

Near Infra-Red Spectroscopy Predicts Crude Protein in Hemp Grain

Ryan Crawford¹, Jamie Crawford¹, Lawrence B. Smart², Virginia Moore³

¹Cornell University, Ithaca, NY,

²Cornell AgriTech, Geneva, NY,

³Cornell 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; Ely & Fike, 2022; **callaway2004?**). 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). 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 used for other purposes 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 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](#)

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.

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 validation 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 25 iterations of bootstrapping and minimized root mean squared error (RMSE) in selecting the number of principal components in the final model .

Source: [Article Notebook](#)

Initially a number 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: 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), 4) standard normal variate fol-

lowing Savitzky-Golay (SNV-SG) (same SG parameters as above), 5) SNV-detrend with second order polynomial, and 6) multiplicative scatter correction.

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 validation 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 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 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 raw spectral data Table 2. 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). Preprocessing methods had -11.6 % lower RMSE, and had 3.1% higher R^2 7.4% higher RPIQ than unprocessed spectra.

The SNV-SG method had the lowest RMSE, highest R^2 , and highest RPIQ averaging over all iterations. SNV-SG RMSE averaged 1.4% lower, while R^2 and RPIQ averaged 0.4% and 2.4% higher respectively than the next best preprocessing method (SG in both cases), but the difference between the best and second best method by metric were only statistically significant at $\alpha < 0.05$ for RPIQ. RPIQ was devised to accurately reflect the spread of data in skewed populations (**bellon-maurel2010?**) and thus offers a robust metric for model assessment in this context, where the CP data are skewed. Therefore the superiority of SNV-SG as measured via RPIQ made it the best choice for the final model.

Table 2: Evaluation of Preprocessing Methods by Metric \pm Standard Error

Preprocessing Method	RMSE	R^2	RPIQ
Standard Normal Variate following	1.02 \pm	0.84 \pm	3.97 \pm
Savitzky-Golay	0.012	0.004	0.076
Savitzky-Golay	1.03 \pm	0.83 \pm	3.88 \pm
	0.012	0.004	0.072
First Derivative	1.07 \pm	0.82 \pm	3.77 \pm
	0.013	0.004	0.075
Standard Normal Variate	1.12 \pm	0.80 \pm	3.61 \pm
	0.016	0.005	0.081
Gap-segment Derivative	1.12 \pm	0.81 \pm	3.60 \pm
	0.018	0.006	0.086
Standard Normal Variate-Detrend	1.13 \pm	0.80 \pm	3.55 \pm
	0.015	0.005	0.079
Multiplicative Scatter Correction	1.17 \pm	0.79 \pm	3.47 \pm
	0.016	0.006	0.080
Raw Spectra	1.22 \pm	0.79 \pm	3.42 \pm
	0.044	0.009	0.105

Source: [Article Notebook](#)

SNV and SNV-detrend correct light scatter. SG is a smoothing filter that regresses on the signal over a series of windows. Derivatives remove noise, but not necessarily light scattering. **cite**

cite: Barnes RJ, Dhanoa MS, Lister SJ. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy*, 43(5): 772-777.

The preprocessing methods examined represent a portion of those available. As well, preprocessing methods tend to have a number of user-adjustable parameters whose various permutations were not tested. This subset of preprocessing methods and parameters nonetheless contained substantial variations in model quality, demonstrating the importance of the selection of an appropriate preprocessing method.

0.3.3 Final model development and summary

Source: [Article Notebook](#)

The model improved most rapidly as the number of principal components increased from 1 to 7, with the inclusion of each additional PC being associated with a decrease in RMSE of 5-12% . From 8 to 12 PCs, model performance continued to improve, although gains were more modest (decrease in RMSE of 0.7-3%). With 13 or more PCs, performance gains were minimal and the relative ranks of the models tended to be stable Figure 1.

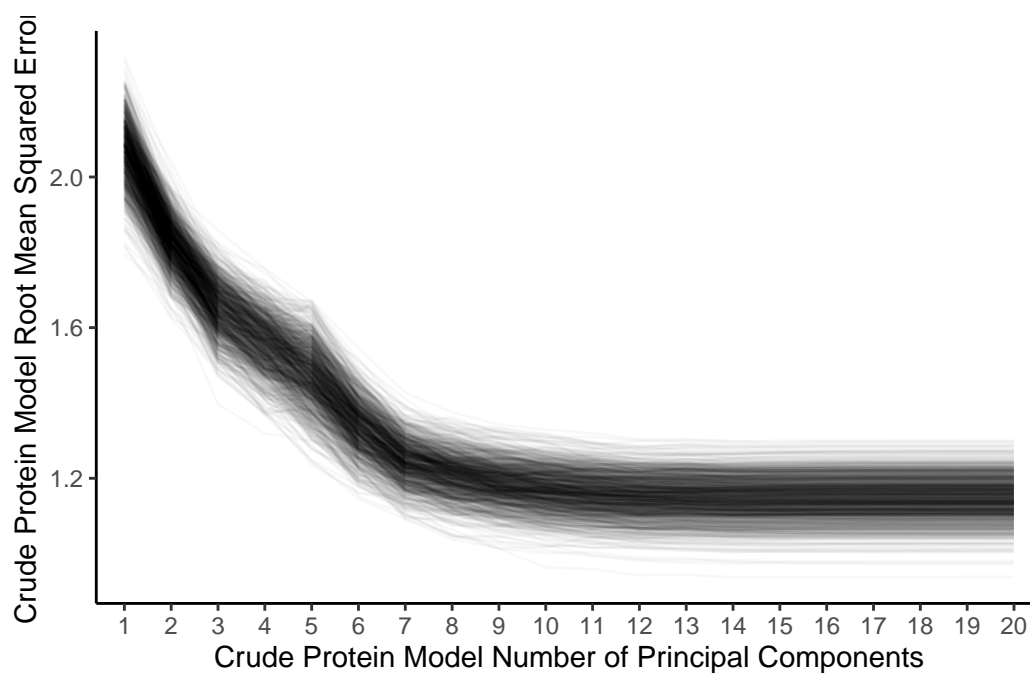


Figure 1: Decreasing RMSE with increasing number of PCs

177 Source: [Article Notebook](#)

178 Source: [Article Notebook](#)

179 Final model performance was similar, but not identical to, that obtained during
 180 the initial comparison of preprocessing methods. The final models' mean RMSE
 181 was 1.03, R^2 was 0.83, and RPIQ was 3.89 (all calculated on the test sets). Despite
 182 the generally good model performance, a subset of poor models can be seen. For
 183 example, Figure 2 shows twenty-one models with R^2 below 0.7. **more comment on**
 184 **poor models?**

185 Source: [Article Notebook](#)

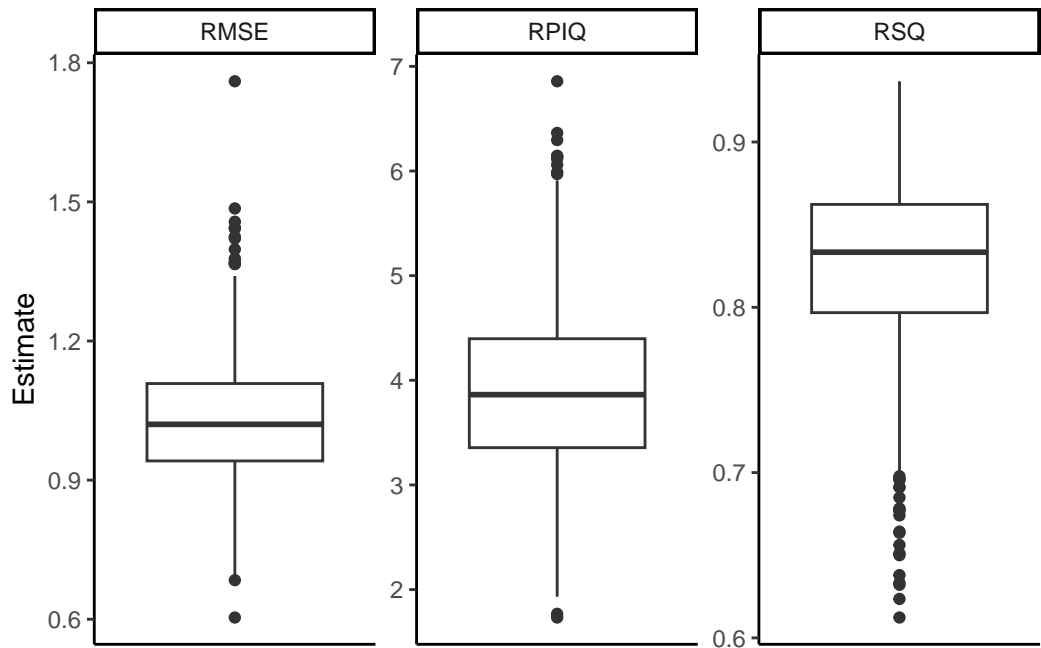


Figure 2: Final model validation set performance (1000 iterations)

```
186 Source: Article Notebook
187 Source: Article Notebook
188 Source: Article Notebook
189
190 Call:
191 lm(formula = difference ~ adj_cp, data = temp_dat)
192
193 Residuals:
194      Min       1Q   Median       3Q      Max
195 -2.51794 -0.58132  0.06936  0.50754  2.74745
196
197 Coefficients:
198             Estimate Std. Error t value Pr(>|t|)
199 (Intercept)  0.80412    0.17827   4.511 1.31e-05 ***
200 adj_cp      -0.15334    0.03051  -5.026 1.44e-06 ***
201 ---
202 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
203
204 Residual standard error: 0.9138 on 147 degrees of freedom
205 Multiple R-squared:  0.1466,    Adjusted R-squared:  0.1408
206 F-statistic: 25.26 on 1 and 147 DF,  p-value: 1.438e-06
207
208      ith_in_data_set      preds crude_protein      cutpoints plot_order
209      <int>          <num>          <num>          <fctr>          <int>
210 1:          50  0.8041160          20.8 (20.8,24.1]          1
211 2:          42  0.7274479          21.3 (20.8,24.1]          2
212 3:          52  0.6967806          21.5 (20.8,24.1]          3
213 4:          83  0.6814470          21.6 (20.8,24.1]          4
214 5:          85  0.6814470          21.6 (20.8,24.1]          5
```

214

215

145:

12

-0.6219116

30.1

(27.5,30.8]

145

216

146:

55

-0.6219116

30.1

(27.5,30.8]

146

217

147:

117

-0.6372452

30.2

(27.5,30.8]

147

218

148:

63

-0.6679125

30.4

(27.5,30.8]

148

219

149:

112

-0.7292470

30.8

(27.5,30.8]

149

220

Source: [Article Notebook](#)

221

Finally, the pattern of errors was examined on a per-sample basis. Figure 3

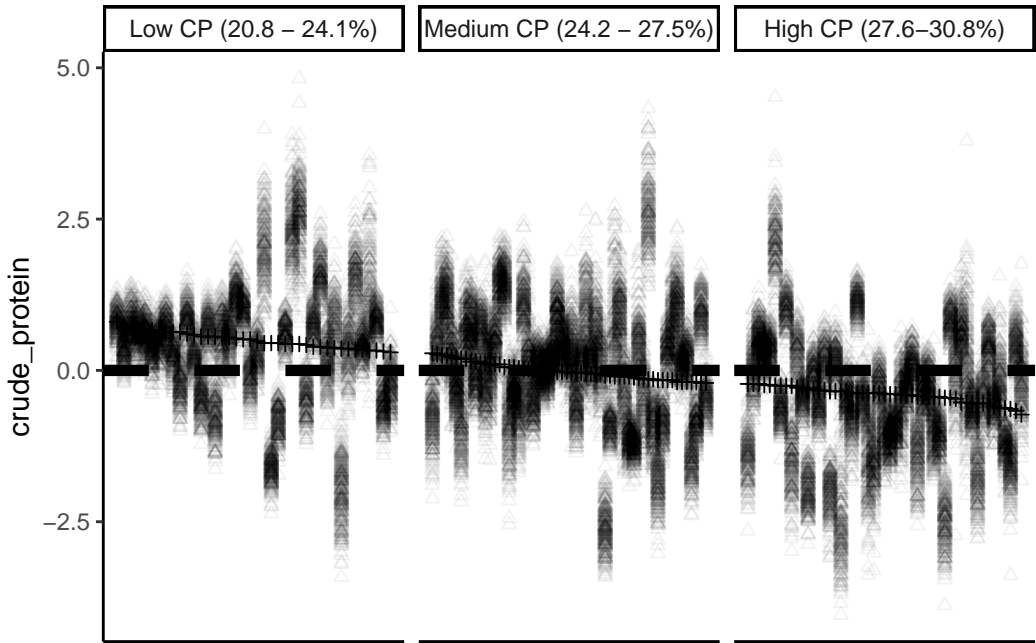


Figure 3: Test set prediction errors on a per-sample basis. Actual sample value set to 0, and samples ranked from least to greatest.

222

Source: [Article Notebook](#)

223

Errors tend to be lower at higher levels of actual CP

224

Key: <loc>

225

loc

ith_in_data_set

crude_protein

difference

tmp_ith

adj_cp

226

<char>

<int>

<num>

<num>

<int>

<num>

227

1:

cnoll

4

27.1

-1.826000348

90

6.3

228

2:

free

91

23.1

-1.608410140

24

2.3

229

3:

free

100

27.5

0.008386084

100

6.7

230

4:

free

103

26.0

-0.095290561

64

5.2

231

5:

freev

66

28.3

0.156805468

117

7.5

232

6:

freev

69

28.7

0.115415937

128

7.9

233

7:

freev

70

27.3

0.123941888

93

6.5

234

8:

ithaca

123

25.9

0.137420230

60

5.1

235

9:

ithaca

124

23.2

2.325322409

27

2.4

236

10:

ithaca

125

23.2

2.582685059

25

2.4

237

11:

ithaca

140

23.5

1.597167816

30

2.7

238

12:

ithaca

146

28.0

-1.965391942

110

7.2

239	13:	ithaca	148	26.4	0.145123025	76	5.6
240	14:	mcg	92	22.7	-0.110406545	20	1.9
241	15:	mcg	94	28.1	-0.067559276	111	7.3
242	16:	mcg	121	28.3	-2.434565554	115	7.5
243	17:	mg4e	28	25.4	0.034158139	53	4.6
244	18:	mg4e	39	26.1	-0.016921027	66	5.3
245	19:	mg4l	41	26.4	0.118021909	74	5.6
246	20:	rn041_gen	108	27.0	2.600884960	86	6.2
247	21:	rn041_gen	109	23.7	-1.980462540	33	2.9
248	22:	rn041_gen	110	29.0	-2.495337360	135	8.2
249	23:	rn041_gen	111	29.6	-1.777305991	140	8.8
250	24:	rn041_gen	113	29.5	0.011781362	139	8.7
251	25:	rn041_gen	114	27.7	2.115039075	107	6.9
252	26:	rn00a	14	29.1	-0.003578789	137	8.3
253	27:	rn00a	18	23.1	2.020690041	22	2.3
254	28:	rn00a	19	25.3	-0.110141285	52	4.5
255	29:	rn00a	25	26.6	-2.603177621	78	5.8
256	30:	rn00a	26	28.3	-2.034162792	114	7.5
257	loc ith_in_data_set crude_protein difference tmp_ith adj_cp						
258	grouping		temp_id	harv_year	cultivar		type
259	<char>		<int>	<int>	<char>		<char>
260	1:	worst_predicted_10_pct	32	2017	cfx-1		
261	2:	worst_predicted_10_pct	111	2018	grandi		grain
262	3:	best_predicted_10pct	128	2019	canda		grain
263	4:	best_predicted_10pct	131	2019	nwg-elite		dual
264	5:	best_predicted_10pct	94	2017	picolo		
265	6:	best_predicted_10pct	97	2017	futura 75		
266	7:	best_predicted_10pct	98	2017	cfx-1		
267	8:	best_predicted_10pct	155	2020	nwg-2730		multistate
268	9:	worst_predicted_10_pct	152	2020	futura 75		dual_replant
269	10:	worst_predicted_10_pct	156	2020	bialobrzeskie		multistate
270	11:	worst_predicted_10_pct	9	2021	logan c2		exp.
271	12:	worst_predicted_10_pct	5	2021	crs-1		grain_dual
272	13:	best_predicted_10pct	10	2021	logan x anka c2		exp.
273	14:	best_predicted_10pct	115	2018	nebraska (feral)		dual
274	15:	best_predicted_10pct	113	2018	picolo		grain
275	16:	worst_predicted_10_pct	149	2019	h-51		multistate
276	17:	best_predicted_10pct	56	2017	cfx-2		
277	18:	best_predicted_10pct	67	2017	futura 75		
278	19:	best_predicted_10pct	69	2017	anka		
279	20:	worst_predicted_10_pct	136	2019	earlina 8		grain
280	21:	worst_predicted_10_pct	137	2019	katani		grain
281	22:	worst_predicted_10_pct	138	2019	joey		grain
282	23:	worst_predicted_10_pct	139	2019	futura 75		dual
283	24:	best_predicted_10pct	141	2019	h-51		dual
284	25:	worst_predicted_10_pct	142	2019	hliana		dual
285	26:	best_predicted_10pct	42	2017	tygra		
286	27:	worst_predicted_10_pct	46	2017	katani		
287	28:	best_predicted_10pct	47	2017	cfx-2		
288	29:	worst_predicted_10_pct	53	2017	cfx-1		
289	30:	worst_predicted_10_pct	54	2017	wojko		
290	grouping		temp_id	harv_year	cultivar		type
291	in_ny	loc2					
292	<lgcl>	<char>					
293	1:	TRUE	geneva				

294 2: TRUE freeville
 295 3: TRUE freeville
 296 4: TRUE freeville
 297 5: TRUE freeville
 298 6: TRUE freeville
 299 7: TRUE freeville
 300 8: TRUE ithaca
 301 9: TRUE ithaca
 302 10: TRUE ithaca
 303 11: TRUE ithaca
 304 12: TRUE ithaca
 305 13: TRUE ithaca
 306 14: TRUE ithaca
 307 15: TRUE ithaca
 308 16: TRUE ithaca
 309 17: TRUE ithaca
 310 18: TRUE ithaca
 311 19: TRUE ithaca
 312 20: TRUE geneva
 313 21: TRUE geneva
 314 22: TRUE geneva
 315 23: TRUE geneva
 316 24: TRUE geneva
 317 25: TRUE geneva
 318 26: TRUE geneva
 319 27: TRUE geneva
 320 28: TRUE geneva
 321 29: TRUE geneva
 322 30: TRUE geneva
 323 in_ny loc2

324 Source: [Article Notebook](#)

325 This study is limited in that it represents the creation of one model based upon
 326 spectra collected from one machine. NIRS calibrations can be unique to a particular
 327 machine, even if the machines compared are of the same model (**Reeves_2012?**) .
 328 As well, the calibration and validation sets are relatively small.

329 This research showed the promise of the use of NIRS in order to make predictions
 330 concerning %CP in hemp grain using PLS. Promising preprocessing methods were
 331 identified and a model was validated. Further research could refine the model by
 332 including more samples or by examining other predictive methods.

333 0.4 ACKNOWLEDGMENTS

334 0.5 SUPPLEMENTAL MATERIAL

Table 3: Tally of hemp cultivars and locations. Private cultivars are labeled “cultivar1”, “cultivar2”, etc.

cultivar2	chazy	freeville	geneva	ithaca	willsboro	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

Table 3: Tally of hemp cultivars and locations. Private cultivars are labeled “cultivar1”, “cultivar2”, etc.

cultivar2	chazy	freeville	geneva	ithaca	willsboro	Total
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
pico		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

Source: [Article Notebook](#)

0.6 OPTIONAL SECTIONS

0.7 REFERENCES

0.8 FIGURES AND TABLES

Source: [Article Notebook](#)

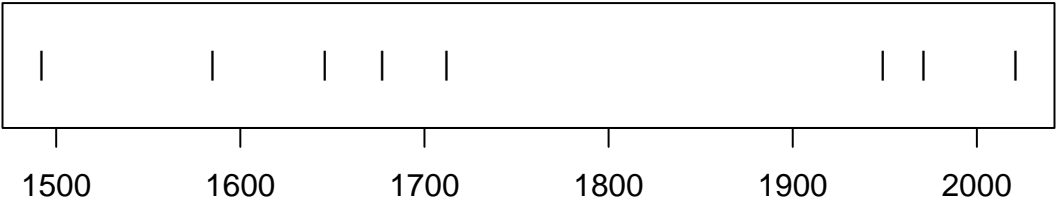


Figure 4: Timeline of recent earthquakes on La Palma

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

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

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 4: 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 4 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).

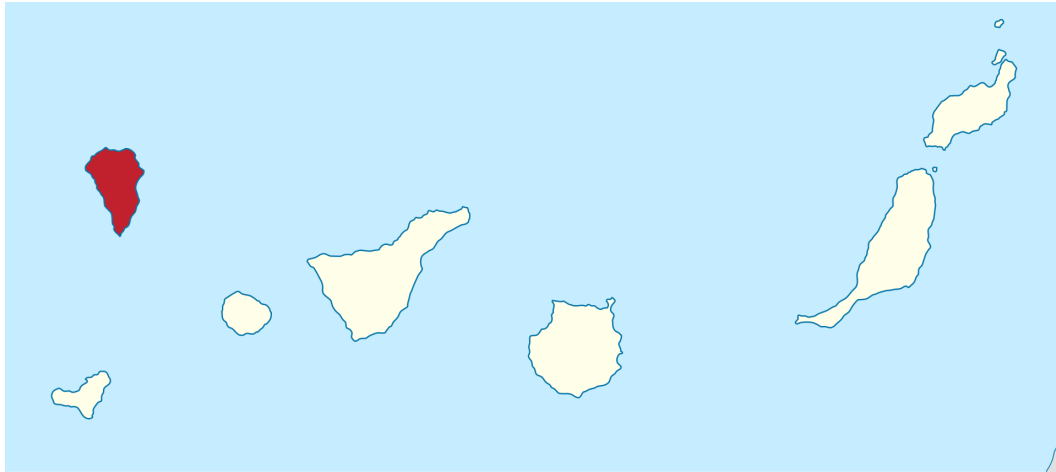


Figure 5: Map of La Palma

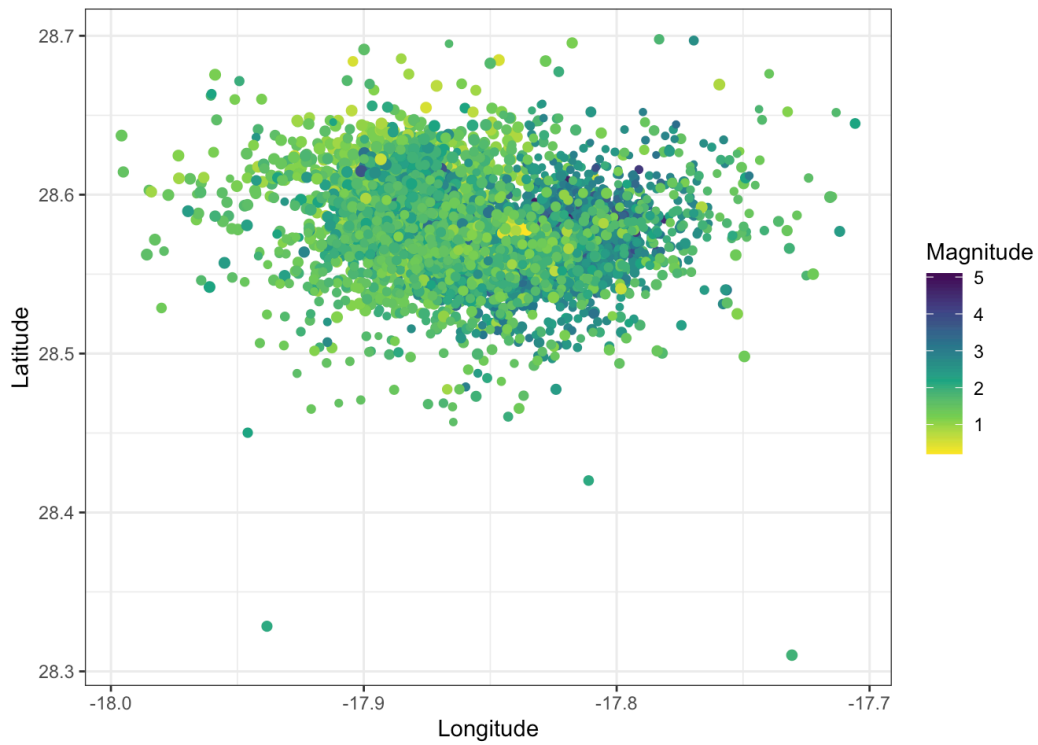


Figure 6: Locations of earthquakes on La Palma since 2017

358 Source: [Explore Earthquakes](#)

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

0.9 Data & Methods

0.10 Conclusion

References

- 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>
- 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>
- Liland, K. H., Mevik, B.-H., & Wehrens, R. (2023). *pls: Partial least squares and principal component regression*. <https://CRAN.R-project.org/package=pls>
- 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. (2023). *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/>
- 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*.

- 415 Waring, E., Quinn, M., McNamara, A., Arino de la Rubia, E., Zhu, H., & Ellis,
416 S. (2022). *skimr: Compact and flexible summaries of data*. [https://CRAN.R](https://CRAN.R-project.org/package=skimr)
417 [-project.org/package=skimr](https://CRAN.R-project.org/package=skimr)
- 418 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R.,
419 Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L.,
420 Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu,
421 V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source*
422 *Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- 423 Williams, P. C. (1975). Application of near infrared reflectance spectroscopy to anal-
424 ysis of cereal grains and oilseeds. *Cereal Chemistry*, 52(4 p.561-576), 576–561.