Project Part 2.1: Refined Exploratory Analysis and PCA

Students: Bahadir Bagci, Hassan Majeed, Zoheb Rahman

Team Name: Spectral Data Soft Sensor (B)

1. Introduction

Data cleaning, initial exploration, and dimensionality reduction are essential steps in any data analysis process. In this part , we applied these steps on a spectral dataset, where the goal is to predict vegetation traits from hyperspectral band values. These will help us building trait prediction models based on spectral information.

2. Methodology

For the communication and code sharing, we decided on using Microsoft Teams and Github.

The dataset was compiled from 42 studies conducted across different regions, climates, and vegetation types. It combines hyperspectral reflectance data with vegetation trait measurements.

In total, it contains 12,180 observations and 1,741 variables: the first 20 are plant traits, and the rest are spectral bands ranging from 400 to 2450 nm — although the official description mentions 450 to 2500 nm.

Bands were preprocessed. No spectral bands were fully missing, so all were retained. Based on missing values, 5 traits were selected for modeling: C, Chl, EWT, LMA, and N content.

An initial check included trait distribution plots and visualizing the first 50 samples by converting selected bands into RGB colors. Afterwards, the spectral data was standardized, and PCA was applied.

We examined explained variance, visualized the loadings, and identified the wavelengths that contributed most to the first few principal components.

3. Results

Figure 1 shows the correlation matrix of wavelengths, which highlights strong multicollinearity across large regions of the wavelengths indicating that PCA is an appropriate method for this dataset. From Figure 2, we can observe that ~90% of variance is captured by the first two components, meaning that only two components are sufficient to represent the majority of the data.

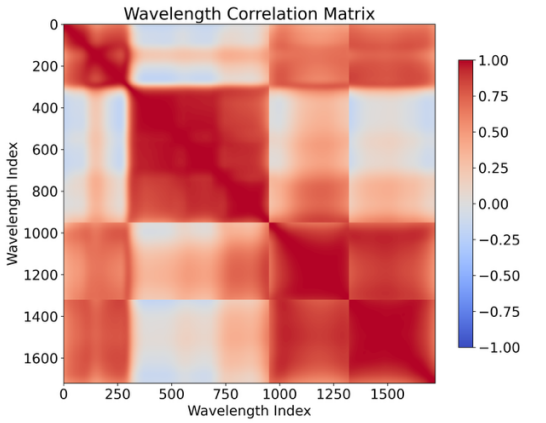

Figure 1: Wavelength Correlation Matrix

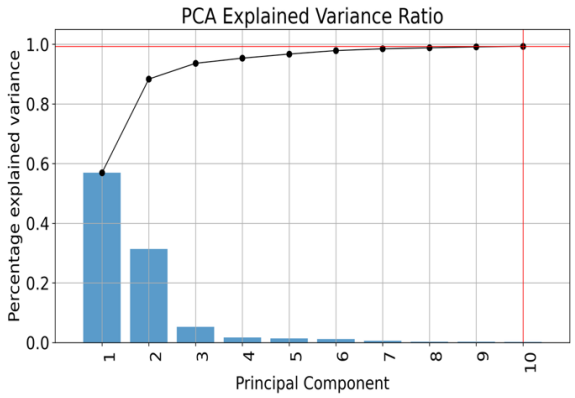

Figure 2: PCA Explained Variance Ratio

The biplots in Figure 3 visualises the distribution of observations and their loading along PC1 and PC2. The 1746nm wavelength can be seen to be highly correlated with PC1 as it has positive sign, and is almost horizontal, while 869nm wavelength shows high correlation with PC2 as it is almost vertical. The bar plot in Figure 4 further illustrates the loadings with respect to PC1 and PC2. In particular, the wavelength region 800-1300 nm has strong correlation with both PC1 and PC2 making it a strong candidate for selection in downstream analysis and representation.

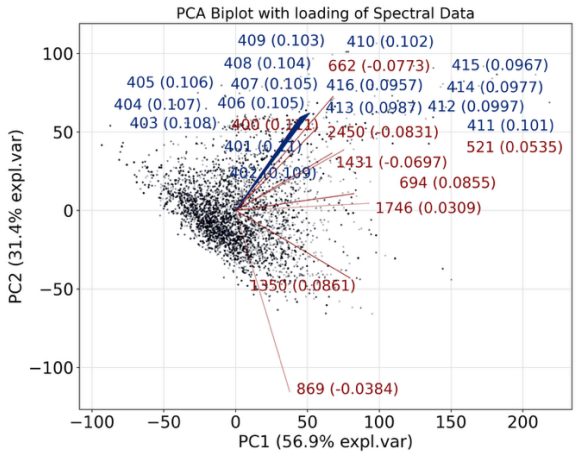

Figure 3: Biplot

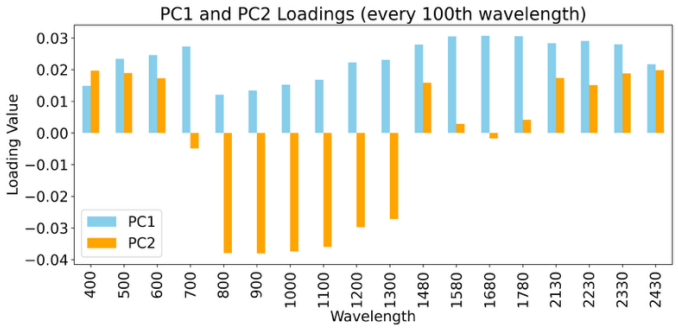

Figure 4: PC1 and PC2 Loadings

4. Pretreatment Steps and Plan

First, we explored the dataset and applied initial cleaning steps. Then, we selected five traits with the least amount of missing data and standardized the spectral variables. Afterwards, PCA was used to examine relationships between bands and identify the most informative wavelengths. The aim going forward is to build a separate regression model for each selected trait, using the spectral data as predictors.