GEOCARTO INTERNATIONAL
2024, VOL. 39, NO. 1, 2387087
https://doi.org/10.1080/10106049.2024.2387087

OPEN ACCESS

Hybrid retrieval of grass biophysical variables based-on radiative transfer, active learning and regression methods using Sentinel-2 data in Marakele National Park

Philemon Tsele[a] and Abel Ramoelo[b]

[a]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa; [b]Centre for Environmental Studies, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa

ABSTRACT

Biophysical variables such as leaf area index (LAI) and leaf chlorophyll content (LCC) are cited as essential biodiversity variables. A comprehensive comparison and integration of retrieval methods is needed for the estimation of biophysical variables such as LAI and LCC over a multispecies grass canopy. This study tested an assortment of five potentially robust, nonparametric regression methods (NPRMs) for inversion of radiative transfer model (RTM) to retrieve grass LAI and LCC in the Marakele National Park (MNP) of South Africa. The NPRMs used were, namely (i) Partial least squares regression (PLSR), (ii) Principle components regression (PCR), (iii) Kernel ridge regression (KRR), (iv) Random forest regression (RFR), and (v) K-nearest neighbours regression (KNNR). Furthermore, the study attempted to constrain the inversion process by using active learning (AL) techniques which ensured the selection of informative samples from a large pool of RTM simulations. Results show the most accurate grass LAI and LCC retrievals had lower relative root mean squared errors (RRMSEs) of 39.87% and 16.58% respectively. These findings have significant implications for the development of transferable rangeland monitoring systems in protected mountainous regions.

ARTICLE HISTORY
Received 14 November 2023
Accepted 26 July 2024

KEYWORDS
Leaf area index (LAI); leaf chlorophyll content (LCC); Sentinel-2 imagery; active learning; PROSAIL

Introduction

Retrieval of vegetation biophysical variables is important for biodiversity monitoring, spanning ecological and agricultural applications. Biodiversity monitoring is a key component of protected area management and planning. Therefore biophysical variables such as leaf area index (LAI), leaf chlorophyll content (LCC) and canopy chlorophyll content (CCC) are cited as essential biodiversity variables (EBVs) that can be used to assess and monitor the vegetation state at varying spatial scales (Skidmore et al. 2021). In particular, LAI is defined as the one-sided leaf area per unit of horizontal surface area (Jonckheere

CONTACT Philemon Tsele philemon.tsele@up.ac.za

et al. 2004). It is an important indicator of vegetation structure and growth, and also forms an essential input in climate models to determine ecosystem productivity. Another biophysical variable is called the LCC, which is usually obtained through averaged SPAD (dimensionless) leaf chlorophyll measurements. LCC carries valuable information about vegetation physiology and could be regarded as a key indicator of plant health status. Accurate measurements of LCC can be helpful for precision management of natural resources and agricultural fields (Bei et al. 2019). Furthermore, the CCC, which refers to the overall amount of chlorophyll *a* and *b* pigments in a compact group of plants per unit ground area (Gitelson et al. 2005) is derived from the product of the LCC, µg.cm^{-2} and the corresponding LAI, m^{2}.m^{-2} in a subplot (Darvishzadeh et al. 2008). CCC is an important indicator of vegetation health condition, plant species diversity and forage quality assessment (Ali et al. 2020).

Remote sensing provides an alternative method to expensive and time consuming field campaigns, particularly for biodiversity monitoring through EBVs over broad spatial extents on a regular basis, spanning a long period of time (Myneni et al. 2002). For example, a number of available global vegetation biophysical products are generated from coarse to moderate spatial resolution satellite sensors such as, Advanced Very High Resolution Radiometer (AVHRR) (García-Haro et al. 2018), Moderate Resolution Imaging Spectroradiometer (MODIS) (Jia et al. 2019, Disney et al. 2016), PROBA-Vegetation (Baret et al. 2013), ENVISAT Medium Resolution Imaging Spectrometer (MERIS) (Dash and Curran 2004; Bacour et al. 2006) on a regular basis over different time periods. However, their relatively coarse spatial resolutions could make it difficult for the products to provide reliable estimations of vegetation biophysical properties, particularly in heterogenous ecosystems on a local scale (Lv et al. 2021). Recently, the global Sentinel-2 Level 2 Prototype Processor (SL2P) allows the generation of vegetation biophysical estimates at high spatial (∼20 m) and temporal (∼5 days) resolution from Sentinel-2 imagery (Weiss and Baret 2020). However, the global SL2P reported inadequate retrievals of LAI, CCC and fractional vegetation cover (FVC) over two large national parks in South Africa characterised by multiple grass species, diversity of land cover and varying terrain slopes (Tsele et al. 2022).

The estimation of vegetation biophysical variables from remote sensing data, is carried out using three approaches, namely the empirical methods, radiative transfer models (RTMs) and hybrid methods (Verrelst et al. 2015). Vast amount of literature is available on using empirical methods such as parametric or non-parametric regression methods due to their inherent simplicity (Verrelst et al. 2015) in obtaining statistical relationships between the biophysical variable of interest and its corresponding reflectance. The RTMs on the other hand, have minimum reliance on in-situ data in that, they use the physical laws (Goel, 1987) to accurately describe the spectral variation of canopy reflectance as a function of viewing and illumination geometry, leaf, canopy and soil background characteristics (Darvishzadeh et al. 2011). However, it was reported that RTMs still require local parameterization in order to simulate multispecies canopies accurately, especially in heterogenous environments (Combal et al. 2003; Darvishzadeh et al. 2008; Bsaibes et al. 2009; Atzberger et al. 2015). The third approach is hybrid retrieval schemes and these entail the integration regression methods with RTM data. Basically, a regression model is trained using large database of RTM-simulated reflectance data, in-order to retrieve the biophysical variable of interest. A major challenge with this approach is that a portion of the RTM-data may contain redundant and potential outliers which do not improve on the prediction accuracy of the resulting regression model (Verrelst et al. 2016). In particular, RTM-data can potential have outliers which are basically, simulated samples of reflectance that are exceedingly higher and/or lower beyond the ideal reflectance range. One of the ways in addressing this issue is through the use of active learning (AL) sample selection algorithms to: (i) disregard the non-diverse and possible outliers from the large pool of RTM-simulated reflectance samples, and (ii) optimise the simulated training dataset to contain only intelligent or informative samples needed for improving the regression model's retrieval accuracy (Pasolli et al. 2012).

A hybrid approach of integrating parametric and/or non-parametric methods with AL algorithms using RTM data has been widely tested in agricultural environments or crop related studies, for example Verrelst et al. (2016); Verrelst et al. (2020); Berger et al. (2021); Candiani et al. (2022); Pascual-Venteo et al. (2022); and Wocher et al. (2022). These aforementioned studies have successfully demonstrated improved retrieval accuracies of biophysical variables such as the LAI and LCC. However, very few studies were found on using this hybrid approach in other natural ecosystems such as mangrove forests, but not in heterogeneous grasslands with combinations of different grass species distributed over a mountainous region. For example, Binh et al. (2022) used an AL-based PROSAIL hybrid model to retrieve mangrove LAI from Sentinel-2 data with high accuracy i.e. RMSE 0.13 $m^2.m^{-2}$.

One of the most widely used RTM is PROSAIL (Baret et al. 1992). PROSAIL has by far, become the most popular model in the scientific community for vegetation characterization due to ease of use, robustness and consistent validation. For example, Masemola et al. (2016) estimated the grassland LAI from Landsat 8 imagery using a combination of the PROSPECT leaf optical RTM (Jacquemoud and Baret 1990) and the SAIL canopy reflectance RTM (Verhoef 1984) in the Mpumalanga region of South Africa. Further, Cho et al. (2014) found accurate estimates of LAI in three South African biomes (grassland, Karoo and Forest) by inverting the PROSAIL on the MODIS 250 m imagery when compared to the acquired MODIS LAI product. As a result, it was suggested that the PROSAIL be applied at national or sub-continental scale to produce LAI time series output for assessing the impact of land use and climate change within a given landscape. Atzberger et al. (2015) compared the PROSAIL based on look-up-tables (LUTs), predictive equations and narrow-band vegetation indices for the accurate estimation of LAI in the Mediterranean grassland within the Majella National park, Italy. Their results found that LUT-based PROSAIL inversion had the highest accuracy of LAI estimation compared to other methods; and thus, could be useful for the monitoring and managing National Parks including endangered habitats. Furthermore, Darvishzadeh et al. (2008) inverted the PROSAIL for the retrieval of LAI in a heterogenous grassland canopy using hyperspectral data and found intermediate accuracies, which suggests PROSAIL does not adapt well to heterogenous grasslands or multi-species canopies. The current study further investigated this notion by using a hybrid retrieval approach that integrates PROSAIL RTM data with AL algorithms and nonparametric regression methods using Sentinel-2 multispectral imagery in the mesic Savanna of Marakele National Park (MNP) in South Africa. To our knowledge, the current study was the first to test such hybrid approach in a multi-species grassland canopy using Sentinel-2 multispectral data for the retrieval of grass LAI and LCC during peak productivity.

The nonparametric regression methods (NPRMs) considered in this study were, namely (i) Partial least squares regression (PLSR), (ii) Principle components regression (PCR), (iii) Kernel ridge regression (KRR), (iv) Random forest regression (RFR), and (v) K-nearest neighbours regression (KNNR). Furthermore, the following AL algorithms were tested in this study: angle-based diversity (ABD), cluster-based diversity (CDB), euclidean distance-based diversity (EBD), pool active learning (PAL), random sampling (RS) and residual active learning (RSAL). The objectives of this paper were to: (i) compare the performance of linear and non-linear NPRMs trained with large database of simulated canopy reflectance samples for the estimation of LAI and LCC, over a multi-species grass canopy located in a protected mountainous region; (ii) apply several AL sample selection algorithms in-order to disregard the non-diverse and possible outliers from the large pool of RTM-simulated reflectance samples; and (iii) optimise the simulated training dataset to contain only intelligent or informative samples needed for improving the regression model's retrieval accuracy. This study has significant implications for the development of transferable rangeland monitoring systems in protected mountainous regions.

Material and methods

Study area description

The study site encompasses the entire Marakele National Park (MNP) which is a South African National Park located between 27°26'30"E, 24°16'30"S and 27°48'30"E, 24°33'0"S in the Waterberg district and mountains of the Limpopo province (Figure 1). The study site was selected based on key location attributes, which encompassed the savanna and grassland biomes and different vegetation communities, according to the national vegetation map (Mucina and Rutherford (2006).

Furthermore, MNP is mountainous and characterized by surface height variation that range between approximately 976 m to 2091 m, estimated from the 30 m resolution Shuttle Radar Topography Mission (SRTM) data acquired from the United States Geological Survey (USGS) Earth Explorer (https://earthexplorer.usgs.gov/). The site falls within the summer rainfall region of South Africa, and can receive average rainfall of up to around 630 mm annually (Van Staden and Bredenkamp 2005).

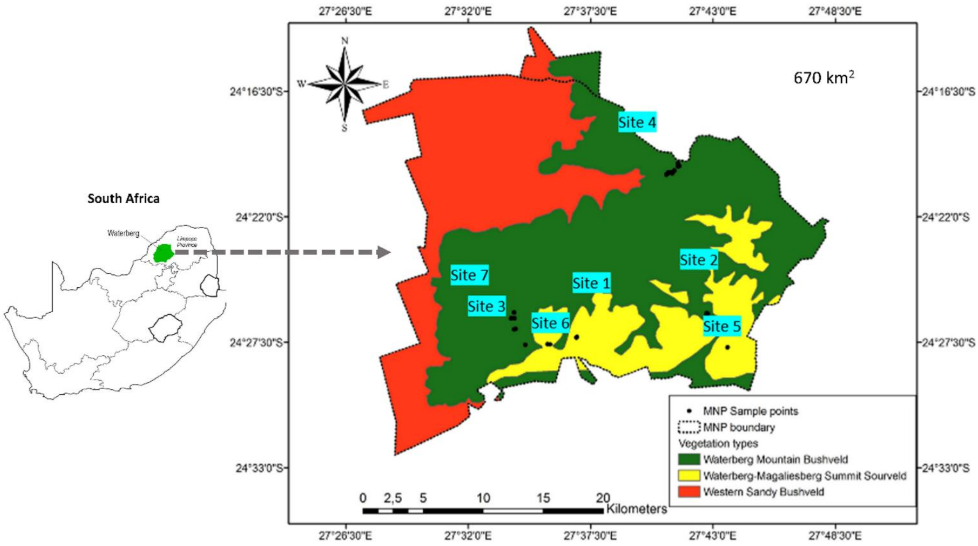

Figure 1. The marakele National Park (MNP) covering an area of about 670 km^2 is located in the mountainous waterberg region of South Africa, where leaf area index (LAI) and leaf chlorophyll content (LCC) field sample measurements across various sites within the park were taken on 8 – 10 april 2021 respectively.

Schematic workflow

Figure 2 show a schematic hybrid retrieval workflow summarising the various phases of the methodology that were implemented in this study. These phases are discussed in subsequent sections of the paper.

Field data collection

Field data collection in the study area took place on 8 – 10 April 2021 during peak productivity for the natural heterogenous grasses. The total number of sampled locations were 68 in Marakele National Park (MNP) respectively. The sampling strategy involved a combination of stratified and purposive sampling methods (Lv et al. 2021). Random samples were initially taken across different grass vegetation communities and varying slope terrains spanning the crests, valleys and low to mid-slopes. However, when in the field, there were certain inaccessible areas, which led to the use of purposive sampling where replacement of the sampled locations was done, close to the randomized points. Each selected sample location represented a plot with a size of 20 m x 20 m and within that plot, two subplots of size 1 m x 1 m spaced apart were taken in-order to capture variability within each plot. A number of recordings were taken in each subplot namely, the (i) subplot number and photo (ii) geographic coordinates using the Global Navigation Satellite System – Real Time Kinematic (GNSS-RTK) method (Schloderer et al. 2011), (iii) leaf area index (LAI) using the ACCUPAR LP-80 ceptometer, (iii) leaf chlorophyll content (LCC) using the SPAD 502 Plus chlorophyll meter, and (iv) Grass height (cm) using the disk pasture meter. Field data collection took place on 8 – 10 April 2021 in MNP.

In this study, LAI readings were performed under generally clear skies with intermittent cloud cover from the late morning hours at about 10:00 until early afternoon at around 14:00 in-order to minify variations of the sun zenith angle among the subplots. Moreover, in each subplot we used the SPAD 502 Plus chlorophyll meter to take

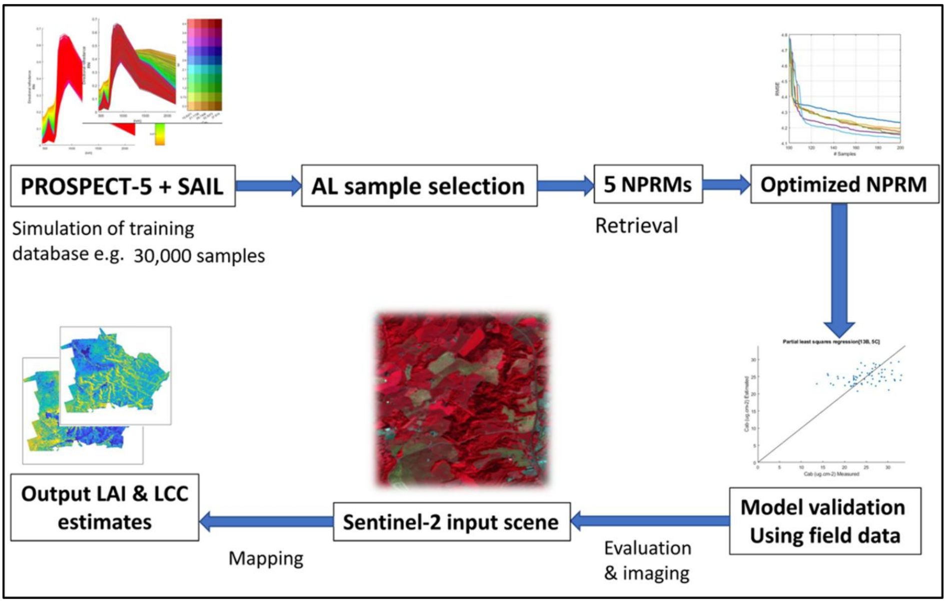
PROSPECT-5 + SAIL
Simulation of training database e.g. 30,000 samples
AL sample selection
5 NPRMs
Retrieval
Optimized NPRM
Model validation Using field data
Evaluation & imaging
Sentinel-2 input scene
Mapping
Output LAI & LCC estimates
Partial least squares regression(130, 5C)

Figure 2. Hybrid retrieval workflow including PROSAIL, nonparametric regression methods (NPRMs) and active learning (AL) sample reduction for estimation of grass leaf area index (LAI) and leaf chlorophyll content (LCC).

dimensionless chlorophyll readings of five randomly selected green leaves, representing the dominant species and recorded the average chlorophyll reading. The average chlorophyll readings (i.e. SPAD measurements) of all subplots were converted into LCC per unit area, µg.cm^{-2} by applying an empirical calibration method described in Markwell et al. (1995). In overall, more than 30 grass species were identified within 68 subplots during field work in MNP such as (to name a few), *Hyperthelia dissoluta, Hyparrhenia hirta, Cymbopogon excavates, Eragrostis lehmanniana, Themeda triandra, Digitaria eriantha, Sporobolus africanus, Miscanthus junceus, Digitaria Brazzae, Aristida diffusa, Eragrostis racemosa, Schizachyrium jeffisi* and *Panicum natalense.*

Remotely-sensed imagery

The acquisition of Sentinel-2 data was free of charge from the European Space Agency data hub (https://scihub.copernicus.eu/dhus/#/home) on the 9th of April 2021. The selection of this image was such that (i) it is free from any cloud obscuration (ii) it covered the study site and (iii) it had the acquisition date that was very close (i.e. $<=$ 6 days) to the field data collection dates. Sentinel-2 data has 13 spectral bands, characterised by fine spatial resolutions in the range 10-60 m, that cover large geographic areas (i.e. 120 km × 120 km per scene) at high a temporal resolution of up-to 5 days (Frampton et al. 2013). The Sentinel-2 image was pre-processed to bottom of the Atmosphere (BOA) reflectance i.e. Level-2A using the Sentinel Application Platform (SNAP) Sentinel-2 atmospheric correction tool, Sen2Cor, version 2.8 (Louis et al. 2016). The Sentinel-2 BOA image was resampled to the spatial resolution of 20 m in order to correspond to single field plots of size of 20 m x 20 m that contained two subplots, each of size 1 m x 1 m.

PROSAIL model parameterisation

The PROSAIL radiative transfer models (RTMs) i.e. PROSPECT and SAIL models Jacquemoud et al. (2009), as well as model inversion and regression algorithms are available to the public in a toolbox named Automated Radiative Transfer Models Operator (ARTMO) (https://artmotoolbox.com/). This tool runs in MATLAB and provides a wide range of functions that are key for executing RTMs and applying inversion algorithms both at the leaf and canopy level. In addition, the Graphical User Interface (GUI) or non-GUI versions of the MATLAB scripts for the aforementioned RTMs and inversion algorithms can be obtained from http://teledetection.ipgp.jussieu.fr/prosail/.

In this study, the PROSAIL model (Jacquemoud et al. 2009) was used for simulating the reflectance of the grassy leaf canopy in Marakele National Park (MNP) based on a combination of adopted and site-specific model parameters (Table 1). In particular, first the parameterisation of the PROSPECT-5 model was done which included the following model inputs: leaf chlorophyll content (LCC), carotenoid content, brown pigments, leaf water content and dry matter content (Table 1). Secondly, the parameterisation of the SAIL model was done, and all its model inputs parameters are shown in Table 1. During parameterisation, the range values for LCC and leaf area index (LAI) were based on actual field measurements reported in Tsele et al. (2022). The gaussian distribution function (Table 1) was adopted for the LCC and LAI field measurements due to the proximity of their measures of central tendency (Tsele et al. 2022). For other parameters, a uniform distribution function was assumed meaning the range of values could have equal or constant probability. Other studies made a similar aforementioned assumption (e.g. Darvishzadeh et al. (2008); Verrelst et al. (2016); Darvishzadeh et al. (2008)) and this,

Table 1. PROSAIL model parameterisation. Ave: average or mean, StDev: standard deviation.

Model parameters	Unit	Range	Distribution	Source
Leaf parameters: PROSPECT-5 model				
Leaf chlorophyll content (LCC)	[µg/cm^2]	13.60 – 33.10	Gaussian (Ave: 24.93; StDev: 4.37)	Tsele et al. (2022)
Leaf structure (N)	Dimensionless	1.5 – 1.9	Uniform	Masemola et al. (2016)
Carotenoids	[µg/cm^2]	0 – 25	Uniform	Masemola et al. (2016)
Leaf water content (LWC)	[g/cm^2]	0.01 – 0.02	Uniform	Masemola et al. (2016)
Brown pigments	Dimensionless	0 – 1	Uniform	Masemola et al. (2016)
Dry matter	[g/cm^2]	0.0025 – 0.0050	Uniform	Masemola et al. (2016)
Canopy parameters: 4SAIL model				
Leaf area index (LAI)	[m^2/m^2]	0.47 – 5.00	Gaussian (Ave: 1.90; StDev: 0.84)	Tsele et al. (2022)
Average leaf angle (ALA)	[▫]	20 – 70	Uniform	Masemola et al. (2016)
Hot spot effect	[m/m]	0.05 – 0.10	Uniform	Masemola et al. (2016), Darvishzadeh et al. (2008)
Ratio of diffuse to downward irradiance	[fraction]	0.1	Fixed	Masemola et al. (2016), Darvishzadeh et al. (2008)
Soil brightness coefficient	Dimensionless	1	Fixed	Masemola et al. (2016)
Solar zenith angle	[▫]	40.71	Fixed	Sentinel-2 image Metadata
View zenith angle	[▫]	4.83	Fixed	Sentinel-2 image Metadata

partly may have been due to the lack of actual field data for some of those parameters. The geometrical parameters such as the solar zenith and view zenith angles were obtained from the Sentinel-2 image metadata file. Furthermore, the range values for the remaining parameters were obtained from published studies that were conducted over similar vegetation type, dominant grass species and environmental setting. In overall, this process is expected to produce parameter-driven, simulated leaf and canopy spectral reflectance, which will be stored in a database for applying and testing nonparametric regression methods (NPRMs) and active learning (AL) techniques.

PROSAIL-simulated spectra

The sensor settings related to Sentinel-2 multispectral imager (MSI) were chosen with 12 bands that range from 443 to 2190 nm. PROSAIL-simulated data were generated containing a large pool of 30,000 samples of synthetic canopy reflectance, stored in a lookup table (LUT) database in ARTMO. Other studies have explored working with a higher number of samples up to 100,000 (Darvishzadeh et al. 2008; Masemola et al. 2016), however this was found to create largely redundant samples for regression (Verrelst et al. 2016). In this study, the generated large pool of RTM-simulated data was used for training the NPRMs for the retrieval of grass LAI and LCC in MNP. Furthermore, the study attempted to constrain the inversion process by using AL methods which ensured the selection of only the best possible samples from a large pool of RTM-simulations for use by the best performing NPRM.

NPRM configuration

This study evaluated five nonparametric regression methods (NPRMs), widely used in the literature for estimating vegetation biophysical variables (Verrelst et al. 2015). The

NPRMs used were: (i) Partial least squares regression (PLSR), (ii) Principle components regression (PCR), (iii) Kernel ridge regression (KRR), (iv) Random forest regression (RFR), and (v) K-nearest neighbours regression (KNNR). These methods are data driven, they define the regression function based on input data, and can optimise the regression model by learning the training data (Verrelst et al. 2019).

In particular, PLSR and PCR are classical linear NPRMs and were chosen in this study based on their simplicity, efficiency and widely reported predictive power in the estimation of vegetation biophysical variables (Atzberger et al. 2010). PLSR uses the matrix inversion algorithm (Geladi and Kowalski 1986) to find important variations in the spectral data that are relevant for estimating the biophysical variable(s) of interest. PLSR reduces data dimensionality by transforming the input features into a small number of statistically independent linear combinations (Okujeni et al. 2014). During parameter setting, this number was set to 5 in ARTMO software package. On the other hand, PCR uses principle components analysis (PCA) to transform the input spectral data into variable components, and thereafter performs linear regression for estimating the regression coefficients of the most relevant components also called PCA scores (Wold et al. 1987). During parameter setting, we used all 13 bands of the Sentinel-2 data as input into the PCR in ARTMO. PCR could handle band redundancy by converting the spectral data to a lower dimensional space. Overall, the linear NPRMs such as the PLSR and PCR may not be flexible particularly when dealing with complex non-linear relations (Verrelst et al. 2019).

As part of the evaluation exercise undertaken in this study, nonlinear NPRMs (popularly known as machine learning methods) were also considered because they (i) are data driven (ii) do not make underlying assumptions on the data distribution, and (iii) optimise the regression model through a learning phase. In addition, based on literature (Verrelst et al. 2015) they have demonstrated their capability in applying nonlinear transformations and enhanced flexibility in capturing nonlinear relationships of image features. In particular, the KRR is a supervised learning model that make use of kernel functions for data analysis and pattern identification (Hastie et al. 2009). KRR is a family of the Least squares support vector machine classifiers (Suykens and Vandewalle 1999) which map the training samples into a higher dimensional feature space and builds a regression function which represents a nonlinear regression in the original input space (Saunders et al. 1998). An optimal function would minimize the squared residuals and lead to improved biophysical variable retrieval. KRR in ARTMO software package required the tuning of the kernel function, regularization and optimization parameters. The optimization was carried out using the standard cross validation procedure.

Another nonlinear NPRM used in this study was the KNNR which in principle, computes the distance between a data record (or new point) and all of the reference data records (or predefined number of training samples) using the traditional Euclidean distance method (Cover and Hart 1967). It looks for the closest number of records (defined by *k*) to the new point, and considers the records that have a majority class in-order to predict a label for the new point. KNNR can yield useful results particularly if the training samples are well distributed in the dimensional feature space (Hardin 1994). Our user specified parameter *k* in ARTMO was set to 5 which was found to make the records selection and label prediction process by KNNR to be stringent and relatively quick.

Lastly, a third nonlinear NPRM that has been used in a variety of remote sensing studies for retrieval of vegetation biophysical properties is the Random forest (RF). The RF method is an ensemble machine-learning algorithm (Breiman 2001) that builds an assortment of multiple decision trees. RF is an extension of the Classification and Regression

Trees (CART) algorithm (Breiman et al. 2017) and it is suited to predict both discrete and continuous variables. RF, hereafter RFR has been found in other studies to be potentially more accurate and relatively robust to outliers, when compared to other nonlinear NPRM methods such as the individual decision trees and neural networks (Mutanga et al. 2012; Rodriguez-Galiano et al. 2012; Chen et al. 2014; Liang et al. 2016). For every tree that is grown in a RFR, a new training set of size m is randomly selected with replacement from the original training set of size M (where $m < M$). The proportion of samples that is not selected in the original training set, is left out-of-bag (OOB) and used to estimate the model performance and variable importance. Furthermore, for each node of the tree, there are X input variables (e.g. spectral bands) from which only x number of variables of out the X are randomly selected for determining the optimal split at that node for growing a forest of trees. The unclassified pixel is run through each of the generated trees, and each tree would then classify this pixel into one of the Y classes (as defined in the training data set). Finally, the pixel would be assigned to the class that had the most classifications i.e. majority vote.

Active learning techniques

Active learning (AL) techniques use selection criterion algorithms (MacKay 1992) to select informative samples from a large synthetic training database in-order to improve the model's estimation accuracy (Pasolli et al. 2012). The AL techniques used in this study falls under two main categories namely, uncertainty criteria algorithms and diversity criteria algorithms. The former, uses variance-based algorithms (Douak et al. 2011) to select from a large pool of samples, only the those with the least confidence (Figure 3). Whereas, the latter uses a variety of distance-related metrics (Demir et al. 2010; Patra and Bruzzone 2012; Douak et al. 2013) to select the most diverse samples and thereby disregarding the redundant samples from a large pool of RTM-simulated reflectance samples.

Figure 3 shows the flow of fundamental stages followed by the uncertainty criteria algorithms used in this study, namely the Pool active learning (PAL: Douak et al. 2013) and Residual active learning (RSAL: Douak et al. 2011). The difference between the two (as depicted in Figure 3) is that, PAL trains a statistical regressor to obtain predictions based a subset of random labelled samples drawn from a large pool of unlabelled RTM reflectance simulations. Thereafter, PAL calculates the variance for each prediction and ranks the different predictions according to the variance. A selection of the samples related to the predictions with the highest variance values is performed. These samples represent greater disagreements between the regressors (Verrelst et al. 2016) and are therefore not considered in optimal final training set. In contrast, RSAL applies the residual model to estimate the prediction error linked to each obtained prediction and ranks the different predictions according to their estimated residual errors (Figure 3). A selection of the samples related to the predictions with the highest prediction errors is performed and are therefore, considered to be the most uncertain samples that will not be considered in optimal final training set (Verrelst et al. (2016)).

Furthermore, the diversity criteria algorithms used in this study were the angle-based diversity (ABD: Demir et al. 2010), Euclidean distance-based diversity (EBD: Douak et al. 2013) and Cluster-based diversity (CDB: Patra and Bruzzone 2012). The ABD algorithm measures the degree of diversity between samples in the initial training set (i.e. subset of n random samples) and those in the RTM-simulated database using the cosine angle distance. The samples with smallest cosine angles are ranked low because they represent samples (also referred to as reflectance-variable pairs) that are redundant and similar to

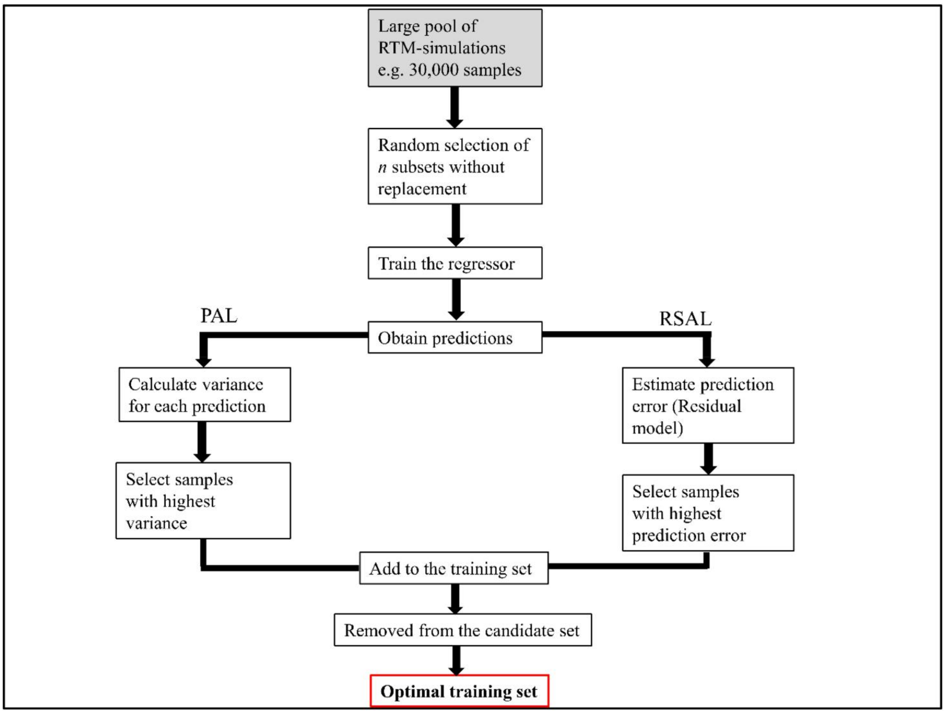
Large pool of RTM-simulations e.g. 30,000 samples
Random selection of *n* subsets without replacement
Train the regressor
PAL
Obtain predictions
RSAL
Calculate variance for each prediction
Estimate prediction error (Residual model)
Select samples with highest variance
Select samples with highest prediction error
Add to the training set
Removed from the candidate set
Optimal training set

Figure 3. Workflow of the uncertainty criteria algorithms used in this study i.e. the Pool active learning (PAL) and residual active learning (RSAL).

those already accounted for in the training set. However, samples with the largest cosine angles are ranked high and added to the training set until it become optimal (Crawford et al. 2013). The EBD algorithm works similarly to ABD, however the difference is that EBD measures the degree of diversity by calculating the Euclidean distances (Douak et al. 2013) between the samples in the initial training set and those in the RTM-simulated database. Samples with the farthest distance are ranked high and added to the training set until it become optimal.

Another diversity criteria algorithm used was the CDB, which is a standard cluster based technique that applies the k-means clustering algorithm (Jain and Dubes 1988) to partition the initial training set into a series of labelled *n* clusters in the feature space. The number of clusters *n*, is set to the number of samples to add in each iteration of the algorithm (Verrelst et al. 2016). The cluster centroid is determined for each cluster and thereafter, iteratively selects the nearest sample (from the large pool of unlabelled synthetic samples) to the cluster centroid. Generally, samples within the same cluster are correlated and (in this case) characterised by minimal variable variations that might produce virtually similar spectra. Therefore, the most informative samples within the clusters would be selected either based on their distribution and/or level of uncertainty (Demir et al. 2010). An improved version of the CDB algorithm couples the diversity measure with uncertainty analysis of the samples (Patra and Bruzzone 2012). Last but not least, the Random sampling (RS) AL algorithm was used in this study. The RS method falls under diversity criteria algorithms and it is considered the most straightforward algorithm in that it gives every sample in the RTM-simulated spectra database equal probability of being selected. Basically, RS selects at random, a pre-defined number of samples within a large pool of unlabelled RTM-simulated spectra, and add them to the training set in-order to obtain an optimised training set.

All six AL algorithms discussed in this section were implemented in ARTMO software package. The key input parameters during the AL phase in ARTMO were the (i) PROSAIL RTM synthetic canopy reflectance database, (ii) training data based on field measurements of LAI and LCC, (iii) AL algorithms, (iv) Selection of the NPRM, and (v) standard cross validation procedure. This process led to a hybrid retrieval approach whereby various optimal training sets obtained by applying the different AL algorithms were applied on the best performing NPRM in-order to test improvement in the retrieval accuracies of the grass LAI and LCC within the MNP.

This hybrid retrieval approach has been successfully tested in several studies, mostly encompassing agricultural applications wherein the environmental setting is largely homogenous e.g. Verrelst et al. (2020); Berger et al. (2021); Candiani et al. (2022); Pascual-Venteo et al. (2022) but very few studies in heterogenous ecosystems i.e. Binh et al. (2022). Therefore, a comprehensive comparison and integration of retrieval methods coupled with AL techniques is needed for a broader understanding of the relative performance of the models over multispecies canopies characterised by diversity of land cover and varying terrain slopes.

Evaluation of model prediction accuracies and performance of the AL algorithms

In this study, the acquired ground observations i.e. 68 field samples in the MNP site of LAI and LCC were used as validation datasets. In particular, the standard cross-validation method (Snee 1977) was used to evaluate the retrieval performance of the NPRMs. In addition, cross-validation was used evaluate the relative performance of the AL algorithms when applied to best performing NPRM. During cross-validation parameterisation in ARTMO software, the LAI and LCC field measurements were randomly divided into $k = 10$ equal-sized sub-datasets. We defined 5 iterative validation steps and, in each step, the k sub-datasets were used only once as a validation dataset for model testing. The hybrid retrieval performance of the NPRMs (with and without the integration of AL algorithms) was evaluated using statistical performance metrics such as the coefficient of determination (R^2), root mean-squared error (RMSE), Relative root mean-squared error (RRMSE) and mean absolute error (MAE). These metrics are widely used in numerous studies involving the estimation of vegetation biophysical and/or biochemical parameters, for example Ali et al. (2021); Kganyago et al. (2021); Verrelst et al. (2015); Ramoelo and Cho (2018); Guerini Filho, Kuplich, and Quadros (2020); Richter, Hank, et al. (2012); Darvishzadeh et al. (2008).

The R^2 shown in Equation (1) was computed for each model to measure the goodness of fit. This was followed by the computation of RMSE shown in Equation (2) which indicate the amount of error expressed in the units of the biophysical variable of interest i.e. m^2.m^{-2} for LAI and μg.cm^{-2} for LCC. RMSE can range from 0 to ∞ and a lower value (closer to 0), indicate an accurate model (Chai and Draxler 2014). Additionally, the RRMSE shown in Equation (3) was used to facilitate comparison of model accuracies between different variables with different data units i.e. LAI and LCC, where model accuracy was regarded as either excellent (RRMSE $<$ 10%), good (10%$<$RRMSE $<$ 20%), fair (20%$<$RRMSE $<$ 30%) or inadequate (RRMSE $>$ 30%) (Jamieson et al. 1991; Heinemann et al. 2012; Richter, Atzberger, et al. 2012). Furthermore, MAE shown in Equation (4) was also used a supplementary metric to RMSE to evaluate model error. The combination of MAE and RMSE metrics gave a representation of the variation in model error distribution, which can be normally- or uniformly distributed (Chai and Draxler 2014).

$$R^2 = 1 - \frac{\sum \left(e_k^N - \bar{e}_k\right)^2}{\sum \left(e_k - \bar{e}_k\right)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N} \left(e_k - m_k\right)^2}{n}} \tag{2}$$

$$RRMSE = \frac{RMSE}{\bar{m}_k} \times 100 \tag{3}$$

$$MAE = \frac{1}{n} \sum\nolimits_{k=1}^{n} \left|e_k - m_k\right| \tag{4}$$

where m_k is the observed biophysical variable i.e. LAI or CCC, and e_k is the model predicted biophysical variable i.e. LAI or CCC, $\bar{m}_k$, and $\bar{e}_k$ denotes the respective means of observed and model predicted biophysical variables, n is the sample size, and N is the number of errors.

Results

Statistical analysis of the field measurements

The field measurements across the 68 subplots (Table 2), resembled an approximately gaussian distribution, which was inferred from the proximity of the respective mean and median values per variable. The leaf area index (LAI) and leaf chlorophyll content (LCC) showed moderate to low variability across the subplots with a coefficient of variation (CV) of about 44% and 16%, respectively. The respective mean and range values of LAI and LCC show that the grasses in the sampled areas were on average green and healthy spanning low to high biomass areas. This variability is important when parameterising the PROSAIL radiative transfer model (RTM) to produce synthetic reflectance that captures a broader range of the grassland vegetation condition across the Marakele National Park (MNP).

Analysis of the PROSAIL RTM data

Figure 4 show the PROSAIL RTM simulations composed of 30,000 samples of the canopy reflectance in MNP. The spectral variation of the canopy reflectance evident across the Sentinel-2 bands, captured the variability of grassland vegetation in-terms of their LAI and LCC within MNP. It is clear that the varying LAI and LCC, influences the grass canopy reflectance within MNP whereby for example, the low LAI and LCC gave a higher spectral response in the visible bands (which could be related to low biomass, sparse and/or dry vegetation) followed by a lower reflectance in the near-infrared (NIR) bands and higher spectral response in the water sensitive bands i.e. 1565-1655 nm and 2100-2280 nm (Figure 4).

Table 2. Summary statistics of measured biophysical variables of grassland sample subplots. The statistical parameters, CV denotes the coefficient of variation, and StDev the standard deviation.

Measured variables	No. of Subplots	Min.	Max.	Mean	Median	StDev	CV
LAI ($m^2.m^{-2}$)	68	0.47	5.00	1.90	1.90	0.84	0.44
LCC ($\mu g.cm^{-2}$)	68	13.6	33.1	24.93	25.0	4.37	0.16

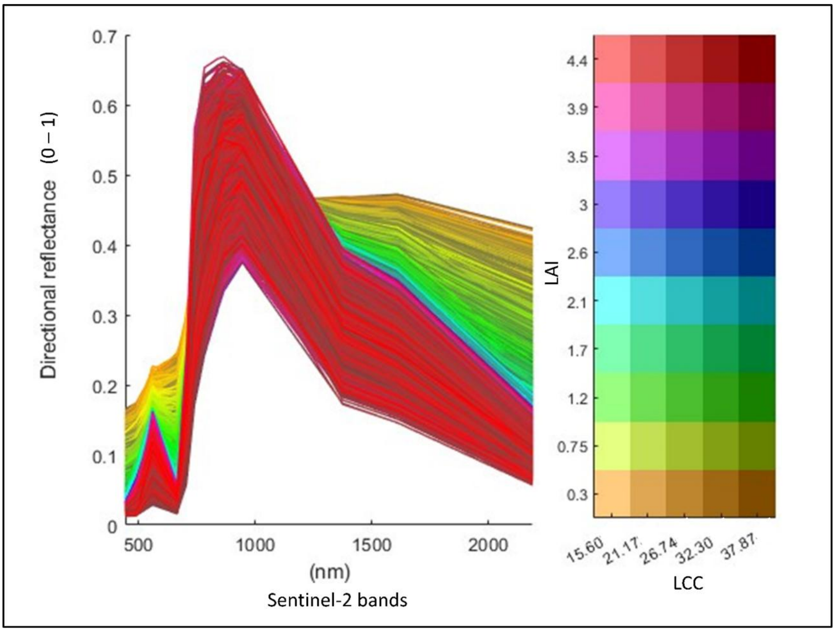

Figure 4. The canopy reflectance in marakele National Park (MNP) simulated using the PROSAIL RTM.

Table 3. The leaf area index (LAI) retrieval performance by various nonparametric regression methods (NPRMs) in marakele National Park (MNP).

NPRM	MAE (m^2/m^2)	RMSE (m^2/m^2)	RRMSE (%)	R^2
PLSR	0.86	1.10	57.60	0.00
RFR	1.05	1.21	63.60	0.08
KRR	1.30	1.81	95.23	0.00
KNNR	1.80	1.93	101.52	0.11
PCR	5.65	6.78	356.90	0.00

In contrast, high LAI and LCC gave lower reflectance in the visible and water sensitive bands and higher NIR reflectance. In overall, these RTM simulations were subsequently used for training the NPRMs for the retrieval of grass LAI and LCC in MNP. Given that these simulations are composed of many samples and not all of them are optimal for improving the model's retrieval accuracy, the AL methods were applied on the RTM simulated data to select informative samples for use by the best performing retrieval model.

NPRMs retrieval performance of LAI and LCC without AL

Table 3 and Table 4 show the retrieval performance of the five NPRMs trained on all PROSAIL RTM simulations for obtaining predictions of the grass LAI and LCC in MNP during peak productivity season of 2021. The LAI retrieval performance for all NPRMs revealed inadequate model accuracies, with RRMSE's generally exceeding 50% (Table 3). The LAI RMSE's for all NPRMs do not fall within the acceptable range typical of a better prediction model i.e. $0.5 \leq \text{RMSE} < 1.0$ (Richter, Atzberger, et al. 2012). Furthermore, the NPRMs explained very little of the LAI variability within MNP according to the low R^2 values approximating zero. Based on the order of RRMSE values, the results suggest PLSR

Table 4. The leaf chlorophyll content (LCC) retrieval performance by various nonparametric regression methods (NPRMs) in marakele National Park (MNP).

NPRM	MAE (μg/cm^2)	RMSE (μg/cm^2)	RRMSE (%)	R^2
KNNR	4.23	5.23	20.96	0.002
PLSR	4.53	5.55	22.28	0.042
RFR	4.57	5.82	23.36	0.016
KRR	7.88	10.26	41.15	0.001
PCR	189.35	190.80	765.41	0.064

Table 5. The leaf area index (LAI) retrieval performance of partial least squares regression (PLSR) based on optimised training samples from various active learning (AL) algorithms i.e. angle-based diversity (ABD), cluster-based diversity (CDB), euclidean distance-based diversity (EBD), Pool active learning (PAL), random sampling (RS) and residual active learning (RSAL).

AL Algorithm	RMSE (m^2/m^2)	RRMSE (%)	MAE (m^2/m^2)	R^2	Time	Samples	Iterations
ABD	0.80	42.08	0.63	0.09	22.30	100	345
CBD	0.77	40.37	0.59	0.21	5.66	100	244
EBD	0.77	40.25	0.59	0.21	4.48	100	261
PAL	0.77	40.32	0.59	0.21	35.27	100	412
RS	0.78	40.98	0.61	0.17	2.18	100	238
RSAL	**0.76**	**39.87**	**0.59**	**0.21**	**15.09**	**100**	**716**

and RFR yielded promising model prediction accuracies compared to KRR, KNNR and PCR (Table 3).

In contrast, the LCC retrieval performance for most NPRMs (i.e. KNNR, PLSR, RFR and KRR) showed better model prediction accuracies with lower RRMSE's in the range of approximately 20% to 41% (Table 4). Although, the aforementioned NPRMs explained very little of the LCC variability within MNP according to the low R^2 values, the models revealed encouraging prediction accuracies of LCC. The highest LCC estimation accuracies based on the order of RMSE's and RRMSE's were achieved by the KNNR and PLSR models. It was interesting to observe that PLSR emerged as a promising regression model for estimation of LAI and simultaneously, gave the best estimation accuracy of LCC in MNP. However, for both LAI and LCC predictions in MNP, the PCR was the worst performing regression model. In overall, PLSR is selected as the best performing NPRM in this study for integration with AL algorithms.

Integrating AL algorithms with the best performing NPRM

The AL algorithms integrated with PLSR showed notable improvement in the estimation of LAI in MNP corresponding to RRMSE's ranging from 39.87% to 42.08% (Table 5). This improvement is also noted in the explained variability (R^2) which moved from 0% (in Table 3) to approximately 20% (Table 5). This improvement is expected to reflect in the spatial patterns during spatial prediction of LAI (later in the subsequent section).

Furthermore, the considerably lower RMSE's and MAE values of LAI indicate the importance of using smaller optimal training set of informative samples, selected by the AL algorithms to obtain a better model (PLSR) estimation accuracy. In particular, each of the AL algorithms started with an initial training set of 100 random PROSAIL-RTM samples and through numerous iterations shown in Table 5, grew this set by adding informative (also referred to as smart or intelligent) samples until it became optimal with a total of 200 training samples (see Figure 5). In overall, PLSR gave the best LAI prediction accuracy (i.e. lowest RRMSE of 39.87%) in MNP when trained with the RSAL AL algorithm. Additionally, the obtained RMSE of 0.76 m^2.m^{-2} falls within the acceptable range

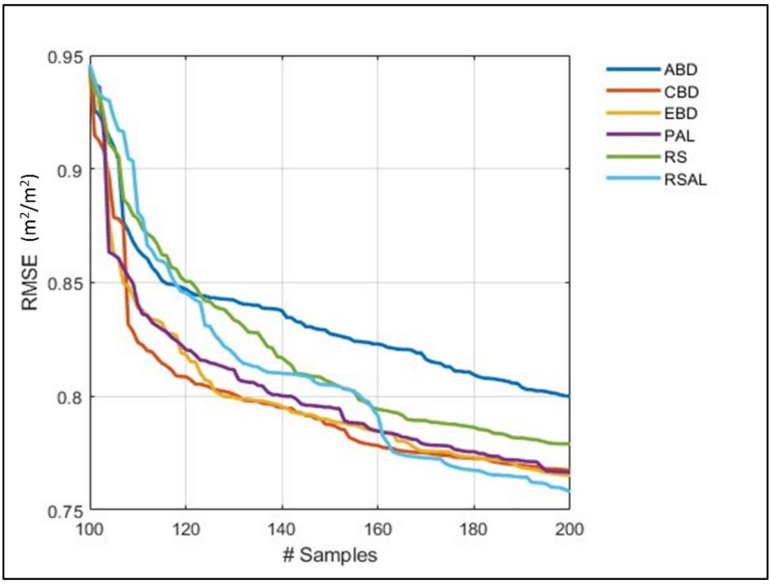

Figure 5. Graphical representation of the root mean squared error (RMSE) for leaf area index (LAI) retrieval by partial least squares regression (PLSR) when trained with only 200 optimised (PROSAIL-RTM) training samples selected by each of the six active learning (AL) algorithms i.e. angle-based diversity (ABD), cluster-based diversity (CDB), euclidean distance-based diversity (EBD), Pool active learning (PAL), random sampling (RS) and residual active learning (RSAL).

Table 6. The leaf chlorophyll content (LCC) retrieval performance of partial least squares regression (PLSR) based on optimised training samples from various active learning (AL) algorithms i.e. angle-based diversity (ABD), cluster-based diversity (CDB), euclidean distance-based diversity (EBD), Pool active learning (PAL), random sampling (RS) and residual active learning (RSAL).

AL Algorithm	RMSE (μg/cm^2)	RRMSE (%)	MAE (μg/cm^2)	R^2	Time	Samples	Iterations
ABD	4.23	16.98	3.29	0.07	143.97	100	1770
CBD	4.17	16.74	3.25	0.09	9.48	100	406
EBD	4.19	16.82	3.28	0.08	36.35	100	1373
PAL	4.15	16.67	3.24	0.10	284.15	100	3768
RS	4.16	16.70	3.24	0.10	6.40	100	799
RSAL	**4.13**	**16.58**	**3.23**	**0.11**	**52.90**	**100**	**2683**

which is representative of a better LAI prediction model i.e. $0.5 \leq \text{RMSE} < 1.0$ (Richter, Atzberger, et al. 2012).

Table 6 show that all AL algorithms integrated with PLSR gave accurate retrievals of LCC in MNP. The accurate retrievals of LCC are evident in the obtained RRMSE's of about 16% (Table 6) which improved from an RRMSE of 22.28% (Table 4). Little improvement is also noted in the explained variability (R^2) which moved from 0% (in Table 4) to approximately 10% (Table 6).

This is also expected to be coupled by little improvement in the spatial patterns during spatial prediction of LCC (later in the subsequent section). In comparison to LAI estimation (Table 5), the LCC estimation by PLSR yielded the best estimation accuracy with RRMSE of 16.58% when trained with the RSAL AL algorithm (Table 6). In particular, the RSAL AL algorithm, similar to other AL algorithms, started with an initial training set of 100 random PROSAIL-RTM samples and through numerous iterations shown in Table 6, grew this set by adding informative samples until it became optimal with a total of 200 training samples (see Figure 6). Furthermore, the results showed that AL algorithms underwent a lot more iterations i.e. with the highest reaching 3768 to find LCC informative

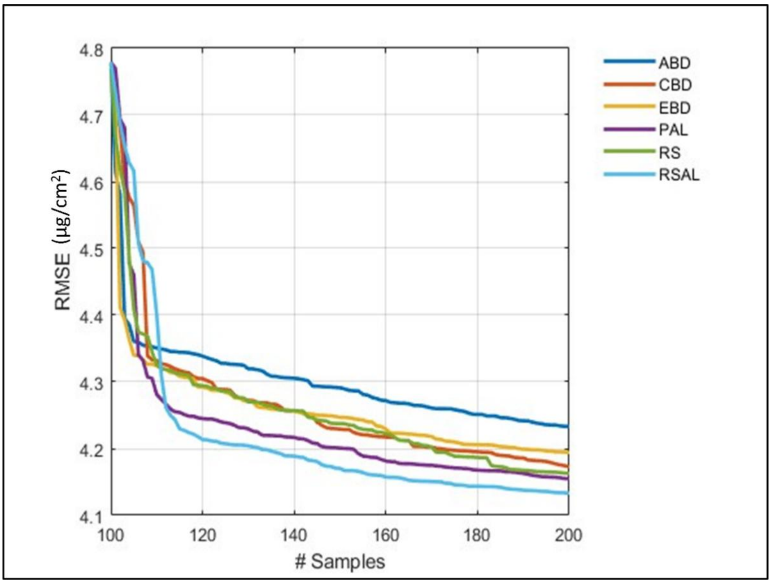

Figure 6. Graphical representation of the root mean squared error (RMSE) for leaf chlorophyll content (LCC) retrieval by partial least squares regression (PLSR) when trained with only 200 optimised (PROSAIL-RTM) training samples selected by each of the six active learning (AL) algorithms i.e. angle-based diversity (ABD), cluster-based diversity (CDB), euclidean distance-based diversity (EBD), Pool active learning (PAL), random sampling (RS) and residual active learning (RSAL).

samples, critical for obtaining the optimised training sets. The integration of PLSR with RSAL yielded the most accurate retrievals of grass LCC and LAI in MNP.

LAI and LCC prediction maps of MNP without the integration of AL algorithms

Figure 7 show the spatial prediction maps of LAI and LCC in MNP generated using the PLSR method during peak productivity. In particular, the LAI map (Figure 7A) showed a general underestimation of LAI across the MNP. This could be seen from the low range of LAI values predicted by PLSR, which appeared to underrepresent the LAI variability. For example, the predicted range values (of approximately $1-1.8\,\mathrm{m}^2/\mathrm{m}^2$) had a notable discrepancy relative to the field data range. In addition, the predicted maximum LAI (Figure 7A) was below the mean LAI from the field measurements (Table 2). Therefore, the LAI map does not suggest realistic patterns of biomass variability in-terms of areas with low, moderate and high biomass.

Besides the biomass, the LAI spatial distribution show patterns across the MNP region which could be influenced by numerous variables such as season, soil type, underlying geology, elevation and vegetation type. For example, in the western part of MNP which is dominated by the sandy bushveld vegetation type (Mucina and Rutherford 2006), clay-rich subsoil (ferric lixisols) and mudstone geology was predicted to have, on average lower LAI values closer to $1\,\mathrm{m}^2/\mathrm{m}^2$. This LAI prediction may suggest the area in the western region has low biomass and could be characterised by large volume grazing, thus subjected to overgrazing. However, the central and eastern parts of MNP that are largely characterised by moderate to high elevation (i.e. ~1024 to 2091 m), mountain bushveld vegetation type (Mucina and Rutherford 2006), sandstone and siltstone geology types and shallow-gravel soil, were modelled to have, on average higher LAI values $> 1.5\,\mathrm{m}^2/\mathrm{m}^2$.

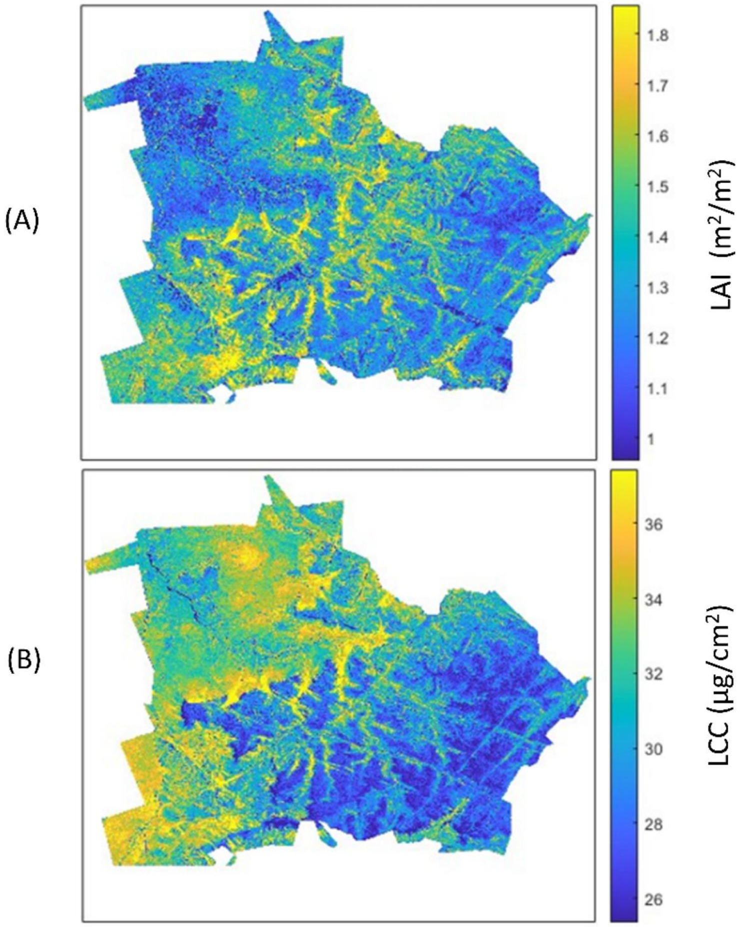

Figure 7. The spatial prediction of leaf area index (LAI): a and leaf chlorophyll content (LCC): B in marakele National Park (MNP) using the partial least squares regression (PLSR) method based on the full PROSAIL RTM simulation database.

Interestingly, the LCC was predicted to be high along the aforementioned areas (Figure 7B). The range of the predicted LCC values came close to the field data range, especially on higher values. The PSLR derived LCC map was able to predict areas within MNP that could possibly have greater LCC than recorded in the field data (Figure 7B). On the other hand, areas predicted to have lower LCC (below 26 μg.cm^{-2}) were mountainous, rocky and characterised by sparse to low cover vegetation. The spatial variability of the predicted LCC showed patterns that are characteristic of the MNP's grass health condition, species diversity and forage quality. Furthermore, both the LAI and LCC spatial prediction maps in Figure 7 could be instrumental in identifying and monitoring potential hotspots where the grazers are most likely to be found. In addition, overgrazed areas coupled with the seasonal and climatic effects on the varying concentrations of vegetation biophysical variables can also be monitored.

LAI and LCC prediction maps of MNP using PLSR integrated with the RSAL algorithm

Figure 8 shows the spatial prediction maps of LAI and LCC in MNP generated using the PLSR method integrated with RSAL algorithm. The effect of this integration could be

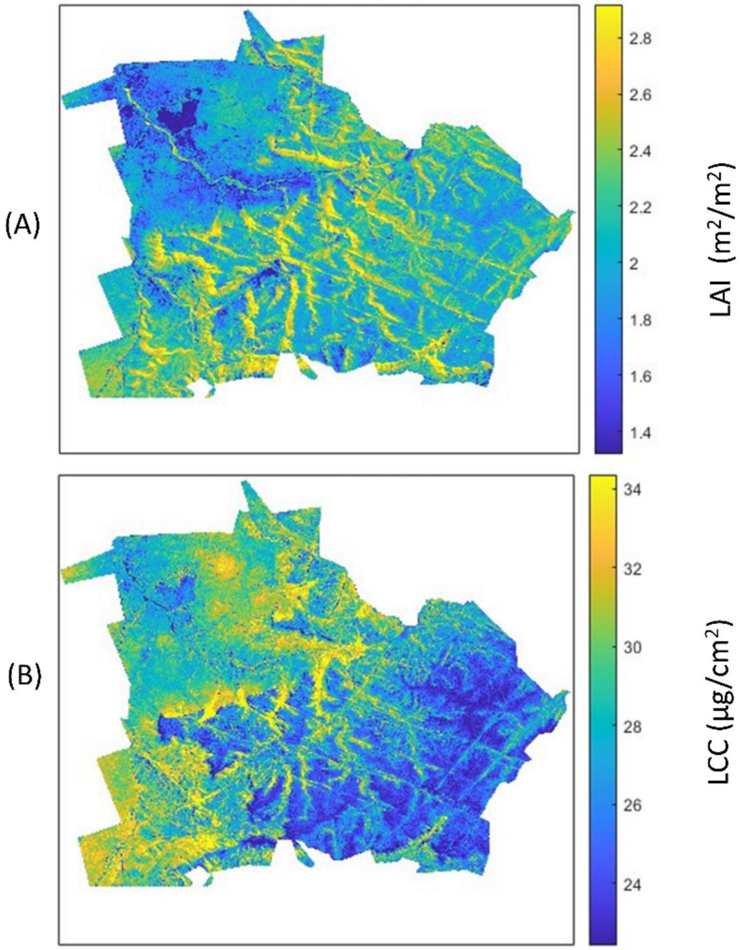

Figure 8. The spatial prediction of leaf area index (LAI): a and leaf chlorophyll content (LCC): B in marakele National Park (MNP) using the partial least squares regression (PLSR) method integrated with the residual active learning (RSAL)-based informative (PROSAIL RTM) samples.

seen in the predicted LAI value range (Figure 8A) with increased spatial variability across MNP. The LAI map showed improved spatial patterns of biomass variability in-terms of areas with low, moderate and high biomass. For example, the areas with predicted LAI values in the range of about $2 - 2.2\,m^2/m^2$ are generally characterised by moderate grass cover and accessible to grazers. The areas that were predicted to have the higher LAI $>$ $2.6\,m^2/m^2$ (Figure 8 A), appeared to have some linear disconnected patterns suggesting that it may be in the water-logged areas and wetlands i.e. valley bottom wetlands. Such areas had high grass biomass but may largely be inaccessible to grazers due to unfriendly terrain, although this observation is yet to be confirmed with animal density data linking it with the map. PLSR gave the most realistic LAI prediction accuracy in MNP when trained with optimal reflectance samples obtained by RSAL AL algorithm (Table 5 and Figure 8A).

Similarly, the prediction of LCC improved when PLSR is integrated RSAL-based informative (PROSAIL RTM) samples (Figure 8B). Although the improvement was minor when comparing the LCC performance retrieval statistics in Table 4 (PLSR without AL) and Table 6 (PLSR with AL); this had a positive result in improving the spatial prediction accuracy of LCC in MNP (Figure 8B).

Discussion

This study successfully tested the integration of nonparametric regression methods (NPRMs) and active learning (AL) algorithms on PROSAIL-RTM simulations using Sentinel-2 data, for improved retrievals of grass leaf area index (LAI) and leaf chlorophyll content (LCC) in Marakele National Park (MNP) during the 2021 peak productivity season. The results showed that the NPRMs particularly partial least squares regression (PLSR), k-nearest neighbours regression (KNNR) and random forest regression (RFR) had the potential to achieve accurate retrievals of grass LCC (Table 4) when trained using many (30,000) samples of RTM reflectance data in a heterogeneous natural ecosystem i.e. MNP characterised by diversity of land cover, varying terrain slopes and multispecies grass canopy. The retrieval accuracy of grass LCC showed improvement reaching a lower relative root mean squared error (RRMSE) of ∼16% when only fewer informative samples of RTM reflectance data was used to train the best performing NPRM i.e. PLSR. Previous studies based on empirical modelling reported RRMSE of about 26.16% of grass CCC in the same region (Tsele et al. 2023). Although, the RTM based results gave better LCC retrievals (as shown in this study) in MNP, more comparative studies between empirical and physically-based approaches are needed to ascertain the consistency of their performance in MNP and in other regions with similar environmental setting.

The LAI retrieval performance by all five NPRMs in MNP was unsatisfactory, corresponding to very high RRMSE values (Table 3). Given the poor performance of the NPRMs, the PLSR in particular, achieved the lowest RRMSE of 57.60% when trained using a large database of RTM simulations. Furthermore, when PLSR was integrated with AL algorithms whereby only few informative samples of synthetic canopy reflectance data were used, the LAI retrieval accuracy showed notable improvement corresponding to the lowest RRMSE of 39.87% (Table 5). Our observation is that grass LAI proved to be a challenging biophysical variable to retrieve in MNP, and the same observation was made when an empirical approach was used in the same region e.g. Tsele et al. (2023) whereby RRMSE's as high as 35.68% were reported. Further investigation is needed to explore varying solution strategies for improving grass LAI retrieval in a heterogenous grassland ecosystem. This could entail collecting more field data samples coupled with additional biophysical and/or biochemical variables, improving RTM parameterisation, and testing more NPRMs and their integration with various AL methods.

For example, previous studies such as Vohland and Jarmer (2008) over the heterogeneous grassland in Rhineland-Palatinate, Germany, have demonstrated that adding field-based structural (such as dry matter content (DMC)) and biochemical (such as leaf water content (LWC)) information during parameterisation of PROSAIL could improve the retrieval accuracy of LAI in the grasslands. Their grass LAI estimation showed improvement in the RMSE values from 0.86 $m^2.m^{-2}$ to 0.74 $m^2.m^{-2}$ corresponding to RRMSE values of 37.94% to 33.48% respectively. Other studies that estimated grass LAI through inversion of PROSAIL reported RMSE's of 0.13 $m^2.m^{-2}$ (Masemola et al. 2016), 0.9 $m^2.m^{-2}$ (Cho et al. 2014), 0.99 $m^2.m^{-2}$ (Darvishzadeh et al. 2008) and 1.09 $m^2.m^{-2}$ (Si et al. 2012). When comparing these errors with the RMSE of grass LAI i.e. 0.76 $m^2.m^{-2}$ (RRMSE of 39.87%) reported in this study, our result (i) compares fairly with the aforementioned similar studies, (ii) revealed marginal differences compared to other reported grass LAI RMSE's and/or RRMSE's and (iii) falls within the acceptable range which is representative of a better LAI prediction model i.e. $0.5\,\text{m}^2.\text{m}^{-2} \leq \text{RMSE} < 1.0\,\text{m}^2.\text{m}^{-2}$ (Richter, Atzberger, et al. 2012).

For example, our result showed improved grass LAI retrieval accuracy of RRMSE of 39.87% across 30 sampled grass species in MNP when compared with RRMSE of 45.55% across 4 sampled grass species in Majella National Park in Italy, reported in Darvishzadeh et al. (2008). In addition, an improvement was observed between our reported RMSE and RRMSE for grass LCC retrieval (i.e. 4.13 μg/cm^2 and 16.58%) in MNP to that reported by Darvishzadeh et al. (2008) of 6.8 μg/cm^2 and 22.61% respectively. Similarly, our findings gave better RRMSE retrievals of grass LAI (39.87%) and LCC (16.58%) compared to those reported by Si et al. (2012) of 51.78% and 46.35% respectively, in the northern part of The Netherlands mainly covered by two grassland types i.e. 70% agricultural grassland (2 species) and 30% semi-natural grassland (5 species). In overall, the hybrid approach of integrating non-parametric PLSR with the Residual active learning (RSAL) active learning (AL) algorithm for the retrieval of grass LAI and LCC in a heterogenous grassland ecosystem gave promising results in MNP. To our knowledge, this study was the first to test such hybrid approach in the grassland ecosystem using Sentinel-2 data. Although, this hybrid approach of integrating parametric and/or non-parametric methods with AL algorithms using RTM data has been widely tested in agricultural environments or crop related studies (e.g. Verrelst et al. 2016; 2020; Berger et al. 2021; Candiani et al. 2022; Pascual-Venteo et al. 2022; Wocher et al. 2022) and successfully demonstrated the potential to improve the retrieval accuracy of biophysical variables such as the LAI and LCC. However, very few studies were found on using this hybrid approach in natural ecosystems.

The active learning (AL) algorithms presents an alternate approach for biophysical parameter estimation when dealing with a large pool of PROSAIL RTM simulations, by ensuring the selection of only the best possible samples from a large pool of RTM simulations for use by the regression model (Pasolli et al. 2012). Further research would be to evaluate the integration of different AL algorithms with other NPRMs in the domain of decision trees, neural networks and kernel-based regression methods. In addition, increasing the sample size of LAI and LCC field measurements could improve the variability of our measurements to be representative across most of the 136 total grass species that exist in MNP. Simultaneously, it may be important to consider measuring additional field-based variables such as carotenoids, LWC and DMC of the grass as well as soil reflectance information. The availability of such measurements would advance the accurate parameterisation of the PROSAIL RTM and constrain the simulations to more realistic reflectance samples. This ultimately, creates an opportunity to further improve the retrieval accuracy of the resulting regression model especially when integrated with AL algorithms. Lastly, the inclusion of uncertainty maps for the spatial predictions is important to consider in future.

Conclusion

Firstly, this paper compared the performance of linear and non-linear nonparametric regression method (NPRMs) trained with large database of simulated canopy reflectance samples for the estimation of leaf area index (LAI) and leaf chlorophyll content (LCC), over a multispecies grass canopy located in a protected mountainous region. Secondly, this paper applied several active learning (AL) sample selection algorithms to: (i) disregard the non-diverse and potential outliers from the large pool of radiative transfer model (RTM)-simulated reflectance samples, and (ii) optimise the simulated training dataset to contain only intelligent or informative samples needed for improving the regression model's retrieval accuracy.

Our findings showed that, before applying AL sample selection techniques, partial least squares regression (PLSR) i.e. linear NPRM followed by random forest regression (RFR) i.e. non-linear NPRM were top performers compared to other regression models in the estimation of grass LAI in Marakele National Park (MNP). Whereas, k-nearest neighbours regression (KNNR) which is a non-linear NPRM, followed by PLSR gave the most accurate retrievals (with a marginal difference) of grass LCC in MNP. Given the consistent performance of PLSR for both LAI and LCC estimations, PLSR was then chosen for integration with AL algorithms to perform hybrid retrieval of the aforementioned biophysical variables. Furthermore, the results of the best performing AL algorithm i.e. residual active learning (RSAL), integrated with PLSR showed the best improvement in the grass LAI retrieval accuracy, corresponding to RMSE of 0.76 $m^2.m^{-2}$, RRMSEs of 39.87% and R^2 of 0.21. On the other hand, the results of RSAL integrated with PLSR revealed the best improvement in grass LCC retrieval accuracy, corresponding to RMSE of 4.13 $\mu g/cm^2$, RRMSEs of 16.58% and R^2 of 0.11. The hybrid models presented in this study gave the most realistic prediction accuracy of LAI and LCC accuracy in MNP when trained with optimal reflectance samples obtained by RSAL AL algorithm. These findings have significant implications for the development of transferable rangeland monitoring systems in protected mountainous regions.

Further investigation is needed to explore varying solution strategies for improving the retrieval accuracy of grass biophysical variables in a heterogenous natural ecosystem. This could entail (i) collecting more field data samples coupled with additional biophysical and/or biochemical variables such as, carotenoids, leaf water content (LWC) and dry matter content (DMC) of the grass as well as soil reflectance information (ii) improving PROSAIL RTM parameterisation by using more actual field-based measurements, (iii) a comprehensive evaluation of linear and non-linear NPRMs and their integration with AL methods.

Acknowledgments

The Sentinel-2 data used in this study were downloaded from the European Space Agency Copernicus Open Access Hub. We sincerely thank the field assistants (namely Phomolo Seriba, Steven Khosa, Brian Mabunda, Madodomzi Mafanya and Mcebisi Qabaqaba) in Marakele National Park for their collaborative effort in collecting grass LAI, FVC and LCC ground measurements. Special thanks to Jochem Verrelst, Juan Pablo Rivera, Jorge Vicent for providing us with the ARTMO license.

Author contributions

Conceptualisation, P.T. and A.R.; methodology, P.T. and A.R.; Formal analysis, P.T; validation, P.T; writing—original draft preparation, P.T.; writing—review and editing, A.R.; project administration, P.T.;

Disclosure statement

The authors declare no conflict of interest.

Funding

This research was funded by the (i) Research development programme of the University of Pretoria, (ii) National Research Foundation (NRF) of South Africa, grant number 118590CSRP23030580913 and (iii) Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL), grant number 118590.

Data availability statement

We understand that the publication of the data is becoming a good practice in research.

References

Ali AM, Darvishzadeh R, Skidmore A, Gara TW, Heurich M. 2021. Machine learning methods' performance in radiative transfer model inversion to retrieve plant traits from Sentinel-2 data of a mixed mountain forest. Int J Digital Earth. 14(1):106–120. doi: 10.1080/17538947.2020.1794064.

Ali AM, Darvishzadeh R, Skidmore A, Gara TW, O'Connor B, Roeoesli C, Heurich M, Paganini M. 2020. Comparing methods for mapping canopy chlorophyll content in a mixed mountain forest using Sentinel-2 data. Int J Appl Earth Observ Geoinform. 87:102037. doi: 10.1016/j.jag.2019.102037.

Atzberger C, Darvishzadeh R, Immitzer M, Schlerf M, Skidmore A, Le Maire G. 2015. Comparative analysis of different retrieval methods for mapping grassland leaf area index using airborne imaging spectroscopy. Int J Appl Earth Observ Geoinform. 43:19–31. doi: 10.1016/j.jag.2015.01.009.

Atzberger C, Guérif M, Baret F, Werner W. 2010. Comparative analysis of three chemometric techniques for the spectroradiometric assessment of canopy chlorophyll content in winter wheat. Comput Electr Agricult. 73(2):165–173. doi: 10.1016/j.compag.2010.05.006.

Bacour C, Baret F, Béal D, Weiss M, Pavageau K. 2006. Neural network estimation of LAI, fAPAR, fCover and LAI × Cab, from top of canopy MERIS reflectance data: principles and validation. Remote Sens Environ. 105(4):313–325. doi: 10.1016/j.rse.2006.07.014.

Baret F, Stéphane J, Guyot G, Leprieur C. 1992. Modeled analysis of the biophysical nature of spectral shifts and comparison with information content of broad bands. Remote Sens Environ. 41(2–3):133–142. doi: 10.1016/0034-4257(92)90073-S.

Baret F, Weiss M, Lacaze R, Camacho F, Makhmara H, Pacholcyzk P, Smets B. 2013. GEOV1: LAI and FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. Part1: principles of development and production. Remote Sens Environ. 137:299–309. doi: 10.1016/j.rse.2012.12.027.

Bei C, Qian-jun Z, Wen-jiang H, Xiao-yu S, Hui-chun YE, Xian-feng Z. 2019. Leaf chlorophyll content retrieval of wheat by simulated RapidEye, Sentinel-2 and EnMAP data. J Integr Agric. 18(6):1230–1245. doi: 10.1016/S2095-3119(18)62093-3.

Berger K, Caicedo JPR, Martino L, Wocher M, Hank T, Verrelst J. 2021. A survey of active learning for quantifying vegetation traits from terrestrial earth observation data. Remote Sens (Basel). 13(2):287. doi: 10.3390/rs13020287.

Binh NA, Hauser LT, Hoa PV, Phuong Thao GT, An NN, Nhut HS, Phuong TA, Verrelst J. 2022. Quantifying mangrove leaf area index from Sentinel-2 imagery using hybrid models and active learning. Int J Remote Sens. 43(15-16):5636–5657. doi: 10.1080/01431161.2021.2024912.

Breiman L. 2001. Random forests. Machine Learning. 45(1):5–32. doi: 10.1023/A:1010933404324.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 2017. Classification and regression trees. New York: Routledge.

Bsaibes A, Courault D, Baret F, Weiss M, Olioso A, Jacob F, Hagolle O, Marloie O, Bertrand N, Desfond V, et al. 2009. Albedo and LAI estimates from FORMOSAT-2 data for crop monitoring. Remote Sens Environ. 113(4):716–729. doi: 10.1016/j.rse.2008.11.014.

Candiani G, Tagliabue G, Panigada C, Verrelst J, Picchi V, Caicedo JPR, Boschetti M. 2022. Evaluation of hybrid models to estimate chlorophyll and nitrogen content of maize crops in the framework of the future CHIME mission. Remote Sens (Basel). 14(8):1792. doi: 10.3390/rs14081792.

Chai T, Draxler RR. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geosci Model Dev. 7(3):1247–1250. doi: 10.5194/gmd-7-1247-2014.

Chen W, Li X, Wang Y, Chen G, Liu S. 2014. Forested landslide detection using LiDAR data and the random forest algorithm: a case study of the Three Gorges, China. Remote Sens Environ. 152:291–301. doi: 10.1016/j.rse.2014.07.004.

Cho MA, Ramoelo A, Math R. 2014. Estimation of leaf area index (LAI) of South Africa from MODIS imager by inversion of PROSAIL radiative transfer model. Paper Presented at the 2014 IEEE Geoscience and Remote Sensing Symposium.

Combal B, Baret F, Weiss M, Trubuil A, Macé D, Pragnère A, Myneni R, Knyazikhin Y, Wang L. 2003. Retrieval of canopy biophysical variables from bidirectional reflectance: using prior information to solve the ill-posed inverse problem. Remote Sens Environ. 84(1):1–15. doi: 10.1016/S0034-4257(02)00035-4.

Cover T, Hart P. 1967. Nearest neighbor pattern classification. IEEE Trans Inform Theory. 13(1):21–27. doi: 10.1109/TIT.1967.1053964.

Crawford MM, Tuia D, Yang HL. 2013. Active learning: any value for classification of remotely sensed data? Proc IEEE. 101(3):593–608. doi: 10.1109/JPROC.2012.2231951.

Darvishzadeh R, Clement A, Andrew S, Martin S. 2011. Mapping grassland leaf area index with airborne hyperspectral imagery: a comparison study of statistical approaches and inversion of radiative transfer models. ISPRS J Photogramm Remote Sens. 66(6):894–906. doi: 10.1016/j.isprsjprs.2011.09.013.

Darvishzadeh R, Skidmore A, Schlerf M, Atzberger C. 2008. Inversion of a radiative transfer model for estimating vegetation LAI and chlorophyll in a heterogeneous grassland. Remote Sens Environ. 112(5): 2592–2604. doi: 10.1016/j.rse.2007.12.003.

Dash J, Curran PJ. 2004. The MERIS terrestrial chlorophyll index.

Demir B, Persello C, Bruzzone L. 2010. Batch-mode active-learning methods for the interactive classification of remote sensing images. IEEE Trans Geosci Remote Sens. 49(3):1014–1031. doi: 10.1109/TGRS.2010.2072929.

Disney M, Muller J-P, Kharbouche S, Kaminski T, Voßbeck M, Lewis P, Pinty B. 2016. A new global fAPAR and LAI dataset derived from optimal albedo estimates: comparison with MODIS products. Remote Sens. 8(4):275. doi: 10.3390/rs8040275.

Douak F, Benoudjit N, Melgani F. 2011. A two-stage regression approach for spectroscopic quantitative analysis. Chemometr Intelligent Lab Syst. 109(1):34–41. doi: 10.1016/j.chemolab.2011.07.007.

Douak F, Melgani F, Benoudjit N. 2013. Kernel ridge regression with active learning for wind speed prediction. Appl Energy. 103:328–340. doi: 10.1016/j.apenergy.2012.09.055.

Frampton WJ, Dash J, Watmough G, Milton EJ. 2013. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. ISPRS J Photogrammetr Remote Sens. 82:83–92. doi: 10.1016/j.isprsjprs.2013.04.007.

García-Haro FJ, Campos-Taberner M, Muñoz-Marí J, Laparra V, Camacho F, Sánchez-Zapero J, Camps-Valls G. 2018. Derivation of global vegetation biophysical parameters from EUMETSAT Polar System. ISPRS J Photogrammetr Remote Sens. 139:57–74. doi: 10.1016/j.isprsjprs.2018.03.005.

Geladi P, Kowalski BR. 1986. Partial least-squares regression: a tutorial. Anal Chim Acta. 185:1–17. doi: 10.1016/0003-2670(86)80028-9.

Gitelson AA, Viña A, Ciganda V, Rundquist DC, Arkebauer TJ. 2005. Remote estimation of canopy chlorophyll content in crops. Geophys Res Lett. 32(8):L08403. doi: 10.1029/2005GL022688.

Guerini Filho M, Kuplich TM, De Quadros FLF. 2020. Estimating natural grassland biomass by vegetation indices using Sentinel 2 remote sensing data. Int J Remote Sens. 41(8):2861–2876. doi: 10.1080/01431161.2019.1697004.

Hardin PJ. 1994. Parametric and nearest-neighbor methods for hybrid classification: a comparison of pixel assignment accuracy. Photogrammetr Eng Remote Sens. 60(12):1439–1447.

Hastie T, Tibshirani R, Friedman JH, Friedman JH. 2009. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. Stanford, CA: Stanford University.

Heinemann AB, Van Oort PAJ, Fernandes DS, Maia AdHN 2012. Sensitivity of APSIM/ORYZA model due to estimation errors in solar radiation. Bragantia. 71(4):572–582. doi: 10.1590/S0006-87052012000400016.

Jacquemoud S, Frédéric B. 1990. PROSPECT: A model of leaf optical properties spectra. Remote Sens Environ. 34(2):75–91. doi: 10.1016/0034-4257(90)90100-Z.

Jacquemoud S, Verhoef W, Baret F, Bacour C, Zarco-Tejada PJ, Asner GP, François C, Ustin SL. 2009. PROSPECT + SAIL models: a review of use for vegetation characterization. Remote Sens Environ. 113: s56–S66. doi: 10.1016/j.rse.2008.01.026.

Jain AK, Dubes RC. 1988. Algorithms for clustering data: Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Jamieson PD, Porter JR, Wilson DR. 1991. A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand. Field Crops Res. 27(4):337–350. doi: 10.1016/0378-4290(91)90040-3.

Jia K, Yang L, Liang S, Xiao Z, Zhao X, Yao Y, Zhang X, Jiang B, Liu D. 2019. Long-term Global Land Surface Satellite (GLASS) fractional vegetation cover product derived from MODIS and AVHRR Data. IEEE J Sel Top Appl Earth Observations Remote Sens. 12(2):508–518. doi: 10.1109/JSTARS.2018.2854293.

Jonckheere I, Fleck S, Nackaerts K, Muys B, Coppin P, Weiss M, Baret F. 2004. Review of methods for in situ leaf area index determination: part I. Theories, sensors and hemispherical photography. Agricult Forest Meteorol. 121(1-2):19–35. doi: 10.1016/j.agrformet.2003.08.027.

Kganyago M, Mhangara P, Adjorlolo C. 2021. Estimating crop biophysical parameters using machine learning algorithms and Sentinel-2 imagery. Remote Sens. 13(21):4314. doi: 10.3390/rs13214314.

Liang L, Qin Z, Zhao S, Di L, Zhang C, Deng M, Lin H, Zhang L, Wang L, Liu Z. 2016. Estimating crop chlorophyll content with hyperspectral vegetation indices and the hybrid inversion method. Int J Remote Sens. 37(13):2923–2949. doi: 10.1080/01431161.2016.1186850.

Louis J, Debaecker V, Pflug B, Main-Knorn M, Bieniarz J, Mueller-Wilm U, Cadau E, Gascon F. 2016. Sentinel-2 Sen2Cor: L2A processor for users. Paper presented at the. Proceedings Living Planet Symposium 2016.

Lv T, Zhou X, Tao Z, Sun X, Wang J, Li R, Xie F. 2021. Remote sensing-guided spatial sampling strategy over heterogeneous surface ground for validation of vegetation indices products with medium and high spatial resolution. Remote Sens. 13(14):2674. doi: 10.3390/rs13142674.

MacKay DJC. 1992. Information-based objective functions for active data selection. Neural Computat. 4(4):590–604. doi: 10.1162/neco.1992.4.4.590.

Markwell J, Osterman JC, Mitchell JL. 1995. Calibration of the Minolta SPAD-502 leaf chlorophyll meter. Photosynth Res. 46(3):467–472. doi: 10.1007/BF00032301.

Masemola C, Cho MA, Ramoelo A. 2016. Comparison of Landsat 8 OLI and Landsat 7 ETM + for estimating grassland LAI using model inversion and spectral indices: case study of Mpumalanga, South Africa. Int J Remote Sens. 37(18):4401–4419. doi: 10.1080/01431161.2016.1212421.

Mucina L, Rutherford MC. 2006. The vegetation of South Africa. Lesotho and Swaziland: South African National Biodiversity Institute.

Mutanga O, Adam E, Cho MA. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int J Appl Earth Observ Geoinform. 18:399–406. doi: 10.1016/j.jag.2012.03.012.

Myneni RB, Hoffman S, Knyazikhin Y, Privette JL, Glassy J, Tian Y, Wang Y, Song X, Zhang Y, Smith GR, et al. 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. Remote Sens Environ. 83(1-2):214–231. doi: 10.1016/S0034-4257(02)00074-3.

Okujeni A, Van der Linden S, Jakimow B, Rabe A, Verrelst J, Hostert P. 2014. A comparison of advanced regression algorithms for quantifying urban land cover. Remote Sens. 6(7):6324–6346. doi: 10.3390/rs6076324.

Pascual-Venteo AB, Portalés E, Berger K, Tagliabue G, Garcia JL, Pérez-Suay A, Rivera-Caicedo JP, Verrelst J. 2022. Prototyping crop traits retrieval models for CHIME: dimensionality reduction strategies applied to PRISMA data. Remote Sens (Basel). 14(10):2448. doi: 10.3390/rs14102448.

Pasolli E, Melgani F, Alajlan N, Bazi Y. 2012. Active learning methods for biophysical parameter estimation. IEEE Trans Geosci Remote Sens. 50(10):4071–4084. doi: 10.1109/TGRS.2012.2187906.

Patra S, Bruzzone L. 2012. A cluster-assumption based batch mode active learning technique. Pattern Recogn Lett. 33(9):1042–1048. doi: 10.1016/j.patrec.2012.01.015.

Ramoelo A, Cho M. 2018. Explaining leaf nitrogen distribution in a semi-arid environment predicted on Sentinel-2 imagery using a field spectroscopy derived model. Remote Sens. 10(2):269. doi: 10.3390/rs10020269.

Richter K, Atzberger C, Hank TB, Mauser W. 2012. Derivation of biophysical variables from Earth observation data: validation and statistical measures. J Appl Remote Sens. 6(1):063557–1. doi: 10.1117/1.JRS.6.063557.

Rodriguez G, Francisco V, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogrammetr Remote Sens. 67:93–104. doi: 10.1016/j.isprsjprs.2011.11.002.

Saunders C, Gammerman A, Vovk V. 1998. Ridge regression learning algorithm in dual variables.

Schloderer G, Bingham M, Awange JL, Fleming KM. 2011. Application of GNSS-RTK derived topographical maps for rapid environmental monitoring: a case study of Jack Finnery Lake (Perth, Australia). Environ Monit Assess. 180(1-4):147–161. doi: 10.1007/s10661-010-1778-8.

Si Y, Schlerf M, Zurita-Milla R, Skidmore A, Wang T. 2012. Mapping spatio-temporal variation of grassland quantity and quality using MERIS data and the PROSAIL model. Remote Sens Environ. 121:415–425. doi: 10.1016/j.rse.2012.02.011.

Skidmore AK, Coops NC, Neinavaz E, Ali A, Schaepman ME, Paganini M, Kissling WD, Vihervaara P, Darvishzadeh R, Feilhauer H, et al. 2021. Priority list of biodiversity metrics to observe from space. Nat Ecol Evol. 5(7):896–906. doi: 10.1038/s41559-021-01451-x.

Snee RD. 1977. Validation of regression models: methods and examples. Technometrics. 19(4):415–428. doi: 10.1080/00401706.1977.10489581.

Suykens JAK, Vandewalle J. 1999. Least squares support vector machine classifiers. Neural Process Lett. 9(3):293–300. doi: 10.1023/A:1018628609742.

Tsele P, Ramoelo A, Qabaqaba M. 2023. Development of the grass LAI and CCC remote sensing-based models and their transferability using sentinel-2 data in heterogeneous grasslands. Int J Remote Sens. 44(8):2643–2667. doi: 10.1080/01431161.2023.2205982.

Tsele P, Ramoelo A, Qabaqaba M, Mafanya M, Chirima G. 2022. Validation of LAI, Chlorophyll and FVC biophysical estimates from Sentinel-2 Level 2 Prototype Processor over a heterogeneous savanna and grassland environment in South Africa. Geocarto Int. 37(26):14355–14378. doi: 10.1080/10106049.2022.2087756.

Van Staden PJ, Bredenkamp GJ. 2005. Major plant communities of the Marakele National Park. Koedoe. 48(2):59–70. doi: 10.4102/koedoe.v48i2.101.

Verhoef W. 1984. Light scattering by leaf layers with application to canopy reflectance modeling: the SAIL model. Remote Sens Environ. 16(2):125–141. doi: 10.1016/0034-4257(84)90057-9.

Verrelst J, Berger K, Rivera-Caicedo JP. 2020. Intelligent sampling for vegetation nitrogen mapping based on hybrid machine learning algorithms. IEEE Geosci Remote Sens Lett. 18(12):2038–2042. doi: 10.1109/lgrs.2020.3014676.

Verrelst J, Camps-Valls G, Muñoz-Marí J, Rivera JP, Veroustraete F, Clevers JG, Moreno J. 2015. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties–a review. ISPRS J Photogrammetr Remote Sens. 108:273–290. doi: 10.1016/j.isprsjprs.2015.05.005.

Verrelst J, Dethier S, Rivera JP, Munoz-Mari J, Camps-Valls G, Moreno J. 2016. Active learning methods for efficient hybrid biophysical variable retrieval. IEEE Geosci Remote Sensing Lett. 13(7):1012–1016. doi: 10.1109/LGRS.2016.2560799.

Verrelst J, Malenovský Z, Van der Tol C, Camps-Valls G, Gastellu-Etchegorry J-P, Lewis P, North P, Moreno J. 2019. Quantifying vegetation biophysical variables from imaging spectroscopy data: a review on retrieval methods. Surv Geophys. 40(3):589–629. doi: 10.1007/s10712-018-9478-y.

Verrelst J, Rivera JP, Veroustraete F, Muñoz-Marí J, Clevers JG, Camps-Valls G, Moreno J. 2015. Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods–a comparison. ISPRS J Photogrammetr Remote Sens. 108:260–272. doi: 10.1016/j.isprsjprs.2015.04.013.

Vohland M, Jarmer T. 2008. Estimating structural and biochemical parameters for grassland from spectroradiometer data by radiative transfer modelling (PROSPECT + SAIL.). Int J Remote Sens. 29(1): 191–209. doi: 10.1080/01431160701268947.

Weiss M, Baret F. 2020. S2ToolBox Level 2 products: LAI, FAPAR, FCOVER.

Wocher M, Berger K, Verrelst J, Hank T. 2022. Retrieval of carbon content and biomass from hyperspectral imagery over cultivated areas. ISPRS J Photogramm Remote Sens. 193:104–114. doi: 10.1016/j.isprsjprs.2022.09.003.

Wold S, Esbensen K, Geladi P. 1987. Principal component analysis. Chemometr Intelligent Lab Syst. 2(1-3):37–52. doi: 10.1016/0169-7439(87)80084-9.