Near Infra-Red Spectroscopy Predicts Crude Protein in Hemp Grain

Ryan Crawford[1], Jamie Crawford[1], Lawrence B. Smart[2], Virginia Moore[3]

[1]Cornell University, Ithaca, NY,
[2]Cornell AgriTech, Geneva, NY,
[3]Cornell University, Ithaca, NY,

Corresponding author: Ryan Crawford, rvc3@cornell.edu

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Plain Language Summary

Earthquake data for the island of La Palma from the September 2021 eruption is found …

incomplete: may contain errors, run-ons, half-thoughts, etc.

0.1 INTRODUCTION

Hemp (Cannabis sativa L.) is an annual crop with potential uses as a source of food or feed from grain, and bast fiber or hurd from the stalk. Hemp cultivars are commonly grown for one or both purposes and a cultivar may be referred to as a grain, fiber, or dual-purpose type. Because of protein's nutritional importance, the protein content of a grain crop is an prime consideration for researchers, producers, and consumers. Whole hemp grain typically contains approximately 20-30% protein (Bárta et al., 2024; Callaway, 2004; Ely & Fike, 2022). Crude protein (CP) is often used as a proxy for the direct measurement of protein concentration and consists of the multiplication of nitrogen concentration by a conversion factor because measuring nitrogen concentration is relatively simple (Hayes, 2020).

Near-infrared spectroscopy (NIRS) technology is rapid, non-destructive, and cheap. It consists of the measurement of NIR radiation reflected and absorbed from a sample (the spectra) and the relation of the spectra to laboratory values for components such as moisture, protein, fat, or fiber (Roberts et al., 2004). NIRS technology has been used since the 1970's to assess forage CP (Reeves, 2012; Williams, 1975). A NIRS calibration set often consists of samples from one species grown in many environments encompassing the range of expected values from the analyte or analytes (Chadalavada et al., 2022). Partial least squares regression (PLSR) is a typical method used in the agricultural and food sciences to relate spectra to analyte (Roberts et al., 2004). PLSR calculates components that maximize covariance between predictor and response variables. PLSR uses some number of components, often selected via cross-validation, in order to fit the regression model and is commonly used in spectroscopy because it tends to work well with highly-correlated, noisy spectral data (Wold et al., 2001).

A NIRS-scanned sample of undamaged grain may used for other purposes besides its scan or it may planted as a seed. In wheat and corn, grain protein content has been shown to be heritable (Geyer et al., 2022; Giancaspro et al., 2019). This suggests (at least potentially) that NIRS technology could serve as resource to rapidly identify high CP hemp germplasm, enabling the screening of more germplam as grain, before planting to the field, and thus enabling the more efficient development of high CP hemp grain cultivars.

For this study, a benchtop NIR spectrometer was used to develop a model to predict CP content based on a data set of hemp grain representing multiple years, locations, and cultivars from grain and dual-purpose hemp types using PLSR.

0.2 MATERIALS AND METHODS

Source: Article Notebook

Source: Article Notebook

0.2.1 Hemp Grain Sample Background

Spectral data were obtained from whole (unground) hemp grain samples, harvested at maturity, collected from 2017 - 2021 from 18 cultivar trials in New York (NY) (NA samples). Grain samples were obtained by hand sampling or mechanical harvest and were cleaned of chaff and dried at 30 C for six days in a forced-air dryer. In total, 38 cultivars were represented in the data set. Cultivars were grain or dual-purpose types and included both commercially available and experimental material.

All cultivar trials were planted in randomized complete block design with each cultivar replicated four times. The 2017 data were comprised of samples from the same thirteen cultivars sampled from six NY locations. For those trials, grain was harvested from each plot individually and aggregated by cultivar within each trial. Four subsamples were drawn from each aggregated sample and scanned separately. These spectra were averaged at each 2 nm increment. All remaining samples from 2018-2021 were collected on a per-plot basis. All possible cultivars and possible locations were represented in 2017, but only a selected subset of cultivars and locations were represented in 2018-2021.

0.2.2 Spectral Data Collection and Preprocessing

A benchtop NIR spectrometer (FOSS/ NIR FOSS/ NIR Systems model 5000) was used to obtain the spectra (FOSS North America, Eden Prairie, MN, USA). Spectra were collected every 2 nm from 1100-2498 nm and the logarithm of reciprocal reflectance was recorded. A 1/4 rectangular sample cup (5.7 cm$\times$ 4.6 cm) was used.

WINISI software version 1.02A (Infrasoft International, Port Matilda, PA, USA) was used to average the spectra in 2017, as well as to select samples for laboratory assay. Samples were selected according to their spectral distance from their nearest neighbor within the calibration data set with a cutoff of a distance of 0.6 H, where H is approximately equal to the squared Mahalanobis distance divided by the number of principal components used in the calculation (Garrido-Varo et al., 2019). Prior to selection selection, spectra were preprocessed using SNV-detrend with settings 1,4,4,1 for the derivative, gap, smooth, and smooth 2 settings respectively.

0.2.3 Laboratory Validation

Laboratory assays were performed by Dairy One Forage Laboratory (Ithaca, NY). For those assays, 1mm ground samples were analyzed by combustion using a CN628 or CN928 Carbon/Nitrogen Determinator. Samples from 2017 were aggregated as described above, but the remaining samples were not aggregated.

0.2.4 Model Development

Calibration and validations sets were created by dividing the laboratory CP values into tertiles according to their percent CP in order to ensure that the range of CP values was present in both calibration and testing sets. Within each tertile, 75% of the samples were randomly assigned to the calibration set and the remaining 25% were assigned to the testing set. For each calibration set, models were developed in the caret package using PLSR. In fitting the model, the number of components was optimized over a grid search from 1-20. Model performance was evaluated with 25 iterations of bootstrapping and minimized root mean squared error (RMSE) in selecting the number of components in the final model .

Source: Article Notebook

Initially a number of common spectral preprocessing methods were tested by creating 100 calibration and testing sets, as described above. Spectral data from those data sets were transformed by each of the following methods: 1) first derivative, 2) Savitzky-Golay (SG) using the first derivative, third order polynomial, and a window of size 5, 3) gap-segment derivative using the first derivative, a gap of eleven, and a segment size of 5, 4) standard normal variate (SNV), 5) standard normal variate fol-

lowing Savitzky-Golay (SNV-SG) (same SG parameters as above), 6) SNV-detrend with second order polynomial, and 7) multiplicative scatter correction. As a comparison, models were also developed using untransformed spectra.

For each of these preprocessing methods, models were fit and predictions were made on the corresponding validation set (since there were 8 preprocessing methods, 8 separate models were fit for each of the 100 sets. The relationship between the predicted and actual values of the testing set were calculated via RMSE, R^2 and Ratio of Performance to InterQuartile distance (RPIQ), three common model assessment metrics. Larger R^2 and RPIQ, and smaller RMSE values are superior. Analyses of variance (ANOVA) were performed for each of these metrics in order to compare the preprocessing methods. For each ANOVA, each data set was considered as a subject and different variances were allowed for each preprocessing method.

Once the most promising preprocessing method was identified, 1000 more data sets were created and analyzed via that method and performance on the testing sets was summarized with RMSE, R^2, and RPIQ.

0.2.5 Additional software used

We used R version 4.3.3 (R Core Team, 2024) and the following R packages: caret v. 6.0.94 (Kuhn & Max, 2008), data.table v. 1.15.2 (Barrett et al., 2024), emmeans v. 1.10.0 (Lenth, 2024), nlme v. 3.1.163 (J. Pinheiro et al., 2023; J. C. Pinheiro & Bates, 2000), pls v. 2.8.3 (Liland et al., 2023), prospectr v. 0.2.7 (Stevens & Ramirez-Lopez, 2024), skimr v. 2.1.5 (Waring et al., 2022), tidymodels v. 1.1.1 (Kuhn & Wickham, 2020), tidyverse v. 2.0.0 (Wickham et al., 2019).

Source: Article Notebook

0.3 RESULTS AND DISCUSSION

0.3.1 Laboratory assay CP values

Laboratory assay percent CP values are summarized in the following table. These are similar to the range of CP values observed in the literature, indicating an reasonable basis for a chemometric model. The CP values are left-skewed and two thirds of the samples contained more than 25% CP.

Table 1: Summary of Laboratory Assayed CP Values (Percent Dry Matter)

Mean	Sd	Minimum	First Quartile	Median	Third Quartile	Maximum
26.1	2.5	20.8	23.9	26.4	28.2	30.8

Source: Article Notebook

0.3.2 Preprocessing methods comparison

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

All preprocessing methods outperformed untransformed spectral data Table **??**. Averaged together, all preprocessed spectra were superior to untransformed spectra, with lower RMSE, and higher R^2 and RPIQ values (significant at α level <0.001).