Revista Ciência Agronômica, v. 55, e20238703, 2024
Centro de Ciências Agrárias - Universidade Federal do Ceará, Fortaleza, CE
www.ccarevista.ufc.br

Scientific Article
ISSN 1806-6690

Near infrared spectroscopy for the classification of vigor level of soybean seed[1]

Martha Freire da Silva[2], Jussara Valente Roque[3], Júlia Martins Soares[4], Lorena de Oliveira Moura[5], André Dantas de Medeiros[6], Felipe Lopes da Silva[7], Laércio Junio da Silva[8]

ABSTRACT - This work aimed at investigating the viability of near infrared spectrometry (NIR), associated with chemometric methods, in order to identify differences at the levels of vigor of naturally and artificially aged soybean seeds. Seeds of six soybean genotypes were analyzed when freshly harvested, after natural aging in storage for eight months, and after artificial aging at the temperature of 41 °C for 96 hours. The seed moisture content, germination potential and vigor were evaluated. Also, there were measured the content of protein, oil and of the fatty acids: palmitic, stearic, oleic, linoleic and linolenic. The NIR spectra were obtained from the freeze-dried and grinded seeds. The natural and artificial aging of the seeds promote deterioration at distinct levels, reflecting in differences in seed vigor. The regions of the electromagnetic spectrum between wavelengths of 1000-1200 nm, 1350-1450 nm, 1800-1900 nm and 2300-2400 nm are important to distinguish the levels of quality of the soybean seeds. The contents of oil and protein have a relationship with the physiological quality of the seeds. Also, the most relevant wavelengths for the classification of seed vigor present a relationship with these compounds. NIR spectroscopy, in combination with chemometric methods, presents potential to be used in the classification of soybean seeds regarding their physiological quality.

Key words: NIR. Seed Storage. Seed deterioration. Physiological quality of seeds. Biochemical composition of seeds.

DOI: 10.5935/1806-6690.20240005
Editor-in-Chief: Prof. Alek Sandro Dutra - alekdutra@ufc.br
*Author for correspondence
Received for publication 01/02/2023; approved on 20/07/2023
[1]Research work
[2]Department of Agricultural Sciences, State University of Maringá (UEM), Umuarama-PR, Brazil, marthafreire86@hotmail.com (ORCID ID 0000-0002-3958-8169)
[3]Institute of Chemistry, Federal University of Goiás, Goiânia-GO, Brazil, jussara_roque@ufg.br (ORCID ID 0000-0002-5220-959X)
[4]Department of Agronomy, Federal University of Viçosa (UFV), Viçosa-MG, Brazil, julia.m.soares@ufv.br (ORCID ID 0000-0001-9432-6009)
[5]Department of Agronomy, Federal University of Viçosa (UFV), Viçosa-MG, Brazil, lorenaomoura@gmail.com (ORCID ID 0000-0002-0052-4651)
[6]Department of Agronomy, Federal University of Viçosa (UFV), Viçosa-MG, Brazil, andre.d.medeiros@ufv.br (ORCID ID 0000-0002-1097-0292)
[7]Department of Agronomy, Federal University of Viçosa (UFV), Viçosa-MG, Brazil, felipe.silva@ufv.br (ORCID ID 0000-0001-9866-9615)
[8]Department of Agronomy, Federal University of Viçosa (UFV), Viçosa-MG, Brazil, laercio.silva@ufv.br (ORCID ID 0000-0001-7202-0420)

INTRODUCTION

The physiological quality of the seeds is one of the main factors responsible for the success of soybean crops in the world. Seeds of high quality have a good performance in the field, quick emergence and establishment of plants, even if under non-favorable environmental conditions (EBONE *et al*., 2020). One the other hand, the use of seeds of a low quality usually causes slow initial growth of plants, failures in the stand of plants in the field and low tolerance to biotic and abiotic stresses (BEWLEY *et al*., 2013).

Routinely, the physiological quality of the seeds is evaluated through tests of germination and vigor in laboratories (AL-AMERY *et al*., 2018). For soybean, the methods developed for the evaluation of vigor include the tests of germination, tetrazolium, accelerated aging, emergence, emergence speed index, electrical conductivity and those based on seedling performance.

The physical and chemical properties of the seeds can also be an indicative of vigor (HAYATI; ANGGASTA; ANWAR, 2020) and the quantification of compounds can help in the discrimination of physiological quality among seed lots. However, all these tests involve laborious procedures, which require a lot of time to prepare, execute and analyze, need labor and experience of the analyst and do not always represent the performance of the seeds in the field (FAN; MA; WU, 2020; ZHANG *et al*., 2020).

In this sense, the use of techniques that require less time and labor, and allow the evaluation of the physiological quality in a fast, precise and assertive way is suggested. In this context, near infrared spectroscopy (NIR) arises as a promising technique for these purposes (FAN; MA; WU, 2020; MAYRINCK *et al*., 2020). The obtainment of NIR spectra is carried out in an instrument of easy operation, able to generate a great amount of information, with a short time for the analysis and little need of labor. Also, it does not generate polluting waste, and makes use of a small amount of samples, with no need of pre-processing (LI *et al*., 2020; MAYRINCK *et al*., 2020).

The radiation of the NIR electromagnetic region (780-2500 nm) is absorbed by water and organic compounds such as carbohydrates, proteins, oils or alcohols (AGELET; HURBURGH, 2014). The absorbed energy by a sample, calculated from the transmitted or diffusely reflected radiation, may be related to compound content (AGELET; HURBURGH, 2014). The spectral information obtained from organic molecules extracted from readings in the NIR equipment, make it possible to highlight differences in the biochemical composition of the seeds, which may be related to viability (KUSUMANINGRUM *et al*., 2018) and vigor (FAN; MA; WU, 2020).

Several works have been published proving the efficiency of the use of NIR to check differences in the quality of seeds (BAZONI *et al*., 2017; FAN; MA; WU, 2020; HUANG *et al*., 2013; KUSUMANINGRUM *et al*., 2018; MAYRINCK *et al*., 2020; YASMIN *et al*., 2019; ZHANG *et al*., 2020). However, studies on NIR viability to distinguish levels of vigor in seeds are still incipient. Thus, this work aimed at analyzing the viability of NIR spectroscopy, in combination with chemometric methods, to identify differences in the vigor levels of naturally and artificially aged soybean seeds.

MATERIAL AND METHODS

The experiment was carried out in the Laboratório de Pesquisa de Sementes and Instituto de Biotecnologia Aplicada à Agropecuária ("Seed research laboratory" and "Institute of Biotechnology Applied to Farming"), both belonging to the Federal University of Viçosa – MG, Brazil.

Seeds of six soybean cultivars were used. The soybean seeds used were from pre commercial strains, and they were called GEN1 (maturity group 5.9, with RR technology); GEN2 (maturity group 5.6, with RR technology); GEN3 (maturity group 5.8, with IPRO technology); GEN4 (maturity group 5.5, with IPRO technology), GEN5 (maturity group 6.8, with IPRO technology) and GEN6 (maturity group 5.5, with IPRO technology). The seeds of the six genotypes were produced under the same cultivation conditions in the town of Passo Fundo, RS, Brazil. The seeds were analyzed when freshly harvested and then submitted to storage (natural aging) and artificial aging.

For natural aging, the seeds were put in Kraft Multifolha paper bags and stored under a non-refrigerated condition, in a shed, in the town of Passo Fundo, RS, Brazil, for eight months.

For artificial aging, the seeds were distributed on metal mesh trays coupled to plastic boxes of the *gerbox* type (11 x 11 x 3.5 cm), containing 40 mL of distilled water on the bottom, and kept under relative humidity of 100% at 41 ^{o}C, for 96 hours. After artificial aging, the seeds were left on a countertop in a laboratory environment for natural drying until they reached their initial moisture content (approximately 12%).

The freshly harvested seeds (initial), the artificially aged ones (art aging) and those stored for eight months (stored) were submitted to the following tests and determinations:

Moisture content (M) - The moisture content of the seeds was determined through the oven method, at 105 ± 3 °C, for 24 h, using four replications of 50 seeds each (BRASIL, 2009).

Germination (G) - Four replications of 50 seeds each were used. The seeds were sown in moistened germination paper with water volume equivalent to 2.5 times the weight of the dry substrate and kept in a germinator at 25 °C. Evaluations were carried out, with the recording of the percentage of the normal seedlings on the 5th and 8th days after sowing (BRASIL, 2009).

Accelerated Aging (AA) - Four replications of 50 seeds each were used. The seeds were distributed on a single layer on a metal mesh tray coupled to a plastic box of the gerbox type (11 x 11 x 3.5cm) which contained, on the bottom, 40 ml of distilled water. The boxes were covered in order to obtain 100% of relative humidity inside them and kept in a BOD chamber at 41 °C for 48 hours. After this period, the seeds were submitted to the germination test and the percentage of normal seedlings was evaluated on the 5th day after sowing.

Seedling growth - Four replications of 20 seeds were used, evenly sown on two sheets of germination paper, moistened with distilled water at the proportion of 2.5 times the weight of the dry paper. The seeds were put to germinate, in moistened paper rolls, and kept in a germinator, at 25 °C, for three days. The seedlings were scanned and, using Software Vigor-S®, root length and the aerial part were measured. The data of length, together with the germination data, were used for the calculation of vigor index (VI) (MEDEIROS; PEREIRA, 2018), obtained by means of package SeedCalc of the software R (SILVA; MEDEIROS; OLIVEIRA, 2019).

Seedling dry mass (SDM) - The seedlings used in the growth test were used to obtain dry mass, obtained after the seedlings were dried in an oven with air-forced circulation, at 70 °C for 72 hours.

Emergence Speed Index (ESI) - Four replications of 50 seeds were sown on polystyrene trays containing 2 liters of sand. The substrate was initially moistened until it reached 60% of water retention capacity and it was irrigated daily. Daily counts of the number of emerged seedlings were carried out until the 12th day after seedling. The count data were used to obtain the emergence speed index, according to what was proposed by Maguire (1962).

Electrical Conductivity (EC) - Four replications of 50 seeds were weighed and put in plastic cups containing 75 mL of distilled water, and they were kept in a BOD chamber at 25° C, for 24 hours. After this period, the electrical conductivity of the solution was determined, by making use of a conductivity meter.

Oil Content - Three replications of 50 seeds each were used. The seeds were freeze-dried and grinded in a cutting mill, and aliquots of 0.9 g of the powder of the grinded seed was harvested from each sample. The samples were weighed, in 'filter bags', and dried in an oven at 105 °C for 3 hours. After this procedure, the samples were weighed again and put in extractor Ankom® XT15, where the extraction of the oil was carried out for 50 min at 90 °C, using petroleum ether as the extractor. After the extraction, the samples were removed from the extractor and dried in an oven at 105 ºC for 30 minutes. Oil content was calculated based on the difference of weights between the samples before and after the extraction, according to the AM 5-04 methodology of the American Oil Chemists' Society (AOCS).

Content of Soluble Protein - Four replications of 10 seeds each were used. The seeds were soaked for 16 hours. The teguments were removed and the embryos of the seeds were freeze-dried and grinded in a ball mill for the obtainment of a fine powder. A subsample of 100 mg of the grinded material was used for each one of the replications of each treatment. The determination of the content of soluble protein was carried out in accordance with the methodology described by Bradford (1976), using BSA as a standard. The reading was carried out in a spectrophotometer at the wavelength of 595 nm.

Content of Fatty Acids - The content of fatty acids stearic, palmitic, oleic, linoleic and linolenic in the fraction of soybean oil was determined through gas chromatography. Ten seeds were used per treatment, and three replications were carried out per treatment. The seeds were freeze-dried and grinded in a cutting mill, and 150 mg of the grinded material was used per sample. The samples were put in microtubes, to which 1 mL of hexane was added, and kept at 4 ºC for 16 h. After this period, the lipid fraction was poured into tubes and the solvent was evaporated by N_2 bubbling. In order to obtain methyl esters, the methodology described by Jham, Teles and Campos (1982) was used. After sample preparation, aliquots were injected into a CG-17^a gas chromatographer, equipped with an automatic sampler (Shimadzu, model AOC-17) and integrator (Shimadzu, model C-R7A). The Carbowax capillary column (30 m x 0,32 mm) was kept at 225 ºC, and the temperatures of the injector and detector were 245 ºC and 280 ºC, respectively. Nitrogen gas was used as a carrier, at a flow of 1.1 mL min^{-1}.

Near Infrared Spectroscopy (NIR) - Three replications of 50 seeds each were used. The seeds were freeze-dried and grinded in a cutting mill, and an aliquot of 10g of the powder of the grinded seed was removed per sample. Three spectra were harvested per sample. Each spectrum was obtained from 32 scans in the range between 10.000 - 4000 cm^{-1} (1.000 – 2.500 nm), with a resolution of 4 cm^{-1}. The mean was carried out and a single spectrum was obtained for each sample. The readings were carried out in spectrometer Thermo Scientific, Antaris II model, in reflectance mode (R), and the results were expressed in log (1/R).

Statistical Analysis **-** The experiment was carried out in a completely randomized design, in a factorial scheme, with the use of six genotypes and three kinds of aging: freshly harvested seeds (initial), naturally aged seeds (stored for eight months), and artificially aged seeds (in BOD, at 41 ºC and 100% RH, for 96 hours), totaling 18 treatments and three replications each. The treatments were separated into classes, and, therefore, only the effects of the different kinds of aging on the physiological performance of the seeds were investigated. The data were submitted to the analysis of variance and the means were compared by using the Tukey test, at 5% probability. To verify the relationship of the aging levels and the physiological and biochemical variables jointly, the principal component analysis (PCA) was carried out (JOLLIFFE; CADIMA, 2016).

The NIR spectra were labelled as high, intermediary and low vigor, according to the physiological performance of the seeds. Initially, an exploratory analysis was carried out with the spectra, by plotting charts from the original spectra per treatment, from the mean of the spectra per class, followed by the PCA (JOLLIFFE; CADIMA, 2016), by making use of the original spectra. Afterwards, the data were submitted to the chemometric methods of pre-treatments and construction of classification models. The following pre-treatments were used: autoscaling, multiplicative scatter correction (MSC), standard normal variate (SNV), and 1st and 2nd derivates through the Savitzky-Golay method, by using a window of 13 variables (SAVITZKY; GOLAY, 1964). The number of latent variables (LV) was determined through a random crossed validation with ten divisions. For each combination of pre-treatment, a classification model was generated by means of the PLS-DA (BARKER; RAYENS, 2003), using 90% of the data for training and 10% for the crossed validation. The performance of the models was evaluated according to their accuracy, sensitivity and specificity for the training and the crossed validation. Accuracy indicates the general performance of the model. Among all the classifications, it indicates how many of them the model classified correctly (BARKER; RAYENS, 2003). Sensitivity is the number of samples predicted to belong to the class divided by the number of samples that belong to the class. Specificity is the number of samples predicted not to be in the class divided by the real number that is not in the class. Accuracy, sensitivity and specificity were calculated according to Equations (1), (2), (3), respectively:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

where, TP is true positive, TN is true negative, FN is false negative and FP is false positive. TP is the number of samples that belong to class i, classified as belonging to class i. TN is the number of samples that do not belong to class i, classified as not belonging to class i. FN is the number of samples belonging to class i, not classified as belonging to class i. FP is the number of samples that do not belong to class i, classified as belonging to class i (BARKER; RAYENS, 2003).

For the construction of classification models via PLS-DA, the entire NIR spectrum was used, comprising the range between 1000 - 2500 nm. The wavelength bands that represented importance above 50% were also identified for each class.

The data analyses were carried out by means of software R Core Team 4.0.2, using the packages *caret*, *signal*, *prospectr*, *nira* and *patchwork*.

RESULTS AND DISCUSSION

There was no significant effect between the interactions of genotypes and types of aging for all analyzed variables. Thus, the average performance results of physiological quality (Figure 1) and chemical composition of the seeds (Figure 2) were presented, when freshly harvested and naturally or artificially aged.

Seed water content was approximately 12% before the performance of the tests. The freshly harvested seeds (initial) presented greater physiological quality if compared to the other treatments, with a higher percentage of germination (G) and of normal seedlings after the accelerated aging test (AA), a higher emergence speed index (ESI), seedling dry mass (SDM), vigor index (VI) and lower electrical conductivity (EC). On the other hand, the artificially aged seeds (Art aging) presented a lower physiological performance, with smaller values of G, AA, SDM, VI, ESI and greater EC. The seeds that were stored for eight months (stored) presented an intermediary performance (Figure 1). Thus, the seeds were classified as having high vigor (initial), intermediary vigor (stored) and low vigor (art aging).

The content of soluble protein (PROT) was lower in the stored seeds, and it was not different between the freshly harvested and the artificially aged seeds. However, oil content was greater in the stored seeds, followed by the artificially aged ones. There was no significant difference in the content of fatty acids palmitic, steric, oleic, linoleic and linolenic among the seeds with different vigor levels (Figure 2).

In the joint analysis of the physiological and biochemical data (Figure 3), about 79% of the total

variation of the data could be explained by the first three principal components. In this analysis, the most important variables to distinguish treatments were content of oleic fatty acid, vigor index, emergence speed index, germination, oil content, seedling dry mass, content of linolenic acid, content of protein and accelerated aging, respectively (Figure 3).

It was possible to point out the separation between the freshly harvested seeds (concentrated on the positive scores of Dim1) from the seeds that were artificially and naturally aged (negative scores of Dim1). However, there was no great distinction between the stored seeds and the ones that were artificially aged, when vigor and the results of the biochemical analyses were analyzed jointly (Figure 3).

By means of the NIR spectra, it was possible to pinpoint differences among the three levels of the physiological quality of the seeds according to the kinds of aging (Figure 4).

More than 99% of the variation of the spectra data could be explained by the first two principal components in the PCA analysis by using the original spectra, which highlighted the distinction among the vigor levels of the seeds (Figure 5).

Figure 1 - Germination (G). accelerated aging (AA). vigor index (VI). seedling dry mass (SDM). electrical conductivity (EC) and emergency speed index (ESI) of freshly harvested soybean seeds (initial). After storage for eight months (stored) and artificially aged at 41 °C for 96 hours (Art. Aging)

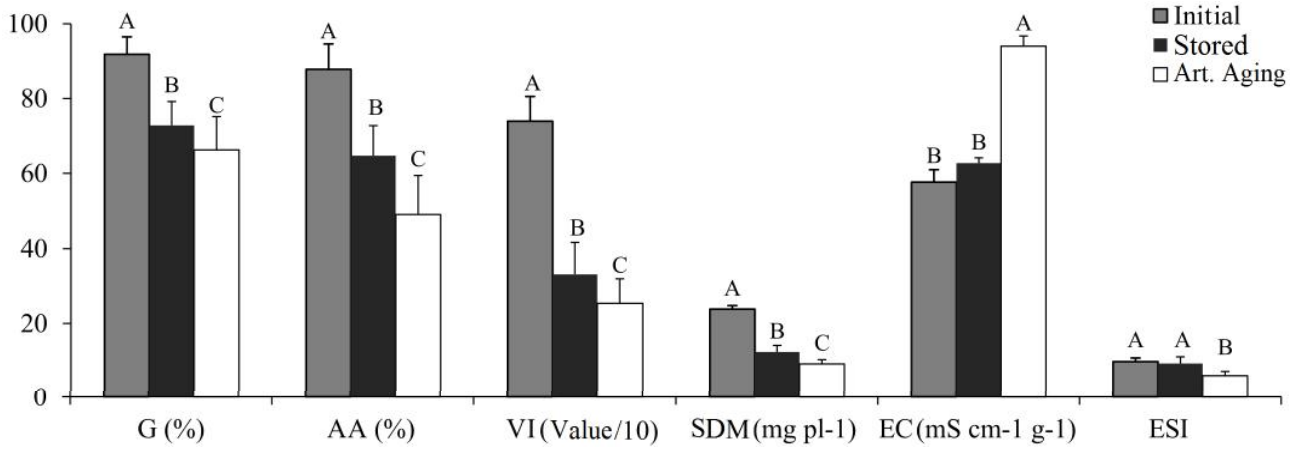


Figure 2 - Contents of protein (PROT). oil and fatty acids palmitic (PALM). estearic (ESTEA). oleic (OLEIC). linoleic (LINLC) and linolenic (LINLN) in freshly harvested soybeans seeds (initial). after storage for eight months (stored) and artificially aged at 41 °C for 96 hours (Art. Aging)

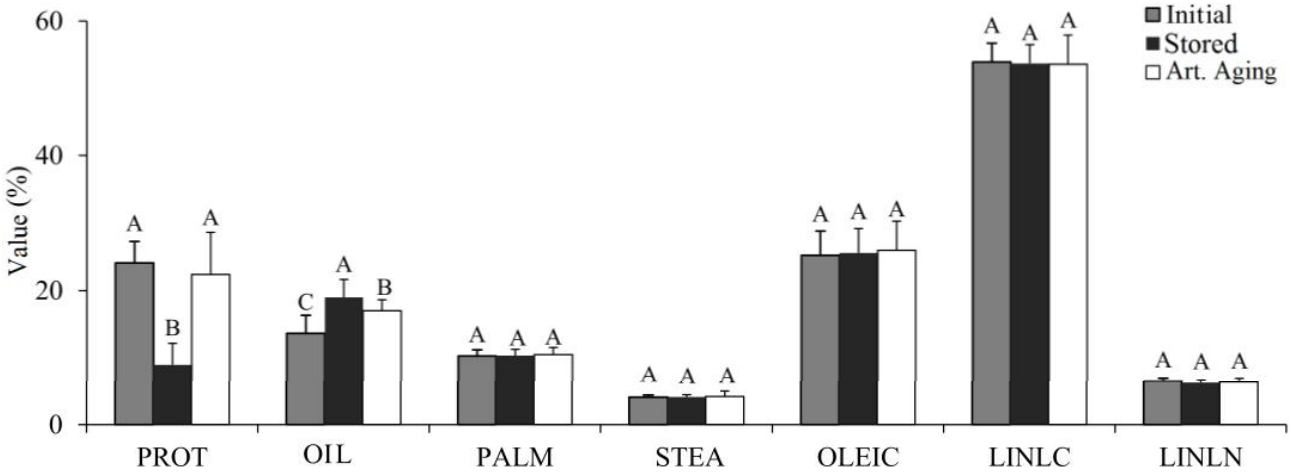


Figure 3 - Biplot of the principal component analysis (PCA) **(A)** and contribution of the physiological and biochemical **(B)** variables for the distinction among treatments. Oleic = oleic acid; VI = vigor index; ESI= emergency speed index; G= germination; SDM = seedling dry mass; LINLN = linolenic acid; PROT = protein content; AA = accelerated aging; EC = electrical conductivity; LINLC = linoleic acid; PALM = palmitic acid; STEA = stearic acid

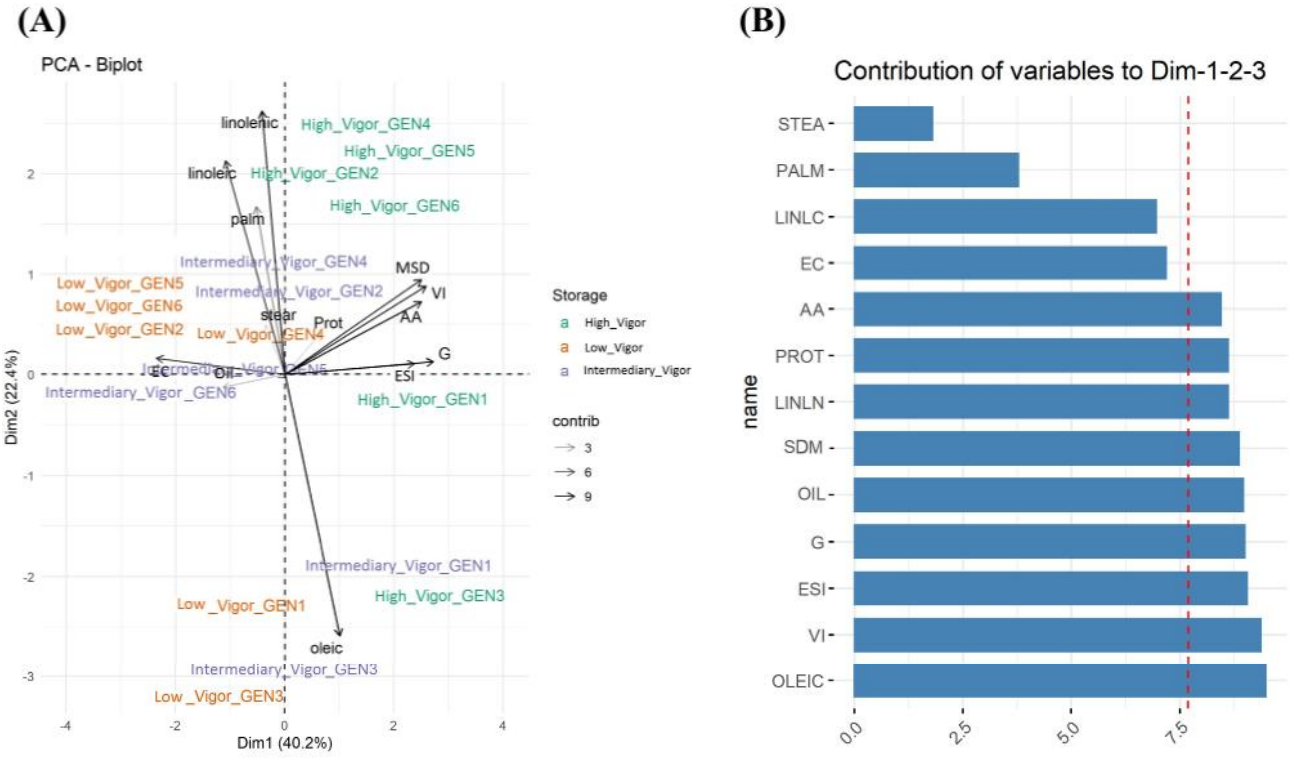
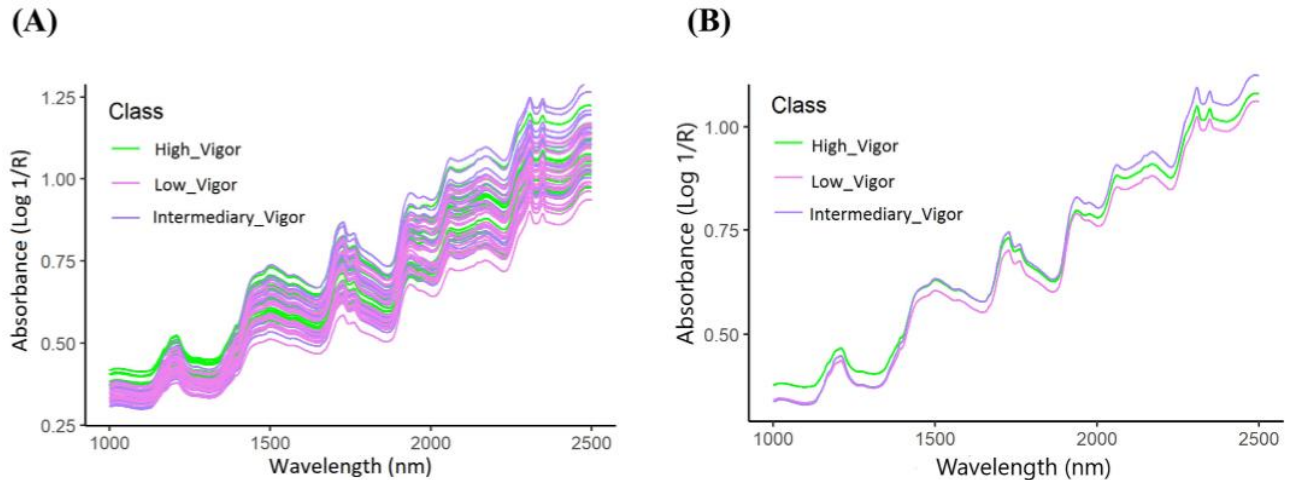


Figure 4 - Original spectra (A) and means of the spectral data per class (B). In purple. freshly harvested seeds (high vigor); in green. seeds stored for eight months (intermediary vigor); and. in pink. artificially aged seeds 41°C for 96 hours (low vigor)



The spectra were submitted to different kinds of pre-treatments and, by means of the PLS-DA method, models of classification of the three levels of the physiological quality of the seeds were obtained: high vigor (freshly harvested seeds), intermediary vigor (stored seeds) and low vigor (artificially aged seeds) (Table 1).

All the models presented high accuracy, sensitivity and specificity (Table 1). The classification model obtained through the original spectra, without any kind of data pre-treatment, demonstrated high accuracy (0.98), sensitivity (0.98) and specificity (0.99). After the application of pre-treatments 1st SG derivate and 2rd SG derivate,

followed by the auto-scaling procedure, the metrics of the classification models presented increment, with similar or very close to 1 values of accuracy, sensitivity and specificity, both in the training and in the crossed evaluation, using a smaller number of latent variables (8 and 5, respectively) (Table 1).

Considering the model in which pre-treatments 2nd SG derivate followed by auto-scaling were used, it was possible to observe a clear separation of the classes viewed in the score chart of the PLS-DA (Figure 6). Based on this classification model, graphs of the importance of the variables for each class were also constructed (Figure 7).

Figure 5 - Principal component analysis (PCA). **(A)** Scores and **(B)** PCA loadings of the original spectral data of soybean seeds of high vigor (freshly harvested). intermediary vigor (stored for eight months). and low vigor (artificially aged at 41°C for 96 hours). The scores refer to the grouping of the samples and the loadings refer to the variables

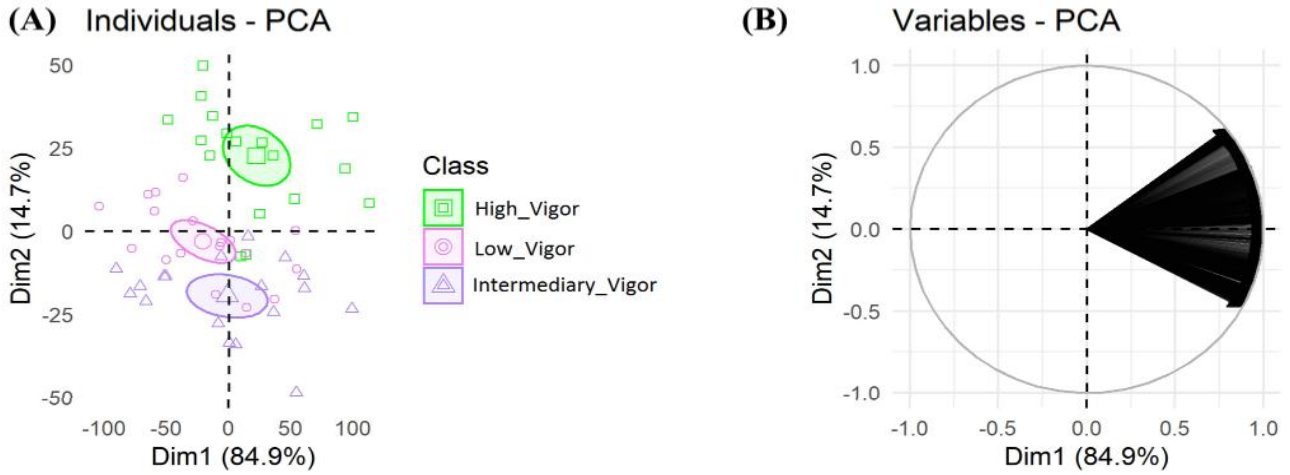


Table 1 - Number of latent variables (LV). accuracy. sensitivity (Sn) and specificity (Sp) for the data set of training and crossed validation of the classification models generated by means of the PLS-DA analysis with different pre-treatments

Pre-treatments	LV	Training			Crossed Validation		
		Accuracy	Sn	Sp	Accuracy	Sn	Sp
Original spectrum	10	1.0000	1.0000	1.0000	0.9815	0.9833	0.9908
Sc	10	1.0000	1.0000	1.0000	0.9831	0.9850	0.9916
MSC	10	1.0000	1.0000	1.0000	0.9291	0.9333	0.9652
SNV	9	0.9814	0.9814	0.9907	0.9271	0.9316	0.9641
1st SG Der	8	1.0000	1.0000	1.0000	0.9901	0.9900	0.9961
2nd SG Der	5	1.0000	1.0000	1.0000	0.9896	0.9900	0.9950
MSC + Sc	10	1.0000	1.0000	1.0000	0.9358	0.9416	0.9683
SNV + Sc	10	1.0000	1.0000	1.0000	0.9251	0.9300	0.9627
1st SG Der + Sc	8	1.0000	1.0000	1.0000	0.9866	0.9883	0.9936
2nd SG Der + Sc	5	1.0000	1.0000	1.0000	0.9903	0.9916	0.9950
1st SG Der + MSC	10	1.0000	1.0000	1.0000	0.9711	0.9733	0.9861
2nd SG Der + MSC	8	1.0000	1.0000	1.0000	0.8765	0.8766	0.9366
1st SG Der + SNV	10	1.0000	1.0000	1.0000	0.9691	0.9716	0.9850
2nd SG Der + SNV	8	1.0000	1.0000	1.0000	0.8745	0.8800	0.9383

Sc = auto-scaling; SG Der = Savitzky-Golay Derivate; MSC = Multiplicative Scatter Correction; SNV = Standard Normal Variate

Considering the variables that presented contribution above 50% for the distinction of the classes, it is noted that the wavelengths in the range between 1.350-1.450, 1.800-1.900 and 2.300-2.400 nm were important for the classification of the seeds of high vigor. For the seeds with intermediary vigor, the most important regions comprised the wavelengths between 1.000-1200; 1.350-1.450 and 2.300-2.400 nm. As for the seeds of low vigor, the regions of the most important spectrum were located between 1.350-1.450; 1800-1.900 and 2.300 a 2.400 nm.

In this work, seeds of six genotypes were analyzed initially (freshly harvested), after natural aging, in storage for eight months, and after artificial aging, in which they were exposed to the temperature of 41 °C for 96 hours under 100% relative humidity. It was observed that the different kinds of aging promoted to the seeds had distinct reflexes in their physiological quality.

Figure 6 - PLS-DA scores of the spectral data of soybean seeds of high vigor (freshly harvested). intermediary vigor (stored for eight months) and low vigor (artificially aged at 41 °C for 96 hours)

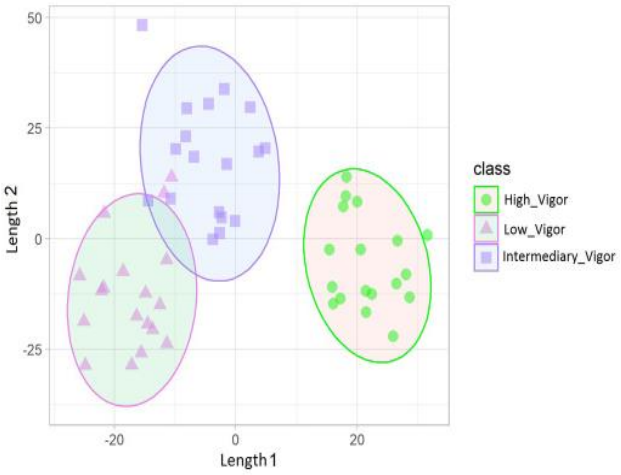
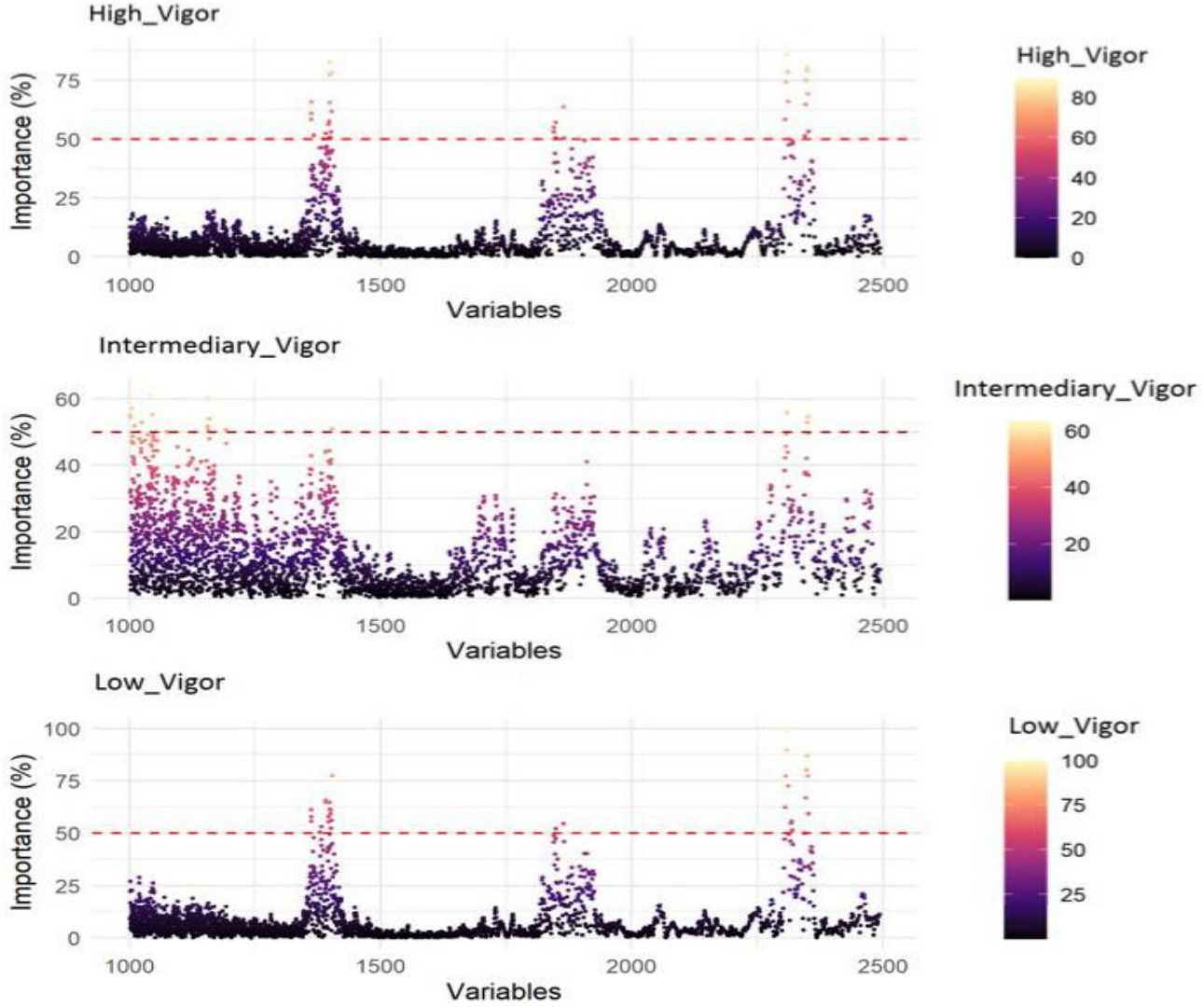


Figure 7 - Importance of the wavelength variables for classification via PLS-DA of the vigor levels of naturally and artificially aged soybean seeds. Seeds of high vigor (freshly harvested). intermediary vigor (stored for 8 months) and low vigor (artificially aged at 41 °C for 96 hours)



Vigor tests demonstrated that the freshly harvested seeds presented the highest physiological potential, followed by the seeds that were stored and by those artificially aged, respectively. The highest physiological quality of the seeds is achieved at physiological maturity point. From this moment, the deterioration process starts (BEWLEY *et al*., 2013), which justifies the reduction in the physiological quality of the seeds with the aging processes, both natural and artificial (Figure 1).

When the data of vigor and of biochemical composition of the seeds were analyzed jointly (Figure 3), it was not possible to point out a clear difference between the artificially and naturally aged seeds. However, there was a clear distinction between the aged and non-aged seeds. The freshly harvested seeds presented higher vigor and content of soluble protein, while the aged seeds presented higher electrical conductivity, higher oil content and lower physiological performance (Figures 1 and 2).

These results show the positive relationship between the content of soluble protein and seed vigor and the correlation between the highest content of oil and of electrical conductivity, with aging. Gao *et al*. (2015) also highlighted a positive relationship between oil content and a negative one of protein content with the aging of soybean seeds. Castellión *et al*. (2010) and Mathias, Coelho e Garcia (2019) point out that the content of soluble protein can be used as a positive indicator of seed quality. On the other hand, the high content of oil and poli-unsaturated fatty acids lead to a greater tendency to non-enzimatic and enzimatic peroxidation, which results in a quick decrease of seed quality (SINGH; PAROHA; MISHRA, 2017). The peroxidation of membrane lipids is considered to be the main mechanism of deterioration of oily seeds and its consequence is the increase of electrical conductivity, as it could be observed for the aged seeds of this work (Figure 1).

The NIR spectra obtained from the seeds pointed out a clear distinction among the levels of seed quality (Figure 4). In addition, in the exploratory analysis, the PCA using the original spectra allowed to highlight the distinction among the three classes of seed quality, which represented more than 99% of the total variation of the data with the first two components (Figure 5). It was noted that the PCA analysis of the original spectra data allowed a better differentiation of the kinds of aging than the PCA of the physiological and biochemical data, which demonstrated the separation of only two classes, aged and non-aged (Figures 3 and 4). This is probably due to the fact that only the contents of protein, oil and its fractions were quantified, and that the variation in the content of oil and of fatty oils was small between the naturally and artificially aged seeds (Figure 2).

However, the radiation of the near infrared electromagnetic region can be absorbed by several other compounds contained in the seeds, such as carbohydrates (GUO *et al*., 2011), soluble sugars (GUO *et al*., 2011), cellulose, hemicellulose, lignin (HUANG; YU, 2019), among others, in addition to oil and protein (AGELET; HURBURGH, 2014; GUO *et al*., 2011). Thus, the NIR spectra might contain valuable qualitative and quantitative information about organic compounds of seeds and allows greater sensitivity for the detection of differences among samples, if compared to other traditional methodologies of quantification of isolated compounds (VENKATESAN *et al*., 2020).

Although the calibration of the classification model has been satisfactory with the data of the original spectra, it was pinpointed that the pre-treatment of the spectra data by the second derivate of Savitzky-Golay, followed by the application of the auto-scaling, promoted a better performance in the calibration and crossed validation of the model (Table 1). Therefore, these pre-treatments were applied to the original spectra before the construction of the classification model and for further identification of the most relevant wavelengths. The application of pre-treatments to spectra data is common to minimize the effects of instrument noise and to improve the efficiency of the calibration models. The PLS-DA classification models showed high values of accuracy, sensitivity and specificity (Table 1), which highlights the viability of NIR, together with the application of chemometric methods, for the classification of seeds according to their vigor level. PLS-DA is a commonly used technique for the construction of classification models, allowing the separation among observation groups, so that the maximum separation among the classes is obtained (BRERETON; LLOYD, 2014). It has been often used in the classification of data of high dimensionality and with highly correlated variables, such as the spectra data obtained from NIR (BRERETON; LLOYD, 2014). For the prediction of seed quality, it is described as the most efficient method (VENKATESAN *et al*., 2020).

By means of the PLS-DA, it is also possible to identify and/or select variables in a discriminatory way (BRERETON; LLOYD, 2014). The results in this work pointed out that the most important spectral regions for the classification of the level of seed vigor, via PLS-DA, were the ranges of wavelengths between 1.350-1.450 nm, 1.800-1.900 nm and 2.300-2.400 nm for the seeds of high vigor; 1.000-1.200 nm, 1.350-1.450 nm and 2.300-2.400 nm for the seeds of intermediary vigor; and 1.350-1.450 nm; 1.800-1.900 nm and 2.300-2.400 nm for the seeds of low vigor (Figure 7).

It was observed that the ranges of the spectrum that comprise wavelengths of 1.350-1.450 nm and 2.300-2.400 nm were coincidently important for the classification of all the three classes (Figure 7). The range

between 1.350 -1.430 nm is represented by combinations of characteristic bands of C-H of the CH_2 molecule, and peaks 1.392 nm and 1.414 nm correspond to the bands of combination of C-H vibration, associated with lipids (HOURANT *et al*., 2000). Also, range 1.330-1.392 nm also represents bands associated with lipids (MUKASA *et al*., 2019). Gislum *et al*. (2018) describe peaks 1.379, 1.400 and 1.424 nm with a high correlation with oil content in seeds. On the other hand, Xu *et al*. (2020) also describe range 1.325-1.500 nm with peaks of high absorption of molecules associated with protein. Wavelengths 1.390 and 1.440 nm correspond to the second stretching overtone of C-H, and to the 1.420 nm stretching overtone of O-H. In addition, wavelengths 1.365 nm, 1.375 nm and 1.400 nm are described as important to differentiate the content of soluble protein in seeds.

Our results pointed out differences in the contents of oil and soluble protein among the different vigor levels of the seeds (Figure 2), and it is known that the aging process causes some changes in these compounds (BEWLEY *et al*., 2013). Thus, the wavelengths highlighted as the most important for the distinction of the vigor level of the seeds confirm the changes in these compounds.

The region of the electromagnetic spectrum that corresponds to the wavelength range of 2.230-2.500 nm is typical of the stretching combination and of other vibration modes of C-H (HOURANT *et al*., 2000). In the range between 2.280-2.330 nm there are combinations of stretching bands of C-H and of CH_2 deformation which might be related to carbohydrates, sugars and cellulose (WORKMAN; WEYER, 2007) and combination bands of O-H and C-O stretching, associated with polymers (KUSUMANINGRUM *et al*, 2018). Combination bands of stretching and of deformation of C-H are also associated with peptide groups associated with proteins in the range between 2.270-2.532 nm (WORKMAN; WEYER, 2007). Range 2.240-2.470 is characterized by combination bands of C-H of CH_2 and CH_3 molecules associated with oil and fatty acids (HOURANT *et al*., 2000; MUKASA *et al*., 2019). Furthermore, the absorption peaks at 1.450, 1.940 and 2.300 nm were described as important to demonstrate physicochemical changes in stored soybean seeds (BAZONI *et al*., 2017).

Therefore, although the contents of the different carbohydrates in the seeds, with both natural and artificial aging, were not quantified in this work, changes occur in these compounds (BEWLEY *et al*., 2013). During storage, sugars, especially reducing sugars, react with proteins and, consequently, inactivate vital enzymes for seed metabolism (CASTELLIÓN *et al*., 2010), which is connected to the reduction of the levels of soluble proteins observed in the stored seeds in this work (Figure 2) and to the changes in the NIR spectrum, which confirmed this process.

Considering only the classification of the freshly harvested and artificially aged seeds, range 1.800-1.900 nm also had a relative importance (Figure 7). The range between 1.880-1.930 nm is associated with combination bands of O-H vibration (HOURANT *et al*., 2000). There is also a combination of vibration of stretching of O-H and of the third overtone of C-O associated with cellulose with absorption at 1.820 nm (WORKMAN; WEYER, 2007). Absorption peak at 1.860 nm is reported by the combination of asymmetrical stretching of N-H with amide II, related to proteins (KUSUMANINGRUM *et al*., 2018). The range between 1.850-2.050 nm has little information about oil and fats (HOURANT *et al*., 2000).

For the class of stored seeds, the region between wavelengths 1.000-1.200 nm stood out (Figure 7). Peaks 972 and 1.009 nm correspond to the third O-H overtone associated with saccharides (WORKMAN; WEYER, 2007). The range between 1.090-1.180 nm corresponds to the second C-H overtone of the CH_2 molecule; between 1.100-1.200 nm, corresponds to the C-H second overtone of CH_3, and the range between 1.150-1.260 nm corresponds to the C-H second overtone of CH=CH (HOURANT *et al*., 2000). Wavelengths of 1.145 and 1.190 nm were described as lengths associated with the absorption of protein molecules (ARMSTRONG, 2006). Xu *et al*. (2020) describe the absorption at the wavelength of 990 nm due to the third overtone of O-H stretching; 1.020 nm as the second overtone of NH stretching; 1.130 and 1.165 nm associated with the third overtone of NH stretching, and wavelengths 980 nm, 1.180 nm, 1.190 nm, 1.230 nm and 1.235 nm were important to distinguish the content of soluble protein in soybean seeds. In addition, the peak at 1.210 nm and ranges close to 1.180 nm were described as associated with the C–H second overtone, related to fatty acids (HOURANT *et al*., 2000). Besides, oils rich in mono or poly-unsaturated fatty acids present greater absorption in the wavelength range of 1.164 nm than fats rich in saturated fatty acids (HOURANT *et al*., 2000).

The regions of the electromagnetic spectrum considered to be important for the classification of vigor in soybean seeds in this work (Figure 7) have already been described as important in the biochemical composition of different products. According to Ozaki, McClure and Christy (2006), an absorption peak at 1002 nm is associated with amorphous sucrose. The wavelength range between 1362-1480 nm is considered to be one of the most important for the calibration of models of mono-saturated, poly-unsaturated and saturated fatty acids (OZAKI; MCCLURE; CHRISTY, 2006). Absorption at the length of 1390 nm is described for lipids (AL-AMERY *et al*., 2018) and this same wavelength is

associated with proteins (XU *et al*., 2020). Absorption at 1400 and 1403 nm was reported as important to differentiate the viability in castor bean seeds (GISLUM *et al*., 2018) and wheat seeds (FAN; MA; WU, 2020), respectively. Absorption peak at 1396 nm was one of the most important wavelengths for the prediction model of content of protein in soybean (ARMSTRONG, 2006). Absorption at 1860 nm was described for molecules associated with proteins (KUSUMANINGRUM *et al*., 2018). The wavelength range around 2300 was important to evaluate the physical-chemical changes in stored soybean (BAZONI *et al*., 2017). The range between 2308–2348 nm is associated with lipids and wavelengths 2308 and 2346 nm represent ranges of important absorption for oil (OZAKI; MCCLURE; CHRISTY, 2006); absorption peak of 2.308 nm can be used to differentiate mono-unsaturated fatty acids from poly-unsaturated ones (HOURANT *et al*., 2000). As for peaks at 2306 and 2346 nm, they were described as important to distinguish soybean oil from other kinds of oil (OZAKI; MCCLURE; CHRISTY, 2006). Wavelength 2345 nm was also related to lipids and it was important to distinguish the viability of watermelon seeds (LOHUMI *et al*., 2013).

Therefore, the changes observed in the spectra of naturally and artificially aged seeds in this work (Figure 4) reflected the changes in the lipids and proteins in the seeds (Figure 2); with artificial aging, these changes were more drastic and led to a greater decrease in seed quality (Figure 1). Although artificial aging is often used as a predictor of natural aging, which occurs during seed storage, these two kinds of aging have their peculiarities (GAO *et al*, 2015). Natural aging usually occurs slowly and gradually, unlike artificial aging, which occurs in a fast and drastic way (GAO *et al*, 2015). During storage, the metabolic rates of the seeds are low and the predominant deterioration mechanisms are linked to the oxidation of fatty acids (SINGH; PAROHA; MISHRA, 2017) and to protein degradation by means of the Amadori and Maillard reactions (CASTELLIÓN *et al*., 2010). These reactions promote aggregation, loss of solubility and consequent reduction of the content of soluble proteins, which could be observed in this work (Figure 2). On the other hand, artificial aging leads to the increase of the metabolic rate of the seeds, due to their exposure to high temperatures and relative humidity (BEWLEY *et al*., 2013). This increase in seed metabolism promotes the formation of species reactive to oxygen, protein synthesis, including those which act in the process of reserve mobilization (BEWLEY *et al*., 2013; SINGH; PAROHA; MISHRA, 2017), affecting the size, uniformity and dry mass of the seedlings (Figure 1). Commonly, it also induces lipid peroxidation (BEWLEY *et al*., 2013), which leads to the increase of electrical conductivity and decrease of seed vigor, as it was verified in the results of this work (Figure 1).

Even though the identification of specific chemical compounds in seeds is very difficult due to the overlapping of spectral bands that might be associated with different compounds (KUSUMANINGRUM *et al*., 2018), it was pointed out, in this work, that there are regions of the electromagnetic spectrum which are more important to distinguish vigor level in soybean seeds determined by the natural and artificial aging of the seeds (Figure 7). It was also observed that the different kinds of aging promote deterioration at distinct levels which reflect in the quality of the seeds (Figure 1). In addition, it was possible to demonstrate the relationship between the content of soluble protein, and content of oil and its fractions of fatty acids with the quality of the seeds (Figures 1 and 2) and verify the relationships between the content of oil and protein with the most important wavelengths for the classification of the vigor levels of naturally and artificially aged soybean seeds.

CONCLUSIONS

1. The regions of the electromagnetic spectrum between wavelengths of 1000-1200 nm; 1350-1450 nm; 1800-1900 nm and 2300-2400 nm are important to distinguish the level of quality of the seeds;
2. The contents of oil and protein have a relationship with the physiological quality of the seeds and the most relevant wavelengths for the classification of seed vigor also presented a relationship with these compounds;
3. NIR spectroscopy, in combination with chemometric methods, has a potential to classify soybean seeds according to their vigor.

ACKNOWLEDGEMENTS

We would like to thank the Research Support Foundation of the State of Minas Gerais (FAPEMIG), the Coordination for the Improvement of Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPq) for the financial support.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

REFERENCES

AGELET, L. E.; HURBURGH, C. R. Limitations and current applications of Near Infrared Spectroscopy for single seed analysis. **Talanta**, v. 212, p. 288-299, 2014.

AL-AMERY, M. *et al*. Near-infrared spectroscopy used to predict soybean seed germination and vigour. **Seed Science Research**, v. 28, p. 245-252, 2018.

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

BARKER, M.; RAYENS, W. Partial least squares for discrimination. **Journal of Chemometrics**, v. 17, p. 166-173, 2003.

BAZONI, C. H. *et al.* Near-infrared spectroscopy as a rapid method for evaluation physicochemical changes of stored soybeans. **Journal of Stored Products Research**, v. 73, p. 1-6, 2017.

BEWLEY, J. D. *et al*. **Seeds**: physiology of development, germination and dormancy. 3. ed. New York: Springer, 2013. 407 p.

BRADFORD, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. **Analytical Biochemistry**, v. 72, p. 248-254, 1976.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Secretaria de Defesa Agropecuária. **Regras para análise de sementes**. Brasília: Mapa/ACS, 2009. 395 p.

BRERETON, R. G.; LLOYD, G. R. Partial least squares discriminant analysis: taking the magic away. **Journal of Chemometrics**, v. 28, p. 213-225, 2014.

CASTELLIÓN, M. *et al*. Protein deterioration and longevity of quinoa seeds during long-term storage. **Food Chemistry**, v. 121, p. 952-958, 2010.

EBONE, L. A. *et al.* Soybean seed vigor: uniformity and growth as key factors to improve yield. **Agronomy**, v. 10, n. 4, 545, p. 1-15, 2020.

FAN, Y.; MA, S.; WU, T. Individual wheat kernels vigor assessment based on nir spectroscopy coupled with machine learning methodologies. **Infrared Physics and Technology**, v. 105, p. 1-7, 2020.

GAO, H. W. *et al*. Differences in properties of seed vigor between artificially and naturally aged soybean seeds. **Seed**, v. 34, n. 1, p. 14-9, 2015.

GISLUM, R. *et al*. Characterization of castor (*Ricinus communis* L.) seed quality using Fourier transform near-infrared spectroscopy in combination with multivariate data analysis. **Agriculture**, v. 59, n. 8, p. 1-10, 2018.

GUO, J. *et al*. NIR calibrations for soybean seeds and soy food composition analysis total carbohydrates, oil, proteins and water contents. **Nature Precedings**, v. 6, p. 1-40, 2011.

HAYATI, P. K. D.; ANGGASTA, G. N.; ANWAR, A. Physical and chemical properties of dura and pisifera genotypes of oil palm seed and its viability and vigor. **International Conference of Bio-Based Economy and Agricultural Utilization**, v. 497, p. 1-8, 2020.

HOURANT, P. *et al*. Oil and fat classification by selected bands of near-infrared spectroscopy. **Applied Spectroscopy**, v. 54, p. 1168-1174, 2000.

HUANG, J.; YU, C. Determination of cellulose, hemicellulose and lignin content using near-infrared spectroscopy in flax fiber. **Textile Research Journal**, v. 89, p. 1-9, 2019.

HUANG, Z. *et al*. Feasibility study of near infrared spectroscopy with variable selection for non-destructive determination of quality parameters in shell-intact cottonseed. **Industrial Crops and Products**, v. 43, p. 654-660, 2013.

JHAM, G. N.; TELES, F. F. F.; CAMPOS, L. G. Use of aqueous HCl/MeOH as esterification reagent for analysis of fatty acids derived from soybean lipids. **Journal of the American Oil Chemists' Society**, v. 59, p. 32-133, 1982.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A**, v. 374, 2065, p. 1-16, 2016.

KUSUMANINGRUM, D. *et al*. Non-destructive technique for determining the viability of soybean (*Glycine max*) seeds using NIR spectroscopy. **Journal of the Science of Food and Agriculture**, v. 98, p. 1734-1742, 2018.

LI, X. *et al*. Review of NIR spectroscopy methods for nondestructive quality analysis of oilseeds and edible oils. **Trends in Food Science & Technology**, v. 101, p. 172-181, 2020.

LOHUMI, S. *et al*. Nondestructive evaluation for the viability of watermelon (*Citrullus lanatus*) seeds using fourier transform near infrared spectroscopy. **Journal of Biosystems Engineering**, v. 38, n. 4, p. 312-317, 2013.

MAGUIRE, J. D. Speed of germination-aid selection and evaluation for seedling emergence and vigor. **Crop Science**, v. 2, p. 176-177, 1962.

MATHIAS, V.; COELHO, C. M. M.; GARCIA, J. Soluble protein as indicative of physiological quality of soybean seeds. **Revista Caatinga**, v. 32, n. 3, p. 730-740, 2019.

MAYRINCK, L. G., *et al.* Use of near infrared spectroscopy in cotton seeds physiological quality evaluation. **Journal of Seed Science**, v. 42, p. 1-11, 2020.

MEDEIROS, A. D.; PEREIRA, M. D. SAPL®: a free software for determining the physiological potential in soybean seeds. **Pesquisa Agropecuária Tropical**, v. 48, n. 3, p. 222-228, 2018.

MUKASA, P. *et al.* Determination of viability of retinispora (*Hinoki cypress*) seeds using FT-NIR spectroscopy. **Infrared Physics and Technology**, v. 98, p. 62-68, 2019.

OZAKI, Y.; MCCLURE, W. F.; CHRISTY, A. A. **Near-infrared spectroscopy in food science and technology**. New Jersey: Wiley-Interscience, 2006. 406 p.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry**, v. 36, n. 8, p. 1627-1639, 1964.

SILVA, L. J.; MEDEIROS, A. D.; OLIVEIRA, A. M. S. Seedcalc, a new automated R software tool for germination and seedling length data processing. **Journal of Seed Science**, v. 41, p. 250-257, 2019.

SINGH, J.; PAROHA, S.; MISHRA, R. P. Factors affecting oilseed quality during storage with special reference to soybean (*Glycine max*) and niger (*Guizotia abyssinica*) seeds. **International Journal of Current Microbiology and Applied Sciences**, v. 6, n. 10, p. 2215-2226, 2017.

VENKATESAN, S. *et al*. Role of near - Infrared spectroscopy in seed quality evaluation: a review. **Agricultural Reviews**, v. 41, p. 106-115, 2020.

WORKMAN, J. JR.; WEYER, L. **Practical guide to interpretive near-infrared spectroscopy**. Boca Raton: CRC Press: Taylor & Francis Group, 2007. 346 p.

XU, R. *et al.* Use of near-infrared spectroscopy for the rapid evaluation of soybean [*Glycine max* (L.) Merril.] water soluble protein content. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 224, 117400, p. 1-8, 2020.

YASMIN, J. *et al*. Classification method for viability screening of naturally aged watermelon seeds using NIR spectroscopy. **Sensors**, v. 19, n. 5, p. 1-14, 2019.

ZHANG, T. *et al.* Non-destructive analysis of germination percentage, germination energy and simple vigour index on wheat seeds during storage by Vis/NIR and SWIR hyperspectral imaging. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 239, 118488, p. 1-11, 2020.