Spectral Data Soft Sensor (A2) - Advanced Data Analysis and Machine Learning

Intermediary Report 1: Data Exploration, Transformation and Visualization
Team Members: Théo Deniel, Bohao Xing, Arbind Yadav

Communication Channel and code sharing strategy

We set up a Microsoft Teams chat for communication and a GitHub repository for code sharing and collaboration. Meetings and a WhatsApp group were used to coordinate plans, distribute workload, and track progress.

Data Overview and Preliminary Visualization

The goal of this project is to develop and evaluate prediction models, and analyze why some perform better than others. We use the “Hyperspectral Soft Sensor” dataset, which includes 12,180 observations and 1,741 variables. These are divided into two sets: (1) Hyperspectral data comprising 1,721 wavelengths (400–2450 nm, 1 nm interval, with water absorption bands removed), and (2) 20 vegetation traits, such as carbon content and equivalent water thickness, which serve as the target variables for prediction. The dataset does not present any temporal aspect. Therefore, we assume that we do not deal with time series.

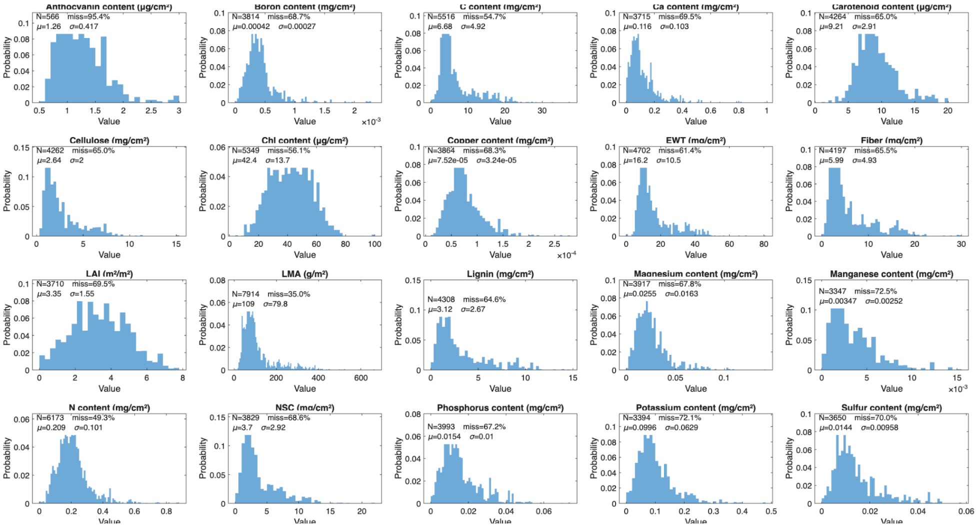

Figure 1: Distribution of response variables

Figure 1 shows the distribution of the 20 traits. Most traits are right-skewed, while a few are closer to normal. Missing rates vary strongly, from ~35% to over 90%. The dataset is not uniform. Some traits are skewed and many have missing values, so pretreatment is needed.

Figure 2 shows the mean reflectance spectrum with ±1 standard deviation. A sharp rise near 700 nm marks the vegetation red edge. High reflectance appears in the NIR (800–1300 nm), while strong drops at 1400 nm and 1900 nm correspond to water absorption. Variability is largest in the NIR regions. As you can see, the spectrum is continuous even considering the removal of water absorption bands, plotted in MATLAB. This is for visualisation convenience: we do not consider interpolation during any analysis computation.

Figure 3 shows the Pearson correlation matrix among spectral bands. Neighboring bands are highly correlated (mostly >0.9), forming large yellow blocks that reveal strong redundancy in hyperspectral data. This supports the need for dimensionality reduction (e.g., PCA or band selection).

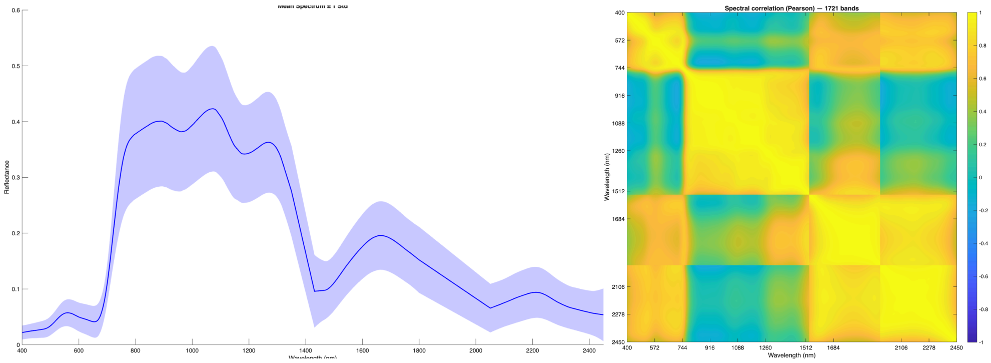

Figure 2&3 : Wavelength means (±1 SD) and Correlation between variables

Exploratory Data Analysis with PCA

Figure 4 shows the explained variance of the first 20 principal components and their cumulative contribution. PC1 alone explains about 56% of the variance, and PC2 adds about 31%, so together they explain nearly 90%. With PC3&PC4, the cumulative variance exceeds 95%. This indicates that only a few principal components are needed to capture most of the spectral variability.

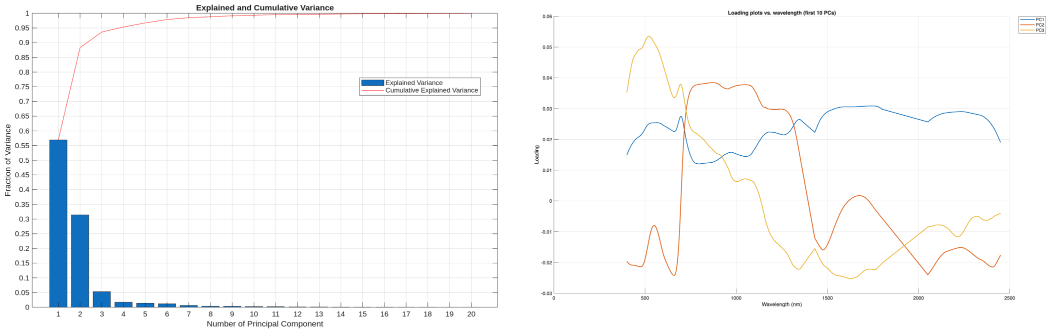

Figure 4&5: Cumulative explained Variance and Loadings of 3 first Principal Components

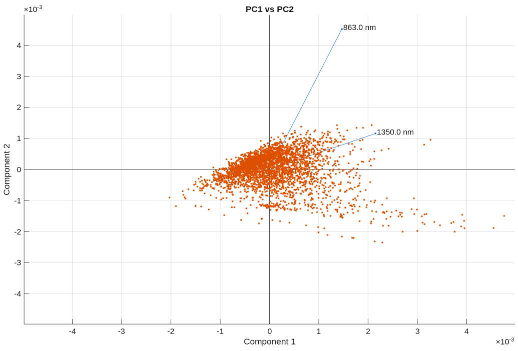

Figure 5 shows the loadings of the first three principal components. PC1 is smooth and contributes across the full spectrum, representing overall reflectance level. PC2 and PC3 have stronger localized peaks and sign changes, suggesting links to leaf color features.

Figure 6 shows the biplot of PC1 vs PC2. Most samples cluster near the origin, indicating that the first two PCs capture the main spectral variation. Additionally, the two bands with the largest loadings (863nm & 1350nm) are displayed, showing the directions where these wavelengths contribute most strongly to PC1 and PC2.

Data Pretreatment

The main pretreatment steps include checking and imputing missing values (e.g., interpolation), standardizing the spectral data (e.g., z-scores), identifying outliers with PCA statistics (e.g., T² and SPE), and reducing redundancy through PCA or band selection.