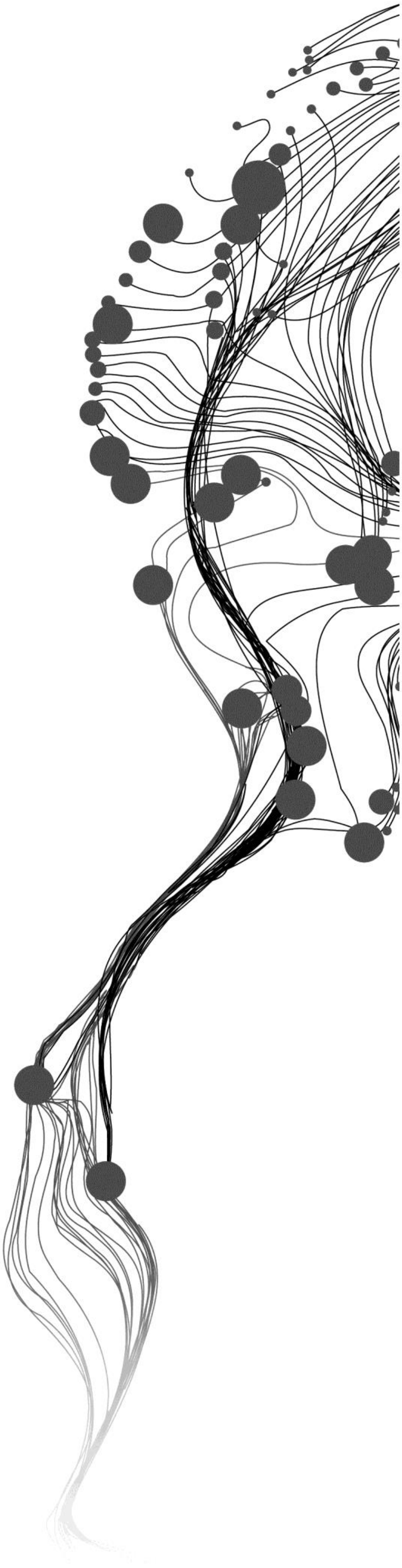
Multivariate Statistical Analysis for Estimating Grassland Leaf Area Index and Chlorophyll Content using Hyperspectral Data

HADI

June 2015

SUPERVISORS:
Dr. R. (Roshanak) Darvishzadeh
Prof. dr. A.K. (Andrew) Skidmore



Multivariate Statistical Analysis for Estimating Grassland Leaf Area Index and Chlorophyll Content using Hyperspectral Data

HADI

Enschede, The Netherlands, June 2015

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geo-information Science and Earth Observation for Environmental Modelling and Management

SUPERVISORS:
Dr. R. (Roshanak) Darvishzadeh (First supervisor)
Prof. Dr. A.K. (Andrew) Skidmore (Second supervisor)

THESIS ASSESSMENT BOARD:
Dr.Ir. C.A.J.M. (Kees) de Bie (Chair)
Dr. J. Clevers (External examiner, University of Wageningen)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Grassland habitat covers about one-quarter of the Earth's land surface, providing significant contribution to the world's total agricultural production, plant biodiversity, and carbon sequestration. Remote sensing (RS) provides a practical and cost-effective means for quantifying grassland biophysical and biochemical properties. However, grassland presents a challenge for RS due to the complexity of their spectral response. The advent of hyperspectral RS and the future launch of planned spaceborne hyperspectral missions will open up new possibilities over conventional multispectral RS to better quantify grassland characteristics. In this regard, hyperspectral data, while rich in information, presents a challenge for analysis due to its high dimensionality and multicollinearity. This present study investigated four selected high dimensional multivariate regression methods namely partial least squares regression (PLSR), regularization and shrinkage method Lasso, nonparametric Random Forest (RF) regression, and ensemble method Bayesian model averaging (BMA) to predict grassland leaf area index (LAI) and chlorophyll using field canopy hyperspectral measurements (n=185). For each regression model, three spectral transformations namely continuum-removal, first-derivative, and pseudo-absorbance were evaluated.

The results showed that relatively good predictive accuracy could be obtained for canopy-integrated chlorophyll content (cross-validated R^2=0.760; relative RMSE=32.1% or 0.28 $\mathrm{g\,m^{-2}}$) and LAI (R^2=0.719; relative RMSE=28.9% or 0.81 $\mathrm{m^2 m^{-2}}$), whereas leaf chlorophyll content could be predicted with relatively low accuracy (R^2=0.492; relative RMSE=14.8% or 4.45 $\mathrm{\mu g\,cm^{-2}}$). Multivariate methods utilizing all wavebands (whole spectral analysis) outperformed Lasso which performed waveband selection (optimal spectral analysis), suggesting some loss of information in the latter. Compared to the gold-standard model PLSR, no significant improvement in accuracy was obtained by the alternative multivariate regression models. Further, the spectral transformations in general did not significantly improve the accuracy either. This could suggest that the prediction errors were likely the results of grassland canopy spectral complexity due to heterogeneity such as the presence of different grass species having different canopy architecture. Therefore, approaches that explicitly account for structural differences such as model stratification based on species, incorporation of multiple structural parameters as in 3-D radiative transfer model for heterogeneous canopy, and data integration with radar or lidar capable of extracting the structural parameters are potentially useful.

Analysis of the identified important wavebands revealed the usefulness of wavebands in the far near-infrared and shortwave-infrared region attributed to water and carbon-based compound absorption features, for the prediction of both LAI and chlorophyll. Further, exclusion of wavebands in water absorption region to simulate spaceborne retrieval revealed the high significance of red edge bands. Consequently, our spectral simulation showed that, while not achieving prediction accuracy (CCC) as high as hyperspectral sensors, optical sensors with wavebands placed across the full optical domain (400-2400 nm) and importantly in the relatively narrow red edge region (such as Sentinel-2 MSI) offer a promising upscaling potential given their relatively high spatial resolution, provided that sufficient radiometric calibration and atmospheric correction are performed accordingly.

Overall, this study concluded that utilizing hyperspectral data and high dimensional multivariate statistical analysis allowed for successful estimation of grassland LAI and canopy chlorophyll content, provided useful insights on important wavebands, and concurrently on the upscaling potentials of the retrievals using sensors with different spectral resolutions.

ACKNOWLEDGEMENTS

This 22-month journey has been a dream come true, starting from Lund, Sweden and finishing here in ITC, the Netherlands. I am grateful to the financial support from European Union.

First and foremost, I would like to express my deepest gratitude to my first supervisor Dr. Roshanak Darvishzadeh for her patience, care, and advices in guiding the whole process of this MSc thesis. She redirected me back to the right path whenever I drifted from main goal.

My deep appreciation goes to my second supervisor Prof. Andrew Skidmore for his great advice I will never forget on focusing on the essentials: “less is more”.

I am greatly indebted with Raymond for his flexible approach and sincere help throughout my time here. Also to Dr. Michael Weir for assisting us as we got started here and in choosing research topic. Not to forget the whole GEM course staff here and in Lund: Laura, Marion, Petter to name a few; as well as all the great teachers from whom I greatly expand my knowledge.

I am extremely fortunate to have my GEM companions here at ITC Djaner, Fang, Badal, and those now in other universities Anton, San, Juan, Mieke, Hendrix, Sam. I will forever cherish the friendship we made, the long hours in GIS lab, the lunch in geocentrum, etc. Also the new friends I made here Kasi, Chao^2, Angela, Anni, Koen, Rabiul, Mbak Dewi, Amelia, Aslihan, and many more I could not possibly mention all, with whom I shared the many meaningful conversations and wonderful time.

To my family back home in Indonesia, the thought of them has always given me strength at difficult times during this research. I miss apa, ama, cece, koko, and adek so very much!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

A_i	Absorption feature
R_{cv}^2	Cross-validated coefficient of determination
A	Absorbance
ANN	Artificial neural network
Anth	Anthocyanin
ARD	Automatic relevance determination
BD	Band depth
BMA	Bayesian model averaging
BNA	Band depth normalized to area
BNC	Band depth normalized to center depth
BRDF	Bi-directional reflectance distribution function
Car	Carotenoid
CART	Classification and regression tree
CCA	Canonical component analysis
CCC	Canopy chlorophyll content
Chl	Chlorophyll
CHRIS	Compact High Resolution Imaging Spectrometer
CI	Chlorophyll index
CR	Continuum removal
CV	Cross-validation
EBV	Essential Biodiversity Variable
EeteS	End-to-end simulation tool
EM	Electromagnetic
EnMAP	Environmental Mapping and Analysis Program
EO	Earth observation
ERMES	An Earth Observation Model based information RicE Service
fAPAR	Fraction of absorbed PAR
FDR	First derivative reflectance
FDS	First derivative spectra
FWHM	Full width half maximum
GER	Geophysical and Environmental Research Corporation
GPP	Gross primary productivity
GPR	Gaussian process regression
HNBVI(s)	Hyperspectral narrow band indices
HyMap	Hyperspectral mapping imaging spectrometer
IPBES	Intergovernmental Science Policy Platform on Biodiversity and Ecosystem Services
IPCC	Intergovernmental Panel on Climate Change
KRR	Kernel ridge regression
LAD	Leaf angle distribution
LAI	Leaf area index
LAI-2000	Plant canopy analyzer LAI-2000 (LICOR Inc., Lincoln, NE, USA)
Lasso	Least absolute shrinkage and selection operator
LCC	Leaf chlorophyll content
LOOCV	Leave-one-out cross validation

LUE	Light use efficiency
MCMC	Monte Carlo Markov Chain
MLRA	Machine learning regression algorithm
MS	Multispectral
MSI	Multispectral Instrument (Sentinel-2)
NAOC	Normalized area over reflectance curve
NDVI	Normalized difference vegetation index
NEE	Net ecosystem exchange
NIR	Near-infrared
NPP	Net primary productivity
OLI	Operational Land Imager (Landsat-8)
OLS	Ordinary least squares
OOB	Out-of-bag
OSA	Optimal spectral analysis (feature selection)
PAI	Plant area index
PAR	Photosynthetically active radiation
PCA	Principal component analysis
PCR	Principal component regression
PIP	Posterior inclusion probability
PLSR	Partial least squares regression
PMP	Posterior model probability
R	Reflectance
r	Correlation coefficient
RBF	Radial basis function
REIP	Red edge inflection point
RF	Random Forest
RJ	Reversible jump
RS	Remote Sensing
RTM	Radiative Transfer Model
SLA	Specific leaf area
SMLR	Stepwise multiple linear regression
SPAD	SPAD-502 leaf chlorophyll meter (Minolta, Inc.)
SVR	Support vector regression
SWIR	Shortwave-infrared
VIP	Variable importance for projection (PLSR)
VIS	Visible domain (light)
WSA	Whole spectral analysis
WT	Wavelet transform
$mtry$	RF parameter: number of randomly selected covariates
$nRMSE_{cv}$	Cross-validated relative (to mean) root mean square error
$ntree$	RF parameter: number of trees

1. INTRODUCTION

1.1. Background and motivation

1.1.1. Remote sensing of vegetation: moving towards hyperspectral RS applications

With the advent of space technology, remote sensing (RS)—a technique for gathering information by a device without being in contact with the target—for Earth observation (EO) has provided a fast, efficient, non-destructive, and relatively low cost means (in contrast to traditional ground *in situ* survey methods) to retrieve various land and ocean surface characteristics over a large area all around the planet in the last fifty years since the first environmental satellites were launched in the 1960s (Wang et al., 2005; Tomppo et al., 2008; Jones & Vaughan, 2010, p. 92; Pu & Gong, 2011; Homolová et al., 2013). These techniques have been made possible based on the physical principle that different materials reflect and absorb light differently at different wavelength of the electromagnetic (EM) energy. In other words, objects can be characterized from their unique spectral signature. Among the various types of sensors, the sensors operating in the optical region of the EM (that is, visible and reflective infrared (near infrared and shortwave infrared)) have dominated the Earth observation system. This is especially true for vegetation application as most of the diagnostic absorption features of green vegetation are located in the optical part of the EM spectrum (Kokaly et al., 2009; Ustin et al., 2009).

Initially acquiring light reflectance from targets in only a few broad wavelength intervals (known as broadband or multispectral sensor), further sensor development in the early 1980s (Goetz, 2009) has led to increasingly more detailed measurement at finer spectral resolution—the hyperspectral sensor—recording light reflectance in a large number (typically hundreds and even thousands) of narrow contiguous wavelength intervals (or spectral bands) revealing full spectral signature of targets of interest (Figure 1). Hyperspectral RS increases the number of information (reflectance) collection channels from 3-10 to 100-1000, and increasing the spectral resolution from over 100 nm to 1-10 nm. This improvement in spectral resolution is needed as most terrestrial materials are characterized by spectral absorption features as wide as just 20-40 nm (Hunt, 1980).

Hyperspectral RS has improved the estimations of vegetation parameters and plant traits as compared to previous retrievals from traditional broadband multispectral data (Lee et al., 2004; Goetz, 2009; Zhao et al*.,* 2007). Traditional multispectral data contains limited information in a few broad spectral bands and typically one feature such as the normalized difference vegetation index (NDVI) employing two broad bands (NIR and red) is used for studying all vegetation characteristics. Hyperspectral data with hundreds of narrow bands has offered possibilities to establish unique features such as unique indices (hyperspectral narrowband vegetation indices (HNBVI) employing two, three, or more of the available bands) to study specific vegetation attributes: hyperspectral water/moisture indices to study plant water or moisture, hyperspectral biomass and structural indices to study biomass, hyperspectral biochemical indices to study plant pigments, hyperspectral lignin-celullose index, and so on (Thenkabail et al., 2014). HNBVI has improved the accuracy in modelling and mapping vegetation properties by about 10 to 30 per cent over broadband indices (Haboudane, 2004; Bolton & Friedl, 2013; Thenkabail et al., 2013).

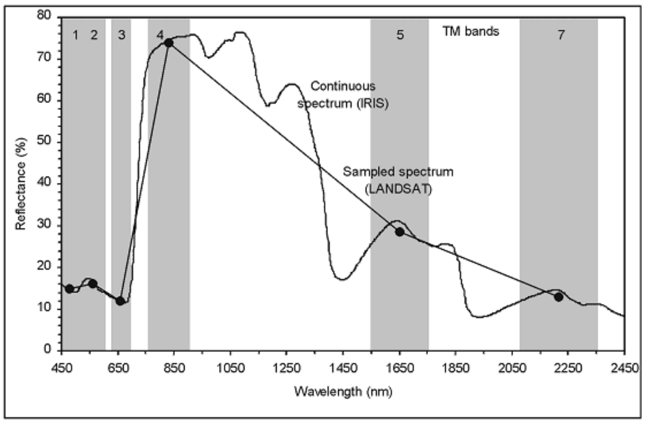


Figure 1. Data content of an example multispectral broadband (Landsat 7) and hyperspectral narrowband (IRIS) sensors (taken from Kumar et al., 2001). Shaded areas represent the broadband widths.

To date, hyperspectral data have been used to retrieve plant biochemical parameters including non-pigment (i.e., nutrient) biochemical such as nitrogen (Huang et al*.,* 2004; Axelsson et al., 2011; Wang et al*.,* 2012, Ramoelo et al., 2013), water content (Casas et al., 2014; Mirzaie et al., 2014), phosphorus (Mutanga et al., 2004; Axelsson et al., 2011), and lignin/cellulose (Daughtry et al., 2004; Zhao et al*.,* 2007); as well as pigment biochemical such as carotenoids (Blackburn, 2007), anthocyanins (Ustin et al., 2009), and especially chlorophyll (Yang et al., 2007; Zhao et al., 2007; Darvishzadeh et al., 2008; Lemaire et al., 2008; Qu et al., 2008; Atzberger et al., 2010; Axelsson et al., 2011; Huang & Blackburn, 2011; Navarro-Cerrillo et al., 2014). Biophysical parameters retrieved from hyperspectral data include fractional vegetation cover/crown closure (Boschetti et al., 2003; Pu & Gong, 2004; Guerschman et al., 2009; Somers et al., 2009), biomass/leaf mass per area (Casas et al., 2014; Schlerf et al., 2005; Ramoelo et al., 2013), and even more extensively, leaf area index (Boschetti et al., 2003; Casas et al., 2014; Schlerf et al., 2005; Lee et al., 2004; Yang et al., 2007; Haboudane, 2004; Pu & Gong, 2004; Darvishzadeh et al., 2008), as well as other structural parameters such as specific leaf area (Wittenberghe et al., 2014), diameter-at breast height and mean tree height (Schlerf et al., 2005; Cho et al., 2009).

Five major planned spaceborne hyperspectral missions are expected for launch in the near future (2015+ and 2020+, see Table 9, Appendix C), demonstrating the increasing recognition of the importance of hyperspectral remote sensing worldwide. The increasingly available airborne and spaceborne hyperspectral data has stimulated and sustained research interest to design new methods or to improve existing methods of retrieving the vegetation parameters from the unprecedented wealth of information in hyperspectral data (Lee et al., 2004).

1.1.2. Methods for vegetation retrieval from hyperspectral RS: statistical *vs* physically-based model

Two general approaches are now both being developed for retrieving vegetation characteristics from hyperspectral RS namely the empirical or statistically-based approach which accounts for a single plant trait at one time, and the physically-based approach which essentially attempts to represent (to model) the complex light scattering regime (the radiative transfer model (RTM)) involving multiple vegetation and other parameters at once (Dorigo et al., 2007).

Between the two approaches, the empirical or statistically-based methods have evidently been dominating and remained a viable approach in the field of hyperspectral RS of vegetation due to being simple, fast, and efficient, despite their lack of robustness and transferability (that is, they are potentially sensor, site, species, and time/season specific) in comparison to the potentially more robust physically-based methods (le Maire et al., 2004; Main et al., 2011). This is due to the still unresolved limitations of the physically-based models mainly the need for accurate auxiliary data on their many parameters, the model assumptions or boundary conditions (simplifications) to represent the scattering regime, the computational demand, and the ill-posed (non-unique solution) nature of the RTM inversion (Combal et al., 2003; Dorigo et al., 2007). The latter is caused by the fact that several combinations of the vegetation canopy biophysical and biochemical parameters result in similar spectral signature (Fang, 2003; Darvishzadeh et al., 2008; Main et al., 2011; Casas et al., 2014; Rivera et al., 2014). For these reasons, statistical approach continues to play an important role in hyperspectral RS of vegetation (Zhao et al., 2013) and improvement in statistically-based retrievals remains a high interest.

1.1.3. Importance of leaf area index (LAI) and chlorophyll

A review of hyperspectral RS studies (Pu & Gong, 2011; Homolová et al., 2013) in the last decade reveals the ever-increasing efforts in estimating two widely-studied critical vegetation parameters, namely the leaf area index (LAI) and chlorophyll. LAI and chlorophyll (which is related to and considered as operational proxy measurement of leaf nitrogen (Homolová et al., 2013)) are among the land surface characteristics important in ecosystem modeling which have been successfully estimated from remote sensing and Earth observation data (Turner, Ollinger, & Kimball, 2004).

In the broader context, LAI is also one of the more than fifty candidates of the essential climate (terrestrial) variables (ECVs) to be implemented in the Global Climate Observing System (GCOS) required to support the work of the United Nations Framework Convention on Climate Change and the Intergovernmental Panel on Climate Change (IPCC) (Bojinski et al., 2014). Plant chlorophyll on the other hand is related to species phenological traits which is a strong candidate of the essential biodiversity variables (EBVs)—an initiative inspired by the ECVs—which are currently under development by the Group on Earth Observations Biodiversity Observation Network as a follow up action to the IPCC-like mechanism for biodiversity known as the Intergovernmental Science Policy Platform on Biodiversity and Ecosystem Services (IPBES) (Larigauderie & Mooney, 2010; Pereira et al., 2013). From practical perspective, both LAI and chlorophyll have the potentials to be fully and directly estimated from remote sensing and Earth observation data.

LAI, generally defined as one-half (one-sided) the total surface area of leaves per unit ground area ($m^2 m^{-2}$; a dimensionless quantity) (Watson, 1947), is an important structural parameter closely related to energy and mass exchange processes between terrestrial ecosystems and atmosphere such as photosynthesis, respiration, transpiration, the carbon and nutrient cycle, and rainfall interception (Pu & Gong, 2011; Verrelst et al., 2012a). Thus, spatially-continuous (map of) LAI is a necessary input to various spatially distributed biogeochemical, ecosystem, and crop growth models to quantify these processes especially over a large area (Fischer et al., 1997; Colombo et al., 2003), for example the FOREST-BGC (Running & Coughlan, 1988), BIOME-BGC (Running & Hunt, 1993), and WOFOST (Diepen et al., 1989). Figure 20 (Appendix A) illustrates (albeit rather simplified) the intricate interrelationship between LAI and chlorophyll plant traits, and ecosystem processes (see text under caption).

Chlorophyll is the most important plant pigment and organic molecule on Earth found in the chloroplasts of green plants, which controls the amount of solar radiation that a leaf absorbs, and hence the photosynthetic potential and consequently primary production (Richardson et al., 2002; Davies, 2004; Gitelson et al., 2006). Therefore, total vegetation (canopy) chlorophyll is the plant trait most directly relevant for estimating plant productivity (such as crop yield) and carbon sequestration potential of vegetation (Gitelson et al., 2006). This leads to the possibility of a new framework to estimate productivity (GPP: gross primary productivity) as the product of total canopy chlorophyll and incoming photosynthetically active sun radiation (Gitelson et al., 2006; Peng et al., 2011). Chlorophyll is useful for diagnosis of plant stress (Zarco-Tejada et al., 2002; Baltzer & Thomas, 2005; Kopačková, 2012), nutrient management and precision agriculture (Schellberg et al., 2008) as it has been increasingly used as operational indicator of leaf nitrogen (Moran et al., 2000; Johnson, 2001; Homolová et al. 2013) Furthermore, the absorption features of chlorophyll along with other biochemicals such as leaf water have been found useful in mapping species composition and distribution (Kokaly et al., 2009; Siebke & Ball, 2009).

1.1.4. Importance of grassland habitat

Grasslands habitat (mainly pastures) covers some 26 per cent (3.44 billion hectares) of the Earth's land surface which is about twice that of arable land, and therefore contributes considerably to the world's total agricultural production (FAO, 2008; Schellberg et al., 2008). In some areas in temperate climate zones of Central Europe and in Northern America, intensively managed grassland adds more than 80 per cent to the agricultural land and hence substantially supports the production and output of milk and beef. Therefore, grassland (forage) production (yield) and quality are strongly linked to animal husbandry (Schellberg et al., 2008). In addition, grassland also accounts for almost half of 234 Centers of Plant Diversity (CPDs), and together stores 34 per cent of global terrestrial carbon stock (White, Murray, & Rohweder, 2000). Most of the precision agriculture research and development have focused on application in arable crops rather than on grassland (Schellberg et al., 2008).

In RS domain, grassland, especially mixed-species grassland, still presents a challenge for prediction of biophysical and biochemical properties due to the complexity of their spectral response. Grassland reflectance is complicated by the presence of a high fraction of non-photosynthetic vegetation (NPV) and exposed soil (He, Guo, & Wilmshurst, 2006; Beeri et al., 2007), grazing impact (Numata et al., 2007), and species heterogeneity creating complex canopy architecture (Cho et al., 2007; Darvishzadeh et al., 2008a; Darvishzadeh et al., 2008b). The unique spectral complexity of grassland canopies requires local studies at field level (proximal, using field spectrometer) to understand their basic spectral characteristics as a necessary step to assess the potential for upscaling the remote sensing retrieval to broader spatial scales using imaging spectrometer at airborne or spaceborne level (Numata, 2012).

1.2. Research problem and significance

Review of the literature (Table 8, Appendix B) reveals that a majority of hyperspectral studies for LAI and chlorophyll estimation has been carried out in agricultural cropland (15 out of 29 studies) and forest ecosystem (13 out of 29). There seems to be still limited number of studies in grassland ecosystem (4 studies). In addition, as was reviewed in more detail in Chapter 2, hyperspectral data is characterized by high dimensionality and multicollinearity and hence its utilization presents a challenge. Various statistical methods have been employed, and we have observed the following methodological trend: (1) The move from univariate methods based on hyperspectral narrowband indices towards multivariate methods; (2) The need for both optimal-spectral-analysis (band selection) and whole-spectral-analysis methods; and (3) The recent adoption of non-parametric machine learning regression algorithm.

Therefore, this present study addresses a two-fold research problem in the realm of hyperspectral RS of LAI and chlorophyll, namely (1) the apparent lack of hyperspectral RS studies of grassland LAI and chlorophyll; and (2) the need for methodological inter-comparison studies concerning hyperspectral data analysis using multivariate statistical methods. Based on the methodological review (presented in Chapter 2), the following methods known for their ability to cope with high dimensional multicollinear nature of hyperspectral data and for their interpretability (i.e., providing a measure of predictor (band) importance) have been selected for inter-comparison purpose:

- Partial least squares regression (PLSR) (the gold standard, linear, whole spectral analysis) which provides variable importance for the projection
- Least absolute shrinkage and selection operator (Lasso) (linear, optimal spectral analysis) which performs variable selection
- Random Forest (RF) regression (non-parametric/non-linear, ensemble (tree)-based whole spectral analysis) which provides permutation-based variable importance known as out-of-bag (OOB) error
- Bayesian model averaging or BMA (linear, ensemble-based whole spectral analysis) which provides posterior inclusion probability (PIP)

To our knowledge, these selected (justification in Chapter 2) potentially useful high-dimensional regression methods have not been compared in hyperspectral studies. Moreover, to our knowledge, Lasso and RF have not been tested for retrieval of LAI and chlorophyll from hyperspectral data, while only one study has used BMA (Table 8, Appendix B). The comparative analysis in this present study allows us to gain an insight on the performance of optimal spectral analysis *vs* whole spectral analysis, and whether the non-parametric (non-linear) model offers significant improvement over the conventional linear parametric methods. The study benefit from field spectral measurements which allow the evaluation of the selected high dimensional regression methods by minimizing other confounding factors (perturbing signals) such as atmospheric noise, mixed pixel effect (different land covers), and viewing geometry, all which affect the canopy signal at airborne or spaceborne measurement.

1.3. Research objectives

The aim of the present study is to evaluate the estimation of LAI and chlorophyll content in Mediterranean heterogeneous grasslands from field hyperspectral measurement using multivariate statistical methods. In particular, the focus is on evaluating the high-dimensional multivariate methods selected from methodological review in Chapter 2. The study area is the Majella National Park, Italy.

The specific objectives are:

1. To estimate LAI, leaf, and canopy chlorophyll content in heterogeneous grassland using field hyperspectral measurement and partial least squares regression (gold standard model), Lasso, Random Forest regression, and Bayesian model averaging.

2. To investigate the influence of spectral transformations namely continuum-removal, first-derivative, and pseudo-absorbance on the accuracy in predicting LAI, leaf, and canopy chlorophyll content using the above-mentioned multivariate regression models.

3. To investigate the effect of spectral resolution on the retrieval accuracy using the “optimum” (highest accuracy) model, and concurrently assess the upscaling potential (spectral domain) to existing and planned optical Earth-observation missions.

1.4. Research questions

The research questions include:

1. To which degree (assessed by predictive accuracy i.e. cross-validated coefficient of determination R_{cv}^2, and relative root mean square error $nRMSE_{cv}$) grassland LAI, LCC, and CCC can be predicted from field hyperspectral measurement?

2. Which of the three grassland variables (LCC, LAI and CCC) can be most accurately predicted (highest R_{cv}^2 and lowest $nRMSE_{cv}$)?

3. Which of the four investigated multivariate regression models (in combination with input spectral transformation) can most accurately predict LCC LAI, and CCC (i.e., which model is the "optimum" model)?

4. Which wavebands in the investigated models (and corresponding absorption features) are characterized to predict grassland LAI, LCC, and CCC?

5. How is the predictive accuracy of the "optimum" model in (3) affected by varying spectral resolution using the existing and planned optical sensors?

1.5. Research hypothesis and anticipated results

The research hypothesis or anticipated results associated with the above research questions are as follows:

1. Utilizing field hyperspectral data, there is high correlation (R_{cv}^2>0.5) and very low $nRMSE_{cv}$ (<10%) between estimated and measured LAI, LCC, and CCC.

2. CCC can be predicted with significantly higher accuracy (higher R_{cv}^2 and lower $nRMSE_{cv}$) than LCC and LAI.

3. Non-parametric Random Forest regression model applied to continuum-removed reflectance achieves the highest predictive accuracy for all grassland variables i.e., LCC, LAI and CCC. The predictive accuracy is significantly higher than the gold standard model PLSR.

4. In the investigated models, wavebands attributed to chlorophyll absorption features in the visible domain are most frequently selected/highest ranked for LCC and CCC retrieval, while wavebands in the red edge and near-infrared domain are most important for predicting LAI.

5. Sensors with higher spectral resolution give relatively higher prediction accuracy than sensors with lower spectral resolution.

2. LITERATURE REVIEW

This chapter introduces the basic physical principle of hyperspectral RS of LAI and chlorophyll, and subsequently reviews the relevant statistical-based methodology applied to hyperspectral data for vegetation application in general, and LAI and chlorophyll estimation in particular. The purpose was to identify the potential promising methods which need further investigation, or new method which has not been tested before for the particular task of estimating LAI and chlorophyll from hyperspectral measurements.

2.1. Hyperspectral RS of LAI and chlorophyll: the physical principles

Solar radiation arriving on a surface is either reflected, absorbed or transmitted. For leaves, solar radiation is either absorbed by leaf biochemical constituents and leaf water, or scattered (reflected or transmitted) by the structural elements such as cell walls (Jacquemoud & Baret, 1990). The nature and amount of reflection, absorption and transmission depend on the wavelength of the EM, incidence angle (which causes either specular or diffuse scattering), surface roughness (leaf cuticular surface), and importantly the differences in the leaf structure and biochemical constituents (Kumar et al., 2001). The main absorbing biochemical in leaves are chlorophyll and other pigments in the visible domain (roughly between 400 and 700 nm), and water as well as various carbon based biochemicals (lignin, cellulose, protein) in the near-infrared (700 to 1300 nm) and shortwave (mid-) infrared (1300 to 2500 nm). This and the fact that leaves and other vegetation elements such as stems and fruits typically contain similar biochemical constituents create a unique overall spectral signature of vegetation as shown in Figure 2 below.

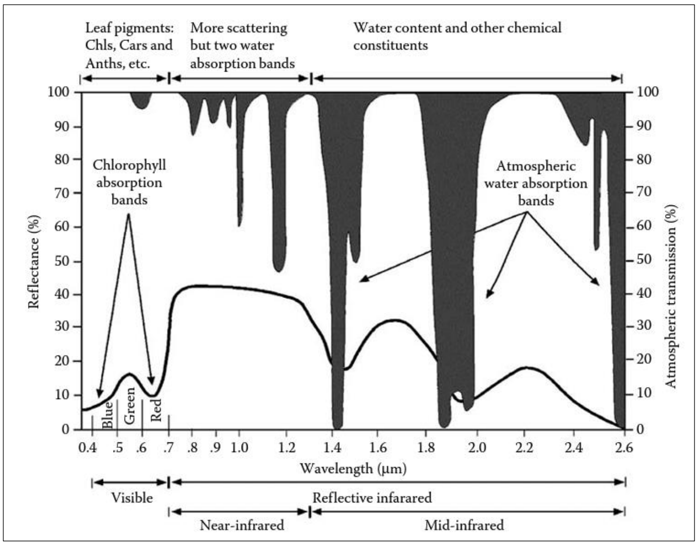


Figure 2. Typical spectral reflectance curve of vegetation (taken from Pu & Gong, 2011, adapted from Jensen, 2007)

Table 1 lists the complete known absorption features associated to the various plant constituents in the optical domain. However, it is important to note that these known absorption features are from controlled laboratory measurement (*in vivo*) of dried (pure) plant compounds which may differ from *in situ* field measurement of fresh leaves (Curran, 1989) where typically the relatively stronger and broader water absorption features tend to mask/obscure the subtler signal from leaf biochemicals in the NIR and SWIR region (Kokaly & Clark, 1999).

Table 1. Known absorption features related to plant compounds (taken from Kumar et al. (2001), compiled from Elvidge (1987), Williams & Norris (1987), Himmelsbach et al. (1988), Curran (1989), and Elvidge (1990); also Horler et al. (1983), Ben-Dor et al. (1997), and Dawson & Curran (1998)). This table was used for waveband interpretation analysis.

No	Wavelength (nm)	Absorbing Compounds	No	Wavelength (nm)	Absorbing Compounds
C1	430	Chl-a	C24	1736	Cellulose
C2	460	Chl-b	C25	1780	Cellulose, sugar, starch
C3	640	Chl-b	C26	1820	Cellulose
C4	660	Chl-a	C27	1900	Starch
C5	800	Lignin, tannin	C28	1924	Cellulose
C6	910	Protein	C29	1940	Water, protein, lignin, cellulose
C7	930	Oil	C30	1960	Starch, sugar
C8	970	Water, starch	C31	1980	Protein
C9	990	Starch	C32	2000	Starch
C10	1020	Protein	C33	2060	Protein, nitrogen
C11	1040	Oil	C34	2080	Starch, sugar
C12	1120	Lignin	C35	2100	Starch, cellulose
C13	1200	Water, cellulose, starch, lignin	C36	2130	Protein
C14	1400	Water	C37	2180	Protein, nitrogen
C15	1420	Lignin	C38	2240	Protein
C16	1450	Starch, sugar, water, lignin	C39	2250	Starch
C17	1490	Cellulose, sugar	C40	2270	Cellulose, sugar, starch
C18	1510	Protein, nitrogen	C41	2280	Starch, cellulose
C19	1530	Starch	C42	2300	Protein, nitrogen
C20	1540	Starch, cellulose	C43	2310	Oil
C21	1580	Starch, sugar	C44	2320	Starch
C22	1690	Lignin, starch, protein	C45	2340	Cellulose
C23	1730	Protein	C46	2350	Cellulose, nitrogen, protein

Although leaf optical properties are well understood (Jacquemoud & Baret, 1990), vegetation canopy reflectance is also influenced by multiple light interactions between canopy elements (Jones & Vaughan, 2010, p. 49). That is, the radiative properties of the canopy are determined by canopy structure/architecture (biophysical attributes) such as the spatial arrangement and orientation of leaves (i.e. leaf angle distribution (LAD) and foliage clumping) which cause shadow and hotspot effects (Asner, 1998). The variable widely used to describe the canopy structure is leaf area index or LAI (Homolová et al., 2013).

Leaf chlorophyll and LAI have a known influence on the vegetation reflectance. Figure 3 shows how increase in leaf chlorophyll decreases overall reflectance in VIS (less in the low-light-penetration blue wavelengths, more in green) and especially rapidly around chlorophyll absorption maxima in red. Chlorophyll-a has absorption maxima *in vivo* around 420, 490, and 660 nm and Chl-b around 435 and 643 nm (Kumar et al., 2001; Blackburn, 2007). However, it is also known that *in situ* Chl-a absorbs at both 450 and 670 nm (Pu & Gong, 2011). Also visible in Figure 3 is the broadening of the Chl absorption in red with increasing amount of chlorophyll, shifting the red edge inflection point—graphically, the point of transition from concave to convex shape, or the point of maximum slope in the reflectance—towards longer wavelengths (Kumar et al., 2001). LAI on the other hand strongly influences the canopy reflectance in NIR. Figure 4 shows the simulated reflectance of varying LAI values (keeping other biochemical and biophysical parameter constant), generally showing increasing NIR reflectance with increasing LAI.

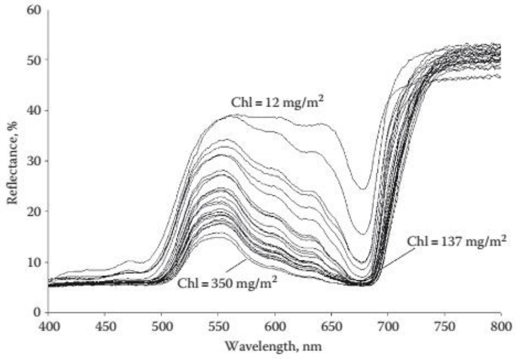


Figure 3. Leaf (beach leaves) reflectance spectra with different chlorophyll content. (taken from (Gitelson, 2012))

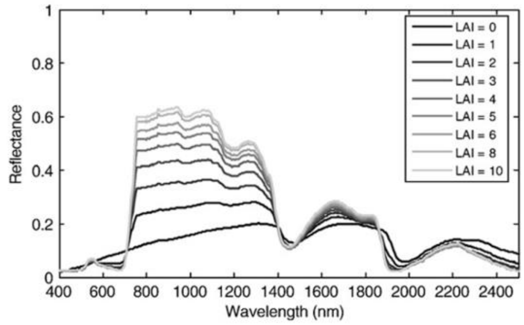


Figure 4. Effect of LAI on canopy reflectance simulated using PROSAIL fixing other leaf and canopy parameters (taken from Jacquemoud et al., 2009).

2.2. Hyperspectral RS of LAI and chlorophyll: a review of statistical methods

In the context of RS of vegetation, the statistical approach models the empirical relationship (regression analysis) between spectral or transformation of spectral data into spectral features and the target vegetation properties. The spectral features extracted from hyperspectral RS include primarily the long developed vegetation indices which are computed by mathematical combination of two (i.e., originally making use the sharp increase in vegetation reflectance from red to NIR in the red edge) or more of the original spectral bands, reviewed in Jones & Vaughan (2010, p. 169-171). The basic form of the spectral indices ranges from simple ratio, simple difference, to the normalized difference form. Further modification made along the way include the soil-line based indices which aims to minimize soil background reflectance from soil below a sparse canopy, atmospherically-resistant indices which purpose is to minimize atmospheric noise/attenuation to the canopy signal by including additional band in the atmospherically-sensitive blue region, and the hybrid of the two.

With the advent of hyperspectral RS, a large variety of hyperspectral narrowband vegetation indices (HNBVI), with carefully selected optimal hyperspectral narrowbands which are sensitive to different vegetation biophysical and biochemical parameters have been formulated, for example as compiled by Pu & Gong (2011), Thenkabail et al. (2011), Roberto et al. (2012), and Roberts et al. (2012). Main et al. (2011) also listed 73 published spectral indices (until 2008) specially formulated for estimating leaf and/or canopy chlorophyll of which a majority of them in principle is based on the red edge feature. The move towards higher spectral resolution data also led to the development of other spectral features such as the red edge inflection position (REIP, e.g. Cho & Skidmore, 2006) using derivative spectra, continuum-removed spectral absorption (band) depths (Kokaly & Clark, 1999), and area under reflectance curves or spectral integration features (Delegido et al., 2010; Li et al., 2014).

Table 8 (Appendix B) lists the studies which use hyperspectral data to estimate LAI and chlorophyll using statistical-based methods in the last two decades.

2.2.1. From univariate to multivariate statistical methods

From reviewing the studies in Table 8 (Appendix B), it is clear that methods based on spectral indices formed with combination of selected narrowbands, the hyperspectral narrow band vegetation indices (HNBVI), have shown their overwhelming dominance (17 out of 29 studies). Spectral indices has always been advocated based on their advantage in that the mathematical transformation (normalization) minimizes the variability in spectral reflectance caused by external factors such as scene illumination differences, soil background reflectance, and atmospheric scattering; as well as internal factors such as leaf angle distribution and canopy structure in relation to the viewing geometry. Indeed, all the efforts to improve the indices revolve around improving the sensitivity (as well as the linearity) of the indices to the biochemical or biophysical quantity (in wide range) of interest and suppressing other unwanted confounding factors (e.g. chlorophyll indices designed to have high sensitivity to foliar chlorophyll but with minimum sensitivity to LAI).

However, despite the development and various proposed modifications of the index forms or optimal wavelengths (the centers and width; although the optical region sensitive to LAI and chlorophyll is somewhat well understood), at present there is still no clear consensus on the best universal HNBVI for robustly predicting LAI and chlorophyll (Ustin et al., 2009; Main et al., 2011; Zhao et al., 2013). The modifications in practice do not generally result in substantial improvements in index performance because although they may emphasize key parts of the response, they also tend to be increasingly sensitive to small errors or noise in spectral measurement (Rivera et al., 2014).

Owing to the drawbacks of the univariate methods based on HNBVI elaborated above, there has been an increasing application of multivariate statistical methods which exploit the full spectra (information) of hyperspectral data instead of the empirically or theoretically (based on knowledge on leaf optical and canopy radiative properties described earlier) selected narrowbands in the visible domain corresponding to absorption features of chlorophyll (Blackburn, 2007), or narrowbands in the red edge and NIR region sensitive to LAI variation. Stepwise multiple linear regression (SMLR), principal component analysis (PCA) and regression (PCR), canonical component analysis (CCA), and partial least squares regression (PLSR) are among the most popular multivariate statistical techniques as shown in Table 8 (Appendix B).

Exploration of all the complete wavelengths often reveals the usefulness of off-absorption-center wavelengths to improve the estimation, especially at canopy scale, in which univariate methods such as HNBVIs based on absorption centers weaken in their performance or sensitiveness due to the effect of complex canopy structure (especially LAI and LAD) on signal propagation from leaf to canopy level (Asner, 1998), and when dealing with multiple species in an attempt to create a more universal/generalized predictive model (Blackburn, 2007; Majeke et al., 2008). The absorption features of pigments in VIS and water and other biochemicals in NIR and SWIR are useful for estimating LAI (Elvidge, 1990). In another study, Main et al. (2011) observes the utility of off-chlorophyll absorption center wavebands (690-730 nm) in estimating LCC for combined species dataset. This can be partly explained by the fact that reflectance at the chlorophyll absorption feature center will saturate even at relatively low chemical concentrations due to the already low light penetration in VIS, as well as the overlapping absorption features of plant compounds which share the same chemical bonds (Kumar et al., 2001). For example, the strong O-H bond is component of absorption feature of water, cellulose, sugar, starch, and lignin. Thus, concerning vegetation reflectance, 1-3 bands may not be enough to represent one specific vegetation biophysical or biochemical property, and incorporating multiple bands (optimal spectral analysis) or even all bands (whole spectral analysis)—requiring high-dimensional statistical techniques—can better represent the vegetation property (Darvishzadeh et al., 2008) and is useful to account for the various sources of spectra variability.

2.2.2. The challenge of hyperspectral data analysis with multivariate methods: the curse of high dimensionality

Hyperspectral data containing hundreds and even thousands of contiguous narrow wavebands, while containing much richer information than multispectral data, presents a real challenge when performing multivariate statistical analysis on them. The reason is many of the bands are redundant i.e., highly or even nearly perfectly correlated (Thenkabail et al., 2013; Thenkabail et al., 2014), thus adding more bands do not always necessarily mean adding information content. In other words, hyperspectral data are said to be high dimensional because there are a large number of predictors or features, often much larger than the observations ($p \gg n$), which precludes the use of classical ordinary least squares methods (designed for $n > p$ problem) for regression analysis simply because when $p > n$ or $p \approx n$ the model will be too flexible and graphically the least squares regression line will perfectly fit (overfit) the data points/observations (James et al., 2013, p. 239).

It is therefore needed to perform dimension reduction to hyperspectral data to remove data redundancy i.e., to extract unique information pertaining to specific vegetation biophysical or biochemical variables. In general, hyperspectral data mining and dimension reduction can be done by two procedures namely (1) optimal-spectral-analysis or OSA methods (following Thenkabail et al. (2014)), and (2) whole-spectral-analysis or WSA methods. OSA (also known as feature selection methods) results in a subset of the original wavebands, whereas WSA makes use of all wavebands and include feature extraction methods which create new features by combination of several wavebands (feature space transformation) such as principal components (Bajwa & Kulkarni, 2011).

An example of optimal-spectral-analysis method is the widely used variable selection method SMLR. However, since hyperspectral data are highly multicollinear (adjacent bands are similar), SMLR procedure has been widely criticized as being vulnerable in this setting mainly due to the problem of over-fitting (Curran, 1989; Blackburn, 2007) in which the large number of wavelengths compared with the number of samples and major plant constituents tends to exaggerate the goodness of fit—due to highly biased unconstrained regression coefficients and risk of selection of non-relevant bands simply because they have noise patterns correlated to the response chemical—of the chemical prediction model calibration (Bajwa & Kulkarni, 2011). Grossman et al. (1996) showed the other problems with SMLR for hyperspectral band selection namely that the selected bands were not related to known absorption bands and bands selected in other similar studies, varied among datasets and chemical expression unit (concentration per mass or content per area), and were sensitive to the samples entered into the regression. Using other model selection criteria such as the popular Akaike's Information Criteria (AIC) to guide SMLR search potentially leads the selection of more variables than necessary in high dimensional setting (Mallick & Yi, 2013).

PCA, PCR, and PLSR are examples of whole-spectral-analysis methods, all in principal work by transforming the feature space into low dimensional latent variable ($t < p$) space, in which the orthogonal (uncorrelated) latent variables (principal components or PLS factors) are simply the linear combination of the original variables (individual bands) (Bajwa & Kulkarni, 2011). The feature space transformation differs in its criterion: PCA and PCR produce components by maximizing the information content (the variance) in the predictor variables space (the hyperspectral narrowbands), whereas PLSR maximizes the information content in both the predictor and response variables space i.e., by maximizing the covariance between them. PCA is an unsupervised method in which the minimum variance threshold is preset to determine the optimal number of PCs, whereas PCR and PLSR retains the number of components/factors that essentially maximize linear relationship with the response variable (James et al., 2013, p. 231-238).

2.2.3. Optimal spectral analysis *vs* whole spectral analysis

Both optimal spectral analysis and whole spectral analysis methods for hyperspectral data analysis have their own drawbacks and advantages. On one hand given the redundancy and high dimensional nature of hyperspectral data, a careful selection of most useful bands for a given application—estimating LAI and chlorophyll in this present study—is called for especially to improve the model interpretability in terms of the physiological importance of selected wavebands, which ultimately can help the design and optimal use of future multi- and super- spectral (10-50 bands (Verrelst et al., 2012a)) sensors devoted for vegetation monitoring. The WSA methods on the other hand are typically performed by projecting the original bands into latent variables (principal components or factors), while advantageous as they essentially make use of the entire hyperspectral bands, suffers from not-as-clear interpretability in terms of which of the original bands are most useful as they have been linearly combined into the latent variables. Therefore, it can be argued that both OSA and WSA methods remain equally valuable for hyperspectral data analysis, and there is a need to compare both OSA and WSA.

With regards to the OSA, there is a need for other variable selection methods as alternative to the criticized SMLR. There seems to be a potential of adopting the well-established regularization/shrinkage and variable subset selection methods for high dimensional multivariate linear regression (Mallick & Yi, 2013). The regularization methods in principle overcome the problem of over-fitting in the presence of multicollinearity and under high dimensional setting by imposing some form of penalty (constraint) on the objective (loss) function (i.e., sum of squared error) to control or regularize (to shrink) the model parameter (regression coefficient) estimates from being inflated and causing over-fitting. Among the penalty functions which have been proposed in the literature, the Lasso penalty (Tibshirani, 1996) has gained popularity given its useful property in effectively shrinking the coefficient estimates of the unimportant predictors to zero, thus performing variable (bands) selection improving the model interpretability in addition to accuracy.

Among the WSA-related methods, recently increasingly PLSR—a technique borrowed from chemometrics—has been shown to outperform the conventional stepwise regression (and univariate methods based on HNBVI) in general for estimating foliar biochemistry (as reviewed in Majeke et al., 2008), and in particular LAI and/or chlorophyll (e.g. Darvishzadeh et al., 2008; Atzberger et al., 2010; Herrmann et al., 2011) from hyperspectral data. Additionally, despite transforming original wavebands into latent variables (PLS factors), PLSR provides a measure of variable importance called variable importance for the projection or VIP (Wold, Sjöström, & Eriksson, 2001).

2.2.4. The recent adoption of machine learning regression algorithm

Another noticeable methodological trend from the review of previous studies in Table 8 (Appendix B) is the increasing adoption of machine learning regression algorithms (MRLAs, e.g. as reviewed in Camps-Valls (2009)) in studies retrieving vegetation variables (including LAI and chlorophyll) from hyperspectral RS data such as the artificial neural network (ANN) and Gaussian process regression (GPR). These methods began to be explored thanks to the present unprecedented computational speed and efficiency. Perhaps the biggest improvement by MRLAs is their non-parametric nature (not assuming particular distribution, e.g. unlike the linear regression which assumes normal distribution of the prediction residuals) and greater flexibility to cope with the strong non-linearity of the functional relationship between the reflectance and the target parameters (Verrelst et al., 2012a).

Previous statistical methods mostly have developed an empirical relationship using simple linear regression, and somehow attempt to consider this non-linearity by a non-linear transformation of the original reflectance values such as logarithmic, inverse logarithmic, and hyperspectral indices (Zhao et al., 2013). Verrelst et al. (2012a) demonstrated the utility of the quite recently introduced kernel family MLRAs namely support vector regression (SVR), kernel ridge regression (KRR), ANN, and GPR for prediction of LAI and chlorophyll of different crop species; of which GPR outperforms the others. However the study used superspectral resolution data simulated at Sentinel-2 and Sentinel-3 configuration, and not the full hyperspectral configuration. Recently, Yi, Shi, & Choi (2011) showed that GPR suffers from large variance of parameter estimation and high predictive errors for high dimensional dataset with correlated covariates. The standard variable (feature) selection approach for GPR using the automatic relevance determination (ARD) covariance function/kernel (Chen & Martin, 2009) can be problematic because the number of hyperparameters (i.e., the lengthscales for each spectral band) will simply be too many in high dimensional setting and consequently can cause over-fitting (Cawley & Talbot, 2010). Thus, despites their flexibility which may improve predictive accuracy to some extent, the emerging MLRA methods are difficult to implement, have high risk of over-fitting, and often lack physical interpretability i.e., behave like a 'black-box' (Liang, 2007; Zhao et al., 2013).

Despite the above-mentioned drawbacks of MLRAs, there still seems to be a need to evaluate the performance of the non-parametric model against the conventional parametric statistical methods long dominated the vegetation studies using RS, in particular hyperspectral RS. One important class of the non-parametric model seemingly not as popular as the above kernel-based methods in remote sensing area, which has the attractive property of handling high dimensionality well (where ANN typically fails) without over-fitting, and better interpretability (Ghasemi & Tavakoli, 2013), is the tree (CART: classification and regression tree)-based Random Forest (RF) method (Breiman, 2001). The basic idea behind RF is to improve prediction accuracy by growing a large number of independent learners (decorrelated trees) and obtaining prediction by averaging (consensus) the prediction values from all these learners (trees) in the ensemble (forest) for each sample (observation). This approach is especially useful for dataset with a large number of correlated predictors (Breiman, 2001; James et al., 2013, p. 320) such as hyperspectral data.

RF offers better model interpretability not only by the simpler mathematical concept of the algorithm (simply averaging predictions from all trees) as compared to the kernel family, but also by providing very useful measure of variable importance called the OOB (out-of-bag) error. The importance of each variable is evaluated based on how much worse the prediction would be if the data for that variable were permuted (shuffled) randomly, assessed by the difference in OOB error between the permuted and non-permuted samples aggregated across the entire forest (Kuhn & Johnson, 2013, p. 202). Yet another RF advantage is that is has only two tuning parameters hence not too difficult implementation. Therefore, RF regression seems to be a good candidate of non-parametric (MLRA) method for estimating vegetation variables from hyperspectral data.

Despite their advantages, and successful application in spectroscopic calibration (Ghasemi & Tavakoli, 2013), RF has been used more in classification problem in general RS (Adam et al., 2014), and hyperspectral RS domain (Chan & Paelinckx, 2008) and rarely for regression problem albeit a few studies such as Mutanga, Adam, & Cho (2012), Abdel-Rahman, Ahmed, & Ismail (2013), and Adam et al. (2014).

Finally, another method arises in the literature presented in Table 8 (Appendix B) is the Bayesian model averaging (BMA) (Zhao et al., 2013), which is attractive for high dimensional correlated hyperspectral data as it addresses the uncertainty in selecting the optimal wavebands (and discarding the rest of the bands which may be useful to some extent albeit not the best predictors) for estimating vegetation parameters. BMA differs from the standard 'single best model' paradigm in that rather than selecting one best model with one best subset of predictors, it seeks to leverage on all the plausible competing models to improve predictive performance (Hoeting et al., 1999; Wintle et al., 2003). Another salient feature of BMA is that it provides information about the relative variable (band) importance as indicated by marginal probability (relative frequency) of that band being included in the top performing models. Zhao et al. (2013) demonstrated the superior performance of BMA in terms of accuracy and identification of important bands as compared to SMLR and PLSR methods examining a large spectral-chemical dataset representing over 80 tree and crop species across the globe.

2.2.5. The role of spectral transformation

Hyperspectral RS measurement providing full continuous spectral reflectance profile has also made possible the use of spectral transformation techniques adopted from chemometrics field, namely the standard derivatives (often first derivative spectra (FDS)), continuum-removal (CR) and pseudo-absorbance (log-transformed (Log (1/Reflectance)). These spectral transformation techniques serve to enhance and isolate the absorption features of foliar biochemicals of interest, while minimizing unwanted perturbing signal from atmospheric, background (e.g. soil), and water absorption effects; as well as reducing data redundancy (Kokaly & Clark, 1999; Ramoelo et al., 2011). The pseudo-absorbance (log (1/R)) is performed due to the almost linear relation between them and the concentration of the absorbing component (Kumar et al., 2001).

2.3. Conclusion

Based on the review of the statistical methods, four multivariate methods have been identified to be compared in this present study, namely the gold-standard partial least squares regression (PLSR), an optimal-spectral-analysis regularization method Lasso, the non-parametric Random Forest (RF) regression, and the ensemble method Bayesian model averaging. These regression methods were applied together with the original and also transformed spectra using continuum-removal, first derivative, and pseudo-absorbance.

"*Everything must be made as simple as possible. But not simpler.*"—Albert Einstein (1879-1955)

3. MATERIALS AND METHODS

This chapter introduces the study area and data, the spectral transformations, multivariate regression models, and model validation procedure.

3.1. Study area

Majella National Park (total area 740.95 km^2) is located in the southern part of Abruzzo region at a distance of 40 km from the Adriatic sea, encompassing 39 municipalities in the provinces of Chieti, Pescara, and L'Aquila, in Italy, approximately at latitude 41°52' to 42°14' N and longitude 13°14' to 13°50' E (Figure 5). It is estimated that 75 per cent of all Europe's flora and fauna species are represented in Abruzzo region, and the park houses over 78 per cent of the mammal species in this region (including the Apennine wolf, Marsicano brown bear, Abruzzo chamois, otter, and roe deer), over 130 bird species, and over 1,700 flora species (of which many are endemic), making the park a significant biodiversity 'hot spot' internationally. The park is characterized by a territory dominated by mountains with 55 per cent of its area situated over 2,000 meters above sea level. Owing to the park's wideness and altitude, many climate types are represented despite the dominant temperate oceanic climate. The park is certified as one of the only 12 parks (having at least 100 km^2 of wilderness/untouched nature) in the PAN Parks network, a Europe-wide non-governmental organisation founded by World Wildlife Fund dedicated to the preservation of Europe's natural habitats and fragile ecosystem.

The grasslands (plant formations consist of herbs) occupy approximately 29.5 per cent of the entire protected area. The grasslands have high species richness and are home to many orchids and other rare and endemic species. Numerous birds (some are rare species) occupy the grasslands during spring snowmelt when the area is temporarily flooded to rest and during summer to feed and nest. The grasslands lie in between the oak woodlands at the lower altitudes (400 m to 600 m) and beech forests (1200 m to 1800 m) at the higher altitudes.The dominant grass species include *Brachypodium genuense*, *Briza media*, *Bromus erectus* and *Festuca sp*. Herbs include *Helichrysum italicum*, *Galium verum*, *Trifolium pratense*, *Plantago lanceolata*, *Sanguisorba officinalis* and *Ononis spinosa* (Cho et al., 2007; Darvishzadeh et al., 2008).

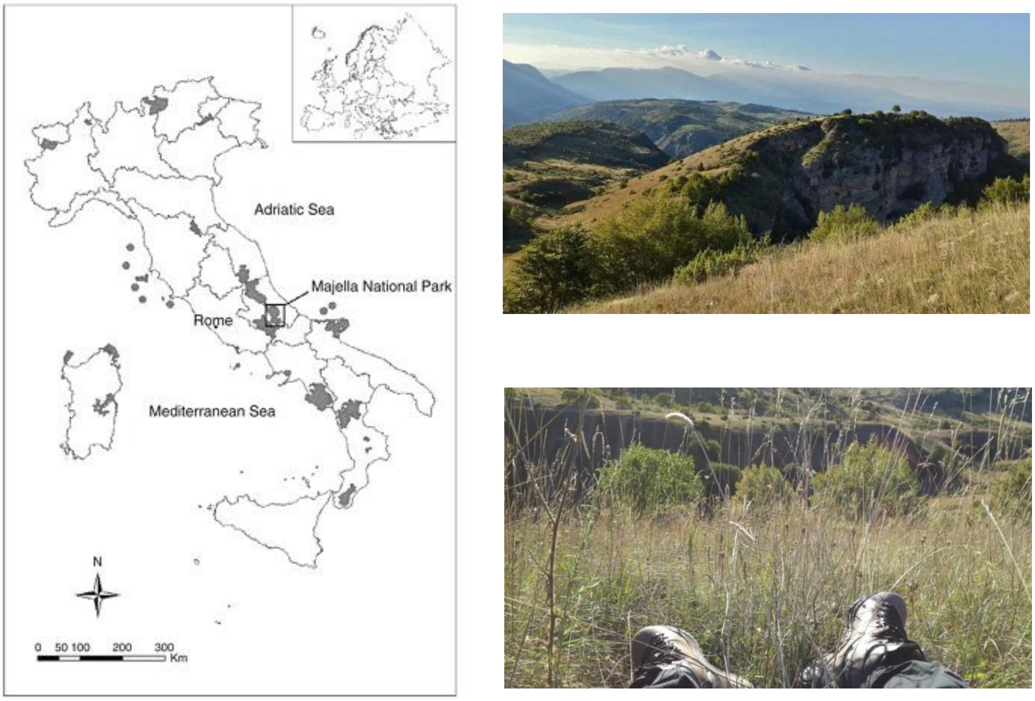


Figure 5. (Left) Location of the study area, Majella National Park, Italy (taken from Darvishzadeh et al., 2008). (Right) Example grassland area in the park (taken by author of the present study in September 2014)

3.2. Data

Data used in this study was collected in a field campaign by Darvishzadeh et al. (2008) between June 15 and July 15, 2005. Field measurement of canopy reflectance, leaf area index (LAI), and leaf chlorophyll was carried out in a total of 191 plots (1 m x 1 m) randomly generated within the grassland strata (based on land cover map provided by the park's management) in the park's area.

From each plot, 15 replicates of canopy spectral measurements were taken using GER 3700 spectroradiometer (Geophysical and Environmental Research Corporation, Buffalo, New York) subject to averaging to suppress much of the noise in spectral measurement. The instrument's wavelength range is 350 nm to 2500 nm, with a spectral sampling of 1.5 nm in the 350 nm to 1050 nm range, 6.2 nm in the 1050 nm to 1900 nm range, and 9.5 nm in the 1900 nm to 2500 nm range. The sensor was held approximately 1 m above the ground at nadir position and observes ground area with 45 cm diameter. To minimize atmospheric perturbations and viewing angle effects, spectral measurements were made on clear sunny days between 11:30 a.m. and 2:00 p.m.

In each plot, LAI was non-destructively measured using Plant Canopy Analyzer LAI-2000 (LICOR Inc., Lincoln, NE, USA). The instrument measures the gap fraction in five zenith angles, using measurements of incoming solar radiation above and below the canopy. The gap fraction is used to estimate effective LAI assuming random spatial distribution of leaves. The measurements were taken either under clear skies with low solar elevation or under overcast conditions, facing away from the sun, on the same day as canopy spectral measurement. In each grass plot reference samples of above-canopy radiation and subsequently five below canopy-samples were taken and averaged. The LAI measured using LAI-2000 corresponds to plant area index (PAI) which includes the photosynthetic and non-photosynthetic components (Chen et al., 1997). However, non-photosynthetic components were almost non-existent in the study area (Darvishzadeh, et al., 2008).

Leaf chlorophyll content (LCC) was non-destructively measured with SPAD-502 Leaf Chlorophyll Meter (Minolta, Inc.) which measures the transmittance in the red (650 nm) and near-infrared (920 nm) wavelength regions. Thirty (30) leaves representing the dominant species were randomly selected in each plot, and their SPAD readings were averaged and converted into LCC (μg cm^{-2}) by empirical calibration function (Markwell et al., 1995). The total canopy chlorophyll content (CCC; in g m^{-2}) for each plot was obtained by multiplying LCC with the corresponding LAI (CCC = LAI * LCC) (Gitelson et al., 2005; He & Mui, 2010). LCC measurement in six plots were recognized as outliers and were excluded, thus 185 plots with all LAI, LCC, and CCC measurements were analyzed in this present study. Table 2 shows the summary statistics of the grassland variables measured in the field hyperspectral campaign.

Table 2. Summary statistics of the measured grassland variables (n=185). LAI is leaf area index; l LCC is leaf chlorophyll content; and CCC is canopy chlorophyll content.

Measured variable	*Min*	*Mean*	*Max*	*StDev*	*Range*	*Coefficient of variation*
LAI (m^2m^{-2})	0.39	2.81	7.34	1.5	6.95	0.53
LCC $(\mu g\ cm^{-2})$	17.1	30.07	49.66	6.12	32.55	0.2
CCC $(g\ m^{-2})$	0.1	0.87	2.7	0.55	2.56	0.63

Darvishzadeh et al. (2008) showed that due to relatively much higher coefficient of variation for LAI than LCC, CCC is highly correlated with LAI (r=0.94). The correlation coefficient between CCC and LCC is 0.50, and between LAI and LCC is 0.24. CCC therefore contains both structural (LAI) and chlorophyll signal.

3.3. Spectral pre-processing and transformation

3.3.1. Savitzky-Golay filter

After removing very noisy bands below 400 nm and above 2400 nm, the plot-average (15 replicates) spectra was further smoothened using a moving Savitzky-Golay filter (Nevius & Pardue, 1984) with a frame size of 15 data points (2nd degree polynomial). Mathematically, the filter operates simply as a weighted sum of neighbouring values as follows:

$$x_{j*} = \frac{1}{N}\sum_{h=-k}^{k} c_h x_{j+h} \tag{1}$$

Where x_{j*} is the new value, N is a normalizing coefficient, k is the number of neighbour values at each side of j and c_h are pre-computed coefficients, that depends on the chosen polynomial order and degree.

Next, three spectral transformations were applied to the smoothed spectra, namely (1) standard first derivative, (2) continuum-removal, and (3) pseudo absorbance.

3.3.2. Standard first derivative

The first derivative (Dawson & Curran, 1998) was calculated using a first-difference transformation of the reflectance spectrum as follows:

$$FDR_{\lambda i} = \frac{R_{\lambda(j+1)} - R_{\lambda(j)}}{\Delta_\lambda} \qquad (2)$$

Where FDR is the first derivative reflectance at a wavelength i, midpoint between wavebands j and $(j+1)$, $R_{\lambda(j)}$ is the reflectance at the j waveband, $R_{\lambda(j+1)}$ is the reflectance at the $(j+1)$ waveband, and Δ_λ is the difference in wavelengths between j and $(j+1)$.

3.3.3. Continuum removal

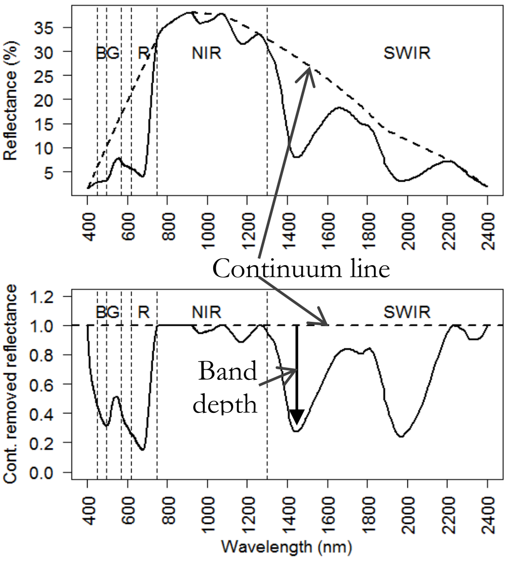


Figure 6. Top: continuum line (dashed bold line) fit on top of reflectance (solid line). Bottom: reflectance is then normalized (divided) by corresponding values of continuum line, enhancing absorption features especially in VIS. Absorption depth (arrow) is 1 minus the CR value.

Continuum removal (CR) was applied to full spectrum from 400 to 2400 nm (Huang et al., 2004; Axelsson et al., 2013) assuming not all the absorption features useful for estimating grassland LAI, LCC, and CCC were exactly known. CR is performed by first approximating the continuum line (the reflectance baseline or albedo) which is the convex hull fitted over the top of a spectrum joining both shoulders of local spectrum maxima. The continuum-removed reflectance of each wavelength λ_i in the absorption band (region bounded by the absorption shoulders), $R'_{(\lambda_i)}$, is then calculated by dividing the original reflectance $R_{(\lambda_i)}$by the corresponding reflectance value of the fitted continuum line $R_{c(\lambda_i)}$ (Figure 6) as:

$$R'_{(\lambda_i)} = \frac{R_{(\lambda_i)}}{R_{c(\lambda_i)}} \qquad (3)$$

The wavelengths at the endpoints (shoulders) which the continuum line connects lie on the hull and therefore have value equal to 1 (the maximum CR value i.e., zero absorption), while the rest of the absorption bands have CR value between 0 and 1 (Mutanga & Skidmore, 2003). This albedo (slope of the continuum line) normalization technique thus enhances/isolates the local absorption features and corrects for the apparent shifts of the band minimum (absorption center) due to wavelength dependent scattering, returning the true absorption band center (Clark & Roush, 1984). CR was performed using the 'prospectr' package (Stevens & Ramirez-Lopez, 2013) in R statistical environment (R Core Team, 2014).

It was also investigated if the band depth ($D_{(\lambda_i)} = 1 - R'_{(\lambda_i)}$) normalization techniques proposed by (Kokaly & Clark, 1999) and later extended by Curran, Dungan, & Peterson (2001) can improve the linear relationship (assessed by simple Pearson's linear correlation r) between band depth and the grassland variables (i.e., LAI, LCC, and CCC). To do this, CR was applied to local absorption features identified visually from the grass reflectance profile. The band depth of each wavelength ($D_{(\lambda_i)}$) was normalized by: (1) the band depth at the absorption center (D_c) into $BNC_{(\lambda_i)}$; or (2) the area of the absorption feature (A_c) into $BNA_{(\lambda_i)}$ as follows:

$$BNC_{(\lambda_i)} = D_{(\lambda_i)}/D_c \qquad (4)$$

$$BNA_{(\lambda_i)} = D_{(\lambda_i)}/A_c \qquad (5)$$

3.3.4. Pseudo-absorbance

The transformation of the original reflectance into pseudo-absorbance is as follows (Kumar et al., 2001):

$$A_{(\lambda_i)} = log_{10}\left(\frac{1}{R_{(\lambda_i)}}\right) \tag{6}$$

3.4. Regression analysis

Four multivariate statistical methods were tested to build a predictive model for LAI, LCC, and CCC (the three response variables y) from field hyperspectral measurement (the independent variable x), without (untransformed) and with the 3 spectral transformations just described. For each method, a general description of how it works is provided below, followed by the necessary summary of the mathematical foundation.

3.4.1. Partial least squares regression

Partial least squares regression (PLSR) is a linear multivariate model useful to analyze data with many (high dimensional), noisy, and collinear predictors (Wold et al., 2001) the like of hyperspectral data. It is a feature-extraction (projection-based) method for dimension reduction similar to the principal component regression (PCR). Both PCR and PLSR work by transforming the original predictors $X_1, X_2, \dots, X_p$ into uncorrelated latent variables $Z_1, Z_2, \dots, Z_M$ with $M < p$ (thus lower dimensionality) where Z is weighted linear combinations (or directions) of the original p predictors

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{7}$$

For some constants $\phi_{1m}, \phi_{2m}, \dots, \phi_{pm}$, $m = 1, \dots, M$. Linear regression model is then fit to the latent variables (i.e., PLS factors) in the more manageable low dimension orthogonal space (M):

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i\,, \qquad i = 1, \dots, n \tag{8}$$

Where θ_0 is the regression intercept, θ_m is the regression coefficients for each of the PLS factor z across the n observations; y_i being the response and ϵ_i the i.i.d. residuals.

The key is in how the transformation is done: while PCR (which is simply linear regression applied to principal components obtained by principal component analysis PCA) 'extracts' the M latent variables by maximizing the information (the variance) only in the p predictors (in other words, to seek the direction of M components that explain most of the variance in the predictors), PLSR improves by maximizing the covariance between the factors Z and response Y during the decomposition of the original predictors P. Thus the M PLS factors importantly also explain most of the variance in Y, improving the prediction

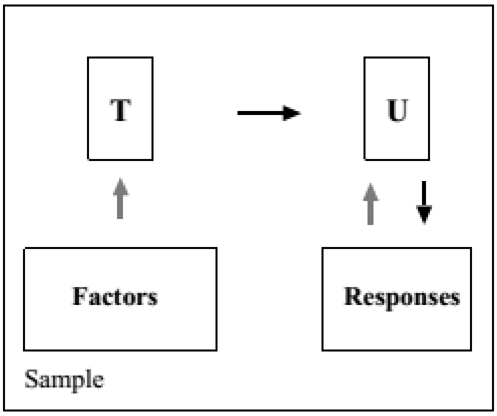


Figure 7. Simplified schematic outline of PLSR model. PLSR directly extracts latent variable *T* (also called *X-scores*) and *U* (*Y-scores*) from the factors (predictors) and responses, respectively. *T* are used to predict *U*, and then *U* are used to construct the predictions for the responses (indirect modelling). Taken from Tobias (1995).

accuracy (Figure 7). The mathematical details on how this is achieved can be found in Wold et al. (2001).

To ensure the best predictive performance, the PLSR model is tuned/calibrated by means of cross-validation (see section 3.5) to obtain the optimum number of PLS factors.

3.4.1.1. PLSR Variable Importance in Projection

A useful summary measure of variable importance in PLSR is the VIP (variable importance for the projection). Predictors (spectral bands) which have high VIP scores are predictors which are important for both the modelling (transformation/projection) of PLS factors, and for modelling the response Y. The VIP score for each predictor (waveband) λ_i is computed as follows:

$$VIP_{\lambda_i} = \sqrt{d} \sum_{k=1}^{h} \boldsymbol{v}_k \left(\boldsymbol{w}_{k_{\lambda_i}}\right)^2 / \sum_{k=1}^{h} \boldsymbol{v}_k \tag{9}$$

Where d is the number of predictors and h is the number of PLS factors. The equation shows VIP is computed as the proportion of the fraction of the explained variance of $\boldsymbol{X}$ expressed by $\boldsymbol{v}_k$ weighted by the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ represented by $\boldsymbol{w}_{k_{\lambda_i}}$ for each waveband λ_i. Typically "VIP scores > 1" rule is used to identify important predictors (Tran et al., 2014). PLSR analysis was implemented using 'caret' (Kuhn et al., 2015) and 'pls' (Bjorn-Helge, Wehrens, & Liland, 2013) packages in R statistical environment (R Core Team, 2014).

3.4.2. Lasso

In high dimensional setting with correlated (redundant) predictors, subset selection such as the conventional stepwise multiple linear regression (SMLR) risks over-fitting the calibration/training data due to inflated (biased) regression parameter (coefficient) estimates, which produces model with low bias but high variance (see Figure 21, Appendix D for further information) i.e., low prediction accuracy when extrapolated to validation/testing data (Kuhn & Johnson, 2013, p. 123). To combat collinearity, it is therefore useful to implement procedure that gives more biased regression model but with lower variance and thus better predictive performance. One way of accomplishing this is by controlling (regularizing or shrinking) the coefficient estimates by adding a "penalty" to the objective/loss function (i.e., sum of squared errors (SSE) in regression). Different penalty functions have been proposed, and this present study used one widely used penalty capable to effectively shrinks some of the coefficients to zero (thus performing variable selection) namely Lasso. That said, other penalties exist (Mallick & Yi, 2013) but they are out of the scope of this present study.

Lasso (Tibshirani, 1996)—short for least absolute shrinkage and selection operator—adds the following penalty function (the second term) called the L_1 penalty to the standard OLS loss function (the first term, sum-of-squares):

$$SSE_{L_1} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\lambda \sum_{j=1}^{P} |\beta_j|}_{penalty\ term} \tag{10}$$

Where as usual y_i is actual/measured response, $\hat{y}_i$ is the estimated/predicted response, with n observations. The penalty is simply constraining the sum of the absolute value of the regression coefficients β_j across all p predictors. Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage (increasing model bias) and more of the smaller coefficients are shrunk to zero (i.e., more candidate predictors are removed leading to more parsimonious model). The tuning parameter λ is chosen by cross-validation (see section 3.5). This procedure of variable (spectral band) selection is continuous (coefficients are shrunk towards each other and "evolves" together as λ increases) and thus more reliable/stable and less computationally intensive than the discrete process of step-by-step retaining or discarding variables as in SMLR (Fan & Li, 2001) especially in high dimensional multicollinear setting. Figure 22 (Appendix E) shows the Lasso procedure graphically. Lasso was implemented using caret (Kuhn et al., 2015) and 'glmnet' package (Friedman, Hastie, & Tibshirani, 2010) in R statistical environment (R Core Team, 2014).

3.4.3. Random Forest regression

Random forest (RF) proposed by Breiman (2001) is a tree-based ensemble machine learning regression method especially useful for the "large p, small n" and correlated predictors problem of high-dimensional dataset such as genomic data (Chen & Ishwaran, 2012; Kursa, 2014), spectral multivariate calibration (Ghasemi & Tavakoli, 2013) and, without exception, hyperspectral data mostly for classification (Ham, Crawford, & Ghosh, 2005; Lawrence, Wood, & Sheley, 2006; Chan & Paelinckx, 2008) and increasingly for regression task e.g. Mutanga, Adam, & Cho (2012), Abdel-Rahman, Ahmed, & Ismail (2013), and Adam et al. (2014).

RF is a recent important improvement to the original CART (classification and regression tree) method introduced earlier (Breiman et al., 1984), and to the related application of the powerful statistical method for predictive variance-reduction called bagging or bootstrap aggregation (Breiman, 1996). Tree-based methods work in principle by stratifying or segmenting the predictor space into a number of simple (decision) regions. This class of learners has evolved as an important non-parametric (no formal distributional assumptions) capable to fit highly non-linear interactions, deal well with irrelevant predictor variables and robust to outliers in the predictor variables (Cutler et al., 2009). From an arbitrary single tree, the idea is then extended to bagging based on the idea of consensus (ensemble) modelling, which aims to improve predictive performance through variance reduction by aggregating (averaging) predictions from a large collection of trees (in a 'forest'), each tree built by taking bootstrap (random sampling with replacement) samples from the dataset. In this way, bagging is a stochastic procedure which differs from the deterministic CART. Random forest was then proposed to further improve bagging by de-correlating the trees i.e., making the trees different. This is done by randomly sampling the predictors in each split of the trees, instead of keeping all predictors which makes similar trees.

RF therefore essentially works by constructing a tree-based ensemble of many independent base learners (trees), by random selection of (1) observations to grow the trees, and (2) predictors at each node of the trees. More precisely, the RF regression procedure can be found in Hastie, Tibshirani, & Friedman (2009, p. 588). Figure 23 (Appendix F) explains in more detail the procedure.

Two parameters need to be tuned for RF namely (1) number of trees T_b ($ntree$), and (2) number of variables (spectral bands) m randomly selected at each split ($mtry$). In this present study, they are tuned by cross-validation (see section 2.5.). The RF regression was implemented using 'caret' (Kuhn et al., 2015) and 'randomForest' package in R statistical environment (R Core Team, 2014). Preliminary experiments testing $ntree$=500, 1000, and 5000 in combination with $mtry$=1,2,3,…583 showed no substantial changes and similar gradual (no apparent local minima) pattern in cross-validated training error, suggesting RF is relatively insensitive to the tuning parameters as claimed (Cutler et al., 2009). Therefore, the maximum $ntree$=5000 (RF does not overfit with increasing $ntree$ (Cutler et al., 2009)), which showed somewhat more stable error evolution with $mtry$ was used in all RF regression runs to ensure adequately large number of trees, along with varying $mtry$=5, 10, 15, 30, 50, 100, and 200 as computational time allowed. To note is that by default the recommended $mtry$ is $p/3$=584 bands/3≈200.

3.4.3.1. RF variable importance

A useful measure of variable importance in RF is the out-of-bag (OOB) error or permutation-based variable importance. In each tree constructed with bootstrap samples (~70 per cent of training dataset), the remaining (~30 per cent) observations not used to grow the regression tree (the OOB) are passed down the tree and the predicted values are computed. To calculate the importance of variable j (spectral band λ_i), the values of variable j are randomly permuted whilst keeping all other predictor variables fixed. These modified OOB data are again passed down the tree and the predicted values are computed. The variable importance is then computed based on the difference in OOB error between the real/non-permuted dataset and permuted dataset across all trees. In other words, the importance of variable j is assessed by how much worse the RF prediction accuracy becomes if the data for variable j is permuted randomly (Prasad, Iverson, & Liaw, 2006; Cutler et al., 2009).

3.4.4. Bayesian model averaging

Statistical inversion of vegetation variables from hyperspectral RS faces two main challenges in building the predictive models: (1) which predictors (spectral bands) and (2) what model forms (structure) to establish the relation between reflectance and the target variables (i.e., grassland LAI and chlorophyll in this study) (Zhao et al., 2013). The lack of consensus on both issues (i.e., common wavebands and spectral indices for estimating the same biochemical, e.g. Féret et al., 2011; Main et al., 2011) puts a considerable risk of model misspecification using the standard "single best model" paradigm (e.g., selecting a single best subset of bands using stepwise multiple regression, or single best narrowband index). Bayesian model averaging (BMA) provides a mechanism to address this model selection uncertainty essentially by instead of selecting one best model, it leverages on the many competing plausible models (our "hypothesis"). To put simply, BMA works by averaging (ensemble learning, similar to bagging as in Random Forest) across a large set of models, with each model weighted—using Bayesian inference framework—based on the probability of it being the true model (or, the PMP: posterior model probability) (Hoeting et al., 1999).

In this present study, BMA was implemented based on the multiple linear regression form following Zhao et al. (2013) and Raftery, Madigan, & Hoeting (1997):

$$\mathbf{y} = \mathbf{X}_{M_i}\boldsymbol{\beta}_{M_i} + \epsilon, \qquad i = 1, \dots, 2^p \tag{11}$$

where $\mathbf{y}$ is the vector of response values (i.e., LAI, LCC, or CCC), $\mathbf{X}$ is the matrix of predictors (hyperspectral narrowbands), $\boldsymbol{\beta}$ is the regression coefficients, and ϵ is the usual assumed i.i.d. residuals. Importantly, the subscript M_i signifies one model i from all 2^p possible model configurations constructed from hyperspectral data with p number of bands. That is, we have a model space $M = \{M_1, M_2, \dots, M_{2^p}\}$. $\mathbf{X}_{M_i}$therefore is the set of selected bands in candidate model M_i with regression coefficients $\boldsymbol{\beta}_{M_i}$.

Following Bayes' theorem, the "usefullness" (i.e., the weights for averaging) of each model M_i—called the posterior model probability PMP = $p(M_i|D)$—is calculated as follows:

$$p(M_i|D) = \frac{p(D|M_i)\pi(M_i)}{\sum_{k=1}^{2^p} p(D|M_k)\pi(M_k)}, \quad i = 1, \dots, 2^p \tag{12}$$

Where in the numerator, $p(D|M_i)$ is the marginal likelihood of model M_i, defined as the probability of observing the data Dgiven model M_i, where $D = \{\mathbf{x}_i, y_i\}_{i=1,\ldots,n}$ i.e., n observations relating spectral bands $\mathbf{x}$ and response grassland variables y; $\pi(M_i)$ is the prior probability (our prior belief, which needs to be elicited beforehand) of model M_i being the true model; and the denominator is simply the normalization term which is the integrated likelihood/probability across all 2^p possible models in model space M as explained above. Obviously, the model space M can be too large to evaluate/enumerate exhaustively, and fortunately there is an efficient way of exploring the model space using a "guided" sampling algorithm called the Monte Carlo Markov Chain (MCMC) sampler. To put simply, the MCMC allows us to sample a finite (N) number of the most important models $\{M^{(t)}\}_{t=1,\ldots,N;\ N \ll 2^p}$ which allows us to derive meaningful inference and prediction from them, provided the sampling chain has converged with enough number of iterations. Once the PMP is computed, the model weighted posterior distribution (e.g., characterized by mean and standard deviation) for any statistic of interest such as the regression coefficients $\boldsymbol{\beta}$, which prior distribution was also elicited beforehand) can be computed. The posterior distribution of the predictions for the target grassland variables is also computed similarly by averaging the individual predictions over the MCMC-sampled models. The detail derivation can be found in Hoeting et al. (1999).

BMA was implemented using the 'BMS' package (Feldkircher & Zeugner, 2009) in R statistical environment (R Core Team, 2014). BMA requires setting the prior model probability (random or fixed), prior model size (i.e., how many spectral bands), the prior of regression coefficients variance (mean assumed zero) defined according to Zellner's g, the MCMC sampler, and the number of burn in (initial iterations to be discarded) and iterations. Our preliminary trials with different configuration of these settings and different starting model (Lasso, stepwise MLR) with the full 584 bands of the hyperspectral data revealed that the MCMC chains always failed to converge (as indicated by low correlation between the iteration counts and analytical PMPs, see Zeugner, 2011) even with combined chains of total 30 million iterations. Therefore, we decided to resample the field hyperspectral data to HyMap spectral resolution (i.e., 119 bands in 400-2400 nm; on average 15 nm spectral resolution up to 1313 nm, 13 nm for 1409-1800 nm, 17 nm beyond that). The final model settings chosen were: (1) 'dilution' prior model probability proposed by George, (2010) for redundant model space, (2) EBL g-prior (based on personal communication with the BMS package author), (3) reversible-jump sampler (RJ-MCMC, described in Madigan & York (1995)), (4) prior model size with 7 bands, and with sufficiently long iterations: (5) half million burn in followed by 5 million iterations.

3.4.4.1. BMA variable importance

BMA also provides information about relative band importance as indicated by the marginal probability (relative occurrence frequency) of that band being included in the models sampled in the MCMC chain, or alternatively the top performing models (Zhao et al., 2013). This is called the posterior inclusion probability (PIP).

3.5. Model calibration and validation

To avoid biased validation due to a single random training-test set partition (Kuhn & Johnson, 2013, p. 78), one way to achieve the above is by cross-validation (CV), and as the data (n=185) and computational capacity allows, this present study implemented a uniform stratified 10-fold CV (recommended as best by Kohavi, 1995) for all models to fairly assess their true predictive performance. "Stratified" here refers to that the folds/partitions were created based on the full range of values of the response variable, to make each fold as representative as possible to the whole dataset. Standard leave-one-out cross-validation (LOOCV) was not used because it would have been too computationally expensive (especially when tuning random forest) and is a high-variance procedure because the training samples ($n - 1$) are similar (removing 1 sample at a time) (Hastie et al., 2009, p. 242). To ensure fair comparison among the models, the same 10-fold training-test set partitions were used.

Additionally, to ensure an independent validation as well as calibration, model calibration and parameter tuning was performed only using the training set (the test set remains "unseen") in each of the ten cross-validation folds/runs. Tuning/optimizing model parameters (model selection) using the complete dataset has been shown to be an overly-optimistic procedure for model assessment (Hastie et al., 2009, p. 245; Cawley & Talbot, 2010). It is therefore recommended to carry out a double/nested CV procedure: the inner 10-fold CV within the training set for model selection/calibration/parameter tuning, followed by the outer 10-fold CV for model assessment/validation. Figure 24 (Appendix G) illustrates the CV procedure employed in this study. The model training and validation was carried out using 'caret' (Kuhn et al., 2015) package in R statistical environment (R Core Team, 2014)

3.5.1. Model comparison

To determine if the differences between the accuracy of the models tested are statistically significant (thanks to identical resampled data sets i.e. the CV folds/partitions), the non-parametric Mann-Whitney *U* test (also known as Wilcoxon rank-sum test) was used to compare the distribution of R^2_{cv} and $nRMSE_{cv}$values not assuming normal distribution. The null hypothesis is that the models compared (pairwise, two-at-a-time) have equivalent accuracies, or analogously, that the mean difference in accuracy for the resampled data sets is zero. The *U* test was performed using the 'stats' package in R statistical environment (R Core Team, 2014).

3.6. Interpreting the importance of spectral wavebands

As described above, all the multivariate methods tested readily give a measure or indicator of variable importance (hence chosen in this study): VIP in PLSR, OOB error in RF regression, PIP in BMA, and the selected wavebands in Lasso. VIP, OOB error, and PIP rank the waveband importance and therefore are to be interpreted relative to other wavebands, except for VIP of which VIP scores greater than 1 indicate important bands (Wold et al., 2001). The average and standard deviation (stability) of the band importance over the 10 CV runs were examined accordingly. For Lasso, the wavebands selected the more number of times (maximum 10 times if selected in all 10 CV runs) indicate the most important ones. Waveband importance analysis was performed to identify the multivariate method which not only gives predictive accuracy (as cross-validated above), but also leads to the selection of spectral wavebands more directly attributed to known absorption features: we seek a model with two salient features: accuracy and interpretability. The band interpretation was based on the known absorption features shown early in Table 1 (Chapter 1), along with the wavebands recently identified as the optimal non-redundant wavebands in vegetation and crop studies by Thenkabail et al. (2014) (see Table 10, Appendix H).

3.7. Spectral simulation of optical sensors with varying spectral resolution

This study was also interested to investigate the need for high spectral resolution for predicting grassland CCC. The field hyperspectral measurement (584 wavebands) were resampled to simulate seven (7) existing and planned optical RS missions of varying spectral resolution and spectral coverage (domain) as shown in Table 13 (Appendix K), namely (with decreasing spectral resolution): (1) EnMap, (2) HyMap, (3) CHRIS (Proba-1) land channel and (4) chlorophyll channel configurations, (5) Worldview-3 Multispectral (MS) & SWIR, (6) Sentinel-2 Multi-Spectral Instrument (MSI), and (7) Landsat-8 Operational Land Imager (OLI).

The purpose was to assess the "spectral" potential of these sensors (especially the multispectral sensors (see Table 14, Appendix K) with limited number of bands sampled in broad wavelengths) for studying grassland variables of interest. That is, the focus is on the effects of bandwidth and band placement, in isolation from other effects of radiometric resolution, spatial resolution, and atmospheric condition (Lee et al., 2004). The spectral resampling was done based on Gaussian spectral response function (SRF) using the band center and width (in full width half maximum, FWHM) (van der Meer, De Jong, & Bakker, 2001). The function is available in the 'prospectr' package (Stevens & Ramirez-Lopez, 2013) in R statistical environment (R Core Team, 2014). The bands within atmospheric vapour/water absorption region (1340-1470 nm and 1800-1970 nm) and atmospheric-purpose bands (shaded in Table 14, Appendix K) were later removed as they typically would not capture land surface reflectance with sufficiently high signal-to-noise level.

3.8. General workflow of the methodology

Figure 8 shows the overall workflow in this present study.

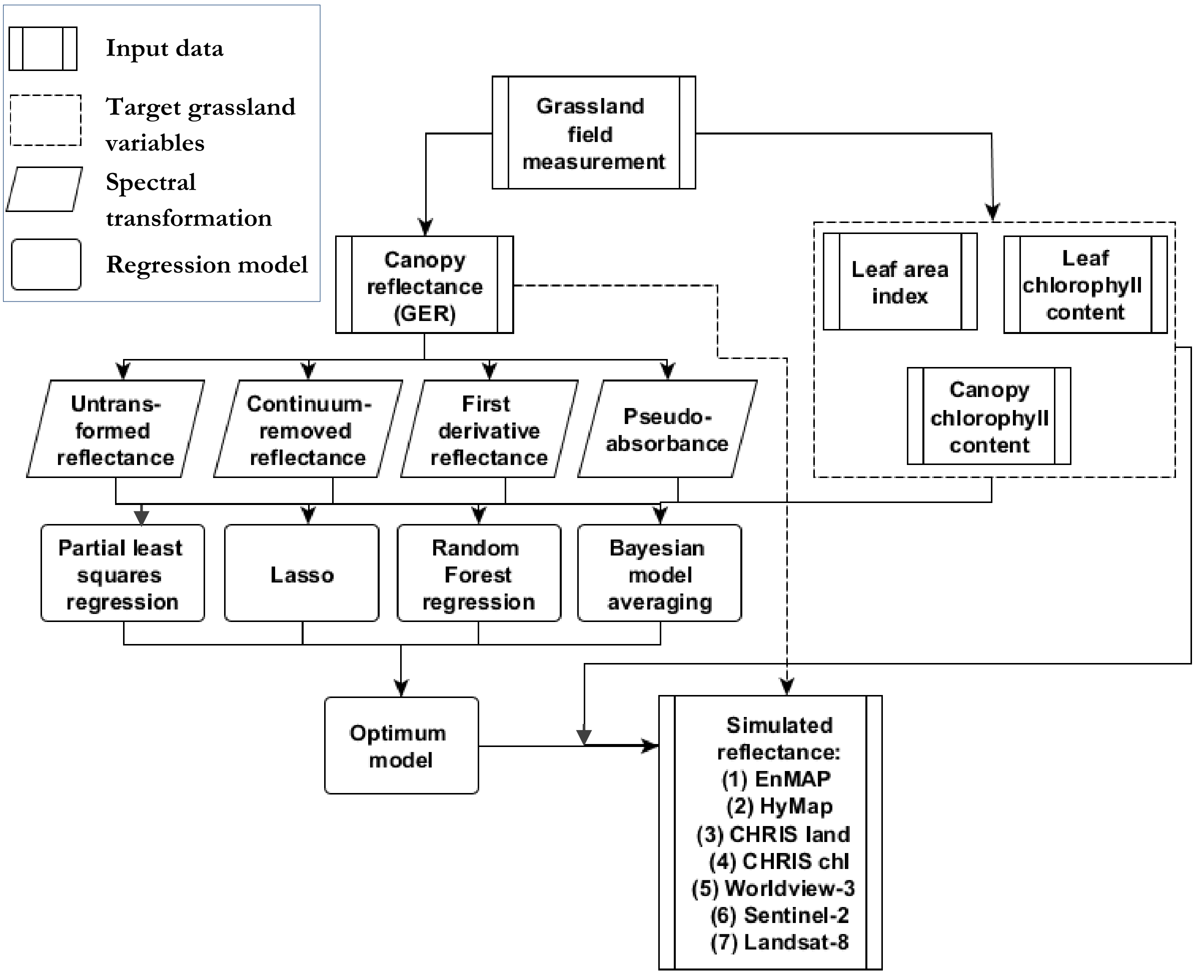


Figure 8. General workflow of the methodology. Four regression models (partial least squares regression, Lasso, Random Forest regression, and Bayesian model averaging) were developed using four input spectra (untransformed, continuum-removed, first-derivative, and pseudo-absorbance) to predict three grassland variables (leaf area index, leaf chlorophyll content, and canopy chlorophyll content (CCC=LAI x LCC)). The optimum model (highest accuracy) was subsequently used to predict grassland variables using the (spectrally) simulated seven optical sensors data.

"*In God we trust. All others bring data.*"—W. E. Deming (1900-1993)

4. RESULTS

This chapter presents the results in the following sequence: Firstly, the influence of spectral transformation on the relationship with the grassland variables was assessed, followed by an investigation on whether local continuum-removal with band depth normalization showed better potential for improvement in predictive accuracy than the full-spectrum continuum-removal. Secondly, the influence of spectral transformations on retrieval with the gold-standard model PLSR was presented. Thirdly, the performance of optimal-spectral-analysis method Lasso was compared against the benchmark PLSR model. Fourthly, all models were assessed together to identify if the two whole-spectral-analysis methods RF and BMA offered significant improvement over PLSR in terms of predictive accuracy and relevant waveband ranking, in seeking the optimum multivariate regression model. Fifthly, the wavebands selected by Lasso were interpreted. Finally, the influence of spectral resolution on the retrieval accuracy using the "optimum" models was shown.

4.1. Preliminary assessment on influence of spectral transformation

Figure 9(a) shows the variability in the (field) measured canopy reflectance from the 185 grassland plots.

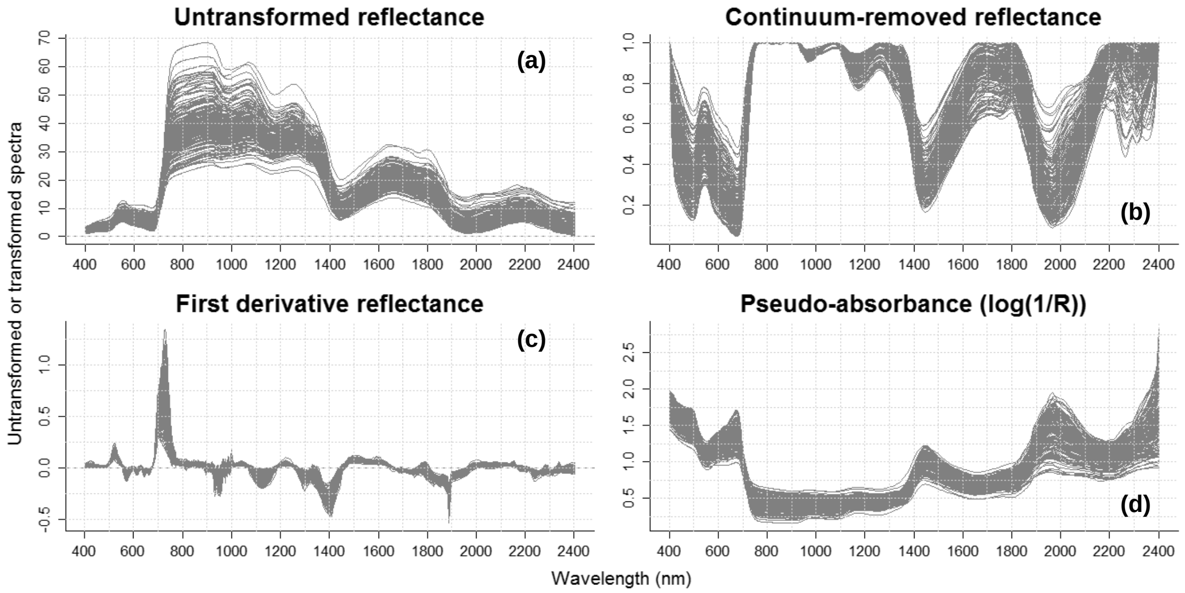


Figure 9. Field hyperspectral measurement (n=185) after smoothing (untransformed or R, a) and the three spectral transformations i.e., continuum-removal or CR (b), first derivative reflectance or FDR (c), and pseudo-absorbance or A (d).

The lowest variation can be seen in the visible (VIS) region (~400 to 700 nm), especially in blue (~450 to 495 nm). Relatively greater variation occurs in the red region (~620 to 750 nm) where chlorophyll has its absorption maxima (i.e., minimum reflectance). The low variability may partly be due to the low variation in leaf chlorophyll content (LCC) as measured (i.e., between 17.1 and 49.66 $\mu g\,cm^{-2}$, with 20% coefficient of variation from the mean; see Table 2, Chapter 3). From VIS, the reflectance varies considerably highly in the near-infrared (NIR) region (~800 to 1300 nm) as the result of light scattering by leaf internal structure. The red edge region (~680 to 730 nm) is consequently sharpened. Two water absorption regions are apparent in NIR centered around 970 nm and 1177 nm which seems to strengthen as NIR reflectance increases likely indicating higher vegetation cover (i.e., higher leaf area index or LAI). The high spectral variability in NIR likely represents, among others, the relatively high LAI variation between 0.39 and 7.34 m^2m^{-2} with 53% variation from the mean. Moving to longer wavelengths in shortwave infrared (SWIR) region (~1300 to 2400 nm) considerably high spectral variability can also be observed with two deep absorption features around 1449 nm and 1966 nm typically predominately caused by leaf water, and to a lesser degree the ligno-cellulose compounds as their absorption features have been identified in this region.

Three spectral transformations were applied to the untransformed (original) reflectance to enhance or isolate the absorption features. The influence of the three spectral transformations on the spectral variability is shown in Figure 9b, c, and d. Continuum removal (CR; Figure 9b) applied to full spectra (~400 to 2400 nm) greatly enhances the six major absorption features namely the foliar pigment (of which chlorophyll typically dominates) absorptions in VIS (blue and red region), two local water-related absorption features in NIR and the strong water absorption features in SWIR. CR suppresses the variability in the feature-less portion of NIR plateau between around 760 and 920 nm (CR value equals 1 meaning zero absorption). First derivative (Figure 9c)—the rate of change in reflectance between adjacent wavelengths—as expected shows the highest value in the red edge where a sharp increase in reflectance takes place, followed by the water-related absorption features. FDR therefore isolates these local features while, similar to CR, suppresses the variability in the remaining wavebands. Finally, the log(1/Reflectance) transformation to the so-called pseudo-absorbance (Figure 9d)—which was known to be more linearly related to absorbing compounds—depicts high absorbance in VIS (pigments) and SWIR (mostly water), with the highest variation in absorbance at the second SWIR water absorption around 1966 nm, and somewhat the upper (wavelength) end of SWIR.

As a preliminary assessment on whether the above spectral transformations potentially improve the prediction of the target grassland variables, Figure 10 shows the individual waveband Pearson's correlation coefficient (r) between the untransformed and transformed spectra and the grassland variables measured in all 185 plots (i.e., leaf area index or LAI and leaf chlorophyll content or LCC; CCC not shown as it essentially integrates both LAI and LCC i.e., CCC = LAI x LCC).

Four behaviours can be observed from the correlation plot. Firstly, comparing LCC and LAI, overall the correlation is higher for LAI (up to almost 0.8) compared to LCC (<0.6). Secondly, with regards to the spectral transformation, it can be seen that CR enhancement (in purple) gives the highest improvement in the relationship for both LCC (in NIR and SWIR) and LAI (in VIS, far NIR, and SWIR). For LAI, CR especially greatly enhances the usefulness of the pigment absorption bands in the low-light-penetration VIS (which may suggest chlorophyll and LAI covariation) and the wide water absorption features in SWIR. CR however dampens the correlation of feature-less (flat) plateau at the beginning of NIR just after the red edge.

Thirdly, FDR shows fluctuation in correlation likely due to its differentiation nature $((R_{\lambda_i} - R_{\lambda_j})/(\lambda_i - \lambda_j))$. This on one hand may in effect reduce adjacent wavebands intercorrelation, and thus reducing information redundancy in hyperspectral data, or enhance the noise in spectral measurement on the other hand. The latter is less likely as fifteen replicates of canopy spectral measurement were averaged, and subsequently smoothened by Savitzky-Golay filter beforehand. FDR improvement is noticeable for LCC in VIS, and for LAI along the red edge and local absorption features in NIR. For both LCC and LAI, FDR decreases the correlation for SWIR bands as compared to the original reflectance. Finally, pseudo-absorbance bands attain similar correlation (except bands around water absorption at 2000 nm) with LAI as the untransformed reflectance, but slightly improve the relationship between SWIR bands and LCC.

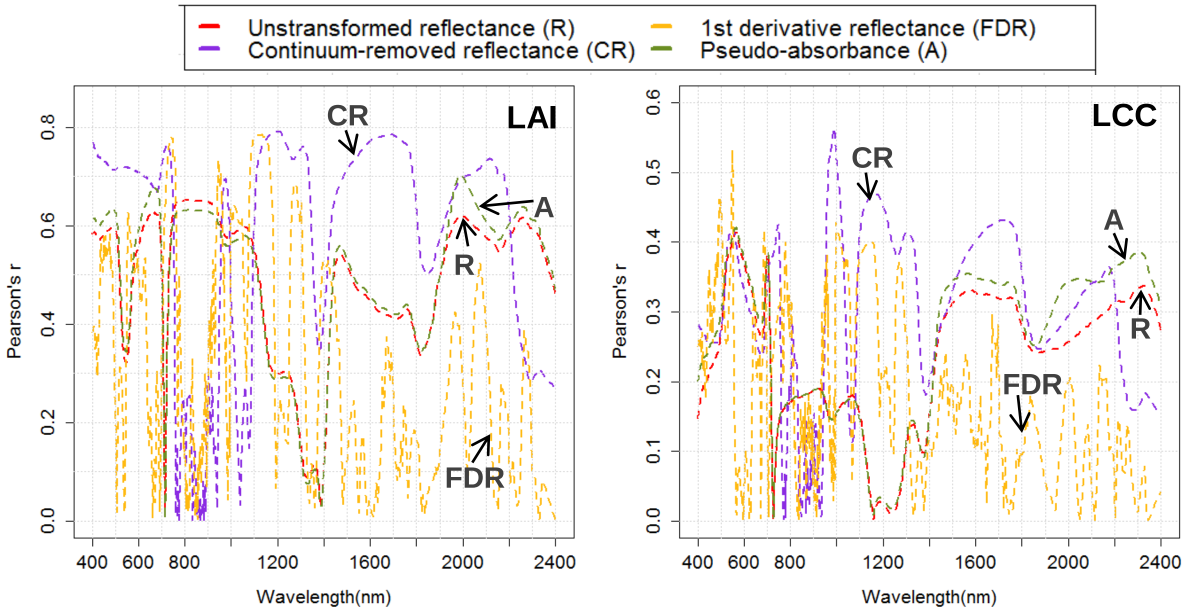
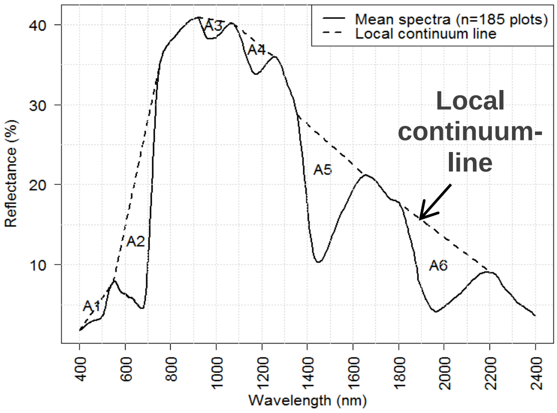


Figure 10. Absolute correlation coefficient (Pearson's r) between untransformed and transformed spectra and LAI (left) and LCC (right) for all plots (n=185). The y-axis differs (overall lower Pearson's r for LCC) to better see the r variation due to spectral transformations.

Thus, the spectral transformations, especially continuum-removal, gave some indications of potential improvement for the task of predicting LAI, LCC, and CCC which were subsequently assessed formally using the multivariate regression analysis.

4.1.1. Band depth normalization

To investigate if the widely used band depth normalization procedure proposed by Kokaly & Clark (1999) potentially improves further the predictive power of the continuum-removed reflectance, the major local absorption features were identified from the measured grass canopy reflectance. Six absorption features were isolated as shown in Figure 11, with the corresponding wavebands interval and center shown on the right. The band depth (1-CR value) values in each absorption region were then normalized by dividing them with the band depth at the respective absorption center, or with the absorption area calculated between the continuum line and the reflectance curve.



Feature	Center (nm)	Lower (nm)	Upper (nm)
Abs. 1	488.07	402.23	541.91
Abs. 2	669.4	550.49	800.27
Abs. 3	968.77	917.07	1068.51
Abs. 4	1176.9	1068.51	1254.95
Abs. 5	1448.89	1354.55	1653.98
Abs. 6	1965.62	1786.71	2205.49

Figure 11. Left: six absorption features (A$_1$-A$_6$) identified from grassland hyperspectral measurement. Band depths in each of A$_i$ were normalized by the absorption center (maximum depth) in A$_i$, or the area of A$_i$. Right: the corresponding absorption waveband center, as well as the lower waveband and upper waveband between which the local continuum-lines (dotted line) were fit to normalize each absorption feature.

Linear correlation analysis between individual band values and target grassland variables was used to compare the relationship between individual bands—with and without normalization—and the target grassland variables LAI and LCC (Figure 12). In general, no substantial improvement was observed following the normalization procedure. The un-normalized band depth values consistently (across wavebands) have higher correlation coefficient than the normalized values for LAI. For LCC, the center-depth-normalization slightly improves the correlation of the red edge wavebands (Absorption 2 in Figure 11) while absorption-area-normalization increases the correlation of pigment absorption in visible blue region (Absorption 1) and slightly the absorption 6 wavebands. Thus, overall it seemed that the predictive power of individual wavebands after local CR (with and without normalizations) was not appreciably higher than the full-spectrum CR. Consequently, the full-spectrum CR was used in the subsequent multivariate regression analysis.

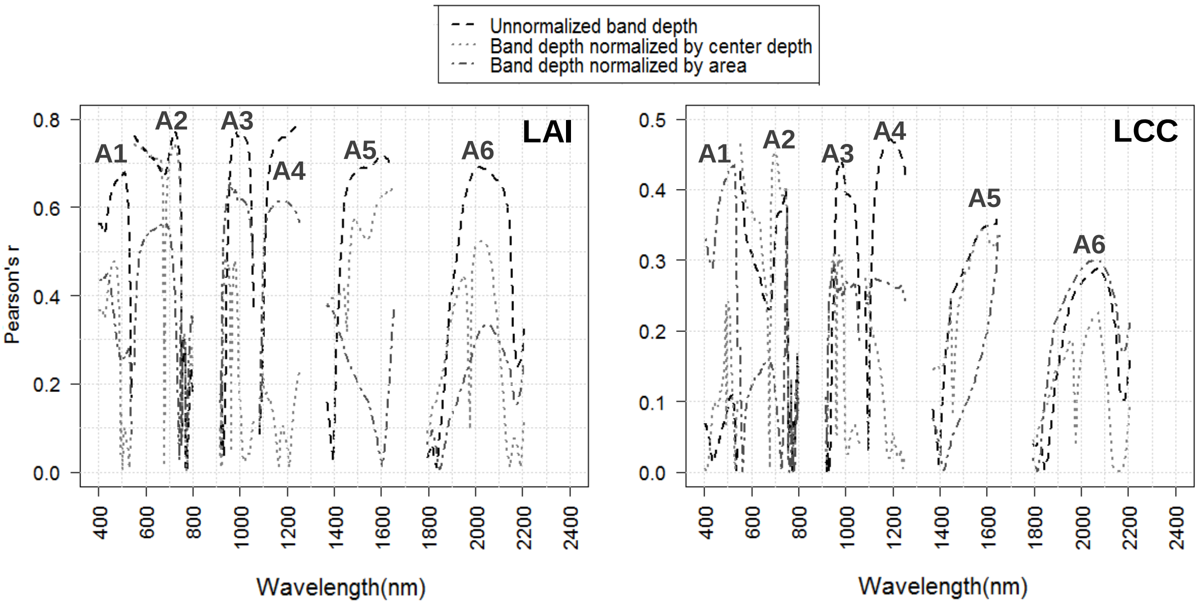


Figure 12. Correlation coefficient (r) between band depth features and grassland LAI (left); and LCC (right)

4.2. Multivariate regression analysis

Sixteen configurations of multivariate regression models (4 regression methods x 4 input spectra) were tested to predict LAI, LCC, and CCC. Table 3 summarizes the results in terms of cross-validated coefficient of determination (R^2_{cv}) and normalized (relative to mean of measured grassland variables) root mean square error ($nRMSE_{cv}$). For ease of comparison the $nRMSE_{cv}$ across the sixteen model configurations is shown in Figure 13. To understand the difference in accuracy as the result of optimally-selected (Lasso) or highly-ranked bands (PLSR), a general overview is given in Figure 14 and the key findings are interpreted along in the text. Interpretation of the precise wavelengths follows in section 4.3.

Table 3. Mean and standard deviation (in parentheses) of the 10 fold cross-validated prediction accuracy for the four multivariate regression models (1-4), to estimate LAI, LCC, and CCC. Also shown in 'parameter' is the mean and standard deviation of the optimum number of PLS components, selected bands by Lasso, optimum $mtry$ parameter of random forest, and posterior model size for BMA. Four spectral transformations (a-d) were tested for each regression model.

Regression model	*Spectral transfor-mation*	*LCC*			*LAI*			*CCC*		
		R^2_{cv}	$nRMSE_{cv}$[5] (%)	Parameter[1]	R^2_{cv}	$nRMSE_{cv}$ (%)	Parameter[1]	R^2_{cv}	$nRMSE_{cv}$ (%)	Parameter[1]
(1) PLSR	a. None	0.416^{3} (0.20)	16.2^{3} (4.5)	2.9 (0.3)	0.662^{3} (0.14)	31.1^{3} (6.5)	4.6 (0.8)	0.712 (0.09)	35.1 (6.9)	5.2 (0.9)
	b. CR	0.353 (0.25)	16.5 (4.9)	2.0 (0.0)	0.610 (0.15)	33.3 (8.3)	1.4 (0.5)	0.727 (0.08)	33.8 (5.6)	2.0 (0.0)
	c. FDR	0.385 (0.23)	16.6 (4.9)	1.0 (0.0)	0.631 (0.16)	31.9 (7.7)	2.0 (0.0)	0.684 (0.09)	36.6 (5.7)	2.1 (0.3)
	d. Abs	0.408 (0.21)	16.2 (4.5)	2.9 (0.3)	0.658 (0.16)	31.1 (6.5)	5.0 (0.5)	**0.760**^{2,3} (0.09)	**32.1**^{2,3} (5.4)	5.7 (0.5)
(2) Lasso	a. None	0.378^{4} (0.25)	16.7^{4} (4.8)	19.8 (3.1)	0.645^{4} (0.14)	31.9^{4} (7.8)	10.8 (2.8)	0.701 (0.09)	35.6 (6.9)	9.8 (2.1)
	b. CR	0.319 (0.20)	17.4 (3.8)	7.3 (2.7)	0.637 (0.15)	32.4 (7.1)	9.2 (2.0)	0.716^{4} (0.08)	34.1^{4} (6.2)	8.5 (2.8)
	c. FDR	0.299 (0.19)	18.0 (3.6)	3.2 (1.1)	0.619 (0.13)	32.9 (6.8)	4.6 (1.2)	0.645 (0.11)	37.5 (7.5)	5.6 (2.8)
	d. Abs	0.331 (0.21)	17.3 (4.0)	11.3 (2.4)	0.603 (0.19)	33.7 (8.9)	9.1 (2.3)	0.689 (0.11)	36.4 (8.4)	13.2 (2.8)
(3) RF	a. None	0.433 (0.20)	15.8 (4.4)	18.5 (8.2)	0.652 (0.16)	31.9 (8.3)	25.0 (61.5)	0.629 (0.14)	39.2 (10.3)	44.0 (82.2)
	b. CR	0.431 (0.21)	15.9 (4.4)	16.0 (8.1)	0.700 (0.13)	29.7 (6.6)	11.0 (3.9)	0.723^{4} (0.10)	33.8^{4} (7.7)	37.5 (58.4)
	c. FDR	0.488 (0.24)	15.1 (4.6)	84.0 (67.0)	**0.719**^{2,4} (0.13)	**28.9**^{2,4} **(6.4)**	66.0 (30.3)	0.701 (0.08)	35.7 (6.5)	53.0 (53.3)
	d. Abs	**0.492**^{2,4} (0.21)	**14.8**^{2,4} (4.3)	7.5 (2.6)	0.705 (0.16)	28.9 (7.1)	25.0 (61.5)	0.628 (0.14)	39.2 (10.4)	44.0 (82.2)
(4) BMA	a. None	0.383 (0.21)	16.6 (4.6)	11.3 (1.2)	0.679^{4} (0.09)	30.7^{4} (6.0)	11.1 (1.3)	0.720 (0.10)	34.2 (4.8)	11.1 (1.3)
	b. CR	0.417 (0.21)	15.8 (4.3)	5.8 (0.9)	0.658 (0.12)	31.5 (6.4)	6.7 (0.7)	0.726 (0.09)	33.5 (6.3)	4.5 (0.8)
	c. FDR	0.436^{4} (0.22)	15.7^{4} (4.5)	6.8 (1.2)	0.665 (0.11)	30.7 (5.8)	9.0 (2.0)	0.717 (0.11)	35.1 (5.6)	6.9 (0.7)
	d. Abs	0.391 (0.20)	16.2 (4.3)	12.2 (2.1)	0.633 (0.12)	32.8 (6.5)	12.2 (2.1)	0.760^{4} (0.07)	32.6^{4} (5.2)	8.2 (0.6)

[1] no. of PLS factors for PLSR; no. of selected bands for Lasso; $mtry$ i.e. no. of bands in each tree split for RF (all $ntree$=5000); posterior model size (number of covariates/bands) for BMA.

[2] Best of all (lowest $nRMSE_{cv}$, highest R^2_{cv}); [3] Best of PLSR (best input); [4] Best input spectral transformation (CR is continuum-removal, FDR is first derivative reflectance, Abs is pseudo-absorbance (log(1/reflectance)).

[5] relative to mean measured LAI (2.81 $m^2 m^{-2}$), LCC (30.07 $\mu g\ cm^{-2}$), and CCC (0.87 $g\ m^{-2}$)

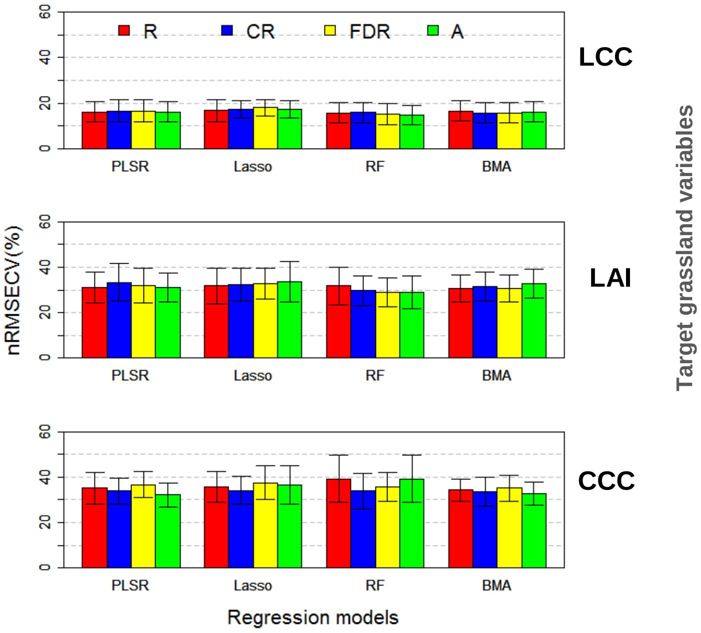


Figure 13. Mean and standard deviation (error bar) of cross-validated $nRMSE$ (as % from mean of measured grassland variables) of all regression models, for the three target grassland variables. R is untransformed reflectance, CR is continuum-removed reflectance, FDR is first derivative reflectance, and A is pseudo-absorbance.

4.2.1. Improving PLSR with spectral transformation

With regards to the spectral transformation methods, looking at the benchmark model PLSR (1a-1d in Table 3), the three transformations namely continuum-removal, first-derivative reflectance, and pseudo-absorbance did not result in improvement in accuracy for predicting LCC and LAI, but did so for predicting CCC (pseudo-absorbance). For CCC, pseudo-absorbance decreased PLSR model $nRMSE_{cv}$ (32.1%) by 3 per cent (despite not statistically significant i.e., Mann-Whitney U test p-value=0.13) and increases the R^2_{cv} (0.760) by 0.05 (p-value=0.17) i.e., the pseudo-absorbance transform allows PLSR model to explain 5 per cent more variability in the measured canopy chlorophyll content. This could indicate that the models which best predict LCC or LAI individually do not necessarily also best predict CCC. In fact, for the other three multivariate regression models it was also found that the best input spectral transformation for CCC prediction differs from the best transformation for LCC and LAI prediction (Table 3).

The lower performance of FDR in PLSR model was likely due to the fact that the transformation serves to enhance/isolate the absorption features at the expense of reducing the importance (lower correlation) of other off-absorption wavebands. This was evident from the fewer number of PLS factors (components) extracted for CR and FDR inputs (1-2 factors compared to 3-6 factors extracted from the untransformed reflectance and similarly pseudo-absorbance). This especially confirmed that including all wavebands can be more useful as the absorption features in plants are rather not exactly attributable to some precise wavebands but influence the reflectance of other wavebands.

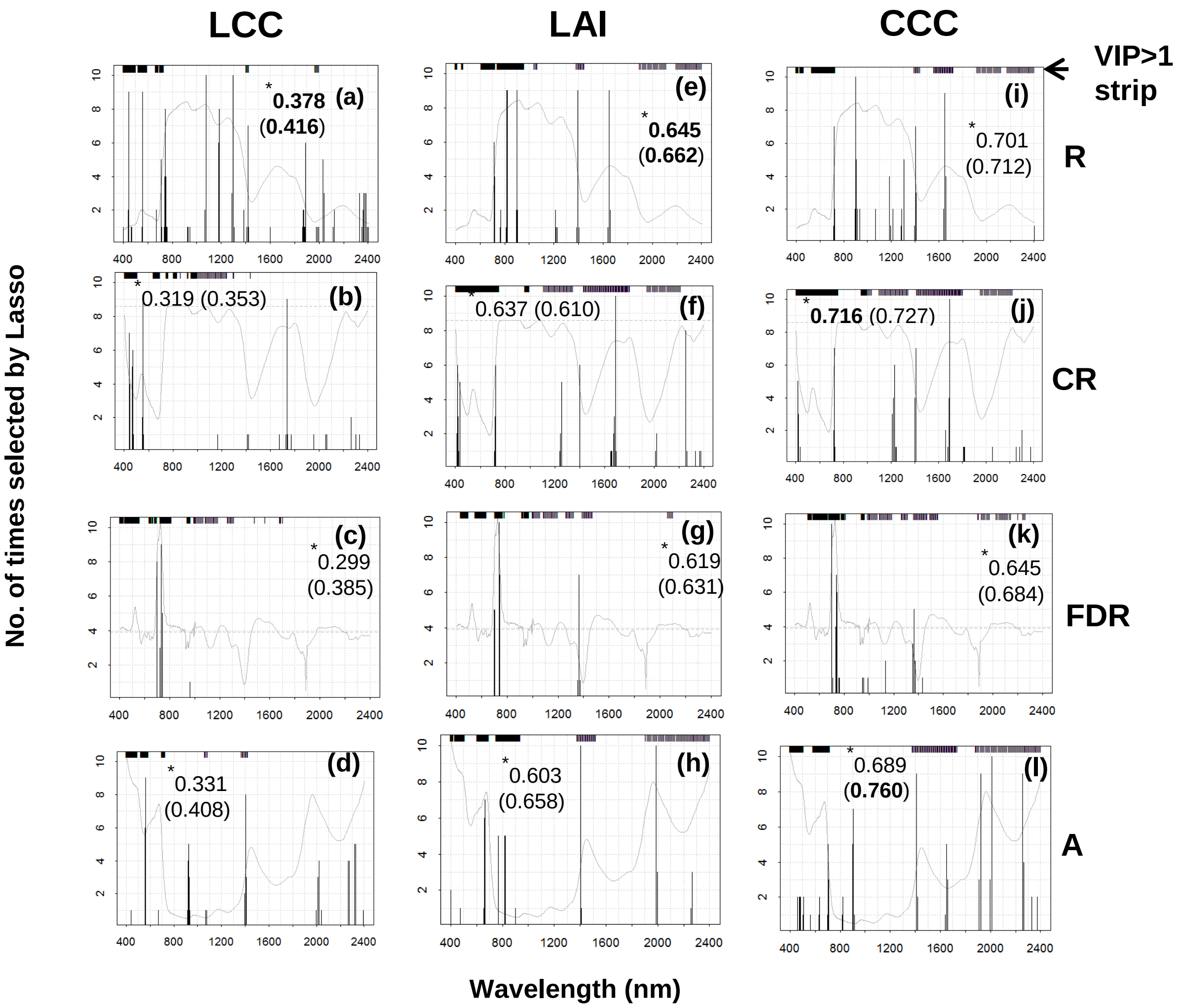


Figure 14. Wavebands (x-axis) selected by Lasso (vertical lines) and the frequency of selection (out of 10 runs of cross-validation/10-fold; shown in y-axis), along with wavebands with PLSR VIP score (cross-validation average) greater than 1 (shown as strips on top), for the prediction of leaf chlorophyll content (LCC; a-d), leaf area index (LAI; e-h), and canopy chlorophyll content (CCC; i-l). Also shown is the influence of the four spectral transformations (plotted in the background for clarity) i.e., R: untransformed reflectance (a, e, i); CR: continuum removal (b, f, j); FDR: first derivative reflectance (e, g, k); and A: absorbance (d, h, l), on the position and frequency of selected bands. * is the cross-validated coefficient of determination (R^2_{cv}; as in Table 3. In bold is the highest) of Lasso model and in parentheses, PLSR model.

Comparing the important bands (VIP>1) for the three different grassland variables of the input spectra that gave highest PLSR accuracy, for LCC prediction, Figure 14a (strip) showed the importance of visible bands, red edge bands, as well as a few NIR-SWIR edge bands (around 1400 nm) and in the second water absorption (around 2000 nm) with none in NIR plateau. The inclusion of NIR plateau bands (800-1300 nm) after CR (Figure 14b) and FDR (Figure 14c) transformation did not increase the accuracy. LAI (Figure 14e, strip) on the other hand showed less importance for VIS bands, more in red edge and the NIR plateau immediately after it, as well as along the wide SWIR absorption around 2000 nm and 2400 nm. For CCC (Figure 14l, strip, pseudo-absorbance), it seemed that a combination of important bands for LCC and LAI was identified by the PLSR model, except the NIR plateau. Compared to using untransformed reflectance (Figure 14i, strip), the inclusion of more SWIR absorption bands by pseudo-absorbance transformation somewhat increased the CCC prediction accuracy (R^2_{cv}=0.760 compared to 0.712 using untransformed reflectance).

4.2.2. Performance of optimal spectral analysis (Lasso)

The optimal spectral analysis Lasso (2a-2d in Table 3 above), as opposed to the benchmark whole spectral analysis PLSR model, performed not as good as PLSR. In particular, applying to the best input spectral transformation (i.e., untransformed reflectance for LCC (Figure 14a) and LAI (Figure 14e) prediction, CR for CCC prediction (Figure 14j)), Lasso attained 0.5 per cent higher $nRMSE_{cv}$ (16.7%), and R_{cv}^2 (0.378) lower by 0.04 (explain 4 per cent less variability in response grassland variable) for LCC; 0.8 per cent higher $nRMSE_{cv}$ (31.9%) and ~0.02 unit decrease in R_{cv}^2 (0.645) for LAI, as well as 2 per cent higher $nRMSE_{cv}$ (34.1%) and 0.04 unit decrease in R_{cv}^2 (0.716) for CCC. On one hand this indicates that selecting a subset of narrowbands deteriorates the predictive power of the information-rich hyperspectral data. However, on the other hand the sacrifice made for accuracy came with significantly reduced number of predictor narrowbands i.e., on average of the CV runs, 20 narrowbands for LCC prediction, 11 narrowbands for LAI prediction, and 9 narrowbands for CCC prediction (Table 3). For example, the red edge and the beginning of NIR plateau bands for LAI prediction (Figure 14e, strips) can seemingly be adequately represented by 2-3 bands (Figure 14e, vertical lines), and similarly for LCC prediction, 2 bands may represent the information in VIS (Figure 14a). Thus, depending on the modelling purpose and user's need (accuracy *vs* interpretability), a tradeoff should be made in choosing between whole-spectral-analysis PLSR and optimal-spectral-analysis Lasso.

Comparing Lasso-selected bands (vertical lines in Figure 14) and those with PLSR VIP greater than 1 (strips in Figure 14), it is apparent that they did not always agree to each other. Using untransformed reflectance, for LCC (Figure 14a) and CCC (Figure 14i) prediction, Lasso selected the NIR bands not considered important by PLSR, while ignoring the VIS and SWIR bands for LAI (Figure 14e) and CCC (Figure 14i) prediction. This somewhat caused a slightly lower prediction accuracy for LAI and CCC using Lasso (R_{cv}^2 decreases by <0.02 compared to PLSR).

Concerning the influence of spectral transformation on the narrowband selection by Lasso, CR and especially FDR led to a fewer number of bands being selected. For example, for LAI prediction using FDR input, on average adding more than 5 narrowbands (Table 3, Parameter; Figure 14g) did not improve the accuracy of Lasso model. Compared to using Lasso with the untransformed reflectance (selecting on average 11 narrowbands), this led to increase in $nRMSE_{cv}$ (32.9%) by 1 per cent and decrease in R_{cv}^2 (0.619) by 0.03 (or compared to best input PLSR model, 1.8 per cent higher $nRMSE_{cv}$ and 0.04 unit increase in R_{cv}^2). Thus, first derivative did not seem to be a beneficial transformation for predicting grassland LAI, LCC, and CCC using feature-selection (optimal spectral analysis) Lasso (although FDR somewhat did benefit RF and BMA (Table 3)). For CCC, selecting (on average) 6 bands from the FDR as opposed to 9 bands from the best input (CR) decreased the predictive accuracy by more i.e., 3.4 per cent higher $nRMSE_{cv}$ (37.5%) and 7 per cent less variability in measured CCC explainable by the model (R_{cv}^2=0.645). Transforming to pseudo-absorbance unnecessarily included 4 more bands than CR with discouragingly lower accuracy in predicting the CCC ($nRMSE_{cv}$=36.4%, R_{cv}^2=0.689).

Just as it changed important region (VIP>1) in PLSR model, spectral transformations also led to different bands selected by Lasso (Figure 14). In general, continuum-removal and pseudo-absorbance put more emphasize on pigments absorption in VIS and water absorptions (more so for pseudo-absorbance) in SWIR, while removing the NIR plateau, while FDR contained the selection in the steep (high slope of reflectance over wavelength) red edge region and NIR-SWIR edge (1400 nm). However, in terms of prediction accuracy, the transformation only helped Lasso to predict CCC, and slightly (R_{cv}^2 increase by 0.02; Figure 14j).

4.2.3. Best overall model: predictive accuracy

Overall, based on firstly the $nRMSE_{cv}$ and secondly R^2_{cv} values (Table 3), Random Forest regression model applied to pseudo-absorbance provided the best predictive performance for LCC (lowest $nRMSE_{cv}$ =14.8% and highest R^2_{cv} =0.492) and again for LAI but with input FDR (lowest $nRMSE_{cv}$ =28.9% and highest R^2_{cv} =0.719). However, for the canopy-integrated CCC variable, the benchmark PLSR model outperformed the rest with both lowest $nRMSE_{cv}$ (32.1%) and highest R^2_{cv} (0.760). Compared to the best input PLSR model, RF-Absorbance model reduced 1.4 per cent in $nRMSE_{cv}$ and explained 7.6 per cent more variability for LCC prediction; while RF-FDR provided 2.2 per cent lower $nRMSE_{cv}$ and 5.7 per cent more explained variability for LAI prediction.

Both in terms of $nRMSE_{cv}$ and R^2_{cv}, BMA model with best input spectra achieved the second best predictive accuracy after RF for LAI and LCC, and after PLSR (similar R^2_{cv} but a mere 0.5 per cent higher $nRMSE_{cv}$ than PLSR) for CCC. Therefore, to some extent, the ensemble regression methods (non-linear RF providing best prediction accuracy for LCC and LAI; and linear BMA achieving comparable accuracy as the best model PLSR) showed some degree of improvement over the non-ensemble models i.e. PLSR and Lasso. Another general remark can be made is that the whole-spectral-analysis methods (PLSR, RF, BMA) all outperformed the optimal-spectral-analysis method Lasso for prediction of all three grassland variables.

Comparing among regression models, input spectral transformations, and the predicted grassland variables however, there was no clear pattern on the best regression model in combination with the best input spectral transformation that can best predict all three grassland variables. The Mann-Whitney U test (Table 4) was subsequently performed to evaluate the statistical significance of the improvement in accuracy provided by the best alternative multivariate regression model configuration (for predicting LCC, LAI, or CCC), against the benchmark PLSR model. The test results showed inadequate evidence (p-value>0.05) to conclude that the improvement is statistically significant at 95% confidence level. More formally, for the task of predicting each of the grassland variables (LCC, LAI, or CCC), the test result informed us that there is no sufficient evidence to reject the null hypothesis that the prediction accuracy (assessed in terms of $nRMSE_{cv}$ (the lower, the more accurate) and R^2_{cv} (the higher, the more accurate)) of the alternative models is the same as (statistically speaking, the 10 values of R^2_{cv} and $nRMSE_{cv}$ come from the same distribution) the gold-standard model i.e., PLSR. However, as the sample size (10 values of R^2_{cv} and $nRMSE_{cv}$) is rather small, the test may actually have little power and there was considerable risk of committing type II error whereby the null hypothesis of no improvement in R^2_{cv} was falsely accepted.

Table 4. One-tailed Mann-Whitney U test applied to the coupled distribution of R^2_{cv} and $nRMSE_{cv}$ between best alternative model and the best-input benchmark model PLSR.

Measured grassland variables	*Coupled subject to one-tailed Mann-Whitney U test*	*Alternative hypothesis*	*p-value*	
			R^2_{cv}	$nRMSE_{cv}$
LCC	RF-Abs *vs* PLSR-Refl	RF-Abs has greater R^2_{cv}, and smaller $nRMSE_{cv}$ than PLSR-Refl	0.46	0.47
LAI	RF-FDR *vs* PSLR-Refl	RF-FDR has greater R^2_{cv}, and smaller $nRMSE_{cv}$ than PLSR-Refl	0.15	0.33
CCC	BMA-Abs *vs* PLSR-Abs	BMA-Abs has greater R^2_{cv}, and smaller $nRMSE_{cv}$ than PLSR-Abs	0.5	0.43

Figure 15 plots the fit between measured and predicted grassland variables using the four multivariate regression models with their respective best input spectral transformation.

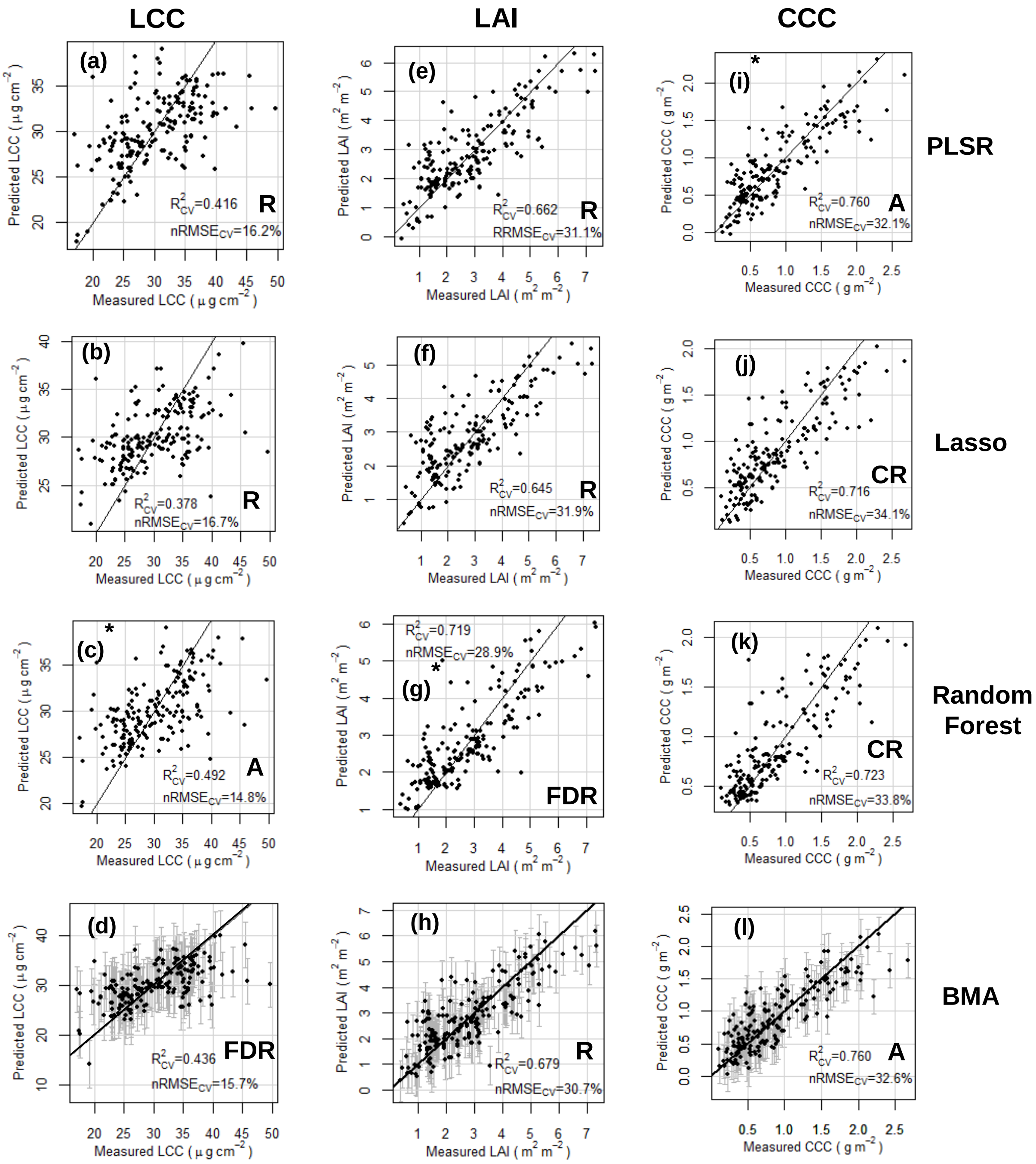


Figure 15. Measured and cross-validated predicted values of leaf chlorophyll content (LCC; a to d), leaf area index (LAI; e to h), and canopy chlorophyll content (CCC; i to l) using partial least squares regression (PLSR; a, e, i), Lasso (b, f, j), Random Forest regression (c, g, k), and Bayesian model averaging (d, h, l) applied to best input spectral transformation (i.e., R: reflectance, CR: continuum-removal, FDR: first derivative reflectance; and A: absorbance) that gave the smallest $nRMSE_{cv}$ (Table 3). * is the best-of-all model configuration. The error bar in BMA (d, h, l) shows the standard error of prediction for each observation (plot). The straight line is 1-to-1 line.

Comparing the three grassland variables, canopy chlorophyll content (CCC) could be predicted with higher accuracy (R_{cv}^2) with leaf chlorophyll content prediction being least accurately predicted. Mann-Whitney U test (Table 5) comparing the 10 R_{cv}^2 values found statistically significant difference (99% confidence) in R_{cv}^2 between the best LCC model (Figure 15c) and best LAI model (Figure 15g); and between the best LCC model (Figure 15c) and CCC (Figure 15i) model. However, the accuracy of best CCC model (Figure 15i) was not significantly different from best LAI model (Figure 15g). In general, all regression models have larger standard deviation of R_{cv}^2 when predicting LAI compared to CCC (Table 3), with LCC prediction being most unstable in this regard. This suggests the accuracy (R_{cv}^2) for predicting LAI and especially LCC varies more widely (than CCC) as the dataset was partitioned (stratified resampling) into the 10 cross-validation folds.

Table 5. One-tailed Mann-Whitney U test applied to coupled distribution of R_{cv}^2 (n=10) values between best LCC, best LAI, and best CCC models.

Alternative hypothesis (null hypothesis: R_{cv}^2 not significantly different)	p-value
R_{cv}^2 LAI > R_{cv}^2 LCC	0.001
R_{cv}^2 CCC > R_{cv}^2 LCC	0.0002
R_{cv}^2 CCC > R_{cv}^2 LAI	0.24

Best LCC model (RF-A) R_{cv}^2=0.492±0.21; Best LAI (RF-FDR) R_{cv}^2=0.719±0.13;
Best CCC (PLSR-A) R_{cv}^2=0.760±0.09

Closer inspection to the scatter of the points from the 1-to-1 line shows considerably normal distribution of the model residuals (for all LCC, LAI, CCC) suggesting linear regression model is appropriate, which is likely the reason of non-significant improvement by the non-linear nonparametric RF regression model (Table 4). No presence of extreme outliers was observed either which suggests the unsuccessful improvement in accuracy was not simply due to no robust treatment of outliers.

The Lasso fit to LAI (Figure 15f) and CCC (Figure 15j) somewhat showed more underestimation of LAI and CCC at the higher values. Selecting best subset of narrowbands and discarding the remaining bands therefore can cause some degree of saturation in relationship between grassland biophysical/biochemical variables and its canopy reflectance (in other words, including more bands may alleviate the saturation problem). RF regression model which best predicts LAI showed improvement in the fitting of low LAI values (~1.5 to 3.5 m^{-2} m^{-2}) but with somewhat greater scatter (higher residuals) in the higher LAI value range. This can also be observed for RF regression fit to CCC (Figure 15k) where a cluster of points lies close to the 1-to-1 line in the low range of CCC values. Thus, despite its accuracy, RF regression model seemed to be not well adapted to the whole range of LAI and CCC values, at least according to the dataset used in this present study. This was also indicated by the larger standard deviation of RF regression error in the cross-validation ($nRMSE_{cv}$=33.8±7.7%) than PLSR ($nRMSE_{cv}$ =32.1±5.4%) for CCC prediction (Table 3, Figure 13). Visually, PLSR model showed the best and most linear fit for CCC (Figure 15i) and LAI (Figure 15e) prediction, with equally good visual fit using BMA models for LAI (Figure 15h) and CCC (Figure 15l).

4.2.4. Best overall model: relevant waveband identification

As the LCC prediction accuracy is low and CCC strongly reflects variability in LAI and to a smaller extent the low variation of LCC (Darvishzadeh et al., 2008), the best overall model in terms of relevant wavebands identification was assessed from the wavebands identified for prediction of canopy-integrated variable CCC (Figure 16). Firstly, looking at PLSR model (Figure 16c), the top five highest ranked wavebands (based on VIP scores) were located around the center of the first water (also lignin but likely masked by water (Ramoelo et al., 2013)) absorption (~1400 nm) in SWIR, and at longer SWIR wavelength around 2250 nm (previously related to moisture and biomass). Despite the many wavebands having VIP>1 in the VIS, their relative importance was lower than the SWIR bands as also evident from their VIP ranks (above 50), despite the relatively highest ranked VIS bands closely match the known absorption center of chlorophyll (Chl a at 660 nm, Chl b at 460 nm). Therefore interpreting the most important wavebands from PLSR model is somewhat difficult, if one's interest is in understanding the most sensitive wavebands (individually, not just the region) in addition to predictive accuracy. Lasso model (Figure 16c) on the other hand intuitively selected a representative waveband in the VIS, red edge (with encouragingly very narrow shift through the cross-validation with different calibration set, suggesting selection stability), and the same water absorption around 1400 nm, despite its lower predictive accuracy (R_{cv}^2 lower by 0.04, $nRMSE_{cv}$ higher by 2% than PLSR). Nevertheless, PLSR seemed to identify the relevant region (VIP>1) and local peaks (highest VIP), and a greater contrast in VIP values (more obvious peaks) can be expected when applying PLSR to spectra resampled to lower spectral resolution to reduce the bands redundancy (as shown in section 4.4).

RF (Figure 16b) somehow put the highest importance in the local weaker water absorption in NIR plateau around 1200 nm where most (including top 5) of the highly ranked bands were located. These wavebands have 11-12 per cent OOB error which means predicting the CCC while permuting randomly the values of these wavebands would cause 11-12 per cent higher prediction error. Unlike PLSR and Lasso (also BMA), relatively high importance (9-10% OOB error, in top 20 highest ranked) was assigned for the SWIR wavebands outside the wide water absorption trough, but the peaks at 1594 and 2009 nm were rather unexplainable (unknown absorption). This suggests RF regression model somehow gave highest importance to non-relevant wavebands (Lasso applied similarly to best input CR led to more relevant bands (as will be shown in section 4.3)), which was perhaps the reason for its lower accuracy (R_{cv}^2 lower by 0.037 than PLSR).

Finally, for BMA (Figure 16a) which offered a competitive accuracy (similar R_{cv}^2) to the gold-standard method PLSR, using the same input pseudo-absorbance, at least 7 wavebands clearly show higher importance (peaks) indicated by the higher PIP values, although at lower than 0.5 i.e., they were included less than 50% of the time in the top models, likely due to multicollinearity. This occurred even after resampling the spectra to HyMap (from GER 584 bands to HyMap 119 bands) resolution to ensure the MCMC chain convergence; the posterior model size is relatively small with not more than 12 wavebands (Table 3, BMA-CCC, parameter) included in each of the top sampled (MCMC) models but the exact wavebands included differ. The high accuracy was somewhat achieved with only red edge, NIR, and NIR-SWIR edge region excluding most of the SWIR bands. Out of the 7 highest ranked wavebands however, only 3 were within 10 nm of known absorption features (red edge 716 nm, 1075 nm to biophysical and biochemical quantities, and 1420 nm to lignin). Thus, the competitive accuracy seemed not accompanied by relevant band ranking.

All in all therefore, for predicting grassland canopy chlorophyll content, the tested whole-spectral-analysis multivariate regression models not only did not provide substantial improvement in predictive accuracy over the gold-standard method PLSR, they also did not lead to relevant spectral wavebands selection and therefore to some extent suffered from the high dimensionality and multicollinearity characteristics of hyperspectral data. Additionally, looking at the standard deviation of the waveband ranking (i.e., PIP for BMA, OOB error for RF), multiple partitioning the grassland dataset in the cross-validation overall showed greater variation in importance (especially for the highly ranked bands, which may indicate instability) for BMA and RF model compared to the rather stable PLSR VIP.

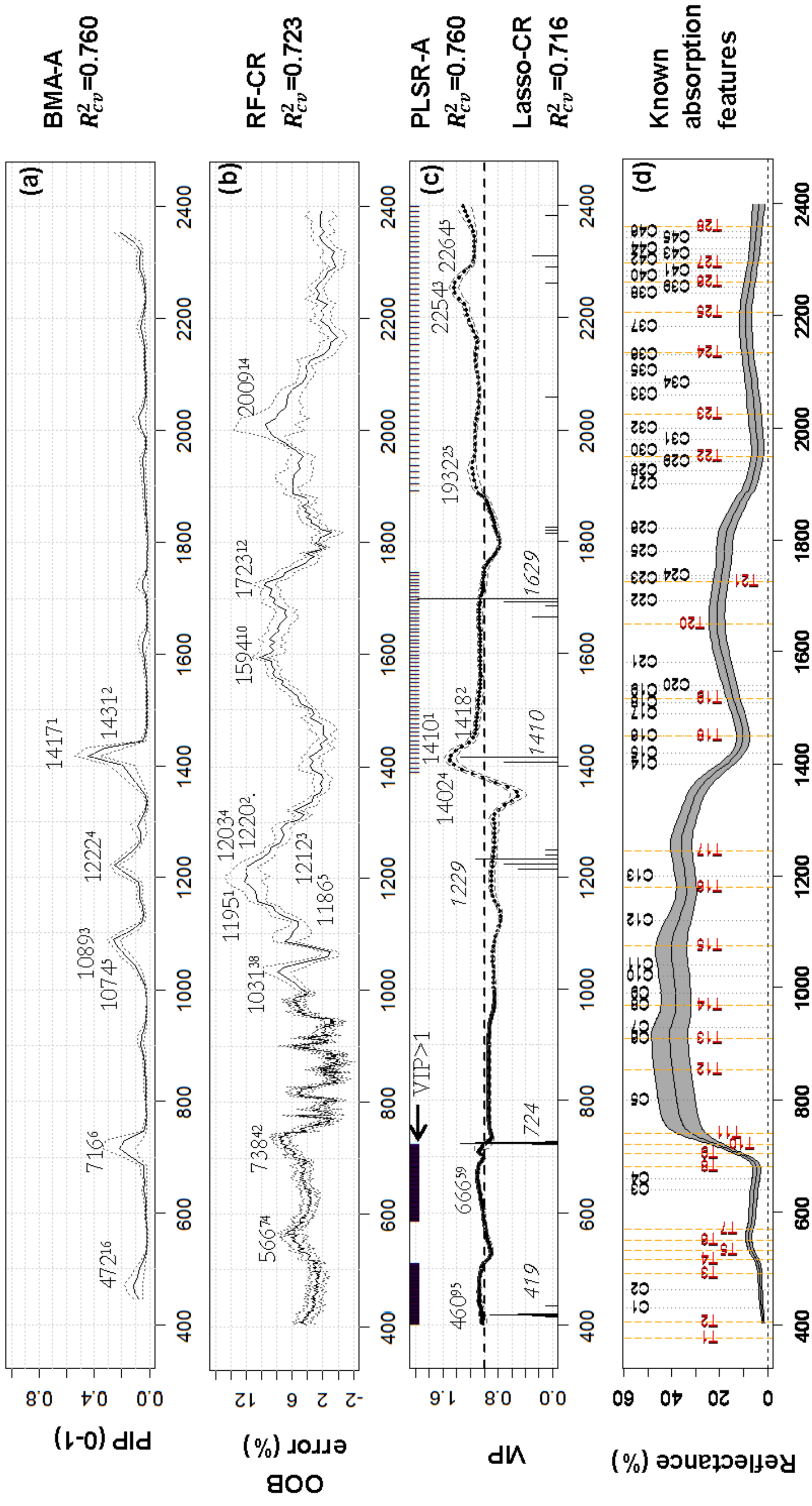


Figure 16. Band selection and ranking (cross-validated mean and 1 standard deviation) for canopy chlorophyll content (CCC) prediction using (a) Bayesian model averaging (PIP: posterior inclusion probability); (b) Random Forest regression (OOB: out-of-bag error); (c) Partial least squares regression (VIP: variable importance for projection) and Lasso (vertical line, higher means more frequently selected); in relation to the (d) known absorption features in plants (C is from Elvidge (1987), Williams & Norris (1987), Himmelsbach et al. (1988), Curran (1989), and Elvidge (1990); also Horler et al. (1983), Ben-Dor et al. (1997), and Dawson & Curran (1998)) and recently identified optimal non-redundant bands for studying vegetation (T is from Thenkabail et al. (2014)). Each regression model is shown with its best input (highest R^2_{cv} and lowest $\boldsymbol{nRMSE_{cv}}$) respectively: A=pseudo-absorbance, CR: continuum-removal. The precise wavebands (* ranking) are shown for interpretation in relation to known features. Note that BMA is applied to HyMap-resolution bands due to non-converging MCMC. See Table 1 section 2.1 and Table 10 (Appendix H) for complete information on the known absorption features.

4.3. Interpreting wavebands selected for predicting grassland variables

Table 6 lists the most frequently selected bands by Lasso using the best input (untransformed for LCC and LAI, CR for CCC).

Table 6. Wavebands selected more than 50% times (>5 out of 10 CV runs) by Lasso for predicting LCC, LAI, and CCC, using best input spectral transformation (untransformed for LCC and LAI, continuum-removal for CCC)

Optical region	***Wavelength (nm) selected***			***Known absorption features***	***Attributed to***
	LCC	***LAI***	***CCC***		
Visible blue	443.17[b]			430[1], 450[3], 435[3],	Chlorophyll-a[1,3], chlorophyll-b[3] (435)
Visible green	557.65[b]			550(±5)[2]	Chlorophyll[2]
Red edge		716.3	723.64 725.11	720(±5)[2], 700-800[1]	Stress and chlorophyll[2], nitrogen[1], protein[1]
	742.75 744.22 745.69			700-740[2], 700-800[1]	Nitrogen[1], protein[1], chlorophyll[2]
NIR		816.57 818.06[b] 819.54[b] 821.03[b]		800[1], 813[1], 855(±20)[2]	Lignin[1], tannin[1], biophysical[c] quantities and yield[2]
		901.93[b]		910[1](±5; peak NIR)[2]	Moisture[2], biomass[2], protein[1,2]
Far NIR	**1077.73[a]**			1075[2]	Biophysical and biochemical[d] quantities[2]
	1176.9 1185.71[b]			1180(±5)[2], 1120[1]	Water absorption band[2], lignin[1]
			1229.25	**Unknown**[*]	Closest is 1200[1] (water, cellulose, starch, lignin), 1215[1] (starch) 1245[2](±5; peak in 1050-1300 nm=water sensitivity[2])
	1297.08[a]			**Unknown**	**Unknown**
		1394.51[b]		1400[1]	Water[1]
Early SWIR	1418.04		1410.23	1400[1], 1420[1], 1450(±5)[2]	Water[1] (1400), lignin[1] (1420), plant moisture[2], 1450[1] (starch, sugar, water, lignin)
		1647.46[b]		1650(±5)[2]	Moisture[2]
			1692.19	1690[1]	Lignin[1], starch[1], protein[1]
Far SWIR	1887.13			**Unknown**	Closest is 1900[1] (starch)

[*] Unknown if no previously known features or not within 10 nm of the known features

[a] Always selected (10 times); [b] Selected 9 times; [c] including LAI and total chlorophyll; [d] include total chlorophyll; [1] from reference 1 (below); [2] from reference 2 (in situ); [3] from reference 3 (in situ)

Reference: **1:** Horler et al. (1983), Elvidge (1987), Williams & Norris (1987), Himmelsbach et al. (1988), Curran (1989), Elvidge (1990), Ben-Dor et al. (1997), and Dawson & Curran (1998)

2: Thenkabail et al. (2014); **3**: Pu & Gong (2011).

Except the three 'unknown' wavebands which are not associated with any previously known biochemicals (or lie outside 10 nm from known features), the other selected wavebands are within 10 nm from known absorption features of plant compounds. Lasso therefore performed appropriate waveband selection. However, only 7 out of the 21 wavebands with known absorption features agree with (within 10 nm of) the recently published optimal non-redundant hyperspectral narrowbands for studying vegetation and agricultural crops (Thenkabail et al., 2014), which may suggest grassland requires a different subset of optimal narrowbands.

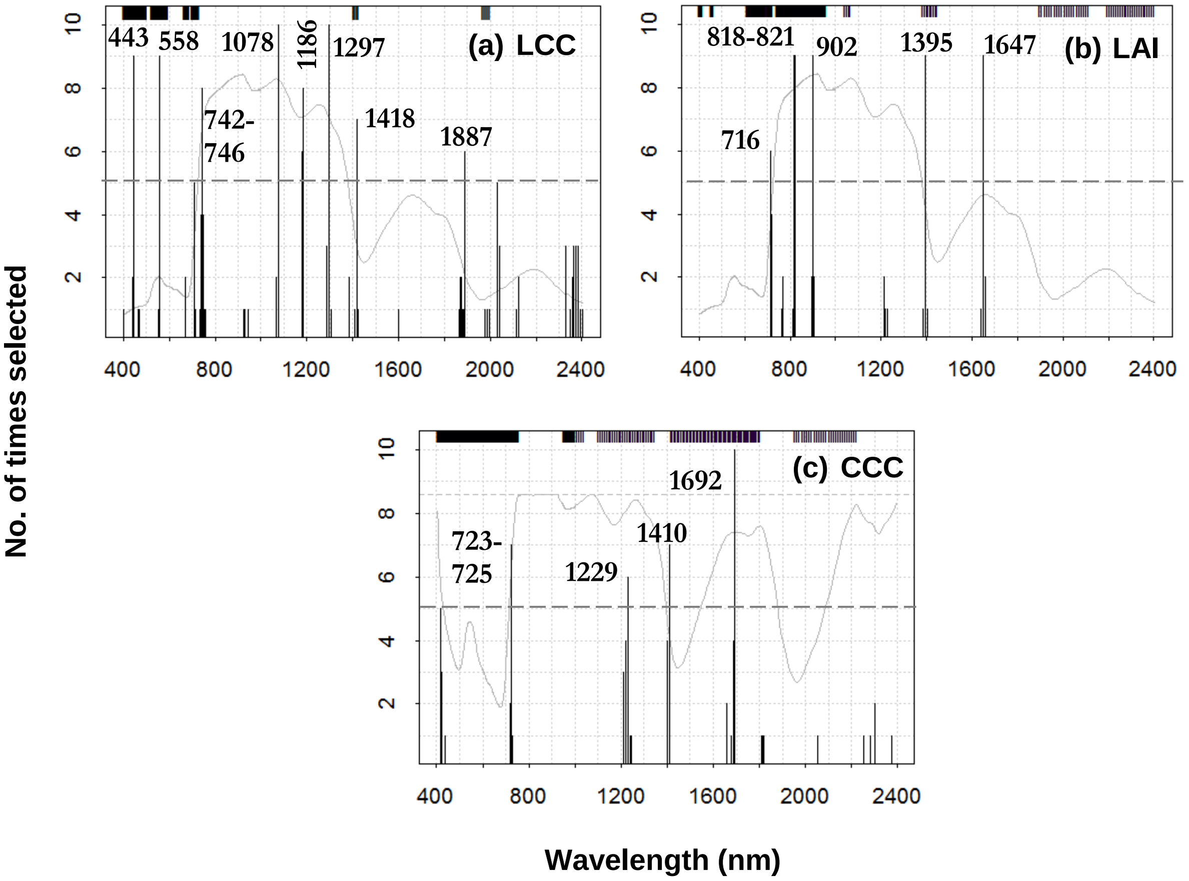


Figure 17. Wavebands selected by Lasso (best input spectral transformation: untransformed for LCC and LAI; continuum-removal for CCC) for prediction of (a) LCC, (b) LAI, and (c) CCC. Bands selected more than 5 times (out of 10 CV runs) were considered important (i.e., not species specific or simply noise) and interpreted in Table 6. On top (strip) is bands with PLSR VIP>1 as in Figure 14a,e,j.

Looking at the grassland variables separately, unexpectedly the most important wavebands for predicting leaf chlorophyll content (Table 6, Figure 17a) include not only the visible (selected 9/10 times; no chlorophyll absorption maxima in red however) and red edge region (~740 nm), but also the far NIR and SWIR wavebands associated with water and carbon-based (lignin, cellulose, protein, starch) biochemicals. In fact, the wavebands always selected in the cross-validation were in far NIR (1077.73 nm) which was widely known for their use to predict total (canopy, instead of leaf) chlorophyll (Thenkabail et al., 2014); with another far NIR waveband (1394.51) somewhat not associated with any known biochemical and likely simply affected by structural NIR scattering.

The most important wavebands (selected 90% times) for LAI prediction (Table 6, Figure 17b) were in the early NIR (as expected) around 820 nm where the reflectance variability is dominated by the leaf internal structure and close to lignin and tannin absorption centers (813 nm); and equally important, the water/moisture-related bands in far NIR (~1400 nm) and in early SWIR (~1650 nm). This indicates the usefulness of water absorption bands to predict grassland LAI as plant moisture content is likely correlated with the amount of biomass (van Wittenberghe et al., 2014). Thus it may seem that while leaf water is known to obscure/mask the signal of the other plant biochemicals of interest in other applications (e.g., nitrogen, protein, ligno-cellulose) especially in the SWIR, in this case water actually helped in predicting LAI through the plant traits covariation. This benefit of leaf water however may not be available in the case of retrievals from airborne or spaceborne hyperspectral imaging whereby the wavebands in this water absorption region can be too noisy to reflect signal from land. Aside from NIR and SWIR, the early red edge (716.3 nm) was also important for LAI prediction although not selected as frequently by Lasso. Overall, the bands selection were remarkably stable suggesting these bands are potentially useful LAI predictors across grass species.

The wavebands selected for CCC (Table 6, Figure 17c) prediction (using the respective best input CR) were located in the red edge (~720 nm), and early SWIR (~1400 nm and 1690 nm) which are sensitive to leaf water and carbon-based compounds, with none in VIS (chlorophyll absorption wavebands). The waveband centered at 1692 nm was always selected which may signify the covariation between carbon-based compounds, leaf water, and CCC. This perhaps indicates that CCC (LAI x LCC) strongly reflects variability in LAI and only to a little extent LCC (as also observed by Darvishzadeh et al., 2008). However, important wavebands in early NIR for LAI prediction were not present for CCC prediction. This could be due to the effect of continuum-removal transformation on the spectra, as in Figure 14i, similar early NIR band was selected for CCC prediction using untransformed reflectance (R_{cv}^2 lower by 0.01 than using best input CR). Nevertheless, in general the more frequently selected wavebands were close between them i.e., CCC prediction using Lasso with untransformed reflectance and CR (Figure 14i and j).

4.4. Effect of spectral resolution

To investigate the upscaling potential of the best multivariate regression model i.e., the PLSR model (as the other tested models did not seem to significantly outperform this gold standard model) in predicting grassland canopy chlorophyll content (CCC), the canopy reflectance measured with field spectroradiometer (GER, 584 narrowbands) was spectrally resampled to the existing and planned optical missions of varying spectral resolution i.e., the number of bands, the band placement and width (Figure 18, decreasing spectral resolution from a to h).

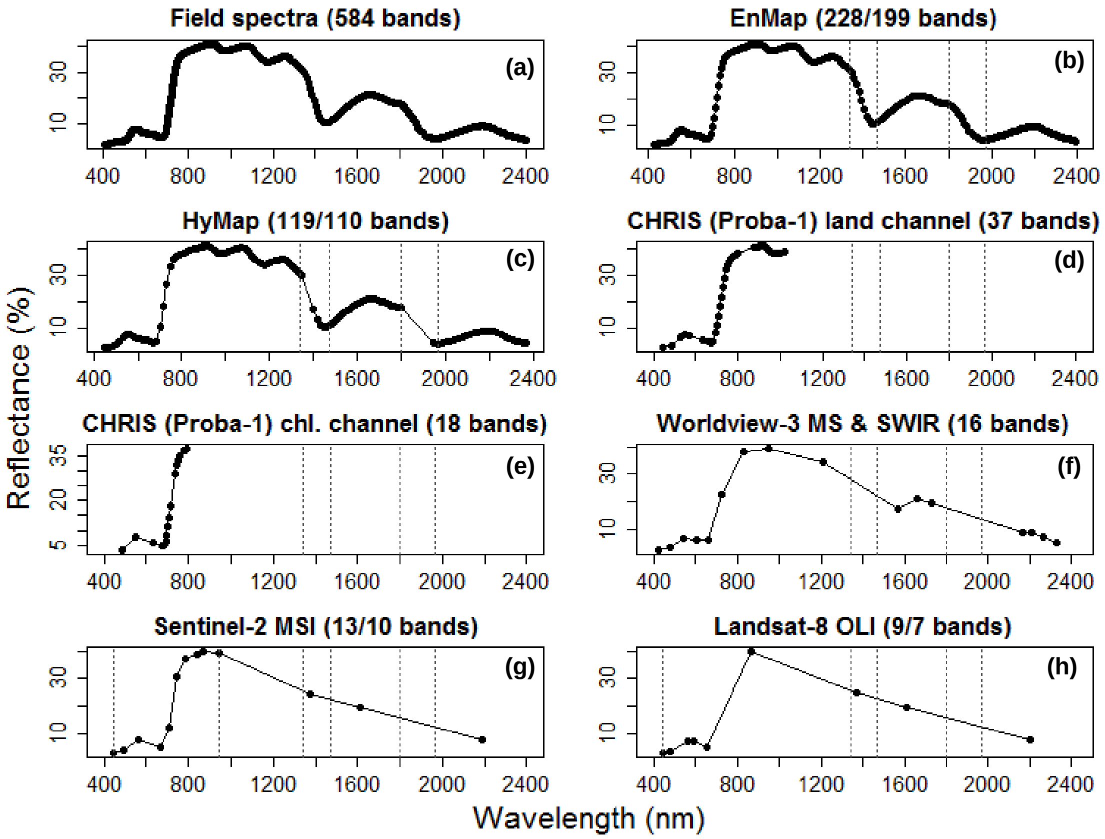


Figure 18. Field spectra (shown is average of n=185 plots) resampled (using Gaussian fit to FWHM) to existing and planned hyperspectral and multispectral optical sensors (points are the band center position). Atmospheric water absorption region around 1400 and 1900 nm (between the dotted lines) and atmospheric-purpose wavebands (at dotted line) were excluded except for original GER, EnMAP, and HyMap (with and without 'atm' bands). The fewer number of bands in parentheses is excluding atmospheric bands.

Table 7. Partial least squares regression applied to resampled/simulated spectra (untransformed reflectance). In parentheses is standard deviation in the cross-validation runs.

No.	*Sensor (# bands)*	*Factors*[1]	R^2_{cv}	$nRMSE_{cv}$ (%)
(a)	Field (GER) (584)	5.2	0.712 (0.09)	35.1 (6.9)
(b1)	EnMAP (228)	5.7	0.748 (0.09)	33.2 (6.2)
(b2)	EnMAP (no atm.[2]) (199)	5.3	0.688 (0.09)	36.3 (7.5)
(c1)	HyMap (119)	5.1	0.722 (0.09)	34.3 (6.6)
(c2)	HyMap (no atm.) (110)	5.3	0.696 (0.09)	35.5 (7.3)
(d)	CHRIS land (37)	3.5	0.653 (0.10)	37.7 (6.2)
(e)	CHRIS chl. (18)	3.3	0.660 (0.10)	37.3 (7.3)
(f)	Worldview 3 (16)	5.2	0.693 (0.10)	35.7 (6.4)
(g)	Sentinel 2 (no atm.) (10)	4.0	0.677 (0.10)	36.9 (7.0)
(h)	Landsat 8 (no atm.) (7)	2.9	0.632 (0.10)	39.0 (7.7)

[1]The average number of optimal number of PLS factors in the cross-validation

[2] no atm. : the atmospheric-affected bands were excluded

Quite surprisingly, the lower spectral resolution hyperspectral sensors namely EnMAP and HyMap provided higher CCC prediction accuracy (both in terms of lower $nRMSE_{cv}$ and higher R^2_{cv}) than the original GER spectral resolution (Table 7). That is, going from GER 584 bands to about half the number of bands, EnMAP sensor (228 bands) increased R^2_{cv} from 0.712 to 0.748 (~0.04 unit higher) and decreased $nRMSE_{cv}$ from 35.1 to 33.2 per cent (1.9% lower). This was then followed by HyMap sensor (119 bands) which still attained slightly higher accuracy than GER (R^2_{cv} higher by 0.01, $nRMSE_{cv}$ lower by 0.8%). Excluding bands in the atmospheric absorption region reduced the prediction accuracy, suggesting CCC retrieval from airborne and spaceborne hyperspectral sensors may lose the benefit of the wavebands associated with leaf water absorption which was found useful for predicting CCC in this present study. This can be seen from the high PLSR VIP (variable importance in projection) scores for the water absorption wavebands around 1400 nm as PLSR was applied to the full GER bands (Figure 19a). While quite similar, excluding the atmospheric absorption bands, it was unexpected that HyMap resolution data gave slightly higher accuracy than EnMAP. This however can partly be explained by the relatively higher VIP scores for the HyMap wavebands than EnMAP wavebands, especially in the visible and red edge domain (~400 to 740 nm) and around the second water absorption trough in SWIR. This indicates the redundantly additional number of narrowbands sampled by EnMAP or at GER spectral resolution may dilute the usefulness (lower the VIP scores) of the wavebands sensitive to variation in grassland canopy chlorophyll (i.e., fewer broader wavebands of HyMap can adequately contain the information). However, interestingly the perhaps too-broad spectral sampling in VIS by non-hyperspectral sensors (Worldview-3 in Figure 19a, Landsat-8 in Figure 19b) on the other hand somehow eliminated the VIS region importance for CCC prediction.

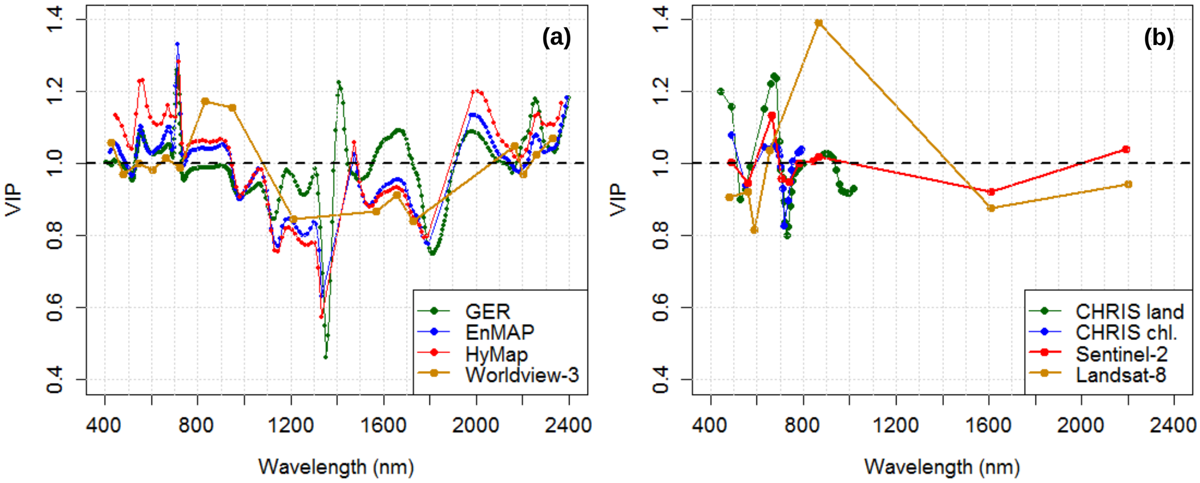


Figure 19. PLSR band importance based on mean VIP score in cross-validation for CCC prediction using reflectance of simulated sensors. Except for the original spectra (GER), the atmospheric absorption wavebands were excluded to represent airborne and spaceborne retrieval situation. Bands with VIP scores higher than 1.0 are regarded as important (Wold et al., 2001).

Simulated data of the other superspectral and multispectral sensors in general did not allow as high prediction accuracy as the hyperspectral sensors, with the exception of the unexpected lower accuracy for EnMAP (atmospheric bands excluded) than Worldview 3. In this case, Worldview-3 seemed to be most benefited by the NIR bands which, as in the VIS, its two bands may well represent the many NIR narrowbands of hyperspectral sensors (Figure 19a). Both the simulated superspectral CHRIS land (37 bands) and CHRIS chlorophyll (18 bands) channel predicted CCC with somewhat lower accuracy than Worldview-3 (16 bands) and even the multispectral Sentinel 2 (10 usable bands), likely due to the absence of SWIR bands in the CHRIS sensor settings. This confirmed the usefulness of far NIR and SWIR wavebands (see the high VIP scores of SWIR absorption wavebands beyond 2000 nm in Figure 19a) in predicting CCC, in which with only two additional SWIR bands (only one at 2200 nm seem to be important however, Figure 19b), Sentinel-2 MSI (10 bands, 3 bands not relevant for land-retrieval application were removed) could more accurately (R_{cv}^2=0.677, $nRMSE_{cv}$=36.9%) predict CCC than either CHRIS land (R_{cv}^2=0.653, $nRMSE_{cv}$=37.7%) or CHRIS chlorophyll (R_{cv}^2=0.660, $nRMSE_{cv}$=37.3%) channel which were designed with more and narrower bands but only placed in the VNIR region. Thus, it may seem that predicting grassland CCC is more benefited by lower spectral resolution (fewer and broader bands) data but with band placement across the full optical electromagnetic spectrum domain (400-2400 nm). This was also indicated by the number of PLS factors extracted for CCC prediction: the 16-band Worldview-3 contains more unique information for explaining CCC variation (5 factors, in fact similar to the hyperspectral sensors) than the CHRIS sensors (3-4 factors) which concentrate only in VNIR.

Finally, among all, it was as expected that the lowest spectral resolution sensor Landsat-8 OLI also provided the lowest accuracy (R_{cv}^2=0.632, $nRMSE_{cv}$=39.0%). As it also has similarly placed SWIR wavebands as Sentinel-2 (which predicted better than CHRIS), the likely reason for the lowest accuracy is simply the inadequate spectral sampling in the red edge and NIR region (i.e., available in CHRIS sensor design for land and chlorophyll channel) widely known for its usefulness in vegetation LAI and chlorophyll retrieval. Indeed, even when using HyMap or EnMAP hyperspectral resolution, the red edge wavebands proved to be most important for grassland CCC prediction (Figure 19a). As expected Landsat-8 showed the NDVI red and NIR bands to be most important (Figure 19b). Compared to the EnMAP sensor which best predicted CCC as shown above, Landsat-8 retrieval accuracy was lower by 5.8% in terms of $nRMSE_{cv}$ and 0.116 unit in terms of R_{cv}^2 (i.e., PLSR model could explain 11.6% less variability in CCC using the grassland reflectance simulated to Landsat-8 spectral resolution).

"*The value of interpretation is in enabling others to fruitfully think about an idea.*"—Andreas Buja

5. DISCUSSION

In the following discussion the key findings were compared with previous studies and suggestions on improving the predictive accuracy were discussed.

5.1. The influence of spectral transformation

In general, only minor improvements were obtained from the three spectral transformations tested (continuum-removal CR, first derivative FDR, and pseudo-absorbance A) in predicting either grassland biochemical (chlorophyll) or biophysical variable (leaf area index). Previously, Ullah et al. (2012) found both grassland biophysical (green biomass) and biochemical (nitrogen density, gm^{-2}) were better estimated with band depth (BD=1-CR value) analysis than original reflectance (however, with comparable accuracy to this present study i.e., R^2=0.73 for biomass (this study, best is 0.72 for LAI)). In particular, the normalized band depth features (i.e., band depth normalized by center depth BNC=BD/D_c, or by absorption area BNA=BD/D_a) gave more accurate estimation than the un-normalized band depth values. This is unlike the finding in this present study where univariate correlation analysis did not show considerable improvement by the band depth normalization procedure. This however may be due to the rather arbitrary definition (based on visual identification of absorption features in the grassland canopy reflectance) of the absorption feature wavelength intervals, instead of more automatic absorption shoulder determination based on e.g. inflection point using the DISPEC 3.2 IDL program as in Girma et al. (2013).

In another study, Ramoelo et al. (2013) predicted N:P ratio using PLSR and found that CR and water-removal performed better than FDR and original reflectance. This may suggest that spectral transformations may work better for estimation of plant biochemicals with subtler signal or narrower absorption feature than the rather broad absorption feature of chlorophyll in the visible region and water absorption in SWIR found useful to estimate LAI in this present study. Indeed, CR and FDR (or both as in first derivative of continuum-removed spectra, CRDR) have been successfully used to estimate nutrient (non-pigment) biochemicals such as nitrogen, phosphorus, calcium, potassium, sodium, and magnesium (e.g., Mutanga, Skidmore, & Prins, 2004; Ferwerda & Skidmore, 2007, Axelsson et al., 2013), or plant phenolics (Kokaly & Skidmore, 2015). Although not substantial, the higher performance of pseudo-absorbance (log(1/Reflectance)) than original reflectance for LCC and CCC prediction using the best model, respectively, RF and PLSR, seemed to be in agreement with Fourty & Baret (1998), Johnson (2001), and Serrano, Peñuelas, & Ustin (2002) as there is a linear relationship between the foliar biochemical and its contribution to the log(1/R) at the wavelength absorbed (Kumar et al., 2001).

However, our finding regarding the lack of improvement offered by spectral transformation agrees with Cho et al. (2007), studying grassland biomass in the same study area (Majella), who found no substantial improvement after FDR transformation as compared to original reflectance using PLSR with full spectrum (R^2increased by just 0.01) and in fact worse performance of CR (R^2decreased by 0.06). They did however found slight improvement by CR (R^2 increase by 0.11) using pre-selected bands. Similarly, also analyzing grassland LAI and chlorophyll, Darvishzadeh et al. (2008) found relatively higher independent test set validation accuracy of PLSR using original reflectance of pre-selected subset of wavebands. This present study however focused on using full spectrum data to identify useful wavebands to predict grassland LAI and chlorophyll. We also tested the PLSR model with internal simultaneous variable selection procedure through regularization similar to Lasso (called sparse PLSR, Chun & Keleş, 2010) and found relatively similar accuracy to full spectrum PLSR (Table 11, Appendix I).

There was also no one spectral transformation that worked best for all regression models to predict either LCC, LAI, CCC, or all of the grassland variables. As a uniform cross-validation procedure with the same dataset partitioning was carefully implemented, this behaviour was not likely caused by a pure randomness in the statistical analysis. This therefore may suggest, as the spectral transformations effectively enhance different parts of the spectrum, different transformations may be needed for different biophysical or biochemical variables, such as the finding by Ferwerda & Skidmore (2007).

5.2. Optimal spectral analysis *vs* whole spectral analysis

All the whole-spectral-analysis methods (PLSR, RF, BMA) in general outperformed optimal-spectral-analysis method Lasso. This showed that multivariate methods that select an optimum subset of narrowbands (also e.g., stepwise multiple linear regression), although can prevent loss of information in otherwise univariate methods commonly based on narrowband indices employing 2-3 narrowbands (Darvishzadeh et al., 2008), may actually still lose some information in hyperspectral data. Therefore, full spectrum (whole spectral) analysis methods capable of exploiting all wavebands and handling multicollinearity such as PLSR and RF may be the preferred approach for maximizing predictive accuracy. BMA suffered from failure of MCMC convergence when we applied full 584 bands data, and therefore was not as adaptable to high dimensional data. Other MCMC samplers could be tested.

However, there will be cases where optimal spectral analysis (bands selection) remains essential especially if the modelling interest is not only in maximizing prediction accuracy, but also interpretation of most important wavebands (i.e., the accuracy *vs* interpretability tradeoff). According to Thenkabail et al. (2014), typically 3 to 8 hyperspectral narrowbands are sufficient to attain best possible accuracy in modelling crop biophysical and biochemical variables (in this present study Lasso selected on average 11 bands for LAI and 9 bands for CCC). Therefore, future studies should continue the exploration of optimal spectral analysis methods such as regularization/shrinkage methods with other established penalty functions. For example, the possible drawback of Lasso in equally shrinking the regression coefficients of all narrowbands (thus shrinking the important wavebands as much as the unimportant ones) could suggest that the bands selection (and thus predictive accuracy) can potentially be improved by the so-called double Lasso procedure, namely relaxed Lasso and adaptive Lasso which in principle impose different/adaptive shrinkage (Zou, 2006; Meinshausen, 2007).

5.3. The utility of non-parametric (machine learning) regression algorithm and importance of model evaluation

This present study did not find strong evidence regarding improvement by non-parametric (and stochastic), non-linear Random Forest regression model over the linear parametric models (PLSR, Lasso, BMA). Coefficient of determination did increase for LAI estimation, however closer inspection on the plot of measured *vs* predicted values revealed a somewhat "localized" improvement in the fitting over the low range values of the grassland variables. This exemplifies the importance of assessing both R^2_{cv} and $nRMSE_{cv}$ as well as the visual fit for a more reliable assessment of prediction accuracy. R^2_{cv} does not measure the exact difference between measured and predicted data (Richter et al., 2012). Model evaluation is therefore just as important as model development in biophysical/biochemical variables retrieval from remote sensing and Earth observation data, and this present study adopted the nested (with independent model calibration step) 10-fold cross-validation taking advantage of the large dataset (n =185) to approximate external validation (see Figure 24, Appendix G).

The RF results suggest that the prediction errors from the linear models were not likely due to non-linear dependency (e.g., saturation of reflectance at dense canopy (Chen & Cihlar, 1996; Mutanga & Skidmore, 2004)) between grassland canopy reflectance and the grassland variables. Partly an attempt to capture the non-linearity, some studies readily transformed the reflectance into spectral indices as input to the multivariate regression analysis, for example Li et al. (2014) using PLSR, Abdel-Rahman et al. (2013) and Adam et al. (2014) using RF regression. However, Verrelst et al. (2012) demonstrated that non-linear method applied to original band can approximate more flexible relationship than when applied to spectral indices. The careful selection of narrowbands using multivariate methods can alleviate the saturation problem and thus linear model is sufficient. Investigating different fitting functions to estimate LCC and LAI, Rivera et al. (2014) concluded that the major impact on retrieval accuracy does not come from the choice of curve fitting, but rather from the choice of narrowbands or the spectral dimension.

Thus, the utility of the non-parametric machine learning regression models (Verrelst et al., 2012) recently advocated for biophysical retrievals from EO ought to be further investigated preferably through a comprehensive methodology intercomparison (linear/parametric *vs* non-linear/non-parametric) studies, especially concerning retrieval from hyperspectral data. This is especially taking into account the need for longer (even more so for hyperspectral data) training time, more computational power, and substantive expertise to implement the machine learning algorithms; all which may require sufficiently higher improvement in accuracy to be worth the extra computational burden. We also tested RF regression with spectra resampled to lower spectral resolution as in section 4.4 and found in general relatively lower accuracy than PLSR for all sensors, suggesting non-linear RF did not outperform linear PLSR either under lower dimensional settings (see Table 15, Appendix L).

5.4. Comparing retrievals for LCC, LAI, and CCC

Overall LAI and CCC could be estimated with good accuracy (best R^2_{cv}>0.5, following Richter et al. (2012)). LCC on the other hand was poorly estimated in all models. CCC and LAI were significantly better predicted than LCC. Previously, Ullah et al. (2012) also found that the canopy integrated variable nitrogen density (green biomass x nitrogen concentration) was better estimated than leaf nitrogen concentration as it was dominated by the effect of green biomass. Concerning the close correlation between CCC and LAI, Blackburn (1998) suggested to instead measure chlorophyll concentration per unit mass which is more independent from the canopy structural development and thus serves as a more useful indicator of plant physiological status as the chlorophyll concentration per mass in stressed plants will decline even when the canopy structure (i.e., LAI) is maintained.

5.5. Likely sources of prediction errors and ways to improve accuracy

The poor retrieval for LCC likely indicates the poor leaf chlorophyll signal propagation to the grassland canopy reflectance (Asner, 1998; Daughtry et al., 2000; Darvishzadeh et al., 2008). Grass leaf signal was likely interfered by reflectance variation induced by structural variability (e.g., LAI and leaf spatial arrangement i.e. leaf angle distribution, LAD), other foliar pigments present, leaf water, background signal from exposed bare soil (which can be neglected for canopies with LAI>3 (Atzberger et al., 2003), in this study plot (185) average LAI=2.81), and non-photosynthetic tissues such as standing litter; all contributing to the canopy signal. The canopy biophysical variation was also likely the reason for seemingly un-improvable accuracy when predicting LAI and CCC in this present study. Numata et al. (2007) observed that the variation in canopy structure within the field of view of a field spectrometer contributes to spectral variability of canopy reflectance even for areas with the same amount of biomass. Darvishzadeh & Skidmore (2008) demonstrated the significant influence of plant architecture on LAI estimation.

This complexity of canopy signal was likely further amplified by the heterogeneous (mixed species) nature (Yoder & Pettigrew-Crosby, 1995; Huber et al., 2008) of the studied grasslands with different species possibly having different canopy architecture. This may be indicated by the relatively high standard deviation of R_{cv}^2 suggesting the models calibrated using each of the realisation of the training set (in the cross-validation step) may not perform well when validated against the independent test set which may contain different grass species. Differences in the canopy physical structure of different species have also been found to obscure absorption features of plant nutrients (Ferwerda & Skidmore, 2007). Incorporating all bands (whole spectral analysis) helped to resolve the reflectance variability due to species differences (van Wittenberghe et al., 2014) but not completely. Additionally, the fact that CR and FDR were not able to improve the prediction accuracy may indicate that the species differences influence the canopy signal not by simply offsetting the reflectance baseline/overall brightness (thus no change in local absorption features (Asner, 1998)) but rather in a more complex manner.

Consequently, potential ways to improve LAI and chlorophyll prediction of grassland may not be offered by solely the statistical modeling part as investigated in this present study. The prediction errors were not likely due to incorrect model (functional) form or predictors (wavebands). We also evaluated the possibility of the influence of outliers using robustified PLSR (partial robust M-regression) and found no substantial improvement in accuracy either (see Table 12, Appendix J). Thus, it may seem that to improve the accuracy, the grassland canopy structural variability should be accounted for more explicitly. One way to do so is by stratifying the predictive models based on grass/herb species. Partitioning the data based on species has been found to improve prediction accuracy for biochemical and biophysical variables (Mutanga et al., 2004; Darvishzadeh et al., 2008). Another way is by accounting for integral effects of canopy structure itself such as in Knyazikhin et al. (2013) or in the practice of physically-based radiative transfer model inversion, although the suitability of the physical models to heterogeneous grassland canopy still needs to be investigated (Darvishzadeh et al., 2008b; He & Mui, 2010) and the ill-posed nature of the model inversion remains a challenge (Dorigo et al., 2007; Rivera et al., 2014).

Finally, despite great care has been exercised during the field measurement (i.e., taking several replicate measurements of canopy reflectance, LAI, and leaf chlorophyll for averaging to minimize error and their measurement under suitable weather and sun illumination condition), other sources of the prediction errors can originate from the uncertainty in the field measurements. For example, scaling up leaf chlorophyll to canopy chlorophyll by non-destructive canopy-integrated approach (Jago, Cutler, & Curran, 1999; Gitelson et al., 2005; Ciganda et al., 2009; Atzberger et al., 2010; Delegido et al., 2010). This may not be suitable if leaf chlorophyll is not uniformly distributed or if a significant amount of non-photosynthetically active components is present in the canopy (He & Mui, 2010). Uncertainty can also come from the assumption of LAI measurement with the LAI-2000 (LICOR Inc., Lincoln, NE, USA) which is that the leaves are randomly distributed (spatially) with no correction for clumping (Chen et al., 2002; Darvishzadeh et al., 2008; Rivera et al., 2014), or the indirect measurement (van Wittenberghe et al., 2014) of LCC using Markwell et al. (1995) empirical calibration function to convert the unit-less SPAD readings to the amount of leaf chlorophyll ($\mu g\ cm^{-2}$). Concerning the latter, laboratory extraction of the leaf chlorophyll to perform specific calibration is also not without uncertainty (Hu, Tanaka, & Tanaka, 2013). Thus, it can be argued that for practical consideration (e.g. time, cost, convenience, scale of application/mapping), often the rapid portable non-destructive approach remains the preferred one. Indeed, recently we see the increasing application of smartphone app-based PocketLAI (Confalonieri et al., 2013; Francone et al., 2014) and PocketN (Confalonieri et al., 2015) for ground validation as in the European FP7 project ERMES.

5.6. Useful wavebands to predict grassland variables

Despite the slightly lower accuracy, Lasso provided a useful insight on the best subset of wavebands to predict LAI and chlorophyll of the multi-species grassland. A noticeable trend from the wavebands selection analysis was the usefulness of wavebands in the far NIR (1000-1300 nm) and early SWIR (1400-1700 nm) region in the prediction of both LAI and chlorophyll. In particular, the wavebands in this region were known to be sensitive to leaf structure, water, and/or carbon-based compounds such as lignin, starch, and protein (see Table 6, section 4.3). Previously, analyzing a multi-species dataset in a field-based experiment, van Wittenberghe et al. (2014) similarly found the importance of the spectral features related to structural and carbon storage functions for both biochemical (leaf chlorophyll) and biophysical (specific leaf area i.e. ratio of fresh leaf area over dry leaf weight) variables.

For LCC prediction, although the highest importance (most frequent selection) of far NIR wavebands sensitive to leaf structure than direct chlorophyll absorption in the visible region was rather unexpected, the other frequently selected wavebands (in order of importance i.e. frequency of selection) in the pigment absorption region in VIS, water absorption in far NIR and SWIR, as well as red edge region are in good agreement with previous studies. Lasso selected green waveband instead of red as it is well known that the wavebands off of chlorophyll absorption maxima provided greater sensitivity to higher range of leaf chlorophyll (Gitelson & Merzlyak, 1994; Sims & Gamon, 2002; Blackburn, 2007). Interestingly however, the blue waveband managed to be selected 90% of the times despite the known strong chlorophyll absorption and interference from other pigments in this region (Merzlyak et al., 2003). Of next importance is the red edge wavebands (~750 nm) commonly used for chlorophyll indices (Main et al., 2011), and the water or lignin absorption waveband around 1420 nm. The latter may suggest that the chlorophyll absorption was additively driven by water absorption overtone (van Wittenberghe et al., 2014).

For LAI, the water absorption bands in far NIR (~1400 nm) and SWIR (~1600 nm) were also important, suggesting covariation between leaf area and leaf water (van Wittenberghe et al., 2014). The red edge is less involved, which was in disagreement with Lee et al. (2004) who concluded that red-edge and SWIR regions were more important that NIR for estimating LAI.

For CCC, the most important waveband was somewhat related to ligno-cellulose absorption (~1690 nm) which also had some importance for chlorophyll estimation in van Wittenberghe et al. (2014), followed by water absorption band (~1400 nm) and the red edge bands (~720 nm). The absence of chlorophyll absorption bands in VIS likely reflects the strong domination of LAI variability in CCC (Darvishzadeh et al., 2008) whereas LCC contributed relatively smaller variation to CCC.

Thus, in general, the far NIR and SWIR bands proved as useful as the more traditionally explored visible, red edge, and early NIR region for predicting LAI and chlorophyll in a heterogeneous system such as the studied multi-species grassland.

5.7. Effect of spectral resolution and upscaling the retrieval

The results of the spectral simulation revealed that the highest spectral resolution sensor field spectroradiometer GER (584 bands) did not necessarily provide the highest accuracy in predicting the grassland canopy chlorophyll content. This may indicate that CCC prediction with PLSR model (concluded as the optimum model in terms of accuracy, visual fit, and interpretability in this present study) is too some extent affected by too-high dimensionality i.e., too many bands (Chun & Keleş, 2010). Concerning the upscaling to hyperspectral sensors at airborne and spaceborne platform, too-narrow spectral sampling (at a given instantaneous field of view) may cause too-low signal-to-noise ratio especially for wavebands in SWIR region (Kruse, Boardman, & Huntington, 2002; Kaufmann et al., 2006; Rivera et al., 2014). Spaceborne sensors orbiting at an altitude of 500 km receive approximately 10000 times less radiance than an aircraft flying at 5 km. (Kumar et al., 2001). One upscaling challenge is therefore perhaps to perform optimum radiometric correction (and thus atmospheric correction) to recover the far NIR and SWIR bands proved useful for predicting grassland LAI and chlorophyll in this present study. Previously Gong et al. (2003) also found bands in SWIR as most important to predict LAI from the spaceborne hyperspectral sensor Hyperion data.

It was found that hyperspectral sensors did provide higher accuracy, with the lowest spectral resolution Landsat-8 OLI performing the worst. This was not in line with Herrmann et al. (2011) who concluded that data simulated to superspectral VENμS and Sentinel-2 can spectrally estimate LAI as good as field hyperspectral sensor. It was also found that placement of broader bands throughout the optical domain (400-2400 nm) as in Worldview-3 MS & SWIR and Sentinel-2 MSI was more advantageous than narrower bands concentrated in the visible, red edge, and early NIR excluding far NIR and SWIR bands (CHRIS). Thus, it may seem that the effect of band position is more detrimental than band width on retrieval accuracy for CCC. However, a compromise has to made between the accuracy and data cost and availability.

One common observation from the PLSR band importance (VIP) of the different sensors (Figure 19, section 4.4) was the significance of red edge bands for predicting canopy-integrated chlorophyll content (CCC), especially for low spectral resolution sensors with no or only a few SWIR bands and as atmospheric water absorption wavebands were excluded. This supports previous simulation studies (Sentinel-2) also showing the red-edge bands significance for estimating LAI and chlorophyll (Delegido et al., 2011; Clevers & Gitelson, 2013). The range of the important red edge bands was also rather narrow, which may suggest sensitivity of the retrieval accuracy to small shift in red edge band placement.

Concerning the upscaling potential, however, in reality upscaling the retrieval from field to airborne and spaceborne measurement does not only carry spectral resolution degradation, but also spatial resolution degradation (e.g., mixed pixel effect), and other confounding factors such as atmospheric noise, BRDF (viewing geometry, shadow) effects, as well as lower radiometric quality (signal-to-noise ratio). Most grassland ecosystems are characterized by discontinuous canopy (especially if grazing activity is present) and thus the signals from non-vegetation components such as exposed soil and standing litter can dominate grass canopy reflectance (Asner et al., 2000; Okin et al., 2001; Numata, 2012) and may set practical limits of retrieval using hyperspectral, or optical RS data in general, if their effects are not explicitly accounted for.

6. CONCLUSION AND RECOMMENDATION

6.1. Conclusion

This most important conclusions from this present study are as follows:

- Grassland LAI and canopy chlorophyll content could be predicted with relatively good accuracy using field hyperspectral measurement of canopy reflectance. However, relatively poor accuracy was obtained for leaf chlorophyll content indicating the poor leaf signal propagation to canopy level signal (question 1 below).

- Whole-spectral-analysis models (PLSR, RF, and BMA) which make use of all wavebands (full information, despite redundancy) performed better than optimal-spectral-analysis model (Lasso) which instead performs wavebands selection. Therefore, there was some loss of information using multivariate methods that perform wavebands selection.

- However when considering model interpretability, despite the relatively small sacrifice in accuracy, Lasso appeared to be a viable alternative as it selected relevant wavebands that could be attributed to known absorption features. The choice between the optimum models therefore depends on the tradeoff between model accuracy and interpretability.

- In general, no substantial and no significant improvement in accuracy (over the gold standard model PLSR) was provided by the other multivariate methods in combination with the spectral transformations. In terms of model interpretability, the alternative whole-spectral-analysis methods (ensemble BMA and non-linear RF) also did not identify the relevant wavebands which can be associated to known absorption features or previously published sensitive wavebands.

- For predicting grassland LAI and chlorophyll, wavebands not directly attributed to absorption features of biochemical variable of interest (e.g. chlorophyll) were useful, such as water absorption wavebands. The far NIR and SWIR region were as important as the more traditionally explored VIS, red edge, and early NIR for grassland.

- Based on the spectral simulation results, although not achieving prediction accuracy (CCC) as high as hyperspectral sensors, there seemed to be a promising upscaling potential (spectrally speaking) for optical sensors with bands placement across the full optical domain (400-2400 nm) and importantly within the relatively narrow red edge region, such as Sentinel-2 which is planned for launch in June 2015.

6.2. Summary to answers to research questions

The followings are brief answers to the research questions based on results in Chapter 4:

Question 1: To which degree grassland LAI, LCC, and CCC can be predicted from field hyperspectral measurement?

Answer: Using the most accurate model (regression model and input spectral transformation, see Question 3 below), leaf area index (LAI) and canopy chlorophyll content (CCC) could be predicted with good accuracy, namely cross-validated coefficient of determination R_{cv}^2=0.719 (cross-validated relative error $nRMSE_{cv}$=28.9% or 0.81 m^2m^{-2}) for LAI prediction; R_{cv}^2=0.760 ($nRMSE_{cv}$=32.1% or 0.28

$g\ m^{-2}$) for CCC prediction. However, leaf chlorophyll content (LCC) could be predicted with poor accuracy (R^2_{cv}=0.492; $nRMSE_{cv}$=14.8% or 4.45 $\mu g\ cm^{-2}$). This is based on Table 3 (section 4.2). To note is that the $nRMSE_{cv}$ is relative to the mean value of the measured grassland variables (i.e., 30.07 $\mu g\ cm^{-2}$ for LCC; 2.81 $m^2 m^{-2}$ for LAI; and 0.87 gm^{-2} for CCC).

Question 2: Which of the three grassland variables can be most accurately predicted?
Answer: The canopy-integrated variable CCC (which contains both structural and chlorophyll signal) could be predicted with the highest accuracy. Statistically, CCC predictive accuracy was significantly higher than LCC, but not significantly higher than LAI. This is based on Table 3 (section 4.2) and Table 5 (section 4.2.3).

Question 3: Which of the four investigated multivariate regression models (in combination with input spectral transformation) can most accurately predict LCC, LAI, and CCC?
Answer: The optimum model (highest accuracy) is non-parametric Random Forest regression with input pseudo-absorbance for LCC; Random Forest regression with input first-derivative reflectance for LAI; and partial least squares regression with input pseudo-absorbance for CCC. Compared to PLSR, the improvement in accuracy by Random Forest regression for LCC and LAI prediction however could be considered not substantial, and statistically not significant. This is based on Table 3 (section 4.2) and Table 4 (section 4.2.3).

Question 4: Which wavebands in the investigated models are characterized to predict grassland LAI, LCC, and CCC?
Answer: This is based on the interpretation of Lasso-selected wavebands (explainable wavebands i.e., within 10 nm of known absorption features (Table 1, section 2.1) or previously published as optimum wavebands (Table 10, Appendix H)).

- Wavebands useful for LCC prediction, in order of importance, were found in far NIR, visible blue and green, followed by red edge and SWIR. While the visible and red edge wavebands have previously been attributed to leaf chlorophyll, the far NIR and SWIR wavebands were related to other biochemical or biophysical variables such as water and lignin.
- For LAI prediction, most useful wavebands were located in NIR peak of structural scattering which was also associated with lignin and tannin absorptions. Of equal usefulness were water absorption bands in far NIR and SWIR. The red edge was less involved.
- For CCC, waveband in SWIR attributed to lignin, starch, protein absorption somewhat appeared as a consistent predictor, followed by water/lignin absorption band and the red edge bands. Chlorophyll absorption wavebands were not present and this may be due to more dominant contribution of LAI than LCC to variability in canopy-integrated CCC (Darvishzadeh et al., 2008).

Question 5: How is the predictive accuracy of the "optimum" model in (3) affected by varying spectral resolution using the existing and planned optical sensors?
Answer: This is based on section 4.4. In general non-hyperspectral sensors could predict CCC with relatively lower accuracy than hyperspectral sensors. However, the highest spectral resolution data did not necessarily provide the highest accuracy. This could suggest PLSR could be affected by noisy bands when spectral sampling is too narrow. Sensors which sample broader wavebands across VNIR-SWIR achieved relatively higher accuracy than sensors with narrower bands but placed solely in VIS, red edge, and early NIR. Excluding the water absorption bands that could be noisy in spaceborne measurements, and for multispectral sensors, the red edge bands proved highly important for CCC prediction and spanned a relatively narrow range, suggesting the retrieval would be sensitive to a small shift in band placement in the

red edge region. Sentinel-2 data was clearly benefited by the inclusion of red edge bands as compared to Landsat-8.

6.3. Suggestion for further studies

This present study has shown that predicting biochemical and biophysical variable of grassland remains a challenging task (advanced and non-parametric multivariate regression models could not significantly improve the accuracy) as the accuracy is likely affected by the heterogeneity-induced spectral complexity of the canopy reflectance. As such, approaches that explicitly account for grass structural differences (spatial distribution of canopy elements such as leaf angle distribution and clumping) especially for different grass species are hereby recommended. This may be achieved by:

(1) Stratifying the predictive model based on grass species. Hyperspectral data can be especially useful to detect spectral differences between grass species (Schmidt & Skidmore, 2001), and therefore prior grass species classification can be performed. However, as Darvishzadeh (2008, p. 136) noted, this influence of heterogeneity (species diversity) is relative to the spatial scale of measurement.

(2) Multisensor approach: data integration/fusion between optical (hyperspectral) and 'structural' sensors such as radar (microwave) and lidar. For example, grassland height and density can be estimated using airborne laser scanner (Straatsma & Middelkoop, 2007). Especially for estimating grassland foliar biochemical variable (e.g., chlorophyll), the canopy signal can be corrected for this structural variability beforehand and therefore biochemical signal can be better isolated in the residual signal (Knyazikhin et al., 2013).

(3) Integration of statistical and physical models. The statistically-retrieved LAI and chlorophyll can be used to parameterize (to regularize) the physical models which account for multiple structural and foliar biochemical parameters simultaneously. The three-dimensional canopy reflectance (radiative transfer) models which are designed for heterogeneous canopy such as DART (Gastellu-Etchegorry, 1996) can be investigated.

(4) Concerning the upscaling potential, as we look forward to the future spaceborne hyperspectral missions (Table 9, Appendix C), further studies are needed to evaluate LAI and chlorophyll retrieval using simulated raw spaceborne imagery data by simulating not only the spectral resolution, but also the spatial and radiometric resolution as well as the atmospheric effects on the reflectance for example using the EnMAP end-to-end-simulation tool (Segl et al., 2012).

However, we acknowledge that our above conclusions naturally apply to the grassland we studied and in particular the range of measured LAI and chlorophyll. That is, considering both statistical and physical models have their own advantages and drawbacks (section 1.1.2), further studies are still needed to investigate the usefulness of statistical inversion from hyperspectral data in other grassland ecosystem and preferably collection of samples with wider range (variation) of chlorophyll content (per leaf area) and/or concentration (per mass). The latter can be done for example by providing nutrient treatments in experimental grassland plots.

The utility of the whole-spectral-analysis methods namely the ensemble methods (RF and BMA in this present study) as well as the non-parametric machine learning regression algorithm (MLRA) models (e.g., RF) should also be investigated further with other datasets, and modifications might be needed to make the models more adaptable to the high-dimensionality and multicollinearity of hyperspectral data. This can be achieved for example by incorporating regularization/shrinkage procedure (e.g., Lasso penalty) to simultaneously perform important waveband selection within the non-linear fitting, as in penalized Gaussian process model (Yi et al., 2011). The recently developed MLRA toolbox in ARTMO (Rivera et al., in press) can be useful to more easily implement the MLRA models. Alternatively, as we found relatively small sacrifice in accuracy as compared to whole-spectral-analysis methods, wavebands selection (optimal-spectral-analysis) methods with other high-dimensional shrinkage procedures (section 5.2) can be tested, to develop interpretable models and ultimately to identify most useful wavebands to predict biophysical and biochemical variables of grassland. Finally, the multivariate statistical methods demonstrated in this present study can in principal be applied (calibrated and validated) for other study area, vegetation types, and vegetation parameters.

LIST OF REFERENCES

Abdel-Rahman, E. M., Ahmed, F. B., & Ismail, R. (2013). Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing*, *34*(2), 712–728. doi:10.1080/01431161.2012.713142

Adam, E., Mutanga, O., Abdel-Rahman, E. M., & Ismail, R. (2014). Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing*, *35*(2), 693–714. doi:10.1080/01431161.2013.870676

Asner, G. P. (1998). Biophysical and Biochemical Sources of Variability in Canopy Reflectance. *Remote Sensing of Environment*, *253*, 234–253.

Asner, G. P., Wessman, C. A., Bateson, C. A., & Privette, J. L. (2000). Impact of Tissue, Canopy, and Landscape Factors on the Hyperspectral Reflectance Variability of Arid Ecosystems. *Remote Sensing of Environment*, *74*(1), 69–84. doi:10.1016/S0034-4257(00)00124-3

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. *Computers and Electronics in Agriculture*, *73*(2), 165–173. doi:10.1016/j.compag.2010.05.006

Atzberger, C., Jarmer, T., Schlerf, M., Koetz, B., & Werner, W. (2003). Retrieval of wheat bio-physical attributes from hyperspectral data and SAILH + PROSPECT radiative transfer model. In *Proceedings of the Third EARSeL Workshop on Imaging Spectroscopy*. Herrsching, Germany.

Axelsson, C., Skidmore, A. K., Schlerf, M., Fauzi, A., & Verhoef, W. (2013). Hyperspectral analysis of mangrove foliar chemistry using PLSR and support vector regression. *International Journal of Remote Sensing*, *34*(5), 1724–1743. doi:10.1080/01431161.2012.725958

Bajwa, S. G., & Kulkarni, S. S. (2011). Hyperspectral Data Mining. In P. S. Thenkabail, J. G. Lyon, & A. Huete (Eds.), *Hyperspectral remote sensing of vegetation* (pp. 93–119). Boca Raton, FL, USA: CRC Press/Taylor and Francis Group.

Baltzer, J. L., & Thomas, S. C. (2005). Leaf optical responses to light and soil nutrient availability in temperate deciduous trees. *American Journal of Botany*, *92*(2), 214–23. doi:10.3732/ajb.92.2.214

Beeri, O., Phillips, R., Hendrickson, J., Frank, A. B., & Kronberg, S. (2007). Estimating forage quantity and quality using aerial hyperspectral imagery for northern mixed-grass prairie. *Remote Sensing of Environment*, *110*(2), 216–225. doi:10.1016/j.rse.2007.02.027

Ben-Dor, E. (1997). The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. *Remote Sensing of Environment*, *61*(1), 1–15. doi:10.1016/S0034-4257(96)00120-4

Benyamin, D. (2012). A Gentle Introduction to Random Forests, Ensembles, and Performance Metrics in a Commercial System.

Bjorn-Helge, M., Wehrens, R., & Liland, K. H. (2013). pls: Partial Least Squares and Principal Component Regression. R package version 2.4.3. Retrieved from http://cran.r-project.org/package=pls

Blackburn, G. A. (1998). Quantifying Chlorophylls and Caroteniods at Leaf and Canopy Scales: An Evaluation of Some Hyperspectral Approaches. *Remote Sensing of Environment*, *66*(3), 273–285. doi:10.1016/S0034-4257(98)00059-5

Blackburn, G. A. (2007). Hyperspectral remote sensing of plant pigments. *Journal of Experimental Botany*, *58*(4), 855–67. doi:10.1093/jxb/erl123

Blackburn, G. A. (2007). Wavelet decomposition of hyperspectral data: a novel approach to quantifying pigment concentrations in vegetation. *International Journal of Remote Sensing*, *28*(12), 2831–2855. doi:10.1080/01431160600928625

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A., & Zemp, M. (2014). The concept of Essential Climate Variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, *95*(9), 1431–1443.

Bolton, D. K., & Friedl, M. A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, *173*, 74–84. doi:10.1016/j.agrformet.2013.01.007

Boschetti, M., Colombo, R., Michele, M., Busetto, L., Panigada, C., Brivio, P. A., … Miller, J. R. (2003). Use of Semi-empirical and Radiative Transfer Model to Estimate Biophysical Parameters in a Sparse Canopy Forest, *4879*, 133–144.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi:10.1007/BF00058655

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

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Camps-Valls, G. (2009). Machine learning in remote sensing data processing. In *2009 IEEE International Workshop on Machine Learning for Signal Processing* (pp. 1–6). IEEE. doi:10.1109/MLSP.2009.5306233

Casas, A., Riaño, D., Ustin, S. L., Dennison, P., & Salas, J. (2014). Estimation of water-related biochemical and biophysical vegetation properties using multitemporal airborne hyperspectral data and its comparison to MODIS spectral response. *Remote Sensing of Environment*, *148*, 28–41. doi:10.1016/j.rse.2014.03.011

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, *11*, 2079–2107.

Chan, J. C.-W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*(6), 2999–3011. doi:10.1016/j.rse.2008.02.011

Chen, J. ., Pavlic, G., Brown, L., Cihlar, J., Leblanc, S. ., White, H. ., … Pellikka, P. K. . (2002). Derivation and validation of Canada-wide coarse-resolution leaf area index maps using high-resolution satellite imagery and ground measurements. *Remote Sensing of Environment*, *80*(1), 165–184. doi:10.1016/S0034-4257(01)00300-5

Chen, J. M., & Cihlar, J. (1996). Retrieving leaf area index of boreal conifer forests using Landsat TM images. *Remote Sensing of Environment*, *55*(2), 153–162. doi:10.1016/0034-4257(95)00195-6

Chen, J. M., Rich, P. M., Gower, S. T., Norman, J. M., & Plummer, S. (1997). Leaf area index of boreal forests: Theory, techniques, and measurements. *Journal of Geophysical Research*, *102*(D24), 29429. doi:10.1029/97JD01107

Chen, T., & Martin, E. (2009). Bayesian linear regression and variable selection for spectroscopic calibration. *Analytica Chimica Acta*, *631*(1), 13–21. doi:10.1016/j.aca.2008.10.014

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99*(6), 323–9. doi:10.1016/j.ygeno.2012.04.003

Cho, M. A., & Skidmore, A. K. (2006). A new technique for extracting the red edge position from hyperspectral data: The linear extrapolation method. *Remote Sensing of Environment*, *101*(2), 181–193. doi:10.1016/j.rse.2005.12.011

Cho, M. A., Skidmore, A. K., & Sobhan, I. (2009). Mapping beech (Fagus sylvatica L.) forest structure with airborne hyperspectral imagery. *International Journal of Applied Earth Observation and Geoinformation*, *11*(3), 201–211. doi:10.1016/j.jag.2009.01.006

Cho, M., Skidmore, A. K., Corsi, F., van Wieren, S. E., & Sobhan, I. (2007). Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *International Journal of …*. Retrieved from http://www.sciencedirect.com/science/article/pii/S030324340700013X

Chun, H., & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *72*(1), 3–25. doi:10.1111/j.1467-9868.2009.00723.x

Chung, D., Chun, H., & Keles, S. (2013). spls: Sparse Partial Least Squares (SPLS) Regression and Classification. R package version 2.2-1. Retrieved from http://cran.r-project.org/package=spls

Ciganda, V., Gitelson, A., & Schepers, J. (2009). Non-destructive determination of maize leaf and canopy chlorophyll content. *Journal of Plant Physiology*, *166*(2), 157–67. doi:10.1016/j.jplph.2008.03.004

Clark, R. N., & Roush, T. E. D. L. (1984). Reflectance Spectroscopy ' Quantitative Analysis Techniques for Remote Sensing Applications, *89*, 6329–6340.

Clevers, J. G. P. W., & Gitelson, A. A. (2013). Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. *International Journal of Applied Earth Observation and Geoinformation*, *23*, 344–351. doi:10.1016/j.jag.2012.10.008

Clevers, J. G. P. W., & Kooistra, L. (2012). Using Hyperspectral Remote Sensing Data for Retrieving Canopy Chlorophyll and Nitrogen Content. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(2), 574–583. doi:10.1109/JSTARS.2011.2176468

Colombo, R. (2003). Retrieval of leaf area index in different vegetation types using high resolution satellite data. *Remote Sensing of Environment*, *86*(1), 120–131. doi:10.1016/S0034-4257(03)00094-4

Combal, B., Baret, F., Weiss, M., Trubuil, A., Macé, D., Pragnère, A., … Wang, L. (2003). Retrieval of canopy biophysical variables from bidirectional reflectance. *Remote Sensing of Environment*, *84*(1), 1–15. doi:10.1016/S0034-4257(02)00035-4

Confalonieri, R., Foi, M., Casa, R., Aquaro, S., Tona, E., Peterle, M., … Acutis, M. (2013). Development of an app for estimating leaf area index using a smartphone. Trueness and precision determination and comparison with other indirect methods. *Computers and Electronics in Agriculture*, *96*, 67–74. doi:10.1016/j.compag.2013.04.019

Confalonieri, R., Paleari, L., Movedi, E., Pagani, V., Orlando, F., Foi, M., … Acutis, M. (2015). Improving in vivo plant nitrogen content estimates from digital images: Trueness and precision of a new approach as compared to other methods and commercial devices. *Biosystems Engineering*, *135*, 21–30. doi:10.1016/j.biosystemseng.2015.04.013

Curran, P. J. (1989). Remote sensing of foliar chemistry. *Remote Sensing of Environment*, *30*(3), 271–278. doi:10.1016/0034-4257(89)90069-2

Curran, P. J., Dungan, J. L., & Peterson, D. L. (2001). Estimating the foliar biochemical concentration of leaves with reflectance spectrometry Testing the Kokaly and Clark methodologies, *76*, 349–359.

Curran, P. J., Dungan, J. L., & Peterson, D. L. (2001). Estimating the foliar biochemical concentration of leaves with reflectance spectrometry: Testing the Kokaly and Clark methodologies. *Remote Sensing of Environment*, *76*, 349–359. doi:10.1016/S0034-4257(01)00182-1

Curran, P. J., Kupiec, J. a., & Smith, G. M. (1997). Remote sensing the biochemical composition of a slash pine canopy. *IEEE Transactions on Geoscience and Remote Sensing*, *35*(2), 415–420. doi:10.1109/36.563280

Cutler, A., Cutler, R. D., & Stevens, J. R. (2009). Tree-Based Methods. In X. Li & R. Xu (Eds.), *High-Dimensional Data Analysis in Cancer Research* (pp. 83–101). Springer Science & Business Media LLC.

Cutler, M., & Kellar-Bland, H. (2008). *CHRIS Data Format* (p. 37). Retrieved from https://earth.esa.int/c/document_library/get_file?folderId=23844&name=DLFE-592.pdf

Darvishzadeh, R. (2008). *Hyperspectral remote sensing of vegetation parameters using statistical and physical models*. University of Twente. Retrieved from http://www.itc.nl/library/papers_2008/phd/roshanak.pdf

Darvishzadeh, R., Atzberger, C., Skidmore, A., & Schlerf, M. (2011). Mapping grassland leaf area index with airborne hyperspectral imagery: A comparison study of statistical approaches and inversion of radiative transfer models. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(6), 894–906. doi:10.1016/j.isprsjprs.2011.09.013

Darvishzadeh, R., Skidmore, A., Atzberger, C., & van Wieren, S. (2008). Estimation of vegetation LAI from hyperspectral reflectance data: Effects of soil type and plant architecture. *International Journal of Applied Earth Observation and Geoinformation*, *10*(3), 358–373. doi:10.1016/j.jag.2008.02.005

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 Sensing of Environment*, *112*(5), 2592–2604. doi:10.1016/j.rse.2007.12.003

Darvishzadeh, R., Skidmore, A., Schlerf, M., Atzberger, C., Corsi, F., & Cho, M. (2008). LAI and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements. *ISPRS Journal of Photogrammetry and Remote Sensing*, *63*(4), 409–426. doi:10.1016/j.isprsjprs.2008.01.001

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). Robust statistics in data analysis — A review. *Chemometrics and Intelligent Laboratory Systems*, *85*(2), 203–219. doi:10.1016/j.chemolab.2006.06.016

Daszykowski, M., Serneels, S., Kaczmarek, K., van Espen, P., Croux, C., & Walczak, B. (2007). TOMCAT: A MATLAB toolbox for multivariate calibration techniques. *Chemometrics and Intelligent Laboratory Systems*, *85*(2), 269–277. doi:10.1016/j.chemolab.2006.03.006

Daughtry, C. S. T., Hunt, E. R., & McMurtrey, J. E. (2004). Assessing crop residue cover using shortwave infrared reflectance. *Remote Sensing of Environment*, *90*(1), 126–134. doi:10.1016/j.rse.2003.10.023

Daughtry, C. S. T., Walthall, C. L., Kim, M. S., Brown de Colstoun, E., & McMurtrey III, J. E. (2000). Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sensing of Environment*, *74*(2), 229–239. doi:10.1016/S0034-4257(00)00113-9

Davies, K. M. (2004). *Plant pigments and their manipulation. Annual plant reviews, Vol. 14.* (K. M. Davies, Ed.). Oxford, UK: Blackwell Publishing.

Dawson, T. P., & Curran, P. J. (1998, November 25). Technical note A new technique for interpolating the reflectance red edge position. Taylor & Francis Group. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/014311698214910

Dawson, T. P., North, P. R. J., Plummer, S. E., & Curran, P. J. (2003). Forest ecosystem chlorophyll content: Implications for remotely sensed estimates of net primary productivity. *International Journal of Remote Sensing*, *24*(3), 611–617. doi:10.1080/01431160304984

Delegido, J., Alonso, L., González, G., & Moreno, J. (2010). Estimating chlorophyll content of crops from hyperspectral data using a normalized area over reflectance curve (NAOC). *International Journal of Applied Earth Observation and Geoinformation*, *12*(3), 165–174. doi:10.1016/j.jag.2010.02.003

Delegido, J., Verrelst, J., Alonso, L., & Moreno, J. (2011). Evaluation of Sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content. *Sensors (Basel, Switzerland)*, *11*(7), 7063–81. doi:10.3390/s110707063

Diepen, C. A., Wolf, J., Keulen, H., & Rappoldt, C. (1989). WOFOST: a simulation model of crop production. *Soil Use and Management*, *5*(1), 16–24. doi:10.1111/j.1475-2743.1989.tb00755.x

Dorigo, W. a., Zurita-Milla, R., de Wit, a. J. W., Brazile, J., Singh, R., & Schaepman, M. E. (2007). A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International Journal of Applied Earth Observation and Geoinformation*, *9*(2), 165–193. doi:10.1016/j.jag.2006.05.003

Elvidge, C. D. (1987). Reflectance characteristics of dry plant materials. In *Proceedings of the 21st International Symposium on Remote sens. environ* (pp. 721–733). Ann Arbor, Michigan. Retrieved from http://ntrs.nasa.gov/search.jsp?R=19890023606

Elvidge, C. D. (1990). Visible and near infrared reflectance characteristics of dry plant materials. *International Journal of Remote Sensing*, *11*(10), 1775–1795. doi:10.1080/01431169008955129

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. doi:10.1198/016214501753382273

Fang, H. (2003). Retrieving leaf area index using a genetic algorithm with a canopy radiative transfer model. *Remote Sensing of Environment*, *85*(3), 257–270. doi:10.1016/S0034-4257(03)00005-1

FAO. (2008). Compendium of agricultural-environmental indicators (1989-91 to 2000). Retrieved April 24, 2008, from http://www.fao.org/es/ess/os/envi_indi/part_15.asp

Feldkircher, M., & Zeugner, S. (2009). Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging. *IMF Working Paper 09-202*.

Féret, J.-B., François, C., Gitelson, A., Asner, G. P., Barry, K. M., Panigada, C., … Jacquemoud, S. (2011). Optimizing spectral indices and chemometric analysis of leaf chemical properties using radiative transfer modeling. *Remote Sensing of Environment*, *115*(10), 2742–2750. doi:10.1016/j.rse.2011.06.016

Ferwerda, J. G., & Skidmore, A. K. (2007). Can nutrient status of four woody plant species be predicted using field spectrometry? *ISPRS Journal of Photogrammetry and Remote Sensing*, *62*(6), 406–414. doi:10.1016/j.isprsjprs.2007.07.004

Fischer, A., Kergoat, L., & Dedieu, G. (1997). Coupling satellite data with vegetation functional models: Review of different approaches and perspectives suggested by the assimilation strategy. *Remote Sensing Reviews*, *15*(1-4), 283–303. doi:10.1080/02757259709532343

Fourty, T., & Baret, F. (1998). On spectral estimates of fresh leaf biochemistry. *International Journal of Remote Sensing*, *19*(7), 1283–1297. doi:10.1080/014311698215441

Francone, C., Pagani, V., Foi, M., Cappelli, G., & Confalonieri, R. (2014). Comparison of leaf area index estimates by ceptometer and PocketLAI smart app in canopies with different structures. *Field Crops Research*, *155*, 38–41. doi:10.1016/j.fcr.2013.09.024

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/

Gastellu-Etchegorry, J. (1996). Modeling radiative transfer in heterogeneous 3-D vegetation canopies. *Remote Sensing of Environment*, *58*(2), 131–156. doi:10.1016/0034-4257(95)00253-7

George, E. I. (2010). Dilution priors : Compensating for model space redundancy, *6*, 158–165. doi:10.1214/10-IMSCOLL611

Ghasemi, J. B., & Tavakoli, H. (2013). Application of random forest regression to spectral multivariate calibration. *Analytical Methods*, *5*(7), 1863. doi:10.1039/c3ay26338j

Girma, A., Skidmore, A. K., de Bie, C. A. J. M., Bongers, F., & Schlerf, M. (2013). Photosynthetic bark: Use of chlorophyll absorption continuum index to estimate Boswellia papyrifera bark chlorophyll content. *International Journal of Applied Earth Observation and Geoinformation*, *23*, 71–80. doi:10.1016/j.jag.2012.10.013

Gitelson, A. A. (2012). Nondestructive Estimation of Foliar Pigment (Chlorophylls, Carotenoids, and Anthocyanins) Contents: Evaluating a Semianalytical Three-Band Model. In P. S. Thenkabail, J. G. Lyon, & A. Huete (Eds.), *Hyperspectral remote sensing of vegetation* (pp. 141–166). Boca Raton, FL, USA: CRC Press/Taylor and Francis Group.

Gitelson, A. A., Vina, A., Ciganda, V., Rundquist, D. C., & Arkebauer, T. J. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters*, *32*(8), L08403. doi:10.1029/2005GL022688

Gitelson, A. A., Viña, A., Verma, S. B., Rundquist, D. C., Arkebauer, T. J., Keydan, G., … Suyker, A. E. (2006). Relationship between gross primary production and chlorophyll content in crops: Implications for the synoptic monitoring of vegetation productivity. *Journal of Geophysical Research*, *111*(D8), D08S11. doi:10.1029/2005JD006017

Gitelson, A., & Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology*, *22*(3), 247–252. doi:10.1016/1011-1344(93)06963-4

Goetz, A. F. H. (2009). Three decades of hyperspectral remote sensing of the Earth: A personal view. *Remote Sensing of Environment*, *113*, S5–S16. doi:10.1016/j.rse.2007.12.014

Gong, P., Ruiliang, P., Biging, G. S., & Larrieu, M. R. (2003). Estimation of forest leaf area index using vegetation indices derived from hyperion hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(6), 1355–1362. doi:10.1109/TGRS.2003.812910

Grossman, Y. L., Ustin, S. L., Jacquemoud, S., Sanderson, E. W., Schmuck, G., & Verdebout, J. (1996). Critique of stepwise multiple linear regression for the extraction of leaf biochemistry information from leaf reflectance data. *Remote Sensing of Environment*, *56*(3), 182–193. doi:10.1016/0034-4257(95)00235-9

Guerschman, J. P., Hill, M. J., Renzullo, L. J., Barrett, D. J., Marks, A. S., & Botha, E. J. (2009). Estimating fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil in the Australian tropical savanna region upscaling the EO-1 Hyperion and MODIS sensors. *Remote Sensing of Environment*, *113*(5), 928–945. doi:10.1016/j.rse.2009.01.006

Haboudane, D. (2004). Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment*, *90*(3), 337–352. doi:10.1016/j.rse.2003.12.013

Ham, J., Crawford, M. M., & Ghosh, J. (2005). Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(3), 492–501. doi:10.1109/TGRS.2004.842481

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed., p. 731). New York: Springer.

He, Y., Guo, X., & Wilmshurst, J. (2006). Studying mixed grassland ecosystems I: suitable hyperspectral vegetation indices. *Canadian Journal of Remote Sensing*, *32*(2), 98–107. doi:10.5589/m06-009

He, Y., & Mui, A. (2010). Scaling up semi-arid grassland biochemical content from the leaf to the canopy level: challenges and opportunities. *Sensors (Basel, Switzerland)*, *10*(12), 11072–87. doi:10.3390/s101211072

Herrmann, I., Pimstein, A., Karnieli, A., Cohen, Y., Alchanatis, V., & Bonfil, D. J. (2011). LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands. *Remote Sensing of Environment*, *115*(8), 2141–2151. doi:10.1016/j.rse.2011.04.018

Himmelsbach, D. S., Boer, H., Akin, D. E., & Barton, F. E. I. (1988). Solid-state 13C NMR, FTIR, and NIRS spectroscopic studies of ruminant silage digestion. *Analytical Applications of Spectroscopy / Edited by C.S. Creaser and A.M.C. Davies*. Retrieved from http://agris.fao.org/agris-search/search.do?recordID=US201302694673

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging : A Tutorial, *14*(4), 382–417.

Homolová, L., Malenovský, Z., Clevers, J. G. P. W., García-Santos, G., & Schaepman, M. E. (2013). Review of optical-based remote sensing for plant trait mapping. *Ecological Complexity*, *15*, 1–16. doi:10.1016/j.ecocom.2013.06.003

Horler, D. N. H., Dockray, M., & Barber, J. (1983). The red edge of plant leaf reflectance. *International Journal of Remote Sensing*, *4*(2), 273–288. doi:10.1080/01431168308948546

Hu, X., Tanaka, A., & Tanaka, R. (2013). Simple extraction methods that prevent the artifactual conversion of chlorophyll to chlorophyllide during pigment isolation from leaf samples. *Plant Methods*, *9*(1), 19. doi:10.1186/1746-4811-9-19

Huang, J. F., & Blackburn, G. A. (2011). Optimizing predictive models for leaf chlorophyll concentration based on continuous wavelet analysis of hyperspectral data. *International Journal of Remote Sensing*, *32*(24), 9375–9396. doi:10.1080/01431161.2011.558130

Huang, Z., Turner, B. J., Dury, S. J., Wallis, I. R., & Foley, W. J. (2004). Estimating foliage nitrogen concentration from HYMAP data using continuum removal analysis. *Remote Sensing of Environment*, *93*(1-2), 18–29. doi:10.1016/j.rse.2004.06.008

Huber, S., Kneubühler, M., Psomas, A., Itten, K., & Zimmermann, N. E. (2008). Estimating foliar biochemistry from hyperspectral data in mixed forest canopy. *Forest Ecology and Management*, *256*(3), 491–501. doi:10.1016/j.foreco.2008.05.011

Hunt, G. R. (1980). Electromagnetic radiation: the communication link in remote sensing. In B. Siegal & A. Gillespie (Eds.), *Remote Sensing in Geology* (p. 702). New York: Wiley.

Jacquemoud, S., & Baret, F. (1990). PROSPECT: A model of leaf optical properties spectra. *Remote Sensing of Environment*, *34*(2), 75–91. doi:10.1016/0034-4257(90)90100-Z

Jacquemoud, S., Verhoef, W., Baret, F., Bacour, C., Zarco-Tejada, P. J., Asner, G. P., … Ustin, S. L. (2009). PROSPECT+SAIL models: A review of use for vegetation characterization. *Remote Sensing of Environment*, *113*, S56–S66. doi:10.1016/j.rse.2008.01.026

Jago, R. A., Cutler, M. E. J., & Curran, P. J. (1999). Estimating Canopy Chlorophyll Concentration from Field and Airborne Spectra, *4257*(98).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (p. 441). New York: Springer.

Jensen, J. R. (2007). *Remote Sensing of the Environment: An Earth Resource Perspective, 2nd ed.* Upper Saddle River, NJ: Prentice Hall.

Johnson, L. F. (2001). Nitrogen influence on fresh-leaf NIR spectra. *Remote Sensing of Environment*, *78*(3), 314–320. doi:10.1016/S0034-4257(01)00226-7

Jones, H. G., & Vaughan, R. A. (2010). *Remote sensing of vegetation: principles, techniques, and applications*. Oxford university press.

Ju, C.-H., Tian, Y.-C., Yao, X., Cao, W.-X., Zhu, Y., & Hannaway, D. (2010). Estimating Leaf Chlorophyll Content Using Red Edge Parameters. *Pedosphere*, *20*(5), 633–644. doi:10.1016/S1002-0160(10)60053-7

Kaufmann, H., Segl, K., Chabrillat, S., Hofer, S., Stuffler, T., Mueller, A., … Bach, H. (2006). EnMAP A Hyperspectral Sensor for Environmental Mapping and Analysis. In *2006 IEEE International Symposium on Geoscience and Remote Sensing* (pp. 1617–1619). IEEE. doi:10.1109/IGARSS.2006.417

Knyazikhin, Y., Schull, M. A., Stenberg, P., Mõttus, M., Rautiainen, M., Yang, Y., … Myneni, R. B. (2013). Hyperspectral remote sensing of foliar nitrogen content. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(3), E185–92. doi:10.1073/pnas.1210196109

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *5*.

Kokaly, R. F., Asner, G. P., Ollinger, S. V., Martin, M. E., & Wessman, C. a. (2009). Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote Sensing of Environment*, *113*, S78–S91. doi:10.1016/j.rse.2008.10.018

Kokaly, R. F., & Clark, R. N. (1999). Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. *Remote Sensing of Environment*, *67*(98), 267–287. doi:10.1016/S0034-4257(98)00084-4

Kokaly, R. F., & Skidmore, A. K. (2015). Plant phenolics and absorption features in vegetation reflectance spectra near 1.66μm. *International Journal of Applied Earth Observation and Geoinformation*, 1–29. doi:10.1016/j.jag.2015.01.010

Kopačková, V. (2012). Utilization of hyperspectral image optical indices to assess the Norway spruce forest health status. *Journal of Applied Remote Sensing*, *6*(1), 063545. doi:10.1117/1.JRS.6.063545

Kruse, F. A., Boardman, J. W., & Huntington, J. F. (2002). Comparison of EO-1 Hyperion and airborne hyperspectral remote sensing data for geologic applications. In *Proceedings, IEEE Aerospace Conference* (Vol. 3, pp. 3–1501–3–1513). IEEE. doi:10.1109/AERO.2002.1035288

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (p. 600). New York: Springer.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Scrucca, L. (2015). caret: Classification and Regression Training. Retrieved from http://cran.r-project.org/package=caret

Kumar, L., Schmidt, K., Dury, S., & Skidmore, A. (2001). Imaging Spectrometry and Vegetation Science. In F. van der Meer & S. M. De Jong (Eds.), *Imaging Spectroscopy: Basic Principles and Prospective Applications* (1st ed., pp. 111–156). Springer Science & Business Media Netherlands.

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods.

Larigauderie, A., & Mooney, H. A. (2010). The Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services: moving a step closer to an IPCC-like mechanism for biodiversity. *Current Opinion in Environmental Sustainability*, *2*(1-2), 9–14. doi:10.1016/j.cosust.2010.02.006

Lawrence, R. L., Wood, S. D., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*, *100*(3), 356–362. doi:10.1016/j.rse.2005.10.014

Le Maire, G., François, C., & Dufrêne, E. (2004). Towards universal broad leaf chlorophyll indices using PROSPECT simulated database and hyperspectral reflectance measurements. *Remote Sensing of Environment*, *89*(1), 1–28. doi:10.1016/j.rse.2003.09.004

Lee, K.-S., Cohen, W. B., Kennedy, R. E., Maiersperger, T. K., & Gower, S. T. (2004). Hyperspectral versus multispectral data for estimating leaf area index in four different biomes. *Remote Sensing of Environment*, *91*(3-4), 508–520. doi:10.1016/j.rse.2004.04.010

Lemaire, G., Francois, C., Soudani, K., Berveiller, D., Pontailler, J., Breda, N., … Dufrene, E. (2008). Calibration and validation of hyperspectral indices for the estimation of broadleaved forest leaf chlorophyll content, leaf mass per area, leaf area index and leaf canopy biomass. *Remote Sensing of Environment*, *112*(10), 3846–3864. doi:10.1016/j.rse.2008.06.005

Li, X., Zhang, Y., Bao, Y., Luo, J., Jin, X., Xu, X., … Yang, G. (2014). Exploring the Best Hyperspectral Features for LAI Estimation Using Partial Least Squares Regression. *Remote Sensing*, *6*(7), 6221–6241. doi:10.3390/rs6076221

Liang, S. (2007). Recent developments in estimating land surface biogeophysical variables from optical remote sensing. *Progress in Physical Geography*, *31*(5), 501–516. doi:10.1177/0309133307084626

Madigan, D., & York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, *63*, 215–232. Retrieved from http://www.jstor.org/stable/1403615?seq=1#page_scan_tab_contents

Main, R., Cho, M. A., Mathieu, R., O'Kennedy, M. M., Ramoelo, A., & Koch, S. (2011). An investigation into robust spectral indices for leaf chlorophyll estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(6), 751–761. doi:10.1016/j.isprsjprs.2011.08.001

Majeke, B., van Aardt, J., & Cho, M. (2008). Imaging spectroscopy of foliar biochemistry in forestry environments. *Southern Forests: A Journal of Forest Science*, *70*(3), 275–285. doi:10.2989/SF.2008.70.3.11.672

Mallick, H., & Yi, N. (2013). Bayesian Methods for High Dimensional Linear Models. *Journal of Biometrics & Biostatistics*, *1*, 005. doi:10.4172/2155-6180.S1-005

Markwell, J., Osterman, J. C., & Mitchell, J. L. (1995). Calibration of the Minolta SPAD-502 leaf chlorophyll meter. *Photosynthesis Research*, *46*(3), 467–72. doi:10.1007/BF00032301

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis*, *52*(1), 374–393. doi:10.1016/j.csda.2006.12.019

Merzlyak, M. N., Gitelson, A. A., Chivkunova, O. B., Solovchenko, A. E., & Pogosyan, S. I. (2003). Application of Reflectance Spectroscopy for Analysis of Higher Plant Pigments. *Russian Journal of Plant Physiology*, *50*(5), 704–710. doi:10.1023/A:1025608728405

Mirzaie, M., Darvishzadeh, R., Shakiba, a., Matkan, a. a., Atzberger, C., & Skidmore, a. (2014). Comparative analysis of different uni- and multi-variate methods for estimation of vegetation water content using hyper-spectral measurements. *International Journal of Applied Earth Observation and Geoinformation*, *26*, 1–11. doi:10.1016/j.jag.2013.04.004

Moran, J. A., Mitchell, A. K., Goodmanson, G., & Stockburger, K. A. (2000). Differentiation among effects of nitrogen fertilization treatments on conifer seedlings by foliar reflectance: a comparison of methods. *Tree Physiology*, *20*(16), 1113–1120. doi:10.1093/treephys/20.16.1113

Mutanga, O., Adam, E., & Cho, M. A. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation*, *18*, 399–406. doi:10.1016/j.jag.2012.03.012

Mutanga, O., Prins, H. H. T., Skidmore, A. K., Wieren, S., Huizing, H., Grant, R., … Biggs, H. (2004). Explaining grass-nutrient patterns in a savanna rangeland of southern Africa. *Journal of Biogeography*, *31*(5), 819–829. doi:10.1111/j.1365-2699.2004.01072.x

Mutanga, O., Skidmore, A. ., & Prins, H. H. . (2004). Predicting in situ pasture quality in the Kruger National Park, South Africa, using continuum-removed absorption features. *Remote Sensing of Environment*, *89*(3), 393–408. doi:10.1016/j.rse.2003.11.001

Mutanga, O., & Skidmore, A. K. (2003). Continuum-removed absorption features estimate tropical savanna grass quality in situ. In *EARSEL WORKSHOP ON IMAGING SPECTROSCOPY* (pp. 13–16).

Mutanga, O., & Skidmore, A. K. (2004). Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing*, *25*(19), 3999–4014. doi:10.1080/01431160310001654923

Navarro-Cerrillo, R. M., Trujillo, J., de la Orden, M. S., & Hernández-Clemente, R. (2014). Hyperspectral and multispectral satellite sensors for mapping chlorophyll content in a Mediterranean Pinus sylvestris L. plantation. *International Journal of Applied Earth Observation and Geoinformation*, *26*, 88–96. doi:10.1016/j.jag.2013.06.001

Nevius, T. A., & Pardue, H. L. (1984). Development and preliminary evaluation of modified Savitzky-Golay smoothing functions. *Analytical Chemistry*, *56*(12), 2249–2251. doi:10.1021/ac00276a061

Numata, I. (2012). Characterization on Pastures Using Field and Imaging Spectrometers. In P. S. Thenkabail, J. G. Lyon, & A. Huete (Eds.), *Hyperspectral remote sensing of vegetation* (pp. 207–225). Boca Raton, FL, USA: CRC Press/Taylor and Francis Group.

Numata, I., Roberts, D. A., Chadwick, O. A., Schimel, J., Sampaio, F. R., Leonidas, F. C., & Soares, J. V. (2007). Characterization of pasture biophysical properties and the impact of grazing intensity using remotely sensed data. *Remote Sensing of Environment*, *109*(3), 314–327. doi:10.1016/j.rse.2007.01.013

Okin, G. S., Roberts, D. A., Murray, B., & Okin, W. J. (2001). Practical limits on hyperspectral vegetation discrimination in arid and semiarid environments. *Remote Sensing of Environment*, *77*(2), 212–225. doi:10.1016/S0034-4257(01)00207-3

Ortenberg, F. (2011). Hyperspectral Sensor Characteristics: Airborne, Spaceborne, Hand-Held, and Truck-Mounted; Integration of Hyperspectral Data with LIDAR. In P. S. Thenkabail, J. G. Lyon, & A. Huete (Eds.), *Hyperspectral Remote Sensing of Vegetation* (pp. 39–68). Boca Raton, FL, USA: CRC Press/Taylor and Francis Group.

Peng, Y., Gitelson, A. A., Keydan, G., Rundquist, D. C., & Moses, W. (2011). Remote estimation of gross primary production in maize and support for a new paradigm based on total crop chlorophyll content. *Remote Sensing of Environment*, *115*(4), 978–989. doi:10.1016/j.rse.2010.12.001

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., … Wegmann, M. (2013). Essential biodiversity variables. *Science*, *339*(6117), 277–278.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, *9*(2), 181–199. doi:10.1007/s10021-005-0054-1

Pu, R. (2004). Wavelet transform applied to EO-1 hyperspectral data for forest LAI and crown closure mapping. *Remote Sensing of Environment*, *91*(2), 212–224. doi:10.1016/j.rse.2004.03.006

Pu, R., & Gong, P. (2011). Hyperspectral Remote Sensing of Vegetation Bioparameters. In Q. Weng (Ed.), *Advances in environmental remote sensing: sensors, algorithm, and application* (pp. 101–142). CRC Press.

Qu, Y., Wang, J., Wan, H., Li, X., & Zhou, G. (2008). A Bayesian network algorithm for retrieving the characterization of land surface vegetation. *Remote Sensing of Environment*, *112*(3), 613–622. doi:10.1016/j.rse.2007.03.031

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, *92*(437), 179–191. doi:10.1080/01621459.1997.10473615

Ramoelo, A., Skidmore, A. K., Cho, M. a., Mathieu, R., Heitkönig, I. M. A., Dudeni-Tlhone, N., … Prins, H. H. T. (2013). Non-linear partial least square regression increases the estimation accuracy of grass nitrogen and phosphorus using in situ hyperspectral and environmental data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *82*, 27–40. doi:10.1016/j.isprsjprs.2013.04.012

Ramoelo, A., Skidmore, A. K., Schlerf, M., Heitkönig, I. M. A., Mathieu, R., & Cho, M. A. (2013). Savanna grass nitrogen to phosphorous ratio estimation using field spectroscopy and the potential for estimation with imaging spectroscopy. *International Journal of Applied Earth Observation and Geoinformation*, *23*, 334–343. doi:10.1016/j.jag.2012.10.009

Ramoelo, A., Skidmore, A. K., Schlerf, M., Mathieu, R., & Heitkönig, I. M. A. (2011). Water-removed spectra increase the retrieval accuracy when estimating savanna grass nitrogen and phosphorus concentrations. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(4), 408–417. doi:10.1016/j.isprsjprs.2011.01.008

Richardson, A. D., Duigan, S. P., & Berlyn, G. P. (2002). An evaluation of noninvasive methods to estimate foliar chlorophyll content. *New Phytologist*, *153*(1), 185–194. doi:10.1046/j.0028-646X.2001.00289.x

Richter, K., Atzberger, C., Hank, T. B., & Mauser, W. (2012). Derivation of biophysical variables from Earth observation data: validation and statistical measures. *Journal of Applied Remote Sensing*, *6*(1), 063557–1. doi:10.1117/1.JRS.6.063557

Rivera, J. P., Verrelst, J., Muñoz-Marí, J., Moreno, J., & Camps-Valls, G. (in press). Toward a Semiautomatic Machine Learning Retrieval of Biophysical Parameters. *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*.

Rivera, J., Verrelst, J., Delegido, J., Veroustraete, F., & Moreno, J. (2014). On the Semi-Automatic Retrieval of Biophysical Parameters Based on Spectral Index Optimization. *Remote Sensing*, *6*(6), 4927–4951. doi:10.3390/rs6064927

Roberto, C., Lorenzo, B., Michele, M., Micol, R., & Cinzia, P. (2012). Optical Remote Sensing of Vegetation Water Content. In P. S. Thenkabail, J. G. Lyon, & A. Huete (Eds.), *Hyperspectral remote sensing of vegetation*. Boca Raton, FL: CRC Press.

Roberts, D. A., Roth, K. L., & Perroy, R. L. (2012). Hyperspectral Vegetation Indices. In P. S. Thenkabail, J. G. Lyon, & A. Huete (Eds.), *Hyperspectral remote sensing of vegetation*. Boca Raton, FL: CRC Press.

Running, S. W., & Coughlan, J. C. (1988). A general model of forest ecosystem processes for regional applications I. Hydrologic balance, canopy gas exchange and primary production processes. *Ecological Modelling*, *42*(2), 125–154. doi:10.1016/0304-3800(88)90112-3

Running, S. W., & Hunt, E. R. (1993). *Scaling Physiological Processes*. *Scaling Physiological Processes* (pp. 141–158). Elsevier. doi:10.1016/B978-0-12-233440-5.50014-2

Schellberg, J., Hill, M. J., Gerhards, R., Rothmund, M., & Braun, M. (2008). Precision agriculture on grassland: Applications, perspectives and constraints. *European Journal of Agronomy*, *29*(2-3), 59–71. doi:10.1016/j.eja.2008.05.005

Schlerf, M., Atzberger, C., & Hill, J. (2005). Remote sensing of forest biophysical variables using HyMap imaging spectrometer data. *Remote Sensing of Environment*, *95*(2), 177–194. doi:10.1016/j.rse.2004.12.016

Schmidt, K. S., & Skidmore, A. K. (2001). Exploring spectral discrimination of grass species in African rangelands. *International Journal of Remote Sensing*, *22*(17), 3421–3434. doi:10.1080/01431160152609245

Segl, K., Guanter, L., Rogass, C., Kuester, T., Roessner, S., Kaufmann, H., … Hofer, S. (2012). EeteS—The EnMAP End-to-End Simulation Tool. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(2), 522–530. doi:10.1109/JSTARS.2012.2188994

Serneels, S., Croux, C., Filzmoser, P., & van Espen, P. J. (2005). Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, *79*(1-2), 55–64. doi:10.1016/j.chemolab.2005.04.007

Serrano, L., Peñuelas, J., & Ustin, S. L. (2002). Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data. *Remote Sensing of Environment*, *81*(2-3), 355–364. doi:10.1016/S0034-4257(02)00011-1

Siebke, K., & Ball, M. C. (2009). Non-destructive measurement of chlorophyll b:a ratios and identification of photosynthetic pathways in grasses by reflectance spectroscopy. *Functional Plant Biology*, *36*(11), 857. doi:10.1071/FP09201

Sims, D. A., & Gamon, J. A. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment*, *81*(2-3), 337–354. doi:10.1016/S0034-4257(02)00010-X

Somers, B., Cools, K., Delalieux, S., Stuckens, J., van der Zande, D., Verstraeten, W. W., & Coppin, P. (2009). Nonlinear Hyperspectral Mixture Analysis for tree cover estimates in orchards. *Remote Sensing of Environment*, *113*(6), 1183–1193. doi:10.1016/j.rse.2009.02.003

Stevens, A., & Ramirez-Lopez, L. (2013). An introduction to the prospectr package. R package Vignette R package version 0.1.3.

Straatsma, M., & Middelkoop, H. (2007). Extracting structural characteristics of herbaceous floodplain vegetation under leaf-off conditions using airborne laser scanner data. *International Journal of Remote Sensing*, *28*(11), 2447–2467. doi:10.1080/01431160600928633

Team, E. S.-2. (2010). *GMES Sentinel-2 Mission Requirement Document* (pp. 25–26). Retrieved from http://esamultimedia.esa.int/docs/GMES/Sentinel-2_MRD.pdf

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Thenkabail, P. S., Gumma, M. K., Teluguntla, P., & Mohammed, I. A. (2014). Hyperspectral Remote Sensing of Vegetation and Agricultural Crops. *Photogrammetric Engineering and Remote Sensing*, *80*(8), 697–709.

Thenkabail, P. S., Lyon, J. G., & Huete, A. (2011). Advances in hyperspectral remote sensing of vegetation and agricultural crops. In *Hyperspectral Remote Sensing of Vegetation* (pp. 3–29). Boca Raton, FL: CRC Press/Taylor and Francis Group.

Thenkabail, P. S., Lyon, J. G., & Huete, A. (2011). Hyperspectral Remote Sensing of Vegetation and Agricultural Crops: Knowlwdge Gain and Knowledge Gap After 40 Years of Research. In P. S. Thenkabail, J. G. Lyon, & A. Huete (Eds.), *Hyperspectral remote sensing of vegetation* (pp. 663–688). Boca Raton: CRC Press.

Thenkabail, P. S., Mariotto, I., Gumma, M. K., Middleton, E. M., Landis, D. R., & Huemmrich, K. F. (2013). Selection of Hyperspectral Narrowbands (HNBs) and Composition of Hyperspectral Twoband Vegetation Indices (HVIs) for Biophysical Characterization and Discrimination of Crop Types Using Field Reflectance and Hyperion / EO-1 Data, *6*(2), 427–439.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved from http://www.jstor.org/stable/2346178

Tobias, R. D. (1995). An introduction to partial least squares regression. In *Proc. Ann. SAS Users Group Int. Conf., 20th* (pp. 2–5). Orlando, FL.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., & Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment*, *112*(5), 1982–1999. doi:10.1016/j.rse.2007.03.032

Tran, T. N., Afanador, N. L., Buydens, L. M. C., & Blanchet, L. (2014). Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). *Chemometrics and Intelligent Laboratory Systems*, *138*, 153–160. doi:10.1016/j.chemolab.2014.08.005

Turner, D. P., Ollinger, S. V., & Kimball, J. S. (2004). Integrating Remote Sensing and Ecosystem Process Models for Landscape- to Regional-Scale Analysis of the Carbon Cycle. *BioScience*, *54*(6), 573–584. doi:10.1641/0006-3568(2004)054[0573:IRSAEP]2.0.CO;2

Ullah, S., Si, Y., Schlerf, M., Skidmore, A. K., Shafique, M., & Iqbal, I. A. (2012). Estimation of grassland biomass and nitrogen using MERIS data. *International Journal of Applied Earth Observation and Geoinformation*, *19*, 196–204. doi:10.1016/j.jag.2012.05.008

Ustin, S. L., Gitelson, a. a., Jacquemoud, S., Schaepman, M., Asner, G. P., Gamon, J. a., & Zarco-Tejada, P. (2009). Retrieval of foliar information about plant pigment systems from high resolution spectroscopy. *Remote Sensing of Environment*, *113*, S67–S77. doi:10.1016/j.rse.2008.10.019

van der Meer, F., De Jong, S., & Bakker, W. (2001). Imaging spectroscopy: basic analytical techniques. In *Imaging Spectroscopy: Basic Principles and Prospective Applications* (1st ed., pp. 17–60). Springer Science & Business Media Netherlands.

van Wittenberghe, S., Verrelst, J., Rivera, J. P., Alonso, L., Moreno, J., & Samson, R. (2014). Gaussian processes retrieval of leaf parameters from a multi-species reflectance, absorbance and fluorescence dataset. *Journal of Photochemistry and Photobiology. B, Biology*, *134*, 37–48. doi:10.1016/j.jphotobiol.2014.03.010

Verrelst, J., Alonso, L., Camps-valls, G., Member, S., Delegido, J., & Moreno, J. (2012). Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques, *50*(5), 1832–1843.

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J. P., Camps-Valls, G., & Moreno, J. (2012). Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sensing of Environment*, *118*, 127–139. doi:10.1016/j.rse.2011.11.002

Wang, F., Huang, J., Wang, Y., Liu, Z., & Zhang, F. (2012). Estimating nitrogen concentration in rape from hyperspectral data at canopy level using support vector machines. *Precision Agriculture*, *14*(2), 172–183. doi:10.1007/s11119-012-9285-2

Wang, Q., Adiku, S., Tenhunen, J., & Granier, A. (2005). On the relationship of NDVI with leaf area index in a deciduous forest site. *Remote Sensing of Environment*, *94*(2), 244–255. doi:10.1016/j.rse.2004.10.006

Watson, D. J. (1947). Comparative psychological studies in the growth of field crops I. Variation in net assimilation rate and leaf area between species and varieties, and within and between years. *Annals of Botany*, *11*, 41–76.

White, R. P., Murray, S., & Rohweder, M. (2000). *Pilot Analysis of Global Ecosystems: Grassland Ecosystems*. Washington, DC.

Williams, P., & Norris, K. (1987). *Near-Infrared Technology in the Agricultural and Food Industries*. (P. Williams & K. Norris, Eds.). St. Paul, MN: American Association of Cereal Chemists.

Wintle, B. A., McCarthy, M. A., Volinsky, C. T., & Kavanagh, R. P. (2003). The Use of Bayesian Model Averaging to Better Represent Uncertainty in Ecological Models. *Conservation Biology*, *17*(6), 1579–1590. doi:10.1111/j.1523-1739.2003.00614.x

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*, 109–130. doi:10.1016/S0169-7439(01)00155-1

Yang, X., Huang, J., Wang, J., Wang, X., & Liu, Z. (2007). Estimation of vegetation biophysical parameters by remote sensing using radial basis function neural network. *Journal of Zhejiang University SCIENCE A*, *8*(6), 883–895. doi:10.1631/jzus.2007.A0883

Yi, G., Shi, J. Q., & Choi, T. (2011). Penalized gaussian process regression and classification for high-dimensional nonlinear data. *Biometrics*, *67*(4), 1285–94. doi:10.1111/j.1541-0420.2011.01576.x

Yoder, B. J., & Pettigrew-Crosby, R. E. (1995). Predicting nitrogen and chlorophyll content and concentrations from reflectance spectra (400–2500 nm) at leaf and canopy scales. *Remote Sensing of Environment*, *53*(3), 199–211. doi:10.1016/0034-4257(95)00135-N

Zarco-Tejada, P. J., Miller, J. R., Mohammed, G. H., Noland, T. L., & Sampson, P. H. (2002). Vegetation Stress Detection through Chlorophyll + Estimation and Fluorescence Effects on Hyperspectral Imagery. *Journal of Environment Quality*, *31*(5), 1433. doi:10.2134/jeq2002.1433

Zeugner, S. (2011). Bayesian Model Averaging with BMS for BMS Version 0.3.0. R BMS Package Documentation.

Zhao, D., Huang, L., Li, J., & Qi, J. (2007). A comparative analysis of broadband and narrowband derived vegetation indices in predicting LAI and CCD of a cotton canopy. *ISPRS Journal of Photogrammetry and Remote Sensing*, *62*(1), 25–33. doi:10.1016/j.isprsjprs.2007.01.003

Zhao, K., Valle, D., Popescu, S., Zhang, X., & Mallick, B. (2013). Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection. *Remote Sensing of Environment*, *132*, 102–119. doi:10.1016/j.rse.2012.12.026

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. doi:10.1198/016214506000000735

APPENDIX

A. Importance of LAI and chlorophyll in ecosystem functioning and precision agriculture

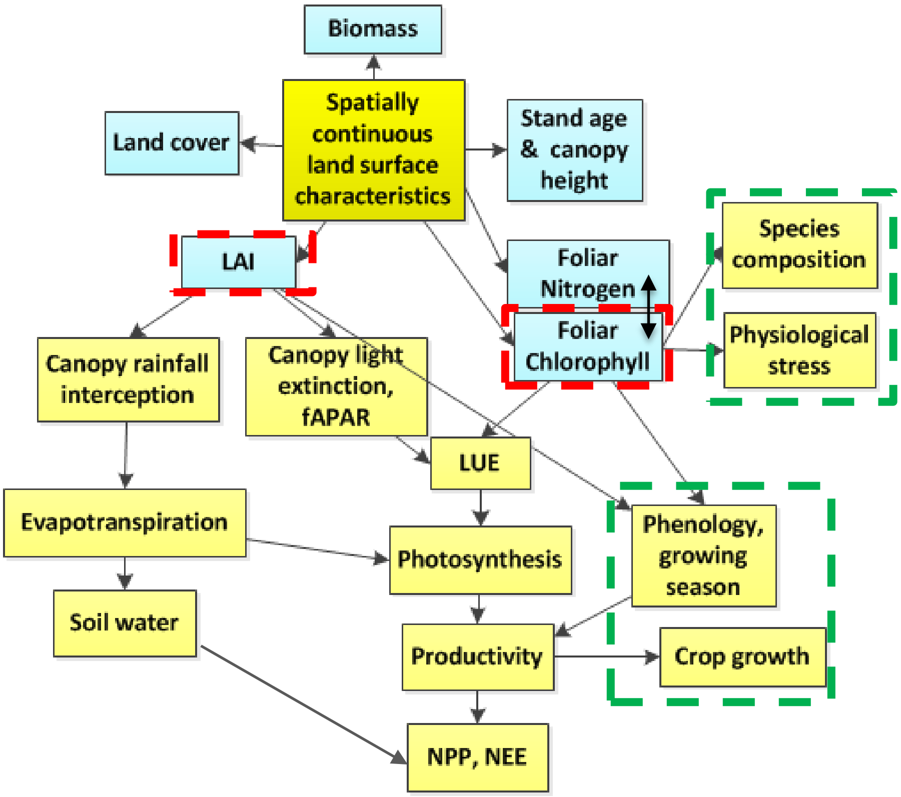


Figure 20. Land surface characteristics (in blue) successfully estimated (in spatially-explicit manner) using remote sensing as input to ecosystem models linking carbon, energy, water, and nutrient balance. Also shown (in yellow) are the key plant functions influenced by LAI and plant chlorophyll, which are associated to ecosystem processes governing the atmosphere-biosphere exchanges/interactions, and useful information for precision agriculture (in green box). (1) LAI on one hand is related to light interception (fAPAR) by plants for photosynthesis and thus the ecosystem primary productivity (NPP), and on the other hand also determines canopy rainfall interception which in turn influences plant transpiration and soil water balance. Importantly, the leaf transpiration is closely linked to CO_2 fixation/uptake from the atmosphere (thus governing photosynthesis rate) in a mechanism called the 'stomatal conductance'. Change in LAI is also useful for monitoring vegetation seasonal growth/timing (phenology). It is therefore not an exaggerated statement to say that leaf area is the grand entrance/interface between biosphere, atmosphere, and hydrosphere. (2) Foliar chlorophyll with its close association to nitrogen very importantly determines the fraction of the intercepted light actually used for photosynthesis (LUE) and thus total canopy chlorophyll (e.g. leaf chlorophyll x LAI) is considered the most directly relevant indicator of productivity. Chlorophyll is also an indicator of plant health status. (3) In climate change studies, the ecosystem carbon balance (net ecosystem exchange (NEE)) is estimated from the net primary productivity (NPP) and soil respiration, which among all, is influenced by soil water. (adapted from Running & Coughlan (1988); Dawson et al. (2003); Turner, Ollinger, & Kimball (2004); and Gitelson et al. (2006))

B. Summary of hyperspectral RS studies estimating LAI and chlorophyll using statistical methods

Table 8. Hyperspectral RS studies of LAI and chlorophyll up to 2014. For abbreviations see under Table.

No	Reference	Parameter(s)	Scale	Statistical methods	Vegetation type(s)	Retrieval accuracy
1	Curran et al. (1997)	Canopy Chl (CCC)	Airborne (AVIRIS)	SMLR of 1st DS	Pine needle	R^2=0.78-0.99
2	Blackburn (1998)	Leaf Chl and Canopy Chl per unit mass (concentration) and area (content: LCC, CCC)	Field (canopy spectra, fresh leaves)	*HNBVI*, REIP, *DS*	Mature bracken	CCC per area r=0.97 (HNBVI); CCC per mass r=0.81-0.91 (2nd DS @ 664.3 nm); LCC per mass r=0.83-0.84 (DS @ 729.3 nm)
3	Jago, Cutler, & Curran (1999)	CCC	Field and airborne (canopy spectra)	REIP	Grassland, winter wheat	Field: r=0.84 (grass) & r=0.80 (wheat); Airborne: nRMSE=12.69% (grass) & nRMSE=16.4% (wheat)
4	Curran et al. (2001)	LCC	Lab (laboratory, leaf spectra, dried compounds)	SMLR of 1st DS, band depth	Pine needle	R^2=0.96 (nRMSE=0.12%)
5	Haboudane et al. (2004)	LAI	Airborne (CASI)	HNBVIs	Crops (soybean, corn, wheat)	R^2=0.74-0.98 (nRMSE=0.28-0.85%)
6	Lee et al. (2004)	LAI	Airborne (AVIRIS)	HNBVIs, *CCA*	Crop, tallgrass prairie, conifer forest	R^2=0.9
7	le Maire et al. (2004)	LCC	Lab (leaf spectra)	*HNBVIs*, REIP, NN	Deciduous tree species + simulated	RMSE=3.7 μg cm^{-2} (HNBVI); 4.2 μg cm^{-2} (NN); N/A cause 'double peak' (REP)
8	Pu and Gong (2004)	LAI	Spaceborne (Hyperion)	SMLR, PCA, *WT*	Mixed conifer forest	Mapped accuracy 75% (WT), 52% (PCA), 51% (SMLR)
9	Gitelson et al. (2005)	CCC (LCCxLAI)	Field (canopy)	HNBVIs (green and red edge)	Crops (maize, soybean)	R^2=0.92

10	Schlerf et al. (2005)	LAI	Airborne (HyMap)	*HNBVIs*, REIP	Forest (Norway spruce)	R^2=0.77 (nRMSE=17%)
11	Blackburn (2007)	LCC	Field, and lab (individual & stack of leaves)	WT with SMLR	Broadleaved trees, bracken, matorral	Combined dataset: R^2=0.63 (nRMSE=57.7%); leaf: R^2=0.75 (nRMSE=28%); stacks: R^2=0.74 (nRMSE=40%); canopy: R^2=0.49 & 0.86 (nRMSE=54% & 25%);
12	Yang et al. (2007)	CCC, LAI	Field (canopy)	RBF-NN of HNBVIs	Rice	R^2=0.66 for LAI, R^2=0.82 for CCC
13	Zhao et al. (2007)	CCC, LAI	Field (canopy)	HNBVIs	Cotton	R^2=0.85 for CCC, R^2>0.8 for LAI
14	Darvishzadeh et al. (2008)	LAI, LCC, CCC (LCCxLAI)	Field (canopy)	*PLSR*, HNBVIs, REIP, SMLR	Mediterranean grassland	PLSR: R^2=0.69 (nRMSE=32%) for LAI, 0.40 (17%) for LCC, 0.74 (34%) for CCC (other results under table[1])
15	Lemaire et al. (2008)	CCC, LAI	Spaceborne (Hyperion)	HNBVI	Broadleaved forest	RMSE=8.2 µg cm^{-2} for CCC; 1.7 m^{2}m^{-2} for LAI
16	Atzberger et al. (2010)	CCC (LCCxLAI)	Airborne (HyMap)	SMLR, PCR, *PLSR*, NDVI	Winter wheat	NDVI: R^2=0.73 (nRMSE=32%); PLSR: R^2=0.82 (nRMSE=21%); PCR: R^2=0.57 (nRMSE=33%); SMLR: R^2=0.79 (nRMSE=24%)
17	Delegido et al. (2010)	LCC, CCC (LCCxLAI)	Spaceborne (CHRIS); airborne (CASI)	NAOC	Different crops	r=0.91 for LCC; r=0.97 (RMSE=4.2 µg cm^{-2}) for CCC
18	Ju et al. (2010)	LCC	Field and lab	Red edge parameters (position, amplitude, area, *symmetry*)	Rapeseed and wheat	Best (symmetry) r>0.8 for both field and lab
19	Darvishzadeh et al. (2011)	LAI	Airborne (HyMap)	HNBVI, PLSR	Mediterranean grassland	HNBVI: R^2=0.85 (nRMSE=21%);

						PLSR: R^2=0.87 (nRMSE=22%)
20	Herrmann et al. (2011)	LAI	Field (canopy)	*PLSR*, REIP, NDVI	Crop (wheat, potato)	Data pooled: r=0.93 (8.5% for potato, 11.5% for wheat) (PLSR), r=0.81 (REIP), r=0.7 (NDVI)
21	Huang & Blackburn (2011)	LCC	Simulated	SMLR of wavelet coefficients, original spectra and 1st DS	Simulated with PROSPECT	R^2=0.99
22	Main et al. (2011)	LCC	Lab (leaf spectra)	73 HNBVIs	3 crop species, 8 savanna tree species	Best (red edge indices) R^2=0.90 (nRMSE=55-57 mg m^{-2}) for combined dataset
23	Clevers & Kooistra (2012)	CCC	Simulated	HNBVI ($CI_{red\ edge}$)	Simulated	R^2=0.94

24	Verrelst et al. (2012)	LCC, LAI	Spaceborne (CHRIS)	GPR applied to single band, HNBVIs, NAOC	9 crop species	Best HNBVI: r=0.87 for LCC, 0.92 for LAI;
						NAOC: r=0.86 for LCC;
						Single band r=0.99 (RMSE=2.24 μg cm^{-2}) for LCC, 0.93 (RMSE=0.57 m^2 m^{-2}) for LAI)

25	Zhao et al. (2013)	LCC	Lab (leaf spectra), LCC extracted from both dry and fresh leaves	*BMA*, PLSR, SMLR	80 species (tree and crop) across globe from 3 spectro-chemical datasets	R^2 of dataset
						ACCP fresh: 0.76 (BMA), 0.71 (PLSR), 0.65 (SMLR);
						ACCP dry: 0.76 (BMA), 0.74 (PLSR), 0.70 (SMLR);
						MM fresh: 0.96 (BMA), 0.93 (PLSR), 0.96 (SMLR)

26	Li et al. (2014)	LAI	Field (canopy)	PLSR applied to spectral features	Wheat	R^2=0.88 (nRMSE=25.5%)
27	Navarro-Cerrillo et	Stand Chl	Airborne (AHS),	HNBVIs	Mediterranean pine	R^2=0.65 for AHS sensor, 0.56 for

	al. (2014)		spaceborne (CHRIS, Hyperion)			Hyperion, 0.57 for CHRIS
28	Rivera et al. (2014)	LAI, LCC	Airborne (HyMap)	HNBVIs	Different crop types	R^2=0.83 (LAI), R^2=0.93 (LCC)
29	van Wittenberghe et al. (2014)	LCC, SLA (Specific Leaf Area)	Field (canopy)	GPR	Multi-species (tree)	R^2=0.84 (nRMSE=9.1%) for LCC; R^2=0.87 (nRMSE=6.0%) for SLA

[1]HNBVI: R^2=0.63 (nRMSE=33%) for LAI, 0.26 (17%) for LCC, 0.68 (35%) for CCC;
REP: R^2=0.52 (nRMSE=38%) for LAI, 0.21 (18%) for LCC, 0.58 (40%) for CCC;
SMLR: R^2=0.66 (nRMSE=33%) for LAI, 0.25 (18%) for LCC, 0.72 (33%) for CCC

Abbreviations: LCC is leaf chlorophyll content and CCC is canopy chlorophyll content; nRMSE is root mean squared error normalized to mean of measured response; HNBVIs is hyperspectral narrowband vegetation indices; REIP is red edge inflection point; SMLR is stepwise multiple linear regression; DS is first derivative spectra; CCA is canonical component analysis; PCA is principal component analysis; NN is neural network; WT is wavelet transform; RBF is radial basis function; BMA is bayesian model averaging; PLSR is partial least square regression; PCR is principal component regression; NAOC is normalized area over reflectance curve; CI is chlorophyll index; GPR is Gaussian process regression.

C. Existing and planned hyperspectral missions and sensor characteristics.

Table 9. Existing and future (planned) hyperspectral RS missions and sensors (Source: Jones & Vaughan (2010, p. 339, 341); Ortenberg (2011); Thenkabail et al. (2014))

No	Sensor	Operational mode	Spatial resolution (m)	Number of bands	Swath (km)	Spectral range (nm)	Band widths (nm)	Frequency of revisit (days)	Launch (date)
1	Hyperion, EO-1 (JPL, USA)	Spaceborne	30	242 (196 calibrated)	7.5	196 calibrated: VNIR (band 8 to 57) 427.55 to 925.85; SWIR (band 79 to 224) 932.72 to 2395.53	10 (approx.) for all 196 bands	16	2000
2	CHRIS, PROBA (ESA, UK)	Spaceborne	18-36 (5 viewing angle)	19-63 (modifiable)	17.5	400-1050	1.25-11	7 (approx.)	2001
3	HyspIRI VSWIR (NASA/JPL, USA)	Spaceborne	60	210	145	210 bands in 380-2500	10 (approx.) for all 210 bands	19	2020+
4	EnMAP (DLR, Germany)	Spaceborne	30	98 130	30	420-1000 900-2450	6.5+/-1.25 10+/-2.5	4 (pointing) & 21	2015+
5	PRISMA (ASI, Italy)	Spaceborne	30	249	30	400-2500	10	7	2015+
6	HISUI/ALOS-3 (JAXA, Japan)	Spaceborne	30	185	30	440-2500	10-12.5	60	2015+
7	HYPXIM-P VSWIR (CNES, France)	Spaceborne	<8	210	16	400-2500	10	19	2020+
8	HyMAP (HyVISTA)	Airborne	3-10 m (depends on flight altitude)	126	(Depends on flight altitude)	436-2485	13-17	N/A	last 17+ years
9	ASD spectroradiometer	Proximal (handheld)	1134 cm2 held at 1.2 m Nadir view 18 degree FOV	2100 effective 1 nm width between 400-2500 nm	N/A	2100 effective bands	1 nm wide (approx.) in 400-2400	N/A	last 30+ years
10	GER 3700 (GER)	Proximal (handheld)	diameter area 45 cm held at 1 m Nadir view 25 degree FOV	584+	N/A	350-2500	1.5-9.5	N/A	N/A

D. Explanation on bias-variance trade-off

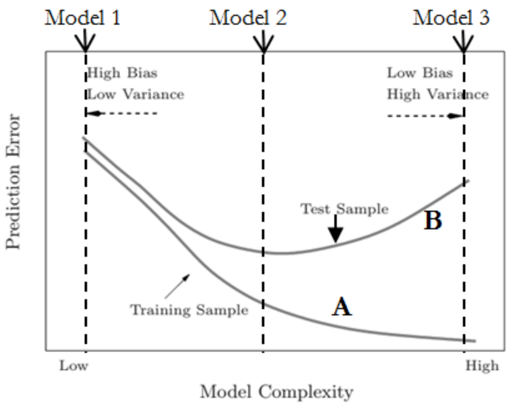


Figure 21. The bias-variance tradeoff (adapted from Hastie et al., 2009, p. 38). A statistical model seeks the optimum model 2 that minimizes test error.

In order to properly evaluate statistical models for the main purpose of making the most accurate predictive model, it is vital to assess the true predictive performance (prediction error, or generalization error) over an *independent* test sample. According to Hastie et al. (2009, p. 37), the true prediction error of a regression model is:

$$Err = Irreducible\ Error + Bias^2 + Variance$$

where the first term is the *irreducible* error associated with natural variability in the system/phenomenon of interest, which is beyond our control. Our models aim to minimize the reducible error which can be decomposed into bias and variance (the second and third term). To put simply, the *bias* term is the difference between the target true function (the reality) we want to recover with statistical models, and our best approximation of that function (i.e., our prediction over the *training* data, shown as curve A in Figure 21). The bias will always decrease as we increase the model complexity simply because the more complex the model is, the more flexible the function can adapt to the training data. An overly complex (low bias) model therefore will risk over-fit the training data (e.g., also fit the noise pattern) and will not generalize well when extrapolated to an independent test sample not used in the training/calibration step. That is, the model will have large test/generalization error (curve B in Figure 21), much larger than the overly-optimistic training error. This difference in performance between datasets (i.e., between training and test set) indicates the *variance* term. Thus, an over-fit/overly complex model (Model 3 in Figure 21) will have low bias but high variance while an under-fit model (Model 1 in Figure 21) will have high bias but low variance: there is bias-variance tradeoff. Objectively, we seek the optimum model with just-enough complexity that minimizes the test error (Model 2 in Figure 21).

E. Graphical explanation of Lasso

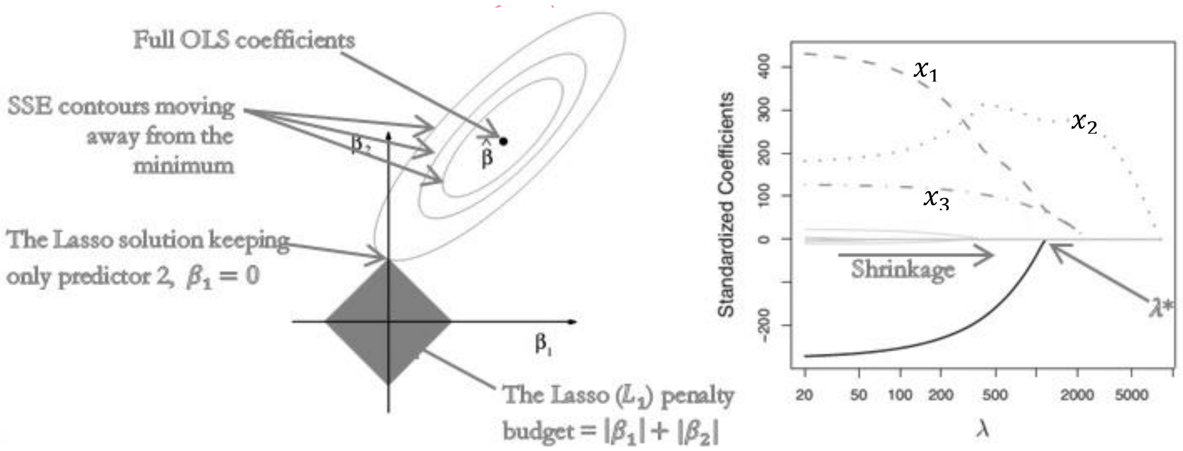

Full OLS coefficients
SSE contours moving away from the minimum
β_2
$\hat{\beta}$
The Lasso solution keeping only predictor 2, $\beta_1 = 0$
β_1
The Lasso (L_1) penalty budget = $|\beta_1| + |\beta_2|$
Standardized Coefficients
400
300
200
100
0
-200
x_1
x_2
x_3
Shrinkage
λ^*
20
50
100
200
500
2000
5000
λ

Figure 22. The Lasso estimates. *Left:* Lasso solution at the point where the contour of errors—(β_1, β_2) combinations which give equal SSE—intersects with the L_1 penalty budget constraint represented by the blue diamond-shaped region. The solution shrinks β_1 to 0 thus discarding predictor 1 and selecting only predictor 2. Note that the full OLS estimate $\hat{\beta}$ keeping both β_1 & β_2 here is local (not global) optimum solution (especially high risk to occur in high dimensional setting) which is over-fitting and therefore does not give optimum prediction accuracy. *Right*: An example of the Lasso coefficient shrinkage 'evolution' showing as the tuning parameter λ increases, the coefficients are shrunk from being too much inflated (over-fitting) and at optimum λ^*(determined by cross-validation) only 3 predictors (x_1, x_2, x_3) are selected in the final model. Figures adapted from Hastie et al. (2009, p. 71) and James et al. (2013, p. 220).

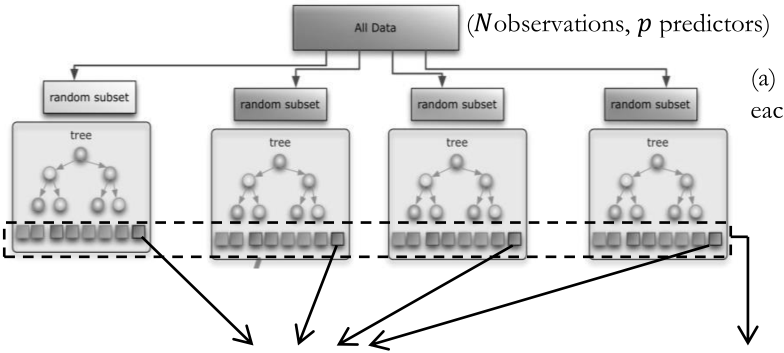
F. Random Forest regression: the algorithm

1. For $b = 1$ to B:
   (a) **Draw a bootstrap sample $\boldsymbol{Z}^*$** (~70%) from the training data.
   (b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each node:
      i. **Select $\boldsymbol{m < p}$ variables at random** from the p variables.
      ii. Pick the best variable/split-point among the m using ***RSS*** criterion:

$$RSS = \sum_{left} (y_i - y_L^*)^2 + \sum_{right} (y_i - y_R^*)^2$$

      where y_L^*= mean y for left node; y_R^*= mean y for right node.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$ i.e., **averaging the prediction of $\boldsymbol{x}$ over all $\boldsymbol{B}$ trees.**



(N observations, p predictors)

(a) Grow B independent trees T_b (here 4), each with bootstrap samples Z^*=~70%

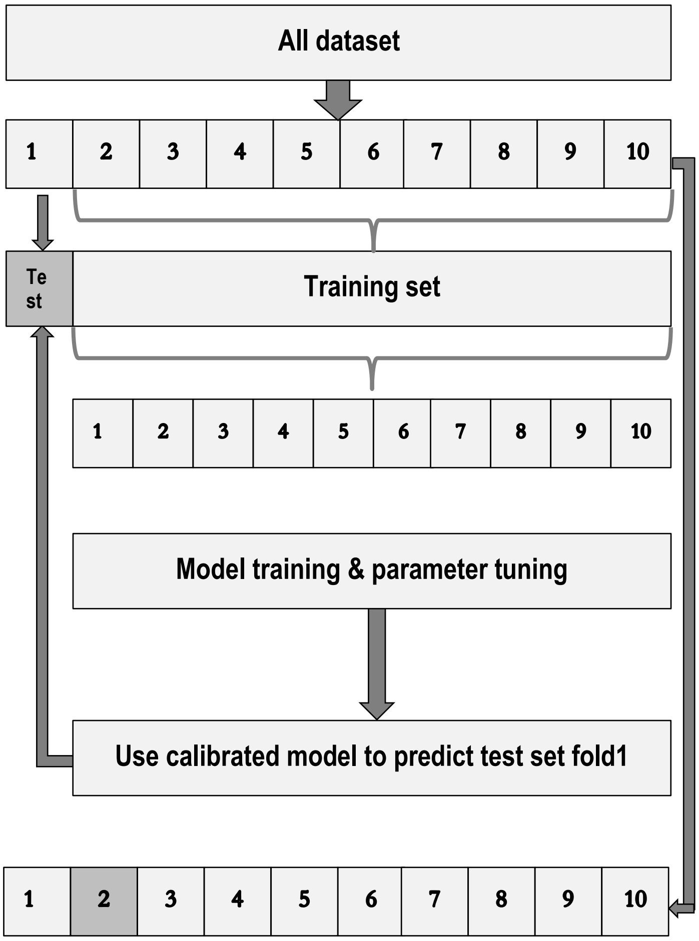
(b) At each tree node, randomly select $m < p$ predictors. Pick the best predictor in m based on RSS to split the node into two daughter nodes.

(c) Use the tree to predict the ~30% out-of-bag observations

(d) Compute predicted value of each observation n by averaging the prediction from all trees

Figure 23. Schematic diagram of Random Forest regression (adapted from Benyamin, 2012)

G. Cross-validation procedure employed in this present study



Split (stratified) data into 10 folds for outer CV, for model assessment.

CV run no. 1 (out of 10) to predict test set fold 1. Fold 2-10 are used for training.

Split training set into 10 folds for inner CV, for model selection (parameter tuning):
(1) PLSR: no. of PLS **factors**
(2) Lasso: regularizer λ
(3) Random forest: *ntree* and *mtry*
(4) BMA: none[1]
Choose optimum parameter that gives cross-validated training RMSE within 1 std. error of the minimum[2]

Use trained model to predict test set fold 1, record R^2 and relative RMSE.

CV run no. 2 (out of 10) to predict test set fold 2. Fold 1, 3-10 are used for training.

Repeat for the remaining CV runs (3-10), each time predicting different fold. Finally, report average (also standard deviation) of the recorded 10 values of R^2 and relative RMSE (as % from mean response value). These accuracy metrics henceforth are called $\boldsymbol{R^2_{cv}}$ and $\mathbf{nRMSE}_{\boldsymbol{cv}}$

Figure 24. Schematic diagram of the nested 10-fold stratified cross-validation procedure employed in this study.

[1]BMA model settings require no tuning, the acceptable setting (to ensure MCMC convergence configured by trial-and-error are used for all runs.

[2]Based on the Breiman et al. (1984) “one-standard error rule” for model parsimony (Kuhn & Johnson, 2013, p. 75)

H. Optimal non-redundant narrowbands for studying vegetation used for band interpretation analysis

Table 10. Optimal non-redundant hyperspectral narrowbands for studying vegetation and crops (Thenkabail et al., 2014)

No.	***Waveband (width)***	***Attributed to***	***No.***	***Waveband (width)***	***Attributed to***
T1	375 (5)	fPAR, leaf water	T14	970 (10)	Water, moisture, biomass
T2	405 (5)	Nitrogen, senescing	T15	1075 (5)	Biophysical and biochemical quantities
T3	490 (5)	Carotenoid, LUE, stress	T16	1180 (5)	Water absorption
T4	515 (5)	Pigments (Car, Chl, Anth), nitrogen	T17	1245 (5)	Water sensitivity
T5	531 (1)	LUE	T19	1518 (5)	Moisture and biomass
T6	550 (5)	Chlorophyll	T20	1650 (5)	Heavy metal stress, moisture sensitivity
T7	570 (5)	Pigments (Anth, Chl), nitrogen	T21	1725 (5)	Lignin, biomass, starch, moisture
T8	682 (5)	Biophysical quantities and yield	T22	1950 (5)	Water absorption
T9	705 (5)	Stress and Chl	T23	2025 (5)	Litter, lignin, cellulose
T10	720 (5)	Stress and Chl	T24	2133 (5)	Litter, lignin, cellulose
T9-T11	700-740	Chl, senescing, stress	T25	2205 (5)	Litter, lignin, cellulose, sugar, starch, protein, heavy metal stress
T12	855 (20)	Biophysical quantities and yield	T26	2260 (5)	Moisture and biomass
T13	910 (5)	Moisture, biomass, protein	T28	2359 (5)	Cellulose, protein, nitrogen

I. Internal variable selection with PLSR: sparse PLSR

Sparse PLSR has been proposed by Chun & Keleş (2010) who showed that the known asymptotic consistency of PLSR estimator for a univariate response does not hold with the large p and small n paradigm. Sparse PLSR imposes "sparsity" in the dimension reduction resulting in a sparse linear combination and thus performs dimensionality reduction and variable selection simultaneously. In addition to the parameter k (number of PLS factors), we need to tune the parameter eta (0-1) which controls the amount of sparsity in the solution; the higher eta the more variables will get zero regression coefficients and thus removed. Compared to standard PLSR model (best R^2_{cv}=0.416, 0.662, and 0.760; best $nRMSE_{cv}$=16.2%, 31.1%, and 32.1%, respectively for LCC, LAI, and CCC), no major improvement in predictive accuracy was observed by the variable selection (*Bands* is the number of bands retained) as shown in Table 11. Sparse PLSR was implemented using 'spls' package (Chung, Chun, & Keles, 2013) in R statistical environment (R Core Team, 2014).

Table 11. Results of sparse PLSR to assess if internal band selection improves predictive accuracy. 'Bands' is the average number of bands retained/selected in the cross-validation runs. 'Factors' is the average number of PLS factors (optimum). In parentheses is the standard deviation.

Spectral transformation	*LCC*				*LAI*			
	R^2_{cv}	$nRMSE_{cv}$ (%)	*Bands*	*Factors*	R^2_{cv}	$nRMSE_{cv}$ (%)	*Bands*	*Factors*
a. None	**0.434** (0.19)	**16.07** (4.3)	159	3.7	**0.665** (0.15)	**31.03** (7.5)	296	4.3
b. CR	0.363 (0.25)	16.36 (4.8)	23	1	0.612 (0.18)	33.09 (8.1)	226	1.8
c. FDR	0.376 (0.22)	16.76 (4.6)	8	1.5	0.636 (0.12)	32.11 (6.1)	46	1.7
d. Abs	0.392 (0.21)	16.61 (4.4)	186	3.7	0.636 (0.16)	32.19 (6.9)	217	6.3

Spectral transformation	*CCC*			
	R^2_{cv}	$nRMSE_{cv}$ (%)	*Bands*	*Factors*
a. None	0.704 (0.11)	35.00 (6.1)	357	4.8
b. CR	0.719 (0.09)	34.10 (7.3)	99	1
c. FDR	0.712 (0.09)	34.17 (5.9)	37	1
d. Abs	**0.744** (0.10)	**33.78** (7.0)	219	7.2

J. Partial robust M-regression (robust PLSR)

Robust M-estimator (Serneels et al., 2005) gives protection against vertical outlier (outliers in error terms) and leverage points (outlying observations in the predictor space) by assigning weights to them as follows:

$$\widehat{\boldsymbol{\beta}}_{RM} = argmin_{\beta} \sum_{i=1}^{n} w_i^r w_i^x \, (y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2$$

where we solve the regression coefficient $\widehat{\boldsymbol{\beta}}_{RM}$ by minimizing the sum of squared residuals and iteratively assign the weights for observations i based on both its residual (weight w_i^r) and its location in predictor space (weight w_i^x; the further from the center of predictor space, the lower the weight i.e., the less importance). For latent variable model $y_i = \boldsymbol{t_i}\gamma + \epsilon_i$, the residual weights w_i^r are simply computed from $r_i = y_i - \boldsymbol{t}_i\gamma$ and the leverage points weights are computed from the scores $\boldsymbol{t}_i$. The model is called Partial Robust M-estimator (PRM). Robust centering ($\boldsymbol{x_c} = \boldsymbol{x} - \boldsymbol{median(x)}$) and standardization ($\boldsymbol{x}_{\boldsymbol{s}_i} = \frac{\boldsymbol{x_c}}{\sigma_{\mathrm{Q_n}}(\boldsymbol{x_c})}$) using multidimensional L1-median and $\mathrm{Q_n}$ estimator were used (Daszykowski et al., 2007).

The model was implemented using the TOMCAT toolbox (Daszykowski et al., 2007) in Matlab 7.13 (The MathWorks, Inc., Natick, Massachussets, United States). We tested the untransformed reflectance and found no substantial improvement either using this robust PLSR method (best standard PLSR R^2_{cv}=0.416, 0.662, and 0.760; best $nRMSE_{cv}$=16.2%, 31.1%, and 32.1%, respectively for LCC, LAI, and CCC), suggesting the problem with outliers was not serious.

Table 12. Results of partial robust M-regression with input untransformed reflectance. In the parentheses is standard deviation.

Grassland variable	$\boldsymbol{R^2_{cv}}$	$\boldsymbol{nRMSE_{cv}}$ **(%)**
LCC	0.465 (0.22)	15.62 (4.3)
LAI	0.695 (0.14)	29.90 (5.9)
CCC	0.743 (0.10)	33.83 (6.2)

K. Spectral characteristics of optical sensors used for simulation in this present study

Table 13. Band settings of the spectrally simulated optical sensors

No	Optical sensors (VNIR-SWIR)	Spectral characteristics
1	EnMAP (DLR)	228 (**199**[1]) bands: 98 bands in 420-1000 nm (6.5±1.25 nm width) + 130 bands in 900-2450 nm (10±2.5 nm width)
2	HyMap (HyVista)	126 (**110**[1]) bands in 436-2485 nm (band width 13-17 nm)
3	CHRIS (Proba-1) land channel	**37**[1] bands in 438-1003 nm (width 6-33 nm)
4	CHRIS (Proba-1) chlorophyll channel	**18**[1] bands in 486-788 nm (width 6-11 nm)
5	Worldview-3 MS & SWIR	**16**[1] bands: 8 VNIR bands (400-1040 nm) + 8 SWIR bands (1195-2365 nm), width 30-180 nm (see Table 5)
6	Sentinel-2 MSI	13 (**10**[1]) bands in 443-2190 nm (4 VNIR, 6 red-edge/SWIR, 3 atmospheric bands), width 20-180 nm (see Table 5)
7	Landsat-8 OLI	9 (**7**[1]) bands (1 coastal band, 1 cirrus band) in 430-1380 nm (width 15-190 nm) (see Table 5)

[1] Number of bands (bold) used for regression analysis excluding atmospheric-purpose bands, atmospheric absorption bands, and bands beyond the wavelength range of field hyperspectral measurement (GER 3700; 402.23-2400.35 nm)

Sources: see Table 9 (Appendix C); http://landsat.gsfc.nasa.gov/?p=5779; ESA Sentinel-2 Team (2010); http://www.satimagingcorp.com/satellite-sensors/worldview-3/; Cutler & Kellar-Bland (2008)

Table 14. Detail band settings of the simulated multispectral sensors

Worldview-3[3] MS & SWIR			Sentinel-2 MSI			Landsat-8 OLI		
Band	Center (nm)	Width (nm)	Band	Center (nm)	Width (nm)	Band	Center (nm)	Width (nm)
Coastal	425	50	B1[2]	443	20	B1 (CA)	442.96	15.98
Blue	480	60	B2	490	65	B2 (Blue)	482.04	60.04
Green	545	70	B3	560	35	B3 (Green)	561.41	57.33
Yellow	605	40				B8 (Pan)	589.5	172.4
Red	660	60	B4	665	30	B4 (Red)	654.59	37.47
Red Edge	725	40	B5	705	15			
			B6	740	15			
			B7	783	20			
NIR1	832.5	125	B8	842	115			
			B8a	865	20	B5 (NIR)	864.67	28.25
NIR2	950	180	B9	945	20			
SWIR-1	1210	30	B10	1375	30	B9 (Cirrus)	1373.43	20.39
SWIR-2	1570	40						
SWIR-3	1660	40	B11	1610	90	B6 (SWIR 1)	1608.86	84.72
SWIR-4	1730	40						
SWIR-5	2165	40						
SWIR-6	2205	40	B12	2190	180	B7 (SWIR 2)	2200.73	186.66
SWIR-7	2260	50						
SWIR-8	2330	70						

[2] shaded bands are atmospheric-purpose bands or bands in atmospheric absorption regions, which are removed from regression analysis. Panchromatic B8 Landsat-8 was excluded as the bandpass overlaps with other bands. [3] Worldview-3 band centers were approximated as the mid-point wavelength (upper wavelength – lower wavelength)

L. Random Forest regression applied to input of varying spectral resolution

Table 15. Random Forest regression applied to simulated reflectance data of varying spectral resolution. Atmospheric-purpose wavebands were excluded in #bands (except the field GER). $mtry$ is parameter number of bands randomly selected for tree split. For hyperspectral GER, EnMAP, and HyMap parameter number of tree $ntree$=5000 was used, for the other sensors $ntree$=1000 (bands<40).

Sensor (# bands)	R^2_{cv}	$nRMSE_{cv}$ **(%)**	$mtry$	$mtry$ **tested**
GER (584)	0.629 (0.14)	39.20 (10.3)	44 (82)	5,10,15,30,50,100,200
EnMAP (199)	0.646 (0.14)	38.52 (10.3)	5 (0)	5,10,15,30,50,100
HyMap (110)	0.649 (0.14)	38.40 (10.2)	5.5 (1.6)	5,10,15,30,50,75
CHRIS land (37)	0.631 (0.12)	38.54 (8.1)	3 (0)	3,5,10,15,20,30
CHRIS chl (18)	0.590 (0.14)	40.56 (10.5)	3.9 (1.4)	3,6,9,12,15
Worldview-3 (16)	0.651 (0.14)	38.18 (9.9)	3 (0)	3,6,9,12
Sentinel-2 (10)	0.641 (0.15)	38.31 (10.3)	3 (0)	3,5,7,9
Landsat-8 (7)	0.606 (0.13)	40.30 (8.9)	1.7 (1.1)	1,2,3,4,5