Fresh Biomass Estimation in Heterogeneous Grassland Using Hyperspectral Measurements and Multivariate Statistical Analysis

Roshanak Darvishzadeh[a], Andrew Skidmore[a], Mojgan Mirzaie[b], Clement Atzberger[c], Martin Schlerf [d]

[a] University of Twente, Faculty of ITC, Enschede, The Netherlands;
[b] University of Calgary, Department of Geography, Alberta, Canada;
[c] University of Natural Resources and Life Sciences, Vienna, Austria;
[d] CRP Gabriel Lippmann, Belvaux, Luxembourg

Abstract

Accurate estimation of grassland biomass at their peak productivity can provide crucial information regarding the functioning and productivity of the rangelands. Hyperspectral remote sensing has proved to be valuable for estimation of vegetation biophysical parameters such as biomass using different statistical techniques. However, in statistical analysis of hyperspectral data, multicollinearity is a common problem due to large amount of correlated hyper-spectral reflectance measurements. The aim of this study was to examine the prospect of above ground biomass estimation in a heterogeneous Mediterranean rangeland employing multivariate calibration methods. Canopy spectral measurements were made in the field using a GER 3700 spectroradiometer, along with concomitant in situ measurements of above ground biomass for 170 sample plots. Multivariate calibrations including partial least squares regression (PLSR), principal component regression (PCR), and Least-Squared Support Vector Machine (LS-SVM) were used to estimate the above ground biomass. The prediction accuracy of the multivariate calibration methods were assessed using cross validated R^2 and RMSE. The best model performance was obtained using LS_SVM and then PLSR both calibrated with first derivative reflectance dataset with R^2_{cv}=0.88 & 0.86 and $RMSE_{cv}$=1.15 & 1.07 respectively. The weakest prediction accuracy was appeared when PCR were used (R^2_{cv} = 0.31 and $RMSE_{cv}$= 2.48). The obtained results highlight the importance of multivariate calibration methods for biomass estimation when hyperspectral data are used.

Introduction

Accurate estimates of grass biomass can provide valuable information about the productivity and functioning of rangelands and grasslands and has been an important focus due to its impact on ecoysystem process and carbon cycles. Remote sensing offers a cost-effective solution for timely and accurate estimation of aboveground green biomass in grasslands from local to regional scale. In recent years, remote sensing techniques have been widely used to estimate aboveground green biomass in grasslands (Anderson, Hanson, & Haas, 1993; Mirik et al 2005; Moreau, Bosseno, Gu, & Baret, 2003; Wylie, Meyer, Tieszen, & Mannel, 2002). Development of hyperspectral sensors, which provide detailed spectral information at higher spectral resolution, has offered unprecedented opportunities to estimate grassland aboveground green biomass using new techniques, such as narrow-band vegetation indices, red-edge position (REP), band depth analysis and partial least square regression. Canopy reflectance is mainly determined by LAI and other properties (e.g. leaf angle, soil optical properties). While LAI shows a close relation to biomass, a good relation between canopy reflectance and biomass can be expected (Gnyp et al., 2014). High correlations between biomass and reflectance are shown in hyperspectral data collected by field spectrometer, airborne, and satellite sensors (Cho et al., 2007; Psomas, Kneubühler, Silvia, Itten, & Zimmermann, 2007). Grassland parameters are generally estimated using standard or multivariate

regression techniques with data derived from spectral reflectance measurements. Several studies have exploited this approach, relating field data with reflectance indices (Cho & Skidmore, 2006; Mutanga & Skidmore, 2004), with first derivative reflectance spectra (Lamb et al., 2002), and with continuum removed spectra (Mutanga, Skidmore, Kumar, & Ferwerda, 2005). These studies underline the complexity of the spectral response of mixed grasslands, especially in presence of a high fraction of non-photosynthetic vegetation (NPV) and exposed soil (Baldocchi et al., 2004), and canopy architecture complexity due to mixed species composition and phenology (Cho et al., 2007; Darvishzadeh et al., 2008c; Numata et al., 2008). This complexity makes the prediction of biophysical and biochemical pasture properties still challenging for the remote sensing community, especially in mixed species environments and keeping in account the temporal variability of vegetation properties during different stages of pasture growth. On the other hand, in statistical analysis of hyperspectral data, multicollinearity is a common problem due to large amount of correlated hyper-spectral reflectance measurements.
Therefore the aim of this study was to examine the prospect of above ground biomass estimation in a heterogeneous Mediterranean rangeland employing multivariate calibration methods.

Materials

The study site is located in Majella National Park, Italy (latitude 41°52' to 42°14' N, longitude 13°14' to 13° 50'E). The park covers an area of 74,095 ha and extends into the southern part of Abruzzo, at a distance of 40 km from the Adriatic Sea. The region is situated in the massifs of the Apennines. Stratified random sampling with clustering was adopted in this study. For this purpose, the area was stratified into grassland, forest, shrubland and bare rock out crops, using the land cover map provided by the management of Majella National Park. Coordinates (x y) were randomly generated in a grassland stratum to select plots. A total of 45 plots (30 m x 30 m) were generated and a GPS (Global Positioning System) was used to locate them in the field. To increase the number of samples in the time available, four to five randomly selected subplots were clustered within each plot. This resulted in a total of 171 subplots being sampled. The 1 m x 1 m subplots differed in species composition and relative abundance while the within-subplot variability was small. Within each sub plot, fresh above ground biomass of each species were measured. The subplot biomass was then calculated using the composition and cover percentage of it's species. Fifteen replicates of canopy spectral measurements were taken from each subplot, using a GER 3700 spectroradiometer (Geophysical and Environmental Research Corporation, Buffalo, New York). The 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. A moving Savitzky-Golay filter (Savitzky and Golay, 1964) with a frame size of 15 data points (2nd degree polynomial) was applied to the averaged reflectance spectra to further smooth the spectra. To remove spectral offset canopy reflectance was transformed to first derivative by calculating differences between adjacent spectra bands.

Methods

PLS and PC model Structure

Partial Least Square and Principal Component are two suitable models when hyper-spectral data is available. The model structure for both methods has the form:

$X = TP^t + E$

$Y = TQ + F$

Where *X* and *Y* are predictors and responses matrixes respectively, which are mean-centered by subtracting the mean of each variable from original measurements, *P* and *Q* are loading matrix that describe how the variables in *T* are related to original, and *E* and *F* are residuals.
The main different between PC and PLS models is that PLS uses factors determined by employing both X and Y in estimation directly. For each PLS regression each component is obtained by maximizing the covariance between y and all possible linear functions of X. Selecting the number of factors for PCR and PLSR was based on percent variant explained in each factors. Cross-validation is a more statistically sound method for choosing the number of components in either PLSR or PCR. It avoids over fitting data by not reusing the same data to both fit a model and to estimate prediction error. Therefore visual inspection of cross-validated Mean Squared Prediction Error (MSPEcv) values versus the number of factor plots was applied as another indicator for choosing the optimal number of factors in both PCR and PLSR.

Support Vector Machine

The idea of SVM have been developed by Vapnik (1995). Svms are known as learning machines that are based on statistical learning theory and minimizing generalization error bound to achieve generalized performance. There are two main categories for support vector machines: support vector classification (SVC) and support vector regression (SVR). Regression version of SVM as the most common application form of SVM has been proposed by Vapnik, Steven Golowich, and Alex Smola in 1997 and is known as support vector regression(SVR).
Least squares support vector machines (LS-SVM) is an alternative method of SVM introduced by Suykens et al. (2002). Ls-svm applied a set of linear equations instead of quadratic programming problem used in classical SVMs to simplify the optimization problem.
LS-SVM equation can be indicated as following equation:

$$f(x) = \sum_{i=1}^{N}(\alpha - \alpha^*)k(X, X_i) + b, \quad \alpha_i \geq 0 \quad , \alpha^* \leq \gamma$$

Where α. And α^* are the Lagrange multipliers, K(X,Xi) is the kernel function, b is the bias value and g is the regularization parameter, determining the trade-off between the fitting error minimization and smoothness of the estimated function. The cross-validation process was employed for finding suitable parameters values. LS-SVR calculations were performed using the MATLAB functions in the free LSSVM toolbox v1.7

Results and discussion

A key step in performing PCR and PLSR is selecting the optimal number of factors or latent variables. Figure (a & b) illustrated cumulative variance explained in reflectance matrix and cross validated MSPE of each factors for original reflectance (not shown for first derivative reflectance). It is clear from Figure (a) that first five PCs include more than 99.65 percent of variation of reflectance matrix, and in Figure b by adding the sixth PC the model error (MSPEcv) started to increasing that indicated the model by adding this factor was overfitted, so 5 factors recognized suitable for PCR.
This figure also indicate that the first 8 PLS factors accounted for 90 percent of the variation in the reflectance matrix and Figure (b) illustrates a rise in cross validated MSPE by adding new factors. It is clear that by adding the 9th factor there are an increase in prediction error of the PLS model. Therefore the model developed with the first eight PLS factor was found to perform the best performance.

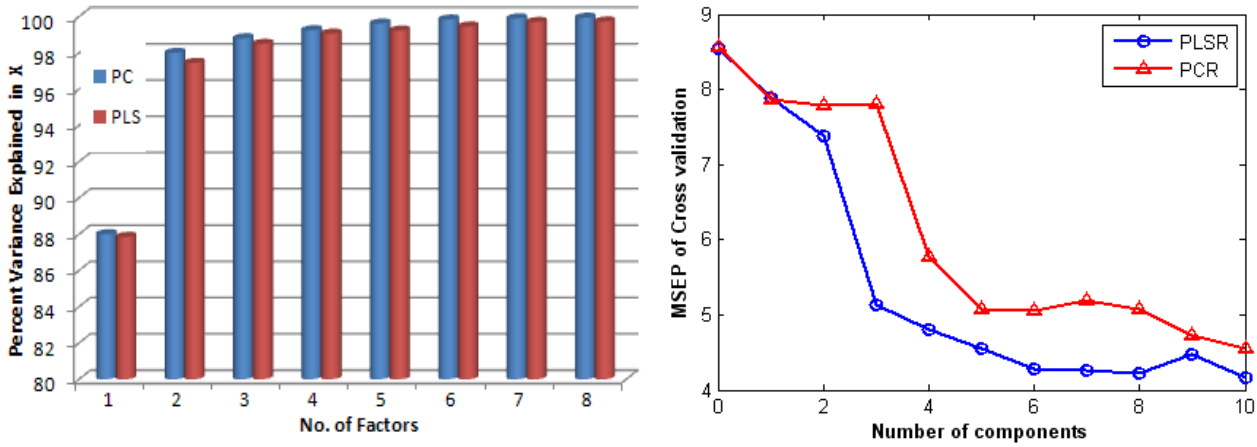


Figure 1. Cumulative variance explained in reflectance matrix for PLS and PCR (a) and cross validated MSPE of each factors for original reflectance (b), used for selection of optimal number of factors or latent variables

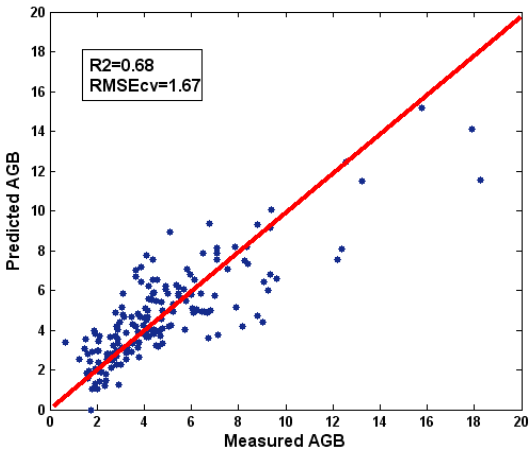
The statistics of PLSR and PCR performance is summarized in Table . The regression results revealed that the PLS model applying 8 factors yield better prediction of AGB compared to PCR. The PC model utilizing 5 principal component didn't achieved reasonable prediction of AGB applying both reflectance matrices ($R^2_{original}$ = 0.40 & $R^2_{1th\ derivative}$ =0.31). While performance of PLS models considerably was better than PC modes ($R^2_{original}$ = 0.68 & $R^2_{1th\ derivative}$ =0.86). The prediction accuracy based on first derivative reflectance significantly was improved in PLSR compared to original reflectance while using the derivative reflectance had negative influence in PCR performance. In terms of cross validated RMSE, the best model was PLSR using derivative reflectance with $RMSE_{CV}$=1.07 and PLSR $_{original}$, PCR $_{original}$ and PCR $_{derivative}$ were in the next orders with $RMSE_{CV}$= 1.67, 2.25 and 2.41 respectively.

From the result of both models performance it is obviously clear that the type of data comparison techniques significantly influence the prediction accuracy of the models. The PLS technique by explaining variation in both predictor matrix (reflectance) and response vector (AGB measurements) showed better result compare to PC model that only summarize variation in predictor matrix. **Error! Reference source not found.** show the predicted AGB versus measured AGB applying PCR and PLSR models using the original and 1st derivative reflectance.

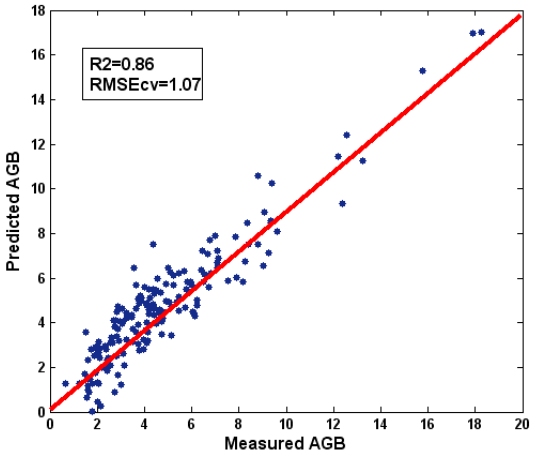
Table 1. Summary statistics of PLSR, PCR and LS-SVM performances

Regression model	Type of input	R^2_{cv}	$RMSE_{cv}$	Regression Parameters
LS-SVM	Original Wavelengths	0.66	1.70	gam=65.49 sig2=4084
	1th derivative	**0.88**	1.15	
	5 PCs	0.71	1.57	gam=10.66 sig2=12.14
PLSR	Original Wavelengths	0.68	1.67	no. Of factor =8
	1th derivative	**0.86**	1.07	no. Of factor=8
PCR	Original Wavelengths	0.40	2.25	no. Of factor =5
	1th derivative	0.31	2.41	no. Of factor=5

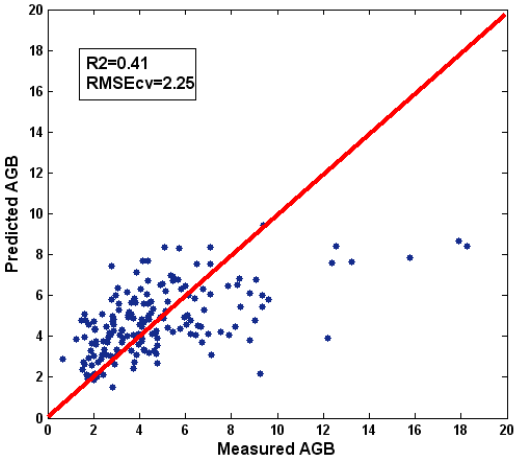
(a): PLSR (original reflectance)



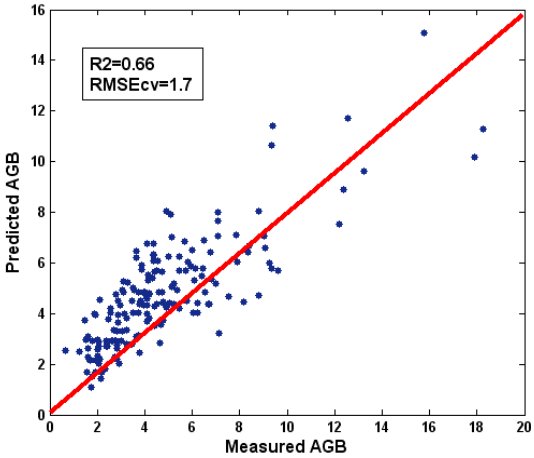
(b):PLSR(1th derivative reflectance)



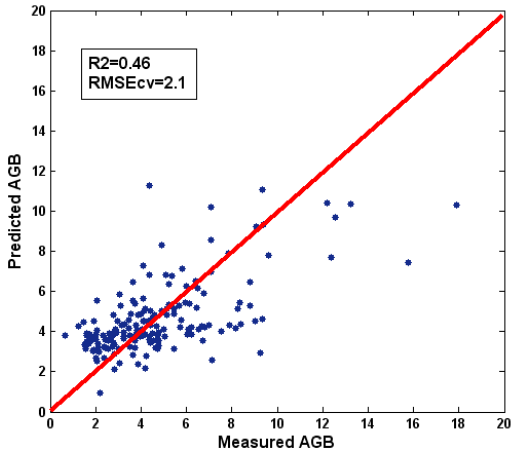
(c):PCR (Original reflectance)



(d):LS-SVM (Original reflectance)



(e): LS-SVM (1th derivative reflectance)



(f): PC-LSSVM (first five factors)

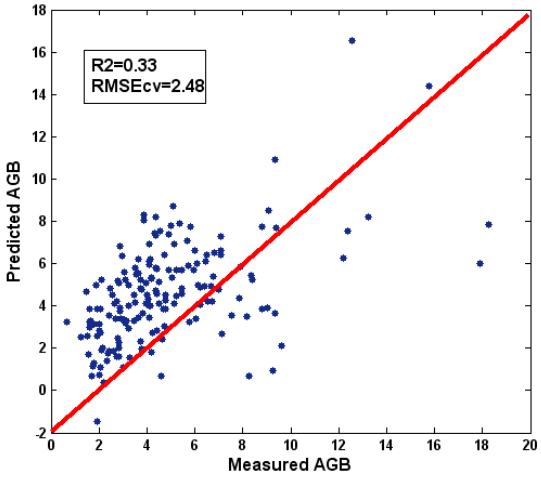


Figure 2. Measured and predicted Biomass using original and first derivative reflectance in PLSR, PCR, and LS-SVM

LS-SVM / PC-LSSVM performance for AGB prediction

The LS_SVM regression was trained using different input data .first SL-SVM was derived using Original wavelengths and first derivative spectra. As statistical regressions, using hyper-spectral wavelengths, often suffer from large dimension of input data, while numerous of them seem to be irrelevant, we applied 5 first PC components derived from PC model to minimize the dimension of independent variable (545 wavelengths) in LS_SVM and then PC -LSSVM have been used to predict the AGB and the performance was compared to other models used in the present study. The criteria for selecting the first five PC factors were explained in pervious section and Figure . In each training the tuning gamma and sig2 was computed from cross validation process. Table depicted a comparative analysis of the prediction accuracy obtained by different trained LS_SVM regressions. It is clear that data comparison technique in order to decrease the dimension of regression inputs was effective in model performance. Prediction of AGB versus measured are illustrated in figure 2 (e &f). From this figure it is clear that measured and prediction samples are distributed along 1 to 1 line especially for LS-SVM using derivative reflectance. As can be observed from Table using spectral transformation techniques (especially derivative transformation) have improved the prediction of AGB compared to the result of LS-SVM model with original wavelength. Cross validated $R^2$and RMSE for SL_SVM increased from 0.66 and 1.7 to 0.88 and 1.15 using the 1^{st} derivative reflectance, respectively.

Conclusions

This study has applied several multivariate regression techniques to predict biomass in heterogeneous Mediterranean grasslands. Validation of the models was done by comparing differences in the coefficient of determination (R^2) and relative root mean square error (RMSE) through cross-validation. In summary, multivariate calibration methods, such as partial least squares regression which previously had provided enhanced estimates of LAI and canopy chlorophyll content in heterogeneous grassland, at the field level, is shown to be important for remote sensing of grasslands. These multivariate methods demonstrated the ability to enhance estimates of different grass variables, and thus present new prospects for mapping and monitoring heterogeneous grass canopies from air- and space-borne platforms.

References

Anderson, G. L., Hanson, J. D., & Haas, R. H. (1993). Evaluating landsat thematic mapper derived vegetation indices for estimating above-ground biomass on semiarid rangelands. *Remote Sensing of Environment*, *45*(2), 165–175.

Baldocchi, D. D., L. K. Xu, and N. Kiang (2004), How plant functionaltype, weather, seasonal drought, and soil physical properties alter water and energy fluxes of an oak-grass savanna and an annual grassland, Agric. Forest Meteorol., 123(1–2), 13– 39.

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.

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

Gnyp, M. L., Bareth, G., Li, F., Lenz-Wiedemann, V. I. S., Koppe, W., Miao, Y., … Zhang, F. (2014). Development and implementation of a multiscale biomass model using hyperspectral vegetation indices for winter wheat in the North China Plain. *International Journal of Applied Earth Observation and Geoinformation*, *33*, 232–242. doi:10.1016/j.jag.2014.05.006

Lamb, D. W., Steyn-Ross, M., Schaare, P., Hanna, M. M., Silvester, W., & Steyn-Ross, A. (2002). Estimating leaf nitrogen concentration in ryegrass (Lolium spp.) pasture using the chlorophyll red-edge: theoretical modelling and experimental observations. *International Journal of Remote Sensing*, *23*(18), 3619–3648.

Mirik, M., Jack E. Norland, Robert L. Crabtree, Mario E. Biondini (2005). Hyperspectral One-Meter-Resolution Remote Sensing in Yellowstone National Park, Wyoming: I. Forage Nutritional Values, Rangeland Ecology & Management,58(5):452-458. DOI: 10.2111/04-17.1

Moreau, S., Bosseno, R., Gu, X. F., & Baret, F. (2003). Assessing the biomass dynamics of Andean bofedal and totora high-protein wetland grasses from NOAA/AVHRR. *Remote Sensing of Environment*, *85*(4), 516–529.

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.

Mutanga, O., Skidmore, A. K., Kumar, L., & Ferwerda, J. (2005). Estimating tropical pasture quality at canopy level using band depth analysis with continuum removal in the visible domain. *International Journal of Remote Sensing*, *V26*(6), 1093–1108.

Psomas, A., Kneubühler, M., Silvia, H., Itten, K., & Zimmermann, N. (2007). Continuum removal in multi-temporal spectrometer data enhances the classification accuracy of grasslands along a drymesic gradient. *Remote Sensing of Environment*, 2007.

Savitzky, A., & Golay, M. (1964). Smoothing and differentiation of data by simplified least square procedure.

Suykens,J.A., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle,J. (2002), Least squares support vector machines, World Scientific.

Vapink, V.P. (1995). The Nature of Statistical Learning Theory. Springer, NewYork

Wylie, B. K., Meyer, D. J., Tieszen, L. L., & Mannel, S. (2002). Satellite mapping of surface biophysical parameters at the biome scale over the North American grasslands a case study. *Remote Sensing of Environment*, *79*(2-3), 266–278.