PROSAIL-Retrieval of biophysical variables (Cab) with Random Forest

Daniele Cecca
Matr. 914358
MSc Artificial Intelligence for Science and Technology
Email: d.cecca@campus.unimib.it

1. Introduction

The estimation of vegetation biophysical parameters from satellite data has become an essential tool for monitoring ecosystem health, agricultural productivity, and the impacts of environmental change. Among these parameters, leaf chlorophyll content (Cab) plays a central role, as it is closely linked to photosynthetic activity, plant vigor, and crop management practices. Accurate retrieval of Cab from satellite imagery enables large-scale and temporally consistent assessments of vegetation condition, offering valuable insights for both scientific research and operational applications.

A key approach to parameter retrieval involves coupling radiative transfer models with machine learning techniques. Radiative transfer models, such as PROSAIL, simulate canopy reflectance spectra based on known combinations of leaf, canopy, and soil parameters. These synthetic spectral libraries can then be used to train statistical or machine learning models capable of mapping spectral data to biophysical variables. In this way, the gap between theoretical radiative transfer simulations and satellite observations can be bridged.

In this work, we develop an end-to-end pipeline for the retrieval and mapping of chlorophyll content (Cab) from Sentinel-2 multispectral imagery. First, synthetic reflectance signatures are generated using PROSAIL and subsequently mapped onto Sentinel-2 spectral bands. A Random Forest regressor is then trained on this dataset to predict Cab values. Once trained, the model is applied to Sentinel-2 imagery over the area of interest, with image collections retrieved through Google Earth Engine. The results are processed locally to produce spatially explicit maps of Cab as well as a multi-year time series (2019–2024). To facilitate exploration and interpretation, the outputs are presented in an interactive format.

This pipeline demonstrates how radiative transfer simulations, machine learning, and cloud-based remote sensing platforms can be combined into a robust framework for vegetation monitoring. While the methodology is general, the focus of this study is on chlorophyll content, with the aim of providing a scalable and reproducible approach for mapping this critical parameter over time and space.

2. Methodologies

2.1. Prosail model

The PROSAIL model couples PROSPECT, which simulates the optical properties of leaves, with SAIL, which models radiative transfer at the canopy scale. In PROSPECT, a leaf is represented as N elementary layers, each described by a refractive index and specific absorption coefficients of its biochemical constituents. The absorption coefficient at wavelength λ is expressed as

$$k(\lambda) = \sum_i C_i \, k_i(\lambda), \tag{1}$$

where C_i is the concentration of constituent i (e.g., chlorophyll Cab, carotenoids Car, water C_w, dry matter C_m), and $k_i(\lambda)$ its specific absorption spectrum. Reflectance (R) and transmittance (T) are then derived using the Beer–Lambert law across the N layers, linking biochemical composition to leaf optical properties.

The SAIL model extends this description to the canopy, solving a simplified radiative transfer equation (RTE) for a turbid medium composed of arbitrarily inclined leaves. Canopy reflectance is determined by integrating leaf scattering and absorption over the leaf area index (LAI), leaf angle distribution, and soil reflectance. The canopy bidirectional reflectance factor (BRF) can be expressed as

$$R_{\text{canopy}}(\theta_i, \theta_v, \phi) = f\big(R_{\text{leaf}}, T_{\text{leaf}}, LAI, \Omega, \rho_{\text{soil}}, \theta_i, \theta_v, \phi\big), \tag{2}$$

where θ_i and θ_v are solar and viewing zenith angles, ϕ is the relative azimuth, Ω is the leaf angle distribution parameter, and ρ_{soil} is soil reflectance. By coupling PROSPECT-derived leaf reflectance and transmittance with the SAIL canopy model, PROSAIL generates synthetic canopy spectra across the visible to shortwave infrared (400–2500 nm). These spectra can then be resampled to satellite sensor response functions, such as those of Sentinel-2, for vegetation parameter retrieval.

3. Dataset

3.1. Synthetic Data Generation

To train the regression model, a dataset of synthetic canopy reflectance spectra was generated using the PROSAIL radiative transfer model. A total of 10,000 synthetic spectral signatures were obtained by randomly sampling the PROSAIL input parameters across realistic ranges of leaf, canopy, and soil variables. The resulting spectra cover the 400–2500 nm domain at 1 nm resolution.

To better approximate real-world satellite measurements and account for instrumental as well as environmental noise, Gaussian noise was added to the simulated reflectance values:

$$R'(\lambda) = R(\lambda) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{3}$$

where $R(\lambda)$ is the noiseless reflectance at wavelength λ, and ϵ is a normally distributed perturbation with variance σ^2. This procedure introduces variability in the training data, improving the generalization ability of the regression model when applied to real Sentinel-2 observations.

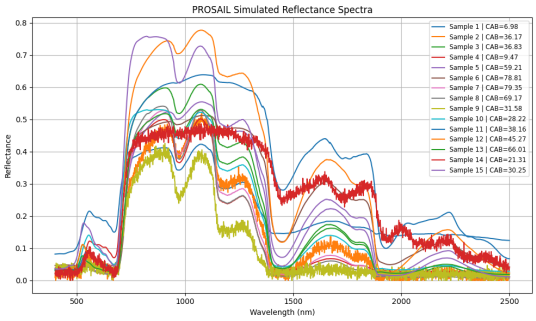

Figure 1: Prosail simulated data

3.2. Mapping to Sentinel-2

While PROSAIL produces continuous spectra , Sentinel-2 provides reflectance measurements only in a discrete set of multispectral bands, each defined by a central wavelength and bandwidth. To make the PROSAIL outputs comparable to Sentinel-2 data, for each Sentinel-2 band with wavelength interval $[\lambda_{\text{low}}, \lambda_{\text{high}}]$, we compute the mean reflectance of the PROSAIL spectrum within that range : This transformation reduces each synthetic signature from ∼2100 wavelengths to a Sentinel-2–like signature containing only the selected bands. For robustness, only bands with spatial resolutions of 10 m or 20 m were used, while the 60 m bands were discarded due to their lower reliability in vegetation monitoring applications.

As a result, the synthetic dataset preserves the biophysical variability simulated with PROSAIL while aligning with the spectral characteristics of Sentinel-2, enabling a consistent training dataset for the regression model.

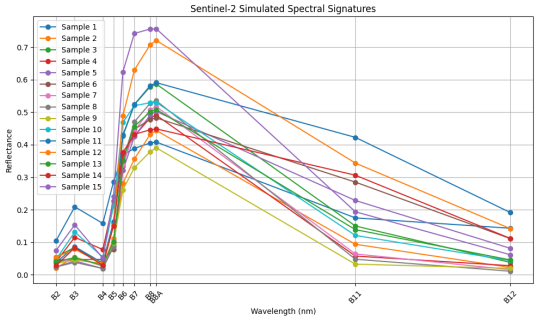

Figure 2: Sentinel-2 simulated spectral signatures

3.3. Final Dataset Construction

After mapping to Sentinel-2, the dataset was organized such that each sample corresponds to a synthetic Sentinel-2–like signature. The input features are given by the reflectance values of the selected Sentinel-2 bands (10 m and 20 m resolution), while the target variable is the chlorophyll content (Cab) used in the PROSAIL simulation.

To verify the suitability of the dataset for regression, the distribution of the target feature Cab was examined. A balanced target distribution is desirable to avoid biasing the regressor toward specific parameter ranges. Visual inspection of the histogram of Cab values confirmed that the dataset is well balanced, ensuring that the regression model will be trained on a representative range of chlorophyll content values.

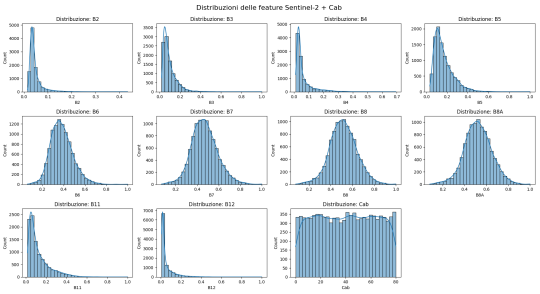

Figure 3: Features distributions

4. Model

4.1. Random Forest Regression

To predict the chlorophyll content (Cab) from the synthetic Sentinel-2–like signatures, we trained a Random Forest regressor. Random Forests are ensemble methods that combine multiple decision trees to improve predictive performance and reduce overfitting. Each tree is trained on a bootstrap sample of the training data, and at each split a random subset of features is considered. The final prediction is obtained by averaging the predictions of the individual trees, which ensures robustness to noise and variability in the dataset.

The model was trained using the synthetic dataset described in the previous section. The data were split into 80% training and 20% test sets. Training was performed with the scikit-learn `RandomForestRegressor` implementation, using default parameters.

4.2. Results

The performance of the model was evaluated on the test set using two standard regression metrics:

- Root Mean Squared Error (RMSE), which quantifies the average magnitude of the prediction error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4}$$

- Coefficient of Determination (R^2), which measures the proportion of variance in the target variable explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{5}$$

On the test set, the Random Forest achieved an $RMSE \approx 8.00$ and an $R^2 \approx 0.88$, indicating good predictive performance and a strong ability to explain the variance of Cab.

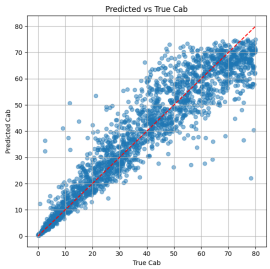

Figure 4: Predicted vs true

Feature importance analysis confirmed that the most informative Sentinel-2 bands for chlorophyll prediction are B5 and B7, both of which are well known for their sensitivity to vegetation properties and are frequently used in vegetation indices and biophysical parameter retrieval.

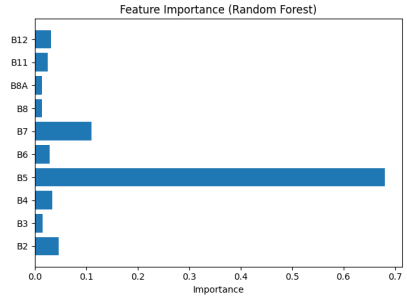

Figure 5: Features importance

5. Application to Sentinel-2 Imagery

5.1. Area of Interest and Data Selection

To evaluate the Random Forest model on real data, we applied it to Sentinel-2 surface reflectance imagery over a selected region of interest (ROI) in southern Italy. The ROI is located near Altamura, in the Apulia region, a predominantly agricultural area characterized by diverse crop types and natural vegetation. This makes it a suitable test site for monitoring chlorophyll dynamics and assessing vegetation health.

The ROI was defined as a polygon with approximate coordinates:

```
"type": "Polygon",
"coordinates": [
  [
    [16.7206, 40.8042],
    [16.7206, 40.7993],
    [16.7284, 40.7993],
    [16.7284, 40.8042],
    [16.7206, 40.8042]
  ]
]
```

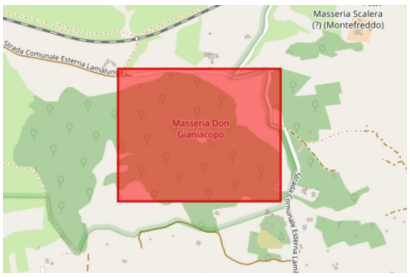

Figure 6: ROI GEE

To analyze temporal changes in chlorophyll content, we selected Sentinel-2 Level-2A images (`COPERNICUS/S2_SR_HARMONIZED`) from 2019 to 2024. For each month within this period, the first image with cloud coverage below 10% was chosen. This ensured a consistent, cloud-free monthly time series for the six-year study period.

5.2. Preprocessing of Sentinel-2 Images

Each selected Sentinel-2 image underwent the following preprocessing steps:

- Clipping: Images were cropped to the ROI, retaining only the relevant pixels within the study area.
- Resampling: Sentinel-2 provides bands at different native spatial resolutions (10 m and 20 m). To harmonize the dataset, all bands were resampled to a common resolution of 20 m using bilinear interpolation.

- Reprojection: The images were reprojected to the UTM Zone 32N coordinate system (EPSG:32632), ensuring spatial alignment across the dataset.
- Normalization: For each image, band reflectance values were normalized on a pixel-by-pixel basis using min–max normalization.

The processed images were exported as GeoTIFF files for subsequent analysis and model application.

5.3. Cab Prediction and Time Series Construction

The trained Random Forest model was applied to the preprocessed Sentinel-2 image collection to produce monthly maps of chlorophyll content (Cab) for the ROI. Each image was transformed into a Cab prediction map, providing a spatially explicit representation of chlorophyll dynamics over time.

The results were organized into a time series spanning January 2019 to December 2024. The image dates were parsed from filenames using regular expressions, and a temporal dataset was built by aligning Cab predictions with acquisition dates. Finally, an interactive time series visualization was generated using `plotly`, displaying the temporal evolution of predicted chlorophyll content across the study area.

One challenge in interpreting the Cab maps is that some images are affected by shadows caused by clouds, making the chlorophyll values less reliable in those regions. This limitation highlights the importance of refining the image selection phase. For example, instead of directly using the first available images from Google Earth Engine, future work could prioritize composite or "main" images that minimize cloud-related artifacts

Figure 7: first image gee

This pipeline enables both spatial (Cab maps) and temporal (time series) monitoring of vegetation chlorophyll content from Sentinel-2 imagery, demonstrating the applicability of the PROSAIL–Random Forest framework to real-world satellite data.

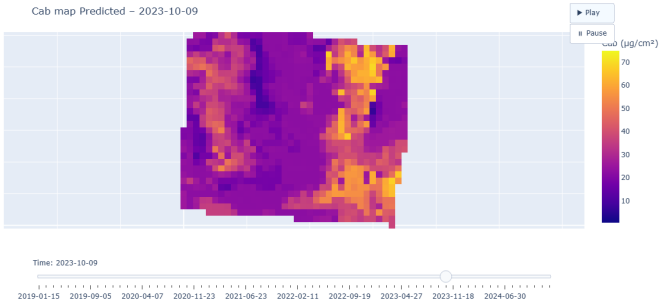

Figure 8: Temporal evolution of predicted chlorophyll

6. Conclusion

In this work, we developed an end-to-end pipeline that links radiative transfer simulations with satellite remote sensing to estimate vegetation chlorophyll content (Cab) from Sentinel-2 imagery. The methodology integrates three main components: (i) synthetic data generation using the PROSAIL model, (ii) regression modeling with a Random Forest trained on Sentinel-2–like reflectance signatures, and (iii) application of the trained model to real Sentinel-2 images for spatial and temporal monitoring.

By generating 10,000 synthetic spectra with PROSAIL and resampling them to Sentinel-2 bands, we created a training dataset that captures realistic canopy reflectance variability. The Random Forest regressor, trained on this dataset, achieved good predictive performance with an RMSE 8.00 and R^2 0.88. Feature importance analysis confirmed that Sentinel-2 bands B5 and B7, both known for their sensitivity to vegetation properties, play a central role in chlorophyll estimation.

The trained model was successfully applied to Sentinel-2 imagery over an agricultural region in Apulia, Italy. A six-year time series (2019–2024) was constructed by selecting monthly cloud-free images, preprocessing them (clipping, resampling, reprojection, and normalization), and generating Cab maps. The results provide both spatial and temporal insights into chlorophyll dynamics, demonstrating the feasibility of combining radiative transfer simulations with machine learning and cloud-based Earth observation platforms for vegetation monitoring.

One limitation encountered was the interpretability of the Cab maps: in some cases, shadows caused by cloud cover led to unrealistic values in the predictions. This suggests that the image selection process could be further refined, for example by relying on composite or representative "main" images rather than the first available ones in Google Earth Engine, to reduce the impact of cloud-related artifacts.

Overall, the proposed framework offers a scalable and reproducible approach for chlorophyll content retrieval from multispectral imagery. Future work will focus on extending the method to additional biophysical parameters, refining uncertainty quantification, and validating predictions against field measurements to further enhance the robustness and applicability of the pipeline.