Project Part 2.2: Modelling Plan

Students: Bahadir Bagci, Hassan Majeed, Zoheb Rahman

Team Name: Spectral Data Soft Sensor (B)

1. Introduction

In the last part, we completed initial data processing and number of principal component selection steps. From now on, we will be working on developing a model in order to predict selected five plant traits: carbon, chlorophyll, equivalent water thickness, leaf mass per area, and nitrogen, those were chosen because they have fewer missing values compared to the others.

2. Methodology

2.1. Tools

Our aim is not just to reduce the number of predictors, but also to ensure that the variation in input variables X is strongly linked with the variation in response variables Y. Therefore, we apply Partial Least Squares (PLS) regression method[i] implemented via MATLAB’s stat toolbox. PLS is supervised method that is well suited for situations with high collinearity among input variables, such as 2200 spectra used here.[ii]

2.2 Model Calibration Strategy

Based on the nature of this dataset, missing trait values do not imply that the corresponding traits are absent or zero. Rather, they indicate that those traits were not measured for that particular sample. Since even the least missing trait has over 30% missing values, we did not apply imputation. Filling in such a large amount of missing data without clear patterns could distort the relationships the model tries to learn. Instead, for each selected trait, we filtered out the rows where the target value is missing. We made sure to use only the real, available values for each trait, instead of filling in missing ones. First of all, we split the dataset into training and test sets (70%-30%) and also perform z-score standardization to get 0 mean and 1 std. Then, for generalization, we check how well the models perform during training. In order to do that, we use 5-fold cross validation on the training set. In brief, this method splits the data into five parts, trains the model on four parts, and validates it on the remaining part—repeating this process five times to help us understand the consistency of the models.

2.3 Model Validation Strategy

We perform the validation to ensure that our model is neither overfitting nor underfitting. In other terms, it generalizes well in learning the relationship between spectra and plant traits. This also helps us determine the right number of Latent Variables, which are primarily responsible for overfitting and underfitting. We perform 5-fold cross-validation using the training samples. The training data is randomly divided into five equal parts. In each iteration, four parts are used for training and the remaining part for validation. This process is repeated five times, with a different subset used for validation each time. Using the coefficient of determination (Q^2) we can know about the fitness of our model.

The cross-validated coefficient of determination, Q^2, measures the model's predictive relevance and is calculated as:

$$\mathbf{Q^2} = \mathbf{1} - \frac{\boldsymbol{PRESS}}{\boldsymbol{TSS}}$$

where:

- **PRESS** (Predictive Residual Error Sum of Squares) $=\sum(\mathbf{y_i} - \hat{\mathrm{y}}_{\mathrm{i}})^2$

  where **y** as the observed vector and $\hat{\mathbf{y}}$ as the predicted vector.

- **TSS** (Total Sum of Squares) is the total variance in the observed **Y** data:

$$\mathrm{TSS} = \sum(\mathbf{y_i} - \bar{\mathrm{y}})^2$$

  where $\bar{\mathrm{y}}$ is the mean of **y**.

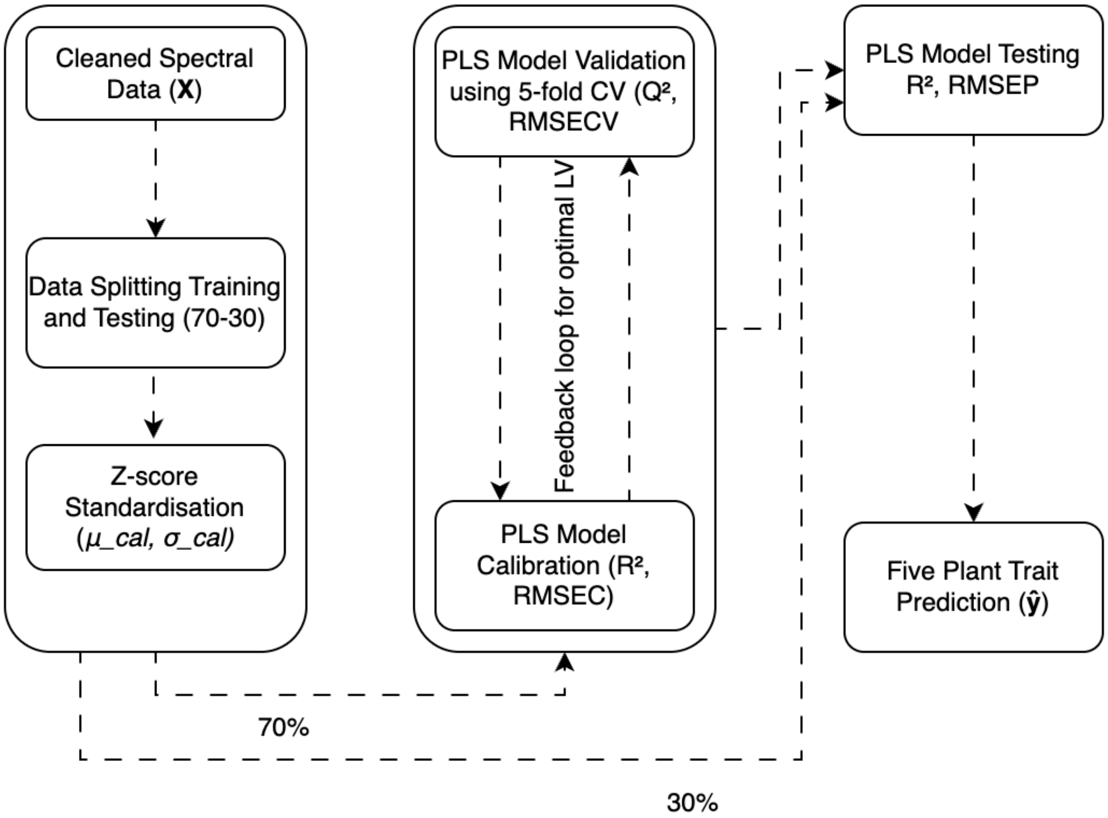
A Q^2 value close to 1 indicates strong predictive power, while $Q^2 < 0$ suggests the model predicts worse than the mean. We use Q^2 to select the optimal number of latent variables ***(k)*** during 5-fold cross-validation, balancing model complexity to prevent overfitting or underfitting.

2.4 Model Testing Strategy

We use the 30% of the dataset that was left out and not seen by the model to test the performance of our model. For benchmarking we compute the RMSEP (Root Mean Squared

Error of Predictions) since this is a regression model. We also compute the R^2 to learn how well the spectra are explaining the variance. To visualize the predictions scatter plots are used to plot predictions vs actual value

2.5 Operations Flowchart


Cleaned Spectral Data (X)
Data Splitting Training and Testing (70-30)
Z-score Standardisation (μ_cal, σ_cal)
PLS Model Validation using 5-fold CV (Q², RMSECV
Feedback loop for optimal LV
PLS Model Calibration (R², RMSEC)
PLS Model Testing R², RMSEP
Five Plant Trait Prediction (ŷ)
70%
30%

[i] https://www.mathworks.com/help/stats/plsregress.html
[ii] https://arxiv.org/pdf/2409.05713