PLS Regression-Based Hyperspectral Soft Sensor Models for Multi-Trait Prediction

Bahadir Bagci, Hassan Majeed, Zoheb Rahman

October 2025

1 Introduction

Spectral data is among the most challenging cases for classical regression models. In spectral data, many wavelength bands are stored as features, and they tend to be highly correlated. As a result of this, classic regression methods often perform poorly, since the models cannot handle the high number of features.

PLS regression is commonly used to deal with multicollinearity in spectral data. Instead of working with the original spectral variables, PLS forms a set of latent variables that summarize the main variation in the spectra and their relation to the traits. In this way, the method simplifies the structure and makes it possible to develop stable regression models.

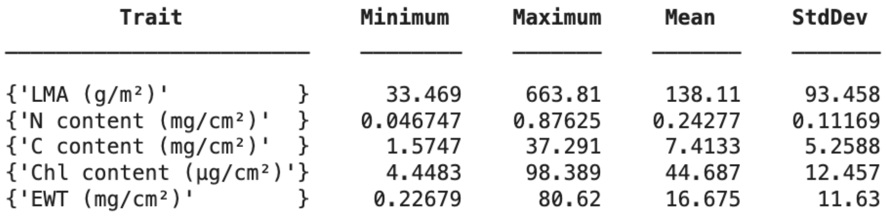
In this work, PLS regression is applied to construct soft sensor models that predict plant traits from hyperspectral data. The aim is to handle the high dimensionality and multicollinearity in the spectra, and to show that PLS provides a stable and accurate framework compared to classical regression methods.

The remainder of this report is organized as follows. Section 2, materials and methods, introduces the data set, the pre-processing steps, and the mathematical background of the PLS regression. Section 3, results and discussion, presents the performance and interprets the results of the models. Finally, section 4, conclusion, provides a short summary of the findings.

2 Materials and Methods

2.1 Data Description

The dataset consists of 12,180 samples collected from 42 studies across various environments. It contains 1,721 spectral bands and 20 vegetation traits. For the regression models, five traits were chosen (C, Chl, EWT, LMA, and N) since they have fewer missing values compared to the rest. An overview of statistical distribution of the selected traits is presented in Fig. 1.



```
         Trait               Minimum     Maximum     Mean       StdDev 
_________________________    ________    _______    _______    _______

{'LMA (g/m²)'           }      33.469     663.81     138.11     93.458
{'N content (mg/cm²)'   }    0.046747    0.87625    0.24277    0.11169
{'C content (mg/cm²)'   }      1.5747     37.291     7.4133     5.2588
{'Chl content (µg/cm²)'}      4.4483     98.389     44.687     12.457
{'EWT (mg/cm²)'         }     0.22679      80.62     16.675      11.63
```

Figure 1: Descriptive Statistics of Selected Vegetation Traits

2.2 Data Preparation

2.2.1 Data cleaning and Splitting

Since we wanted to use the same training and testing set for all traits, rows where all five selected traits were available were retained. This resulted in $2,702$ valid samples. The dataset was then split into training (80%) and testing (20%) using stratified sampling, as the traits have different distributions and needed to be represented fairly in both sets. Following this split, Z-score standardization was performed. The mean (μ) and standard deviation (σ) for both the spectral ($\mathbf{X}$) and trait ($\mathbf{Y}$) were calculated from the training set. These parameters were used to scale both train and test dataset to ensure there is no data leakage. As a result, we had 2163 samples for the train set and 539 samples for the test set. The successful distribution matching between the two subsets is confirmed by the trait histogram shown in Fig. 2

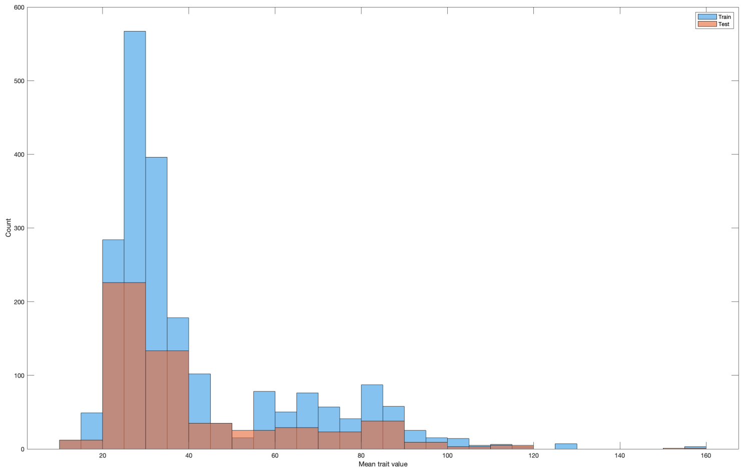


Figure 2: Distribution of Mean Trait Values in Train and Test Sets

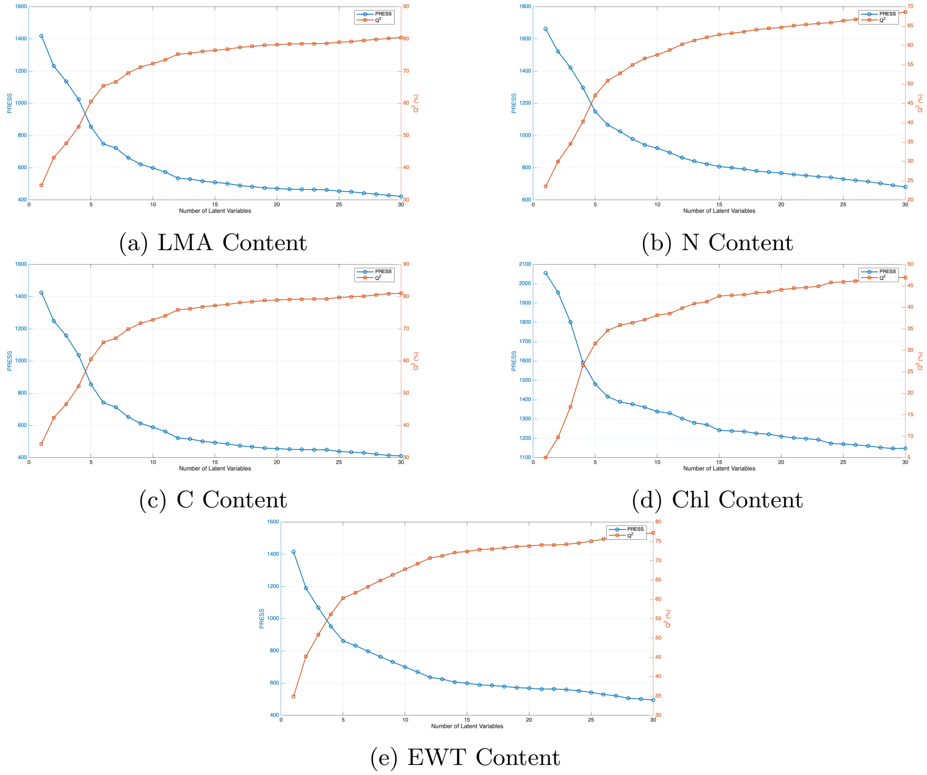
2.3 PLS Regression

To build predictive models from the spectral data, Partial Least Squares (PLS) regression was applied. For each trait, a separate PLS model was trained using the spectral matrix $X \in \mathbb{R}^{n \times p}$ as predictors and a single response vector $y \in \mathbb{R}^n$. The method extracts latent components that are used directly for regression:

$$X = TP^\top + E, \quad y = Tq + f \tag{1}$$

where, T represents the latent scores, P are the loadings of X, and q is the regression weight for the response. Residuals are denoted by E and f. By fitting the model trait by trait, we ensured that each response variable was explained by the spectral information most relevant to it.

The number of latent variables is the most critical hyperparameter in PLS regression, too few components will under fit the data, while too many will lead to overfitting to noise. To find the optimal number, we applied k-fold cross-validation on the training set to decide on the optimal number of latent variables. We tested different cross-validation folds k=5, 10, and 15 to make sure the results are consistent with each other. Fig. 4 shows that the curve flattens after about 10 components for both $PRESS$ and q^2 for one chose trait LMA. A similar pattern is visible in Fig. 3 for all the traits and overall $PRESS$ and q^2. Combining these observations along with the observations from Fig. 5 we decided to go with **10** latent variables as the most optimal one.



(a) LMA Content

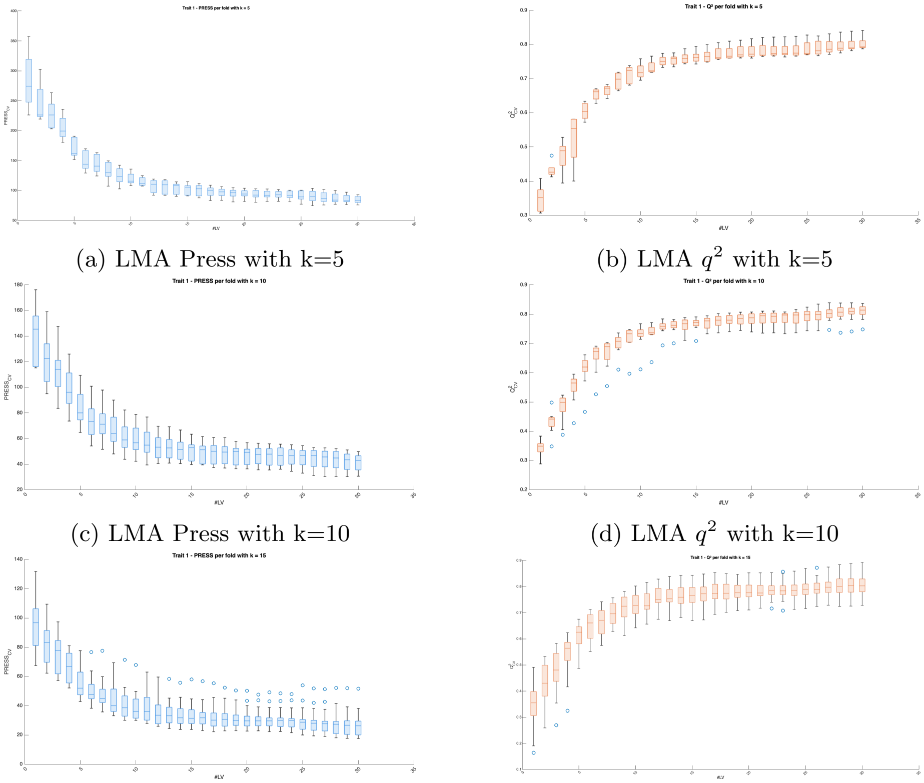
(b) N Content

(c) C Content

(d) Chl Content

(e) EWT Content

Figure 3: Cross-validation results for PRESS and Q^2 across different traits.



(a) LMA Press with k=5

(b) LMA q^2 with k=5

(c) LMA Press with k=10

(d) LMA q^2 with k=10

(e) LMA Press with k=15

(f) LMA q^2 with k=15

Figure 4: Cross-validation results for PRESS and Q^2 across different folds for LMA trait.

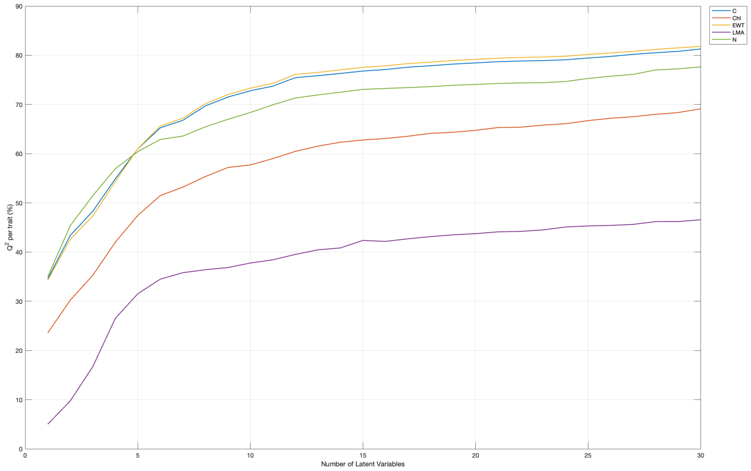


Figure 5: Per Trait Q^2

3 Performance Metrics

Model performance was evaluated using the independent test set (20% of the data). The Root Mean Squared Error of Prediction (RMSEP) was used as the main error metric:

$$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

Since the traits are measured on different scales, the RMSE was also normalized by the trait range to allow comparison across traits:

$$NRMSE = \frac{RMSE}{\max(y) - \min(y)} \tag{3}$$

Predictive ability was assessed using Q^2, computed on the test set:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_{\text{train}})^2} \tag{4}$$

where $\bar{y}_{\text{train}}$ is the mean of the training data. To aid interpretation, scatter plots of predicted vs. observed trait values were also used for visualization.

4 Results and Discussion

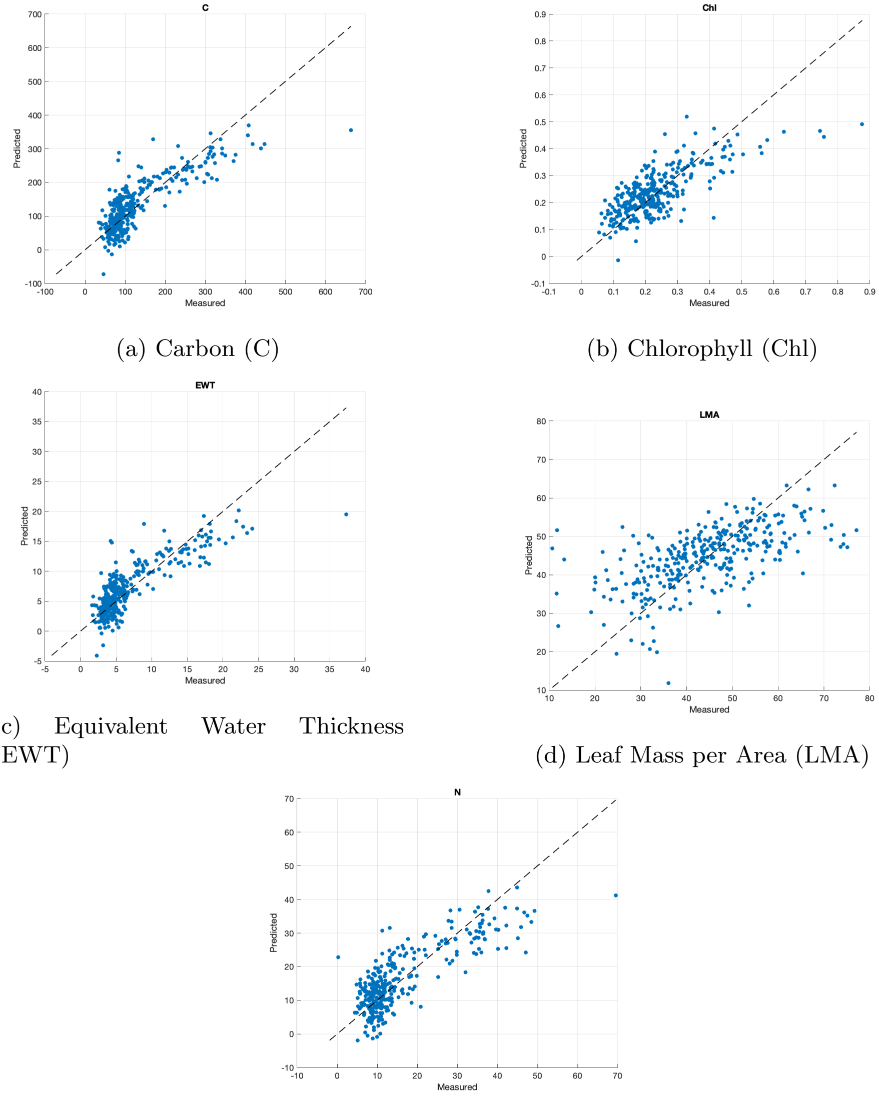
4.1 Predictive Performance of the PLS Models

Table 1 reports the predictive performance of each trait-specific PLS model. The Q^2_{test} values indicate strong predictive ability for Carbon (0.72), EWT (0.73), and Nitrogen (0.69). On the other hand, Chlorophyll (0.56) and particularly LMA (0.33) proved more difficult to estimate. This was expected, as earlier studies have shown that LMA is harder to predict because its spectral signal is weaker and less consistent than that of biochemical traits [1, 2]. The normalized RMSE (NRMSE) values range between 7% and 10% for all traits, demonstrating that once scale differences are removed, all traits have relatively similar relative error levels. The close matching of train and test errors further suggests that the models generalise well and do not suffer from significant overfitting.

Trait	Q^2_{test}	RMSE(Train)	RMSE(Test)	NRMSE(Train)	NRMSE(Test)
C	0.722	47.676	48.934	0.076	0.078
Chl	0.560	0.071	0.076	0.086	0.092
EWT	0.728	2.649	2.749	0.074	0.077
LMA	0.325	9.827	9.734	0.105	0.104
N	0.692	6.498	6.329	0.081	0.079

Table 1: Predictive performance of PLS models for each trait

Figure 6 shows that the scatter plots are in line with the numbers reported in Table 1. Carbon, EWT, and Nitrogen follow the 1:1 line fairly well, while Chlorophyll is a bit more spread out. LMA again comes out as the weakest, with a wide scatter around the line. The plots therefore confirm what we already saw in the table: some traits can be predicted reliably, others much less so.



(a) Carbon (C)

(b) Chlorophyll (Chl)

(c) Equivalent Water Thickness (EWT)

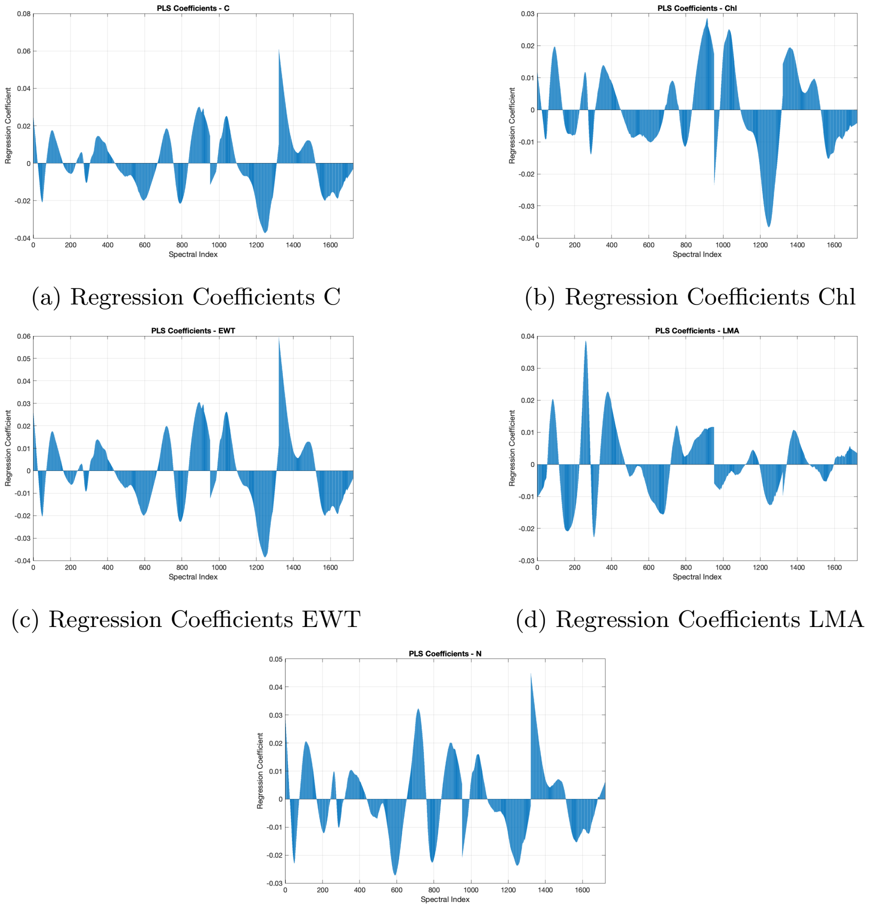
(d) Leaf Mass per Area (LMA)

(e) Nitrogen (N)

Figure 6: Scatter plots of predicted vs. measured traits

4.2 Regression Coefficients

The regression coefficients shown in Figure 7 demonstrate the relationships between spectral wavelengths and each vegetation property, where five subfigures (a–e) represent Carbon (C), Chlorophyll (Chl), EWT, LMA, and Nitrogen (N). The wavelength index is plotted on the x-axis, while the coefficient values are indicated on the y-axis. The highest peaks were considered the more significant wavelengths. The order of strengths of peaks was Carbon and EWT, which were considerably higher than LMA as indicated in the marginal Q^2 value found in Table 1, which touches on wavelength properties of each variable. Several traits have peaks nearly in the same wavelength regions, therefore having the same spectral sensitivity. The overlap between the information implies certain wavelength intervals have spectral for multiple vegetation traits.



(a) Regression Coefficients C

(b) Regression Coefficients Chl

(c) Regression Coefficients EWT

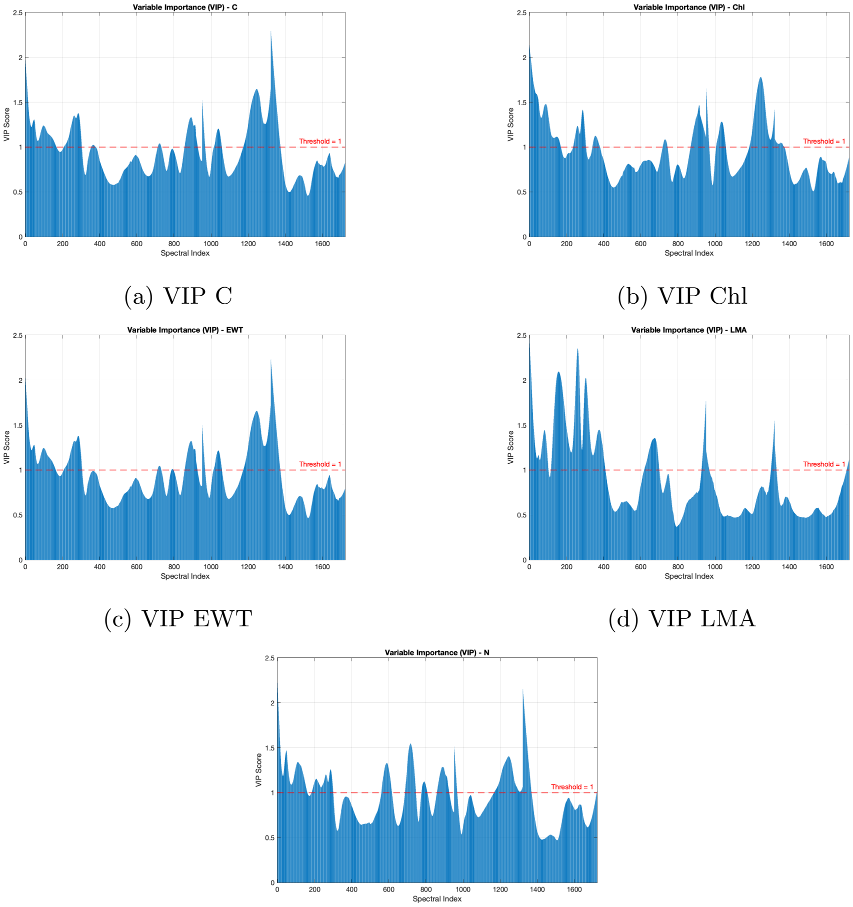
(d) Regression Coefficients LMA

(e) Regression Coefficients N

Figure 7: PLS regression coefficients across different traits.

4.3 Variable Importance

The VIP (Variable Importance in Projection) scores given in Figure 8 evaluate important wavelengths in the PLS models, containing five subfigures (a–e) for Carbon (C), Chlorophyll (Chl), EWT, LMA, and Nitrogen (N), plotting wavelength indices on the x-axis against VIP scores on the y-axis, with a red line representing significant predictors at 1.0. For both Carbon and EWT, the range of wavelengths above 1.0 occurs within the range of 700–1000 nm, where LMA has a smaller number which correlates to its Q^2 value of 0.33 in Table 1, indicating different spectral influences.



(a) VIP C

(b) VIP Chl

(c) VIP EWT

(d) VIP LMA

(e) VIP N

Figure 8: Variable importance plots across different traits.

4.4 Variable Selection and Refined Model Performance

The initial PLS regression models included all spectral variables, some of which were not important in the regression according to the VIP score, by filtering the wavelengths using $VIP > 1$ we were able to reduce the number of wavelengths from 1721 to 1013. The subset of wavelengths is shown in the Fig. 9.

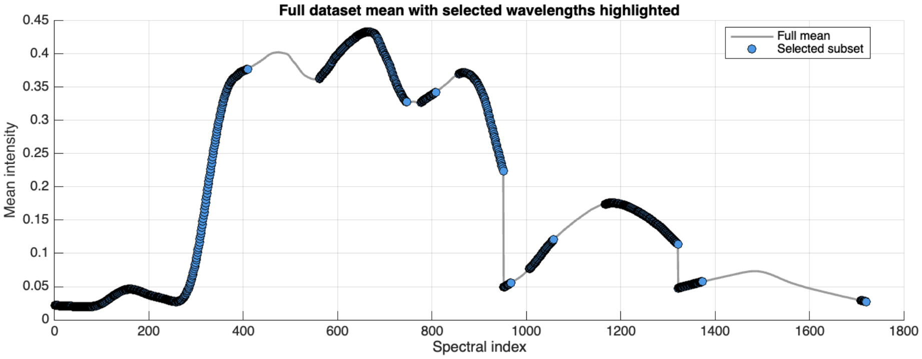


Figure 9: Mean wavelength value of full against subset data

The predictive performance of the reduced model was studied using using Q^2_{test}, RMSE(Train), RMSE(Test) for each trait. As shown in Table 2 the VIP selection process, while successful in reducing the number of spectral wavelengths, had a negligible overall effect on the predictive accuracy across most traits. The Q^2_{test} values were identical or showed only a minute decrease. This indicated that excluded wavelengths were non-informative and their exclusion did not impact the model's predictive power. Both training and test RMSE/NRMSE metrics decreased for LMA and N, confirming that the refined models for LMA and N were slightly more accurate.

Trait	Q^2_{test}	RMSE(Train)	RMSE(Test)	NRMSE(Train)	NRMSE(Test)
C	0.717	48.536	49.417	0.077	0.078
Chl	0.554	0.071	0.076	0.086	0.092
EWT	0.722	2.698	2.778	0.076	0.078
LMA	0.346	9.799	9.584	0.104	0.102
N	0.710	6.339	6.137	0.079	0.076

Table 2: Predictive performance of PLS models using subset of wavelenghts for each trait

5 Conclusion

The PLS regression models effectively predicted important vegetation characteristics. The predictions for Carbon, EWT, and Nitrogen showed strong Q^2 values of 0.72, 0.73, and 0.69, respectively. In contrast, the prediction of Chlorophyll and LMA was more challenging, yielding lower scores of 0.56 and 0.33, likely due to their weaker spectral signatures. By employing 10 latent variables selected through cross-validation and band reduction, the models achieved increased stability and accuracy, which was evidenced by the similarity of training and testing errors. These findings indicate that PLS regression is a valuable tool for forecasting various traits from hyperspectral data.

In this work, we built PLS regression models for vegetation traits and examined how well they capture information from hyperspectral data. Each trait was modeled separately, so five regression models were developed. The study also explored variable importance to identify which wavelength regions had the strongest effect on the predictions. The number of latent variables was set to ten based on the PRESS and Q^2 curves. In terms of model performance, Carbon, EWT, and Nitrogen were predicted quite well, while Chlorophyll and LMA have relatively low scores. For LMA, this was expected since its spectral signal is weaker. Most of the information came from the near-infrared and shortwave-infrared regions, whereas the visible range was mainly related to Chlorophyll. These findings suggest that PLS regression works well for linking spectral data with vegetation traits but still has limits.

References

[1] Serbin, S. P., Singh, A., Desai, A. R., Dubois, S. G., Jablonski, A. D., Kingdon, C. C., Kruger, E. L., & Townsend, P. A. (2014). Spectroscopic determination of leaf morphological and biochemical traits for northern temperate and boreal tree species. *New Phytologist*, 206(1), 129–139. doi:10.1111/nph.12886

[2] Wang, Y., Zhu, Y., Zhao, Y., Zheng, B., & Xu, S. (2021). Retrieval of leaf mass per area (LMA) using PROSPECT model inversion with different spectral bands. *Remote Sensing*, 13(18), 3761. doi:10.3390/rs13183761