Comparative analysis of three regression methods for the winter wheat biomass estimation using hyperspectral measurements

Yuanyuan Fu[1, 2, 3], Guijun Yang[1, 2], Haikuan Feng[1, 2], Xiaoyu Song [1, 2,] Xingang Xu[1, 2], Jihua Wang[1, 2, 3*]
1. Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China;
2. National Engineering Research Center for Information Technology in Agriculture, Beijing 100097.
3. Institute of Applied Remote Sensing & Information Technology, Zhejiang University, Hangzhou 310029, China
*Corresponding author E-mail: wangjh@nercita.org.cn

***Abstract*—Hyperspectal data contain more useful information for characterizing vegetation biomass, compared with multi-spectral data. However, to make full use of the hyperspectral data, the strong multi-collinearity in the data is supposed to be taken into account. With this study we evaluated three multivariate regression methods which are principal component regression (PCR), partial least square regression (PLSR) and stepwise multiple linear regression (SMLR). They are specifically designed to deal with multi-collinearity problem. Furthermore, to identify reliable winter wheat biomass predictive models different types of spectral transformations (continuum removal, first derivative) were combined with the three regression methods, respectively. The comparative analysis was conducted on the data sets collected in 2008 and 2009 field campaigns in Tongzhou and Shunyi district, Beijing, China. Compared with the other combination, the respective combination of three regression methods and continuum removal got the highest estimation accuracy, especially, the combination of PLSR and continuum removal (R^2=0.715, RMSE=0.218kg/m^2). The experimental results demonstrated that the use of PLSR is recommended for highly multi-collinear data sets. The combination of continuum removal and PLSR could improve the estimation accuracy of winter wheat biomass.**

Keywords-winter wheat biomass;hyperspectral;partial least squares regression;principal component regression; stepwise multiple linear regression;spectral transformation

I. INTRODUCTION

Crop aboveground biomass (below referred to as biomass) is an important indicator to reflect crop growth condition. For efficient farmland management, farmers need crop biomass information at early growth stages for guiding the fertilizer supply within fields in order to achieve optimal growth. Towards the end of the growing stage, such information is required for an early yield prediction [1]. The development of hyperspectral sensors, which offer an approximately contiguous spectrum defined by a large number of spectral bands [2], open a new perspective for quantifying physical attributes of vegetation. The vegetation biomass estimation is mainly based on the relationships between the spectral data or their transformations and biomass. In order to take the strong collinearity of spectral data into account, principal component regression (PCR), partial least square regression (PLSR) and stepwise multiple linear regression (SMLR) are widely used to estimate vegetation biophysical variables[3-5] or leaf pigmentation[6-8]. To enhance the features of vegetation spectra, different types of spectral transformations have been proposed, such as first derivative, continuum removal [9]. Research that systematically addresses the combination of the three regression methods and these spectral transformations on their performance in estimating vegetation biomass from hyperspectral remote sensing data is rare.

The objective of the study was to evaluate the performance of the respective combination of three techniques (PCR, PLSR and SMLR) and three spectral transformations (untransformed spectra, first derivative and continuum removal) for modeling the vegetation biomass from field spectrometer data.

II. MATERIAL

A. The study area

The study site is located in Tongzhou district(latitude 39°36’ to 40°02’N, longitude 116°32′ to 116°56′ E) and Shunyi district(latitude 40°00' to 40°18'N, longitude 116°28' to 116°58' E), Beijing, China. Tongzhou district covers an area of 907 square kilometers. This district belongs to continental monsoon climate zone and is influenced by winter and summer monsoon. The average annual temperature is 11.3℃ and average annual precipitation is approximately 620mm. Shunyi district covers an area of 1020 square kilometers. This district has a warm temperate semi-humid continental monsoon climate. The average annual temperature is 11.5℃ and average annual precipitation is approximately 625mm with 75% distributed in summer.

The main cultivars of winter wheat planted are Nongda 211 (erectophile), Zhongyou 206(middle), Jingdong 8 (middle) and Jing 9428 (planophile) in the two districts. Farmlands are decentralized and each farmland is planted by different farmers. It results in difference to some extent in planting cultivar, planting density and farmland management mode.

B. Field data collection

The campaigns were carried out during the typical winter wheat growing season of 2008 and 2009. For the 2008 campaign, the actual dates were from 27 March to 13 May. For the 2009 campaign, the actual dates were from 1 April to 18 May. In every campaign, farmland region in

which winter wheat grew uniformly was selected as a plot for canopy spectral measurements. All canopy plots of 0.5 m × 0.5 m were selected randomly. A total of 108 plots and 102 plots were generated in 2008 and 2009, respectively.

A high spectral resolution spectrometer, ASD FieldSpec Pro spectrometer (Analytical Spectral Devices, Boulder, CO, USA) was used to take in-situ canopy spectral reflectance. The ASD spectroradiometer covers the range from 350 nm to 2500 nm with a spectral sampling of 1.4 nm in the 350-1000 nm range, 2 nm in the 1000-2500 nm. The results were then interpolated by the ASD software to produce readings at every 1 nm. To minimize atmospheric perturbations and BRDF effects, all canopy spectral measurements were taken on a clear day with no visible cloud cover between 10:00 a.m. and 14:00 p.m. (Beijing Local Time). The sensor, with a field of view 25°, was handheld approximately 1.3m above ground (the height of the wheat is 90±3cm at maturity) at the nadir position. The 10 replicate spectral measurements taken from each plot enabled us to suppress much of the measurement noise by averaging the replicate measurements. The spectral regions with wavelength of 400-1350 nm with high signal to noise ratio were used for the analysis. To further minimize noise in the measured reflectance spectra. A moving Savitzky-Golay filter [10] with a frame size of 15 data points (2^{nd} degree polynomial) was applied to the averaged reflectance spectra to denoise the spectra. Prior to each reflectance measurement, the radiance of a white standard panel coated with 40 cm × 40 cm $BaSO_4$ with a known reflectivity was recorded for normalization of the target measurements.

The aboveground biomass was clipped in each plot after ASD measurements and taken back to obtain dry weight using traditional agronomic methods. Aboveground biomass was determined by dividing the weight of the dry winter wheat by the surface area of the plot.

III. METHODS

A. General workflow

The general workflow of the study is showed in Fig.1.The workflow mainly consists of three parts. The first part is field data collection (the collection of canopy spectral data and the corresponding winter wheat biomass). The second part is spectral transformation which includes first derivative and continuum removal. For comparison, original (untransformed) data (Ori) were also analyzed. The third part is regression analysis which contains three methods. They are principal component regression (PCR), partial least square regression (PLSR), and stepwise multiple linear regression (SMLR). The three transformations were combined with the above three regression methods, respectively.

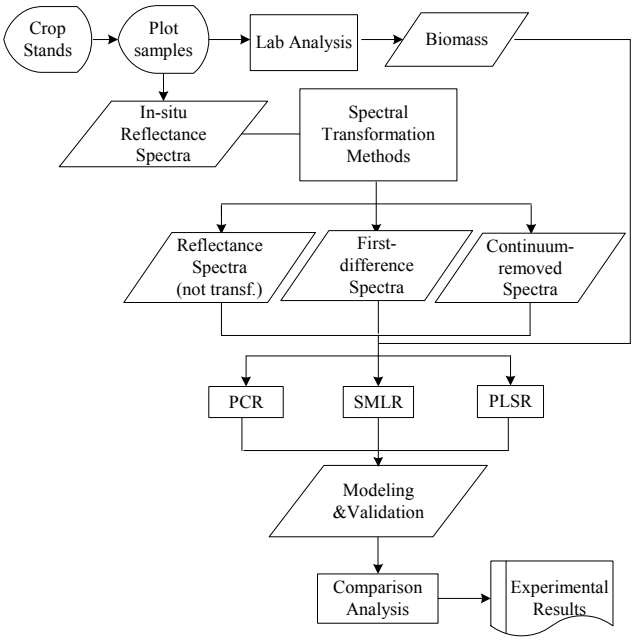

Figure 1. The general workflow of the research.

B. Principal component regression(PCR)

The basis of principal component regression (PCR) is principal component analysis (PCA) which is effective for compressing multi-dimensional data. The main objective of the PCA is to find a lower dimensional representation that can account for most of the variance of the spectral covariance matrix of a dataset [11]. Based on PCA, the main spectral variation is described by several orthogonal regression factors. These factors are then used to establish estimation model.

C. Partial least squares regression(PLSR)

Partial least squares regression (PLSR) is a technique that reduces the large number of measured collinear spectral variables to a few non-correlated latent variables or factors [12]. The factors represent the relevant information in the measured canopy spectral reflectance and are used to predict the dependent variable (here, winter wheat biomass). Although PLSR is closely related to PCR, PLSR actually uses the response variable information during the decomposition process that is different with PCR [13].

D. Stepwise multiple linear regression(SMLR)

A primary hypothesis in stepwise multiple linear regression is that only a subset of all available wavelengths have a significant explanatory effect on the response variable [14].SMLR starts with no predictors (wavelengths) in the regression equation, and at each step it adds the most statistically significant wavelengths (highest F-value or lowest p-value) and remove insignificant wavelengths (lowest F-value or highest p-value) [15].This procedure will stop until no further entry or removal of wavelengths. In this study, the p-values to enter and remove wavelengths were set

at 0.05 and 0.1. The selected wavelengths were used to establish linear regression model.

E. Spectral transformation

Towards to discrete spectral data, first derivative is approximated with first difference by calculating differences in reflectance between adjacent wavelengths. Compared to original spectral data, first derivative spectra are more insensitive to variation in soil background, illumination and so on [16].

The main purpose of continuum-removal is the minimization of effects that extraneous factors may have on reflectance spectra to highlight absorption features. The continuum initially defines a convex hull over the top of a spectrum utilizing straight-line segments that connect local spectral maxima. The continuum-removed reflectance at a certain wavelength is calculated by dividing the original reflectance value by the values of the continuum line at the corresponding wavelength.

F. Validation

The samples collected in 2009 were used calibration data set, and the samples collected in 2008 were used as independent test data set. Regression analyses were performed on the training data set. Empirical validation of the regression models were carried out using the test data set. Using a completely different and independent test data to test the performance of model is important in determining a model's long-term stability. The performances of the various regression models were compared using the coefficient of determination (R^2), root mean square error (RMSE, equation (1)) and mean absolute error (MAE, equation (2)).

$$RMSE = \sqrt{\sum_{i=1}^{n}(Y_{est}-Y_{mea})^2 \Big/ n} \quad (1)$$

$$MAE = \sum_{i=1}^{n}\left|Y_{est}-Y_{mea}\right| \Big/ n \quad (2)$$

where Y_{est} is the estimated biomass; Y_{mea} is the measured biomass, and n is the number of sample.

IV. RESULTS AND DISCUSSION

Before conducting PCR and PLSR, the data including spectral data and measured biomass data were mean-centered. For PCR, the number of component used in model was decided according to the percentage of accumulated variance. In this study, the threshold of this percentage was set to 99%. For PLSR, the leave-one-out cross-validation method was used to select the optimal number of factors to be included in the regression model. To prevent collinearity and to preserve model parsimony, the condition for adding an extra factor to the model was that it had to reduce the root mean square error of cross-validation (RMSECV) by>2% [17] .The RMSECV was determined from the residuals of each cross-validation phase.

The performance of models based on respective combination of the three investigated regression methods and spectral transformations was shown in table Ⅰ. As shown in table Ⅰ ,among the models based on the respective combination of three spectral transformations and PLSR, the model based on the combination of PLSR and continuum removal reached the highest estimation accuracy. The model based on the combination of SMLR and continuum removal achieved the intermediate estimation accuracy. The model based on the combination of PCR and continuum removal got the lowest result. Generally compared with the models based on respective combination of the three regression methods and continuum removal, the models based on respective combination of the three regression methods and first derivative got lower estimation accuracy. The models based on the respective combination of the three regression methods and untransformed spectra which worked as baseline got the lowest estimation accuracy. But the model based on the combination of SMLR and FD did not improve the estimation accuracy, compared with the model based on the combination of SMLR and Ori. The main reason for it is that SMLR is more sensitive to noise in spectra than the other two regression methods, while the first derivative of spectral data is also sensitive to noise in spectra. The combination of SMLR and FD will largely affected by the noise in spectral data, So the model based on the combination of the two got the much lower estimation accuracy.

TABLE I. PERFORMANCE OF THE COMBINATION OF THREE REGRESSION METHODS AND SPECTRAL TRANSFORMATIONS FOR PREDICTING WINTER WHEAT BIOMASS

Method	Calibration			Independent validation		
	R^2	RMSE (kg/m^2)	MAE (kg/m^2)	R^2	RMSE (kg/m^2)	MAE (kg/m^2)
Ori+PCR	0.703	0.175	0.132	0.415	0.283	0.209
FD+PCR	0.734	0.165	0.125	0.644	0.239	0.167
CR+PCR	0.808	0.141	0.105	0.615	0.237	0.163
Ori+PLSR	0.755	0.159	0.123	0.599	0.233	0.173
FD+PLSR	0.774	0.152	0.114	0.586	0.248	0.177
CR+PLSR	0.843	0.127	0.092	0.715	0.218	0.150
Ori+SMLR	0.853	0.123	0.093	0.695	0.294	0.234
FD+SMLR	0.867	0.117	0.086	0.539	0.327	0.238
CR+SMLR	0.858	0.121	0.089	0.634	0.243	0.170

According to the experimental results, it was obvious that continuum removal of spectra was more suitable for establishing winter wheat biomass estimation model than first derivative. First derivative was better than untransformed spectra. Some literature pointed that there is deepening and widening of the red absorption pit with an increase in vegetation biomass [18, 19]. The continuum removal and first derivative enhanced the absorption feature. So the models based on the respective combination of the two transformations and the three regression methods got higher estimation accuracy than the models based on respective combination of untransformed spectra and the three regression methods. Amongst the three regression

methods, PLSR performed best. SMLR performed somewhat better than PCR. It could come to a conclusion that PLSR was useful to deal with multi-collinearity of hyperspectral data and a good choice to establish winter wheat biomass model.

V. CONCLUSION

To make full use of hyperspectral data and establish reliable winter wheat biomass estimation model, the present study has compared and analyzed the predictive performance of the models based on the respective combination of the three multivariate regression methods (PCR, PLSR, and SMLR) and the three spectral transformations (untransformed spectra, first derivative, and continuum removal). The most important conclusions that can be drawn from this study are as follows:

(1). The continuum removal and first derivative of spectra could enhance the absorption feature. The two spectral transformations could reflect more information than original spectra, especially the continuum removal.

(2). Amongst the three multivariate regression methods, PLSR was recommended to deal with the multi-collinearity problem of hyperspectral data.

(3). The model based on the combination of the continuum removal and PLSR got the highest estimation accuracy for winter wheat biomass.

ACKNOWLEDGMENT

This work was supported in part by the State Key Basic Research and Development Program (2011CB311806), the National Natural Science Foundation of China (41071228, 41271345, 41001244), the Beijing Municipal Natural Science Foundation (4102021, 4112022), the Beijing Municipal Talents Training Funded Project (2012D002020000007),the Special Funds for Technology innovation capacity building sponsored by the Beijing Academy of Agriculture and Forestry Sciences (KJCX201104012), and the State Key Laboratory of Remote Sensing Science sponsored by the Institute of Remote Sensing Applications of Chinese Academy of Sciences, through its open funds(OFSLRSS201109).The authors also extend gratitude to Mr. Weiguo Li and Mrs. Hong Chang for data collection.

REFERENCES

[1] J. G. P. W. Clevers, G. W. A. M. Van der Heijden, S. Verzakov, and M. E. Schaepman , “Estimating grassland biomass using SVM band shaving of hyperspectral data,”Data Photogrammetric Engineering & Remote Sensing ,vol.73, no. 10, pp.1141-1148,2007.

[2] J. Xing, S. Symons, M. Shahin, D. Hatcher, “Detection of sprout damage in Canada Western Red Spring wheat with multiple wavebands using visible/near-infrared hyperspectral imaging,”.Biosystems Engineering, vol.106, no.2,pp.188-194, 2010.

[3] H. T. Nguyen, B. W. Lee, “Assessment of rice leaf growth and nitrogen status by hyperspectal canopy reflectance and partial least square regression ,”European Journal of Agronomy, vol.24, no. 4, pp.349-356, 2006.

[4] G. El-Masry, N. Wang, A. El-Sayed, M. Ngadi, “Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry ,”Journal of Food Engineering, vol.81,no.1, pp.98-107, 2007.

[5] C. Atzberger, M. Guerif, B.Frederic, W. Werner, “Comparative analysis of three chemometric techniques for the specroradiometric assessment of canopy chlorophyll content in winter wheat,”.Computers and Electronics in Agriculture vol.73, pp.165-173,2010.

[6] P. M. Hansen, J. K .Schjoerring, “Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression,”. Remote Sensing of Environment, vol.86, pp.542-553,2003.

[7] E. Naesset, O. M .Bollandsas, T.Gobakken, “Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data,”. Remote Sensing of Environment, vol.94, no.4, pp.541-553,2005.

[8] T. Jensen, A. Apan, F. Young, L.Zeller, “Detecting the attributes of a wheat crop using digital imagery acquired from a low-altitude platform,”Computers and Electronics in Agriculture , vol.59, pp.66-77,2007.

[9] R.F.Kokaly, C.A.Hlavka, D.L.Peterson, “Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression,” Remote Sensing of Environment, vol.67,pp.267-287,1994.

[10] A. Savitzky, M. J. E. Golay, “Smoothing and differentiation of data by simplified least square procedure,”Analytical Chemistry, vol.36, no.8, pp.1627-1638, 1964.

[11] Z.Y.Liu , H.F.,Wu , J.F.,Huang, “Application of neural networks to discriminate fungal infection levels in rice panicles using hyperspectral reflectance and principal components analysis,” Computers and Electronics in Agriculture, vol.72, pp.99-106,2010.

[12] C. Atzberger, T. Jarmer, M. Schlerf, B. Kotz, W. Werner, “Spectroradiometric determination of wheat bio-physical variables: comparison of different empirical-statistical approaches,” Remote Ssinsing in Transitions, Proc. 23rd EARSeL symposium, Belgium, pp.463-470 ,2003.

[13] P. Geladi, B. R. Kowalski, “Partial least-squares regression: a tutorial ,”. Anal. Chim. Acta, vol.185, pp.1-17, 1986.

[14] Mathworks,2007. Matlab, The language of Technical Computing. Mathworks Inc., USA.

[15] R.Darvishzadeh, A. Skidmore, M. Schlerf, C.Atzberger, F.Corsi,M. Cho, “ LAI and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements,”Journal of Photogrammetry & Remote Sensing,vol.63,pp.409-426,2008.

[16] M.A.Cho , A.K.Skidmore , C.Atzberger , “ Towards red-edge positions less sensitive to canopy biophysical parameters for leaf chlorophyll estimation using properties optique spectrales des fuilles (PROSPECT) and scattering by arbitrarily inclined leaves (SAILH) simulated data,” IJRS, vol.29, no. 8, pp.2241-2255,2008.

[17] L. Kooistra, E. A. L. Salas, J. G. P. W. Clevers, R. Wehrens, “Exploring field vegetation reflectance as an indicator of soil contamination in river floodplains,”Environmental Pollution, vol.127, pp.281-290, 2004.

[18] S. W. Todd, R. M. Hoffer, D. G. Milchunas, “Biomass estimation on grazed and ungrazed rangelands using spectral indices ,” International Journal of Remote Sensing, vol.19, no.3, pp.427-438, 1998.

[19] J.G.P.W.Clevers, R.Jongschaap, “Imaging spectrometry for agriculture applications,” Imaging Spectrometry: Basic Principles and Prospective Applications. Kluwer Academic, Dordrecht, The Netherlands, pp.157-199, 2001.