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 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, 2012a; 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 principal components (PCs) which relate to the dependent variable and summarize the spectra and uses a subset of PCs in order to fit the regression model. PLSR is commonly used in spectroscopy because it tends to work well with highly-correlated spectral data. Typically the number of principal components is chosen via cross-validation to avoid overfitting. **CITES FOR ALL OF THIS**

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 (Geyer et al., 2022; Giancaspro et al., 2019). This suggests (at least potentially) that NIRS technology could serve as resource to more rapidly identify high CP hemp germplasm, enabling 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 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

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 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 validation sets. Within each tertile, 75% of the samples were randomly assigned to the calibration set and the remaining 25% were assigned to the validation set. For each calibration set, models were developed in caret using PLSR. In fitting the model, the number of principal components was optimized over a grid search from 1-20. Model performance was evaluated with a bootstrapping method using root mean squared error (RMSE) and R^2 statistics in selecting a final model.

Source: Article Notebook

Initially several of common spectral preprocessing methods were tested by creating 100 calibration and validation sets as described above. Spectral data from those data sets were transformed by each of the following methods: First Derivative, Gap-segment derivative, Multiplicative Scatter Correction, Raw Spectra, Savitzky-Golay, SNV-Detrend, Standard Normal Variate, Standard Normal Variate/ Savitzky-Golay. 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 validation set were calculated via RMSE, R^2 and Ratio of Performance to InterQuartile distance (RPIQ). 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 allowing different variances 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 validation sets 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), knitr v. 1.45 (Xie, 2014, 2015, 2023), lme4 v. 1.1.35.1 (Bates et al., 2015), multcomp v. 1.4.25 (Hothorn et al., 2008), 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), randomForest v. 4.7.1.1 (Liaw & Wiener, 2002), rmarkdown v. 2.26 (Allaire et al., 2024; Xie et al., 2018, 2020), 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 and this indicating an appropriate basis for a chemometric model.

Table 1: Summary of Laboratory Assayed CP values (Values are 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

All preprocessing methods outperformed raw spectral data.

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Table 2: Evaluation of Preprocessing Methods

Preprocessing Method	RMSE	R^2	RPIQ
Savitzky-Golay	1.16 ± 0.018	0.77 ± 0.009	3.36 ± 0.081
First Derivative	1.18 ± 0.018	0.76 ± 0.009	3.30 ± 0.081
Standard Normal Variate/ Savitzky-Golay	1.19 ± 0.018	0.76 ± 0.009	3.26 ± 0.082
Gap-segment derivative	1.23 ± 0.019	0.75 ± 0.009	3.17 ± 0.081

Table 2: Evaluation of Preprocessing Methods

Preprocessing Method	RMSE	R^2	RPIQ
SNV-Detrend	1.27 ± 0.021	0.73 ± 0.010	3.10 ± 0.083
Standard Normal Variate	1.29 ± 0.023	0.73 ± 0.010	3.07 ± 0.083
Multiplicative Scatter Correction	1.32 ± 0.020	0.72 ± 0.010	2.96 ± 0.081
Raw Spectra	1.37 ± 0.028	0.69 ± 0.012	2.89 ± 0.086

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

The Savitzky-Golay method had the lowest RMSE, highest R^2, and highest RPIQ averaging over all iterations. Savitzky Golay RMSE averaged 1.7% lower, while R^2 and RPIQ averaged 1.2% and 2.0% higher than the next best preprocessing method (first derivative, according to each metric). However, only at R^2 was the difference from the next best preprocessing method statistically significant at α <0.05. Averaged together, all preprocessed spectra were superior to raw spectra, with lower RMSE, and higher R^2 and RPIQ values (significant at α level <0.001). Averaged together, preprocessing methods had -11.1 % lower RMSE, and had 6.9% higher R^2 9.1% higher RPIQ than unprocessed spectra. Because the Savitzky-Golay method was the best performing under all methods, a final model was developed using that preprocessing method.

0.3.3 Final model development and summary

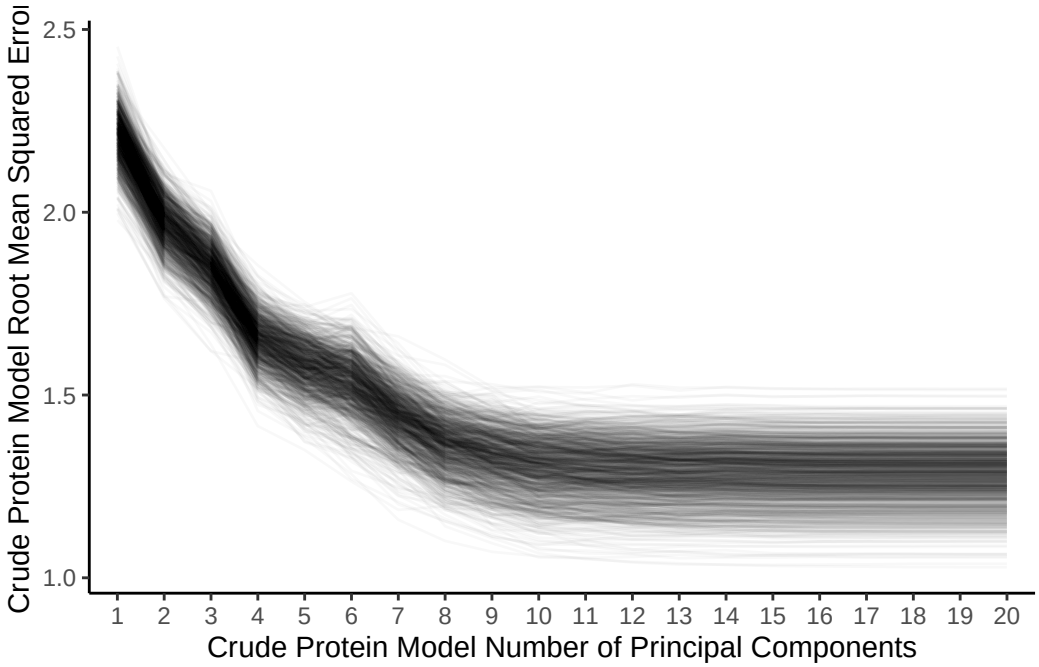

Figure 1: Decreasing RMSE with increasing number of PCs

Source: Article Notebook

calculate relative distance from minima The relative ranks of the models are stable at approximately 14 PCs, with very few models (lines) crossing over one another.

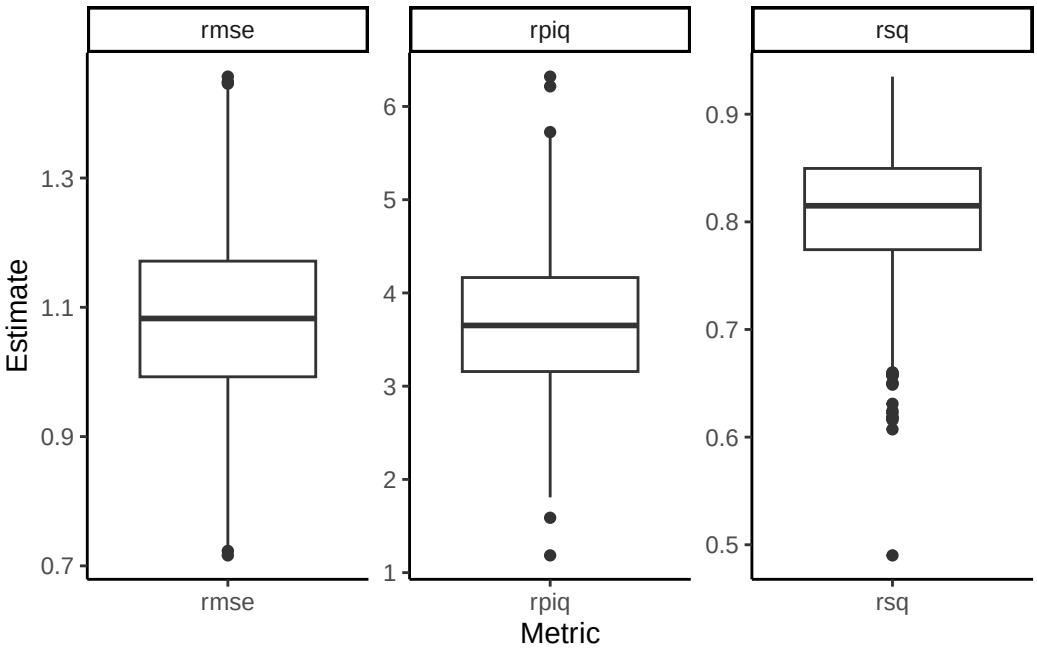

Figure 2: Final model test set evaluations (1000 iterations per plot)

Source: Article Notebook

Source: Article Notebook

Source: Article Notebook

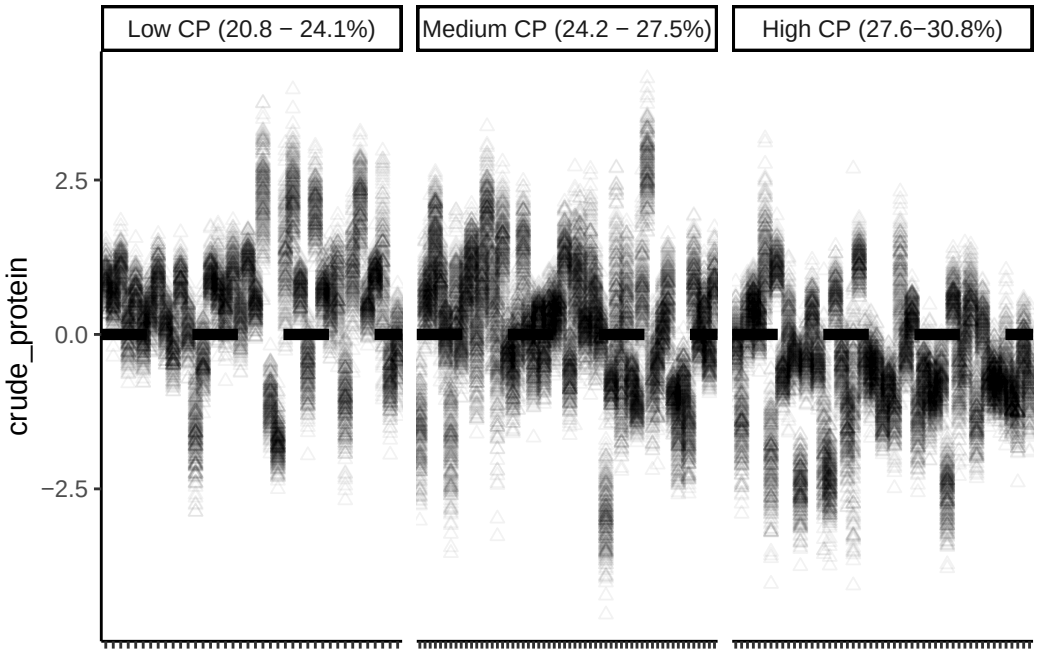

Figure 3: Test set predictions versus normalized scores across hemp samples

Source: Article Notebook

Errors tend to be lower at higher levels of actual CP

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, if the machines compared are of the same model (Reeves, 2012b). As well, the calibration and validation sets are relatively small.

This research showed the promise of the use of NIRS in order to make predictions concerning %CP 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 or by examining other predictive methods.

0.4 ACKNOWLEDGMENTS

0.5 SUPPLEMENTAL MATERIAL

0.6 OPTIONAL SECTIONS

0.7 REFERENCES

0.8 FIGURES AND TABLES

Source: Article Notebook

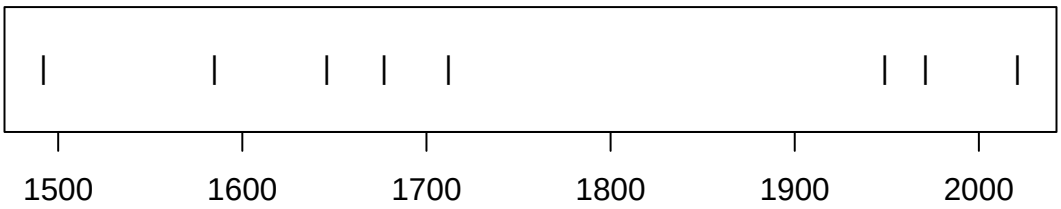

Figure 4: Timeline of recent earthquakes on La Palma

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 4).

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 3: 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 3 summarises the eruptions recorded since the colonization of the islands by Europeans in the late 1400s.

La Palma is one of the west most islands in the Volcanic Archipelago of the Canary Islands (Figure 5).

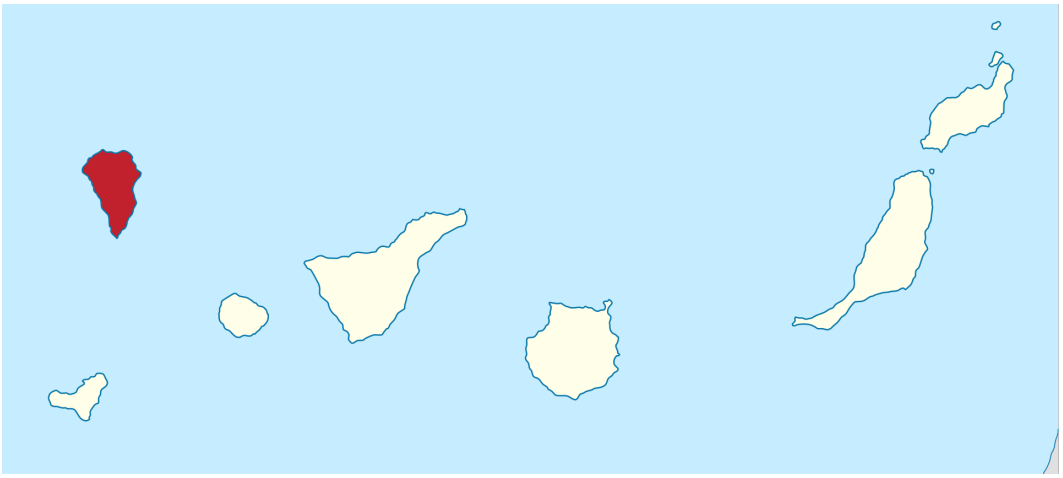

Figure 5: Map of La Palma

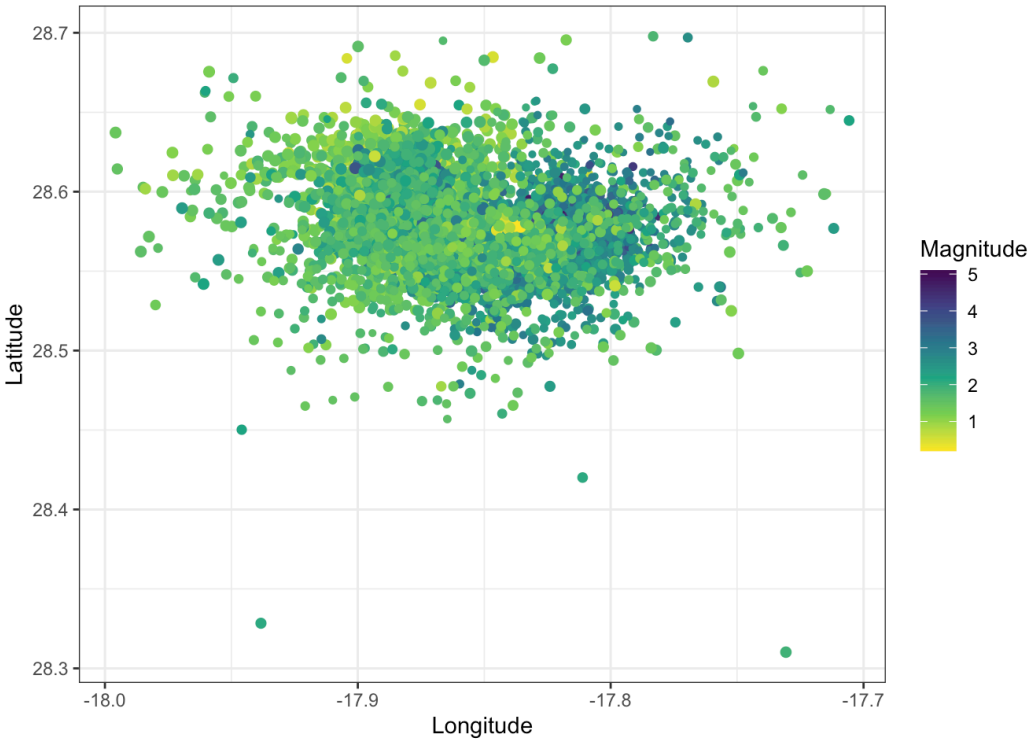

Figure 6: Locations of earthquakes on La Palma since 2017

Source: Explore Earthquakes

Figure 6 shows the location of recent Earthquakes on La Palma.

198 **0.9 Data & Methods**
199 **0.10 Conclusion**
200 **References**

201 Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., et al.
202 (2024). *rmarkdown: Dynamic documents for r*. Retrieved from https://github
203 .com/rstudio/rmarkdown
204 Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., & Hocking, T.
205 (2024). *data.table: Extension of "data.frame"*. Retrieved from https://CRAN
206 .R-project.org/package=data.table
207 Bárta, J., Roudnický, P., Jarošová, M., Zdráhal, Z., Stupková, A., Bártová, V.,
208 et al. (2024). Proteomic Profiles of Whole Seeds, Hulls, and Dehulled Seeds
209 of Two Industrial Hemp (Cannabis sativa L.) Cultivars. *Plants*, *13*(1), 111.
210 https://doi.org/10.3390/plants13010111
211 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects
212 models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://
213 doi.org/10.18637/jss.v067.i01
214 Callaway, J. C. (2004). Hempseed as a nutritional resource: An overview. *Euphytica*,
215 *140*(1), 65–72. https://doi.org/10.1007/s10681-004-4811-6
216 Chadalavada, K., Anbazhagan, K., Ndour, A., Choudhary, S., Palmer, W., Flynn, J.
217 R., et al. (2022). NIR Instruments and Prediction Methods for Rapid Access to
218 Grain Protein Content in Multiple Cereals. *Sensors (Basel, Switzerland)*, *22*(10).
219 https://doi.org/10.3390/s22103710
220 Ely, K., & Fike, J. (2022). Industrial Hemp and Hemp Byproducts as Sustainable
221 Feedstuffs in Livestock Diets. In D. C. Agrawal, R. Kumar, & M. Dhanasekaran
222 (Eds.), *Cannabis/Hemp for Sustainable Agriculture and Materials* (pp. 145–162).
223 Singapore: Springer. https://doi.org/10.1007/978-981-16-8778-5_6
224 Garrido-Varo, A., Garcia-Olmo, J., & Fearn, T. (2019). A note on Mahalanobis
225 and related distance measures in WinISI and The Unscrambler. *Journal of*
226 *Near Infrared Spectroscopy*, *27*(4), 253–258. https://doi.org/10.1177/
227 0967033519848296
228 Geyer, M., Mohler, V., & Hartl, L. (2022). Genetics of the Inverse Relationship
229 between Grain Yield and Grain Protein Content in Common Wheat. *Plants*,
230 *11*(16), 2146. https://doi.org/10.3390/plants11162146
231 Giancaspro, A., Giove, S. L., Blanco, A., & Gadaleta, A. (2019). Genetic Variation
232 for Protein Content and Yield-Related Traits in a Durum Population Derived
233 From an Inter-Specific Cross Between Hexaploid and Tetraploid Wheat Cultivars.
234 *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01509
235 Hayes, M. (2020). Measuring Protein Content in Food: An Overview of Methods.
236 *Foods*, *9*(10), 1340. https://doi.org/10.3390/foods9101340
237 Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general
238 parametric models. *Biometrical Journal*, *50*(3), 346–363.
239 Kuhn, & Max. (2008). Building predictive models in r using the caret package.
240 *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/
241 jss.v028.i05
242 Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for model-*
243 *ing and machine learning using tidyverse principles.* Retrieved from https://
244 www.tidymodels.org
245 Lenth, R. V. (2024). *emmeans: Estimated marginal means, aka least-squares means.*
246 Retrieved from https://CRAN.R-project.org/package=emmeans
247 Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R*
248 *News*, *2*(3), 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/
249 Liland, K. H., Mevik, B.-H., & Wehrens, R. (2023). *pls: Partial least squares and*
250 *principal component regression.* Retrieved from https://CRAN.R-project.org/
251 package=pls

Pinheiro, J., Bates, D., & R Core Team. (2023). *nlme: Linear and nonlinear mixed effects models*. Retrieved from https://CRAN.R-project.org/package=nlme

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

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

Reeves, J. B. (2012a). 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

Reeves, J. B. (2012b). 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*.

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (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. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman; Hall/CRC.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Xie, Y. (2023). *knitr: A general-purpose package for dynamic report generation in r*. Retrieved from https://yihui.org/knitr/

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

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