ABOVEGROUND SHRUB BIOMASS ESTIMATION BASED ON LANDSAT DATA IN MU US SANDY LAND, CHINA

Jian ZHAO[a,c] , Chunxiang CAO[b*] and Peera YOMWAN[a,c,d]

a Ph.D. Student, State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Applications of Chinese Academy of Sciences, Beijing 100101, P.R. China

E-mail: zhaojian2009@irsa.ac.cn, peerayom@hotmail.com

b Professor, State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Applications of Chinese Academy of Sciences,

Postbox 9718-13, Datun Rd., Chaoyang, Beijing 100101, P.R. China;

Tel: +86-139-1161-0226;

E-mail: cao143@irsa.ac.cn

c Graduate School of Chinese Academy of Sciences, Beijing, 100049, P.R. China

d Professional Engineer, Department of Lands, Bangkok, 10210, Thailand

KEY WORDS: Aboveground biomass, Random Forest, remote sensing, shrub, Mu Us sandy land

Abstract: Forest biomass is a crucial parameter in global change and carbon trade. The IPCC reports defined the shrub carbon storage as a part of carbon inventory, but few researches discussed the carbon stock of shrub vegetation. Here we proposed an empirical machine learning model to estimate the shrub aboveground live biomass. The approaches mapping AGB from satellite observations were divided into directly and indirectly ways. Indirectly biomass mapping from parameters like cover, vegetation index was widely used, but the parameter itself and the two step modeling increased the uncertainty and reduced the accuracy. Here, we discusses the directly method based on reflectance and ancillary data. On the basis of Landsat TM images, a regression tree-based model (Random Forest) was built to simulate the AGB. The Landsat images acquired in 30th, June, 2009 were firstly re-projected, geographically and radiometric corrected. Water bodies, clouds and their shadows were identified and masked. Field data was measured from the 39 field plots in 2009 and 2011, which was used in modeling and validation. Ancillary data includes the topographic data. The empirical, non-parametric model Random Forest was successfully applied for forest biomass estimation, which captures non-linear relationship between satellite data and biomass density. Our model had a good performance with the explained variance 75.46%, RMSE 1.55 Mg ha-1. In this study, we demonstrate that Landsat data provide the capability to produce accurate and detailed estimation of biomass distribution in Mu Us sandy land. Adding ancillary data could improve shrub biomass prediction.

INTRODUCTION

Forest biomass is a crucial indicator in carbon sequestration capacity and forest carbon budget evaluation (Dixon, Brown et al. 1994; Fang, Chen et al. 2001; Dong, Kaufmann et al. 2003). The IPCC reports defined the shrub carbon storage as a part of carbon inventory, but few researches discussed the carbon stock of shrub vegetation. Attention has been paid recently to the reforestation possibility of sandland , and the role of shrub-dominant sandlands in the carbon budget (Xu, Cao et al. 2010), where research shows sandy desertified land rehabilitation improves soil carbon sequestration (Su, Wang et al. 2010). In addition, sandland, as the fragile ecosystem with poor resilience, paves the way for a switch to catastrophic shifts in climate, nutrient loading, habitat fragmentation or biotic exploitation (Scheffer, Carpenter et al. 2001; Gao, Hu et al. 2002). Remote sensing from space is the only feasible method for mapping and monitoring the extent and density of woody shrub canopy parameters over these remote and inaccessible tracts of land(Chopping, Moisen et al. 2008). Canopy structure parameters can be extracted directly from high resolution images such as IKONOS, Geoeye-1, et al. These data sources have high measuring precision but have limitations such as costliness, coverage limitation, and are time consuming in image processing. Also without an NIR band, they are relatively less useful in applications (Chen, Gu et al. 2009; Gonzalez, Asner et al. 2010). Low resolution images have great advantages in large-scale applications, although their accuracy is relatively low. Chopping et.al explored the vegetation parameter retrieval by adding angle information (Chopping,

Nolin et al. 2009). Middle resolution images are the most widely used data source as they balance the observing range, resolution and cost(Poulos 2009; Gasparri, Parmuchi et al. 2010; Soenen, Peddle et al. 2010) .

The approaches mapping AGB from satellite observations were divided into directly and indirectly ways. Indirectly biomass mapping from parameters like cover, vegetation index was widely used, but the parameter itself and the two step modeling increased the uncertainty and reduced the accuracy. Especially in arid and semi-arid areas, low vegetation cover and the bright background of the dryland made the vegetation indices less effectiveness (Ni and Li 2000; Chopping, Su et al. 2008). Several studies have discussed the capabilities and limitations of optical sensors for direct biomass estimation, demonstrating the sensitivity of visible and short wave infrared wavelengths to vegetation density and structure(Baccini, Friedl et al. 2004; Avitabile, Baccini et al. 2011).

Breiman proposed random forests an effective tool in prediction, which change how the classification or regression trees are constructed. It performs very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against overfitting (Breiman 2001; Friedman, Hastie et al. 2001) . Advantages of RF compared to other statistical classifiers include (1) very high classification accuracy; (2) a novel method of determining variable importance; (3) ability to model complex interactions among predictor variables; (4) flexibility to perform several types of statistical data analysis, including regression, classification, survival analysis, and unsupervised learning; and (5) an algorithm for imputing missing values(Cutler, Edwards Jr et al. 2007).

Here we proposed an empirical machine learning model to estimate the shrub aboveground live biomass. Here, we discusses the directly method based on reflectance and ancillary data. On the basis of Landsat TM images, a regression tree-based model (Random Forest) was built to simulate the AGB, shown in Figure 1.

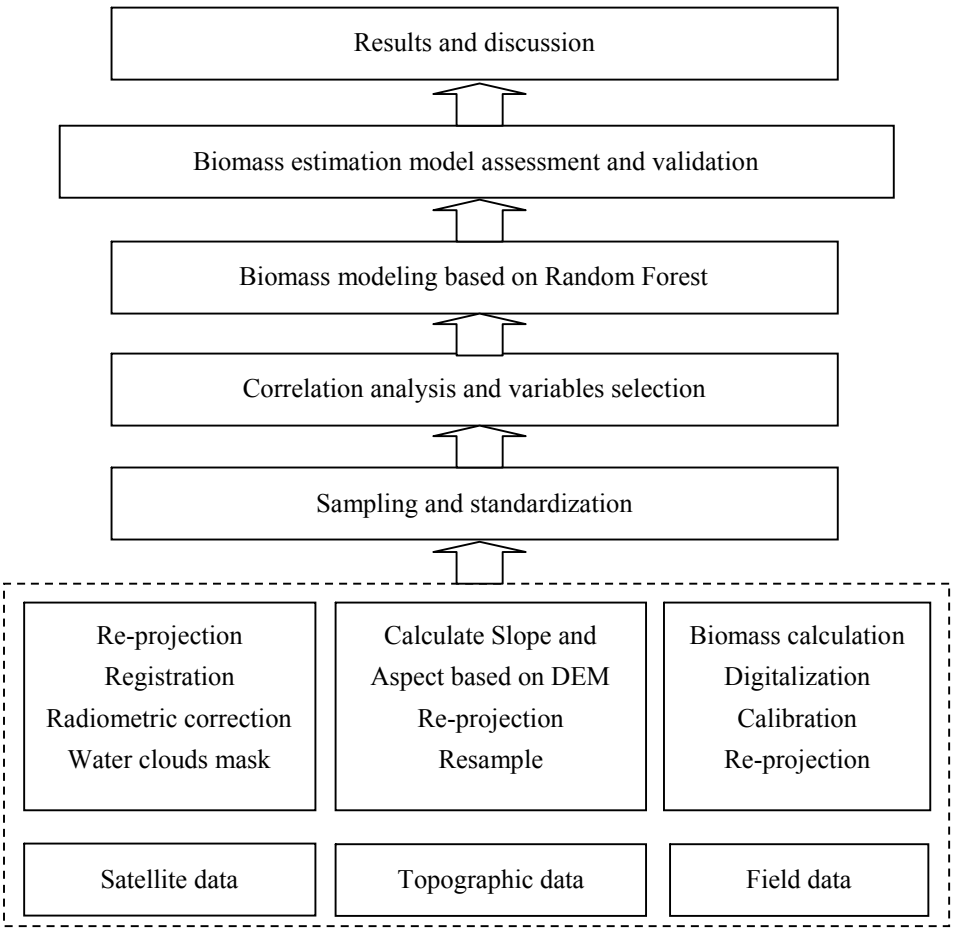

Figure 1: Flowchart of our study

DATA AND METHODS

Study area

Our study area, Mu Us Sandland, is located in the junction of Ningxia, Inner Mongolia and Shanxi Province, China (Fig. 2) between the range of 37°27.5′ ~ 39°22.5′ N and 107°20′ ~ 110°30′ E. This region, between the Ordos Plateau and the Loess Plateau, has an extreme temperate continental climate with low rainfall, drought, wind, strong evaporation and abundant sunshine. The annual precipitation is 350 to 400mm, annual evaporation is about 2592mm, annual average temperature is 6 °C, annual average wind speed is 3.4m/s, and the average number of days of above gale force 8 is about 24 days. The dominant vegetation is shrub, including *tamarix, salix psammophila, hippophae rhamnoides, Pekin willow, caragana, scoparium, artemisia, hedysarum, salix cheilophila* , etc. We selected the smaller study area north of Mu Us Sandland, which covered 36 measure fields shown in Figure 2 in red rectangle.

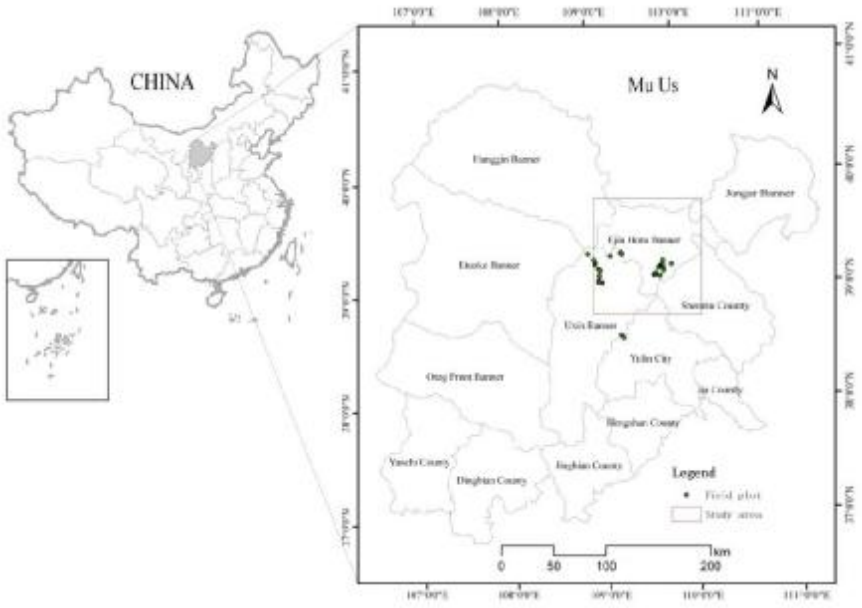

Figure 2: Study area of Mu Us sandland in China and field plots

Field datas

Field data from 39 field plots were collected during July 2009 and 2011. These plots contain main shrub species in the hinterland of Mu Us Sandland. Plot size was set at least 0.09 ha (30m x 30m) to ensure that 30m image pixel would spatially coincide with each plot, because, based on individual pixels image, data was associated with field plots. The latitude and longitude were recorded by a high-accuracy global positioning system (GPS). These points were loaded and reprojected in ArcGIS Desktop ver9.3 software. Above Ground Biomass (AGB) validation data were calculated for each plot from field measurements, according to the different shapes of different species, plots were separated into two types: 1) grasslike shrubs; 2) arborlike shrubs.

For the grasslike shrubs such as *salix monogolica, artemisia arenaria*, etc., shrubs were divided into three classes: big, middle and small clusters, according to the actual measurements of vertical canopy diameter and height for each shrub and for each grasslike shrub plot. Shrub heights and vertical canopy diameters were measured using a surveyor's rod accurate to 0.1 centimeter. A part of each cluster was harvested as a sample of AGB. Samples were weighed immediately after they were harvested, and then put into an oven for 72 hours at a temperature of 120 . Dried samples were weighed to calculate the proportion of dry matter for each plot. The AGB of each plot (Xu, Cao et al. 2010) is formulated as

$$AGB = a \times W_B + b \times W_M + c \times W_S \quad (1)$$

Where W_B, W_M and W_S are dry weights of the big, middle and small cluster samples respectively and a, b and c are the counted numbers of corresponding clusters in the plot.

For arborlike shrubs such as *hankowwillow*, AGB was calculated from diameter at breast height(dbh) and tree height. The dbh was derived from the perimeter measured at the 1.2m height and the tree height was measured directly from the surveyor's rod for each of the trees in the plot.

$$AGB = a(D^2H)^b \qquad (2)$$

Where D and H are the dbh and tree height respectively; a and b are regression parameters.

Satellite data

The Landsat TM 5 image was acquired 30th, June, 2009, which was geographically and atmospherically corrected. All the cloud and water part was masked artificially. Landsat TM 5 was launched March 1, 1984 by NASA. It has a Worldwide Reference System-2 (WRS-2) path/row system, Circular, sun-synchronous, near-polar orbit at an altitude of 705 km, with a repeat cycle of 16 days. This sensor has seven spectral bands descried in Table 1.

Table 1 Spectral of Landsat TM 5

Band	Spectral range	Wavelength (μm)	Resolution(m)
Band 1	Blue	0.45 ~ 0.52	30
Band 2	Green	0.52 ~ 0.60	30
Band 3	Red	0.63 ~ 0.69	30
Band 4	Near-Infrared	0.76 ~ 0.90	30
Band 5	Near-Infrared	1.55 ~ 1.75	30
Band 6	Thermal	10.40 ~ 12.50	120
Band 7	Mid-Infrared	2.08 ~ 2.35	30

The DEM was derived from ASTER Global Digital Elevation Model (ASTER GDEM), a joint product developed by the Ministry of Economy, Trade, and Industry (METI) of Japan and the United States National Aeronautics and Space Administration (NASA) (http://www.jspacesystems.or.jp/ersdac/GDEM/E/index.html). It was generated from data collected from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) and covers the entire land surface of the Earth. The ASTER GDEM is in GeoTIFF format with geographic lat/long coordinates and a 1 arc-second (30 m) grid of elevation postings. It is referenced to the WGS84/EGM96 geoid.

All these spectral and topographic variables were projected to Albers equal area coordinate system, with the central meridian set as 105°E, two standard parallels set as 25°N and 47°N. As this paper only tested the model suitability in the small area, the images was clipped as the red retangle area in Figure 1.Then they were transferred into ASCII format.

Modeling process

Empirical, non-parametric models do not assume any a-priori statistical distribution of the input data nor any specific form in the relation (e.g.linear) between the predictors and there sponse variable. RandomForest, an extension of tree-based models, has been successfully applied for biomass estimation using remote sensing data in several different contexts (Baccini, Friedl et al. 2004; Avitabile, Baccini et al. 2012). This algorithm was implemented in the opensource software R (R Development Core Team 2011). The statistic regression model was produced and the variable importance was predicted(Liaw and Wiener 2002). The percent of variance explained (R^2) and Root Mean Squared Error (RMSE) were conducted to evaluate the regression performance.

RESULTS

We developed a non-parameter tree-based model Random forest to simulate biomass of the test area. The explained variance was 75.46%, which means the predicted value and the observation has relatively high correlation. The RMSE was 1.55 Mg ha^{-1}, which means the error of OOB was not very low. Especially for this arid and semi-arid area, low rain fall along with the low biomass (most of them was lower than 10 Mg ha^{-1}) made it more difficult to predict from spectral reflectance and presented slightly larger errors.

The variable importance was computed as the average increase of node purity(i.e.residual sum of squares) on the OOB data that results from including each variable(Liaw and Wiener 2002). The Near-Infrared band(0.76 ~ 0.90 μm) was predicted as the most important virable, which seens different from other forest biomass estimation(Avitabile, Baccini et al. 2012). In their study, SWIR spectral band provided the maximum contribution.

The spatial continuous biomass density was mapped as in Figure 4. The map shows the distribution of AGB across the study area as well as the spatial variability of AGB. The shrub biomass density ranges from 1.3 to 8.2 Mg ha^{-1}. The frequency distribution of predicted biomass is consistant to that of the training image. The map indicates that most of the country presents low biomass density, but the contribution of these areas to the total AGB stock is important. Most of the high values of biomass are concentrated around the lakes. The biomass values without the use of land cover information were reasonable.

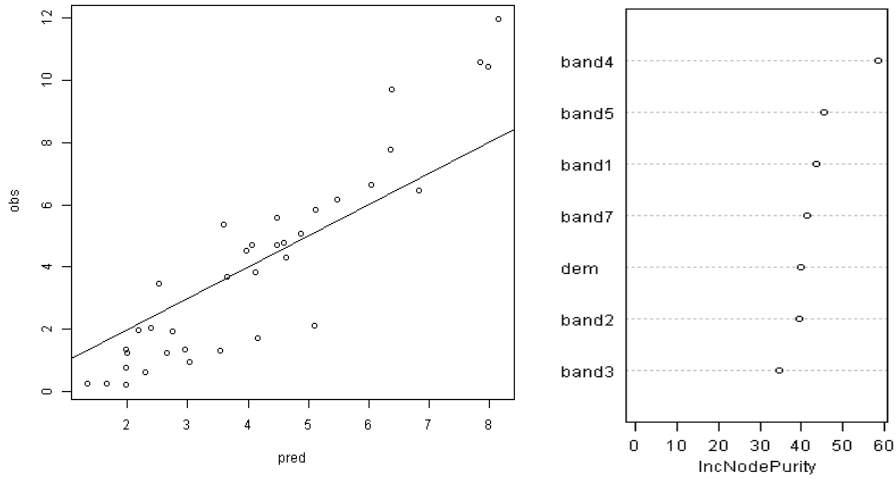

Figure 3 Regression result of predicted and observed biomass, the black line represents 1:1(left); Variable importance plots for biomass model, IncNodePurity(x axis) represents the variable importance(right).

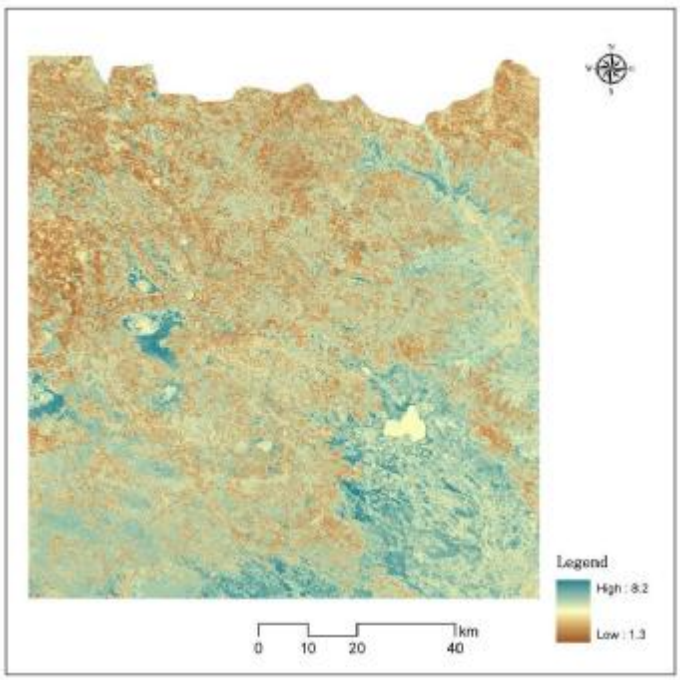

Figure 4 Prediction map of shrub biomass in test area of Mu Us sandland

DISCUSSION

For the plot data acquirement, due to the limitation of field work, we could not get the ideal randomly or systematic distributed data, but we tended to enlarge the field work area and the dispersed plots distribution. To ensure an adequate representation of biomass conditions within the study area, we selected field plots with large range of canopy cover and different dominant vegetation. In addition, plot size seems relatively small to ensure that 30m image pixel would spatially coincide with each plot because of the bios of images and positioning. The number of measured plots was limited.

Regarding the variable importance, the maximum contribution variable was the NIR band in our study. While in other studies it was SWIR band (Baccini, Laporte et al. 2008; Avitabile, Baccini et al. 2012). It's said that SWIR

was more important because they allowed effective separation between high and low biomass data. However, in our arid and semi-arid sandland areas, the vegetation cover was low and the background of the dryland was bright.

In the following study, we should pay attention to more calibration measurements to test the robust of this method. The input of other reasonable and easy to get variables would be considered. Further more, we could compare our model with other linar and non-parameter models.

CONCLUSION

In this study, we acquired 39 field plot measurments through 2009 and 2011 in Mu Us sandland, China. Based on the non-parameter Random Forest model, we produced the spatially continuous biomass density map of a test area in Mu Us sandland using 30m resolution remote sensing observations (Landsat). The high explained variance and low RMSE indicate that our Random Forest model had a good performance of AGB estimation in the arid and low cover area. NIR band had the maximum contribution in the variable impartance evaluation. Our model still need more field work to calibration and assessment.

ACKNOWLEDGMENT

This paper has been supported by Natural Science Foundation of China (Grant No. 41171330) and the State Key laboratory foundation(Y1Y00245KZ). We are grateful to all helped in the experimentation and English writing.

REFERENCES

Avitabile, V., A. Baccini, et al., 2012. Capabilities and limitations of Landsat and land cover data for aboveground woody biomass estimation of Uganda. Remote Sensing of Environment 117(0): 366-380.
Baccini, A., M. Friedl, et al., 2004. Forest biomass estimation over regional scales using multisource data. Geophysical research letters 31(10): L10501.
Baccini, A., N. Laporte, et al., 2008. A first map of tropical Africa's above-ground biomass derived from satellite imagery. Environmental Research Letters 3(4): 045011.
Breiman, L., 2001. Random forests. Machine Learning 45(1): 5-32.
Chen, J., S. Gu, et al., 2009. Estimating aboveground biomass of grassland having a high canopy cover: an exploratory analysis of n situ hyperspectral data. International Journal of Remote Sensing 30(24): 6497 - 6517.
Chopping, M., G. G. Moisen, et al., 2008. Large area mapping of southwestern forest crown cover, canopy height, and biomass using the NASA Multiangle Imaging Spectro-Radiometer. Remote Sensing of Environment 112(5): 2051-2063.
Chopping, M., A. Nolin, et al., 2009. Forest canopy height from the Multiangle Imaging SpectroRadiometer (MISR) assessed with high resolution discrete return lidar. Remote Sensing of Environment 113(10): 2172-2185.
Chopping, M., L. Su, et al., 2008. Remote sensing of woody shrub cover in desert grasslands using MISR with a geometric-optical canopy reflectance model. Remote Sensing of Environment 112(1): 19-34.
Cutler, D. R., T. C. Edwards Jr, et al., 2007. Random forests for classification in ecology. Ecology 88(11): 2783-2792.
Dixon, R. K., S. Brown, et al., 1994. Carbon Pools and Flux of Global Forest Ecosystems. Science 263(5144): 185-190.
Dong, J., R. K. Kaufmann, et al., 2003. Remote sensing estimates of boreal and temperate forest woody biomass: carbon pools, sources, and sinks. Remote Sensing of Environment 84(3): 393-410.
Fang, J., A. Chen, et al., 2001. Changes in Forest Biomass Carbon Storage in China Between 1949 and 1998. Science 292(5525): 2320-2322.
Friedman, J., T. Hastie, et al., 2001. The elements of statistical learning, Springer Series in Statistics.
Gao, L. F., Z. A. Hu, et al., 2002. Genetic diversity of rhizobia isolated from Caragana intermedia in Maowusu sandland, north of China. Letters in Applied Microbiology 35(4): 347-352.
Gasparri, N. I., M. G. Parmuchi, et al., 2010. Assessing multi-temporal Landsat 7 ETM+ images for estimating above-ground biomass in subtropical dry forests of Argentina. Journal of Arid Environments 74(10): 1262-1270.
Gonzalez, P., G. P. Asner, et al., 2010. Forest carbon densities and uncertainties from Lidar, QuickBird, and field measurements in California. Remote Sensing of Environment 114(7): 1561-1575.
Liaw, A. and M. Wiener, 2002. Classification and Regression by randomForest. R news 2(3): 18-22.
Ni, W. and X. Li, 2000. A Coupled Vegetation-Soil Bidirectional Reflectance Model for a Semiarid Landscape. Remote Sensing of Environment 74(1): 113-124.
Poulos, H. M., 2009. Mapping fuels in the Chihuahuan Desert borderlands using remote sensing, geographic information systems, and biophysical modeling. Canadian Journal of Forest Research 39: 1917-1927.

R Development Core Team 2011, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
Scheffer, M., S. Carpenter, et al., 2001. Catastrophic shifts in ecosystems. Nature 413(6856): 591-596.
Soenen, S. A., D. R. Peddle, et al., 2010. Estimating aboveground forest biomass from canopy reflectance model inversion in mountainous terrain. Remote Sensing of Environment 114(7): 1325-1337.
Su, Y. Z., X. F. Wang, et al., 2010. Effects of sandy desertified land rehabilitation on soil carbon sequestration and aggregation in an arid region in China. Journal of Environmental Management 91(11): 2109-2116.
Xu, M., C. X. Cao, et al., 2010. Remote sensing based shrub above-ground biomass and carbon storage mapping in Mu Us desert, China. Science China-Technological Sciences 53: 176-183.