Agrosystems, Geosciences & Environment OPEN ACCESS ASA | CSSA

Near Infrared Spectroscopy Predicts Crude Protein Concentration in Hemp Grain

Journal:	*Agrosystems, Geosciences & Environment*
Manuscript ID	AGE-2024-10-0358-OA
Wiley - Manuscript type:	Original Article
Date Submitted by the Author:	31-Oct-2024
Complete List of Authors:	Crawford, Ryan; Cornell University Crawford, Jamie; Cornell University Hansen, Julie; Cornell University Smart, Larry; Cornell University College of Agriculture and Life Sciences, Horticulture Moore, Virginia; Cornell University
Keywords:	Hemp, Grain, Spectroscopy

Core Ideas

As part of the submission process, we ask authors to prepare highlights of their article. The highlights will consist of 3 to 5 bullet points that convey the core findings of the article and emphasize the novel aspects and impacts of the research on scientific progress and environmental problem solving.

The purpose of these highlights is to give a concise summary that will be helpful in assessing the suitability of the manuscript for publication in the journal and for selecting appropriate reviewers. If the article is accepted the highlights may also be used for promoting and publicizing the research.

Core Idea 1: Models were developed to predict crude protein concentration in hemp grain using near infrared spectroscopy.

Core Idea 2: Most models were able to distinguish between high and lower concentrations of crude protein.

Core Idea 3: Models could be further optimized by including more samples and rectifying class imbalances between environments.

Core Idea 4: CUST_CORE_IDEA_4 :No data available.

Core Idea 5: CUST_CORE_IDEA_5 :No data available.

Plain Language Summary

As part of the submission process, we ask authors to prepare a plain language summary (1000-character limit). A ninth-grade reading comprehension level is suggested for ease of reading. We suggest you structure your plain language summary with four key elements: subject overview, research purpose, key results, and key takeaways. See detailed instructions here.

The purpose of the plain language summary is to explain the article's research and relevance in clear language free from jargon. If the article is accepted, the summary will appear with the abstract and may be used to further promote your article.

The protein concentration of hemp grain is important for researchers, producers, and consumers. This study was conducted to determine whether hemp grain can be non-destructively tested for protein concentration. Hemp grain samples were obtained from research trials conducted in New York from 2017-2021. The samples were scanned with a near-infrared spectroscopy machine and then the samples were ground and analyzed by a commercial laboratory. Many models were developed to compare the results of the scans with those obtained from the laboratory. Approximately 75 % of those models had, at minimum, the ability to distinguish between high and low protein concentration. This may be useful to plant breeders developing high and low protein concentration plant populations.

Models were developed to predict crude protein concentration in hemp grain using near infrared spectroscopy.

Most models were able to distinguish between high and lower concentrations of crude protein.

Models could be further optimized by including more samples and rectifying class imbalances between environments.

Near Infrared Spectroscopy Predicts Crude Protein Concentration in Hemp Grain

Ryan V. Crawford, Jamie L. Crawford, Julie L. Hansen, Lawrence B. Smart,Virginia M. Moore

Ryan V. Crawford, Jamie L. Crawford, and Julie L. Hansen, Cornell University, 126 Medicago Drive, Ithaca NY, USA

Lawrence B. Smart, Cornell AgriTech, 102 Hedrick Hall, Geneva, NY, USA

Virginia M. Moore, Cornell University, 162 Emerson Hall, Ithaca, NY USA

Correspondence: Ryan V. Crawford,126 Medicago Drive, Ithaca NY 14853, USA. Email: rvc3@cornell.edu

Abbreviations: CP, crude protein; NIR, near-infrared; NIRS, near-infrared spectroscopy; NY, New York; PLSR, partial least squares regression; RPD, relative predicted deviation, RPIQ, ratio of performance to interquartile distance; SG, Savitzky-Golay; SNV, standard normal variate, SNV-SG, standard normal variate following Savitzky-Golay

ABSTRACT

This study was conducted to determine whether hemp grain can be non-destructively assayed for crude protein (CP) concentration using spectra obtained from near-infrared spectroscopy (NIRS) to build a prediction model for CP concentration using partial least squares regression (PLSR). Hemp grain samples were obtained from cultivar trials in New York (NY) from 2017-2021. The samples' NIRS spectra were collected and the samples were assayed for validation by a commercial laboratory. Seven potential preprocessing methods, as well as untransformed spectra, were tested on 100 training/ testing splits of the data set and the best method was selected. A preprocessing method consisting of the standard normal variate transformation following a Savitzky-Golay filter had the lowest RMSE and the highest R^2, RPD and RPIQ, with RPD and RPIQ. That method was applied to 1000 additional splits of the data set and predictive performance on the testing sets was examined. Optimal final models typically consisted of 12 components. Seventy-four percent of the final models had the ability to distinguish between high and low values of CP concentration and 49% of the models were capable of approximating quantitative prediction. The worst-predicted samples tended to come from Geneva, NY, possibly as a result of the models' class imbalance (half of the samples were from Ithaca, NY while 28% were from Geneva). The research shows the promise that NIRS offers in the non-desctructive assay of CP concentration in hemp grain.

1 INTRODUCTION

Hemp (*Cannabis sativa* L.) is an annual crop with potential uses as a source of food or feed, derived from the grain, and fiber (bast or hurd), derived from the stalk. Hemp cultivars are commonly grown for one or both purposes and a cultivar may be called a grain, fiber, or dual-purpose type. Because of its nutritional importance, the protein concentration of a grain crop is a prime consideration for researchers, producers, and consumers. Whole hemp grain typically contains approximately 200-300 g kg^{-1} protein (Bárta et al., 2024; Callaway, 2004; Ely & Fike, 2022; Liu et al., 2023). 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, often 6.25 (Hayes, 2020).

Near-infrared (NIR) spectroscopy (NIRS) technology is rapid, non-destructive, and inexpensive. It consists of the measurement of NIR radiation reflected and absorbed from a sample (the spectra) and the relation of the spectra to primary analytical values, typically obtained using wet chemistry assays, 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 concentration (Reeves, 2012; Williams, 1975). A NIRS calibration set often consists of samples from diverse genotypes of 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). Partial least squares regression calculates components that maximize covariance between predictor and response variables. Partial least squares regression uses some number of components, often selected via cross-validation, 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 whole grain may be used for other purposes besides the scan, including planting 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 that NIRS technology could serve as a resource to rapidly identify high concentration CP hemp germplasm, enabling the screening of germplasm as seed, before planting to the field, and facilitating the efficient development of high concentration CP hemp populations.

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

2 MATERIALS AND METHODS

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) (149 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. All CP values were expressed as concentration dry matter. In total, 149 samples from 38 cultivars were represented in the data set. Cultivars were grain or dual-purpose types and included both commercially available and experimental material. Seventy-eight samples were scanned and assayed in 2017, 19 in 2018, 24 in 2019, and 28 in 2021. More information about hemp cultivars and locations is available in Supplemental Table S1.

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 13 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 cultivars and locations were represented in 2017, but only a selected subset of cultivar-location combinations were represented in 2018-2021 because not all cultivars were planted everywhere and only a portion of these cultivar-location combinations were sampled, scanned, and assayed due to logistical constraints.

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 × 4.6 cm) was used to scan the samples.

WINISI software version 1.02A (Infrasoft International, Port Matilda, PA, USA) was used to calculate the mean spectra in 2017 and to select samples for laboratory assay in all years. Samples were selected according to their spectral distance from their nearest neighbor within the 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, spectra were preprocessed using SNV (standard normal variate) -detrend with settings 1,4,4,1 for the derivative, gap, smooth, and smooth-two settings respectively.

2.3 Laboratory Validation

Laboratory assays were performed by Dairy One Forage Laboratory (Ithaca, NY). For those assays, 1 mm 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.

2.4 R software and packages used

We used R version 4.4.1 (R Core Team, 2024) and the following R packages: caret v. 6.0.90 (Kuhn, 2021), data.table v. 1.16.0 (Barrett et al., 2024), emmeans v. 1.10.4 (Lenth, 2024), nlme v. 3.1.165 (J. Pinheiro et al., 2024; J. C. Pinheiro & Bates, 2000), pls v. 2.8.0 (Liland et al., 2021), prospectr v. 0.2.7 (Stevens & Ramirez-Lopez, 2024), skimr v. 2.1.5 (Waring et al., 2022), tidymodels v. 1.2.0 (Kuhn & Wickham, 2020), tidyverse v. 2.0.0 (Wickham et al., 2019).

2.5 Model Development

Training and testing sets were created by dividing samples by their laboratory CP concentration values into tertiles to ensure that a representative range of values was present in both training and testing sets. Within each tertile, 75% of the samples were randomly assigned to the training set and the remaining 25% were assigned to the testing set. For each training 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 RMSE in selecting the number of components in the final model.

Initially a number of common spectral preprocessing methods were tested by creating 100 training and testing sets, as described above. Spectral data 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 five; 3) gap-segment derivative using the first derivative, a gap of 11, and a segment size of five; 4) SNV; 5) standard normal variate following Savitzky-Golay (SNV-SG) using the same SG parameters as above; 6) SNV-detrend with second order polynomial; and 7) multiplicative scatter correction. For comparison, models were also developed using untransformed spectra.

For each of these preprocessing methods, models were fit and predictions were made on the corresponding testing set. Since there were seven preprocessing methods as well as untransformed spectra, eight separate models were fit for each of the 100 sets. The relationship between the predicted and actual values of the test set were calculated via RMSE, R^2, relative predicted deviation (RPD), and ratio of performance to interquartile distance (RPIQ), four common model assessment metrics. Larger R^2, RPD and RPIQ values and smaller RMSE values are best. The answer to the question of exactly which values constitute a "good" model varies depending upon the reference consulted, but for the sake of simplicity, the standard established for an acceptable model was $R^2 > 0.80$, an RPD greater than 2.5 and ideally greater than 3 ("good" to "excellent" quantitative prediction), and an RPIQ greater than 2.3 but ideally greater than 4.1 prediction on the testing set (Chadalavada et al., 2022; Luce et al., 2017; Rawal et al., 2024).

Analyses of variance were performed for each of these metrics in order to compare 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 training and testing sets were created, and models were developed with that method. Performance on the testing sets was summarized with RMSE, R^2, RPD, and RPIQ. The pattern of errors, expressed as the difference between the actual and predicted values for a given sample, was examined.

3 RESULTS AND DISCUSSION

3.1 Laboratory assay CP values

Laboratory assay CP concentration values are summarized in Table1. These are similar to the range of values observed in the literature, indicating an reasonable basis for a chemometric model. The values were left-skewed (skewness of -0.29) and two thirds of the samples contained more than 250 g kg $^{-1}$ CP.

Table 1. Summary of Laboratory Assayed CP Values.

Mean	Sd	Minimum	First Quartile	Median	Third Quartile	Maximum
g kg $^{-1}$						
261	25	208	239	264	282	308

3.2 Preprocessing methods comparison

All preprocessing methods outperformed untransformed spectral data, as shown in Table 2. Averaged together, all preprocessed spectra were superior to untransformed spectra, with lower RMSE and higher R^2, RPD and RPIQ values (significant at α level <0.001). Preprocessing methods had 11.6 % lower RMSE and had 3.1% higher R^2, 6.3% higher RPD and 7.4% higher RPIQ than unprocessed spectra. Preprocessed spectra also had lower standard errors than untransformed spectra.

The SNV-SG method had the lowest RMSE and the highest R^2, RPD and RPIQ averaging over all iterations. SNV-SG's RMSE was 1.4% lower than the next best preprocessing method (SG), while SNV-SG's R^2, RPD, and RPIQ were 0.4%, 2.1%, and 2.4% higher than SG respectively. However, the differences between the best and second-best methods by metric were only statistically significant at α < 0.05 for RPD and RPIQ. There is a long history of using RPD to evaluate chemometric models although the statistic has been criticized as inadequately reflecting the distribution of skewed populations, a situation which RPIQ was designed to address (Bellon-Maurel et al., 2010). In this study, the data were somewhat but not heavily skewed and RPD and RPIQ metrics agreed. The superiority of SNV-SG by these metrics made it the best choice for the final model.

Table 2. Evaluation of Preprocessing Methods by Metric ± Standard Error.

Preprocessing Method	RMSE	R^2	RPD	RPIQ
Standard Normal Variate following Savitzky-Golay	1.02 ± 0.012	0.84 ± 0.004	2.49 ± 0.032	3.97 ± 0.076
Savitzky-Golay	1.03 ± 0.012	0.83 ± 0.004	2.44 ± 0.029	3.88 ± 0.072
First Derivative	1.07 ± 0.013	0.82 ± 0.004	2.36 ± 0.032	3.77 ± 0.075
Standard Normal Variate	1.12 ± 0.016	0.80 ± 0.005	2.26 ± 0.036	3.61 ± 0.081
Gap-segment Derivative	1.12 ± 0.018	0.81 ± 0.006	2.26 ± 0.040	3.60 ± 0.086
Standard Normal Variate-Detrend	1.13 ± 0.015	0.80 ± 0.005	2.22 ± 0.035	3.55 ± 0.079
Multiplicative Scatter Correction	1.17 ± 0.016	0.79 ± 0.006	2.17 ± 0.035	3.47 ± 0.080
Untransformed Spectra	1.22 ± 0.044	0.79 ± 0.009	2.17 ± 0.052	3.42 ± 0.105

From the literature, these results are readily explained. Standard normal variate and SNV-detrend both correct light scatter, which is often a function of differences in particle size and sample packing density, although SNV-detrend is often used for densely-packed, powdered samples (Barnes et al., 1989). SG is a smoothing filter that regresses on the signal over a series of windows, removing noise while preserving the signal's shape and features (Li et al., 2020; Luo et al., 2005). Derivatives, here including SG, gap-segment, and first derivatives pretreatments may remove additive and multiplicative effects, but not necessarily light scatter; as well, derivatives may increase spectral noise (Rinnan et al., 2009). Here, hemp grain was neither powdered nor densely packed but samples were subject to light scatter and noise due to differences in particle size in the hemp grain.

3.3 Final model development and summary

The model improved most rapidly as the number of components increased from one to seven, with the inclusion of each additional component being associated with a decrease in RMSE of 5%-12%. From eight to 12 components, model performance continued to improve, although gains were more modest: there was a decrease in RMSE of 0.7%-3% with the inclusion of each additional component. With 13 or more components, performance gains were minimal and the relative ranks of the models tended to be stable (Figure 1).

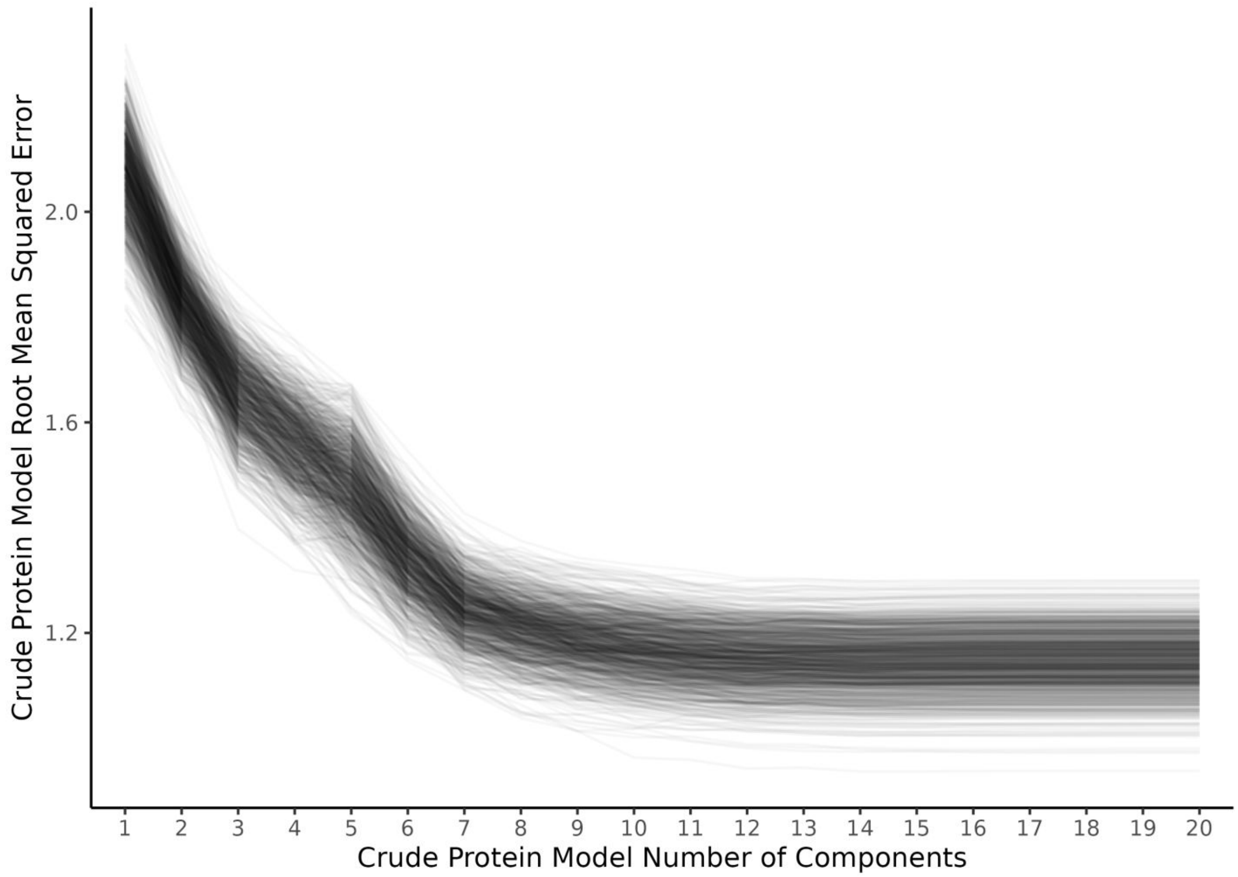

Figure 1. Decreasing RMSE with increasing number of components for 1000 training sets.

The performance of the final models on the test sets were similar, but not identical to, those obtained during the initial comparison of preprocessing methods. The means of the final models were: RMSE = 1.03, R^2 = 0.83, RPD = 2.44, and RPIQ = 3.89. Five percent of the models were "excellent" for quantitative prediction by both metrics, with RPD > 3 and RPIQ > 4.1, while an additional 11% of the models were "good" by both metrics (RPD range from 2.5–3.0, RPIQ range from 2.3–4.1). Forty-nine percent of the models had the ability to approximate quantitative prediction (RPD range from 2.0–2.5), and nine percent of the models were able to distinguish between high and low concentration CP values (RPD range from 1.5–2.0). Therefore, 74% of the models had, at minimum, the ability to distinguish between high and low CP concentration values with 65% having, at minimum, the ability to approximate quantitative prediction. Despite the generally good model performance, a subset of poor models can be seen. For example, Figure 2 shows 21 models with R^2 below 0.7.

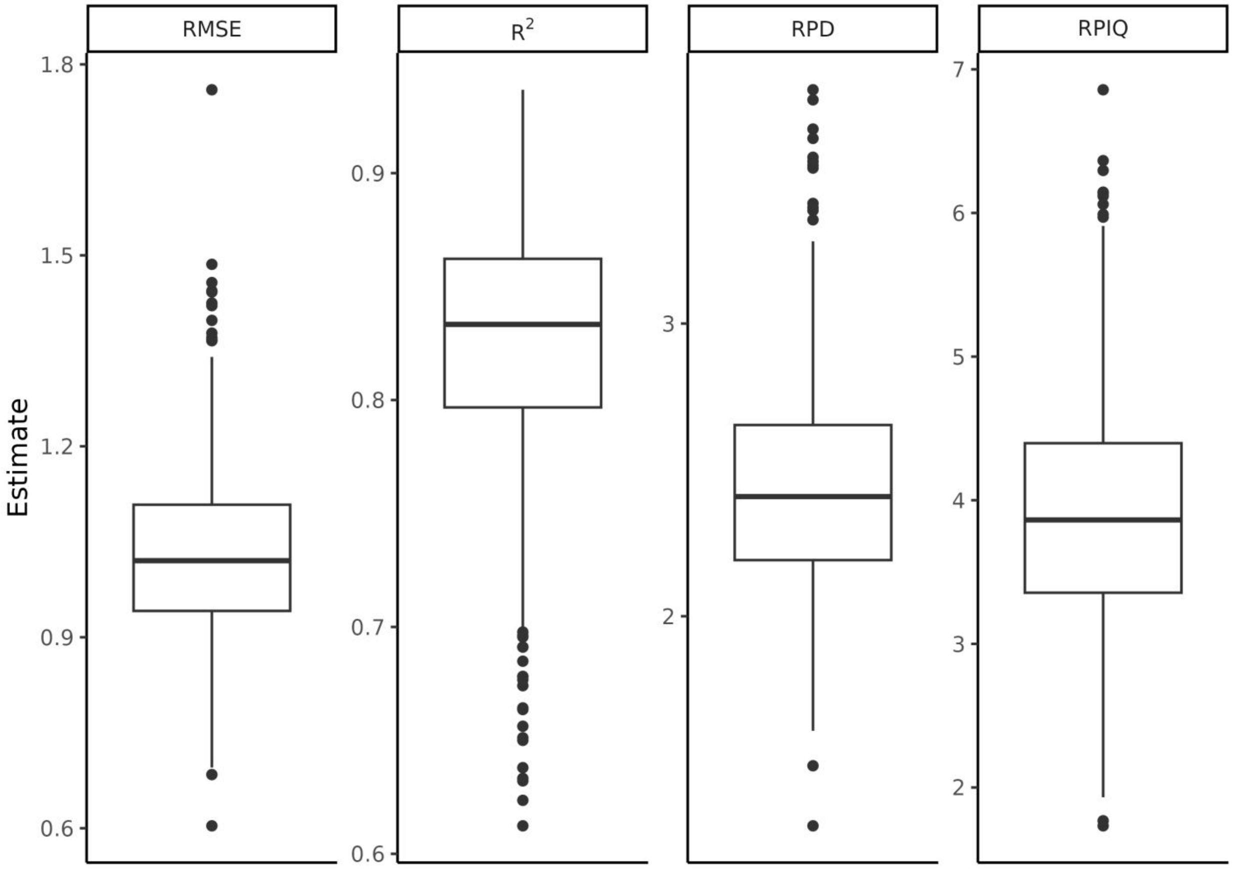

Figure 2. Final model testing set performance over 1000 iterations.

Finally, the pattern of test set errors was examined on a per-sample basis by calculating the difference between the actual and predicted values for the samples in the test sets Figure 3. A linear model was fit considering the mean estimated error for each sample where that sample was in the test set as compared to the sample's actual value. The models overestimated CP concentration by approximately 0.5% in the lowest tertile and underestimated percentage CP concentration by -0.01% and -0.41% in the middle and highest tertile, respectively. The variance of the errors did not increase appreciably as CP concentration increased.

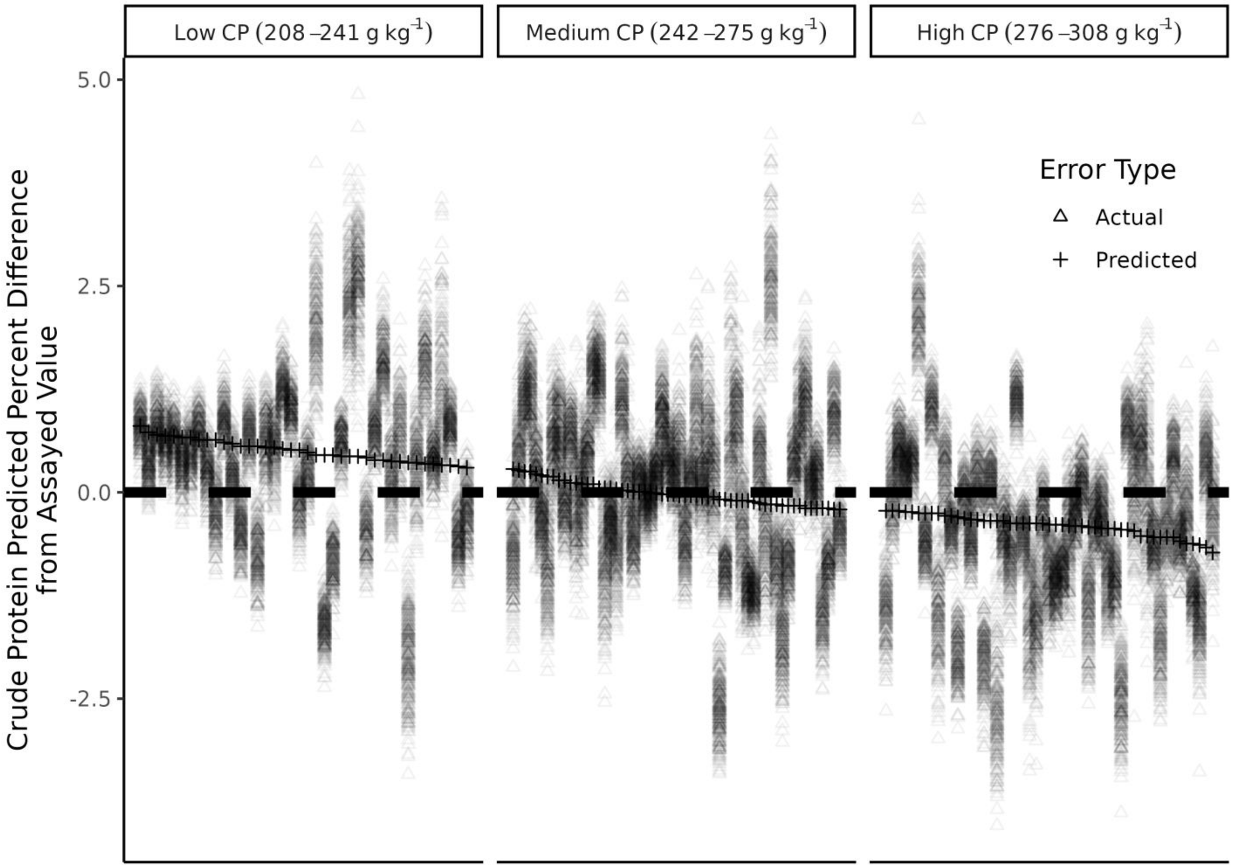

Figure 3. Testing set prediction errors on a per-sample basis. Actual sample value set to zero and samples ranked from least to greatest actual CP concentration value.

The 15 (10%) best and 15 worst predicted samples as measured by the mean absolute error of prediction were identified and their backgrounds examined. Overall, half of the samples in the data set came from Ithaca, NY ("Ithaca"), while 28% were collected from Geneva, NY ("Geneva") Table 3. However, of the 15 worst-predicted samples, nine were from Geneva, while three of the 15 best-predicted samples were from Geneva (by contrast, seven of the best-predicted and five of the worst-predicted samples came from Ithaca). Overall, samples from Geneva had the highest mean absolute error of prediction among locations, 61% greater than samples from Ithaca and 155% greater than samples from Freeville, NY (the only locations where more than 20 samples were assayed).

This study is limited in that it represents the creation of one model based upon spectra collected from one machine. NIRS calibrations can be unique to a particular machine, even if the machines compared are of the same model (Reeves, 2012). As well, the testing and training sets are relatively small.

This research showed the promise of the use of NIRS in order to make predictions concerning CP concentration in hemp grain using PLS. Promising preprocessing methods were identified and a model was validated. Further research could refine the model by including more samples, particularly by rectifying the class imbalance between Geneva and Ithaca, identifying promising spectral regions, or by examining other predictive methods.

ACKNOWLEDGMENTS

This work would not have been possible without the efforts of the field staff, undergraduate, and graduate students who planted, maintained, monitored and harvested these trials. Funding was provided by New York State through a grant from Empire State Development (AC477). We are grateful to those who provided seed for this project, including: Uniseeds, Verve Seeds, Winterfox Farms, International Hemp, Fiacre Seeds, and KonopiUS.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

name: Ryan V. Crawford; orcid: 0009-0006-3052-3269

name: Jamie L. Crawford; orcid: 0009-0002-2523-3479

name: Julie L. Hansen; orcid: 0000-0001-7247-9186 name: Lawrence B. Smart; orcid: 0000-0002-7812-7736

name: Virginia M. Moore; orcid: 0000-0001-7888-3366

SUPPLEMENTAL MATERIAL

A table of numbers of samples from hemp cultivars and locations is included.

REFERENCES

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. Applied Spectroscopy, 43(5), 772–777. https://doi.org/10.1366/0003702894202201

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., & Schwendinger, B. (2024). data.table: Extension of “data.frame”. https://CRAN.R-project.org/package=data.table

Bárta, J., Roudnický, P., Jarošová, M., Zdráhal, Z., Stupková, A., Bártová, V., Krejčová, Z., Kyselka, J., Filip, V., Říha, V., Lorenc, F., Bedrníček, J., & Smetana, P. (2024). Proteomic Profiles of Whole Seeds, Hulls, and Dehulled Seeds of Two Industrial Hemp (Cannabis sativa L.) Cultivars. Plants, 13(1), 111. https://doi.org/10.3390/plants13010111

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC Trends in Analytical Chemistry, 29(9), 1073–1081. https://doi.org/10.1016/j.trac.2010.05.006

Callaway, J. C. (2004). Hempseed as a nutritional resource: An overview. Euphytica, 140(1), 65–72. https://doi.org/10.1007/s10681-004-4811-6

Chadalavada, K., Anbazhagan, K., Ndour, A., Choudhary, S., Palmer, W., Flynn, J. R., Mallayee, S., Pothu, S., Prasad, K. V. S. V., Varijakshapanikar, P., Jones, C. S., & Kholová, J. (2022). NIR Instruments and Prediction Methods for Rapid Access to Grain Protein Content in Multiple Cereals. Sensors (Basel, Switzerland), 22(10). https://doi.org/10.3390/s22103710

Ely, K., & Fike, J. (2022). Industrial Hemp and Hemp Byproducts as Sustainable Feedstuffs in Livestock Diets. In D. C. Agrawal, R. Kumar, & M. Dhanasekaran (Eds.), Cannabis/Hemp for Sustainable Agriculture and Materials (pp. 145–162). Springer. https://doi.org/10.1007/978-981-16-8778-5_6

Garrido-Varo, A., Garcia-Olmo, J., & Fearn, T. (2019). A note on Mahalanobis and related distance measures in WinISI and The Unscrambler. Journal of Near Infrared Spectroscopy, 27(4), 253–258. https://doi.org/10.1177/0967033519848296

Geyer, M., Mohler, V., & Hartl, L. (2022). Genetics of the Inverse Relationship between Grain Yield and Grain Protein Content in Common Wheat. Plants, 11(16), 2146. https://doi.org/10.3390/plants11162146

Giancaspro, A., Giove, S. L., Blanco, A., & Gadaleta, A. (2019). Genetic Variation for Protein Content and Yield-Related Traits in a Durum Population Derived From an Inter-Specific Cross Between Hexaploid and Tetraploid Wheat Cultivars. Frontiers in Plant Science, 10. https://doi.org/10.3389/fpls.2019.01509

Hayes, M. (2020). Measuring Protein Content in Food: An Overview of Methods. Foods, 9(10), 1340. https://doi.org/10.3390/foods9101340

Kuhn, M. (2021). caret: Classification and regression training. https://CRAN.R-project.org/package=caret

Kuhn, M., & Wickham, H. (2020). Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org

Lenth, R. V. (2024). emmeans: Estimated marginal means, aka least-squares means. https://CRAN.R-project.org/package=emmeans

Li, Y., Huang, Y., Xia, J., Xiong, Y., & Min, S. (2020). Quantitative analysis of honey adulteration by spectrum analysis combined with several high-level data fusion strategies. Vibrational Spectroscopy, 108, 103060. https://doi.org/10.1016/j.vibspec.2020.103060

Liland, K. H., Mevik, B.-H., & Wehrens, R. (2021). pls: Partial least squares and principal component regression. https://CRAN.R-project.org/package=pls

Liu, M., Toth, J. A., Childs, M., Smart, L. B., & Abbaspourrad, A. (2023). Composition and functional properties of hemp seed protein isolates from various hemp cultivars. Journal of Food Science, 88(3), 942–951. https://doi.org/10.1111/1750-3841.16467

Luce, M. S., Ziadi, N., Gagnon, B., & Lévesque, V. (2017). Prediction of total carbon, total nitrogen, and pH of organic materials using visible near-infrared reflectance spectroscopy. Canadian Journal of Soil Science, 98(1), 175–179. https://doi.org/10.1139/cjss-2017-0109

Luo, J., Ying, K., He, P., & Bai, J. (2005). Properties of Savitzky–Golay digital differentiators. Digital Signal Processing, 15(2), 122–136. https://doi.org/10.1016/j.dsp.2004.09.008

Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects models in s and s-PLUS. Springer. https://doi.org/10.1007/b98882

Pinheiro, J., Bates, D., & R Core Team. (2024). nlme: Linear and nonlinear mixed effects models. https://CRAN.R-project.org/package=nlme

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rawal, A., Hartemink, A., Zhang, Y., Wang, Y., Lankau, R. A., & Ruark, M. D. (2024). Visible and near-infrared spectroscopy predicted leaf nitrogen contents of potato varieties under different growth and management conditions. Precision Agriculture, 25(2), 751–770. https://doi.org/10.1007/s11119-023-10091-z

Reeves, J. B. (2012). Potential of Near- and Mid-infrared Spectroscopy in Biofuel Production. Communications in Soil Science and Plant Analysis, 43(1-2), 478–495. https://doi.org/10.1080/00103624.2012.641844

Rinnan, Å., Berg, F. van den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry, 28(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Roberts, C. A., Workman, J., & Reeves, J. B. (2004). Near-infrared spectroscopy in agriculture. American Society of Agronomy.

Stevens, A., & Ramirez-Lopez, L. (2024). An introduction to the prospectr package.

Waring, E., Quinn, M., McNamara, A., Arino de la Rubia, E., Zhu, H., & Ellis, S. (2022). skimr: Compact and flexible summaries of data. https://CRAN.R-project.org/package=skimr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Williams, P. C. (1975). Application of near infrared reflectance spectroscopy to analysis of cereal grains and oilseeds. Cereal Chemistry, 52(4 p.561-576), 576–561.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Near Infrared Spectroscopy Predicts Crude Protein Concentration in Hemp Grain

Ryan V. Crawford, Jamie L. Crawford, Julie L. Hansen, Lawrence B. Smart,Virginia M. Moore

One page excluding cover sheet

One table

Supplemental Table S1. Table of numbers of samples from hemp cultivars and locations. Private cultivars are labeled "Cultivar1", "Cultivar2", etc., while experimental cultivars are labeled "Experimental1", "Experimental2", etc.

Cultivar	**Chazy, NY**	**Freeville, NY**	**Geneva, NY**	**Ithaca, NY**	**Willsboro, NY**	**Total**
ALTAIR				1		1
ANKA		1	3	5	2	11
BIALOBRZESKIE		1	3	4	1	9
CANDA		1	1	1		3
CFX-1		1	2	5		8
CFX-2		1	2	4		7
CRS-1	1	1	2	5		9
CULTIVAR1		1				1
CULTIVAR2				1		1
CULTIVAR3				1		1
CULTIVAR4				1		1
EARLINA 8			1			1
EXPERIMENTAL1				1		1
EXPERIMENTAL2				1		1
FELINA 32		1	2	3		6
FUTURA 75		1	3	4		8
GRANDI		3	3	4		10
H-51			1	2		3
HAN-FN-H				1		1
HAN-NW				1		1
HELENA		1				1
HENOLA				2		2
HLESIA				3		3
HLIANA			1	1		2
JOEY		1	1	1		3
KATANI		2	3	4		9
NEBRASKA (FERAL)	1			1		2
PEWTER RIVER		1				1
PICOLO		1	2	5		8
PORTUGAL			1			1
ROCKY HEMP			1			1
STERLING GOLD			1			1
SWIFT	1	1		1		3
TYGRA		1	3	4		8
USO-31	2	1	2	4		9
WOJKO		1	3	4		8
X-59		2		1		3
TOTAL	5	24	41	76	3	149