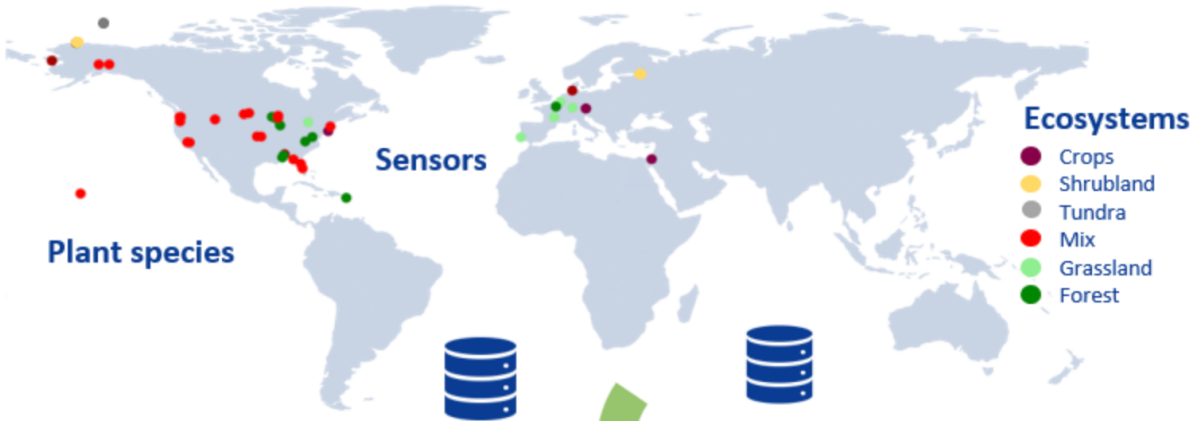
Hyperspectral Soft Sensor



Contents

---

1. Dataset description

The dataset originates from a multi-sensor study, where spectral data and vegetation properties from 42 datasets from various continents, climate and vegetation types have been combined. T

The hyperspectral data (input variables) have wavelengths ranging from 450 to 2500 nm, in 1 nm increments. The data has been processed to (a) remove the water absorption bands (1351-1430; 1801-2023;2451-2501) (b) smooth with Savinsky-Golay filter (c) interpolate the bands so they have the same resolution.

The leaf and canopy traits (response variables) include leaf pigments, leaf area index, equivalent water thickness, etc.. They are presented in the table below.

Trait name	Trait description	Unit	N	N Data sets	Mean	Std	Min	Max
Anth	Anthocyanin content	($\mu g/cm^2$)	644	2	1.27	0.41	0.56	2.98
Boron	Boron content	($\mu g/cm^2$)	1086	14	0.39	0.26	0.01	2.34
C	Carbon content	(mg/cm^2)	1876	23	5.84	4.44	0.10	37.29
Ca	Calcium content	($\mu g/cm^2$)	1045	16	107.25	101.97	0.69	988.73
Car	Carotenoid content	($\mu g/cm^2$)	1859	21	8.75	2.77	1.18	40.44
Cellulose	Cellulose	(mg/cm^2)	1402	15	2.35	1.87	0.35	15.22
Chl	Chlorophyll content	($\mu g/cm^2$)	2141	24	38.57	14.53	4.45	229.50
Copper	Copper content	($\mu g/cm^2$)	1101	14	0.07	0.03	0.01	0.28
EWT	Equivalent Water Thickness	(mg/cm^2)	1918	19	15.65	9.27	0.23	80.62
Fiber	Fiber	(mg/cm^2)	1385	15	5.23	4.57	0.14	29.81
LAI	Leaf Area Index	(m^2/m^2)	1643	15	3.35	1.64	0.06	7.67
Lignin	Lignin	(mg/cm^2)	1415	16	2.69	2.41	0.05	14.58
LMA	Leaf Mass per Area	(g/m^2)	3328	32	92.05	68.08	5.72	663.81
Magnesium	Magnesium content	($\mu g/cm^2$)	1099	15	24.09	16.16	0.25	141.54
Manganese	Manganese content	($\mu g/cm^2$)	894	14	3.09	2.31	0.01	15.19
N	Nitrogen content	(mg/cm^2)	2193	26	0.19	0.10	0.01	0.95
NSC	Non-Structural Carbohydrates	(mg/cm^2)	1093	14	3.21	2.85	0.28	21.83
Phosphorus	Phosphorus content	($\mu g/cm^2$)	1289	16	14.42	9.45	0.29	73.43
Potassium	Potassium content	($\mu g/cm^2$)	1008	15	102.64	62.73	0.40	470.07
Sulfur	Sulfur content	($\mu g/cm^2$)	1039	14	13.31	9.13	0.62	57.23

2. **Modelling goal**
    a. **Level B** (20p) Predicting traits from hyperspectral data.

Choose 5 out of the 20 traits. Calibrate, cross-validate and test regression models for estimating the chosen traits from the hyperspectral data. Reduce the model to the most important bands and evaluate the prediction accuracy.

b. **Level A** (30p) Predicting traits and trait correlations: band selection, missing data estimation and variational profiles of traits.

Create individual prediction models for all the traits. Predict the traits using the multivariate regression model, where there is missing data for a trait, and express a certainty for the prediction. After having a complete trait matrix, apply PCA on the trait matrix and interpret variational profiles of the traits.

c. **Level A2** (30p) Predicting traits and investigating anomalies.

Create individual prediction models for all the traits. Evaluate your prediction and rank the models. For five of the models that have the worst performance, investigate the input matrix using control charts (SPEx and T2). Impose control limits. If there are values exceeding the imposed limits, investigate the wavelength contributions to the control charts. Are the models performing better if the exceeding samples are excluded from calibration?

d. **Level A3** (30p) Predicting traits: selecting a proper technique

Multivariate regression can be performed through multiple techniques. Compare the performance of predicting the traits in case of MLR, PCR, PLS and k-PLS models. Which are the strengths and the weaknesses of each model? Some criteria of comparison is (a) the test partition prediction performance (b) the interpretability of the model (c) the time elapsed for training and prediction.

---

4. **References**

[1] - From spectra to plant functional traits: Transferable multi-trait models from heterogeneous and sparse data.
https://www.sciencedirect.com/science/article/pii/S0034425723001311?dgcid=author