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 25-30% protein (Bárta et al., 2024; 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 easy and cheap via laboratory assay (Hayes, 2020).

Near-infrared spectroscopy (NIRS) technology is rapid, non-destructive, and cheap, and consists of the measurement of NIR radiation reflected from a sample (Roberts et al., 2004). NIR spectra from many samples are related 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).

A NIRS-scanned sample of undamaged grain may subsequently be grown, an important consideration for a plant breeder. In wheat and corn, grain protein content has been shown to be heritable [Giancaspro et al. (2019); Geyer et al. (2022)]. This suggests (at least potentially) that NIRS technology could serve as resource to more rapidly identify high CP hemp germplasm, senabling the delivery of higher CP hemp grain cultivars faster.

For this study, a benchtop NIR spectrometer was used to develop a model to predict CP content based on a data set representing multiple years, locations, and cultivars using PLSR.

0.2 MATERIALS AND METHODS

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) (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.

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.

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 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), lme4 v. 1.1.35.1 (Bates et al., 2015), prospectr v. 0.2.7 (Stevens & Ramirez-Lopez, 2024), randomForest v. 4.7.1.1 (Liaw & Wiener, 2002), rmarkdown v. 2.26 (Allaire et al., 2024; Xie et al., 2018, 2020), tidymodels v. 1.1.1 (Kuhn & Wickham, 2020), tidyverse v. 2.0.0 (Wickham et al., 2019).

Source: Article Notebook

0.2.4 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.

0.2.5 Preprocessing

Multiplicative scatter correction (MSC)

standard normal variate (SNV) transformation

Calibration and validations sets were created by dividing the laboratory CP values into tertiles according to their percent CP. Within each tertile, 75% of the samples were randomly assigned to the calibration set and the remaining 25% were assigned to the validation set.

0.3 RESULTS AND DISCUSSION

Laboratory assay

0.4 ACKNOWLEDGMENTS

0.5 SUPPLEMENTAL MATERIAL

0.6 OPTIONAL SECTIONS

0.7 REFERENCES

0.8 FIGURES AND TABLES

Source: Article Notebook

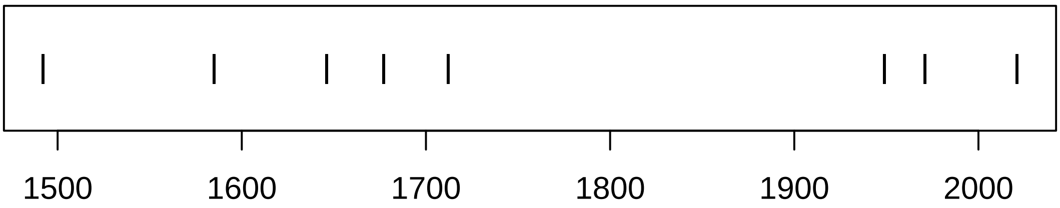

Figure 1: Timeline of recent earthquakes on La Palma

Source: Article Notebook

Source: Article Notebook

Based on data up to and including 1971, eruptions on La Palma happen every 79.8 years on average.

Studies of the magma systems feeding the volcano, such as (**marrero2019?**), have proposed that there are two main magma reservoirs feeding the Cumbre Vieja volcano; one in the mantle (30-40km depth) which charges and in turn feeds a shallower crustal reservoir (10-20km depth).

Eight eruptions have been recorded since the late 1400s (Figure 1).

Data and methods are discussed in Section 0.9.

Let x denote the number of eruptions in a year. Then, x can be modeled by a Poisson distribution

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{1}$$

where λ is the rate of eruptions per year. Using Equation 1, the probability of an eruption in the next t years can be calculated.

Table 1: Recent historic eruptions on La Palma

Name	Year
Current	2021
Teneguía	1971
Nambroque	1949
El Charco	1712
Volcán San Antonio	1677
Volcán San Martin	1646
Tajuya near El Paso	1585
Montaña Quemada	1492

Table 1 summarises the eruptions recorded since the colonization of the islands by Europeans in the late 1400s.

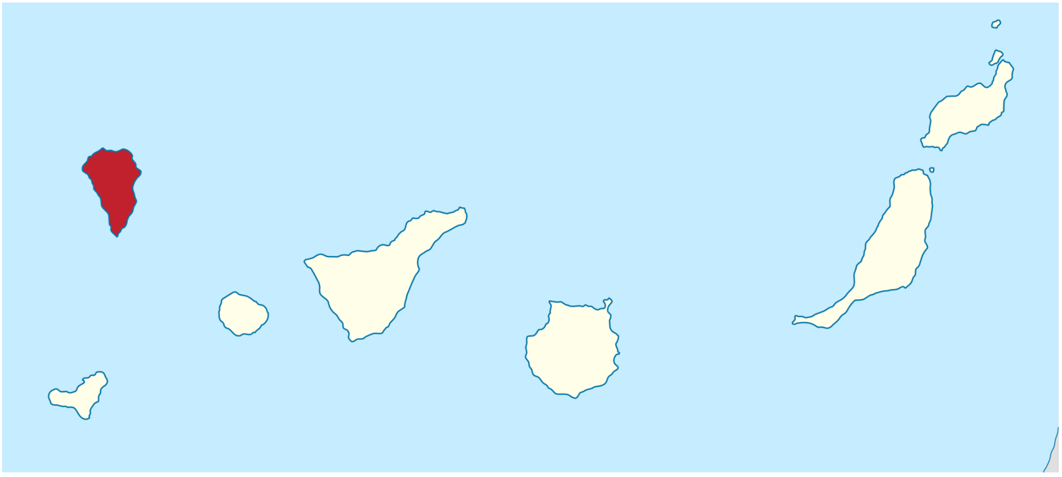

Figure 2: Map of La Palma

La Palma is one of the west most islands in the Volcanic Archipelago of the Canary Islands (Figure 2).

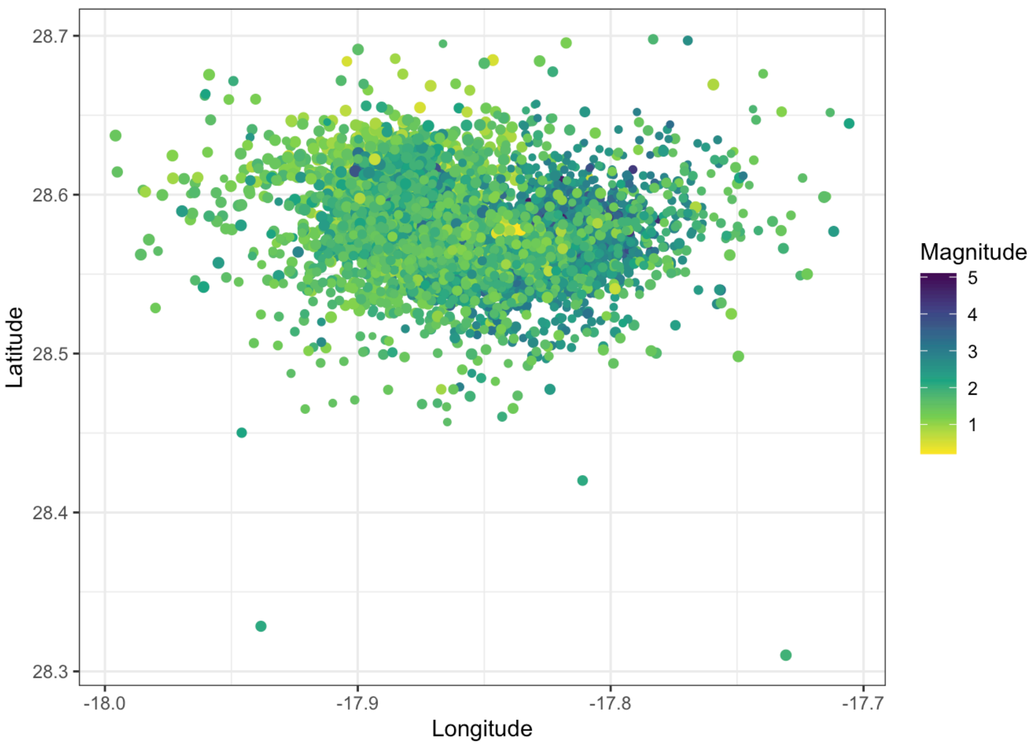

Figure 3: Locations of earthquakes on La Palma since 2017

Source: Explore Earthquakes

Figure 3 shows the location of recent Earthquakes on La Palma.

0.9 Data & Methods

0.10 Conclusion

References

Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2024). *rmarkdown: Dynamic documents for r.* https://github.com/rstudio/rmarkdown

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., & Hocking, T. (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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Kuhn, & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

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

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. https://CRAN.R-project.org/doc/Rnews/

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

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

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*.

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.

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook