Advances in Engineering Research, volume 206
Proceedings of the Environmental Science and Technology International Conference (ESTIC 2021)

Application of Random Forest Approach to Biomass Estimation Using Remotely Sensed Data

Nyamjargal Erdenebaatar[1,*], Batbileg Bayaraa[2], Amarsaikhan Damdinsuren[1]

[1]*Institute of Geography and Geoecology, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia*
[2]*Mongolian University of Life Science, Ulaanbaatar, Mongolia*
[*]*Corresponding author. Email: nyamjargale@mas.ac.mn*

ABSTRACT
The aim of this study is to investigate the application of RS-based vegetation indices to biomass estimation, perform a random forest (RF) classification for estimating biomass and compare the performance of RF method for high resolution and medium resolution images. As data sources, orthorectified QuickBird (QB) and Landsat 8 images acquired over Bornuur soum of Tov province, Mongolia are used. Firstly, the spectral indices were calculated for both images and the correlation between field measured biomass and spectral indices was estimated using partial least square regression. Then, the RF classification was performed to estimate the biomass. For all vegetation indices, VARVI yielded the highest correlation coefficient value for the Landsat data, while SR was considered the highest correlated index for the QB data. For both imageries, G-RVI and VARI were the best vegetation indices to explain the ground biomass. The relationship between the measured biomass and QB derived vegetation indices resulted in an r^2 value of 0.337 and RMSE=83.435 g/m^2, while the vegetation indices from Landsat performed relatively well in predicting the groundcover with a r^2 value of 0.617 and RMSE=50.881 g/m^2. This could be explained by the fact that high spatial resolution images have lots of shadows from trees and terrain, resulting in errors for AGB estimation.

Keywords: *Biomass, RF classification, Landsat, QuickBird*

1. INTRODUCTION

Generally, remote sensing (RS)-based above ground biomass (AGB) estimates use three types of remotely sensed data, namely, optical, radar, and LiDAR images, depending on the spatial resolution required and the application purposes. Optical RS datasets are the most widely available type of data and commonly used optical data include low-resolution AVHRR and MODIS [1], medium-resolution Landsat and SPOT data [2; 3] and high-resolution IKONOS, Quickbird (QB), Worldview, and drone data. High spatial resolution images have a significant number of shadows from trees and terrain, resulting in errors for AGB estimation [4]. In contrast, medium-resolution Landsat (30 m) data are widely used in combination with sample plot data for AGB estimations because of their free of charge availability.

Biomass is defined as the dry weight of both AGB and belowground biomass living mass of vegetation, such as wood, bark, branches, twigs, stumps, or roots as well as dead mass of litter associated with the soil [5]. Therefore, it can be considered as a measure of objects' structure and function. Thus, by knowing the spatial distribution of biomass, it is possible to calculate the new flow of terrestrial carbon, nutrient cycling, forest productivity, biomass energy, and carbon storage and sequestration by the forest, reducing the uncertainty of carbon emission and sequestration measures to support climate change modelling studies [6].

In the past 20 years, a great number of studies aiming at AGB estimation using RS data have been published. One of the most common methods to estimate biomass is the regression analysis, which is a statistical technique to investigate and model the relationship between dependent and independent variables. Some estimation methods have been established as a nonparametric alternative to the use of regression approaches for biomass modeling: k-nearest neighbor, artificial neural network, regression tree, random forest, support vector machine, and

maximum entropy [7]. Regression tree and random forest (RF) are a family of tree-based models; in the first one, data are stratified into homogeneous subsets by decreasing the within-class entropy, whereas in the second one, a large number of regression trees are constructed by selecting random bootstrap samples from the discrete or continuous datasets. In fact, the RF algorithm is now widely used for biomass estimation [8; 9]. It is efficient in dealing with large input datasets while requiring few parameters [10].

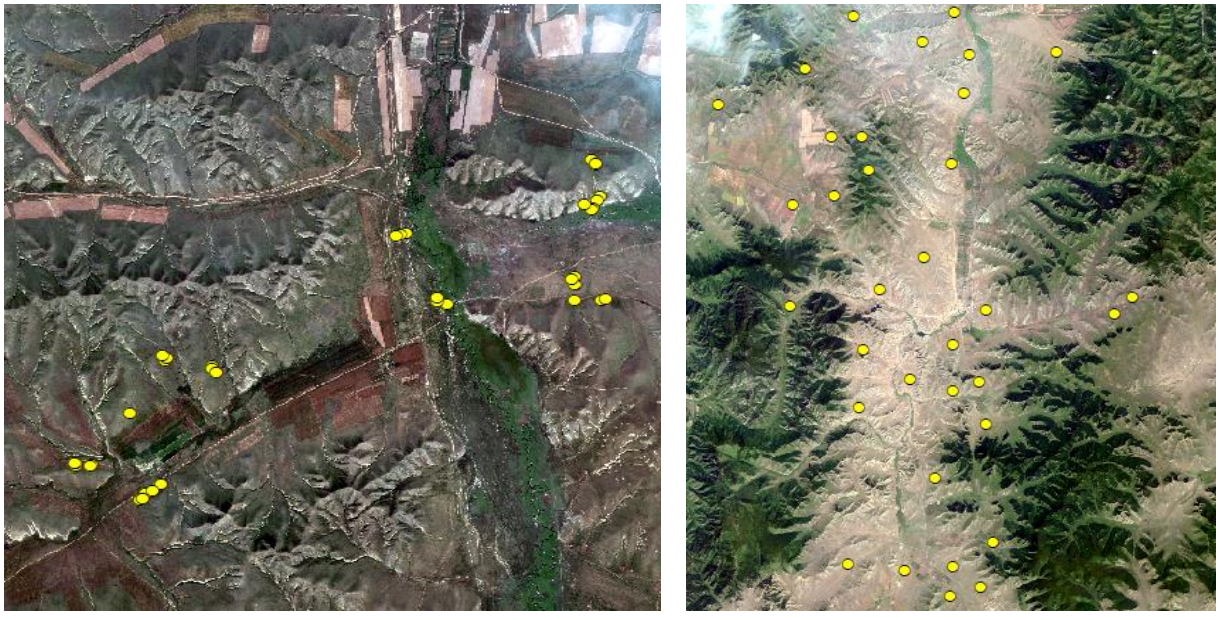
The main objective of this study is to investigate the applicability of RS-based vegetation indices to biomass estimation. For this aim, we applied a RF classification for estimating biomass and compared the performance of RF for high resolution and medium resolution images.

2. STUDY AREA AND DATA

The study was conducted over an area about 2420 square.km located in Bornuur soum of Tov province, Mongolia. The area belongs to a forest-steppe zone and has an elevation of 1000-1500 m above sea level. In terms of physical geography, the top of the ridge is included in the Khentii Mountains. The mean annual precipitation is between 250 and 350 mm and a mean annual temperature in July is +20°C, while it is -30°C in January (http://ldi.nda.gov.mn/).

We selected an orthorectified QB and Landsat 8 images acquired in August of 2009 and 2014 to estimate AGB. The Landsat 8 image was downloaded from the U.S. Geological Survey Earth Explorer web site (https://earthexplorer.usgs.gov/). Using the metadata associated with both of satellite imagery, radiometric correction was applied to convert digital numbers into reflectance and mitigate the impact of scene illumination and viewing geometry. Dark object subtraction was applied for atmosphere correction, which was intended to remove the effects of atmosphere scattering and absorption. Both radiometric and atmosphere corrections were performed using ENVI 5.2.

Fieldworks were conducted in Aug of 2009 and Aug of 2014. A total of 40 sample plots for 2009 measurement with data range of 12–318.8 g/m^2 and a mean value of 147.43 g/m^2 and a total of 34 sample plots for 2014 measurement with data range of 44.38-407.08 g/m^2 and a mean value of 150.40 g/m^2 were collected and used in this research. The distribution of the field measured biomass is shown in Figure 1.



a. b.

Figure 1. Location of field measured biomass. a) QB and b) Landsat.

3. METHODOLOGY

First of all, the spectral indices were calculated for both of imagery and stacked as one layer. Second, the correlation between field measured biomass and spectral indices was estimated using partial least square regression. Third, multiple vegetation indices were selected if their correlation coefficient are more than 0.5. Finally, the RF classification was performed to estimate biomass using these indices.

A myriad of vegetation indices have been developed and researched over the years, we calculated 25 indices commonly used for vegetation analysis in this study (Table 1). Testing multiple indices is useful, because at low fractional vegetated groundcover factors such as soil reflectance may interfere with the vegetation signal, and different indices are more sensitive in different ranges of biomass and groundcover.

Partial least squares (PLS) regression is a quick, efficient and optimal regression method based on covariance. The principle of PLS regression is to firstly decompose explanatory variables into a few non-correlated latent variables or components using information contained in the response variable; then to regress the new components against the response variable [7]. In the present study, we used XLSTAT software and the correlation coefficients (r), the root means square error (RMSE) and the coefficients of determination (r^2) between predicted and measured biomass were two criteria used to select the best model with optimal number of components.

RF is a non-parametric machine learning algorithm that was implemented in this study using the "RandomForest" package [11] within the R software environment (http://www.R-project.org). RF can be used for regression or classification depending on the type of variable to be estimated. Compared with linear regression techniques, RF has lower bias and avoids overfitting [12]. The advantage of RF is that it can run effectively on large and multi-source data sets and it is relatively robust to outliers, a reduction of training data and noise [13; 14; 15]. For each tree, approximately two-thirds of the original data was randomly chosen to build the tree, and the remaining data was used for estimating out-of-bag error and calculating variable importance. In this study, RF was applied to develop biomass estimation using filed measured biomass as reference data and Landsat and QB derived spectral indices with an r value of 0.5< as predictors.

Table 1. Vegetation indices used for the study

No.	Vegetation index	Formula	Name	References
1	DVI	$\rho_{NIR} - \rho_{red}$	Difference vegetation index	[16]
2	EVI	$2.5 * (\rho_{NIR} - \rho_{red}) / (\rho_{NIR} + 6 * \rho_{red} - 7.5 * \rho_{blue} + L)$	Enhanced vegetation index	[17]
3	G-RVI	$(\rho_{green} - \rho_{red}) / (\rho_{green} + \rho_{red})$	Greed-red vegetation index	[18]
4	GARI	$\rho_{NIR} - (\rho_{green} - \gamma * (\rho_{blue} - \rho_{red}) / \rho_{NIR} + (\rho_{green} + (\rho_{blue} - \rho_{red})$	Green atmospherically resistant vegetation index	[19]
5	GDVI	$\rho_{NIR} - \rho_{green}$	Green difference vegetation index	[20]
6	GNDVI	$(\rho_{NIR} - \rho_{green}) / (\rho_{NIR} + \rho_{green})$	Green NDVI	[21]
7	GRVI	$\rho_{NIR} / \rho_{green}$	Green ratio vegetation index	[20]
8	GVI	(-0.2848 * TM1) + (-0.02848 * TM2) + (-0.5836 * TM3) + (0.7243 * TM4) + (-0.1800 * TM5)	Green vegetation index, Tasseled cap	[22]
9	IPVI	$\rho_{NIR} / (\rho_{NIR} + \rho_{red})$	Infrared percentage vegetation index	[23]
10	LAI	3.618 * EVI - 0.118	Leaf area index	[24]
11	MCARI1	$1.2 * (2.5 * (P_{800} - P_{670}) - 1.3 * (P_{800} - P_{550}))$	Modified chlorophyll absorption	[25]
12	MSAVI	$((2 * (P_{800} + 1) - ((2 * P_{800} + 1)^2 - 8 * (P_{800} - P_{670})))^{0.5}) / 2$	Modified soil-adjusted vegetat	[26]
13	MSR	$(P_{800} / P_{670} - 1) / (P_{800} / P_{670} + 1)$	Modified simple ratio	[27]
14	NDVI	$(\rho_{NIR} - \rho_{red}) / (\rho_{NIR} + \rho_{red})$	Normalized difference vegetation index	[28]
15	NLI	$(\rho_{NIR}^2 - \rho_{red}) / (\rho_{NIR}^2 + \rho_{red})$	Non-linear vegetation index	[29]
16	OSAVI	$(\rho_{NIR} - \rho_{red}) / (\rho_{NIR} + \rho_{red} + 0.16)$	Optimization of SAVI	[30]
17	RDVI	$(\rho_{800} - \rho_{670}) / (\rho_{800} + \rho_{670})^{0.5}$	Re-normalized differen vegetation index	[31]
18	SAVI	$(\rho_{NIR} - \rho_{red}) / (\rho_{NIR} + \rho_{red} + L)$	Soil adjusted vegetation index	[32]

19	SR	ρ_{NIR} / ρ_{red}	Ratio vegetation index, simple ratio	[33]
20	TDVI	$(0.5 +(\rho_{NIR} - \rho_{red}) / (\rho_{NIR} + \rho_{red}))^{0.5}$	Transformed vegetation index	[34]
21	TVI	$0.5 * (120 * (\rho_{NIR} - \rho_{green}) / (\rho_{NIR} + \rho_{green}))$	Triangular vegetation index	[35]
22	VARI	$(\rho_{green} - \rho_{red}) / (\rho_{green} + \rho_{red} - \rho_{blue})$	Visible atmospherically resistant vegetation index	[36]
23	VARVI	$(\rho_{green} - \rho_{red}) / (\rho_{green} + \rho_{NIR} + \rho_{red})$	Visible Atmospherically Resistant Index (VARI)	[36]
24	WDVI	$\rho_{NIR} - a*\rho_{red}$ 0.2	Weighted difference vegetation index	[37]
25	WV-VI	$(\rho_{NIR} * 2 - \rho_{red}) / (\rho_{NIR} * 2 + \rho_{red})$	World view NDVI	[38]

4. RESULTS AND DISCUSSION

The correlation coefficients estimated for satellite data are presented in Table 2. For all vegetation indices, VARVI (r=0.712) yielded the highest correlation coefficient value for Landsat data, while SR (r=0.522) was considered the highest correlated index for QB data. In terms of both of the imagery, G-RVI (r=0.502 and 0.700) and VARI (r=0.513 and 0.696) were the best vegetation indices to explain ground biomass. On the other hand, EVI (r=0.049 and -0.383) and LAI (r=0.049 and -0.383) showed the lowest performance to estimate measured biomass. As a result of the PLS, five vegetation indices, namely, IPVI, NDVI, OSAVI, SAVI and TVI, had multi-collinear correlation.

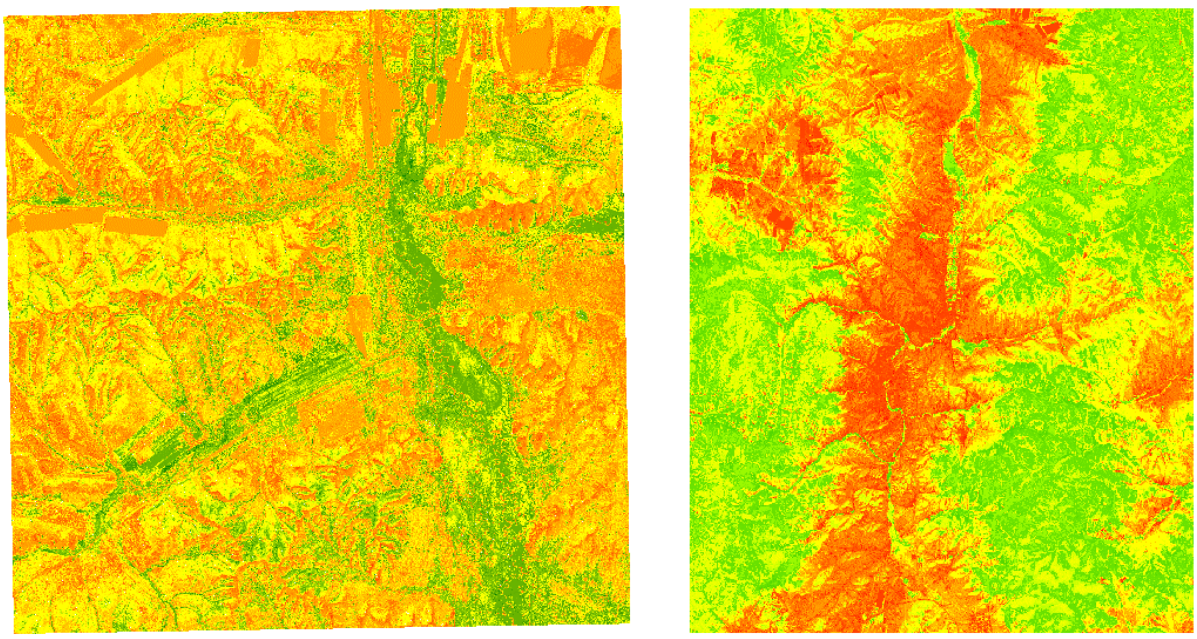
Table 2. The results of correlation coefficients.

	QB	Landsat		QB	Landsat
DVI	0.471	0.454	NDVI	0.508	0.525
EVI	0.049	-0.383	NLI	0.441	0.444
G-RVI	0.502	0.700	OSAVI	0.508	0.525
GARI	0.510	0.508	RDVI	0.490	0.488
GDVI	0.478	0.382	SAVI	0.508	0.525
GNDVI	0.500	0.429	SR	0.522	0.507
GRVI	0.506	0.424	TDVI	0.505	0.527
GVI		-0.541	TVI	0.500	0.429
IPVI	0.508	0.525	VARI	0.513	0.696
LAI	0.049	-0.383	VARVI	0.305	0.712
MCARI1	0.463	0.508	WDVI	0.423	0.335
MSAVI	0.497	0.529	WV-VI	0.441	0.444
MSR	0.518	0.517			

The relationship between measured biomass and QB derived vegetation indices resulted in an r^2 value of 0.337 and RMSE=83.435 g/m^2. On the other hand, vegetation indices from Landsat performed relatively well in predicting the groundcover with a r^2 value of 0.617 and RMSE=50.881 g/m^2. This may demonstrate that high spatial resolution images have a significant number of shadows from trees and terrain, resulting in errors for AGB estimation.

The biomass map was generated using vegetation indices with a correlation coefficient value of more than 0.5. For QB image, 11 vegetation indices (G-RVI, GARI, GRVI, IPVI, MSR, NDVI, SAVI, SR, TDVI, TVI and VARI) were used as predictors to estimate biomass, whereas 13 indices (G-RVI, GARI, IPVI, MCARI, MSR, NDVI, OSAVI, SAVI, SR, TDVI, TVI, VARI and VARVI) derived from Landsat data were selected to train the RF. The results of RF classification are presented in Figure 2.

As could be seen from the Figure 2, the most applicable vegetation index for predicting the biomass is the G-RVI. Then, the GARI, GRVI, IPVI, MSR, NDVI, SAVI, SR, TDVI, TVI and VARI indices could be successively used for the estimation of the AGB. Nevertheless, thorough comparison illustrated that the difference between these vegetation indices was not very high. Meanwhile, the comparison between the QB and Landsat datasets showed that the finer spatial resolution led to lower performance than Landsat 8 imagery.



a. b.

Figure 2. The results of biomass estimation for a) QB and b) Landsat data

5. CONCLUSION

This study aimed to investigate applicability of vegetation indices to remotely estimate biomass and to test the possibility of using high resolution and medium resolution satellite images for the biomass estimation in Bornuur soum of Tov Province using the RF algorithm. Among the 25 vegetation indices, VARVI (r=0.712) had the strongest correlation with biomass of the Landsat data, while SR (r=0.522) yielded the highest correlation coefficient value for the QB data. The most useful vegetation index for predicting biomass was G-RVI followed by GARI, GRVI, IPVI, MSR, NDVI, SAVI, SR, TDVI, TVI and VARI; however, the difference between these vegetation indices was not very high. Also, the study compared QB and Landsat data for biomass estimation. Our results showed that the finer spatial resolution led to lower performance than Landsat 8 imagery.

REFERENCES

[1] Li, X., Du, H., Mao, F., Zhou, G.M., Chen, L., Xing, L.Q., Fan, W.L., Xu, X.J., Liu, Y.L., Cui, L., Li, Y.G., Zhu, D.E. & Liu, T.Y. 2018. Estimating bamboo forest aboveground biomass using EnKF-assimilated MODIS LAI spatiotemporal data and machine learning algorithms. *Agricultural* and Forest *Meteorology*, 256–257:445–457. DOI: https://doi.org/10.1016/j.agrformet.2018.04.002

[2] Gasparri, N.I., Parmuchi, M.G., Bono J., Karszenbaum, H. & Montenegro, C.L. 2010. Assessing multi-temporal Landsat 7 ETM+ images for estimating above-ground biomass in subtropical dry forests of Argentina. *Journal of Arid Environments* ,74:1262–1270. DOI: https://doi.org/10.1016/j.jaridenv.2010.04.007

[3] Zhu, X. & Liu, D. 2015. Improving Forest aboveground biomass estimation using seasonal Landsat NDVI time-series. ISPRS Journal of Photogrammetry and Remote Sensing, 102:222–231.

[4] Vaglio, L.G., Chen, Q., Lindsell, J., Coomes, D., Del, F.F., Guerriero, L., Pirotti, F. & Valentini, R. 2014. Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data. ISPRS Journal of Photogrammetry and Remote Sensing. 89:49–58.

[5] Lu, D. 2006. The potential and challenge of remote sensing-based biomass estimation. International Journal of Remote Sensing. 27:1297-1328.

[6] José, P., Juan, H. & Nirani, R. (2016) Remote sensing-based biomass estimation.

[7] Lu, D., Chen, Q., Wang, G., Liu, L., Li, G. & Moran, E. 2014. A survey of remote sensing-based aboveground biomass estimation methods

in forest ecosystems. International Journal of Digital Earth. 1–43.

[8] Pflugmacher, D., Cohen, W.B., Kennedy, R.E. & Yang, Z. 2014. Using Landsat-derived disturbance and recovery history and lidar to map forest biomass dynamics. Remote Sensing of Environment. 151:124-137. DOI: https://doi.org/10.1016/j.rse.2013.05.033

[9] Tanase, M., Panciera, R., Lowell, K., Tian, S., Kacker, J. & Walker, J. 2014. Airborne multi-temporal L-band polarimetric SAR data for biomass estimation in semi-arid forests. Remote Sensing of Environment. 145:93-104. DOI: https://doi.org/10.1016/j.rse.2014.01.024

[10] Belgiu, M. & Drăguţ, L. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing. 114: 24-31. DOI: https://doi.org/10.1016/j.isprsjprs.2016.01.011

[11] Liaw, A. & Wiener, M. 2002. Classification and Regression by Random Forest. R News, 2: 18-22.

[12] Tian, X., Yan, M., van der Tol, C., Li, Z., Su, Z., Chen, E., Li, X., Li, L., Wang, X., Pan, X. & Gao, L. 2017. Modeling Forest above-ground biomass dynamics using multi-source data and incorporated models: A case study over the qilian mountains. Agricultural and Forest Meteorology, 246:1–14. DOI: https://doi.org/10.1016/j.agrformet.2017.05.026

[13] Breiman, L. 2001. Random Forests. Mach. Learn. 45:5–32.

[14] Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction; *Springer Science + Business Media: New York, NY, USA*.

[15] Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J.P. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing. 67:93–104. DOI: https://doi.org/10.1016/j.isprsjprs.2011.11.002

[16] Tucker, C. 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment. 8:127-150.

[17] Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao X. & Ferreira L.G. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment. 83 (1-2):195-213.

[18] Motohka, T., Nasahara, K., Hiroyuki, O. & Satoshi, T. 2010. Applicability of Green-Red Vegetation Index for Remote Sensing of Vegetation Phenology. Remote Sensing. 2:2369-2387.

[19] Gitelson, A., Kaufman, Y.J. & Merzlyak, M.N. 1996, Use of a green channel in remote sensing of global vegetation from EOS-MODIS, Remote Sensing of Environment, 58:289-298.

[20] Sripada, R., Heiniger, R., White, J. & Weisz, R. 2005. Aerial Color Infrared Photography for Determining Late-Season Nitrogen Requirements in Corn. Agronomy Journal - AGRON J. 97. 10.2134/agronj2004.0314.

[21] Gitelson, A.A. & Merzlyak, M.N. 1998. Remote Sensing of Chlorophyll Concentration in Higher Plant Leaves. Advances in Space Research, 22:689-692.

[22] Kauth, R. J. & Thomas, G. S. 1976. The Tasselled-Cap—A Graphic Description of the Spectral-Temporal Development of Agricultural Crops as Seen by Landsat. Proceedings, Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, IN, 29 June-1 July 1976, 41-51.

[23] Barati, S., Rayegani, B., Saati, M., Sharifi, A. & Nasri, M. 2011. Comparison the accuracies of different spectral indices for estimation of vegetation cover fraction in sparse vegetated areas. The Egyptian Journal of Remote Sensing and Space Science, 14: 49-56.

[24] Boegh, E., Soegaard, H., Broge, N., Hasager, C., Jensen, N.O., Schelde, K. & Thomsen, A. 2002. Airborne multispectral data for quantifying leaf area index, nitrogen concentration, and photosynthetic efficiency in agriculture. Remote Sensing of Environment. 81:179-193.

[25] Haboudane, D., Miller, J., Pattey, E., Zarco-Tejada, P. & Strachan, I. 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. 90:337-352.

[26] Qi J., Chehbouni A., Huete A.R., Kerr Y.H. & Sorooshia S. 1994. A modified soil adjusted vegetation index. Remote Sensing of Environment, 48:119-126.

[27] Chen, J. M. & Cihlar, J. 1996. Retrieving Leaf Area Index of Boreal Conifer Forests Using Landsat TM Images, Remote Sensing of Environment, 55:153-162.

[28] Rouse, J. W., Haas, R. H., Deering, D. W. & Sehell, J. A. 1974. Monitoring the vernal advancement and retrogradation (Green wave effect) of natural vegetation. Final Rep. RSC 1978-4, Remote Sensing Center, Texas A&M University, College Station.

[29] Goel, N. S., & Qin, W. 1994. Influences of canopy architecture on relationships between various vegetation indices and LAI and FPAR: A computer simulation. Remote Sensing of Environment, 10: 309–347.

[30] Rondeaux, G., Steven, M. & Frederic, B. 1996. Optimization of Soil-Adjusted Vegetation Indices. Remote Sensing of Environment. 55:95-107.

[31] Roujean J.L. & Breon, F.M. 1995. Estimating PAR Absorbed by Vegetation from Bi-Directional Reflectance Measurements. Remote Sensing of Environment, 51:375-384.

[32] Huete, A.R. 1988 A Soil Adjusted Vegetation Index (SAVI). Remote Sensing of Environment. 25:295-309.

[33] Birth, G.S. & McVey, G.R. 1968. Measuring the Color of Growing Turf with a Reflectance Spectrophotometer. American Society of Agronomy, 60:640-643.

[34] Abdou, B., Asalhi, H. & Teillet, P.M. 2002. Transformed difference vegetation index (TDVI) for vegetation cover mapping. 5:3053-3055.

[35] Broge, N.H & Leblanc, E. 2001. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. Remote Sensing of Environment. 76:156-172.

[36] Gitelson, A.A., Kaufman, Y.J., Stark, R. & Rundquist, D., Novel algorithms for remote estimation of vegetation fraction, Remote Sensing of Environment, 80:76-87.

[37] Clevers, J.G.P.W. 1991. Application of the WDVI in estimating LAI at the generative stage of barley. ISPRS Journal of Photogrammetry and Remote Sensing, 46:37-47.

[38] Wolf A. 2010. Using WorldView 2 Vis-NIR MSI imagery to support land mapping and feature extraction using normalized difference index ratios. DigitalGlobe 8-Band Research Challenge.1-13.