Research Project:
Retrieval of plant biophysical and biochemical variables from remote sensing data using a hybrid machine learning method

Trier University

Zavud Baghirov

Environmental Sciences

University of Trier

Summer Semester 2021

Abstract

This will be the abstract at the end [TO BE UPDATED]

Contents

List of Figures

List of Tables

List of Abbreviations

3D . . . . . Three-dimensional

INFORM . . . Invertable Forest Reflectance Model

RTM . . . . . Radiative Transfer Model

SAIL . . . . . Scattering by Arbitrary Inclined Leaves

PROSAIL . . The combination of PROSPECT and SAIL models

FLIM . . . . . Forest Light Interaction Model

LAI . . . . . Leaf Area Index

MLRA . . . . Machine Learning Regression Algorithms

ML . . . . . Machine Learning

DT . . . . . Decision Trees

ANN . . . . . Artificial Neural Networks

KBMLRM . . Kernel-Based Machine Learning Regression Methods

RF . . . . . Random Forest

RFR . . . . . Random Forest Regression

LUT . . . . . Look-Up-Table

NN . . . . . Neural Networks

SVR . . . . . Support Vector Regression

SVM . . . . . Support Vector Machines

GPR . . . . . Gaussian Process Regression

GP . . . . . Gaussian Process

VI . . . . . Vegetation Index

DR . . . . . Dimensionality Reduction

WT . . . . . Wavelet Tranform

PCA . . . . . Principal Component Analysis

AL . . . . . Active Learning

NIR . . . . . Near Infrared

SWIR . . . . . Short Wave Infrared

PC . . . . . Principal Component

1 Methods

This section explains the methods used in this research.

1.1 Local sensitivity analysis

Local sensitivity analysis was performed to assess the effect of each of the main 6 plant biochemical and biophysical variables on the PRISMA image bands. In the local sensitivity analysis simulation is performed by keeping all the variables constant at their determined fixed or default values except the parameter of interest. This way the effect of a specific parameter on the simulated spectra can be assessed. In this research the plant parameters C_{ab}, C_w, C_m, LAI_s, CD and SD were varied each 15 times (Table (1.1)), while keeping the rest of the variables at their default values (Table (1.2)). The default and varied values were chosen based on the literature (e.g. Darvishzadeh et al. (2019); Laurent et al. (2011); Schlerf and Atzberger (2012)) where similar RTM method used to simulate reflectance for Spruce trees.

Table 1.1 shows the 6 parameters that were varied, their units, minimum and maximum values. Each parameter was varied 15 times, meaning 15 different spectra were simulated for each variable.

Table 1.1: INFORM Parameters varied in local sensitivity analysis (each parameter were varied 15 times)

Parameter	Abbrev.	Unit	Min	Max
Chlorophyll content	C_{ab}	$\frac{\mu g}{cm^2}$	20	60
Equivalent water thickness	C_w	$\frac{g}{cm^2}$	0.0035	0.035
Leaf dry matter content	C_m	$\frac{g}{cm^2}$	0.008	0.03
Leaf area index (single)	LAI_s	$\frac{m^2}{m^2}$	0	7
Stem density	SD	ha^{-1}	200	5000
Crown diameter	CD	m	1.5	8.5

Table 1.2 shows the determined default values for each INFORM parameter that were kept during the sensitivity simulation while one of the parameter was varied (Table 1.1).

Table 1.2: INFORM Parameters that were kept constant while one parameter was varied at a time

Parameter	Abbr	Unit	Value
Leaf structure parameter	N	$-$	3
Chlorophyll content	C_{ab}	$\frac{\mu g}{cm^2}$	40
Leaf cartenoid content	C_{ar}	$\frac{\mu g}{cm^2}$	8
Brown Pigment Content	C_{brown}	$-$	0.001
Equivalent water thickness	C_w	$\frac{g}{cm^2}$	0.0117
Leaf dry matter content	C_m	$\frac{g}{cm^2}$	0.03
Average leaf inclination angle	$ALIA$	$^\circ$	65
Leaf area index (single)	LAI_s	$\frac{m^2}{m^2}$	6
Leaf area index (understorey)	LAI_u	$\frac{m^2}{m^2}$	0.5
Hot spot parameter	Hot	$\frac{m}{m}$	0.02
Solar zenith angle	tts	$^\circ$	45.43
Observer zenith angle	tto	$^\circ$	0
Sun-sensor azimuth angle	psi	$^\circ$	181.41
Soil brightness	α_{soil}	$-$	0.5
Stem density	SD	ha^{-1}	700
Crown diameter	CD	m	5
Mean Height	H	m	20
Fraction of diffuse incoming	$skyl$	$-$	0.1
Soil reflectance spectrum	B_g	$-$	default

Solar zenith angle and *Sun-sensor azimuth angle* were calculated based on the PRISMA image acquisition parameters (date, lat/long etc.) using the *solar position calculator* at `https://www.esrl.noaa.gov/gmd/grad/solcalc/azel.html`.

RTM models PROSPECT5, 4SAIL and FLIM were coupled (INFORM) in order to simulate canopy reflectance. Simulations were carried out using the *ccrtm* package (Visser, 2021) in *R* (R Core Team, 2021). The default soil spectra provided by the the *ccrtm* package (Visser, 2021) was used for the simulations. Spectral resampling was performed in order to resample the INFORM output spectra (1nm

resolution) into PRISMA image bands. For spectral resampling the *R* package *hsdar* (Lehnert et al., 2019) was utilized.

1.2 RTM simulation (INFORM)

PROSPECT5, 4SAIL and FLIM RTM models were coupled (INFORM) to simulate forest canopy reflectance based on different values of plant biophysical and biochemical parameters. The 6 parameters that were mentioned in the previous chapter were varied and spectra was simulated based on each combination of these variables. The number of combinations increase exponentially, which in turn requires increased computational power. Therefore, the trade-off must be taken into account between computational power or time and accurate simulation.

Different authors suggest different number of LUT size for RTM simulation. For example, Danner et al. (2021) mention that LUT size of minimum 50,000 is recommended. Ali et al. (2020) and Darvishzadeh et al. (2019) created a LUT size of 100,000 and 500,000 respectively.

In this research, LUT size of 316,800 was created based on each combination of different plant biophysical and biochemical parameters. The range of the varied parameters and parameters that were kept constant were determined based on the suggestions of the studies that were mentioned in the previous chapter. These studies used similar methods to simulate canopy reflectance for mainly Spruce forests/trees.

Table 1.3 shows the variables that were used to simulate forest canopy parameters. Table 1.3 also contains information about the range of the values and how many times each parameter was varied.

Table 1.3: Range of full input parameters that were used to create a LUT size of 316800

Parameter	Abbr	Unit	Min	Max	Steps
Leaf structure parameter	N	–	3	3	–
Chlorophyll content	C_{ab}	$\frac{\mu g}{cm^2}$	20	60	15
Leaf cartenoid content	C_{ar}	$\frac{\mu g}{cm^2}$	8	8	–
Brown Pigment Content	C_{brown}	–	0.001	0.001	–
Equivalent water thickness	C_w	$\frac{g}{cm^2}$	0.0035	0.035	10
Leaf dry matter content	C_m	$\frac{g}{cm^2}$	0.008	0.03	11
Average leaf inclination angle	$ALIA$	$^\circ$	65	65	–
Leaf area index (single)	LAI_s	$\frac{m^2}{m^2}$	0	6.5	16
Leaf area index (understorey)	LAI_u	$\frac{m^2}{m^2}$	0.5	0.5	–
Hot spot parameter	Hot	$\frac{m}{m}$	0.02	0.02	–
Solar zenith angle	tts	$^\circ$	45.43	45.43	–
Observer zenith angle	tto	$^\circ$	0	0	–
Sun-sensor azimuth angle	psi	$^\circ$	181.41	181.41	–
Soil brightness	α_{soil}	–	0.5	0.5	–
Stem density	SD	ha^{-1}	200	5000	4
Crown diameter	CD	m	1.5	8.5	3
Mean Height	H	m	20	20	–
Fraction of diffuse radiation	$skyl$	–	0.1	0.1	–
Soil reflectance spectrum	B_g	–	default	default	–

All simulations were performed using the library *ccrtm* (Visser, 2021) in *R* programming language (R Core Team, 2021) using the most recent version 4.1.0. Generating a LUT size of 316,800 is an expensive process from a computational standpoint (depending on how much computer resources and time are available this might change). Also, all simulations are independent of each other, meaning simulation of one spectra has no effect on the other, as every simulated spectra is simulated based on a different combination of parameters. These two factors make the generation of such a large LUT good candidate for parallel computation. Therefore, the software packages *doParallel* (Corporation and Weston, 2020) and *foreach* (Microsoft and Weston, 2020) were utilized for parallel computation (using all the available cores) in *R* programming language (R Core Team, 2021). This significantly reduced the computational time. All of the simulations were computed on a Lenovo Thinkpad E480 running under Windows 10 operating system with a

processor Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, 2001 Mhz, 4 Core(s), 8 logical processor(s).

1.3 Spectral resampling

The output of INFORM simulations have 1nm spectral resolution within the range of 400nm-2500nm and needs to be spectrally resampled to PRISMA image bands. In this research, the spectral response function of the PRISMA image was used. Band center wavelengths and full width half maximum values were extracted from the PRISMA image metadata and used for spectral resampling. For spectral resampling, the *R* package *hsdar* (Lehnert et al., 2019) was utilized.

1.4 Statistics of simulated data and PRISMA image

Statistical information such as standard deviation and mean were calculated for the simulated (and resampled to PRISMA bands) data and all the pixels of the PRISMA image within the study area. Pixels that are out of the study area boundary were masked out. Then, average spectra in the LUT (synthetic database) and PRISMA image (only study area) were compared to each other. LUT contains 316,800 simulated spectra, the number of pixels within the study area in the PRISMA image is only 95517. Statistical information were extracted using the libraries in the *tidyverse* package (Wickham et al., 2019) and the plots for visualization were produced using *ggplot2* (Wickham, 2016).

1.5 Gaussian noise

Simulated reflectance data usually do not contain any noise. This is, however, not the case with remote sensing data as they are commonly found to contain various types of noise (Rivera-Caicedo et al., 2017). In this study, 3% Gaussian noise was added to each simulated spectra in the LUT in order to make the simulated data

more similar to the real remote sensing data. In order to assess the effect of adding 3% Gaussian noise to the simulated data, one spectra from the LUT and one pixel from the PRISMA image were randomly picked and plotted.

1.6 Defining training, validation and testing sets

The data in the LUT was divided into training, validation and testing sets. Model building and training will be done using only the training set. Validation set will be used to validate the model (e.g. assessing the impact of different hyper-parameters) and the performance of the final model will be tested using the testing set. This step is important because it will allow us to monitor whether the model can generalize to the data (e.g. testing set) it was not trained on.

First, the full data set was shuffled and about 20% of the data was randomly sampled and assigned to validation and testing sets (10% validation, 10% testing sets). Random sampling ensures that there is no any pattern contained in any of the divided data sets.

1.7 Data processing

First, the simulated canopy reflectance in the training data set was normalized and standardized using the Equation (1.1):

$$Band_{n_{scaled}} = \frac{Band_n - \mu_{Band_n}}{\sigma_{Band_n}} \tag{1.1}$$

Here $Band_n$ refers to the reflectance values in the nth simulated band and μ_{Band_n} and σ_{Band_n} are mean and standard deviation of the reflectance values in the nth simulated band. $Band_{n_{scaled}}$ is a transformed version of $Band_n$ that has a mean of 0 and standard deviation of 1. This step ensures that all simulated bands have the same mean and standard deviation.

Data normalization and standardization were only performed using training data set. Mean and standard deviation of the training set were then used to transform the validation and testing data sets.

1.8 Principal Component Analysis (PCA)

Hyperspectral remote sensing data can contain many highly correlated bands. Dimensionality reduction techniques can be efficiently used to reduce the dimensions of hyperspectral remote sensing data. Benefits of reducing the dimensions of simulated data in plant biophysical variable retrieval studies have been demonstrated (Danner et al., 2021; Rivera-Caicedo et al., 2017). In this study, one of the most commonly used DR technique Principal Component Analysis (PCA) was performed. In general, PCA tries to capture as much variation as possible with smaller number variables compared to the original data. PCA produces new variables called Principal Components and each Principal Component (PC) contains certain amount of variation available in the original data. Typically first PC contains the most variation, the second PC contains the second most variation and so on (Bro and Smilde, 2014).

Like in the processing step, PCA was only applied to the training data and the PCA result in the training data was used to transform the validation and testing sets. Cumulative sum of the variations the PCs contain was calculated in order to assess the proportion of the variation that can be explained with fewer variables than the original data (LUT). PCA and data processing performed using the package *recipes* (Kuhn and Wickham, 2021) in *tidymodels* (Kuhn and Wickham, 2020).

2 Results

2.1 Local sensitivity analysis

Figure 2.1 shows the result of sensitivity analysis. Chlorophyll content (C_{ab}) appears to almost exclusively impact the visible spectra. Some effect can also be noticed in the red-edge, but there is not a significant effect of varying C_{ab} on the simulated spectra within the near-infrared (NIR) and short wave infrared (SWIR) (Figure 2.1.a). Conversely, equivalent water thickness (C_w) (Figure 2.1.b) and leaf dry matter content (C_m) (Figure 2.1.c) both have large effects on simulated spectra within the NIR and SWIR but no significant effect within the visible spectra. Leaf Area Index (single) (LAI_s) (Figure 2.1.d), Crown diameter (CD) (Figure 2.1.e)) and Stem density (Figure 2.1.f) all have noticeable effect on the simulated canopy reflectance almost all over the spectra.

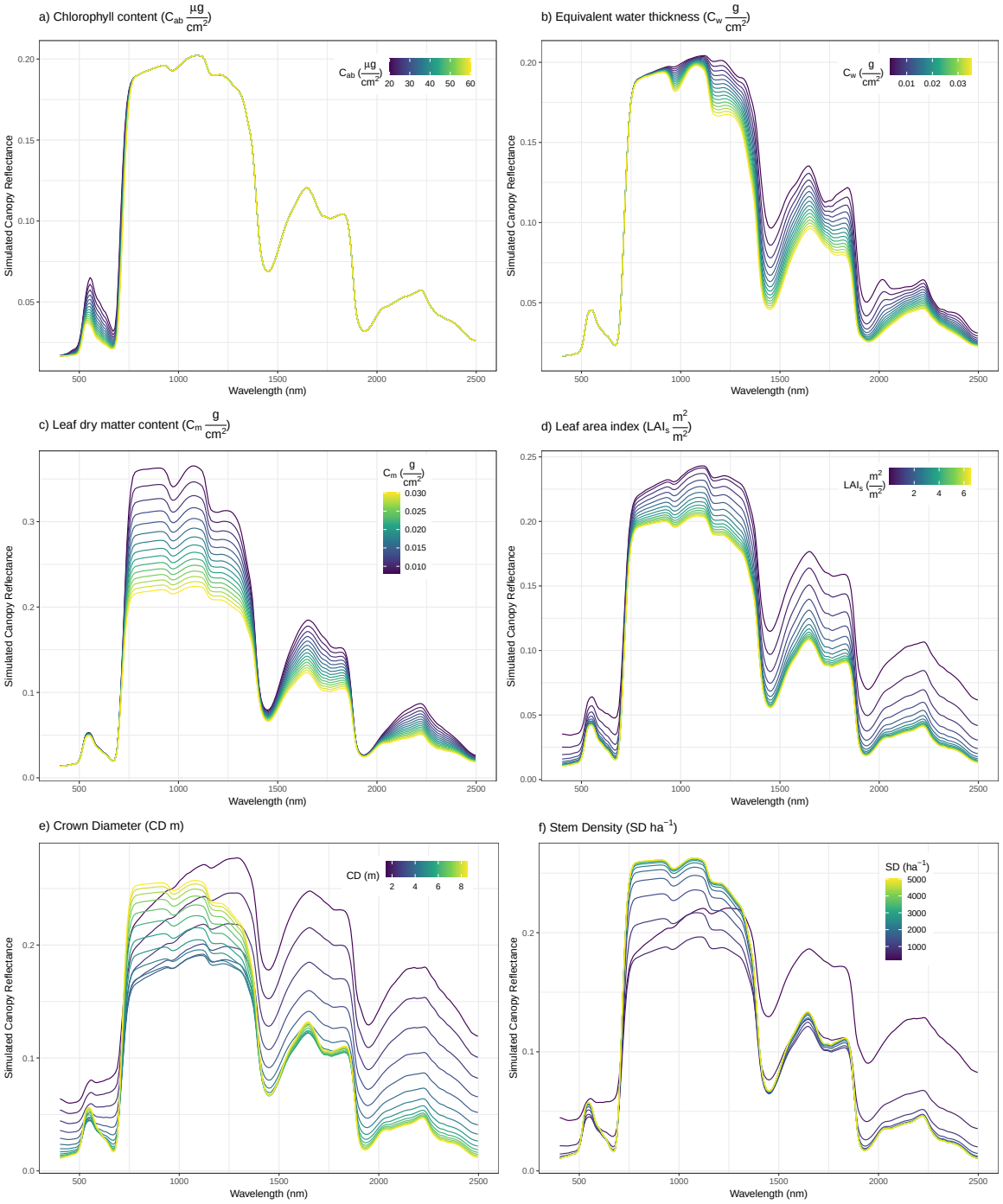

Figure 2.1: Effects of varying the chosen parameters on the simulated spectra

2.2 RTM simulation (INFORM)

Synthetic canopy reflectance data set were produced and stored in a LUT containing all 316,800 simulations. In this research, LUT was defined as a matrix. Each row of this matrix is a different simulated spectra and columns are simulated reflectance of wavelengths with the range of 400nm-2500nm with 1nm spectral resolution and 6 additional columns containing values of the corresponding variables C_{ab}, C_w, C_m, LAI_s, CD and SD that were used for each simulation. Hence the dimensions of the LUT matrix is 316,800 rows (number of simulations) by 2107 columns (2101 simulated "bands" + 6 INFORM variables):

$$\begin{bmatrix} 400nm_1 & \dots & 2500nm_1 & Cab_1 & Cw_1 & Cm_1 & LAIs_1 & CD_1 & SD_1 \\ 400nm_2 & \dots & 2500nm_2 & Cab_2 & Cw_2 & Cm_2 & LAIs_2 & CD_2 & SD_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 400nm_{316,800} & \dots & 2500nm_{316,800} & Cab_{316,800} & Cw_{316,800} & Cm_{316,800} & LAIs_{316,800} & CD_{316,800} & SD_{316,800} \end{bmatrix}$$

In this matrix, $400nm_n$, ..., $2500nm_n$ refer to the simulated reflectance for the corresponding wavelength in the simulation number n. Cab_n, Cw_n, Cm_n, $LAIs_n$, CD_n and SD_n are values of the INFORM parameters that were used in the nth simulation.

2.3 Spectral resampling

The output of INFORM simulations were resampled to 231 PRISMA bands. The LUT matrix was used for spectral resampling and the resulting matrix has a dimension of 316,800 rows (number of simulations) by 237 columns (231 PRISMA image bands + 6 INFORM variables):

$$\begin{bmatrix} Band1_1 & \dots & Band231_1 & Cab_1 & Cw_1 & Cm_1 & LAIs_1 & CD_1 & SD_1 \\ Band1_2 & \dots & Band231_2 & Cab_2 & Cw_2 & Cm_2 & LAIs_2 & CD_2 & SD_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Band1_{316,800} & \dots & Band231_{316,800} & Cab_{316,800} & Cw_{316,800} & Cm_{316,800} & LAIs_{316,800} & CD_{316,800} & SD_{316,800} \end{bmatrix}$$

In this matrix, $Band1_n$, ..., $Band231_n$ correspond to the simulated reflectance for the corresponding image band in the simulation number n. Cab_n, Cw_n, Cm_n, $LAIs_n$, CD_n and SD_n refer to the values of the INFORM parameters that were used in the nth simulation.

2.4 Statistics of simulated data and PRISMA image

The Figure 2.2 shows statistical information (mean and mean ± standard deviation) calculated from the LUT and PRISMA image:

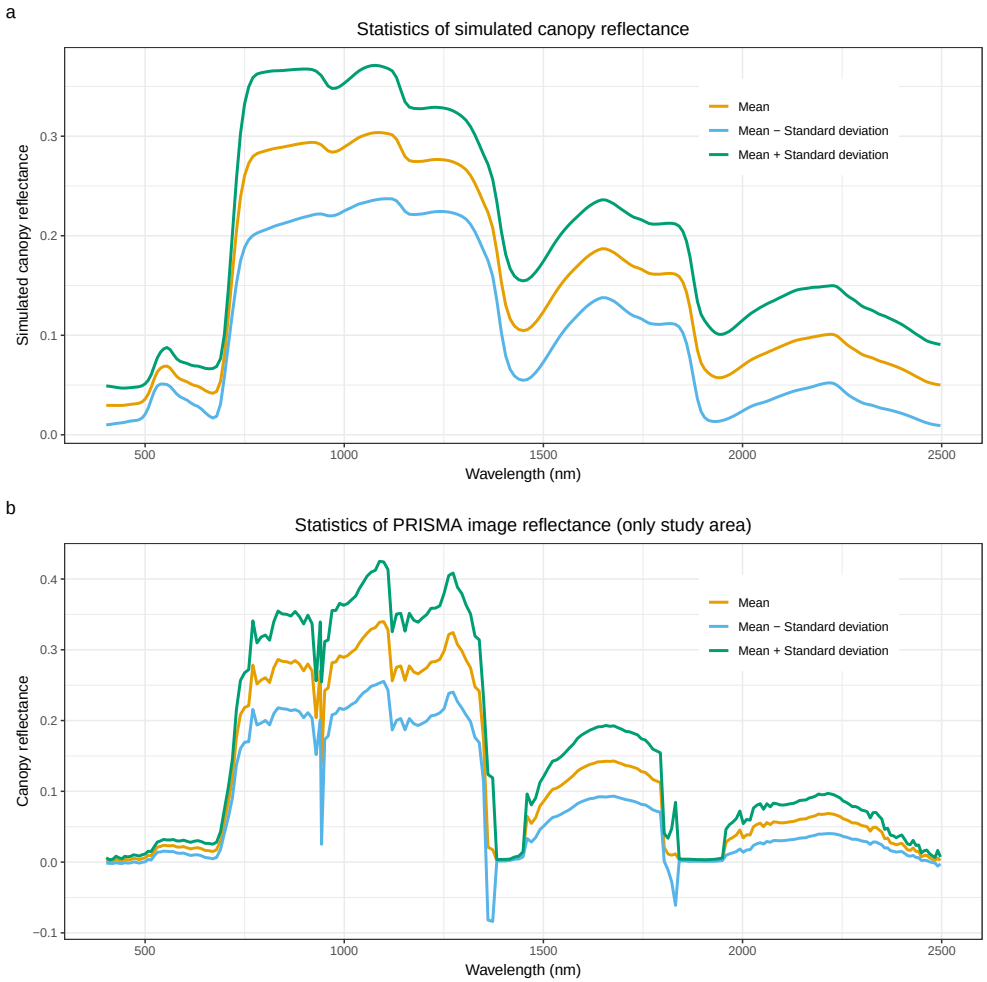

Figure 2.2: Mean and mean ± standard deviation in the a) LUT and b) PRISMA image

Mean and standard deviation within the LUT are much smoother compared to mean and standard deviation within the PRISMA image spectra. This is due to the fact that INFORM model does not add noise during the simulation which can commonly exist in remote sensing images. There is a noticeable amount of noise in the PRISMA image spectra. Some of the noise in the image spectra could potentially be due to the fact that the PRISMA image contained cloud and shadow

within the study area and although most of the cloud and shadow pixels were masked, the nearby pixels could still be affected.

The Figure 2.3 shows the difference between averaged reflectance within the simulated database (LUT) and PRISMA image.

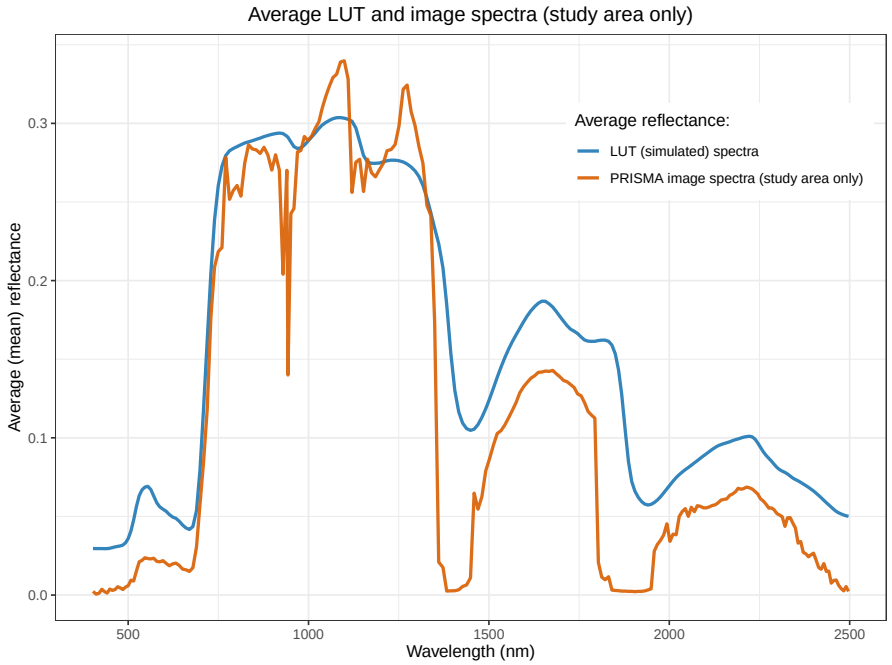

Figure 2.3: Difference between averaged LUT and PRISMA image reflectance

The LUT appears to have higher average reflectance within the visible spectra compared to the PRISMA image spectra. Differences within the water absorbtion bands can also be clearly seen. There is relatively good agreement within the NIR spectrum.

2.5 Gaussian noise

The Figure 2.4 shows the effect of adding 3% Gaussian noise to the simulated data.

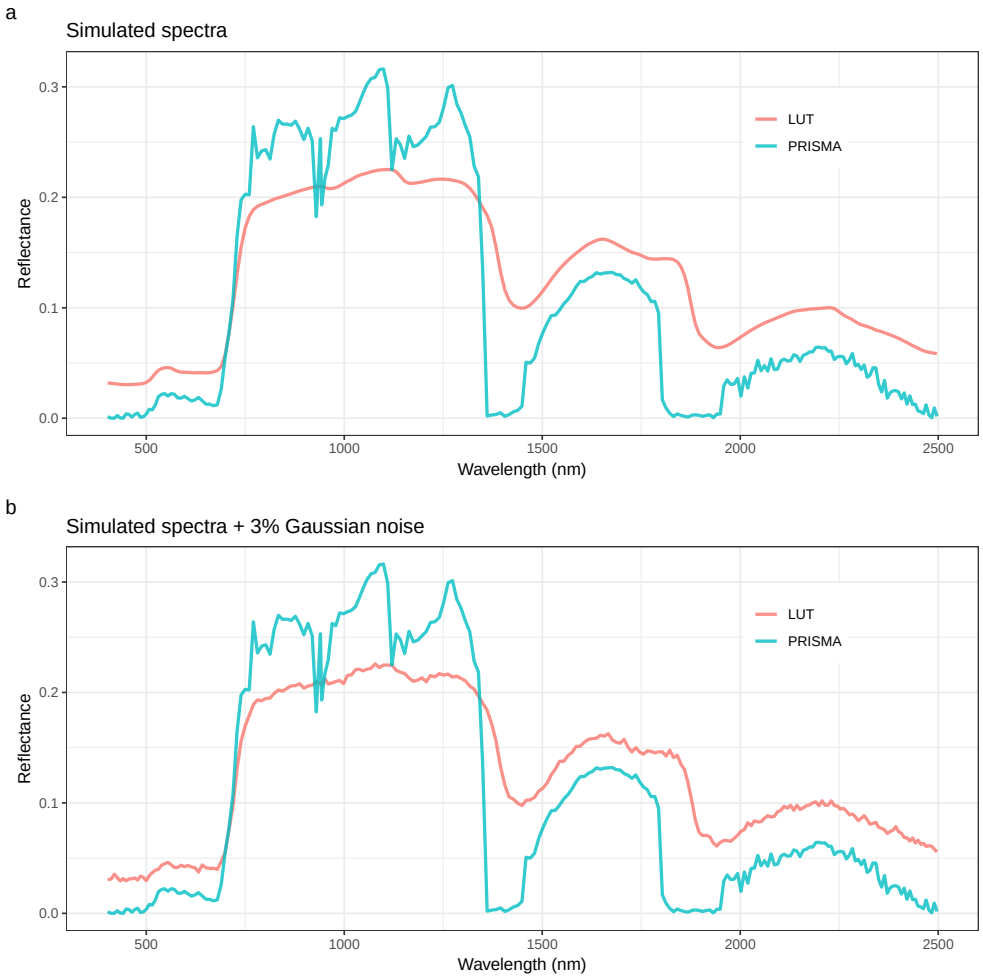

Figure 2.4: Effect of adding 3% Gaussian noise to the simulated spectra. The randomly chosen pixel from the PRISMA data was plotted to illustrate the noise found typically in the image

The Figure 2.4.a shows a simulated spectra that seems perfectly smooth. However, after adding 3% Gaussian noise, the simulated spectra is not as smooth anymore and contains random noise all over the whole spectra (Figure 2.4.b). This also makes the simulated spectra more similar to the pixel extracted from the PRISMA image.

2.6 Principal Component Analysis (PCA)

The result of PCA showed that most of the variation in the simulated data can be explained by much fewer variables (PCs):

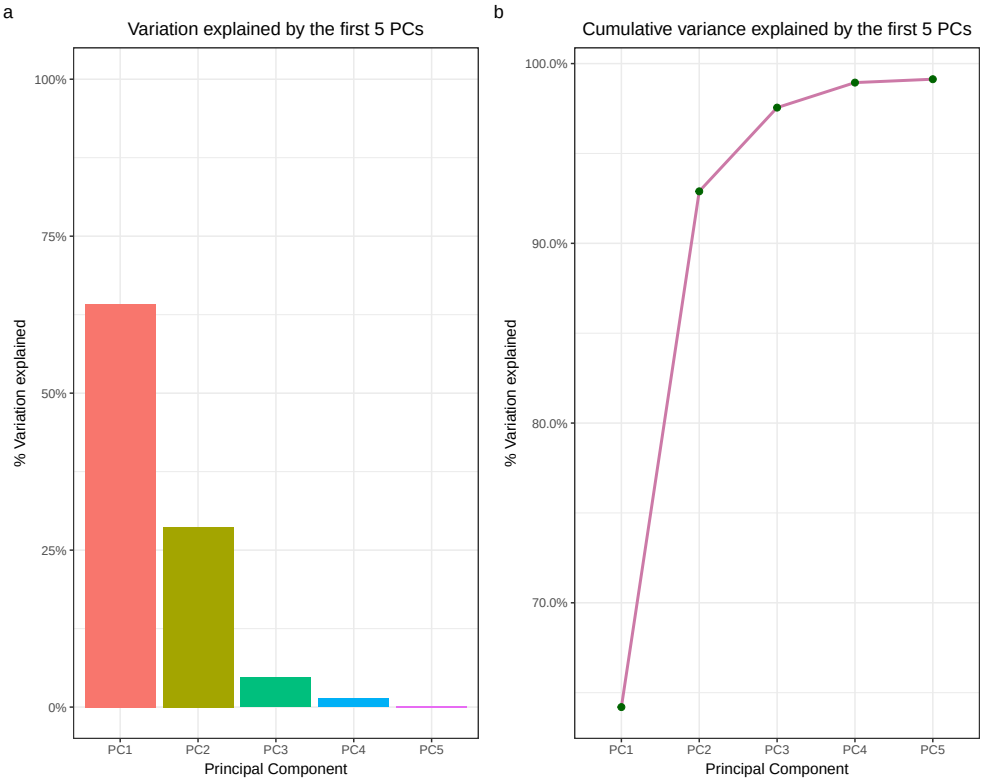

Figure 2.5: Principal Component Analysis: a) Screeplot, b) Cumulative variance explained by the first 5 PCs

The Figure 2.5.a shows the screeplot of the PCA result. The first PC explains the most of the variation and together with the next 4 PCs we can capture more than 99% of the variation that is present in the original data (Figure 2.5.b). This, once again shows the multicollinearity problem with hyperspectral remote sensing data.

References

Ali, A. M., Darvishzadeh, R., Skidmore, A., Gara, T. W., and Heurich, M. Machine learning methods' performance in radiative transfer model inversion to retrieve plant traits from sentinel-2 data of a mixed mountain forest. *International Journal of Digital Earth*, pages 1–15, 2020.

Bro, R. and Smilde, A. K. Principal component analysis. *Analytical methods*, 6(9): 2812–2831, 2014.

Corporation, M. and Weston, S. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2020. URL `https://CRAN.R-project.org/package=doParallel`. R package version 1.0.16.

Danner, M., Berger, K., Wocher, M., Mauser, W., and Hank, T. Efficient rtm-based training of machine learning regression algorithms to quantify biophysical & biochemical traits of agricultural crops. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:278–296, 2021.

Darvishzadeh, R., Skidmore, A., Abdullah, H., Cherenet, E., Ali, A., Wang, T., Nieuwenhuis, W., Heurich, M., Vrieling, A., O'Connor, B., et al. Mapping leaf chlorophyll content from sentinel-2 and rapideye data in spruce stands using the invertible forest reflectance model. *International Journal of Applied Earth Observation and Geoinformation*, 79:58–70, 2019.

Kuhn, M. and Wickham, H. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL `https://www.tidymodels.org`.

Kuhn, M. and Wickham, H. *recipes: Preprocessing Tools to Create Design Matrices*, 2021. URL `https://CRAN.R-project.org/package=recipes`. R package version 0.1.16.

Laurent, V. C., Verhoef, W., Clevers, J. G., and Schaepman, M. E. Inversion of a coupled canopy–atmosphere model using multi-angular top-of-atmosphere radiance data: A forest case study. *Remote Sensing of Environment*, 115(10):2603–2612, 2011.

Lehnert, L. W., Meyer, H., Obermeier, W. A., Silva, B., Regeling, B., Thies, B., and Bendix, J. Hyperspectral data analysis in R: The hsdar package. *Journal of Statistical Software*, 89(12):1–23, 2019. doi: 10.18637/jss.v089.i12.

Microsoft and Weston, S. *foreach: Provides Foreach Looping Construct*, 2020. URL `https://CRAN.R-project.org/package=foreach`. R package version 1.5.1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

Rivera-Caicedo, J. P., Verrelst, J., Muñoz-Marí, J., Camps-Valls, G., and Moreno, J. Hyperspectral dimensionality reduction for biophysical variable statistical retrieval. *ISPRS journal of photogrammetry and remote sensing*, 132:88–101, 2017.

Schlerf, M. and Atzberger, C. Vegetation structure retrieval in beech and spruce forests using spectrodirectional satellite data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):8–17, 2012.

Visser, M. D. *ccrtm: Coupled Chain Radiative Transfer Models*, 2021. URL `https://CRAN.R-project.org/package=ccrtm`. R package version 0.2.

Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.