Multivariate statistical methods for estimating grassland leaf area index and chlorophyll content using hyperspectral measurements

Hadi[(1)] (hadi.hadi@aalto.fi), R. Darvishzadeh[(2)], A. K. Skidmore[(2)]

[(1)]Department of Real Estate, Planning, and Geoinformatics, Aalto University
Vaisalantie 8, PO Box 15800, 02130 Espoo, Finland
[(2)]Faculty of Geo-information Science and Earth Observation, University of Twente
Hengelosestraat 99, PO Box 217, 7500 AE Enschede, the Netherlands

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. The advent of hyperspectral remote sensing and the future launch of planned spaceborne hyperspectral missions will open up new possibilities over conventional multispectral RS to better quantify grassland characteristics. Hyperspectral data, while rich in information, presents a challenge for analysis due to its high dimensionality and multicollinearity. This present study investigated three promising high dimensional multivariate regression models namely partial least squares regression (PLSR), regularization and shrinkage method Lasso, and nonparametric Random Forest (RF) regression, to estimate 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/CCC (cross-validated R^2=0.760) and LAI (R^2=0.719), whereas leaf chlorophyll content/LCC could be predicted with relatively low accuracy (R^2=0.492). Multivariate regression models 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. The findings may suggest that the retrieval accuracy cannot be improved by solely the statistical modelling exercise. The prediction errors were likely the results of grassland canopy spectral complexity due to the presence of different grass species having different canopy architecture. Thus, approaches that explicitly account for the structural heterogeneity such as model stratification based on species, or the use of 3-D radiative transfer model suitable for heterogeneous canopy were recommended.

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, multispectral 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 (10-20 m).

Keywords: remote sensing, hyperspectral, grassland, LAI, chlorophyll, multivariate statistical methods