# PLS Regression-Based Soft Sensor Models for Multi-Trait Prediction from Spectral Data

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. In Materials and Methods, the dataset and preprocessing steps are introduced, followed by the mathematical background of PLS regression. Results and Discussion then presents the predictive performance of the models and interprets the outcomes. Finally, Conclusion provides a short summary of the findings and their implications.

## 2 Materials and Methods

### 2.1 Materials and Methods

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.

## 2.2 Data Preperation

Since we wanted to use the same training and testing set for all traits, only 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. As a result, we had 2163 samples for the train set and 539 samples for the test set.

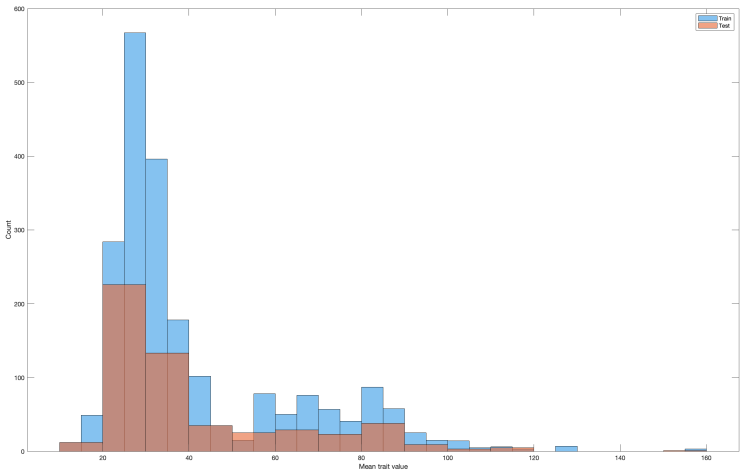



Figure 1: Distribution of Mean Trait Values in Train and Test Sets

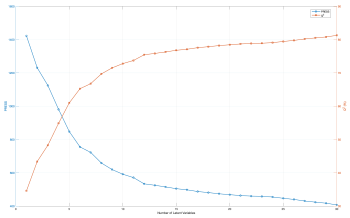
## 2.3 PLS Regression

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:

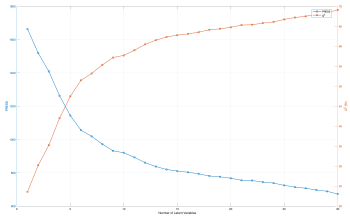
$$X = TP^\top + E, \quad y = Tq + f$$

Here, $T$ 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.

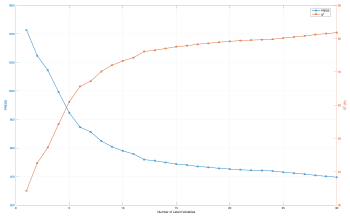
We tested different cross-validation folds (k=5, 10, 15) to decide the number of latent variables. Figure 2 shows that the error curve flattens after about 10 components. A similar pattern is visible in Figure 3 for explained variance. For this reason, 10 latent variables were used in the models.



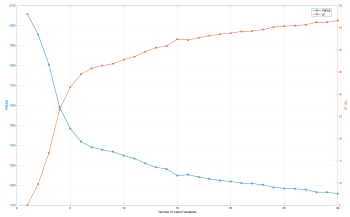
(a) C Content



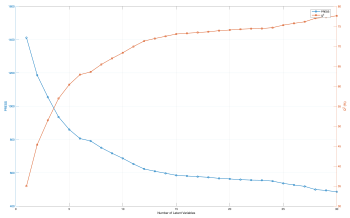
(b) Chl Content



(c) EWT Content



(d) LMA Content



(e) N Content

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

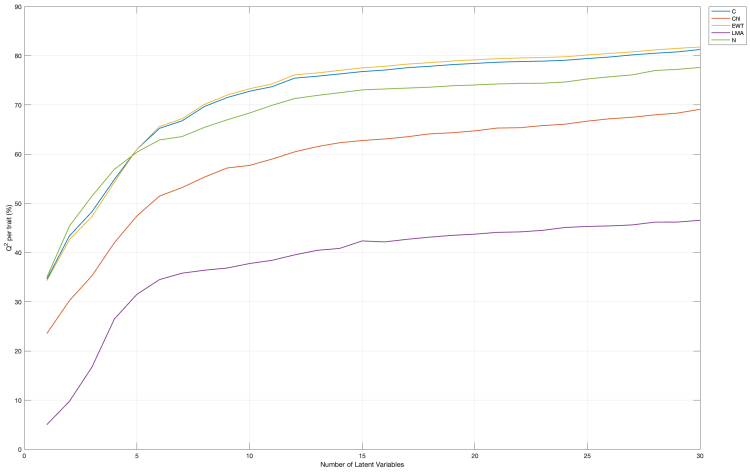



Figure 3: Per Trait $Q^2$

# 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:

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

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

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

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

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

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

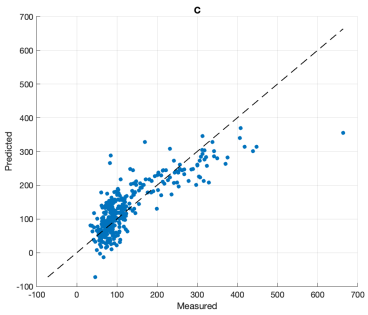
## 4.1 Predictive Performance of the PLS Models

Table 1 reports the predictive performance of each trait-specific PLS model. The $Q^2_{\text{test}}$ 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.

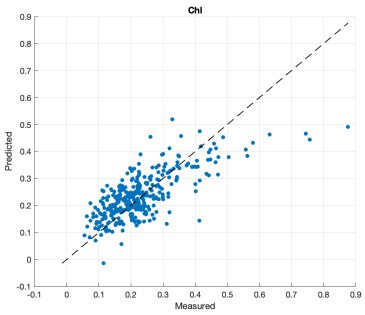
| Trait | $Q^2_{\text{test}}$ | RMSE(Train) | RMSE(Test) | NRMSE(Train) | NRMSE(Test) |
|---|---|---|---|---|---|
| C | 0.722 | 47.676 | 48.934 | 0.076 | 0.078 |
| Chl | 0.560 | 0.071 | 0.076 | 0.086 | 0.092 |
| EWT | 0.728 | 2.649 | 2.749 | 0.074 | 0.077 |
| LMA | 0.325 | 9.827 | 9.734 | 0.105 | 0.104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

Table 1: Predictive performance of PLS models for each trait

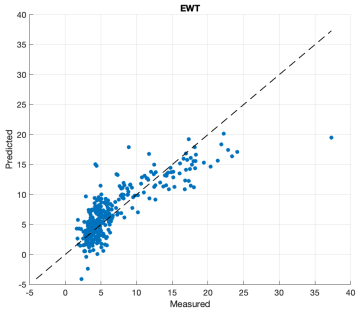
Figure 4 shows that the scatter plots are in line with the numbers reported in Table 1. Carbon, EWT, and Nitrogen follow the 1:1 line fairly well, while Chlorophyll is a bit more spread out. LMA again comes out as the weakest, with a wide scatter around the line. The plots therefore confirm what we already saw in the table: some traits can be predicted reliably, others much less so.



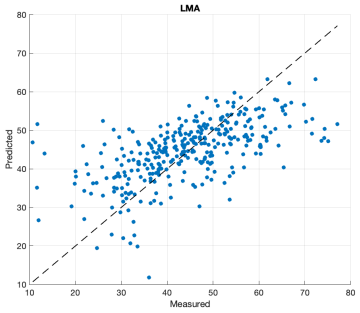

(a) Carbon (C)



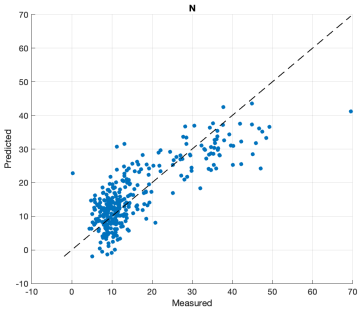

(b) Chlorophyll (Chl)




(c) Equivalent Water Thickness (EWT)




(d) Leaf Mass per Area (LMA)




(e) Nitrogen (N)

Figure 4: Scatter plots of predicted vs. measured traits

### 4.2 Regression Coefficients

## 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). Spectro-

scopic 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