Integrating active learning and machine learning regression methods for hybrid retrieval of grass biophysical variables in a protected mountainous region using Sentinel-2 data

Philemon Tsele and Abel Ramoelo

Department of Geography, Geoinformatics and Meteorology
University of Pretoria
Pretoria 0028, South Africa

National Space Conference: 30 August - 01 September 2023
CSIR, International Convention Centre, Tshwane

01 September 2023

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Introduction

- Biodiversity monitoring is a key component of protected area management and planning.

- Estimation of vegetation biophysical variables is important for understanding vegetation health condition, structure, growth status and gross primary productivity.

    - Leaf area index (LAI): defined as half the total area of green elements of the canopy per unit horizontal ground area
    - Leaf chlorophyll content (LCC): refers to the overall amount of chlorophyll *a* and *b* pigments in a leaf
    - Fractional vegetation cover (FVC): corresponds to the fraction of combined photosynthetic and non-photosynthetic vegetation separated from the exposed soil background within the total study area in the nadir direction
    - Fraction of Absorbed Photosynthetically Active Radiation (FAPAR): quantifies the fraction of the solar radiation absorbed by live leaves for the photosynthesis activity

Essential biodiversity variables

- Ecosystem productivity

- Facilitate effective monitoring and management of natural vegetation at different spatial scales

- Natural heterogeneous canopies like the grasslands of South Africa, are characterized by native grasses of different mixture of species..

Rangeland

- it is critical to (i) assess areas where there is a change in response to climate and/or anthropogenic effects, (ii) quantify the amount of aboveground biomass and vegetation cover, and (iii) monitor the functional status and diversity of the rangeland vegetation communities in-order to enhance ecosystem productivity and stability, guided by effective resource management strategies and policies.

- Some models (empirical or physically based) have made it to the status of operational processing chain

    - Sentinel 2 level 2 prototype processor (SL2P)(Weiss M, Baret F, 2020)

GEOCARTO INTERNATIONAL
2022, VOL. 37, NO. 26, 14355–14378
https://doi.org/10.1080/10106049.2022.2087756

Taylor & Francis
Taylor & Francis Group

Check for updates

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

Philemon Tsele[a], Abel Ramoelo[b], Mcebisi Qabaqaba[b], Madodomzi Mafanya[a,c] and George Chirima[a,d]

[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; [c]Department of Geography, University of South Africa, Pretoria, South Africa; [d]Geoinformation Science Division, Agricultural Research Council Institute for Soil, Climate and Water, Pretoria, South Africa

Problem Statement

- There is a need to develop locally-parameterized & transferrable models
- Biodiversity monitoring tools
- The current hybrid schemes for retrieval of land products such as LAI, LCC, FVC and FPAR yield inadequate retrieval accuracies especially in heterogenous environments characterized by diversity of land cover, species diversity and varying terrain slopes (Tsele, Ramoelo et al. 2022)
- For example, hybrid retrieval schemes such as the Sentinel-2 Level-2 Prototype Processor ((Weiss, Baret et al. 2020) reportedly gave inadequate retrievals of LAI, LCC, CCC and FVC over a heterogeneous grassland environment in South Africa (Tsele, Ramoelo et al. 2023)
- Generally, hybrid schemes rely on retrieval methods, trained with large amount of simulated data.
- There is a need to select only the best possible samples from a large pool of simulations for use by the retrieval method (Berger, Rivera Caicedo et al. 2021)

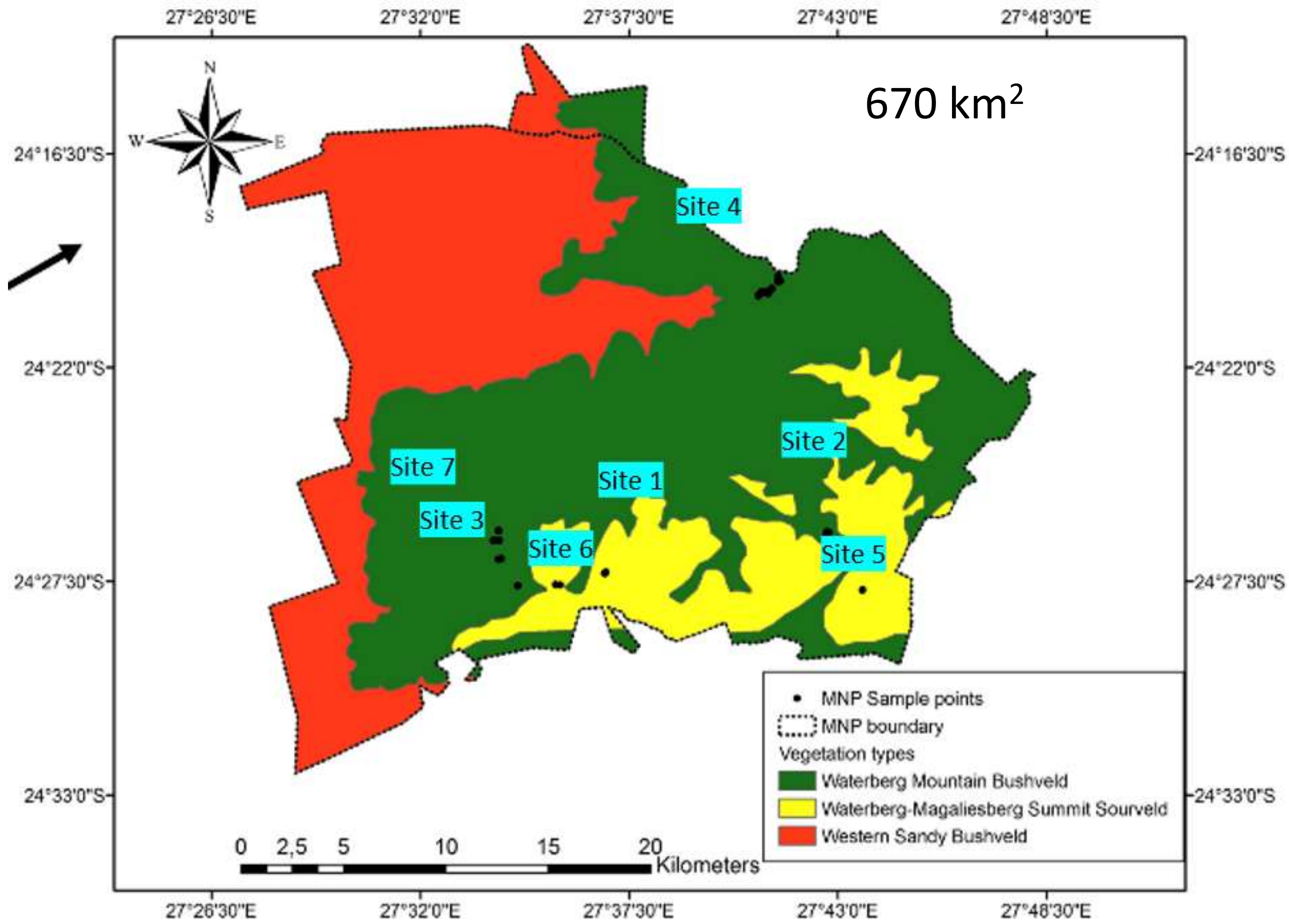
Aim

- To compare various non-parametric regression algorithms (NPRAs) and their integration with active learning (AL) methods, for the improved estimation of leaf area index (LAI) and leaf chlorophyll content (LCC) over a multispecies grass canopy in Marakele National Park.

Study Area

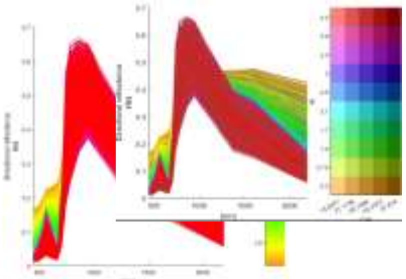
- Encompasses the entire Marakele National Park (MNP)

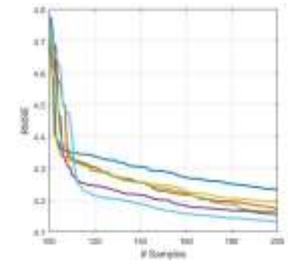
South Africa



- Varying elevation (1307.59m - 1893.29m)
- Slopes 0.34° - 14.12°
- 136 total grass species
- Peak wet season of 2021

Methodology Overview:
Hybrid retrieval workflow including AL sample reduction

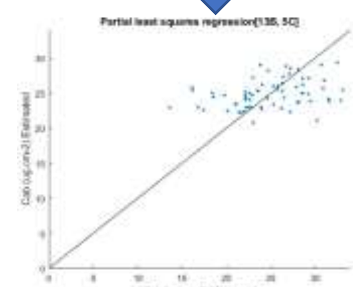


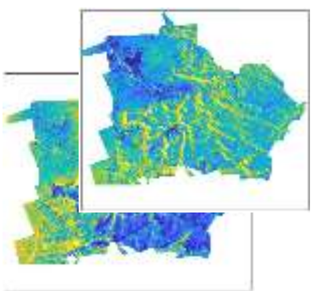


PROSPECT-5 + SAIL → **AL sample selection** → **6 NPRAs** → **Optimized NPRA**

Simulation of training database e.g. 10,000 samples

Retrieval





Model validation Using field data → **Sentinel-2 input scene** → **Output LAI & LCC estimates**

Evaluation & imaging

Mapping

PROSAIL model parameterization

PROSAIL is one of the vegetation Radiative Transfer Models (RTMs) that use physical laws to accurately describe the spectral variation of canopy reflectance ((Jacquemoud, Verhoef et al. 2009)

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)	Unitless	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	Unitless	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	Unitless	1	Fixed	Masemola et al. (2016)
Solar zenith angle	[◦]	40.71	Fixed	Sentinel-2 image Metadata
View zenith angle	[◦]	42.02	Fixed	Sentinel-2 image Metadata

Active learning techniques

- Active learning (AL) techniques use selection criterion algorithms to select informative samples (MacKay 1992)

Large simulated training database e.g. 30,000 samples → **Start**: subset of labelled training samples → Apply selection criteria → Select samples from Larger unlabeled data pool → Selected samples are **labelled** And added to the **training set (Opt)**

- Diversity criteria algorithms:
  - Angle Based Diversity (ABD) (Demir, Persello et al. 2010)
  - Clustering-Based Diversity (CBD) (Patra and Bruzzone 2012)
  - Euclidean Diversity (EBD) (Douak, Melgani et al. 2013)
  - Random Sampling (RS)
- Uncertainty criteria algorithms:
  - Pool Active Learning (PAL) (Douak, Melgani et al. 2013)
  - Residual Active Learning (RSAL) (Douak, Benoudjit et al. 2011)

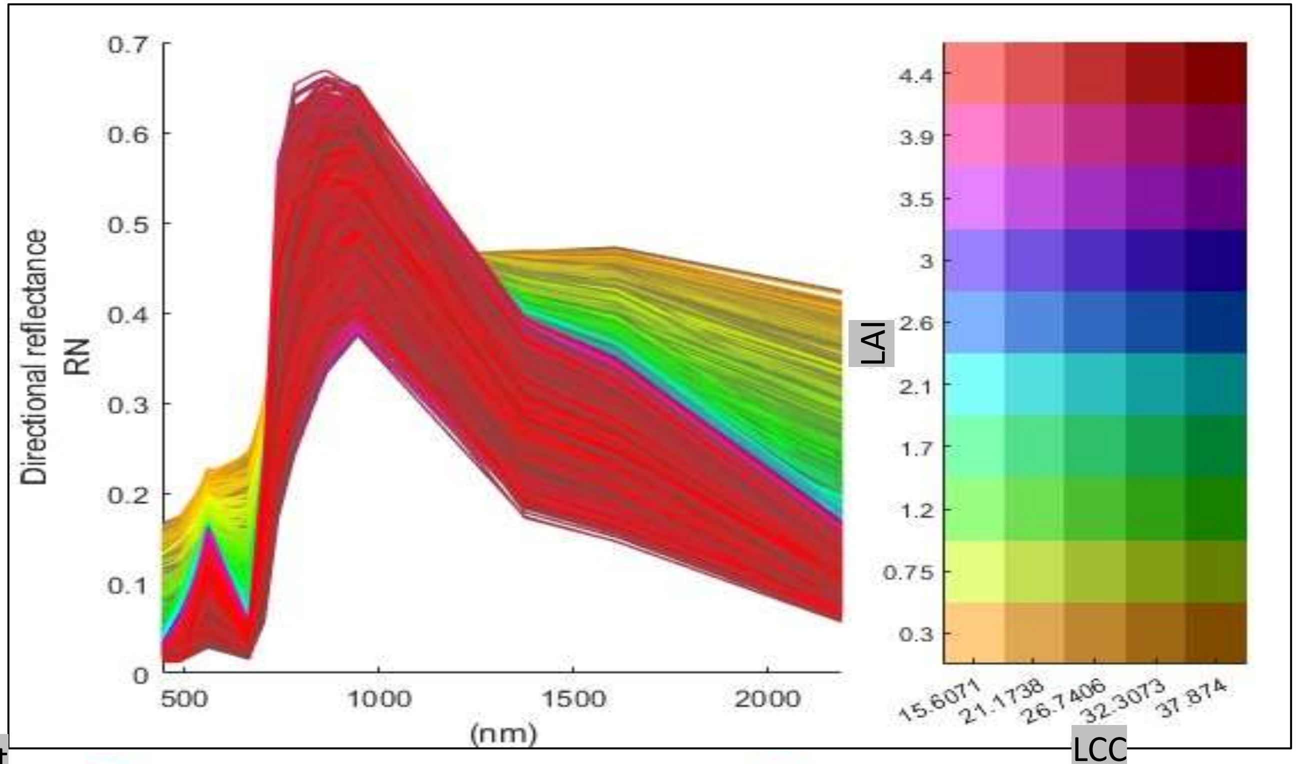
Non-Parametric regression algorithms

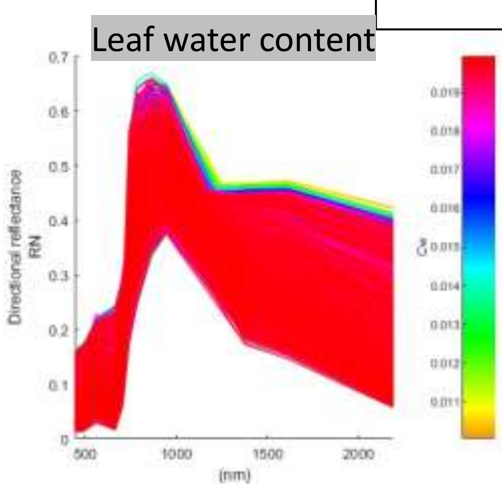
- We evaluated 6 non-parametric regression algorithms (NPRAs), widely used in the literature for estimating vegetation biophysical variables (Verrelst, Camps-Valls et al. 2015)
    - Linear non-parametric regression algorithms
        - Partial least squares regression (PLSR)
        - Principal components regression (PCR)
    - Non-linear non-parametric regression algorithms
        - Gaussian processes regression (GPR)
        - Kernel ridge regression (KRR)
        - Random forest regression (RFR)
        - K-nearest neighbors regression (K-NNR)
- Data driven; Define regression function, Optimize regression model through learning

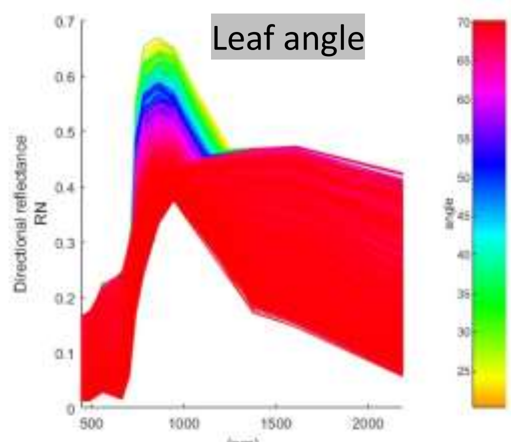
Results

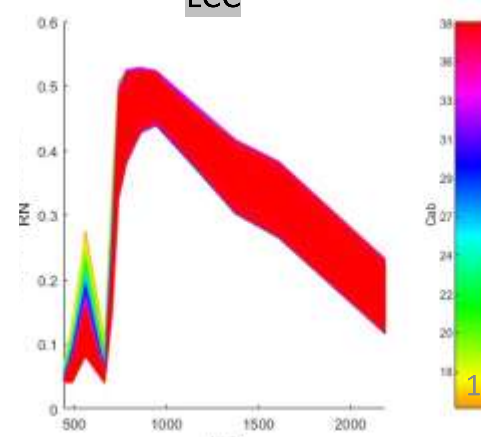
- PROSAIL (30,000 Simulations)

Spectral variation of canopy reflectance









Results…

- NPRAs retrieval performance (without AL)

LAI

NPRAs	MAE (m^2/m^2)	RMSE (m^2/m^2)	RRMSE (%)	NRMSE (%)	R	R^2
PLSR	**0.86**	**1.10**	**57.60**	**24.17**	**-0.06**	**0.00**
RFR	**1.05**	**1.21**	**63.60**	**26.69**	**0.29**	**0.08**
KRR	1.30	1.81	95.23	39.96	0.04	0.00
K-NNR	1.80	1.93	101.52	42.60	0.33	0.11
PCR	5.65	6.78	356.90	149.77	0.02	0.00
GPR	1296.65	1416.25	9999	9999	0.43	0.18

Most accurate retrievals were achieved by the following top 2 methods: partial least-squares (**PLSR**), random forest (**RFR**)

LCC

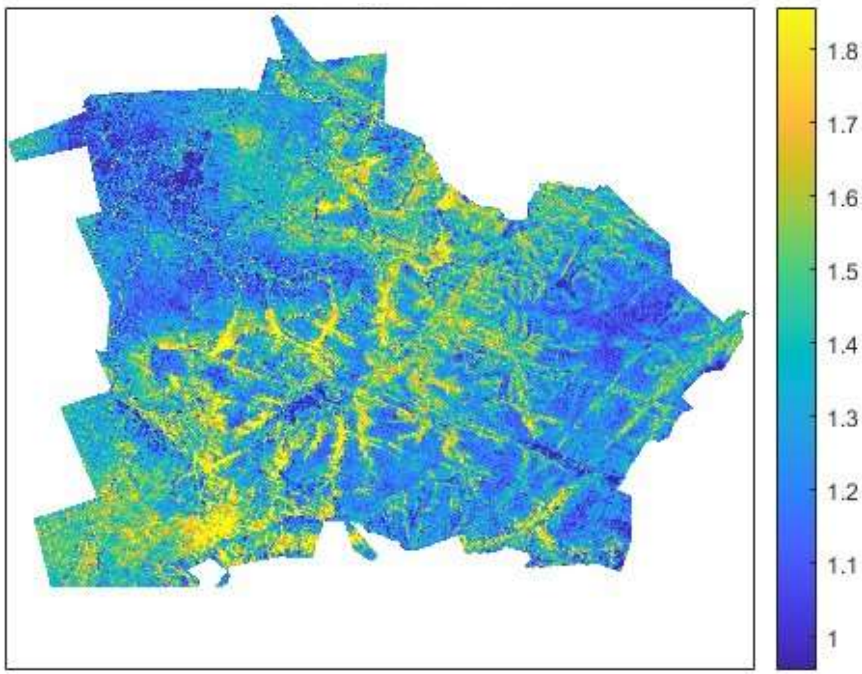
NPRAs	MAE (µg/cm²)	RMSE (µg/cm²)	RRMSE (%)	NRMSE (%)	R	R^2
K-NNR	**4.23**	**5.23**	**20.96**	**26.80**	**0.04**	**0.002**
PLSR	**4.53**	**5.55**	**22.28**	**28.48**	**0.20**	**0.042**
RFR	4.57	5.82	23.36	29.86	-0.12	0.016
KRR	7.88	10.26	41.15	52.61	-0.04	0.001
PCR	189.35	190.80	765.41	978.46	0.25	0.064
GPR	9999	9999	9999	9999	0.18	0.03

Most accurate retrievals were achieved by the following top 2 methods: **K-NNR** and **PLSR**

Results…

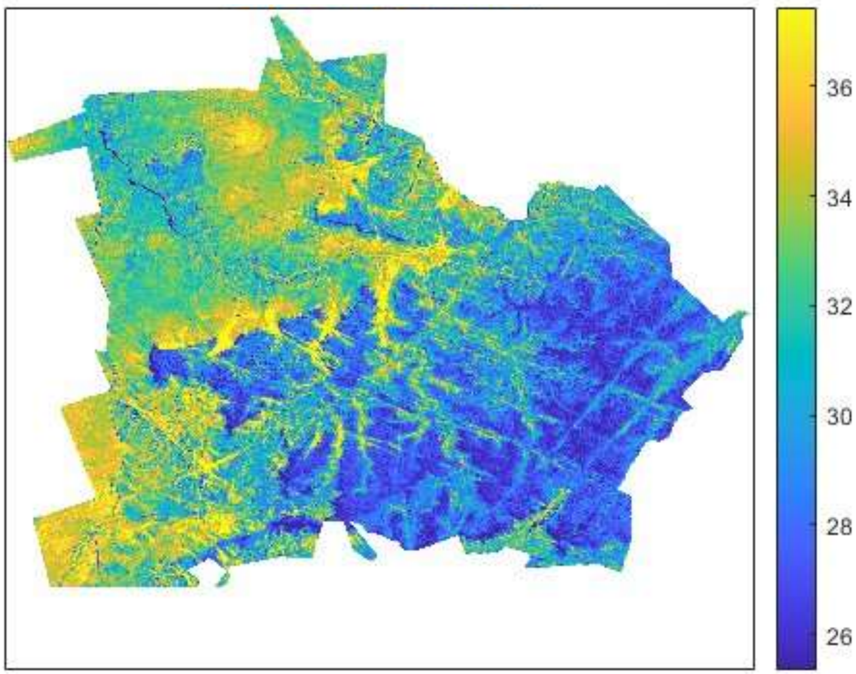
- LAI and LCC estimates (without AL) during peak productivity
- PLSR was chosen for estimation, as one of the top performing retrieval NPRAs

Map of LAI estimated



- Underestimation
- Unrealistic patterns of LAI
- Low biomass

Map of LCC estimated



- Reasonable estimation
- Realistic patterns of LCC
- Forage quality; species diversity

Results…

- NPRAs retrieval performance (with AL)
- AL methods integrated with PLSR showed improvement in both LAI and LCC retrievals

LAI [AL + PLSR]

Algorithm	RMSE (m^2/m^2)	RRMSE (%)	MAE (m^2/m^2)	R	R^2	NRMSE (%)
ABD	0.80	42.08	0.63	0.30	0.09	17.66
CBD	0.77	40.37	0.59	0.46	0.21	16.94
EBD	0.77	40.25	0.59	0.46	0.21	16.89
PAL	0.77	40.32	0.59	0.46	0.21	16.92
RS	0.78	40.98	0.61	0.41	0.17	17.20
RSAL	**0.76**	**39.87**	**0.59**	**0.46**	**0.21**	**16.73**

When PLSR is integrated with Residual active learning (RSAL) method – gave the best retrievals

LCC [AL + PLSR]

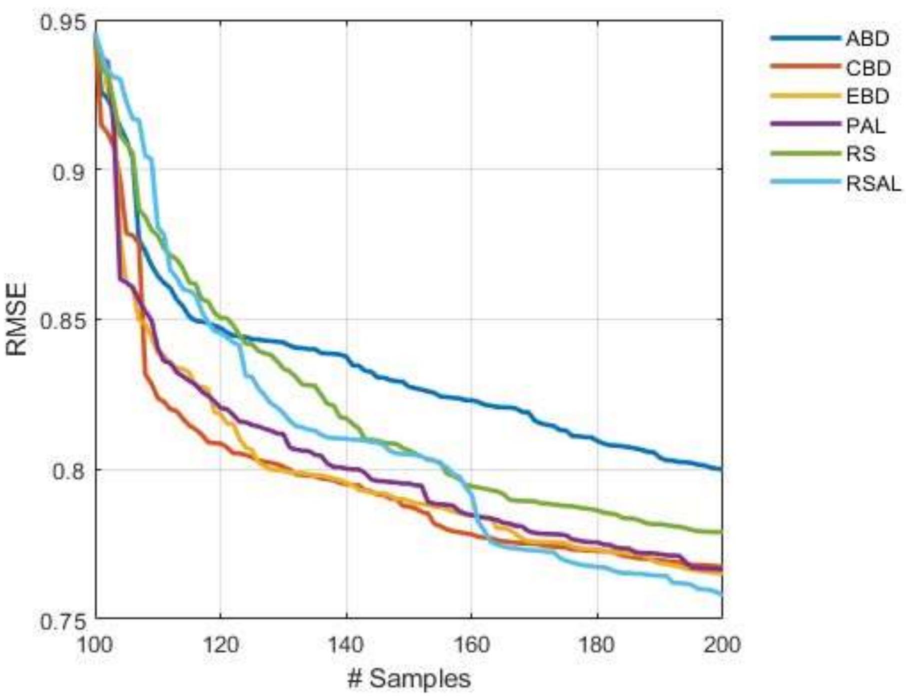
Algorithm	RMSE ($\mu g/cm^2$)	RRMSE (%)	MAE ($\mu g/cm^2$)	R	R^2	NRMSE
ABD	4.23	16.98	3.29	0.27	0.07	21.71
CBD	4.17	16.74	3.25	0.30	0.09	21.40
EBD	4.19	16.82	3.28	0.29	0.08	21.51
PAL	4.15	16.67	3.24	0.31	0.10	21.31
RS	4.16	16.70	3.24	0.31	0.10	21.35
RSAL	**4.13**	**16.58**	**3.23**	**0.33**	**0.11**	**21.19**

• AL performance:

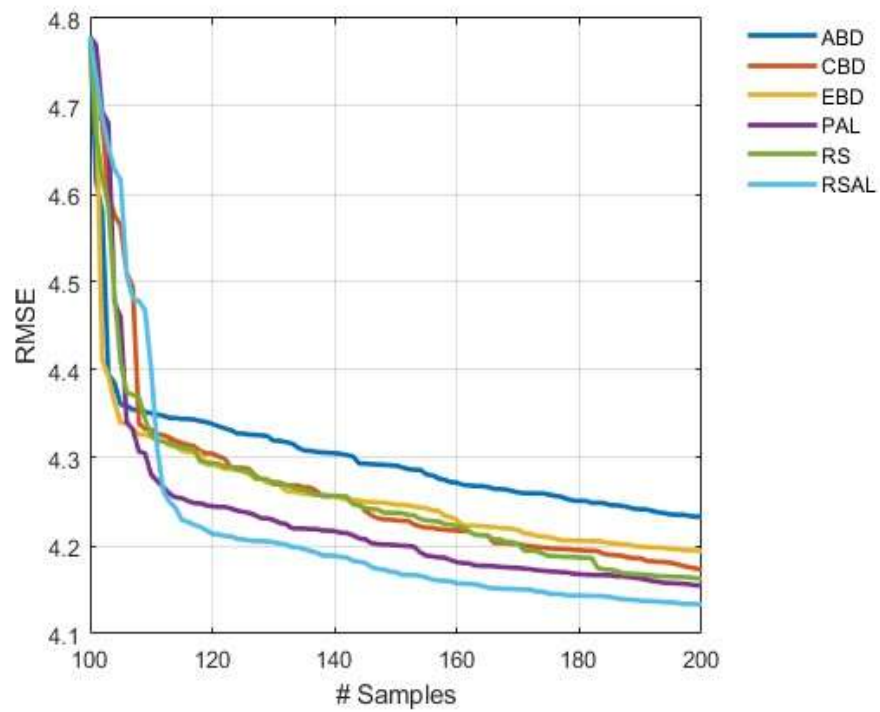
- **Angle Based Diversity (ABD)**
- **Clustering-Based Diversity (CBD)**
- **Euclidean Diversity (EBD)**
- **Random Sampling (RS)**
- **Pool Active Learning (PAL)**
- **Residual Active Learning (RSAL)**

Start: subset of labelled training samples → Apply selection criteria → Select samples from Larger unlabeled data pool → Selected samples are **labelled** And added to the **training set (Opt)**

LAI



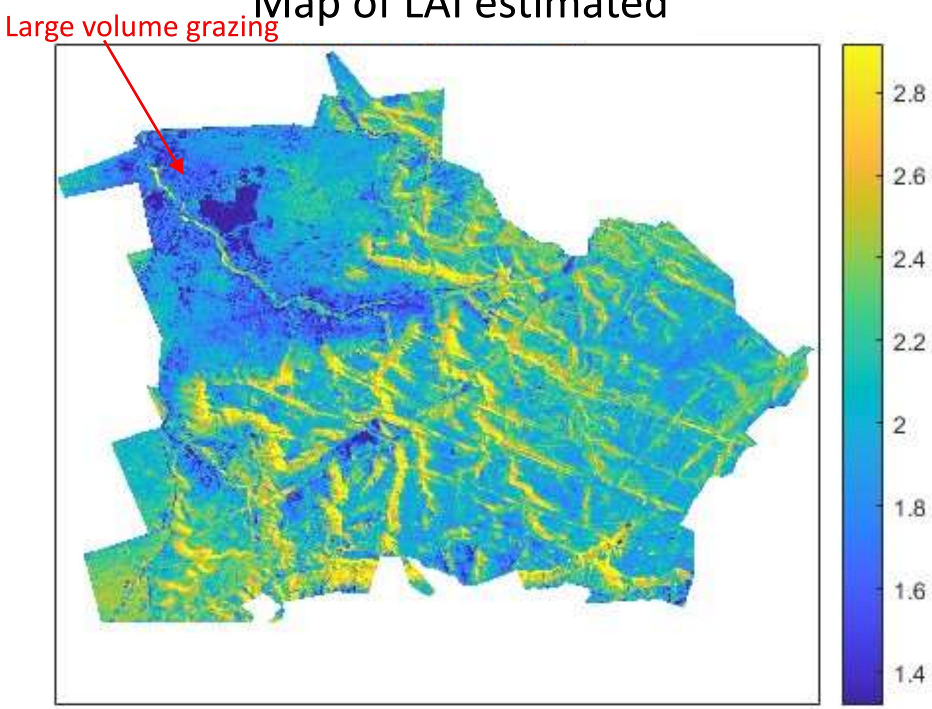
LCC



Results…

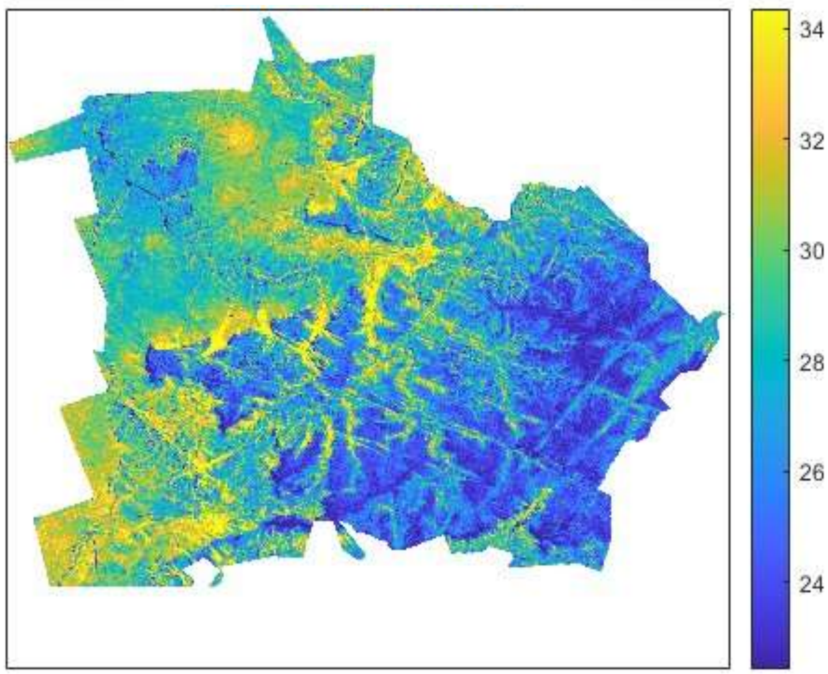
- LAI and LCC estimates (with AL) during peak productivity
- RSAL + PLSR was chosen for estimation
- The spatial prediction maps of LAI and LCC appeared more realistic and the range values came close to the field data range

Map of LAI estimated



- Improved estimation
- More realistic patterns of LAI
- Moderately-high to low biomass

Map of LCC estimated



- Reasonable estimation
- Realistic patterns of LCC
- Forage quality; species diversity

Concluding remarks

- The results of the AL methods integrated with PLSR showed improvement in both LCC and LAI retrievals, corresponding to lower RRMSEs of 16.58% and 39.87%.

- The spatial prediction maps of LAI and LCC appeared more realistic and the range values came close to the field data range

- These findings highlight that RTMs may require local parameterization in order to simulate multispecies canopies accurately, especially in heterogenous environments

- These findings have significant implications for the development of transferable rangeland monitoring systems in protected mountainous regions

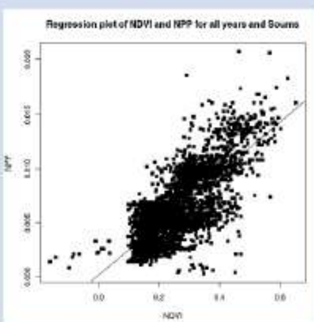
Thank you.

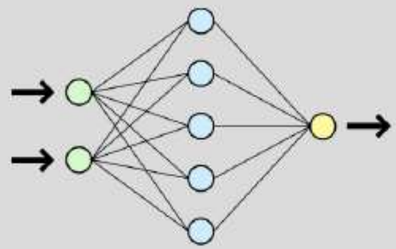
Philemon.tsele@up.ac.za

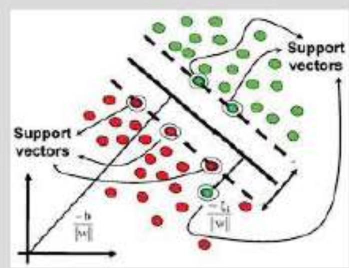
Parametric regression	Non-parametric regression	RTM inversion
Spectral relationships that are sensitive to specific vegetation properties	Advanced techniques that search for relationships between spectral data and biophysical variables	Models that simulate interactions between vegetation and radiation

$$NDVI = \frac{(\rho_{NIR} - \rho_{RED})}{(\rho_{NIR} + \rho_{RED})}$$

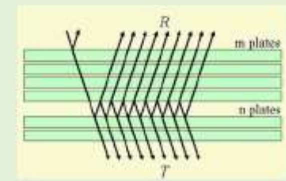
Normalized Difference Vegetation Index

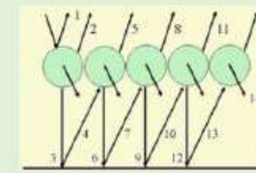




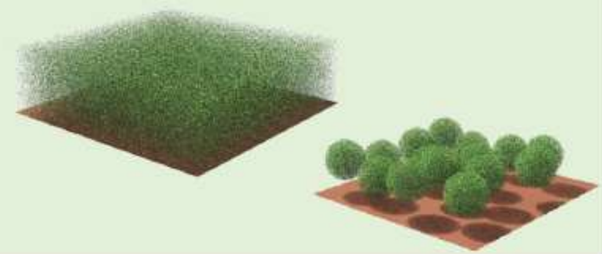


leaf

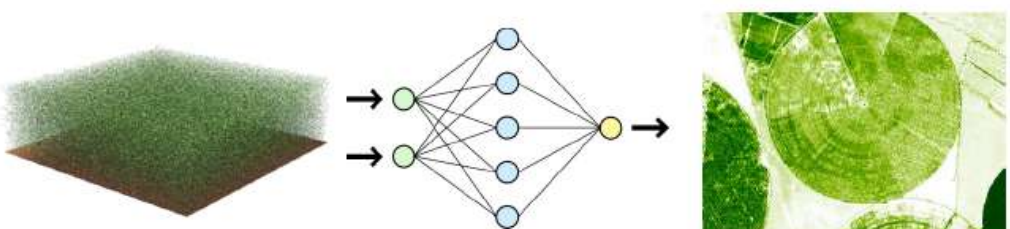




canopy



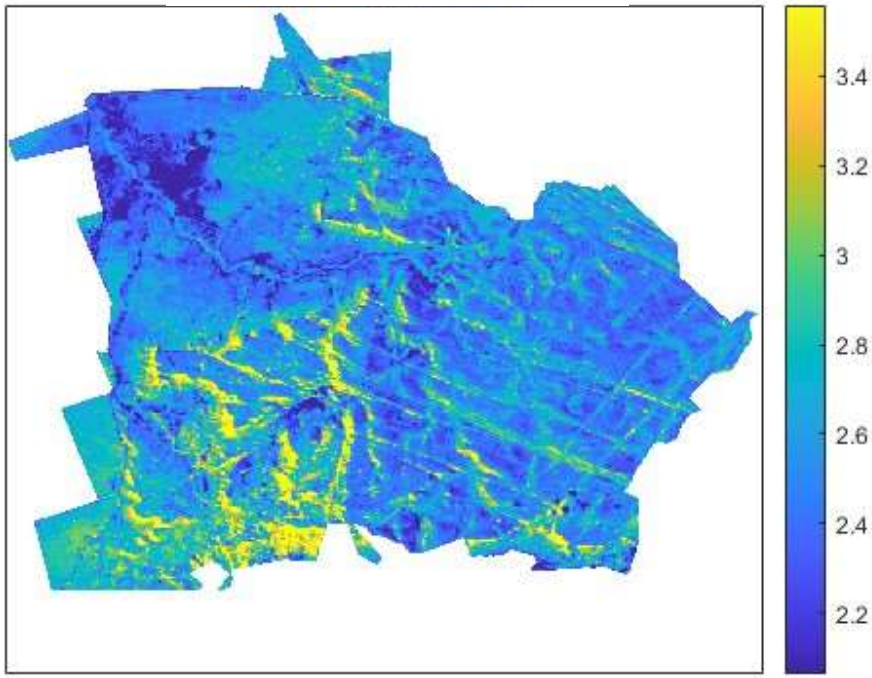
Methods of these different families can be combined: *hybrid methods*



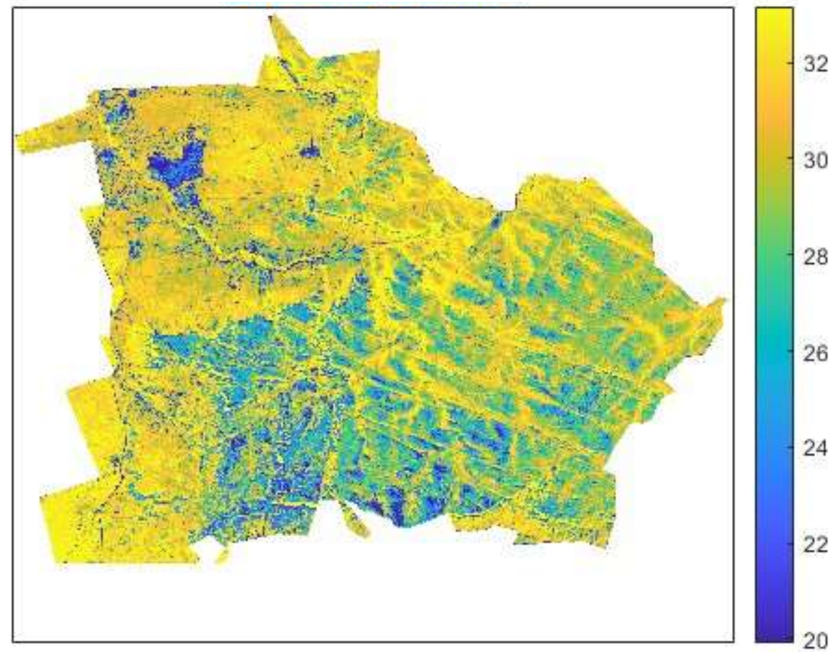
Verrelst and Rivera, 2018

- RFR estimations (without AL)

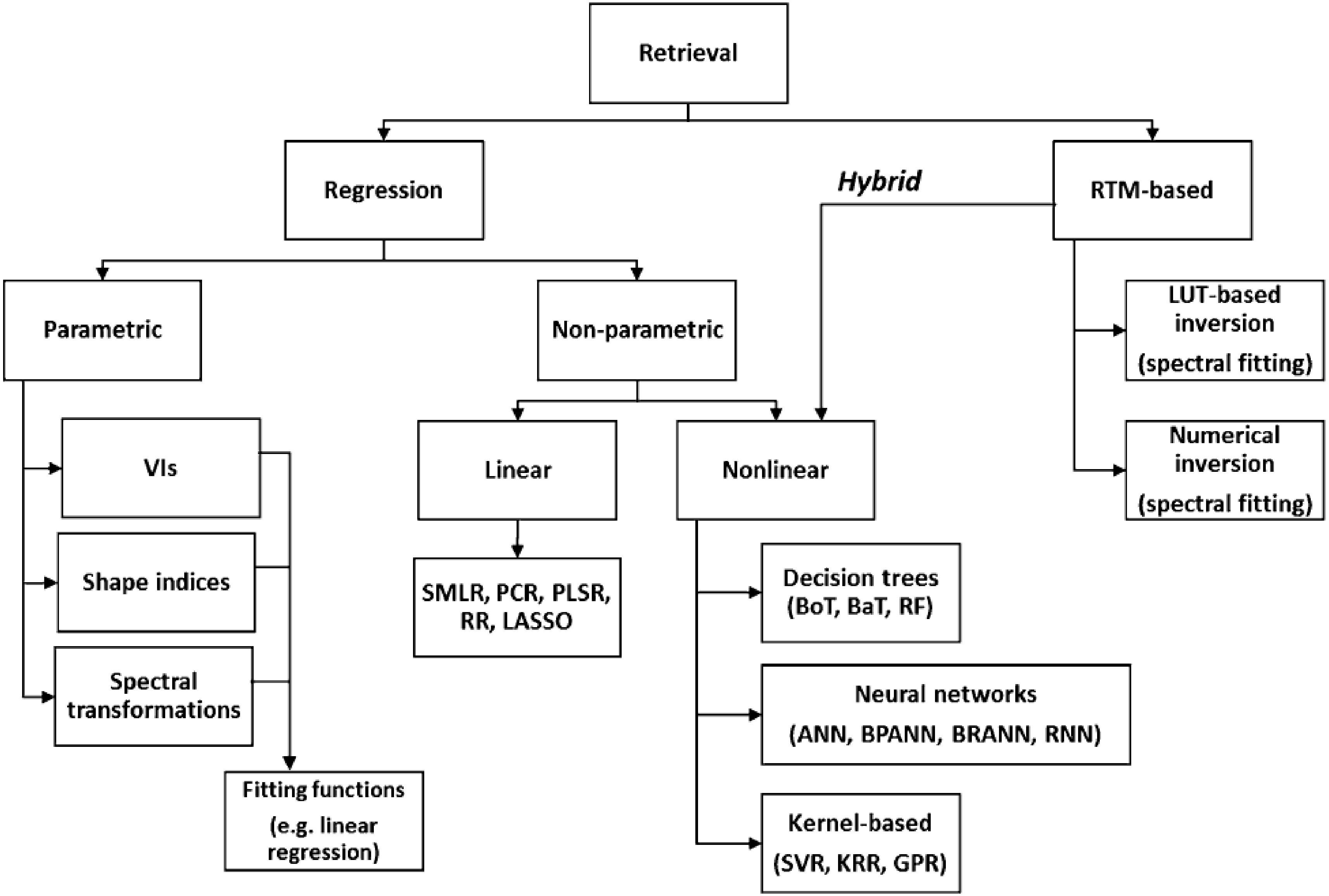
Map of LAI estimated



Map of LCC estimated



Overview of the main retrieval methods



Verrelst, J., Malenovský, Z., Van der Tol, C., Camps-Valls, G., Gastellu-Etchegorry, J.P., Lewis, P., North, P. and Moreno, J., 2019. Quantifying vegetation biophysical variables from imaging spectroscopy data: a review on retrieval methods. *Surveys in Geophysics*, *40*(3), pp.589-629.