Project Part 3: PLS Regression Results Report
**Students:** Bahadır Bagci, Hassan Majeed, Zoheb Rahman
**Team Name:** Spectral Data Soft Sensor (B)

## 1. Introduction

In the earlier submissions of report, we discussed about the data exploration techniques, Principal Component Analysis, data preprocessing steps including data filtering, data standardisation, data modelling steps. This submission will discuss about the results of the PLS regression run to predict five traits: Leaf Mass per Area ($g/m^2$), Nitrogen Content ($mg/cm^2$), Carbon Content ($mg/cm^2$), Chlorophyll Content ($\mu g/m^2$), Equivalent Water Thickness ($mg/cm^2$), all of which were chosen based on number of lowest missing values.

## 2. Modelling and Results

### 2.1. Data Splitting

The data was split using Stratified Sampling which ensures that both the training and testing dataset contains the same distirbution as of mean trait value (K. G. & Khan, 2023). Figure 1. Shows that both training and testing samples follow the same distribution to ensure the model generalises well and there is no bias towards one particiular group.

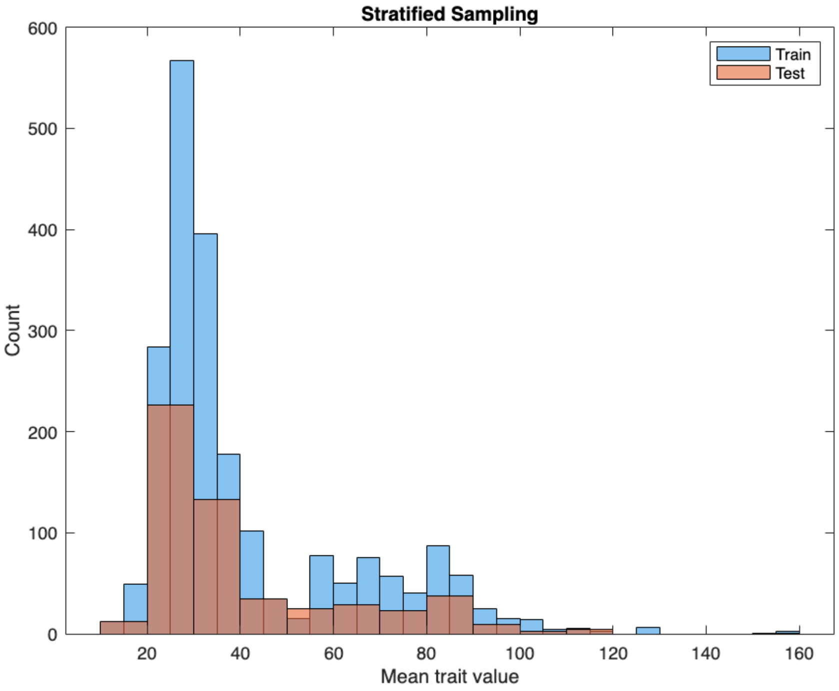


Figure 1

### 2.2. Cross Validation

We tested against different cross validation folds to observe if the results behave in similar manner or not. It can be seen from Figure 2. That for k=5, k=10, k=15 all of them point to similar outcome i.e. the model has an elbow point at around latent variable 10 after which the PRESS and $Q^2$ becomes stable.

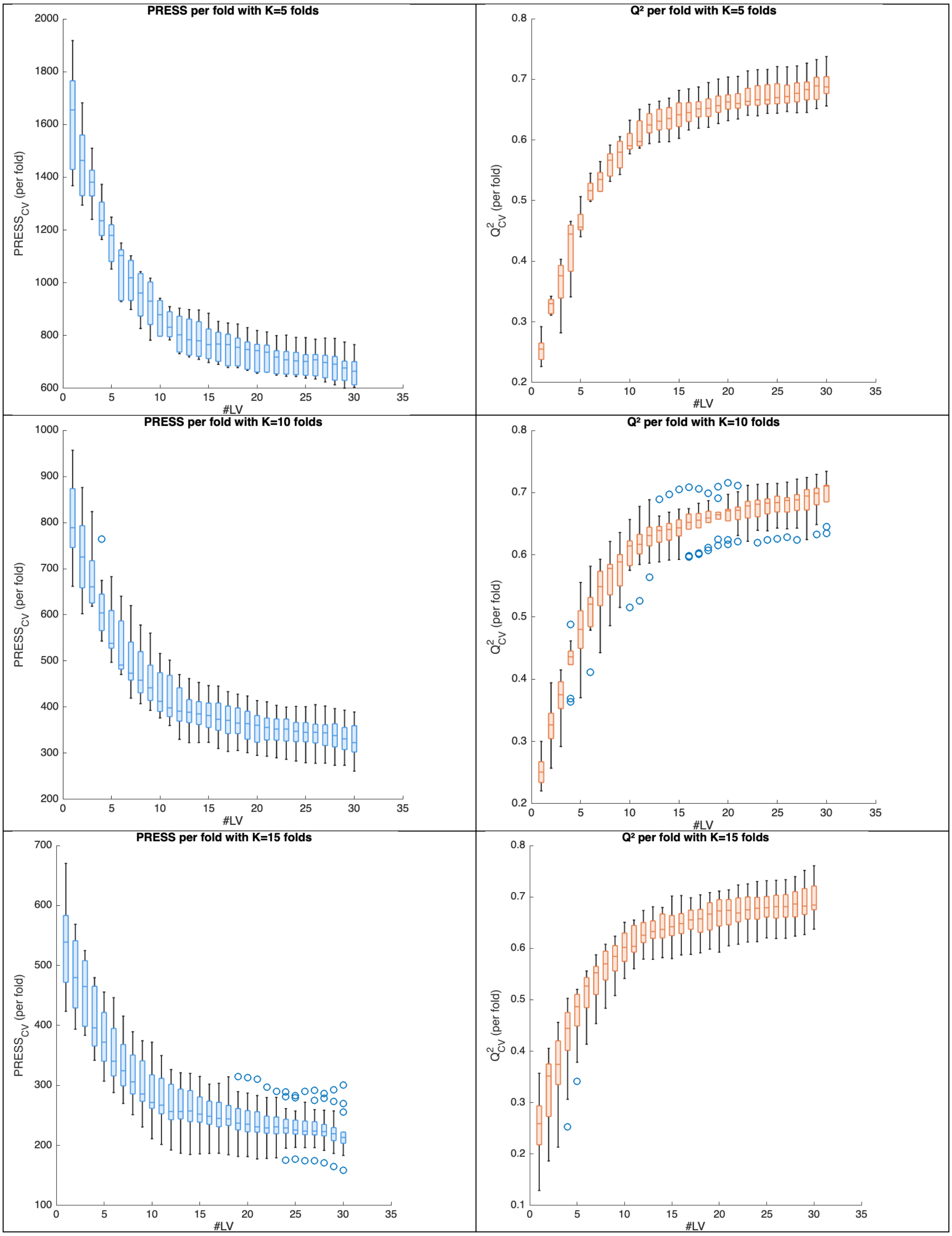


Figure 2.

If we take a look at Figure 3. Which shows the predictive power of latent variable in explaining each trait, it increases sharply until 10 after which the change is gradual. We can also note that the model struggles in predicting Chlorine Content than predicting other traits one of the reasons could be since the measurement scale is in ($\mu g/m^2$), experiemental errors can greatly impact the predictive performance.

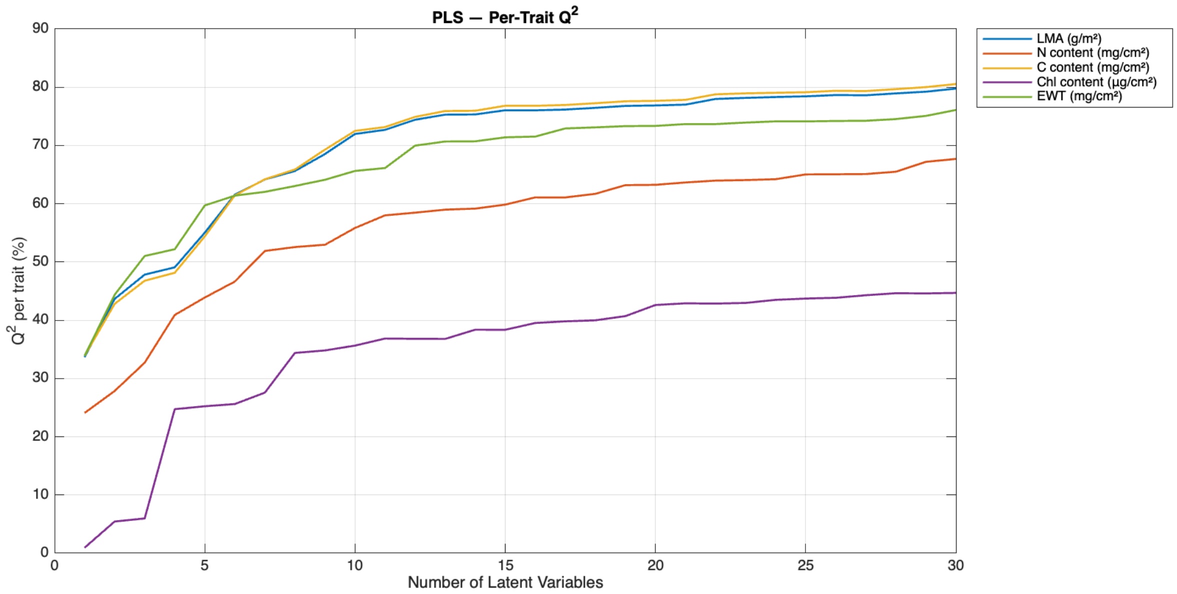
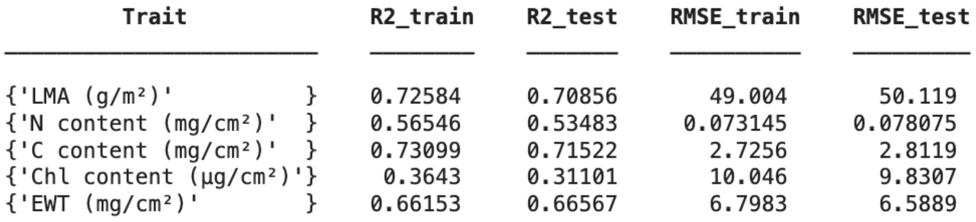


Figure 3.

## 2.3. Testing PLS Model

We tested the model with 10 LV and from Table 1. it can be seen that the model generalises good and there is no severe overfitting with RMSE test and train being close to each other. Traits like LMA, Carbon Content, EWT are predicted well where as the model struggles a little with Nitrogen Content and it performs the worst for Chlorine Content. This can further be corroborated by Figure 4. which shows the predictions vs actual measurements with points having linear relationship with predicted and measured for Carbon, LMA, but they get more scatterd for the rest of the plots.



```
          Trait              R2_train    R2_test    RMSE_train    RMSE_test
    _______________________  ________    _______    __________    _________

    {'LMA (g/m²)'          }  0.72584    0.70856      49.004        50.119
    {'N content (mg/cm²)'  }  0.56546    0.53483     0.073145     0.078075
    {'C content (mg/cm²)'  }  0.73099    0.71522      2.7256        2.8119
    {'Chl content (µg/cm²)'}   0.3643    0.31101      10.046        9.8307
    {'EWT (mg/cm²)'        }  0.66153    0.66567      6.7983        6.5889
```

Table 1.

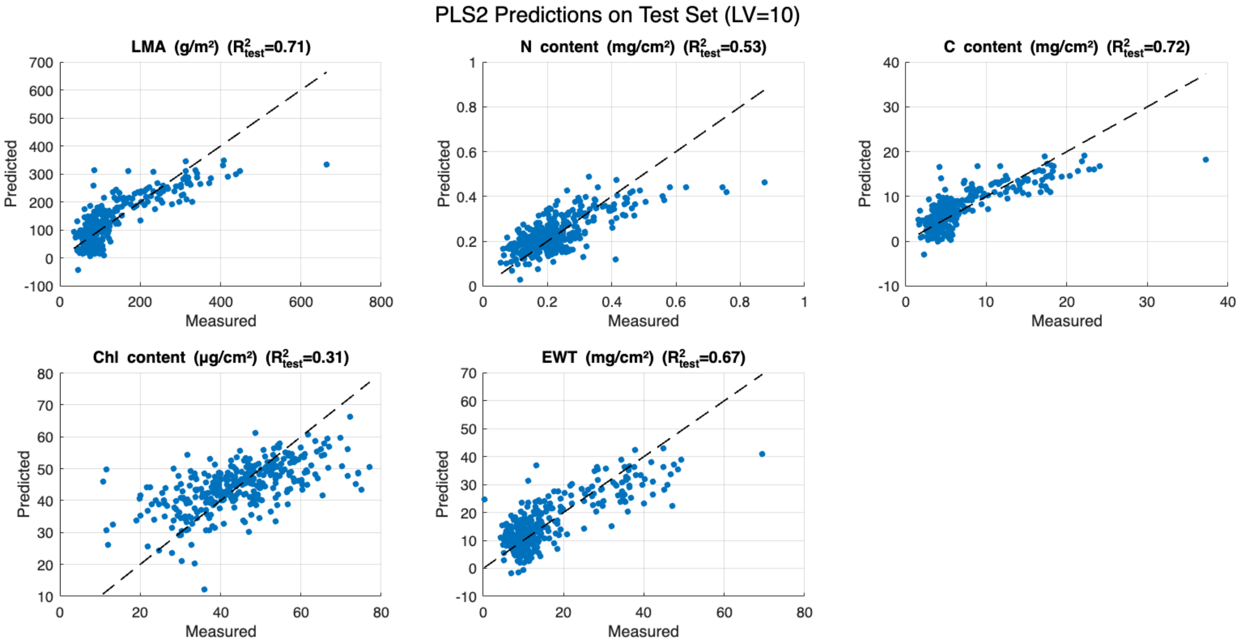


Figure 4.

We now look at the normalised PLS coefficients and from this we can see that most of the coefficiants share similar patterns overlapping spectral information in NIR regions. The Water Thickness shows very large coeeficiants at 1200-1400 index that is 1600-2000nm which corresponds to water absorbtion band (Salinas, Reichel, & Witte, 2021), the Chlorophyll content behaves differently showing peaks in the visible range i.e. 450-700nm (0-300 band index).

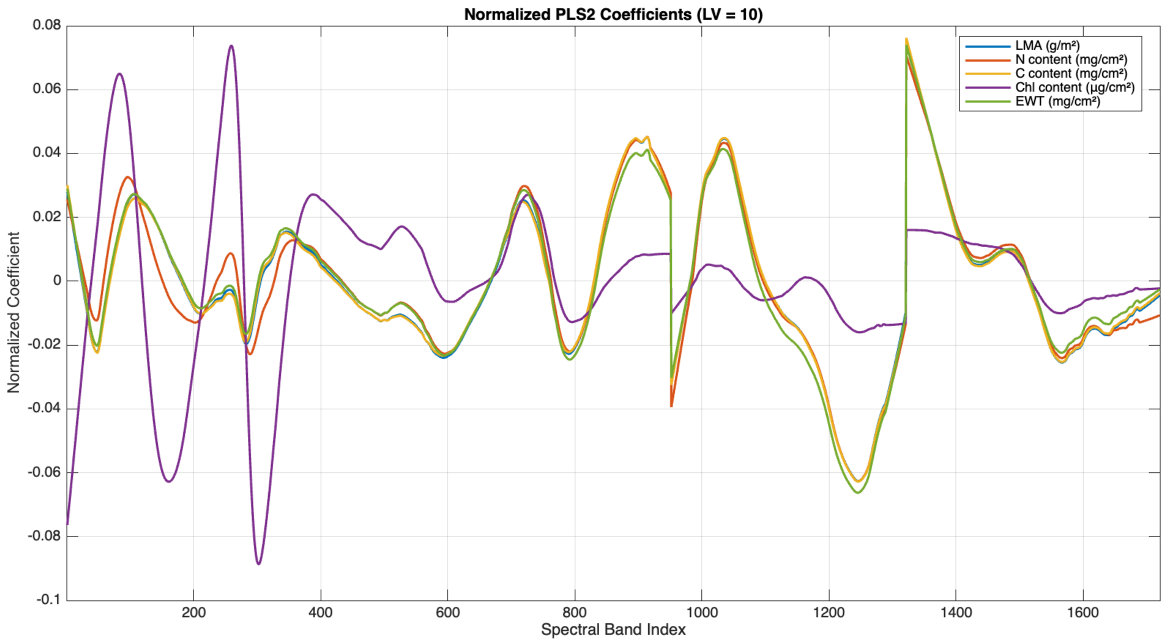


Figure 5.

The PLS2 Scores shown in Figure 6. Indicates that the dataset is largely homogeneous, with most of the samples form a cluster at the origin with a few outliers. LV1 and LV2 captured a high percentage of the variance which confirms the utility of the chosen LV in representing the underlying structure of spectra.

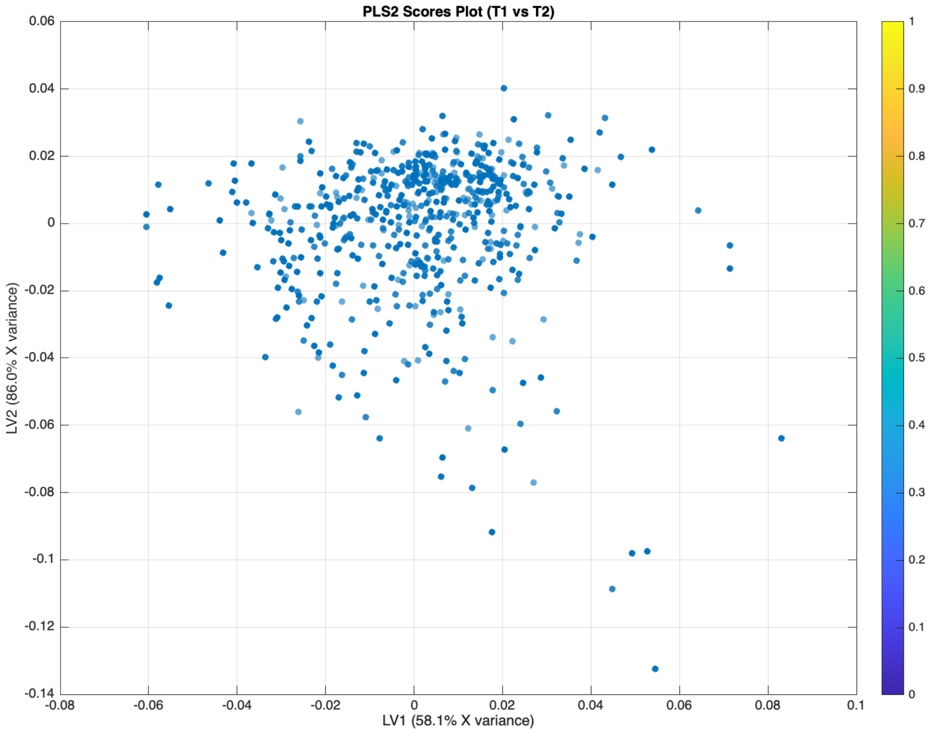


Figure 6.

From Figure 7. which shows the PLS2 Loadings and normalised loadings it can seen that Chl is strongly seperated in the positive LV2 space which indicates that the information unique to chlorphyll is captured by LV2 and is essentially uncorrelated to other traits which can also be seen from Figure 5. Nitrogen content, Carbon content, LMA, EWT are tightly clustered in the negative LV1/LV2 space confirming their strong dependence on structural and water-related absorbtion features.

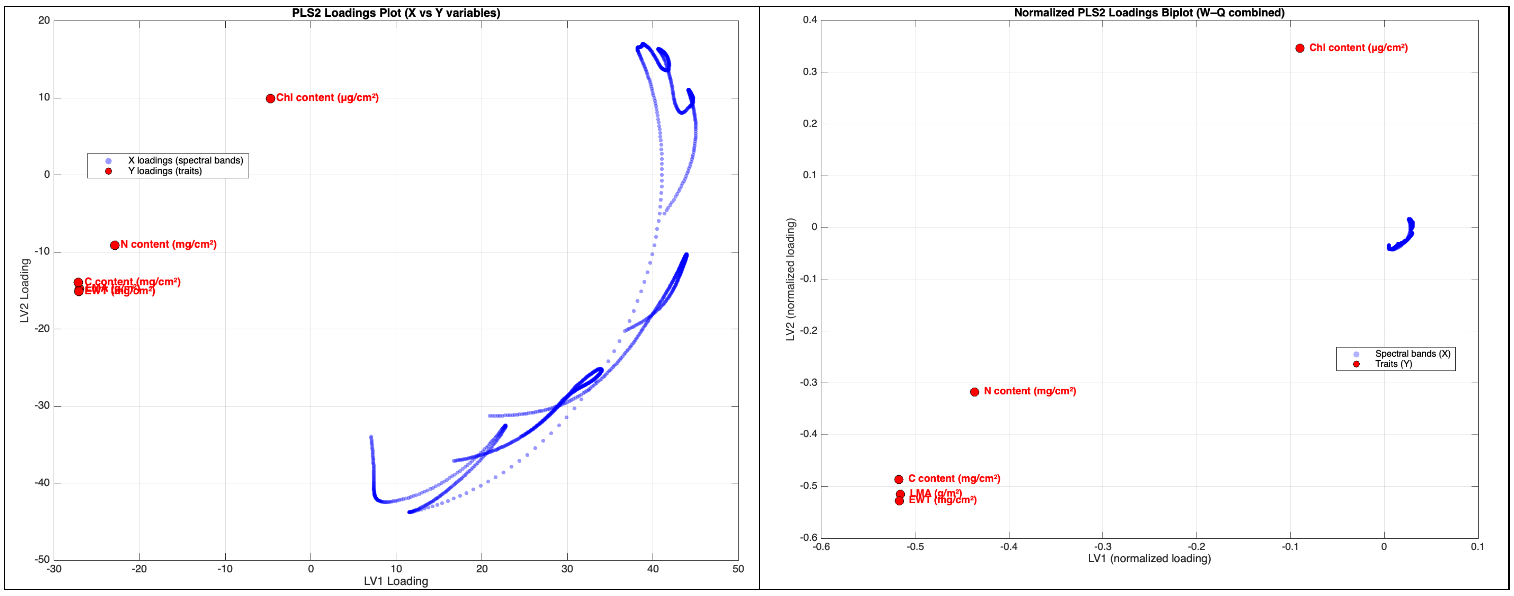


Figure 7.

The VIP (Variable Importance Plot Figure 8.) clearly identifies the region at index 950 as most critical which is also supported by the normalised PLS loading graph from Figure 7. which strongly seperates the Chl content from the cluster of structural/water traits.

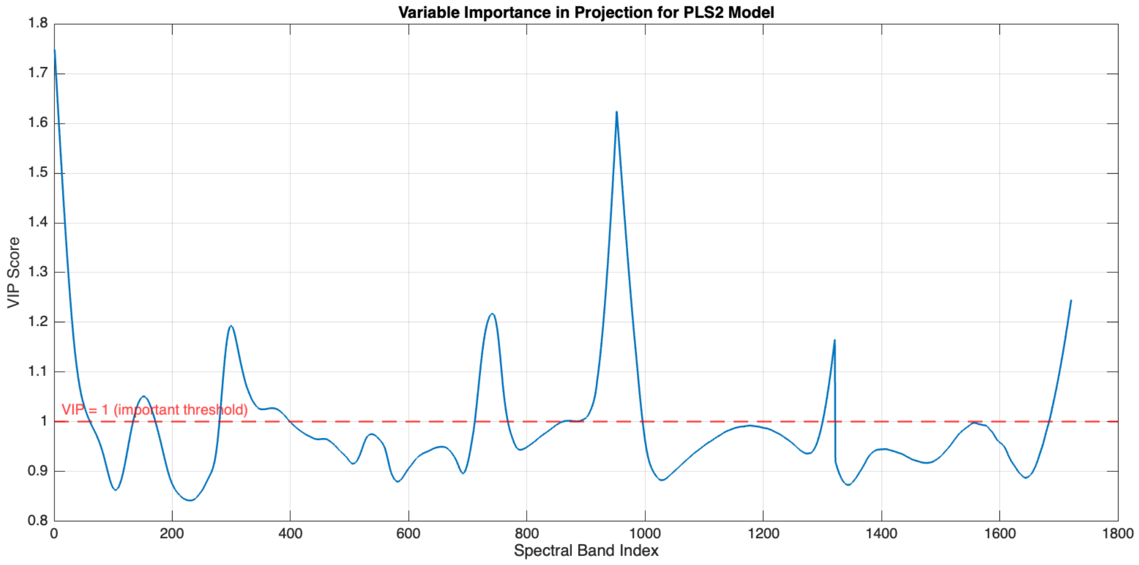


Figure 8.

# Bibliography

K. G., R., & Khan, M. (2023, October 12). Constructing efficient strata boundaries in stratified sampling using survey cost. *Heliyon*, e21407. From National Institutes of Health: https://pmc.ncbi.nlm.nih.gov/articles/PMC10641212/

Salinas, C. M., Reichel, E., & Witte, R. S. (2021). Short-wave Infrared Photoacoustic Spectroscopy for Lipid and Water Detection. *2021 IEEE International Ultrasonics Symposium (IUS)* (pp. 1-4). Xi'an, China: IEEE.