Ciência Rural, Santa Maria, v.53:6, e20220071, 2023
http://doi.org/10.1590/0103-8478cr20220071
ISSNe 1678-4596
CROP PRODUCTION

Path analysis and near-infrared spectroscopy in canola crop

Alexsandro Carvalho Santiago[1*] Guilherme Vieira Pimentel[1] Adriano Teodoro Bruzi[1] Inara Alves Martins[1] Paulo Ricardo Gherardi Hein[2] Michael Douglas Roque Lima[2] Dyanna Rangel Pereira[3]

[1]Departamento de Agricultura, Universidade Federal de Lavras (UFLA), 37200-900, Lavras, MG, Brasil. E-mail: alexsandrocarvalho14@gmail.com. [*]Corresponding author.
[2]Departamento de Ciências Florestais, Universidade Federal de Lavras (UFLA), Lavras, MG, Brasil.
[3]Departamento de Biologia, Universidade Federal de Lavras (UFLA), Lavras, MG, Brasil.

ABSTRACT: This study measured the effect of the association between agronomic traits related to the yield of canola grains grown at different sowing dates through path analysis. Another objective was to obtain a method to predict the oil content in the grains, fitting a multivariate model through near-infrared (NIR) spectroscopy analysis. The experiment was conducted in the field using a randomized block design in plots subdivided by time, with four plots (sowing dates), six subplots (canola hybrids), and four replicates. In each hybrid, phenological observations were performed, and the grain yield was determined. The data were subjected to analysis of variance in the R environment using the F test at 5% probability. The oil content in the grains was determined by the traditional chemical method, and based on the NIR spectral signature of the grain samples, partial least squares regression (PLS-R) was established to estimate the oil content in the canola grains. The sowing dates influenced the production components and oil content of the grains of all hybrids. The trait number of grains in five plants (0.6857) and their height (0.4943) had greater estimates of positive correlations with grain yield, as well as higher values of positive direct effects on yield (0.2494 and 0.1595, respectively). The NIR technique combined with PLS-R was able to predict the oil content in the grains, resulting in good predictive models (R^2 of 0.86 and root mean square error (RMSE) of 1.56 in external validation).
Key words: *Brassica napus* L. *var. oleifera*, oleaginous, association between traits, genetic breeding, NIR.

Análise de trilha e espectroscopia no infravermelho próximo na cultura da canola

RESUMO: Objetivou-se mensurar o efeito da associação entre caracteres agronômicos relacionados à produtividade de grãos de canola cultivada em diferentes épocas de semeadura, através da análise de trilha. Assim como também objetivou-se obter um método para predizer o teor de óleo nos grãos, ajustando um modelo multivariado através da análise por espectroscopia na região do infravermelho próximo. O experimento foi conduzido em campo, utilizando-se o delineamento de blocos ao acaso, em parcelas subdivididas no tempo, sendo quatro parcelas (épocas de semeadura) e seis subparcelas (híbridos de canola), com quatro repetições. Em cada híbrido foram realizadas observações fenológicas e determinada a produtividade de grãos. Os dados foram submetidos à análise de variância em ambiente R pelo teste F, a cinco de probabilidade. O teor de óleo nos grãos foi determinado pelo método químico tradicional, e com base na assinatura espectral no infravermelho próximo de amostras dos grãos foi estabelecida regressão dos mínimos quadrados parciais (PLS-R) para estimar o teor de óleo nos grãos de canola. As épocas de semeadura influenciaram os componentes de produção e o teor de óleo dos grãos de todos híbridos. Os caracteres número de grãos em cinco plantas (0,6857) e altura (0,4943) apresentaram maiores estimativas de correlação positiva com a produtividade de grãos, assim como os maiores valores de efeito direto positivo sobre a produtividade, 0,2494 e 0,1595 respectivamente. Entretanto, o ciclo total (-0,7848), juntamente com dias em florescimento (-0,4520) apresentou correlação significativa negativa com a produtividade. A técnica NIR associada à PLS-R foi capaz de predizer o teor de óleo nos grãos, resultando em bons modelos preditivos (R^2 de 0,86 e RMSE de 1,56 na validação externa) que podem ser usados com sucesso na análise da qualidade das amostras após colheita e nos programas de melhoramento genético.
Palavras-chave: *Brassica napus* L. *var. oleifera*, oleaginosas, associação entre caracteres, melhoramento genético, NIR.

INTRODUCTION

Canola is considered the third most important oilseed in the world, accounting for approximately 14% of the world production of edible oils (USDA, 2021). In Brazil, canola cultivation stands out owing to its high production potential and plant traits, and it is of great interest for crop expansion due to its drought tolerance and the possibility of using it in rotation with soybean, corn, wheat, and beans (TOMM, 2007). However, investments in research for the development of management technologies in

Received 02.12.22 Approved 07.03.22 Returned by the author 08.30.22
CR-2022-0071.R1
Editors: Leandro Souza da Silva Marcos Toebe

the cultivation of canola have been scarce. Thus, it is necessary to improve the technical recommendations, as well as investments in genetic breeding programs, in order to provide significant gains in yield and consequent expansion of the cultivated area.

Thus, within breeding programs, the breeding of the main trait is sought while simultaneously maintaining or improving the expression of other traits. Thus, it is important to know the relationship between traits (LOPES et al., 2002) and study these correlations jointly with path analysis (WRIGHT, 1921), enabling the identification of traits that can be used as indirect selection criteria for yield (CARVALHO et al., 2002), which is a complex trait of low heritability.

In the case of postharvest evaluations in these plant breeding programs, it is necessary to seek alternative methods to the conventional process used to quantify oil content because it is a time-consuming method that uses large amounts of chemical products (CHENG et al., 2017). Near-infrared (NIR) spectroscopy has emerged as an ideal tool for tracking traits (FONT et al., 2006). The NIR spectrum provides information about the chemical composition of the samples, such as the oil content (PASQUINI, 2003). All this information, obtained in a short period of time, allows for immediate processing of the grain samples after harvest and thus results in selection of quality materials (PETISCO et al., 2010), making the tool very advantageous in accelerating evaluations. However, to quantify such properties through NIR spectroscopy, it is necessary to develop multivariate models (PASQUINI, 2003).

In this context, studies of canola culture are crucial to expanding its cultivation. Given the above, this study seeks to provide a foundation for future genetic breeding programs of the crop in the search for more productive genotypes under Cerrado conditions, enabling the expansion of canola cultivation in the country. Furthermore, the study aims to accelerate evaluations of the oil content of different genotypes in breeding programs and obtain an alternative tool to the conventional chemical process in order to quantify the oil content in the grains rapidly and nondestructively.

Thus, this study measured the effect of the association between agronomic traits related to the yield of canola grains grown at different sowing dates with regard to the direct and indirect effects obtained by path analysis. In addition, a multivariate model fit that was able to predict the oil content in grains through NIR spectroscopy analysis.

MATERIALS AND METHODS

Site of experimentation and hybrids

The canola crop was grown in a no-tillage system in a clayey-red–yellow latosol in a region where the climate classification according to Köppen is the Cwa type (rainy temperate) with dry winters and rainy summers, with an average annual temperature of 20.4 °C and average annual rainfall of 1433.3 mm.

Six commercial canola hybrids (ALHT B4, Diamond, Hyola 433, Hyola 571 CL, Hyola 575 CL, and Nuola 300) were used, which were chosen according to the availability/dominance of such genotypes in the market. Prior to sowing, a germination test was performed to evaluate germination and vigor.

Conducting the experiment and experimental design

A randomized block experimental design was used in subdivided plots, with four plots (sowing dates: February 15, February 28, March 20, and April 9) and six subplots (canola hybrids), presenting four replicates. The experimental area was created in the no-tillage system, with subplots of five rows (spacing of 0.20 m) 5 m in length, totaling 5 m^2, and an evenly distributed population of 40 plants/m^2.

Traits evaluated

Based on criteria adopted in Canada and Australia (CANOLA COUNCIL OF CANADA, 2020), phenological observations were performed, and the following variables were evaluated: a) onset of flowering: date when 50% of the plants had at least one flower; b) end of flowering: date when no flowers remained, except in atypical plants; c) plant height: average height of five representative plants of each plot, measured at harvest, considered from the base to the upper end of the branches with siliques; and d) maturation date: date when 50% of the seeds changed to a dark color in the siliques, located in the middle of the main raceme of the plants.

Thus, the days in flowering (DIF) and total cycle (TC) of these plants were evaluated. The number of plants per area (stand) was evaluated at harvest by counting all plants in the useful area. In addition, grain yield was evaluated based on the harvest of three central rows of plants 4 m in length, excluding 0.5 m from the edges of the border, totaling 2.4 m^2 of usable area. The harvest was performed manually, and the plants were kept in bags to dry in the air until reaching approximately 10% moisture. In five representative plants of each plot, the number of siliques (NS), number of grains, (NG), and number of grains per silique (NGS) were

counted. Next, the grain samples were cleaned with the aid of a set of sieves, and the weight of 100 grains (P100, g) was measured.

For the analysis of the oil content in the grains, the uniform seed samples were dried in an oven with forced air ventilation at 65 °C for 48 h to standardize the moisture. After drying, the seeds were milled with the husks, and the oil content was determined. The bran of the seeds was packaged in paper cartridges at 2 g per cartridge in duplicate per experimental unit. For the extraction, the methodology described in IUPAC (1979) was adopted using the Soxhlet system.

Near-infrared (NIR) spectroscopy

The biological material used in the study consisted of samples of grains from the different sowing dates cultivated in the given experiment, and to give greater robustness to the adjusted model, grains from the hybrids Hyola 50, Hyola 61, Hyola 76, Hyola 433, ALHT B4, Hyola 571, Hyola 575, and Diamond, from the 2018 harvest of Embrapa Agroenergy, were also included, totaling eight different canola genotypes.

Spectral acquisition was performed in a Bruker MPA spectrometer, together with the computer program Opus version 7.5. The spectra were obtained from diffuse reflection mode in the range of 12,500 to 3,600 cm^{-1}, with a spectral resolution of 8 cm^{-1}. The spectrum of each sample was obtained by the mean of 16 scans, performed in three different portions of each grain sample.

Statistical analysis

The yield and phenology trait data were subjected to analysis of variance in the R environment (R CORE TEAM 2017) using the F test at 5% probability, as reported by STEEL et al. (1997) when using subdivided plots.

Multivariate statistics for spectral data

Regarding the spectral data collected from the canola grains, multivariate analyses were performed using Unscrambler® (version 9.7) software. Principal component analysis (PCA) was initially adopted to explore the data and evaluate the spectral similarity between the genetic materials studied.

To estimate the canola oil content using spectral data, multivariate models were fitted using the partial least squares regression (PLS-R) method, relating the NIR spectral data obtained and the oil content previously determined by the traditional chemical method. The number of latent variables adopted was chosen based on the lowest standard error of validation and the highest coefficient of determination of the validation (R^2cv).

The models were validated by cross-validation and independent (external) validation methods. Cross-validation was performed by the random method, considering 96 data points from 32 grain samples. In this type of validation, one calibration data point is removed at a time, the model is constructed, the retained sample is estimated, and the process is repeated for all other data. The independent validation was based on two datasets, using 24 samples (72 data points) for the calibration lot and 8 samples (24 data points) for the validation lot.

The calibrations were performed using original spectra and spectra treated mathematically with the first derivative, aiming to improve the signal/noise ratio. To calibrate the models to estimate the oil content of the samples, the selection of spectral ranges was determined by the Martens uncertainty test (WESTAD & MARTENS, 2000). Thus, the spectral range adopted in this study was from 3600 cm^{-1} to 9000 cm^{-1}. The anomalous samples (*outliers*) were detected using the *student* x *leverage* residual plot and were removed from the models.

The models were evaluated by the coefficient of determination (R^2cv and R^2p), the root of the mean standard error (RMSECV and RMSEP), and by graphical representation.

Path analysis

Path analysis was performed using Genes software (CRUZ, 2013). The correlation matrices between the involved traits were estimated, and their significance was evaluated by the Mantel test. Multicollinearity was tested based on the condition number of the matrix (MONTGOMERY & PECK, 1981). This method considers the number of conditions (NC) obtained by the ratio of the highest eigenvalue to the lowest eigenvalue of the correlation matrix.

The analysis was performed from a chain causal diagram, with the objective of unfolding the correlations into direct and indirect effects of explanatory variables on the main variable: grain yield. The decomposition of the correlation between the explanatory variables and the basic variable was given by CRUZ (2006). The coefficient of determination of the model and the effect of the residual variable on the main variable was also obtained in this analysis.

RESULTS AND DISCUSSION

Path analysis

Based on the test suggested by MONTGOMERY & PECK (1981), moderate multicollinearity (100 < NC < 1000) occurred. With this issue, one way to minimize the problem is the removal of redundant traits. When analyzing the correlations between the traits, a high correlation was observed between the number of grains in five plants and the number of siliques (0.96). Therefore, it was decided to remove the number of siliques as a trait. Subsequently, when performing the test again, weak multilinearity was observed. Accordingly, the estimates of the coefficients of simple or phenotypic correlations evaluated for the seven traits of agronomic importance for canola culture are presented in table 1.

In general, the values of the correlations with YIELD ranged from 0.0339 to 0.6857, with the traits NG (0.6857) and HT (0.4943) showing higher estimates of positive and significant correlation (Table 1). Such positive values were higher than those estimated for the other traits evaluated in this study, suggesting that these traits contributed to increased canola yield.

Conversely, the traits TC (-0.7848) and DIF (-0.4520) were negatively associated with YIELD and most of the other traits, indicating that a reduction in the cycle and the flowering phase in canola could result in higher yields. Different results obtained by COIMBRA et al. (2004) showed a positive association between the total cycle and grain yield. KRÜGER et al. (2014) observed that the variable duration of flowering had a negative relationship with grain yield in the three spacings studied, as noted in the present study.

Canola genotypes with short cycles have become an excellent economic alternative since the crop can benefit from the final rainy periods, reducing the need for irrigation and favoring the crop's incorporation into the production system. However, despite the higher demand for early hybrids because they enable better management of time and resources (ROSA et al., 2020), according to TOMM et al. (2010), medium- and late- cycle genotypes tend to have higher yield potential, precisely because they have a longer period to take greater advantage of environmental resources and to perform more photosynthesis. They have more time to compensate for conditions that might limit their production.

However, a high correlation does not imply a cause and effect relationship between the variables analyzed (VENCOVSKY & BARRIGA, 1992). As a result, path analysis was performed to study the unfolding of these correlation coefficients in direct and indirect effects of traits on a basic variable (CRUZ & CARNEIRO, 2006).

The estimates of the direct and indirect effects of the explanatory variables on the YIELD trait are shown in table 2. The sum of the direct and indirect effects yields the correlation coefficient. The sum of the direct effects multiplied by their respective correlations results in the coefficient of determination (R^2), and the root of the difference ($1-R^2$) results in the residual variable effect (P_E), which equals 0.6974 and 0.5501, respectively, indicating that the explanatory variables partially determined the variation in the basic variable (YIELD). However, it is worth noting; that although, the coefficient of determination was not considered low, the residual effect was somewhat high.

Table 1 - Correlation coefficients among seven agronomic traits evaluated in canola hybrids in Lavras - MG in 2019. Lavras - MG, 2021.

Trait	NG	NGS	HT	TC	DIF	ST	YIELD
NG	1	0.4426*	0.3001	-0.7614*	-0.2949	0.1644	0.6857*
NGS		1	-0.0893	-0.3404	-0.2132	0.0339	0.1361
HT			1	-0.4033*	-0.3356	0.2014	0.4943*
TC				1	0.6074*	-0.5476*	-0.7848*
DIF					1	-0.5936*	-0.4520*
ST						1	0.3346
YIELD							1

*Significant at the 5% level by the Mantel test

YIELD: yield, kg ha^{-1}; NG: number of grains in five plants; NGS: number of grains per silique; HT: height, m; TC: total cycle, days; DIF: days in flowering, days; ST: stand, number of plants/ha.

Table 2 - Direct and indirect effects of the explanatory variables on grain yield for the evaluation of canola hybrids in Lavras/MG in the 2019 agricultural year. Lavras - MG, 2021.

----------Effect on YIELD----------							
Variable	via NG	via NGS	via HT	via TC	via DIF	via ST	TOTAL
NG	0.2494	-0.0780	0.0479	0.4688	0.0136	-0.0160	0.6857
NGS	0.1104	-0.1762	-0.0142	0.2096	0.0098	-0.0033	0.1361
HT	0.0748	0.0157	0.1595	0.2483	0.0155	-0.0196	0.4943
TC	-0.1899	0.0600	-0.0643	-0.6157	-0.0280	0.0532	-0.7848
DIF	-0.0735	0.0376	-0.0535	-0.3740	-0.0461	0.0576	-0.4520
ST	0.0410	-0.0060	0.0321	0.3372	0.0274	-0.0971	0.3346
----------COEFFICIENT OF DETERMINATION (R^2)----------							0.6974
----------EFFECT OF RESIDUAL VARIABLE ((P_E)----------							0.5501

YIELD: yield, kg ha^{-1}; NG: number of grains in five plants; NGS: number of grains per silique; HT: height, m; TC: total cycle, days; DIF: days in flowering, days; ST: stand, number of plants/ha.

The traits that, in general, resulted in estimates of a positive direct effect on grain yield were NG (0.2494) and HT (0.1595). Although, of low magnitude, they showed significant positive correlation estimates. VENCOVSKY & BARRIGA (1992) stated that if the correlation coefficient is positive but the direct effect is nonsignificant or negative, this correlation is caused by indirect effects, so in the selection process, these effects should be considered simultaneously. Thus, in the selection practiced for these traits, they can directly contribute to the increase in grain yield. The indirect variable TC, which has the greatest indirect effect on NG and HT, should be considered simultaneously.

Studies performed by ROCHA et al. (2019) stated that when decomposing the direct and indirect effects of the production components regarding the grain yield of canola genotypes, there was a greater effect of the variables number of grains per silique and oil yield, with a positive direct influence on grain yield. COIMBRA et al. (2004) found that the plant population per unit area and number of grains per plant had the greatest direct effects on the grain yield variable and that the number of grains per silique had the greatest secondary effect.

In turn, the remaining traits (NGS, TC, DIF, and ST) had negative direct effect estimates regarding grain yield. For TC, in addition to having the highest estimate (-0.6157) and a high negative phenotypic correlation with YIELD, it still has a value greater than that observed in the residual effect, demonstrating that this trait can be used in indirect selection. According to VENCOVSKY & BARRIGA (1992), when the correlation between a causal trait and the main trait is equal or similar to its direct effect in sign and magnitude, the indirect selection of the causal trait will be efficient since this correlation expresses the true association between such traits.

It is noteworthy that the DIF variable, despite having a low direct negative effect, has a significant negative correlation with yield, suggesting that a reduction in the flowering period could contribute to increased yield in canola. However, flowering is the most critical phase that influences grain yield (DIEPENBROCK, 2000); thus, this trait inspires caution.

It is important to note that in the present study, under initial planting conditions at different sowing times, higher initial temperatures were observed, which might have led to a shortening of the cycle and a shorter duration of flowering without very significant losses in yield. The sowing date has a decisive effect on the duration of the vegetative phase and flowering (ROSA et al., 2020), showing a strong influence from the air temperature (LUZ et al., 2012). EDWARDS & HERTEL (2011) confirmed this information by observing that canola under temperatures in the range of 20 °C produced a new leaf every six to ten days. Conversely, at temperatures greater than 27 °C, there was a reduction in this time to four days.

Therefore, under the earlier planting conditions in the region with a high-altitude tropical climate, with the goal of avoiding water deficits and the coincidence of flowering with low temperatures, the strategy of selecting canola hybrids with a shorter

cycle can be used when seeking higher yields, and as a form of adaptation to productive windows. However, further studies are needed to verify the extent to which such a reduction in the total cycle is feasible so that there is no loss in grain yield and oil content.

When analyzing the negative indirect effects via the variables, such behavior was observed in NG via NGS and via ST; NGS via HT and ST; HT via ST; TC via NG, HT, and DIF; DIF via NG, HT, and TC; and ST via NGS. These observations reduced the direct effect of these traits, as well as the total effect of the other traits on yield. Thus, if an increase in these traits occurred, it could cause a reduction in other traits, even if they expressed a direct effect on grain yield.

However, it is noteworthy that the high effect of the residual variable (0.5501) indicated that the set of six variables does not fully explain the variation in yield since its value exceeds most estimates of direct and indirect effects. Thus, the observed variation in yield is also due to other traits not measured in the present study (CRUZ & CARNEIRO, 2006), such as silique length, silique mass, and number of branches, as well as effects associated with the experimental error arising from random variations.

Near-infrared (NIR) spectroscopy

Table 3 shows the oil content in the grains of canola hybrids, determined by traditional chemical methods in the laboratory by adopting the methodology described in IUPAC (1979) for extraction. These values were used as references to construct the NIR calibration model.

The oil content in the grains varied between hybrids and between sowing dates (Table 3). This behavior was expected because the oil concentration of canola seeds is influenced by the genotype (TOMM et al., 2009b) and by environmental factors (LONG et al., 2012), such as temperature and rainfall (TOMM et al., 2009a).

The mean spectra collected via the integration sphere consisted of the grains of hybrids of each production environment shown in table 3. Since the presence of considerable noise was observed in the range of 12,500 to 9000 cm^{-1}, making it difficult to obtain useful information for the analyses, this range was eliminated, and the range of 9000 to 4000 cm^{-1} was used to construct the model.

Based on the spectral signatures of the samples of canola grains, calibrations and validations (cross and external) were performed using PLS-R multivariate calibration. Table 4 shows the statistics of the models created by PLS-R and cross-validation.

In the present study, able 4 shows that the use of the 96 original spectral data with 5 latent variables (chosen by the program due to the lower residual variance) resulted in a less efficient model, obtaining an R^2c of 0.64 and an R^2cv of 0.45, with high RMSE values. Low values for the coefficient of determination were also reported by KAUR et al. (2017), who obtained an R^2c of 0.4147, an R^2cv of 0.3932, an RMSEC of 1.7105, and an RMSECV of 1.7394 in the development of equations to estimate

Table 3 - Oil content in the grains of different canola hybrids, determined by traditional chemical methods in the laboratory. Lavras - MG, 2021.

----------Oil content (%)----------					
Hybrid	----------Sowing dates (2019 harvest)----------				2018 harvest of Embrapa Agroenergy
	15/02	28/02	20/03	09/04	
Nuola	29.79	30.83	29.07	27.74	-
Hyola 433	30.29	31.75	28.97	24.16	36.09
Hyola 571	28.94	30.88	29.21	27.56	31.39
Hyola 575	29.22	31.95	28.07	25.34	34.60
ALHT B4	30.32	35.72	28.32	24.92	35.28
Diamond	22.98	24.62	27.93	24.75	35.48
Hyola 50	-	-	-	-	26.75
Hyola 61	-	-	-	-	29.84
Hyola 76	-	-	-	-	35.27
Mean	28.59	30.96	28.60	25.75	33.08

Table 4 - Calibration and cross-validation of the canola oil content by PLS-R based on NIR spectra.

Models	Database	Treatment	R^2c	RMSEC	R^2cv	RMSECV	LV
1	96	osd	0.64	2.127	0.45	2.655	5
2	93	osd	0.84	1.426	0.80	1.635	7
3	96	1d	0.79	1.618	0.53	2.449	7
4	93	1d	0.85	1.385	0.70	1.993	7

R^2c - coefficient of determination for calibration; RMSEC - root of the mean standard error for calibration; R^2cv - coefficient of determination for cross-validation; RMSECV - root of the mean standard error for cross-validation; LV- latent variable; osd - original spectral data; 1d - first derivative.

the oil content in seeds of *Brassica napus*. However, the authors observed better results when working with *Brassica juncea*, which had an R^2c of 0.8335, an R^2cv of 0.7410, an RMSEC of 0.9226, and an RMSECV of 1.1560.

According to FERREIRA et al. (1999), when verifying the quality of the calibration set, it is necessary to ensure that the samples form a homogeneous set and to remove data considered outliers. Thus, from the *student* x *leverage* residuals graph, three anomalous samples (outliers) that could reduce the quality of the model were identified and removed. After removing these three outliers (Model 2), cross-validation with 93 original spectral data resulted in the best model, increasing the R^2c to 0.84 and R^2cv to 0.80 and reducing RMSEC to 1.426 and RMSECV to 1.635 (Table 4). However, for this model, 7 latent variables were used, resulting in the lowest residual variance (Figure 1A). For screening purposes, good models should have $R^2c \geq 0.85$, $R^2cv \geq 0.80$, RMSEC ≤ 5, and RMSECV ≤ 10, where RMSECV is the most significant parameter and where low values indicate better implementation potential (SANDAK et al., 2016). Among the cross-validation models, this model has the lowest RMSECV.

Another way of presenting the results is through a graph with the values obtained in the laboratory and predicted by the NIR spectra (Figure 1B). This figure shows the distribution of the calibration (blue) and validation (red) points of the best model for the canola oil content. A strong association is observed between the values measured

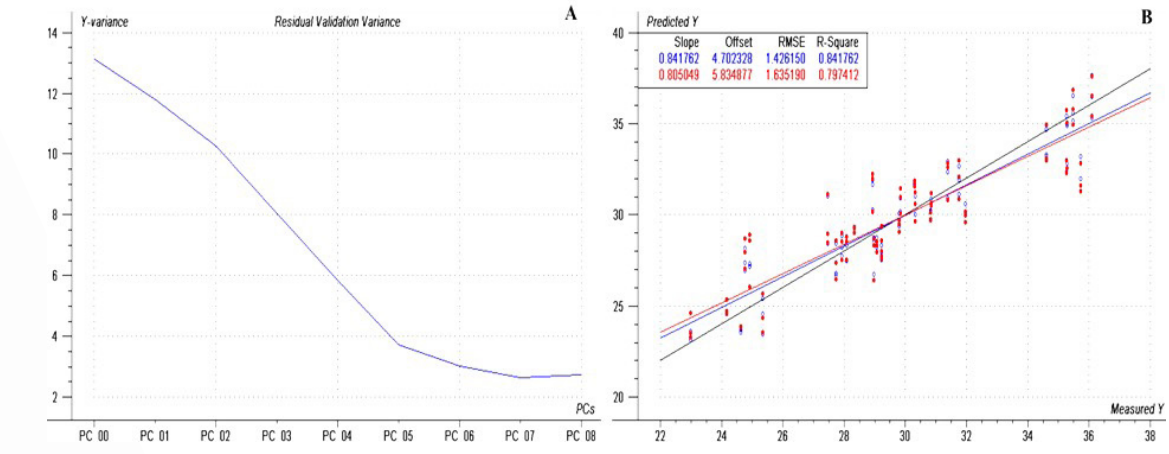

Figure 1 - *Student* x *leverage* residuals graph of the canola grain spectra (A), without three outliers, regression graph of the cross-validation with values obtained in the laboratory and predicted by NIR spectra for oil the content (B), blue color refers to calibration points and red validation, both in percentage (Model 2 of Table 4).

and those predicted by the model, indicating the possibility of using the NIR technique to estimate the oil content in canola grains.

SEN et al. (2018) used modified partial least squares (MPLS) regression to develop new NIR calibration models and predict the oil content and other constituents in two *Brassica* oleaginous species, and they found a higher value of R^2c (0.992) but the lowest value of R^2cv (0.743) for *Brassica napus*. However, better results were reported by WAN et al. (2018) when exploring the possibility of determining the canola oil content using NIR technology, with R^2c, R^2cv, RMSEC, and RMSECV values of 0.98, 0.97, 0.65, and 0.64, respectively. Similarly, BARTHET et al. (2020) developed calibrations for a portable spectrometer to determine the composition of canola seeds in terms of oil and other compounds, finding an R^2c of 0.988, R^2cv of 0.961, RMSEC of 0.27, and RMSECV of 0.49.

During model construction, it is common to pretreat the spectral data, such as by using the first derivative, to remove distortions and errors in the NIR spectra (PASQUINI, 2018). Thus, it is observed in table 4 that when treating all the data with the first derivative (Model 3), there was some breeding in the statistics, reducing the mean square error of the calibration and validation in relation to the data without treatments (osd) and increasing the values of R^2c and R^2cv to 0.79 and 0.53, respectively. When applying the first derivative to the 93 data points (removing outliers - Model 4), the statistics did not improve because despite the reduction in RMSEC and increase in R^2c, there was also a reduction in R^2cv to 0.70 and an increase in RMSECV to 1.993 (Table 4). According to Neto et al. (2017), a lower RMSE indicated that the model will be more accurate and that the dataset will be less dispersed.

According to PASQUINI (2003), the use of external validation is recommended because using it generate more realistic results that are independent of any data used in the construction of the model. For this external set of samples, the performance of the model is usually evaluated by the root mean square error prediction (RMSEP) (PASQUINI, 2003), and the lower the RMSEP is, the better the accuracy of the model (PASQUINI, 2018). Table 5 shows the statistics of the models established by PLS-R regression and external validation, in which the first column shows the dataset used to validate the model.

Table 5 shows that by manually selecting and using the last 24 raw spectral data points (samples 73 to 96, corresponding to Embrapa grains - Model 5) for external validation, good calibration was obtained, with an R^2c of 0.81 and RMSEC of 1.270 but a low R^2p of 0.36 and a high RMSEP of 4.479, resulting in an inadequate validation. This low precision reported might be due to the conditions of the last samples, from Embrapa, which might have differed from those used in the calibration (sowing days). Proper conditioning of the samples before measurement is very important to minimize this undesirable variation between samples (SANDAK et al., 2016). However, it is difficult to keep all measurement conditions and parameters constant, and the robustness of the calibration model is influenced by several external factors, such as temperature, humidity, and other uncertain variables (XU et al., 2019).

To improve the model, 24 raw spectral data points were randomly selected for external validation (Table 5), resulting in a reduction in the calibration quality (R^2c of 0.68 and RMSEC of 1.904) and breeding validation, with R^2p increasing to 0.67

Table 5 - Calibration and external validation of the canola oil content by PLS-R based on NIR spectra.

Model	Data validation	Treatment	R^2c	RMSEC	R^2p	RMSEP	LV
5	Embrapa	osd	0.81	1.270	0.36	4.479	7
6	24 random	osd	0.68	1.904	0.67	2.329	7
7	23 random	osd	0.83	1.406	0.86	1.561	7
8	24 random	1d	0.86	1.285	0.75	2.053	8

R^2c - coefficient of determination for calibration; RMSEC - root of the mean standard error for calibration; R^2p - coefficient of determination for external validation; RMSEP - root of the mean standard error for external validation; LV - latent variables; osd - original spectral data; 1d - first derivatives.

and RMSEP decreasing to 2.329; however, Model 6 was still not satisfactory.

Therefore, to further improve the model, the first derivative was applied to the 24 random raw spectral datasets for external validation (Table 5). There was an increase in the number of latent variables to 8, unlike the other models that used 7. However, there was also an increase in the values of R^2c and R^2cv to 0.86 and 0.73, respectively. Moreover, there was a reduction in RMSEC and RMSEP to 1.285 and 2.053, respectively. Despite the higher R^2 values, the RMSE values were not very low, rendering Model 8 inefficient in predicting the canola grain content.

Subsequently, two outliers were identified through *student* x *leverage* residual analysis with the 24 random raw spectral data that could be causing the reduced model efficiency. By removing these two outliers, excluding one data point from calibration and another from validation (Model 7), external validation was performed with 23 random raw spectral data points, which resulted in the best model, with R^2c and R^2p increasing to 0.83 and 0.86, respectively, and RMSEC and RMSEP decreasing to 1.406 and 1.561, respectively (Table 5). Figure 2A shows that 7 latent variables were used in the model fit. Furthermore, in figure 2B, the results are presented in a graph with the values obtained in the laboratory and predicted by the NIR spectra, showing an intense association between the values measured and those predicted by the model.

These values are close to thosereported by SIDHU et al. (2012), who evaluated NIR calibration models to predict the oil content from 3 g of canola seeds, obtaining an R^2c of 0.82, R^2p of 0.84, RMSEC of 1.39, and RMSEP of 0.61.

ROSSATO et al. (2013) sought to establish a calibration equation and estimate the efficiency of NIR spectroscopy to evaluate the canola oil content in southern Brazil; the authors found different results from the present study, with an R^2 of 0.92, RMSEC of 0.78, and RMSEP of 1.22. Higher values were also obtained by PETISCO et al. (2010), who reported an R^2c of 0.98, R^2p of 0.98, RMSEC of 0.51, and RMSEP of 0.54. These outcomes might be due to the inclusion of greater variability in the study, which used intact seeds of four varieties of *Brassica* (*B. napus* ES Hydromel, *B. napus* ES Nectar, *B. napus* ES Betty and *B. carinata* Line C-101) collected in two different crop years.

Despite the better models of the present study, presenting lower R^2 values than most approaches reported in the literature, are considered satisfactory and show good predictability. R^2 values between 0.66 and 0.81 indicated approximate quantitative predictions, values between 0.82 and 0.90 showed good prediction and values greater than 0.91 indicated optimal calibration for the models (WILLIAMS et al., 2019).

Therefore, the best model is independent validation using 23 random raw spectral data, with an

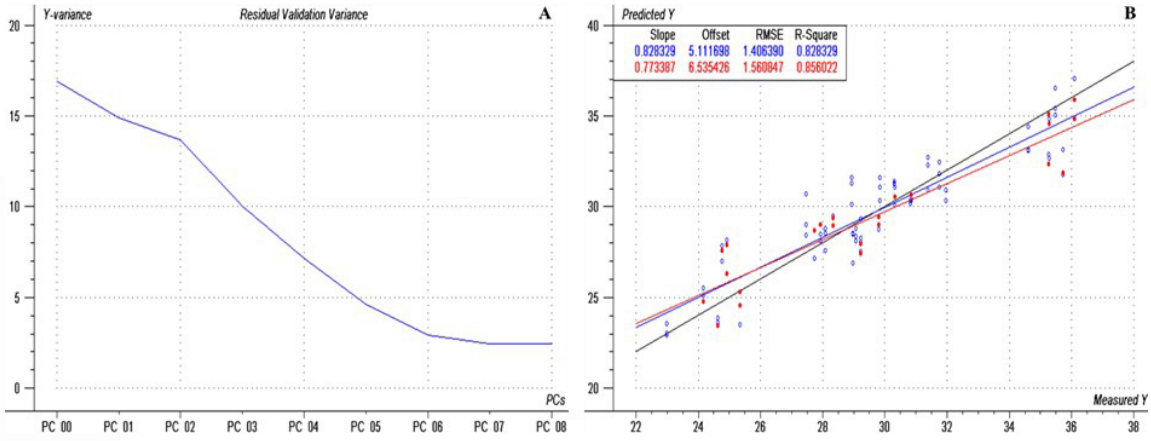

Figure 2 - *Student x leverage* residuals plot of the spectra of canola grains (A) without two outliers; regression plot of random external validation with values obtained in the laboratory and predicted by the NIR spectra for the oil content (B), blue color refers to calibration points and red validation, both in percentage (Model 7 of Table 5).

R^2 of 0.83 during calibration and an R^2 of 0.86 during validation. Moreover, the RMSE is lower than those obtained during calibration with cross-validation, corresponding to 1.56% of the sample range, which indicated low scattering and accurate predictions for the dataset. Thus, Model 7 is satisfactory because it is possible to predict the oil content in thousands of samples with an error of only 1.56%.

Accordingly, Model 7 can be successfully used to predict the oil content in canola grains and to analyze the quality of samples after harvest, thus performing hybrid selection more rapidly. In the future, this approach can be updated, adding samples from different environments, which can further increase the range of variation in the oil content present in the genotypes, ensuring greater application and robustness to the multivariate model. Thus, the development of NIR equations is the first step to replace the traditional chemical process used to quantify oil with nondestructive methods and later the use of NIR spectroscopy in breeding programs (SEN et al., 2018).

CONCLUSION

Two traits, i.e., the number of grains in five plants (0.6857) and their heights (0.4943), had higher estimates of positive correlations with grain yield, as well as higher values of positive direct effects on yield of 0.2494 and 0.1595, respectively. The TC (-0.7848), along with DIF (-0.4520), showed a significant negative correlation with the yield variable, with the cycle having highly negative effect on yield (-0.6157). Therefore, such traits in the canola crop deserve greater attention when practicing selection in breeding programs to increase grain yield, especially the crop cycle.

The present study also allowed for the development of good predictive models for the oil content in canola grains by NIR spectroscopy, in which the best model had an R^2 of 0.86 and RMSE of 1.56 in external validation.

ACKNOWLEDGEMENTS

We would like to thank the “Fundação de Amparo e Pesquisa do Estado de Minas Gerais” (FAPEMIG), for funding the research project. And was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil - Finance code 001. And also Embrapa Agroenergia for supplying the grains.

DECLARATION OF CONFLICT OF INTEREST

We have no conflict of interest to declare.

AUTHORS’ CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [Guilherme Vieira Pimentel], [Adriano Teodoro Bruzi], [Alexsandro Carvalho Santiago] and [Paulo Ricardo Gherardi Hein]. The first draft of the manuscript was written by [Alexsandro Carvalho Santiago] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

REFERENCES

BARTHET, V. J., PETRYK, M. W. P., SIEMENS, B. Rapid nondestructive analysis of intact canola seeds using a handheld near-infrared spectrometer. **Journal of the American Oil Chemists’ Society**, v.97, p.577-589, 2020. Available from: <https://doi.org/10.1002/aocs.12335>. Accessed: Nov. 18, 2021. doi: 10.1002/aocs.12335.

CANOLA council of Canada. **Canola Grower’s Manual**, 2021. Available from: <http://www.canolacouncil.org/publication-resources/print-resources/crop-production-resources/archived-crop-production-publications/canola-growers-manual/>. Accessed: Oct. 20, 2021.

CARVALHO, C. G. P. et al. Correlation and path analyses in soybean lines sowed at different sowing dates. **Pesquisa Agropecuária Brasileira**, v.37, p.311-320, 2002. Available from: <https://www.scielo.br/j/pab/a/M6SxFv8HWY5MZCyKjnR4Vhf/?format=pdf&lang=pt>. Accessed: Dec. 10, 2021. doi: 10.1590/S0100-204X2002000300012.

CHENG, J. et al. NIR hyperspectral imaging with multivariate analysis for measurement of oil and protein contents in peanut varieties. **Analytical Methods**, v.9, p.6148-6154, 2017. Available from: <https://doi.org/10.1039/C7AY02115A>. Accessed: Jun. 17, 2021. doi: 10.1039/C7AY02115A.

COIMBRA, J. L. M. et al. Path analysis of grain yield components in canola genotypes. **Ciência Rural**, v. 34, p.421-1428, 2004. Available from: <https://www.scielo.br/j/cr/a/NZHBTQYR7FbYp4hNScPvtzP/?lang=pt&format=pdf>. Accessed: Aug. 11, 2021. doi: 10.1590/S0103-84782004000500015.

CRUZ, C. D. **Programa GENES**: estatística experimental e matrizes. Viçosa: UFV, 2006. 285 p.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: UFV, 2006, 585 p.

CRUZ, C. D. GENES: a software package for analysis in esperimental statistics and quantitative genetics. **Acta Scientiarum**, v.35 (3), p.271-276, 2013. Available from: < https://doi.org/10.4025/actasciagron.v35i3.21251>. Accessed: Oct. 02, 2021. doi: 10.4025/actasciagron.v35i3.21251.

DIEPENBROCK, W. Análise de rendimento de colza oleaginosa de inverno (Brassica napus L.): uma revisão. **Field Crops Research**, v.67, p.35-49, 2000. Accessed: Dec. 02, 2021. doi: f10.1016/S0378-4290(00)00082-4

EDWARDS, J., HERTEL, K. **Canola growth and development**. Australia: Department of Primary Industries, 2011, 87 p.

FERREIRA, M. M. C. et al. Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, v.22, p.724-731, 1999. Available from: <https://doi.org/10.1590/S0100-40421999000500016>. Accessed: Dec. 08, 2021. doi: 10.1590/s0100-40421999000500016.

FONT, R. et al. The use of near-infrared spectroscopy (NIRS) in the study of seed quality components in plant breeding programs. **Industrial Crops and Products**, v.24, p.3007-313, 2006. Available from: <https://doi.org/10.1016/j.indcrop.2006.06.012>. Accessed: Nov. 15, 2021. doi: 10.1016/j.indcrop.2006.06.012.

IUPAC - International Union of Pure and Applied Chemistry. **Standard methods for the analysis of oils, fats and derivatives**, 1979, p.136.

KAUR, B. et al. Development of near-infrared reflectance spectroscopy (NIRS) calibration model for estimation of oil content in *Brassica juncea* and *Brassica napus*. **Food Analytical Methods**, v.10, p.227-233, 2017. Available from: <https://doi.org/10.1007/s12161-016-0572-9>. Accessed: Dec. 20, 2021. doi: 10.1007/s12161-016-0572-9.

KRÜGER, C. A. M. B. et al. Relations of environments variables and subperiods in yield and content oil in canola. **Ciência Rural,** v.44, p.1671-1677, 2014. Available from: <https://doi.org/10.1590/0103-8478cr20121331>. Accessed: Nov. 22, 2021. doi: 10.1590/0103-8478cr20121331.

LONG, D. S. et al. In-stream measurement of canola (*Brassica napus* L.) seed oil concentration using in-line near infrared reflectance spectroscopy. **Journal of Near Infrared Spectroscopy**, v.20, p.387-395, 2012. Available from: <https://doi.org/10.1255/jnirs.993>. Accessed: Nov. 12, 2021. doi: 10.1255/jnirs.993.

LOPES, A. C. D. A. et al. Variability and correlations among traits in soybean crosses. **Scientia Agricola**, v.59, p.341-348, 2002. Available from: <https://doi.org/10.1590/S0103-90162002000200021>. Accessed: Nov. 28, 2021. doi: 10.1590/S0103-90162002000200021.

LUZ, G. L. D. et al. Baseline temperature and cycle of canola hibrids. **Ciência Rural,** v.42, p.1549-1555, 2012. Available from: <https://doi.org/10.1590/S0103-84782012000900006>. Accessed: Nov. 28, 2021. doi: 10.1590/S0103-84782012000900006.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to linear regression analysis.** New York: J. Wiley, 1981, 504 p.

NETO, A. J. S. et al. Non-destructive prediction of pigment content in lettuce based on visible–NIR spectroscopy. **Journal of the Science of Food and Agriculture**, v.97, p.2015-2017. Available from: <https://doi.org/10.1002/jsfa.8002>. Accessed: Dec. 23, 2021. doi: 10.1002/jsfa.8002.

PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v.14, p.198-219, 2003. Available from: <https://doi.org/10.1590/S0103-50532003000200006>. Accessed: Oct. 20, 2021. doi: 10.1590/S0103-50532003000200006.

PASQUINI, C. Near infrared spectroscopy: A mature analytical technique with new perspectives - A review. **Analytica Chimica Acta**, v.1026, p.8-36, 2018. Available from: <https://doi.org/10.1016/j.aca.2018.04.004>. Accessed: Dec. 03, 2021. doi: 10.1016/j.aca.2018.04.004.

PETISCO, C. et. al. Measurement of quality parameters in intact seeds of Brassica species using visible and near-infrared spectroscopy. **Industrial Crops and Products**, v.32, p.139-146, 2010. Available from: <https://doi.org/10.1016/j.indcrop.2010.04.003>. Accessed: Oct. 10, 2021. doi: 10.1016/j.indcrop.2010.04.003.

R Core Team (2017). **R:** a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: <https://www.R-project.org/>. Accessed: Aug. 05, 2021.

ROCHA, L. de S. et al. Análise de trilha para produtividade de grãos em canola no Cerrado. **In:** 7° Congresso da rede brasileira de tecnologia e inovação de biodiesel. Florianópolis: Embrapa Agroenergia. (2019).

ROSA, W. B., et al. Influence of sowing times on subperiods and agronomic performance of canola hybrids. **Brazilian Journal of Development**, v.6, p.65774-65788, 2020. Available from: <https://doi.org/10.34117/bjdv6n9-126>. Accessed: Nov. 13, 2021. doi: 10.34117/bjdv6n9-126.

ROSSATO, R. et al. Predicting rapeseed oil content with near-infrared spectroscopy. **Pesquisa Agropecuária Brasileira**, v.48, p.1601-1605, 2013. Available from: <https://doi.org/10.1590/S0100-204X2013001200010>. Accessed: Dec. 09, 2021. doi: 10.1590/S0100-204X2013001200010.

SANDAK, J. et al. Assessing trees, wood and derived products with near infrared spectroscopy: hints and tips. **Journal of Near Infrared Spectroscopy**, v.24, p.485-505, 2016. Available from: <https://doi.org/10.1255/jnirs.1255>. Accessed: Sept. 10, 2021. doi: 10.1255/jnirs.1255.

SEN, R. et al. Near-infrared reflectance spectroscopy calibrations for assessment of oil, phenols, glucosinolates and fatty acid content in the intact seeds of oilseed Brassica species. **Journal of the Science of Food and Agriculture**, v.98, p.4050-4057, 2018. Available from: <https://doi.org/10.1002/jsfa.8919>. Accessed: Sept. 03, 2021. doi: 10.1002/jsfa.8919.

SIDHU, H. K. et al. Nondestructive analysis of single plant canola (Brassica napus) seeds using near infra-red spectroscopy. **American Society of Agricultural and Biological Engineers**, 2012. Available from: <https://elibrary.asabe.org/abstract.asp?aid=41762>. Accessed: Dec. 17, 2021. doi: 10.13031/2013.41762.

STEEL, R. G. D.; et al. **Principles and Procedures of Statistics**: A Biometrical Approach. New York: MacGraw-Hill Book Company, 1997, 688 p.

TOMM, G. O. **Indicativos tecnológicos para produção de canola no Rio Grande do Sul**. Passo Fundo: Embrapa Trigo, 2007. 32p. (Sistema de Produção INFOTECA-E).

TOMM, G. O. et al. **Tecnologia para a produção de canola no Rio Grande do Sul**. Passo Fundo: Embrapa Trigo, 2009a. 86p. (Documentos INFOTECA-E).

TOMM, G. O. et al. **Panorama atual e indicações para aumento de eficiência da produção de canola no Brasil**. Passo Fundo: Embrapa Trigo, 2009b. 34p. (Documentos INFOTECA-E).

TOMM, G.O. et al. **Efeito de épocas de semeadura sobre o desempenho de genótipos de canola de ciclo precoce e médio**.

Passo Fundo: Embrapa Trigo, 2010. (Boletim de Pesquisa e Desenvolvimento INFOTECA-E).

USDA - United States Department of Agriculture. **Oilseeds**: World Markets and Trade, 2021. Available from: <https://apps.fas.usda.gov/psdonline/circulars/oilseeds.pdf.>. Accessed: Oct, 20, 2021.

VENCOVSKY, R., BARRIGA P. **Genética biométrica no fitomelhoramento.** Ribeirão Preto: Revista Brasileira de Genética, 1992. 504 p.

WAN, L. et al. Rapid determination of oil quantity in intact rapeseeds using near-infrared spectroscopy**. Journal of Food Process Engineering**, v.41, 12594. 2018. Available from: <https://doi.org/10.1111/jfpe.12594>. Accessed: Aug. 11, 2021. doi: 10.1111/jfpe.12594.

WESTAD, F., MARTENS, F. Variable selection in near infrared spectroscopy based on significance testing in partial least square regression. **Journal of Near Infrared Spectroscopy**, v.8, p.117-124, 2000. Available from: <https://doi.org/10.1255/jnirs.271>. Accessed: Sept. 17, 2021. doi: 10.1255/jnirs.271.

WILLIAMS, P., et al. **Near-infrared technology**: Getting the best out of light. África do Sul: African Sun Media, 1992, 311 p.

WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, v.20, p.557-585, 1921. Available from: <https://doi.org/10.2307/2287275>. Accessed: Sept. 12, 2021. doi: 10.2307/2287275.

XU, X., et al. Factors influencing near infrared spectroscopy analysis of agro-products: a review. **Frontiers of Agricultural Science and Engineering**, v.6, p.105-115, 2019. Available from: <https://doi.org/10.15302/J-FASE-2019255>. Accessed: Dec. 03, 2021. doi: 10.15302/J-FASE-2019255.