Advanced Data Analysis Project

Spectral Dataset

Week 3

LUT University

Alejandro Lopez Vargas
Sina Saeedi
Pau Ventura Rodríguez

Feedback incorporation

We have a dataset containing 20 traits, and our objective is to predict each of these traits using wavelength data. To achieve this, we can create a list of 20 vectors $y=[y_1, y_2, ..., y_{20}]$, where each vector represents the values of one trait without any missing data. Additionally, for each y_i, we will have a corresponding matrix X_i , which serves as the dataset for predicting y_i. Therefore, we will end up with a list of matrices $X=[X_1, X_2, ..., X_{20}]$.

Next, we need to divide these datasets into training, validation, and test data. Essentially, we will have 20 distinct sets for training, validation, and testing.

Finally, it's essential to normalize each dataset (X_i) based on its respective mean and standard deviation. This normalization process will be performed separately for each X_i.

For this week, we have decided to split and normalize the data as a whole dataset and also as a list of datasets for each trait.

Calibration, validation and test partitions

We have decided to split the dataset into 3: Train, which will be used to feed the models, Validation, which will be used to fine-tune the models and Test, which will be used to check if the models perform well in unseen data.

The splits will be 70%, 15% and 15% for train, validation and test respectively, this way we'll have enough data to rigorously train a model and also enough data to test its performance and get consistent results.

To avoid including bias when splitting the dataset (as we have seen, first lines have a lot of null values on most of the response variables), we have applied a random shuffle before performing the split.

Data centering and scaling techniques

Although all the input variables have the same units (reflection percentage), some of them take a different range of values. For example, wavelengths from 750 to 1300 tend to have a much higher value than others. To avoid giving some wavelength more importance than others in some algorithms, we must scale our dataset accordingly.

For doing so, we will use a Standard Scaler (Z-Score STD), which uses the formula:

$$z = \frac{x - \mu}{\sigma}$$

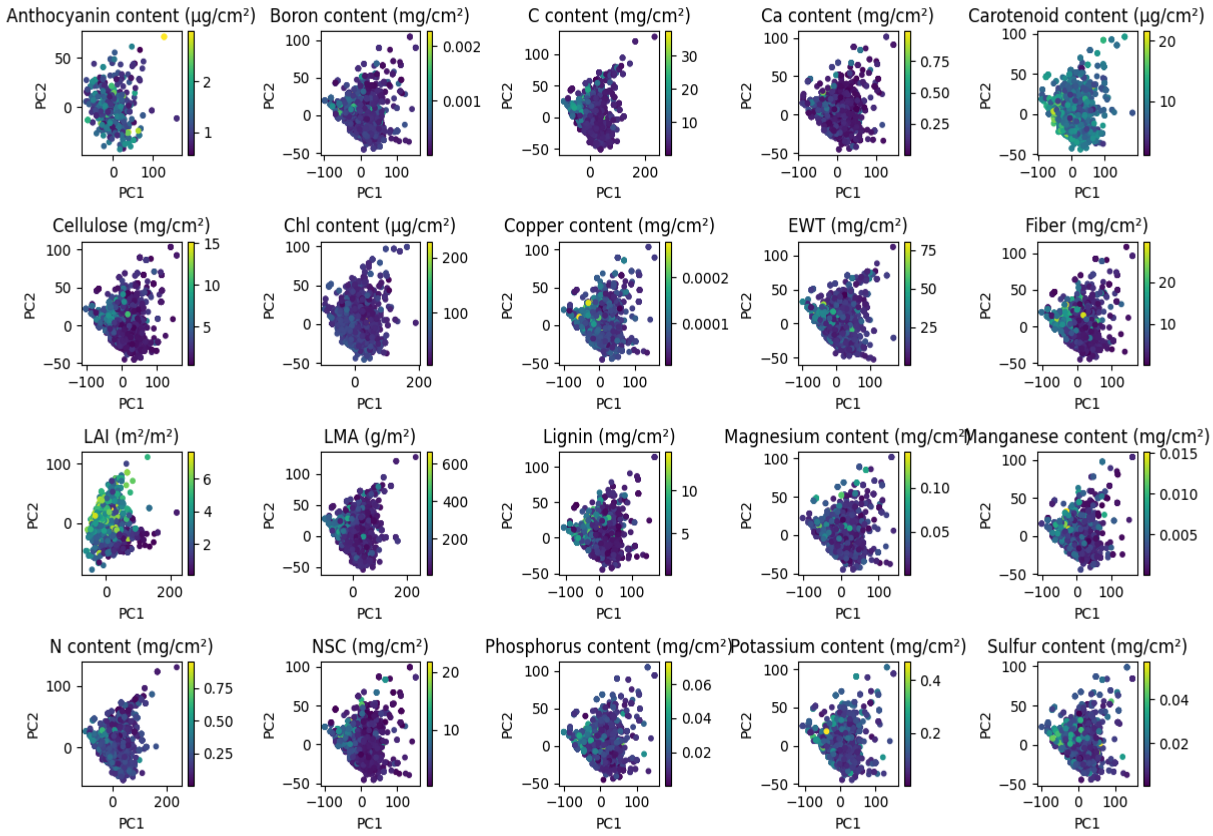
to get a distribution centered in zero and with unit variance for all features.

Another method for normalization, is to use Z-Score Robust method which uses the median value of the data and in that case we will use the following formula:

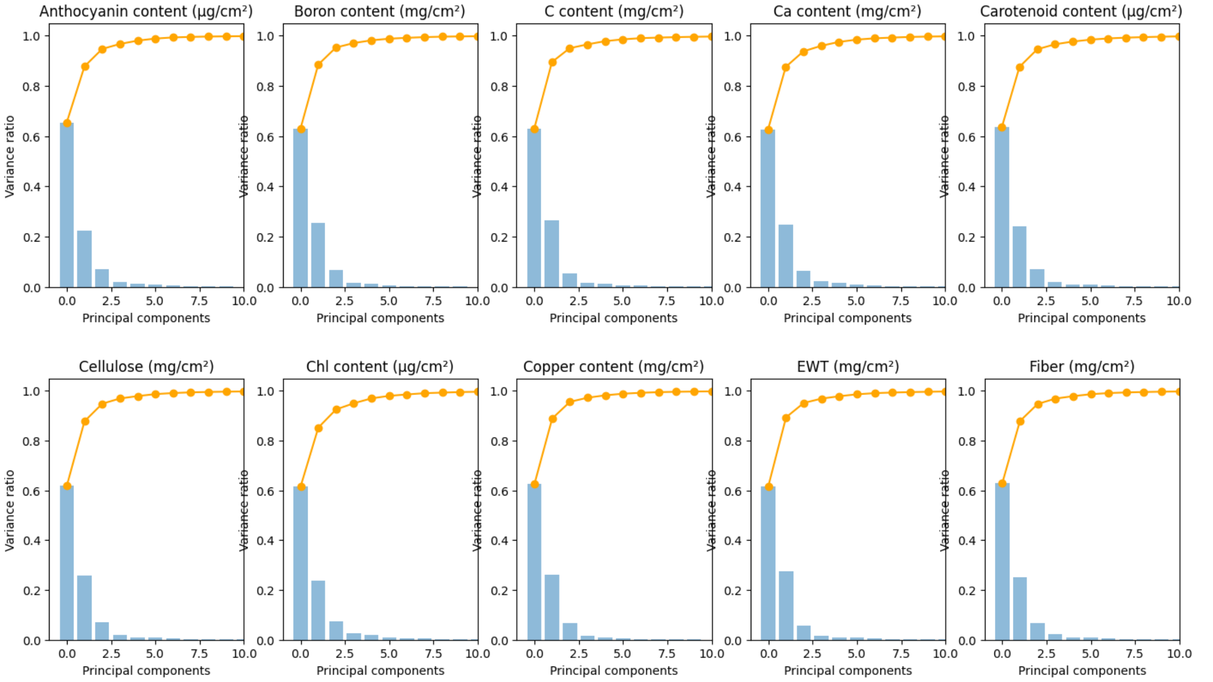
$$z = \frac{0{,}6745(Xi - Median(X))}{MAD} \text{ ,} where\ MAD = Median(|Xi\ -\ Median(X)|)$$

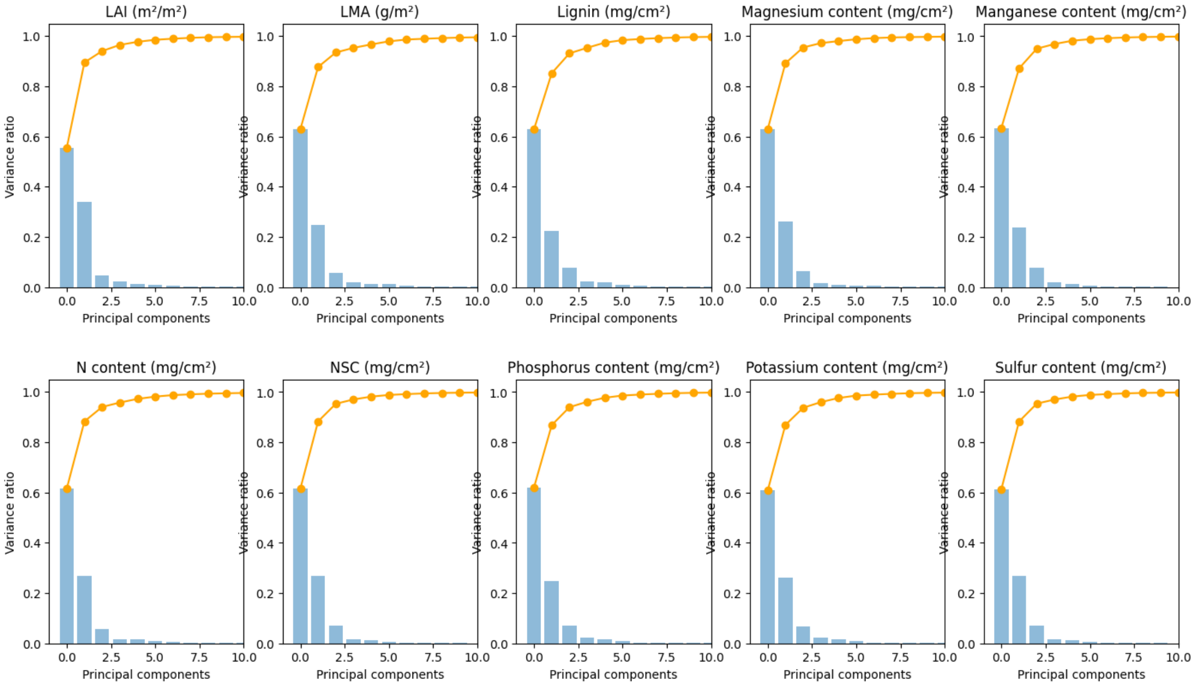
List of datasets (y and X) and PCA

If we divide the data into 20 datasets for each trait, we can plot PC1 versus PC2 plot for each trait based on its own dataset. This means, we run PCA function for each training dataset after the normalization.



Furthermore, we have plotted the variance for each trait's dataset after normalization, with different numbers of principal components. Based on the figures, it can be concluded that using only the first two principal components captures more than 80% of the variance, almost reaching 90%.



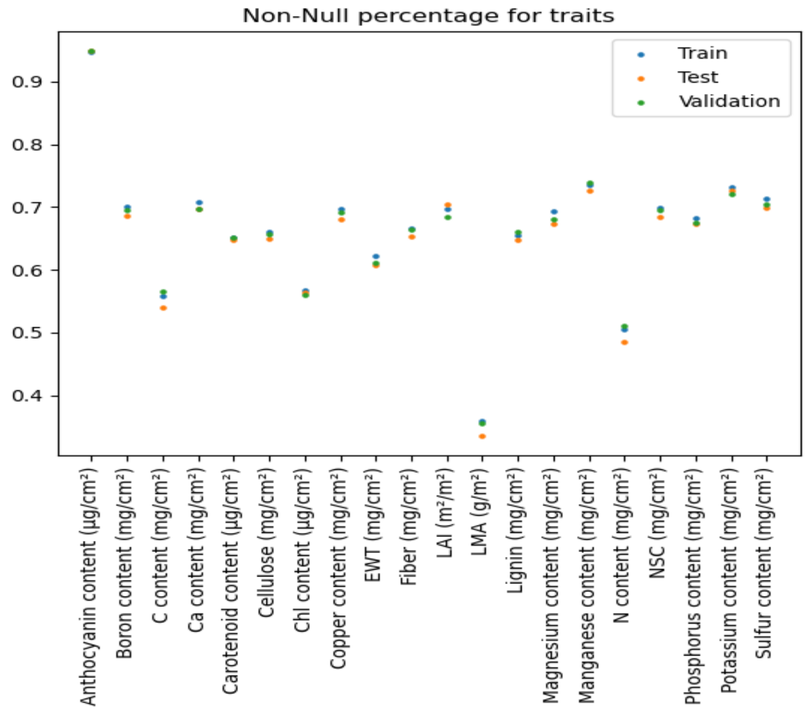


Non-Splitting approach

Another approach is not to split the data into 20 different arrays, and treat the data together. This could allow us to make multi-response prediction, and get a better generalization of our data (as we are using more samples, not only those that do not have null values on the response variable). Although this can be another approach, we will keep going with the other one as it fits better for our objective.

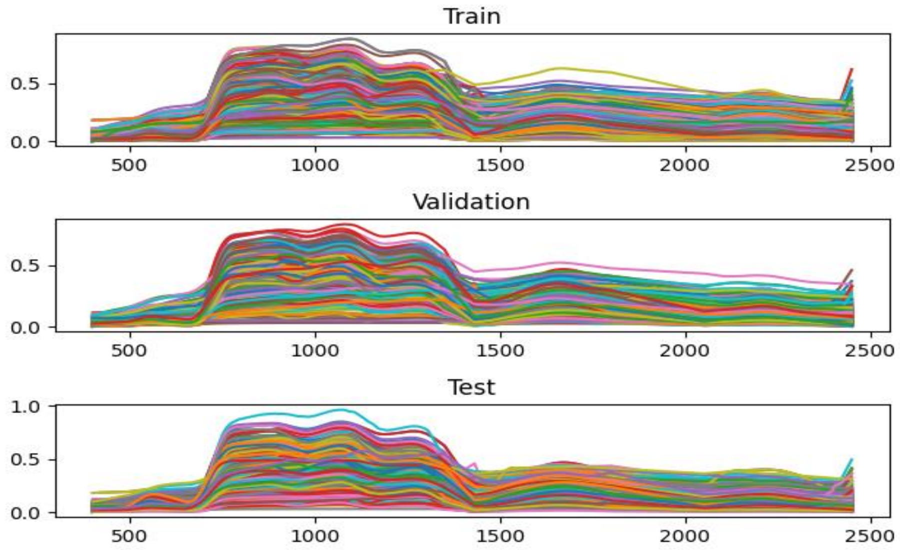
Visualizing data

After splitting the dataset on train, test validation, we will make sure that every split of the data has a representative part of it.



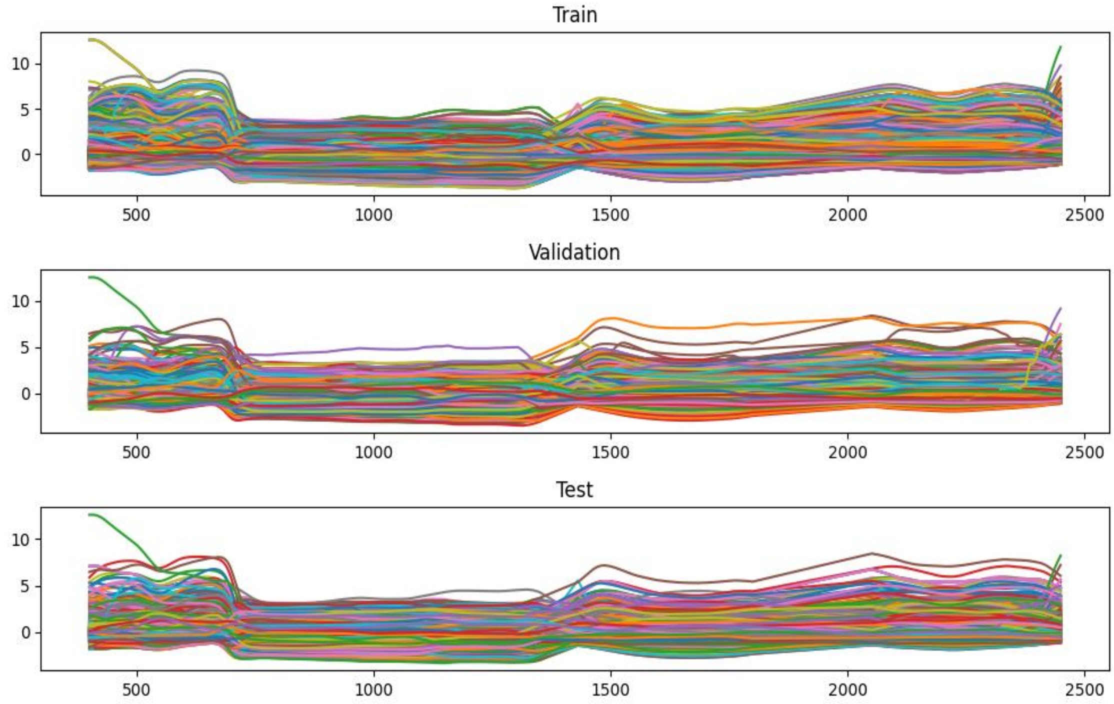
The target variables in all splits contain approximately the same amount of null values, which indicates that the partitions are balanced and they have the same percentage of non-null values.

Let’s also check the distribution of the wavelength in all three partitions:

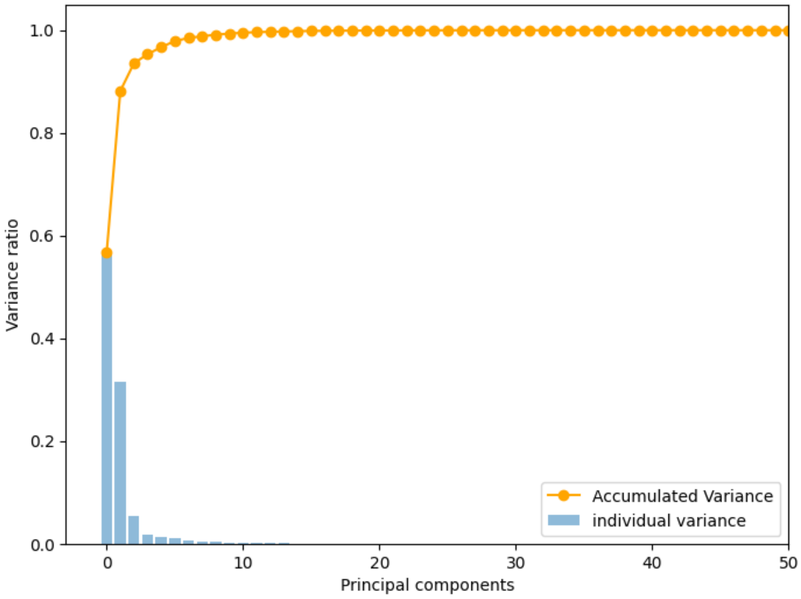


The three datasets seem to have the same distribution on input values, which, along with the target variables visualization, corroborates that the split has been done correctly.

Also, we have plotted the train, validation, and test data after they have been normalized by the Standard Scaler method.

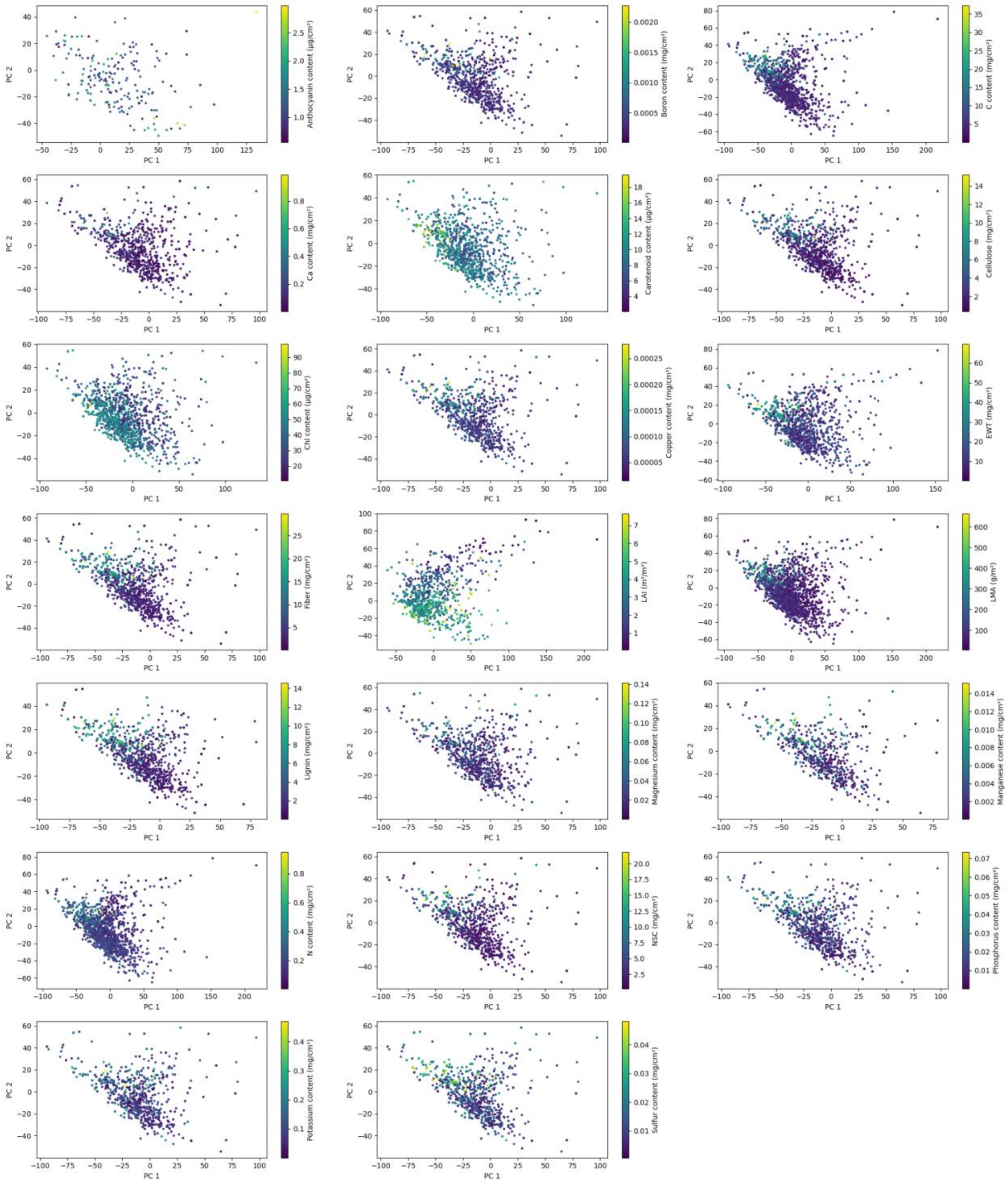


PCA will be very useful to drastically reduce the amount of features, detecting outliers and checking feature importance.



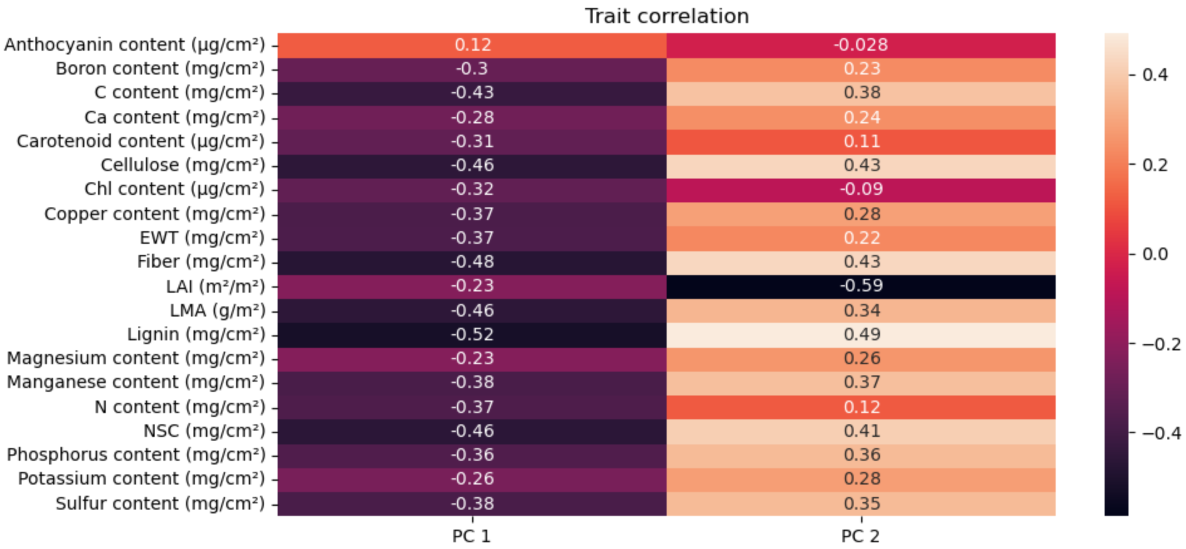
Surprisingly, with two components we accumulate a total of 88% of variance, which is really high. Furthermore, with three components we surpass the 90% threshold, reaching 93% variance. Having this much variance accumulated in a few features lets us have a representative dataset with many less features.

Having only two features lets us plot the samples with the target as a color dimension.



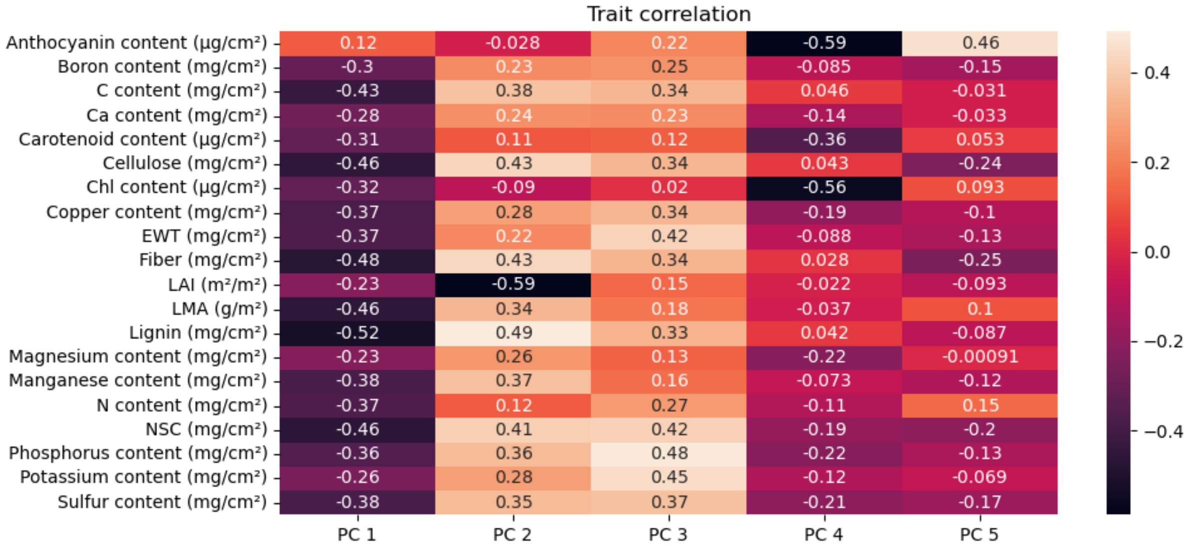
Most of the targets seem to be well explained and easy to predict with both components, such as LAI. Other targets seem not to be that correlated, for instance Anthocyanin (which might be caused due to the lack of data).

Having only two components allows us to check the individual correlations with the targets:



Surprisingly, most of the targets are very correlated with both components. First component seems to have a negative correlation, while the second component seems to have a positive one. This will result in higher values of the traits being situated at the top-left part of the graph. Keep in mind that component 1 and component 2 are perpendicular, so the correlation between them is zero, which means that each one gives different information.

Let’s see how the correlations behave while increasing the number of components.



While the higher components have less overall correlation, notice that component 4 and 5 have high correlation with Anthocyanin, which we could not get from the first 3 components.

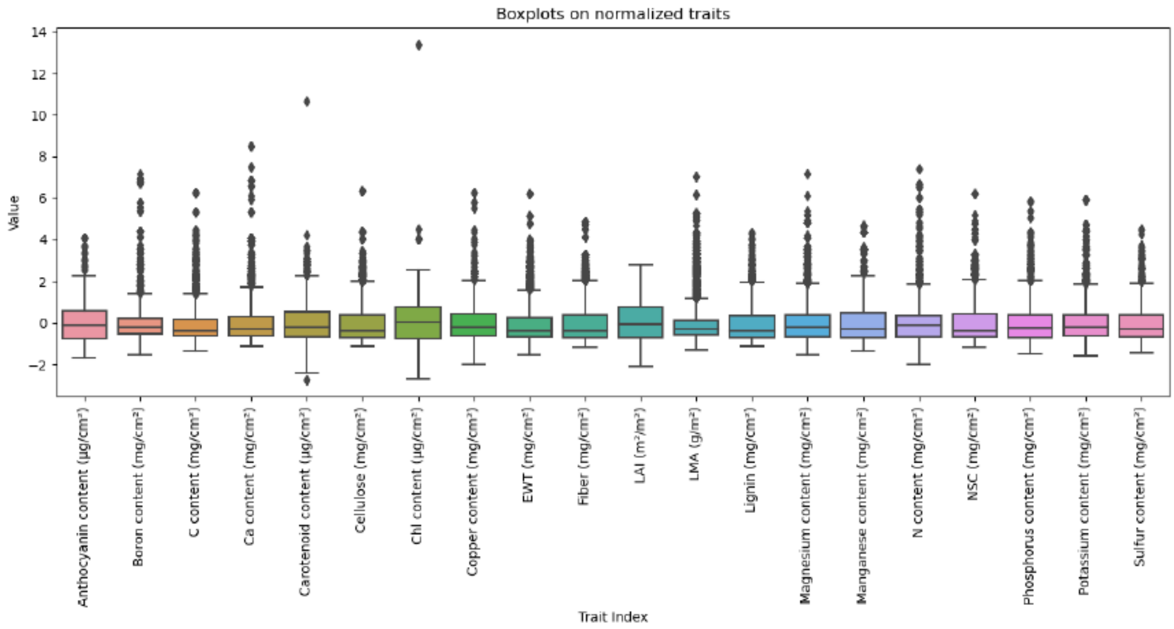
The PCA model was trained on training split, and the results shown are on validation and test dataset.

Mitigating actions and data synchronization

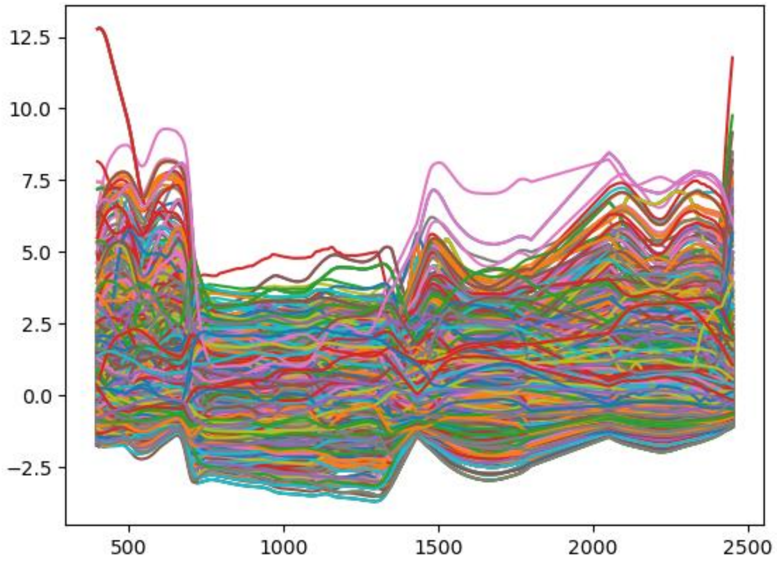
As we do not have time series data, we will only perform outlier detection.

Z-Score outlier detection:

Z score is calculated like in standardization: centering data to zero and imposing unit variance. Then those samples that have a feature values higher than a given threshold (usually around 3) are considered outliers, as the probability of a feature taking a higher number than 3 is extremely low.



For the traits, most of them would have outliers, specially Boron, Ca, Carotenoid, Chl LMA, Magnesium and N. Notice that there is even a sample with a value of 14.

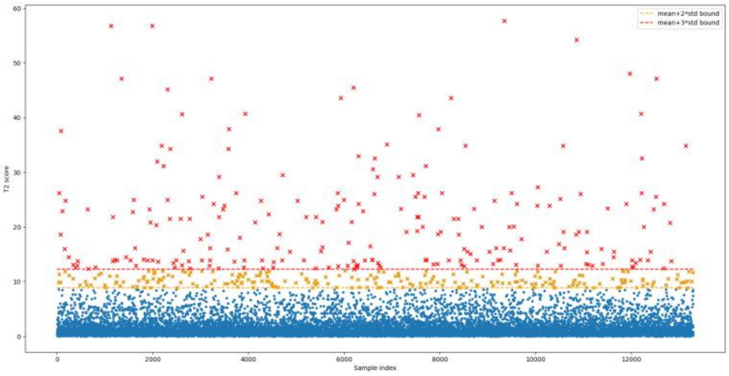


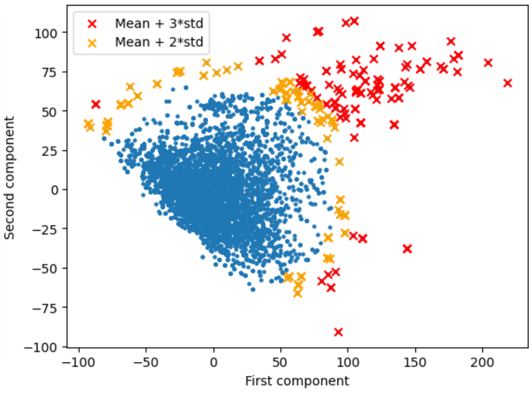
For wavelength, there are a lot of samples that exceed the 3 value threshold and, although there are some samples that clearly outstand (like the upper pink one or the upper red one, we cannot get rid of these samples.

All these results must be investigated, as for using the Z-score method there is a normality assumption, which we cannot ensure in our data, especially when we know that most of the variables take positive values.

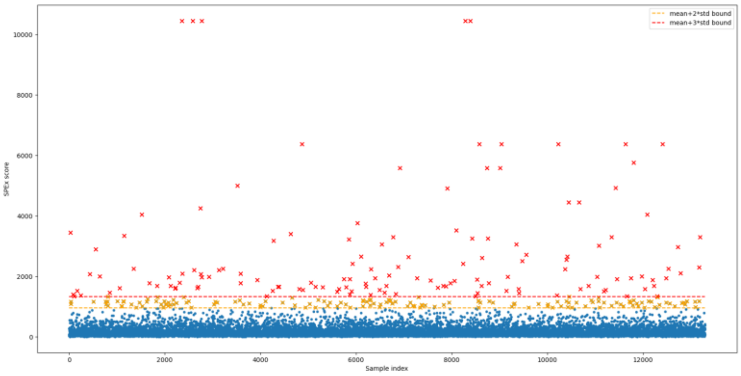
PCA Outlier detection:

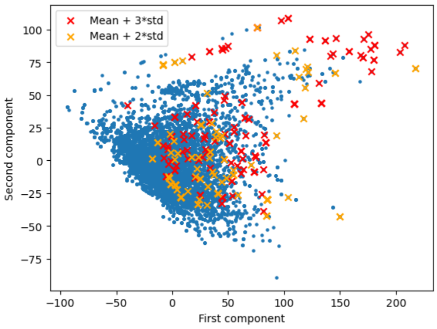
Following the previous work on PCA, we have used the 2 components model on all data to detect possible outliers.





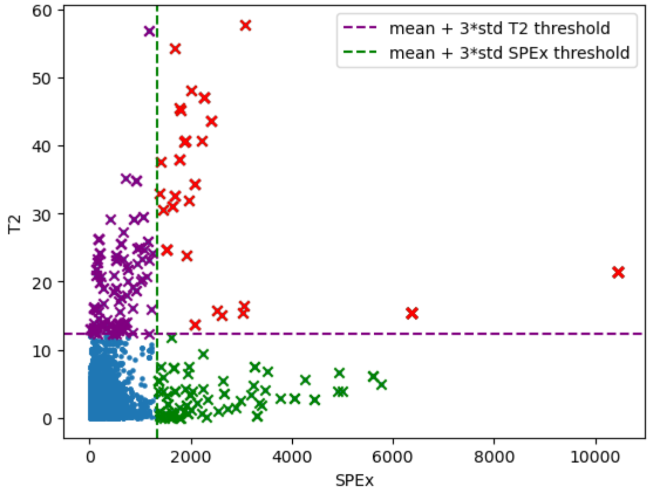
The T2 chart reveals a large number of samples that are out of the bounds, but this does not immediately imply that those samples are outliers. Let's check on SPEx, which measures how reconstructable a sample is.

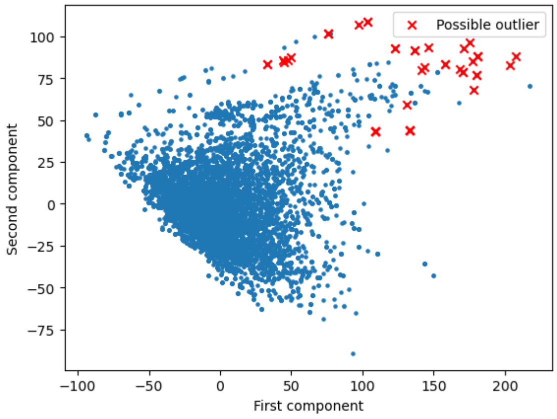




A low number of samples are detected out of the bounds in this case, but some of them have a value 5 times higher than the limit.

If we intersect both sets, we have an indicator of possible outliers.





Those are the samples that have a higher chance of being outliers.

However, we cannot eliminate them as we need to gather more information (probably from experts) to deduce if those samples are either mismeasurements (which we would then drop) or just extreme values of the data distribution (which we would then keep).