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.

1.1.Tools and Path

We will be using libraries scikit-learn(for PLS regression) and matplotlib(for visualization) in Python

1.2.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 not overfitting nor underfitting, in other term it generalizes well in learning the relationship between spectra and plant traits, it also helps us pick the right number of Latent Variables which mainly responsible for overfitting and underfitting. We perform 5-fold cross validation using the training samples. The idea is to split the training data into 5 random parts, with training being done on 4 parts and validation on 5-1. This process then gets repeated 5 times, each time using a different subset for validation. Using the coefficient of determination (R^2) we can know about the fitness of our model.

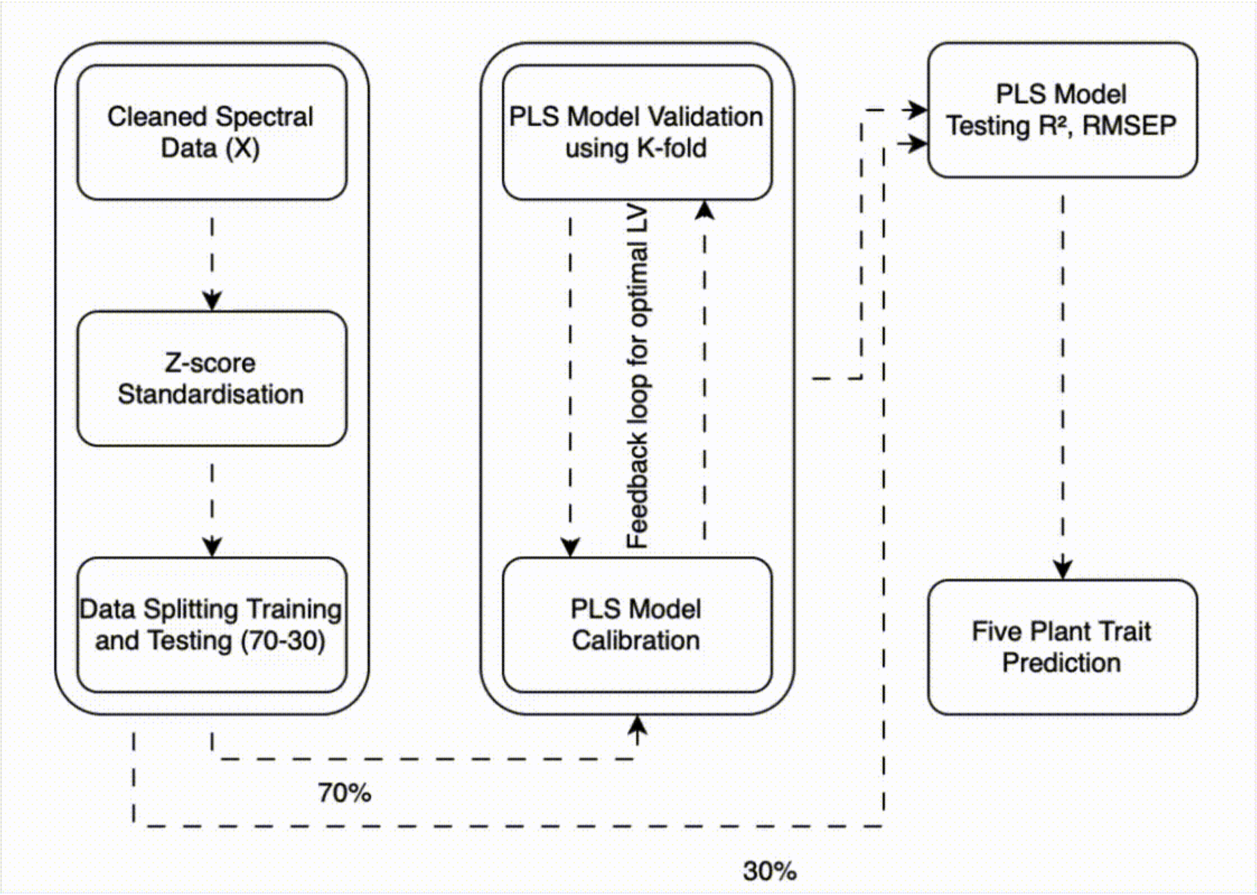
2.4 Model Testing Strategy

We use the 30% of the dataset that was left out and not seen by the model to learn 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 values.

2.5 Mathematical Methods

The data is three-dimensional, with spectral wavelengths (X), pixels (samples), and plant traits (Y). Our aim is not just to reduce the number of predictors, but also to ensure that the variation in X is linked with the variation in Y. Hence, Partial Least Squares (PLS) regression is applied as it is supervised method and takes into account both the spectral wavelengths and plant traits, furthermore, the method is preferrable in cases where we have high collinearity between input variables (spectras).[1]

2.6 Operations Flowchart


Cleaned Spectral Data (X)
Z-score Standardisation
Data Splitting Training and Testing (70-30)
PLS Model Validation using K-fold
Feedback loop for optimal LV
PLS Model Calibration
PLS Model Testing R², RMSEP
Five Plant Trait Prediction
70%
30%

[1] https://arxiv.org/pdf/2409.05713