R.; VOLDENG, H.D. Developing high-protein, high-yield soybean population and lines. **Crop Science**, v.40, p.39-42, 2000.

CRUZ, C. D.; REGAZZI, A. J. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: Imprensa Universitária da UFV, 1994. 390 p.

DALAROSA, L. E. et al. **Parâmetros genéticos e índice de seleção MGIDI na identificação de genótipos superiores de mandioca**. 2021.

DEMIRBAS, A. **Biodiesel.** London: Springer, 2008.

DE VILLEMEREUIL, Pierre et al. Fixed effect variance and the estimation of the heritability: Issues and solutions. **BioRxiv**, p. 159210, 2017.

FEDERER, W.T. **Experimental design:** Theory and application. New York: MacMillan, p.544, 1955.

FEHR, W. R.; CAVINESS, C. E. **Stages of soybean development**. 1977.

FERREIRA, J. M. S. **Índices de seleção baseados em valores genotípicos aplicados ao melhoramento da soja para aumento do teor de proteína.** 2020.

GESTEIRA, G. de S. et al. Seleção fenotípica de cultivares de soja precoce para a região Sul de Minas Gerais. **Revista Agrogeoambiental**, v. 7, n. 3, 2015.

GRAEF, G. et al. A high-oleic-acid and low-palmitic-acid soybean: agronomic performance and evaluation as a feedstock for biodiesel. **Plant Biotechnol J**, v. 7, n. 5, p. 411-21, 2009.

GRANATE, M. J.; CRUZ, C. D.; PACHECO, C. A. P. Predição de ganho genético com diferentes índices de seleção no milho pipoca CMS-43. **Pesquisa Agropecuária Brasileira.** v. 37, n. 7, p. 1001-1008, 2002.

HALLAUER, A.R.; MIRANDA FILHO, J.B. **Quantitative genetics in maize breeding.** Ames: Iowa State University, 468p. 1981.

HARTWIG, E.E.; HINSON, K. Association between chemical composition of seed and seed yield of soy beans. **Crop Science**, v.12, p.829-830, 1972.

HAYATI, R.; et al. Carbon and nitrogen supply during seed filling and leaf senescence in soybean. **Crop Science**, v.35, p.1063-1069, 1995.

HAYATI, R.; EGLI, D.B.; CRAFTS-BRANDNER, S.J. Independence of nitrogen supply and seed growth in soybean: studies using an in vitro culture system. **Journal Experimental Botany**, v.47, p.33-44, 1996.

HELMS, T.C.; ORF, J.H. Protein, oil, and yield in soybean lines selected for increased protein. **Crop Science**, v.38, p.707-711,1998.

HILL, J.E.; BREIDENBACH, R.W. Proteins of soybean seeds. II. Accumulation of the major protein components during seed development and maturation. **Plant Physiology**, v.53, p.747-751, 1974.

HAZEL, L. N. The genetic basis for constructing selection indexes. **Genetics,** v. 28, n. 1, p. 476-490, 1943.

LI, Y. et al. Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. **Plant Science**. v. 266, p. 95-101, 2018.

MIRANDA, F. D. **Produção, conteúdo de proteína e óleo no grão da soja: herdabilidades, correlações e seleção de genótipos superiores**. 2006.

NAOE, A. M. de L. **Efeito do déficit hídrico e épocas de semeadura sobre os teores e rendimentos de óleo e proteína em cultivares de soja no Tocantins**. 2015.

NUNES, J. L. S. Características da Soja (Glycine max). **Agrolink**, 2015.

OLIVOTO, T.; NARDINO, M. MGIDI: a novel multi-trait index for genotype in plant breeding. Bioinformatics, 2020. Disponível em < https://doi.org/10.1093/bioinformatics/btaa981> Acesso em: 20 junho de 2025.

PIMENTEL, A.J.B.; et al., Estimação de parâmetros genéticos e predição de valor genético aditivo de trigo utilizando modelos mistos. **Pesquisa Agropecuária Brasileira.** Brasília, v.49, n.11, p.882-890, 2014.

PÍPOLO, A.E. **Influência da temperatura sobre as concentrações de proteína e óleo em sementes de soja (Glycine max (L.) Merril).** 128p. 2002. (Tese) Doutorado - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2002.

PIPOLO, A. E. et al. **Teores de óleo e proteína em soja: fatores envolvidos e qualidade para a indústria.** 2015.

R CORE TEAM**. R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2020. URL https://www.R-project.org/

RAMALHO, M.A.P.; VENCOVSKY, R. Estimação dos componentes da variância genética em plantas autógamas. **Ciência e Prática**, v.2, n.2, p.117-140, 1978.

RANGEL, M.A.S. et al. Efeito do genótipo e do ambiente sobre os teores de óleo e proteína nos grãos de soja, em quatro ambientes da Região Sul de Mato Grosso do Sul, safra 2002/ 2003. Dourados: **Embrapa Agropecuária Oeste** (Boletim de pesquisa e desenvolvimento, 17, 2004.

REIS E.F.; REIS, M.S.; CRUZ, C.D.; SEDIYAMA, T.; Comparação de procedimentos de seleção para produção de grão em soja. **Ciência Rural**, v.34, p.685-692, 2004.

Resende M. D. V., Duarte J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical** 37:182–194. 2007.

RIZZO, G.; BARONI, L.. Soy, soy foods and their role in vegetarian diets. **Nutrients,** v. 10, n. 1, p. 43, 2018.

ROCHA, J. R. A. S. MACHADO, J. C. AND CARNEIRO, P. C. S., Multitrait index based on factor analysis and ideotype-design: proposal and application on elephant grass breeding for bioenergy. GCB Bioenergy. v. 95, n.1, p. 27-32, 2017.

ROSTAGNO, H. S.; et al., **Tabelas brasileiras para aves e suínos: composição de alimentos e exigências nutricionais.** Viçosa: UFV Imprensa Universitária,187p., 2005.

ROTUNDO, J. L.; WESTGATE, M. E. Meta-analysis of environmental effects on soybean seed composition. **Field Crops Research**, v. 110, n. 2, p. 147-156, 2009.

SILVA, F. L. DA et al. **Soja: do plantio à colheita**. 2. ed. São Paulo, SP: **Oficina de Textos**, 2022. 312p.

SMITH, H. F. A. discriminant function for plant selection. **Annals of Eugenics**, v. 7, n. 1, p. 240-250, 1936.

SOUZA, C. A. et al. Arquitetura de plantas e produtividade da soja decorrentes do uso de redutores de crescimento. **Bioscience Journal**, v. 29, n. 3, p. 634-643, 2013.

WilSON, R.F. Seed composition. In: BOERMA, H.R.; SPECHT, J.E. (Ed.). **Soybeans**: improvement, pro duction and uses. 3.ed. Madison: American Society of Agronomy: Crop Science Society of America: Soil Science Society of America, 2004. p. 621-677, 2004.

Capítulo II. Calibração de curva de predição utilizando espectrometria no infravermelho próximo e caracterização do teor de proteína em soja.

RESUMO

A cultura soja se destaca como uma das principais culturas agrícolas do mundo, sendo o Brasil o maior produtor e exportador mundial. A crescente demanda por produtos com alto valor nutricional e industrial, como proteínas e óleos vegetais, reforça a importância de tecnologias que possibilitem análises rápidas e confiáveis da composição dos grãos. Dentre as metodologias disponíveis, o método de Kjeldahl é amplamente utilizado como referência para quantificação de proteína, embora exija mão de obra especializada e uso de reagentes nocivos à saúde humana e meio ambiente. Como alternativa, a espectroscopia no infravermelho próximo (NIR) surge como alternativa ao método tradicional, a espectroscopia via NIR é uma técnica não destrutiva, rápida e ambientalmente mais sustentável. Neste trabalho, foram desenvolvidos modelos de regressão multivariada do tipo Partial Least Squares (PLS) para predizer o teor de proteína em sementes de soja, utilizando espectros obtidos via NIR e valores de proteína determinados pelo método de Kjeldahl. Foram avaliadas amostras de 110 cultivares, analisadas nas formas de grãos inteiros e moídos. A calibração e validação dos modelos consideraram diferentes combinações amostrais. Os resultados mostraram que os modelos gerados a partir de grãos moídos apresentaram maior acurácia (r^2 = 0,87; RPD = 5,75), em comparação aos modelos com grãos inteiros (r^2 = 0,82; RPD = 4,17), demonstrando que a forma com que o grão é avaliado no NIR e a uniformidade da amostra influencia diretamente na qualidade preditiva dos modelos. Conclui-se que a espectroscopia NIR, associada a modelos de calibração robustos, é uma ferramenta promissora para análises não destrutivas do teor de proteína em soja, com potencial de aplicação em rotinas laboratoriais e programas de melhoramento genético.

Palavras chaves: *Glycine max*, Modelagem multivariada, NIR.

Chapter II. Calibration of prediction curve using near-infrared spectrometry and characterization of protein content in soybeans.

ABSTRACT

Soybean stands out as one of the main agricultural crops in the world, with Brazil being the largest producer and exporter. The growing demand for products with high nutritional and industrial value, such as proteins and vegetable oils, reinforces the importance of technologies that enable rapid and reliable analyses of grain composition. Among the available methodologies, the Kjeldahl method is widely used as a reference for protein quantification, although it requires specialized labor and the use of reagents that are harmful to human health and the environment. As an alternative, near-infrared spectroscopy (NIR) emerges as an alternative to the traditional method; NIR spectroscopy is a non-destructive, rapid and environmentally more sustainable technique. In this work, multivariate regression models of the Partial Least Squares (PLS) type were developed to predict the protein content in soybean seeds, using spectra obtained via NIR and protein values determined by the Kjeldahl method. Samples of 110 cultivars were evaluated, analyzed in the form of whole and ground grains. The calibration and validation of the models considered different sample combinations. The results showed that the models generated from ground grains presented greater accuracy (r^2 = 0,87; RPD = 5,75), compared to models with whole grains (r^2 = 0,82; RPD = 4,17), demonstrating that the way the grain is evaluated in NIR and the sample uniformity directly influence the predictive quality of the models. It is concluded that NIR spectroscopy, associated with robust calibration models, is a promising tool for non-destructive analysis of protein content in soybeans, with potential for application in laboratory routines and genetic improvement programs.

Keywords: Glycine max, Multivariate modeling, NIR.

1 INTRODUÇÃO

Devido ao seu alto teor de proteína e óleo, a cultura da soja assume grande impacto social e econômico (RANGEL, 2004). Os cultivares de soja encontrados hoje no mercado possuem cerca de 21% e 38% de óleo e proteína, respectivamente (ROESSING e GUEDES, 1993). A demanda por soja de alta qualidade aumentará nas próximas décadas devido quantidade de produtos que surgirão para atender o crescimento populacional e pela relação positiva entre o crescimento da renda e o aumento do consumo de proteínas, sejam elas de origem vegetal ou animal (BHEEMANAHALLI, 2022).

Desta forma, é notório que a produção de proteína precisa aumentar para atender o mercado. Empresas de processamento de soja tendem a bonificar produtores que entregam soja com níveis de proteína acima de 35% (WILLIAM et al., 2019). Da mesma maneira, existe requisitos mínimos a serem cumpridos para teor de óleo e proteína para exportação (HERTSGAARD et al., 2019). Os grãos de soja destinados à exportação devem apresentar teor de proteína superior a 34%. Já o farelo de soja é classificado em três categorias conforme o teor proteico: HyPro (≥ 48%), Normal (46%) e LowPro (≤ 43,5%) (MORAES et al., 2006).

Assim, existe a necessidade de realizar análises para verificar a qualidade bioquímica dos grãos de soja, bem como os teores de proteína (VIANA, 2022). Uma das técnicas tradicionais mais utilizadas para quantificar o teor de proteína em grãos de soja foi proposta em 1883 por Kjeldahl, baseia-se na determinação de nitrogênio total (NT) presente em amostras biológicas (NOGUEIRA e SOUZA 2005). O método de Kjedhal é muito usado em rotinas laboratoriais. Além de ser confiável sofreu poucas modificações desde sua origem (GALVANI, 2006). Ele se baseia na decomposição da matéria orgânica através de reagentes químicos, como ácido sulfúrico (H2SO4), ácido bórico (H3BO3) e ácido clorídrico (HCl) (NOGUEIRA e SOUZA 2005).

Embora o método de Kjeldahl seja muito utilizado, para sua realização é necessário assim como qualquer método analítico profissionais qualificados para sua execução (VIANA, 2022). Associado a isso existe uma preocupação com a quantidade de produtos químicos que são liberados no meio ambiente como resultado de analises de rotina em diversos laboratórios no mundo inteiro, fator esse que pode gerar grandes riscos à saúde humana (GALVANI, 2006).

Diante disso, uma alternativa é a utilização de metodologias mais tecnológicas e sustentáveis, que permita alcançar os mesmos resultados de forma rápida e eficaz.

Uma alternativa aos métodos tradicionais é a utilização da Espectroscopia do Infravermelho Próximo (NIR). A espectroscopia é um método utilizado para analisar a estrutura físico-química de compostos inorgânicos, identificar grupos funcionais de substancias orgânicas ou detectar elementos químicos em sua forma simples (BARCELOS, 2007). O método do NIR é muito utilizado no setor agrícola visto que é uma técnica eficiente para determinação de gorduras, proteínas, fibras, açucares e umidade.

A região do NIR abrange a absorção de radiação em comprimentos de onda que variam de 780 a 2.500 nm (SILVA, 2006). A análise espectral nessa faixa permite identificar picos de absorção associados aos grupos funcionais C-H, N-H e O-H, encontrados na água e em compostos orgânicos como: carboidratos, óleos, álcoois e proteínas (LARIOS et al., 2020). Diversos estudos têm explorado a aplicação do NIR na cultura da soja, com objetivo de desenvolver metodologias para predição de características de interesse agronômico e industrial. Silva et al. (2020), avaliaram o uso do NIR para classificar o vigor de sementes. De forma semelhante, Soares (2023) utilizou o NIR para avaliar o potencial fisiológico de sementes de soja. Marchese (2017) empregou a técnica na construção de modelos preditivos que associassem pré-tratamentos do grão ao teor de proteína, indicando que o processamento prévio pode influenciar significativamente a acurácia das estimativas.

A utilização da espectroscopia via NIR para quantificação de proteína em soja já é estabelecida no meio científico. No entanto, é essencial que os programas de pesquisa desenvolvam modelos de regressão específicos, de acordo com seu germoplasma. Isso porque o teor de proteína varia de acordo com o genótipo e influencia os padrões espectrais e consequentemente a performance dos modelos calibrados (ZARKADAS et al., 2007). Logo, obter modelos específicos para os materiais utilizados em programas de pesquisa contribui para a obtenção de resultados mais consistentes, representativos e com maior capacidade preditiva.

Além disso, aspectos relacionados a forma da amostra também impactam diretamente na acurácia dos modelos. Embora estudos com grãos moídos sejam mais frequentes devido à maior homogeneidade espectral proporcionada pela moagem (WILLIAMS e THOMPSON, 1978). A análise de sementes inteiras vem ganhando interesse por permitir uma abordagem verdadeiramente não destrutiva (PANERO,2006). No entanto, ainda são escassos os trabalhos que comparam, de forma sistemática, o desempenho preditivo de modelos construídos a partir

de sementes inteiras em relação aos obtidos com grãos moídos, especialmente dentro de um mesmo germoplasma. Dessa forma, investigar o efeito da forma da amostra sobre a qualidade dos modelos de predição do teor de proteína pode ser considerado um avanço metodológico relevante, ainda pouco explorado na literatura.

Diante do exposto, este trabalho teve como objetivo a obtenção de modelos de regressão para predição do teor de proteína em sementes de soja, bem como, verificar se a obtenção de espectros em sementes de soja inteiras ou moídas influencia na precisão dos modelos.

2 MATERIAL E MÉTODOS

2.1 Execução Experimental

Para este trabalho foram utilizados 110 cultivares de soja, representativos de todas as macrorregiões sojícolas do Brasil, com grupo de maturidade relativa que variam de 5.3 a 8.6, conforme mostrado na Tabela 1.

Para a execução deste trabalho, foi necessária a multiplicação prévia das sementes. A multiplicação ocorreu durante o período compreendido entre os meses de junho a dezembro, em 2024, na Unidade de Ensino Pesquisa e Extensão Horta Nova, (20º45'14" S e 42º52'55" O, altitude de 648 m) pertencente ao Departamento de Agronomia da Universidade Federal de Viçosa, em Viçosa, Minas Gerais.

O plantio foi realizado no dia 12/06/2024, seguindo as práticas culturais recomendadas para soja, em sistema de plantio direto. O delineamento experimental utilizado foi o de Blocos Aumentados de Federer (Federer, 1955), considerando dois blocos, com 2 cultivares comuns e 108 não comuns. Cada parcela experimental foi composta de uma linha de um metro de comprimento, com espaçamento de 0,5 metros entre linhas e densidade de semeadura 16 sementes por metro.

Foram realizadas todas as operações de manejo conforme as exigências da cultura e o controle de pragas, doenças e plantas daninhas foi feito de forma preventiva e à medida que se fez necessário intervenção conforme estabelecido por Silva et al. (2022).

Após a colheita das plantas e obtenção dos grãos limpos, estes foram destinados separados em duas amostras. A primeira amostra foi determinada por grãos inteiros. A segunda amostra foi determinada de grãos moídos. Para tanto, os grãos foram moídos foram obtidos por meio da moagem em moinho de lâminas MARCONI MA020 LÂMINA CYCLONE, peneira 0,7 mm, visando obter amostras finas e homogêneas.

Tabela 1 - Características dos 110 cultivares de soja avaliados quanto ao grupo de maturação relativa (GMR) e ao tipo de crescimento (TC) (continua)

CULTIVAR	GMR	TC	CULTIVAR	GMR	TC	CULTIVAR	GMR	TC
H05310_IPRO	5.3	IND	ST700_I2X	7	IND	79I81RSF_IPRO	7.9	IND
CZ15B40_IPRO	5.4	IND	TMG7067_IPRO	7	SD	790_IPRO	7.9	IND
FPS1755_IPRO	5.5	IND	71_E	7	IND	BRS267	7.9	IND
57K58RSF_CE	5.7	IND	VLP	7.1	SD	NS_7901_RR	7.9	IND
56I59RSF_IPRO	5.9	IND	74I77_RSF_IPRO	7.2	IND	80I84RSF_IPRO	8	IND
B5595_CE	5.9	IND	720_I2X	7.2	IND	80K80RSF_CE	8	DET
B5802_CE	5.9	IND	CD2728_IPRO	7.2	IND	81I84RSF_IPRO	8	IND
BRS1074_IPRO	6	IND	730_RR	7.3	IND	81I85RSF_IPRO	8	IND
NEO610_IPRO	6.1	IND	CZ37B39_I2X	7.3	IND	80I85RSF_IPRO	8	IND
TMG2375_IPRO	6.1	IND	73I70RSF_IPRO	7.3	IND	80IX81RSF_I2X	8	IND
PEKING	6.2	IND	BRS_7482_RR	7.4	IND	NS8400_IPRO	8	IND
6260RSF_IPRO	6.2	IND	TMG_123	7.4	DET	FTR_4280_IPRO	8	IND
63I64RSF_IPRO	6.3	IND	NEO740_IPRO	7.4	IND	80I79RSF_IPRO	8	IND
FPS2063_IPRO	6.3	IND	74H0112_TP_IPRO	7.4	DET	81IX82RSF_I2X	8.1	IND
NS6990_IPRO	6.3	IND	CZ37B43_IPRO	7.4	IND	81H0110_IPRO	8.1	IND
630_IPRO	6.3	IND	NEO750_IPRO	7.5	IND	M8210_IPRO	8.2	DET
64H0114_IPRO	6.4	IND	8473RSF	7.5	IND	82I86_RSF_IPRO	8.2	IND
ADV4672_IPRO	6.4	IND	IAC_100	7.5	IND	80E87RSF_E	8.2	IND

(conclusão)

64IX60RSF_I2X	6.4	IND	770_I2X	7.5	IND	82I78RSF_IPRO	8.2	IND
64H0133_IPRO	6.4	DET	7869RSF	7.6	IND	8121_IPRO	8.2	IND
64I63RSF_IPRO	6.4	IND	75I76RSF_IPRO	7.6	IND	CZ58B28IPRO	8.2	IND
NS6433_I2X	6.4	IND	CZ_47B90_IPRO	7.6	IND	M8210_IPRO	8.2	DET
65I65RSF_IPRO	6.5	IND	CZ_48B32_IPRO	7.6	IND	8321_CE3	8.3	IND
BRS_316RR	6.5	DET	76IX77RSF_I2X	7.6	IND	83H0113_TP_IPRO	8.3	IND
65K67RSF_CE	6.5	IND	76IX78RSF_I2X	7.6	IND	TMG_115	8.3	SD
TMG2165_IPRO	6.5	IND	74I78RSF_IPRO	7.7	IND	820_IPRO	8.3	IND
BRAGG	6.5	IND	BRS_133	7.7	DET	DM5.9i	8.3	IND
65IX67RSF_I2X	6.5	IND	CZ47B74_I2X	7.7	IND	AGN8019_IPRO	8.3	IND
NEO660_IPRO	6.6	DET	77E78RSF_E	7.7	IND	M8001	8.3	IND
NEO 661_I2X	6.6	IND	77H0111_I2X	7.7	IND	84I86_RSF_IPRO	8.4	IND
CD_202	6.6	DET	ST777_IPRO	7.7	IND	84I86_RSF_IPRO	8.4	IND
7166RSF_IPRO	6.6	IND	77I79RSF_IPRO	7.7	IND	840_IPRO	8.4	IND
CONQUISTA	6.8	IND	75I77RSF_IPRO	7.7	IND	84I86RSF_IPRO	8.4	IND
68I70RSF_IPRO	6.8	SD	NS7780_IPRO	7.8	DET	8576RSF	8.5	IND
68_XTD	6.8	IND	78IX80RSF_I2X	7.8	IND	C2860_E	8.6	IND
69IX69RSF_I2X	6.9	IND	TMG2378_IPRO	7.8	SD	9086RSF_IPRO	8.6	IND
NS6906_IPRO	6.9	IND	7921_IPRO	7.9	IND			

IND: indeterminada; DET: determinada; SD: semi determinada.

2.2 Determinação do Teor de Proteína por Meio do Método Kjeldahl

A estimativa do teor de proteínas foi realizada pelo método de Kjeldahl por meio da quantificação do teor de nitrogênio total (NT), conforme descrito pela Association of Official Analytical Chemists (1975), para isso, parte dos grãos moídos foram utilizados.

Após a quantificação do NT determinou-se o teor de proteína bruta (TPB), em porcentagem, da amostra. Por meio da equação (1):

$$TPB = NT \times Fn \quad (1)$$

Em que Fn é o fator de correção, convencionalmente utilizado para soja de valor igual a 6,25

2.3 Obtenção dos Espectros via NIR

As amostras dos grãos de soja moídos e inteiros foram utilizadas para obtenção dos espectros via NIR. Para isto, utilizou-se o espectrômetro Antaris™ II FT-NIR Analyzer, Thermo Scientific.

Para cada amostra, foram coletados três espectros na faixa espectral de 10.000 a 4.000 cm^{-1}, equivalente ao intervalo de 1.000 a 2.500 nm, com resolução espectral de 4 cm^{-1}. Foi utilizado um recipiente com superfície em quartzo, fornecido pelo fabricante para alocar as amostras no equipamento e melhorar a passagem de luz. Posteriormente, foi realizado um média ponto a ponto em cada comprimento de onda, para gerar um único espectro representativo por amostra.

Os dados foram registrados em modo de refletância (R), sendo os resultados expressos em log (1/R), forma padrão de apresentação na espectroscopia NIR (Figuras 1 e 2). Para assegurar estabilidade e confiança das medições ao longo das análises, foi realizado o background spectrum (branco) a cada 10 amostras analisadas no NIR, afim de corrigir desvios instrumentais.

Figura 1 - Conjunto dos espectros originais obtidos dos grãos inteiros via Espectroscopia do Infravermelho Próximo (NIR)

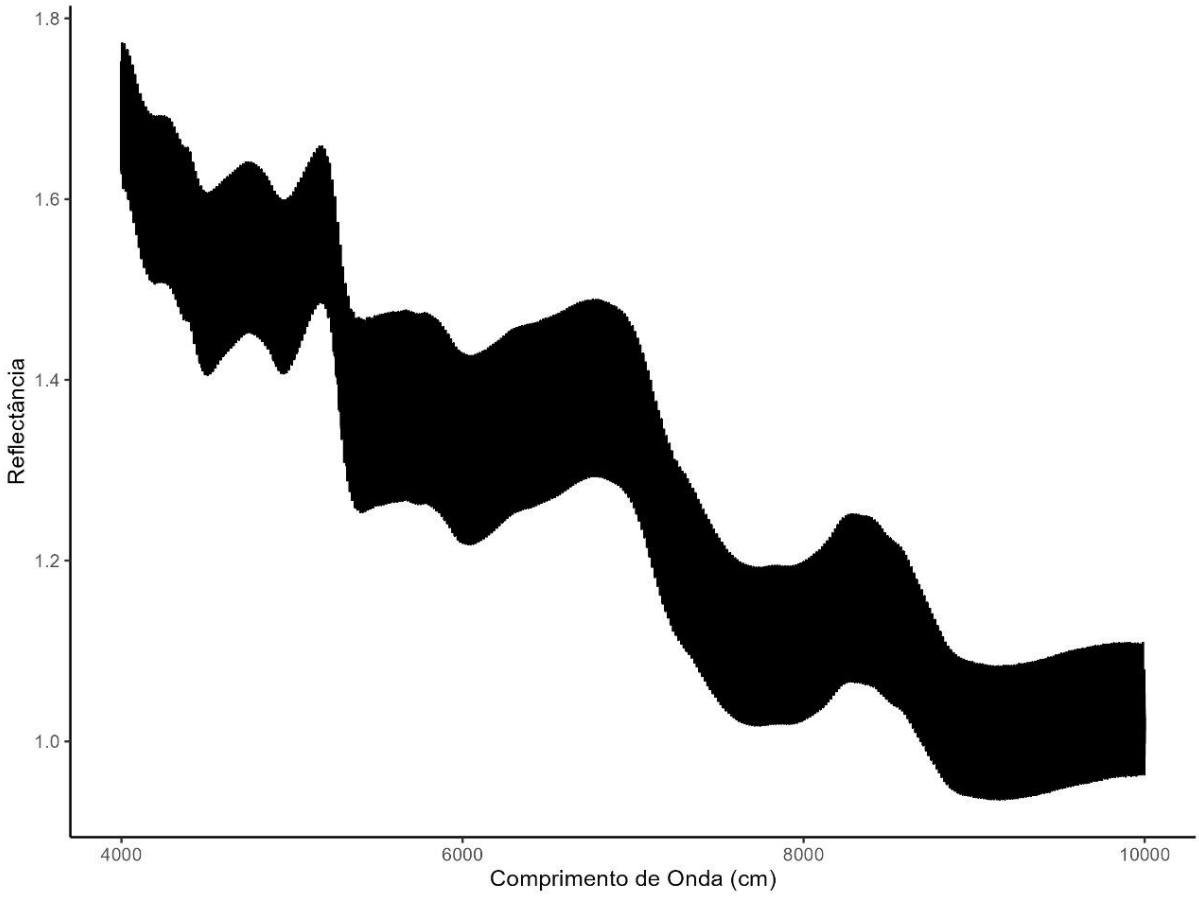
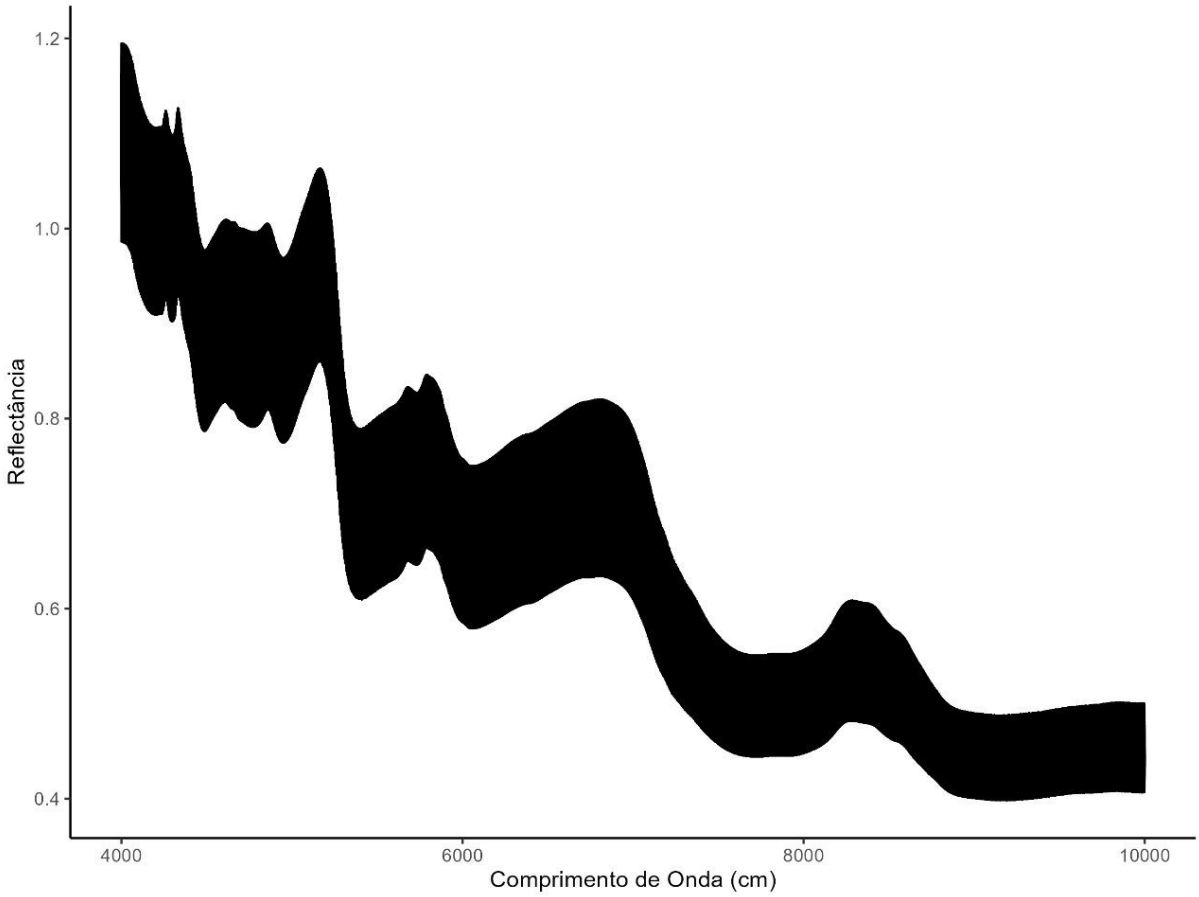


Figura 2 - Conjunto dos espectros originais obtidos dos grãos moídos via Espectroscopia do Infravermelho Próximo (NIR)



Apesar da diferença, os espectros de grãos inteiros e moídos exibem um comportamento espectral semelhante, com picos de absorbância nas mesmas regiões espectrais, porém com intensidades diferentes.

A redução da granulometria aumenta a intensidade dos picos de absorbância em deferentes regiões do espectro, indicando que a redução do tamanho das partículas favorece uma maior exposição dos constituintes químicos do grão à radiação eletromagnética, impactando diretamente o perfil espectral obtido (NOGUEIRA, 2023).

2.4 Análise de dados

O método de regressão multivariada utilizado para tratamento dos dados foi o de regressão por quadrados mínimos parciais (do inglês, *Partial Least Squares*) – PLS utilizando o pacote PLS do software R 4.2.2 (R CORE TEAM, 2020). Os dados dos espectros foram submetidos aos métodos quimiométricos de pré-tratamentos. Foram utilizados os pré-tratamentos: centragem dos dados na média (CM), auto escalar (AE), correção multiplicativa do sinal (MSC), primeira e segunda derivada. Este método de regressão é feito em duas etapas, a primeira é encontrar a relação entre matriz (X) contendo os espectros das amostras de soja do conjunto de calibração (variáveis independentes) e o vetor (Y) que armazena seus respectivos teores de proteína (variáveis dependentes) e a validação, cujo objetivo é aprimorar a capacidade do modelo em descrever com maior precisão as propriedades de interesse (SIMAS, 2005). Este método é reconhecido pela comunidade cientifica e indicado quando a matriz contém variáveis que são fortemente correlacionadas (MARTENS e NAES, 1989). Além disso, esse método tem a vantagem de poder ser utilizado em amostras que contem multicolinearidade, como é o caso do conjunto de calibração (MORGANO, 2007).

O conjunto completo de espectros foi dividido de maneira aleatória em dois subconjuntos, calibração (contendo 80% dos cultivares avaliados) e validação externa (contendo 20% dos cultivares avaliados), para determinar a robustez dos modelos gerados. Esse procedimento foi repetido dez vezes, de forma independente, para garantir maior estabilidade e confiabilidade das estimativas obtidas. A seleção dos cultivares foi realizada priorizando a representação da variabilidade genética presente no conjunto total. Essa estratégia permitiu avaliar a robustez dos modelos frente a diferentes composições amostrais, assegurando que a diversidade de respostas espectrais e composicionais estivesse adequadamente contemplada. Assim, para cada subconjunto de dados, foi realizado uma calibração, com objetivo de identificar qual deles apresenta o melhor desempenho na predição do teor de proteína. Assim, foram ajustados modelos de regressão por PLS para cada subconjunto de calibração, considerando separadamente os grupos de amostras de grãos inteiros e de grãos moídos, bem como para cada combinação de pré-tratamento espectral aplicada.

Na construção de um modelo PLS é de suma importância a escolha correta do número de variáveis latentes (VL) a ser usada. Para a determinação do número de VL utilizados em cada modelo, foi realizada uma validação cruzada do tipo Leave-One-Out (LOO), ou seja, uma

amostra do conjunto de calibração é retirada, o modelo é construído e, o teor de proteína é estimado. O processo se repete até que todas as amostras sejam previstas para 1, 2, ... VL. Para isso foi utilizado o pacote PLS (SÁIZ-ABAJO et al., 2006) do software R. O número de VL ideal é aquele que a redução do erro obtida pelo aumento da complexidade do modelo de regressão deixa de ser vantajosa (ROCHA, 2009). Assim, o número de VL se correlacionado ao modelo com maior poder preditivo é, geralmente, o escolhido.

Os dados foram analizados por meio do software R 4.2.2 (R Core Team, 2020), com o auxílio dos pacotes caret, signal, prospectr, nira e patchwork.

A capacidade do modelo de calibração de predizer o TP baseados nos dados espectrais do NIR foi avaliada usando: o erro quadrático médio, a raiz quadrada do erro quadrático médio, coeficiente de determinação e a razão de desempenho por desvio (do inglês, *Ratio of Performance to Deviation*).

Os parâmetros apresentados acima estão apresentados nas Equações (2), (3), (4), (5) respectivamente:

Erro quadrático médio (EQM):

$$EQM = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \qquad (2)$$

Em que:

$y_{i:}$ valor de referência da amostra i;

$\hat{y}_i$: valor estimado pelo modelo pela para amostra i;

n: número de amostras do conjunto de calibração.

Raiz quadrada do erro quadrático médio (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{EQM} \qquad (3)$$

Segundo Ferreira et al. (2014), quanto menor o valor de RMSE mais confiável é o modelo de predição.

Coeficiente de determinação (r^2):

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (4)$$

Em que:

$\bar{y}$: média dos valores observados.

Em um modelo de regressão o r^2 indica a proporção da variabilidade dos dados que o modelo é capaz de explicar. O seu valor pode variar de 0 a 1, em que valores próximos a 1 indica um modelo bem ajustado aos dados.

Razão de desempenho por desvio (RPD):

$$RPD = \frac{SD}{RMSE} \qquad (5)$$

Em que:

SD: desvio padrão dos valores observados no conjunto de validação.

A RPD representa a relação que existe entre o desvio padrão das amostras e o RMSE da validação cruzada. O RPD oferece embasamento para a padronização do erro padrão da predição (WILLIAMS e SOBERING, 1993). Schimleck e Evans (2004) relata que valores de RPD maior ou igual a 2,5 são considerados satisfatórios, para previsão quantitativa.

2.5 Validação do modelo

Dessa forma, após a etapa de calibração, utiliza-se o subconjunto de validação (Validation Set) para verificar a capacidade preditiva do modelo calibrado. Essa validação é realizada por meio das variáveis independentes (X), em conjunto com os coeficientes de regressão obtidos na calibração, permitindo o cálculo dos valores estimados das variáveis dependentes (Y) (MENDHAM et al., 2002). Uma vez concluída a etapa de regressão, o modelo torna-se apto para ser aplicado na predição de novas amostras (FERREIRA et al., 1999; SIMAS, R., 2005).

3 RESULTADOS E DISCUSSÃO

3.1 Modelos de predição para grãos inteiros

Os resultados dos modelos PLS calibrados para grãos inteiros, com base nos espectros obtidos por espectroscopia no infravermelho próximo (NIR) são apresentados na Tabela 2.

Tabela 2 - Desempenho dos modelos de regressão PLS na predição do teor de proteína bruta (%) em grãos de soja inteiros, com base nos espectros obtidos por espectroscopia no infravermelho próximo (NIR) e nos dados do conjunto de calibração

Modelo	VL	Pré- tratamentos utilizados					EQM	RMSE	r²	RPD
		cm	ae	MSC	der1	der2				
PLS 1	6	X					4,29	0,98	0,66	3,73
PLS 2	8		X				4,22	0,97	0,67	3,51
PLS 3	6	X	X				1,64	0,60	0,82	4,17
PLS 4	6	X	X	X			8,77	1,40	0,68	2,50
PLS 5	5	X	X		X		9,80	1,48	0,52	2,97
PLS 6	3	X	X			X	5,92	1,15	0,23	2,48

VL: variáveis latentes; cm: centralização dos dados na média; ae: auto escalar; MSC: correção multiplicativa do sinal; der1: primeira derivada; der2: segunda derivada; EQM: Erro quadrático médio; RMSE: Raiz quadrada do erro quadrático médio; r²: Coeficiente de determinação; RPD: Razão de desempenho por desvio.

O número de variáveis latentes (VL) utilizadas em cada modelo PLS foi definido com base na análise do erro quadrático médio da validação cruzada (RMSE), aplicada ao conjunto de calibração. Essa abordagem permitiu selecionar, para cada modelo, a quantidade ideal de VLs que equilibrasse o erro de predição e a robustez estatística. Os valores finais de VL utilizados nos modelos estão apresentados na Tabela 2, variando entre 3 e 8, de acordo com o desempenho obtido em cada amostragem.

Para Liu et al. (2009), um modelo preditivo com bom desempenho deve apresentar valores altos de r² e RPD e valores baixos de RMSE e EQM. O r² tem papel importante na calibração de curvas de predição via NIR, ele mede qual a proporção da variabilidade da variável dependente (Y) que pode ser explicada pela variável independente (X). No NIR o r²

também serve para avaliar o desempenho do modelo preditivo em correlacionar os espectros com variáveis de interesse.

Para grãos inteiros, o modelo de regressão que obteve maior r^2 foi o PLS 3, com valor de 0,82. E o modelo que obteve menor r^2 foi o PLS 6, apresentando valor de 0,23 (Tabela 2). Johnson et al. (2019) propõem que a qualidade dos modelos preditivos seja avaliada com base no r^2, em que valores acima de 0,75 são indicativos de modelos que fornecem um nível de predição alto. Nesse contexto, o modelo PLS 3 apresentara desempenho compatível com essa classificação. Os modelos PLS 1, PLS 2, PLS 4 e PLS 5 apresentaram valores de r^2 variando de 0,52 a 0,68, sendo classificados como modelos satisfatórios segundo Johnson et al. (2019), com desempenho preditivo moderado e passíveis de aprimoramento. Tais modelos demonstram potencial para aplicação em triagens rápidas ou etapas iniciais de seleção em programas de melhoramento genético.

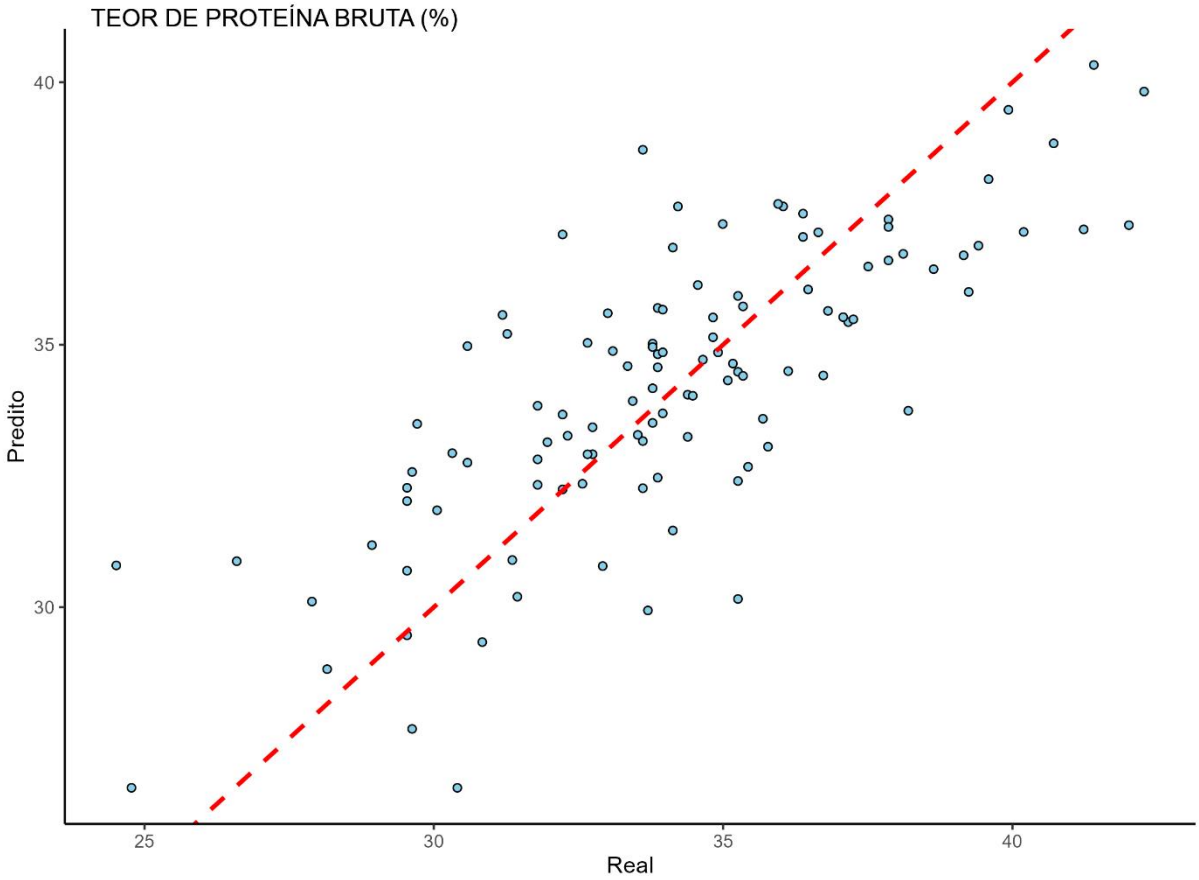
Estudos anteriores mostram que predições via NIR com grãos inteiros tendem apresentar um r^2 com valores entre 0,06 e 0,80 (PAZDERNIK et al., 1997). Essa grande variação é devida aos efeitos de espalhamento da radiação, que surgem de superfícies irregulares e da heterogeneidade das amostras analisadas, como ocorre em sementes e grãos inteiros.

O RMSE representa a raiz quadrada do erro médio quadrático da predição, ou seja, a diferença média entre os valores observados, via métodos de referência, e os valores preditos. Ele indica o quanto os valores preditos variam para mais ou para menos em relação aos valores reais, por isso, quanto menor o RMSE, maior é a acurácia do modelo. Sendo que, valores próximos de zero indicam maior precisão e confiança do modelo para fins de predição (MATSIMBE et al., 2015). Dentre os modelos ajustados, o PLS 3, 2 e 1 foram os que apresentaram menor RMSE, sendo seus valores 0,60, 0,97 e 0,98, respectivamente (Tabela 2). Esses modelos foram construídos com a aplicação de pré-tratamentos de centralização dos dados na média e auto escalamento. Os resultados obtidos neste estudo são consistentes com os valores relatados na literatura, Oliveira et al. (2013) relataram resultados semelhantes ao investigarem a variação nos teores de proteína em grãos de soja inteiros por meio de modelos PLS com espectroscopia NIR, obtendo valores de RMSE de 0,64. Armstrong et al. (2011) relataram valor de RMSE igual a 0,97 para grãos inteiros de soja avaliados via NIR, reforçando a confiabilidade dos modelos desenvolvidos e sua aplicabilidade na predição do teor de proteína em soja em grãos inteiros via NIR.

A razão RPD também foi utilizada para avaliar o desempenho e qualidade dos modelos preditos. O valor de RPD é a razão entre o desvio padrão dos valores de referência e o RMSE do modelo de predição. Ou seja, ele indica o quão bem o modelo calibrado explica a variabilidade natural existente nos dados. Um valor de RPD alto indica modelo mais preciso. Zhang et al. (2015) indicam que modelos ideais possuem RPD acima de 3 apresentando excelente previsibilidade do modelo. Contudo, RPD maior que 2,5 já é aceitável para algumas aplicações como controle de qualidade e para quantificação. Os modelos com maior RPD observados foram o PLS3, PLS1 e PLS 2, cujos valores foram 4,17, 3,73 e 3,51 (Tabela 2). Aulia et al. (2022), ao avaliarem a predição não destrutiva do teor de proteína em sementes de soja por meio de imagens espectrais no infravermelho próximo, obtiveram valores de RPD variando de 3,08 a 4,22 nos modelos ajustados. Panero (2007), obteve valor de RPD de 2,93 quantificação do teor proteína e controle de qualidade em grãos de soja.

De modo geral, dentre os resultados apresentados o modelo PLS 3 apresentou melhor desempenho quando comparado aos outros modelos calibrados para grãos inteiros, o que determinou a maior capacidade preditiva, para teor de proteína nos grãos (Figura 3). Para avaliar o desempenho do modelo ajustado, foi gerado um gráfico de dispersão com os valores observados, obtidos por método de referência e os valores preditos pelo modelo de regressão PLS 3, utilizando os espectros NIR obtidos a partir de grãos de soja inteiros. A reta tracejada representa a linha de 1:1, indicando o ideal de predição perfeita. A avaliação foi realizada com base nos 110 cultivares estudados, configurando uma avaliação interna do ajuste do modelo.

Figura 3 - Dispersão dos valores observados versus preditos para o teor de proteína bruta (%) em grãos de soja inteiros, utilizando o modelo PLS 3 ajustado com espectros obtidos por espectroscopia no infravermelho próximo (NIR)



Os teores de proteína predito, utilizando o modelo com melhor ajuste PLS 3, apresentou variações em relação aos valores de referência obtidos via método Kjeldahl entre 0,26% e 8,43%. Isso mostra que o modelo é moderadamente eficaz para fins de predição em análises não destrutivas. Embora seja fundamental cautela em sua aplicação, é recomendada a atualização contínua do modelo com a incorporação de novos dados, visando aprimorar a acurácia preditiva, reduzir os erros e explorar o uso de diferentes técnicas de pré-processamento.

3.2 Modelos de predição para grãos moídos

Os resultados dos modelos PLS calibrados para grãos moídos, com base nos espectros obtidos por espectroscopia no infravermelho próximo (NIR) são apresentados na Tabela 3.

Tabela 3 - Desempenho dos modelos de regressão PLS na predição do teor de proteína bruta (%) em grãos de soja moídos, com base nos espectros obtidos por espectroscopia no infravermelho próximo (NIR) e nos dados do conjunto de calibração

Modelo	VL	Pré- tratamentos utilizados					EQM	RMSE	r^2	RPD
		cm	ae	MSC	der1	der2				
PLS 1	5	X					4,22	0,97	0,78	4,49
PLS 2	5		X				2,84	0,79	0,76	4,41
PLS 3	5	X	X				1,93	0,66	0,81	3,99
PLS 4	3	X	X	X			4,03	0,95	0,75	4,13
PLS 5	2	X	X		X		1,51	0,58	0,87	5,75
PLS 6	1	X	X			X	4,08	0,95	0,55	3,20

VL: variáveis latentes; cm: centralização dos dados na média; ae: auto escalar; MSC: correção multiplicativa do sinal; der1: primeira derivada; der2: segunda derivada; EQM: Erro quadrático médio; RMSE: Raiz quadrada do erro quadrático médio; r^2: Coeficiente de determinação; RPD: Razão de desempenho por desvio.

Conforme apresentado na Tabela 3, os modelos calibrados com os espectros de grãos moídos utilizaram entre 1 e 5 variáveis latentes, a depender do desempenho obtido em cada subconjunto. Os resultados indicam que os espectros NIR de grãos de soja moídos fornecem alta qualidade de informação espectral, permitindo a construção de modelos preditivos eficientes com baixo número de VL. A moagem contribuiu para a redução da heterogeneidade das amostras, favorecendo maior correlação entre os espectros e os teores de proteína determinados por método de referência.

No modelo PLS 6 foi selecionado apenas uma VL. Embora isso possa indicar que a maior parte da variação explicativa da variável resposta está concentrada em um único componente, também pode refletir uma estrutura de dados simplificada, possivelmente decorrente da homogeneidade espectral promovida pela moagem. No entanto, o uso de apenas uma VL pode resultar em subajuste, pois não incorpora toda a complexidade e variabilidade potencialmente presente nos dados (LAMMERTYN et al. 1998). Isso pode limitar a capacidade

preditiva do modelo, principalmente quando aplicado a novos conjuntos de dados. De acordo com Rocha (2009), o número ótimo de variáveis latentes é aquele que equilibra a redução do erro de modelagem com o risco de sobreajuste. Portanto, valores muito baixos, como observado neste caso, devem ser interpretados com cautela, podendo indicar que o modelo está excessivamente simplificado e sensível a pequenas variações dos dados.

Para amostras de grãos moídos, os valores de r^2 variou de 0,55 a 0,87 (Tabela 3). O modelo que obteve o melhor desempenho foi o PLS 5 e o modelo PLS 6 apresentou o pior ajuste, com baixo poder explicativo e maior erro preditivo entre os modelos avaliados. Johnson et al. (2019), propôs a classificação da qualidade dos modelos preditivos com base no r^2, indicando valores de r^2 acima de 0,75 indicam modelos com alto nível de predição, dentre os modelos calibrados para grãos moídos o PLS 1, PLS 2, PLS 3, PLS 4 e PLS 5 estão dentro dessa classificação. Esses resultados demonstram o melhor desempenho dos modelos calibrados com grãos moídos em comparação aos obtidos com grãos inteiros, refletindo o aumento da eficiência espectral proporcionada pela homogeneização física da amostra.

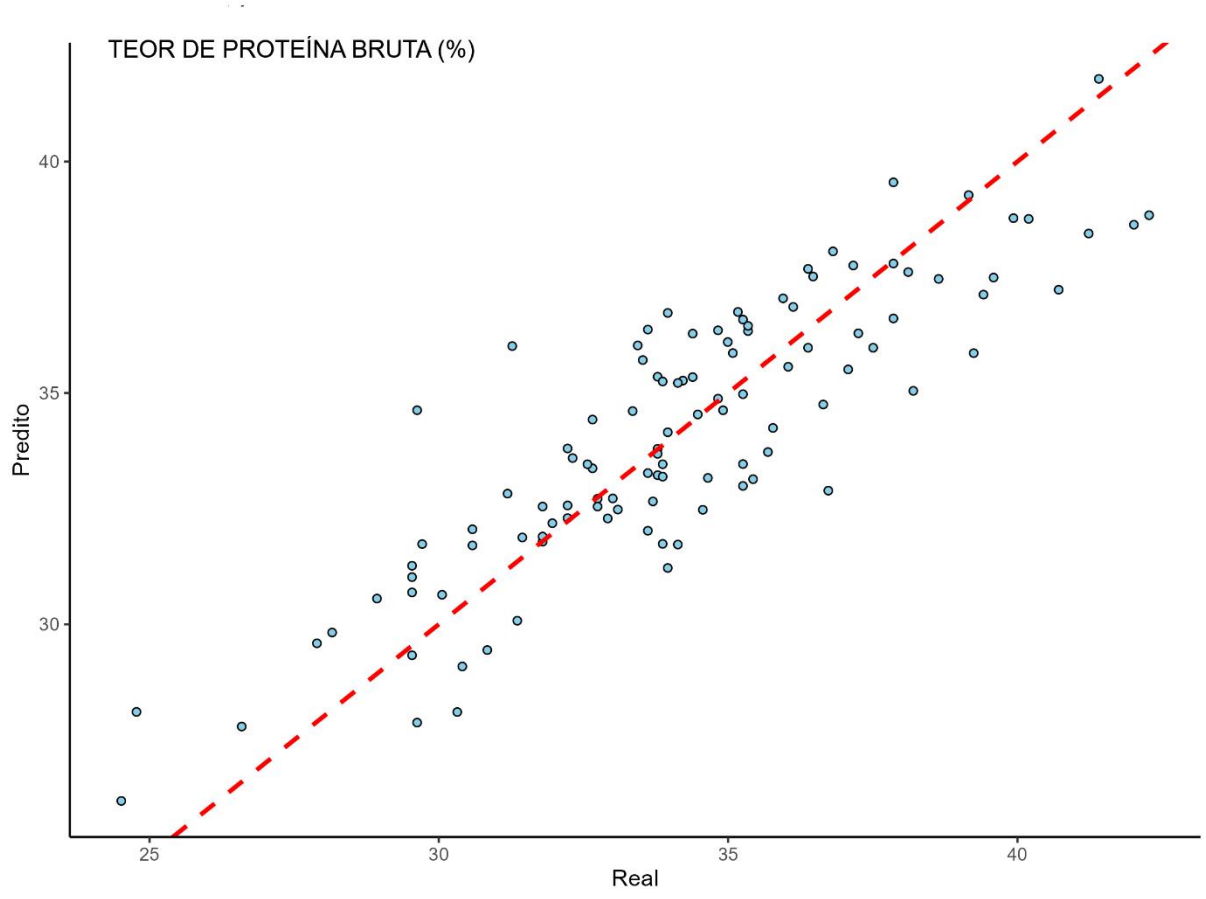
Nogueira (2023) avaliou o efeito da granulometria na análise de materiais de origem vegetal por espectroscopia no infravermelho próximo (NIR) e constatou que a utilização de partículas de menor tamanho tende a proporcionar resultados mais precisos. De forma semelhante, Pazdernik et al. (1997) relataram valores de r^2 variando de 0,40 a 0,85 ao utilizarem grãos de soja moídos, reforçando a tendência de que amostras mais finas e homogêneas resultam em modelos preditivos mais robustos. Tibola et al. (2018), ao utilizarem a espectroscopia NIR para avaliar indicadores de qualidade tecnológica em soja, empregaram modelos de calibração multivariada do tipo PLS, obtendo r^2 de 0,90, o que corrobora os resultados obtidos neste estudo.

Dentre os modelos preditos, os PLS 5, 3 e 2 foram os que apresentaram baixos valores de RMSE, sendo 0,58, 0,66 e 0,79, respectivamente (Tabela 3). Valores semelhantes foram obtidos por Tibola et al. (2018), que analisou espectros de 180 amostras de grãos de soja moída, e desenvolveram um modelo de calibração para predição do teor de proteína bruta por meio da PLS, obtendo valor de RMSE de 0,773, indicando boa precisão na estimativa dos teores de proteína. Os resultados obtidos no presente estudo evidenciam que o uso de amostras de grãos moídos contribuiu para reduzir a variabilidade física dos grãos e aumento da acurácia preditiva, como relatado também por Pazdernik et al. (1997) e Morgano et al. (2007).

Os maiores valores da RPD foram obtidos nos modelos PLS 5, 1 e 2, cujos valores foram 5,75, 4,49 e 4,41 (Tabela 3). Valores de RPD superiores a 3 indicam modelos com excelente capacidade preditiva (ZHANG et al., 2015). Para os grãos moídos, todos os modelos ajustados apresentaram RPD acima desse limite, mostrando-se adequados para análises quantitativas. Os modelos PLS 3 e PLS 5 também são considerados eficientes para predição da característica teor de proteína em grãos de soja conforme critério estabelecido por Chang et al. (2001), que sugeriram a classificação dos melhores modelos baseada na razão RPD > 2 e no $r^2 > 0,80$.

O modelo PLS 5 apresentou a melhor capacidade preditiva entre os modelos ajustados para grãos moídos (Figura 4). Para visualizar seu desempenho, foi construído um gráfico de dispersão entre os valores observados, obtidos por método de referência, e os valores preditos pelo modelo, com base nos espectros NIR gerados a partir de grãos de soja moídos. A reta tracejada representa a linha ideal de predição (1:1). A avaliação foi realizada utilizando os 110 cultivares incluídos na calibração, caracterizando uma análise interna do ajuste do modelo.

Figura 4 - Dispersão dos valores observados versus preditos para o teor de proteína bruta (%) em grãos de soja moídos, utilizando o modelo PLS 5 ajustado com espectros obtidos por espectroscopia no infravermelho próximo (NIR)



Os teores de proteína preditos utilizando o modelo com melhor ajuste, PLS 5, apresentou variações em relação aos valores reais determinados via método Kjeldahl, entre 0,0085% e 4,05%. Observa-se, alta concordância entre os teores determinados por extração química de nitrogênio e aqueles preditos pelo modelo PLS 5, a partir dos espectros obtidos por espectroscopia no infravermelho próximo (NIR). É importante observar que a diferença entre os valores preditos de proteína, em comparação com os obtidos por métodos tradicionais, foi menor nas amostras de grãos moídos do que nas de grãos inteiros, indicando maior uniformidade e precisão na predição para as amostras moídas.

3.3 Considerações finais

A diferença do desempenho dos modelos preditivos para grãos inteiros e moídos está relacionada principalmente a maior homogeneidade e uniformidade dos dados obtidos pelos grãos moídos (NOGUEIRA, 2023). Enquanto os grãos inteiros apresentam variabilidade física, como diferenças no tamanho e forma, que afetam a penetração e absorção da radiação NIR, as amostras moídas oferecem uma matriz mais uniforme, reduzindo a interferência de fatores físicos e melhorando a correlação entre os dados espectrais e os valores de referência, resultando em modelos mais precisos, estáveis e com menor RMSE (FERREIRA, 2013).

A calibração de curvas para grãos inteiros se mostra desafiante por muitos motivos, principalmente devido a heterogeneidade das amostras: cor, tamanho, superfície de contato, forma, rugosidade, etc. Superfícies irregulares e a estrutura completa do grão pode causar mais espalhamento da luz quando comparado a análises realizadas com as amostras devidamente tratadas antes da coleta de seus espectros (NOGUEIRA, 2023). Porém, analises com grãos inteiros também apresentam vantagens como rapidez e natureza não destrutiva, o que é crucial para algumas aplicações, como seleção de sementes para melhoramento genético ou triagem rápida em linha de produção (HOLLUNG et al., 2005). Embora o uso de grãos moídos eleve a precisão dos modelos, essa abordagem demanda um processo destrutivo e etapas adicionais de preparo, o que pode limitar sua aplicação.

Dentre os pré-processamentos avaliados, a centralização na média e o auto escalamento foram os que apresentaram melhor desempenho preditivo, tanto para grãos inteiros quanto moídos. No caso específico dos grãos moídos, a adição da primeira derivada como pré-tratamento dos espectros resultou em ganho adicional de desempenho, evidenciando que a combinação entre homogeneização física da amostra e aprimoramento matemático do sinal espectral contribui significativamente para a robustez dos modelos preditivos. Balabin e Smirnov (2011) e Xu et al. (2008) destacam que a aplicação de pré-processamentos espectrais é fundamental para a construção de modelos preditivos robustos, sendo necessário testar diferentes abordagens e combinações até se identificar aquelas que proporcionem os melhores resultados.

Ao comparar o desempenho dos modelos PLS obtidos a partir dos espectros NIR de grãos inteiros e moídos, observa-se uma superioridade dos modelos gerados com grãos moídos. Considerando a classificação estabelecida por Chang et al. (2001), os modelos que apresentam razão RPD > 2 e r^2 > 0,80 são enquadrados na categoria A, sendo considerados satisfatórios para fins preditivos, na categoria B estariam os modelos com capacidade limitada, com 1,4 < RPD < 2 e 0,50 < r^2 < 0,80 e na categoria C estão os modelos com RPD < 1,4 e r^2 < 0,50, classificado como inadequados para predição. Nos modelos com grãos moídos (Tabela 2), dois modelos se enquadram na categoria A o PLS 3 e PLS 5, apresentando RPD > 2 e r^2 > 0,80, sendo considerados confiáveis para fins preditivos. O modelo PLS 5, destacou-se com o melhor desempenho, apresentando r^2 igual 0,87 e RPD de 5,75, o maior valor dentre todos os modelos avaliados. Os modelos ajustados com grãos inteiros (Tabela 1) apresentaram desempenho ligeiramente inferior em comparação aos modelos obtidos com grãos moídos. O modelo PLS 3 alcançou a categoria A, os modelos PLS 1, 2, 4, 5 categoria B com desempenho considerado limitado, mas útil para seleção preliminar em programas de pesquisa e o PLS 6 categoria C, por apresentarem r^2 < 0,50 e RPD < 1,4, o que os torna inadequados para predição.

Os modelos desenvolvidos apresentaram desempenho moderado a alto, demonstrando potencial para aplicação em rotinas laboratoriais voltadas à pesquisa e à valoração da cadeia produtiva da soja.

Comparando o melhor modelo para grãos inteiros e moídos, observou-se uma redução de 3,3% no RMSE, um aumento de 0,32 unidades no RPD e um incremento no r^2 que passou de 0,82 para 0,87, quando se utilizou materiais moídos. Assim, apesar da necessidade de processamento adicional, o uso de grãos moídos se mostra uma estratégia vantajosa quando o objetivo é maximizar a acurácia das predições do teor de proteína via espectroscopia NIR.

A eficácia da espectroscopia NIR combinada a modelos PLS na predição do teor de proteína tem sido demonstrada em diversos estudos. Shi et al. (2022), obtiveram r^2 de 0,90 na estimativa do teor de proteína bruta em soja, utilizando os pré-processamentos de correção multiplicativa de sinal (SNV) e remoção de tendências (DET). De forma semelhante, Chadalavada et al. (2022) desenvolveram modelos robustos aplicáveis a múltiplas espécies de grãos, com desempenho preditivo dado por r^2 de 0,90, RMSE $\leq$ 0,91 e RPD $\geq$ 3,08. Os resultados obtidos neste estudo, apresentam desempenho compatível com os critérios de qualidade da literatura, reforçando a aplicabilidade dos modelos desenvolvidos para fins preditivos em programas de melhoramento ou controle de qualidade da soja.

Dessa forma, observa-se que os modelos preditivos ajustados com grãos inteiros apresentam desempenho satisfatório, preservando o caráter não destrutivo da espectroscopia NIR, um dos principais atributos dessa técnica. No entanto, essa abordagem pode resultar em modelos menos robustos quando comparados àqueles obtidos a partir de amostras previamente moídas. O uso de grãos moídos, portanto, constitui uma alternativa eficaz para situações em que se busca maior precisão e estabilidade nas predições.

4 CONCLUSÕES

A forma como a amostra é avaliada, seja com grãos inteiros ou moídos, influência a precisão dos modelos de regressão.

A utilização da espectroscopia no infravermelho próximo (NIR), aliada a métodos de calibração multivariada, permite desenvolver modelos preditivos eficientes para determinar o teor de proteína bruta em soja.

Os modelos de predição do teor de proteína bruta em soja, desenvolvidos a partir de espectros de grãos moídos, apresentaram desempenho superior em relação àqueles obtidos com grãos inteiros.

5 REFERÊNCIAS BIBLIOGRÁFICAS

ARMSTRONG, Paul R. et al. Development of single-seed near-infrared spectroscopic predictions of corn and soybean constituents using bulk reference values and mean spectra. **Transactions of the ASABE,** v. 54, n. 4, p. 1529-1535, 2011.

ASSOCIATION OF OFFICIAL ANALYTICAL CHEMISTS (Arlington, Estados Unidos). **Official methods of analysis.** Washington, DC,1094p, 1975.

AULIA, R. et al. Non-destructive prediction of protein contents of soybean seeds using near-infrared hyperspectral imaging. **Infrared physics & technology**, v. 127, p. 104365, 2022.

BALABIN, R. M.; SMIRNOV, S. V. Melamine detection by mid-and near-infrared (MIR/NIR) spectroscopy: A quick and sensitive method for dairy products analysis including liquid milk, infant formula, and milk powder. Talanta, v. 85, n. 1, p. 562-568, 2011.

BARCELLOS, D. C. **Caracterização do carvão vegetal através do uso de espectroscopia no infravermelho próximo**. 2007.

BHEEMANAHALLI, R. et al. Fenotipagem de cultivares de soja do sul dos Estados Unidos para composições de peso potencial de sementes e qualidade de sementes. **Agronomy.** v.12, 839, 2022.

CARRÃO-PANIZZI, M. C.; SILVA, J. B. da. Soja na alimentação humana: qualidade na produção de grãos com valor agregado. In: **Congreso de la soja del mercosur-mercosoja**, p. 1-3. 2011.

CHANDALAVADA, K. et al. NIR instruments and prediction methods for rapid access to grain protein content in multiple cereals. Sensors, Basel, v. 22, n.10, p.3710, 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

GALVANI, F. GAERTNER, E. Adequação da Metodologia Kjeldahl para Determinação de Nitrogênio total e Proteína Bruta. 2006. **Embrapa Pantanal.** Circular Técnica, v. 63, 2024.

FEDERER, W.T. **Experimental design:** Theory and application. New York: MacMillan, 1955. 544p.

FERREIRA, D. S. **Aplicação de espectroscopia no infravermelho e análise multivariada para previsão de parâmetros de qualidade em soja e quinoa = Application of infrared spectroscopy and multivariate analysis to predict quality parameters in soybean and quinoa.** Tese de Doutorado. 2013.

FERREIRA, D. S. et al. 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. et al. Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, v. 22, p. 724-731, 1999.

HERTSGAARD, D.J. et al.. Custos e riscos de testes e misturas para aminoácidos essenciais em soja. **Agribusiness** 35, 265–280, 2019.

HOLLUNG, K., et al. **Journal of agricultural and food chemistry**, v. 53, p. 9112–21, 2005.

Johnson, J.M. et al. Near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice fields in sub-Saharan Africa. **Geoderma,** 354, 113840, 2019.

LAMMERTYN, J. et al. Non-destructive measurement of acidity, soluble solids, and firmness of Jonagold apples using NIR-spectroscopy. **Transactions of the ASAE**, v. 41, n. 4, p. 1089-1094, 1998.

LARIOS, G. et al. Soybean seed vigor discrimination by using infrared spectroscopy and ma chine learning algorithms. **Analytical Methods**, v. 12, n. 35, p. 4303–4309.

LIU, K. Selected topics in the analysis of lipids: modification of an AOCS official method for crude oil content in distillers grain. **Urbana: The AOCS Lipid Library**. (2011).

MARCHESE, Natalia Regina et al. Espectroscopia de infravermelho próximo e metodologia de mínimos quadrados parciais para análise de soja (Glycine Max.(L) Merril) inativada termicamente. **Dissertação de Mestrado**. Universidade Tecnológica Federal do Paraná. 2017.

MARTENS, H.; NAES, T.; **Multivariate calibration**, J. Wiley & Sons Ltd.: Chichester, 1989.

MATSIMBE, S.F.S. et al. Prediction of oil content in the mesocarp of fruit from the macauba palm usin spectrometry. **Revista Ciência Agronômica,** 46(1), 21-28, 2015.

MENDHAM, J. DENNEY, R. C.; BARNES, J. D.; THOMAS, M. J. K. Vogel : Análise Química Quantitativa. In:____. **Estatística: Introdução à Quimiometria**. 6ª Ed. Rio de Janeiro. LTC. Cáp. 4. 2002.

MORAES, R. M. A. de et al. Caracterização bioquímica de linhagens de soja com alto teor de proteína. **Pesquisa Agropecuária Brasileira**, v. 41, p. 715-729, 2006.

MORGANO, M. A. et al. Determinação de açúcar total em café cru por espectroscopia no infravermelho próximo e regressão por mínimos quadrados parciais. **Química Nova**, v. 30, p. 346-350, 2007.

NOGUEIRA, A. R. A.; SOUZA, G. B. **Manual de Laboratórios: Solo, Água, Nutrição Vegetal, Nutrição Animal e Alimentos.** São Carlos: Embrapa Pecuária Sudeste, p. 313, 2005.

NOGUEIRA, T. A. P. C. **Efeito da granulometria de amostras de serragem de Eucalyptus na análise de densidade e pentosanas via NIR para as indústrias de celulose.** 2023.

PANERO, J. S. et al. **Determinação de proteína, óleo e umidade por espectroscopia NIR em grãos de soja do estado de Roraima.** 2007.

PAZDERNIK, D. L.; KILLAM, A. S.; ORF, J. H. Analysis of amino and fatty acid composition in soybean seed, using near infrared reflectance spectroscopy. **Agronomy Journal**, v. 89, n. 4, p. 679-685, 1997.

OLIVEIRA, M. A. de et al. Teores de óleo e proteína em grãos de soja, com diferentes manejos de percevejo, da colheita ao armazenamento, utilizando a espectroscopia no infravermelho próximo (NIR). **Brasília: Embrapa**, 2013.

R CORE TEAM**. R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2020. URL https://www.R-project.org/

RANGEL, M. A. S. et al. **Efeito do genótipo e do ambiente sobre os teores de óleo e proteína nos grãos de soja, em quatro ambientes da região sul de Mato Grosso do Sul, safra 2002/2003**. 2004.

ROCHA, J.T.C. **Emprego de espectrometria no infravermelho e métodos quimiométricos para a identificação e quantificação de petróleos a partir de misturas de frações de diesel.** 122p. Tese (Mestrado) - Universidade Federal do Espírito Santo, Vitória, 2009.

ROESSING,A.C.;GUEDES,L.C.A. **Cultura da Soja no Cerrado.** Piracicaba: Associação Brasileira para a Pesquisa da Potassa e do Fosfato, 1993.

SÁIZ-ABAJO, M.J.et al. Ensemble methods and data argumentation by noise addition applied to the analysis of spectroscopic data. **Analytica Chimica Acta**, 553, 147-159, 2006.

SCHIMLECK, L. R.; EVANS, R. Estimation of Pinus radiata D. Don tracheid morphological characteristics by near infrared spectroscopy. **Holzforschung**, v. 58, n. 1, p.66-73, 2004.

SHI, D. et al. 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**, Orlando, v. 111, p. 104601, 2022.

SILVA, F. L. DA et al. **Soja: do plantio à colheita**. 2. ed. São Paulo, SP: **Oficina de Textos**, 2022. 312p.

SILVA, M. F. da et al. Armazenabilidade, envelhecimento e classificação do vigor de sementes de soja por espectroscopia do infravermelho próximo. 2020.

SILVA, P. R. V. **Histórico da Espectroscopia do Infravermelho Próximo (NIR - Near Infrared)**. 2006. Acessado em 28/04/2025: http://www.angelfire.com/ab/prvs/ . 2006.

SIMAS, R. C. **Determinação de proteína bruta e aminoácidos em farelo de soja por espectroscopia no infravermelho próximo.** 2005. Dissertação (Mestrado), Universidade Estadual de Campinas - UNICAMP. Campinas, São Paulo.

SOARES, J. M. Espectroscopia no infravermelho próximo e métodos quimiométricos para classificação de sementes de soja quanto ao potencial fisiológico. 2023.

TIBOLA, C.S. et al. Espectroscopia no infravermelho próximo para avaliar indicadores de qualidade tecnológica e contaminantes em grãos. **Brasília: Embrapa**, 2018.

VIANA, Valdomiro Teixeira. **Comparativo entre os métodos nitrogênio de Kjeldahl e NIRS para análise de proteína bruta em farelo de soja**. 2022.

WILLIAMS, P. C., NORRIS, K. H., SOBERING, D. C. Determination of protein and moisture in wheat and barley by near-infrared transmission. **Journal of Agricultural ans Doos Chemistry.** v. 33, p. 239-244. 1985

WILLIAMS, P. C.; SOBERING, D. C. Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. **Journal of Near Infrared Spectroscopy**, v. 1, n. 1, p. 25-33, 1993.

WILLIAMS, P. C., THOMPSON, B. N. Influence of whole meal granularity on analysis of HSR wheat for protein and moisture by near infrared reflectance spectroscopy. **Cereal chemistry,** v. 55, p.1014-1037.1978.

WILLIAM, W., Dahl, B., Hertsgaard, D.J. **Diferenciais de qualidade da soja, mistura, testes e arbitragem espacial.** J. Commod. Mark. v. 18, 100095, 2019.

XU, L. et al. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. **Analytica chimica acta**, v. 616, n. 2, p. 138-143, 2008.

ZARKADAS, C.G. et al. Assessment of the protein quality of fourteen soybean [Glycine max (L.) Merr.] cultivars using amino acid analysis and two-dimensional electrophoresis. **Food Research International**, 40, 129–146, 2007.

ZHANG, Chu et al. Application of visible and near-infrared hyperspectral imaging to determine soluble protein content in oilseed rape leaves. **Sensors,** v. 15, n. 7, p. 16576-16588, 2015.