Estimation of Forest Above-Ground Biomass Using Random Forest Algorithm Based on ALOS PALSAR and Landsat 5TM Imageries.

Chathumal M.W. Arachchige, S. Nashrrullah, Kavinda Gunasekara, Manzul K. Hazarika

Geoinformatics Centre (GIC), Asian Institute of Technology (AIT), PO Box 04, Klong Luang, Pathumthani, Thailand
Email: chathumal@ait.asia

KEY WORDS: Biomass Estimation, Nepal, ALOS PALSAR, Landsat 5TM

ABSTRACT: Accurate estimation of forest above-ground biomass (AGB) is a crucial factor for sustainable forest management and mitigating climate change. Satellite Remote sensing technology has proved to be an effective method in large scale forest monitoring as well as forest biomass estimation. In this study, focus has been given for a remote sensing approach for estimate the above-ground biomass of the western Tiger Landscape of Nepal using Radar and Optical remote sensing data.

This study was based on the integration of ALOS PALSAR Radar data with Landsat 5TM optical data. Image pre-processing together with a multitemporal approach was carried out for optical and radar imageries, in order to minimize the effects of backscatter noise of Radar imageries as well as the climatic variations of the region. Forest field inventory data was collected from 2010 to 2011 and obtained from the Forest Resource Assessment FRA) Nepal. The emerging Random Forest (RF) machine-learning algorithm is regarded as one of the most precise prediction methods for regression modeling. The objective of this study was to investigate the applicability of the RF regression algorithm for radar ,optical, and combination of radar and optical data for predicting the forest biomass and test the performance of the RF regression models. In this process ALOS PALSAR radar backscatter, radar texture parameters, Vegetation indexes form Landsat 5TM optical data, and digital elevation model data were taken into the random forest algorithm to predict the results. The performance of the model was compared with optical data, radar data, and combination of both optical and radar data. The results showed that the RF model produced more accurate estimates of the forest biomass results when it based on optical data (R^2 =0.443, RMSE = 101.46 t/ha). A combination of radar and optical data (R^2 =0.391, RMSE = 103.52 t/ha) gives good results than it only based on radar data. Using L band data also has a constraint in terms of biomass saturation (around 100-150 t/ha). Level of the Biomass saturation of the Radar data and the mountainous topography of the region are two key factors to be considered when assessing the results. The RF algorithm provides a better solution for estimating forest aboveground biomass on a large scale in western Tiger Landscape of Nepal.

1. INTRODUCTION

Forest ecosystems act as the most massive carbon sinks on land about 80% of terrestrial biosphere carbon storage, Make a significant impact on mitigating climate change (Duchelle *et al.*, 2018). Identification of the carbon pools in the forest ecosystem helps to manage the forests and access forest health. This explicit spatial measurement of forest above ground biomass also supports REDD+ (reducing emissions from deforestation and forest degradation, plus the sustainable management of forests, and the conservation and enhancement of forest carbon stocks) and due to these reasons, rapid and large-scale accurate monitoring and estimation of forest biomass is an essential need in strategic forest management and carbon stock assessment(Nashrrullah *et al.*, 2012).

Satellite remote sensing plays a significant role in monitoring and mapping the forest since the launch of Landsat mission in 1970s. Integration of Synthetic aperture (SAR) along with the optical remote sensing data have been recently utilized for forest applications. optical data acquired from satellites always give the information about the topmost canopy layer. It is also profoundly affected by atmospheric conditions, haze, and clouds. In that case Synthetic Aperture Data has the benefit of weather independent observations with vegetation penetration capabilities(Vaglio *et al.*, 2017). When considering the SAR data characteristics L band data can penetrate the canopy layer and ALOS PALSAR data has been used in so many biomass estimation studies for this reason. Due to these capabilities of SAR data, most biomass estimation studies use SAR data and combining optical data to make it more informative.

After acquiring information from the space, this information should be modeled in a way to provide the best estimations. When it comes to modeling the data, there are a lot of machine learning algorithms that can be effective in their own way. Among various machine-learning algorithms, the emerging Random Forest (RF) algorithm has been regarded as one of the most precise prediction methods for classification and regression(Wang *et al.*, 2016). Most of the satellite-based biomass estimation studies use Random forest algorithm to model the data, and it works efficiently on large datasets, and it is not sensitive to noise or over-fitting.

In this study, the ability of ALOS PALSAR and Landsat 5 imagery for retrieve and predict forest Above ground biomass random forest algorithm is evaluated.
The objectives of this study include the following:

- to model the relationship between field-measured forests AGB and ALOS PALSAR radar backscatter
- to evaluate and compare the accuracy of the random forest biomass prediction models, based on optical, radar and combination of optical and radar AGB predictors
- to map forest AGB spatial distribution by the best model

The novelty of this paper is the use of radar backscatter with texture parameters along with the optical vegetation indexes in the mapping of AGB, AGB model development, and their comparison.

2. STUDY AREA

This study is based on the area of western Terai Landscape, focusing two districts, Kailali and Kanchanpur, of Nepal. With the elevation range from 130 to 1900m above the mean sea level, this area stretches from the lowland of Terai in the south and touches a bit portion of the Siwalik region in the northern part. It is also home to the coexistence of the few remaining habitats of three endangered large mammals as well as many other outstanding flora and faunas which cover a part of conservation area of Nepal (Nashrrullah *et al.*, 2012).

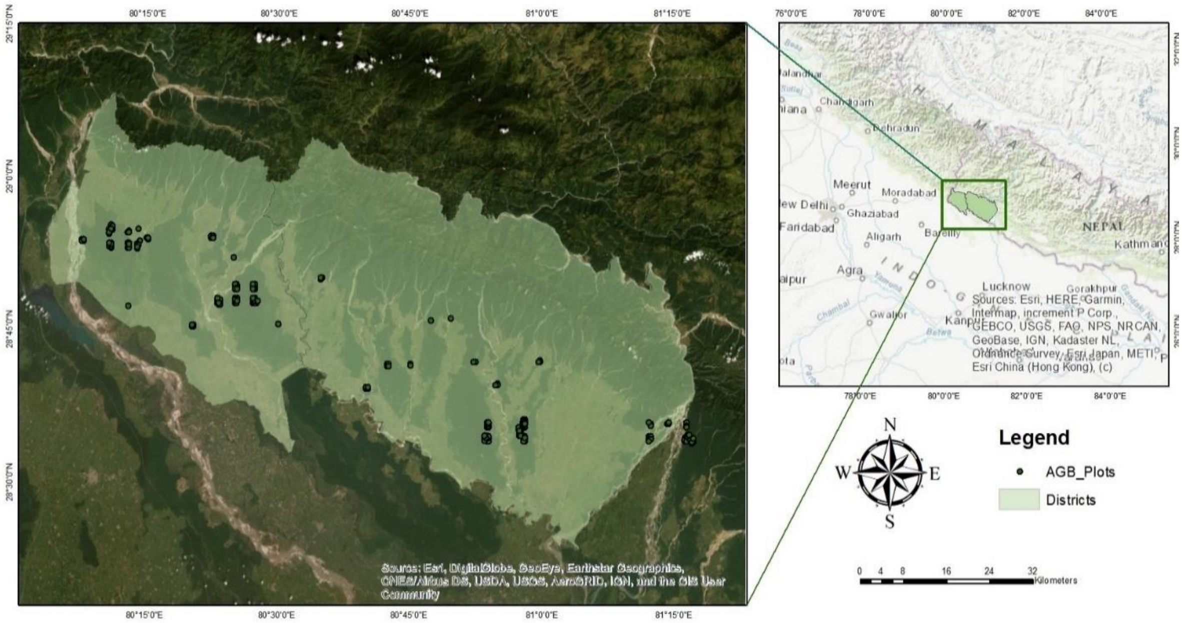


Figure 1: Location of the study area overlay with the ground plots

3. METHOD

3.1 Satellite Data Used

ALOS/PALSAR data from different time intervals ranging from 2008-2010, were obtained, covering the whole study area. All data were acquired in fine beam dual mode (HH and HV polarization) at off-nadir angle of 34.3° and delivered in Level 1.5 processing level as a geocode multi-look image with a pixel spacing of 12.5m.

Landsat 5 TM imagery was included as the optical data used in the study and downloaded from Google Earth Engine. Atmospherically corrected surface reflectance products of Landsat 5 TM are available from March 1984 - May 2012 on that platform.

Table 1: Acquired Satellite Data

Satellite Sensor	Acquisition Date	Resolution	Polarization/Bands	Remarks
Landsat 5TM	2010.10.14	30 m	1,2,3,4	ATM corrected
ALOS PALSAR	2008.05.02	12.5 m	HH/HV	Path 560 Frame 518
	2009.08.05	12.5 m	HH/HV	Path 560 Frame 518
	2010.06.23	12.5 m	HH/HV	Path 560 Frame 518
	2010.08.08	12.5 m	HH/HV	Path 560 Frame 518
	2008.05.19	12.5 m	HH/HV	Path 560 Frame 519
	2009.08.22	12.5 m	HH/HV	Path 560 Frame 519
	2010.07.10	12.5 m	HH/HV	Path 560 Frame 519
	2010.08.25	12.5 m	HH/HV	Path 560 Frame 519

3.2 Ground Truth Data

Ground truth forest data used for this study were collected from the Forest Resource Assessment (FRA), Nepal. As a part of the bilateral co-operation project between Governments of Nepal and Finland, these data were measured in a field campaign conducted in 2010-2011. Several forest type including Sal - Shorea robusta (S), Acacia catechu and Dalbergia sisso (KS/SK), Tropical mixed hardwoods (TMH), Lower mixed hardwood (LMH), Pinus roxburghii (Pr) and Agricultural area (Ag) were the classes collected among the 358 plots measured in the above mentioned field camp.

Tree parameters including tree height, diameter at breast height (DBH), and tree species were collected within a circular plot of 20m radius for each plot in order to estimation the forest biomass using the allometric equations which gives a plot-wise AGB in tons per hectare(t/ha).

3.3 Proposed Methodology

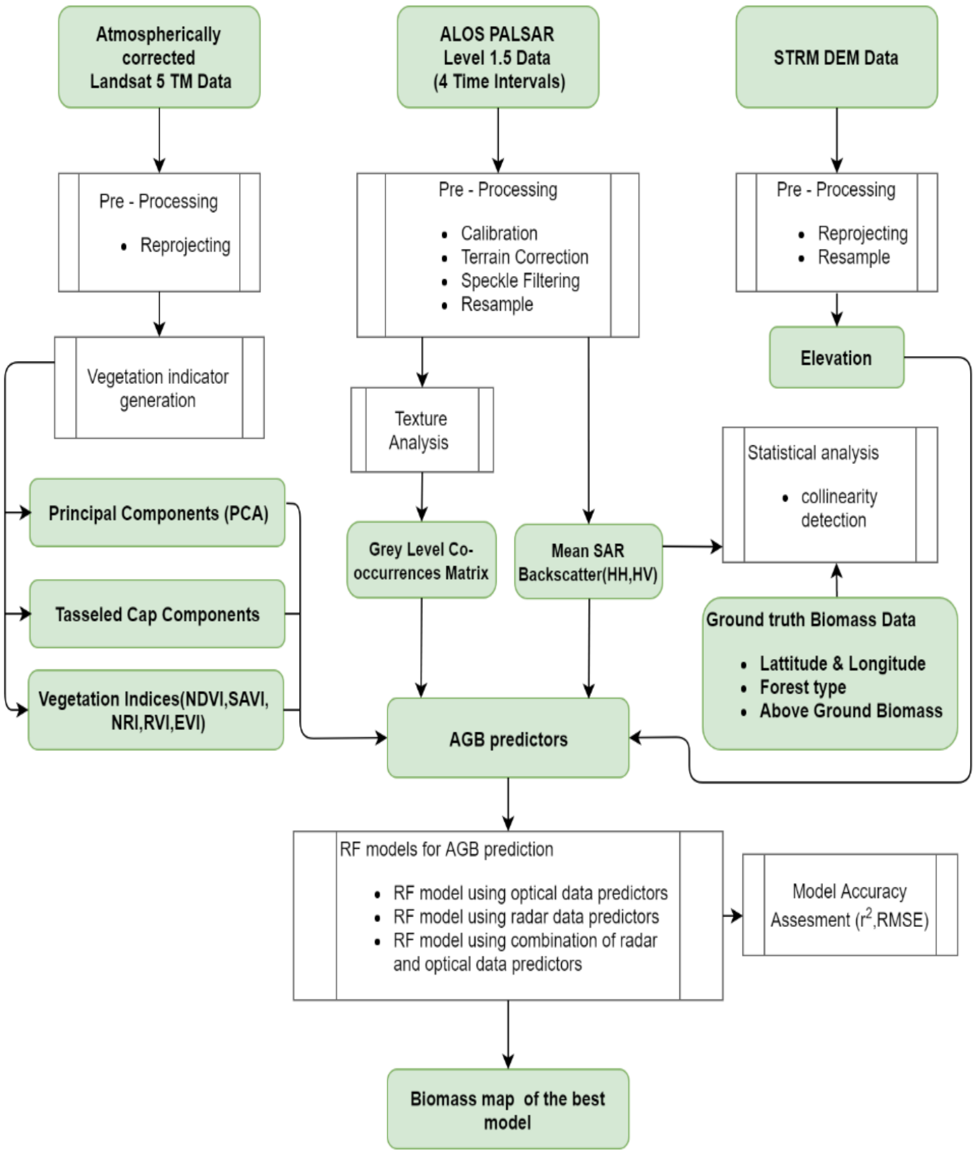


Figure 2: Flow chart of the utilized methodology

3.3.1 Radar Data Processing

Each ALOS PALSAR scene was preprocessed in order to get the Sigma Naught Backscatter value. Terrain correction, co-registration, radiometric calibration, and speckle filtering are the necessary processing involved. ASF MapReady software was used to execute the terrain correction to the radar imageries using SRTM 90m DEM. Radiometric calibration was done using the same software. SNAP

6.0 software was used for the speckle filtering process. Radar texture parameters for this study were generated from the Grey Level Co-occurrence Matrix (GLCM) option in SNAP software.

Geocoded and pre-processed multitemporal imageries were combined and obtain a mean Radar backscatter image for every two mainframes (518,519). These final image products were resampled into the same pixel size (30m) to match with the optical data as well as the AGB field plot size.

3.3.2 Google Earth Engine (GEE) Analysis with Optical Data

Atmospherically corrected and a cloud-free Landsat 5TM image was selected from the google earth engine. All the vegetation indices (see table 2), tasseled cap components and principle components of Landsat 5 ETM sensor bands were generated from google earth engine web platform.

The principal components (PC) transform (also known as the Karhunen-Loeve transform) is a spectral rotation that takes spectrally correlated image data and outputs uncorrelated data. The PC transform accomplishes this by diagonalizing the input band correlation matrix through Eigen-analysis. This process was carried out for the Landsat 5TM image in GEE and obtained PCA1 ,PCA2, and PCA3 which can spectrally characterize vegetation and other classes(Deus, 2016). Landsat 5 tasseled cap (TC) coefficients were also input to the same GEE platform, and according to the spectral characteristics of the optical bands, the brightness, greenness and wetness components were derived. All the vegetation indices mentioned in table 2 are generated using simple band math operations, and all these optical parameters were downloaded after clipping all the products according to our study area.

Table 2: AGB predictors for the Random Forest Models

Data	Relevant Predictors	Description
ALOS PALSAR	Polarization	HH HV
	Texture	HH_Contrast HH_Dissimilirity HH_Homogenity HH_Angular Second Moment HH_Energy HH_Maximum Probability HH_Entropy HH_GLCM Mean HH_GLCM Variance HH_GLCM Correlation HV_Contrast HV_Dissimilirity HV_Homogenity HV_Angular Second Moment HV_Energy HV_Maximum Probability HV_Entropy HV_GLCM Mean HV_GLCM Variance HV_GLCM Correlation
Landsat 5TM	Vegetation Indices	NDVI (NIR - RED)/(NIR + RED) SAVI ((NIR–RED) / (NIR + RED + 0.5) × 1.5) NRI (GREEN - RED)/(GREEN + RED) RVI (NIR/RED) EVI (2.5(NIR- RED)/(NIR + 2.4 RED + 1))
	Tasseled Cap Indices	Brightness Greenness Wetness
	Principle Components(PCA)	PCA1 PCA2 PCA3
SRTM DEM	Digital Elevation Model	Elevation

3.3.2 Model generation from Random Forest algorithm

Random Forest (RF) is an efficient machine learning algorithm that can be used both for classification and regression applications. RF has proven to obtain highly accurate results, find the robust outliers and noise and show the relative importance of input variables(Wang *et al.*, 2016).
In this process bagging algorithm is used to subset training dataset into sub-training datasets which are called bootstrap datasets. Each bootstrap sample based on a decision tree using a classification or regression tree algorithm. All these decision trees are merged to generate the random forest. In general, two-thirds of the total samples from the training dataset should be included in these bootstrap datasets and called 'in bag' data while the remaining data is called 'out-of-bag' (OOB) data and is used to evaluate the RF model.

In this study, focus has been given for three different random forest models. Model based on optical data (12 predictor variables), model based on radar data(22 predictor variables) and model based on combination of radar and optical data(34 predictor variables) are the main models mentioned above. There were 208 sample data for each random forest model which includes relevant AGB predictors according to the random forest model.
All these models were executed in R and RandomForest, and CARET packages were used in this study. First the data set were separated into training and test datasets. Then random forest parameters (mtry, ntree, etc.) were changed accordingly and tuned parameters in order to get the maximum out of these models.

4. RESULTS AND DISCUSSION

4.1 Above Ground Biomass Field Data

In this study, ALOS PALSAR and Landsat 5 data were combined with the forest plot biomass data acquired from the field. Below boxplot represents the selected AGB plot data distribution which is used in the random forest models. The most significant number of plots lies in the TMH (Tropical Mixed Hardwoods) forest type with 161 plots, followed by Sal forest (40 plots), KS/SK (4 plots), agriculture (2plots), and others (1 plot). The minimum AGB value is 0.987 tons/ha, and maximum is 599.65 tons/ha, with mean value is 231.58 ton/ha. From Figure 3, AGB in Sal forest is the highest (average of more than 250 tons/ha), followed by TMH, KS/SK, Agriculture, and others. For this study AGB plots were chosen according to the highest biomass value and the plots which have more than 600 t/ha AGB quantity, were removed before using for the model training and correlation estimations. Finally, there were 208 forest field plots covering the study area.

.

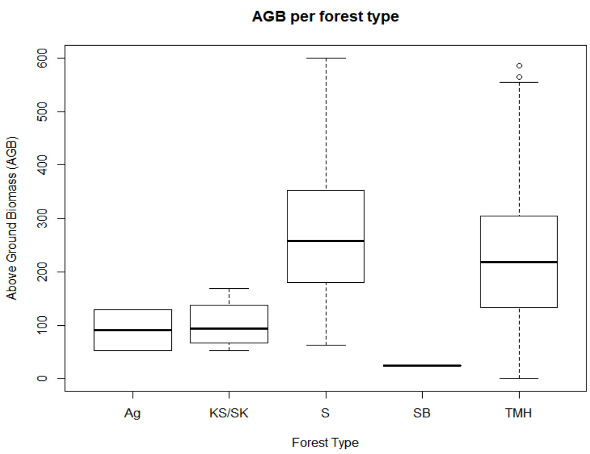
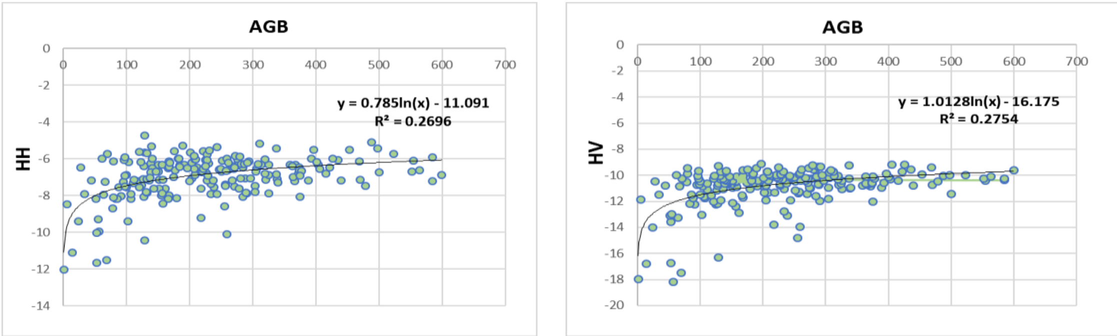


Figure 3: Box-plot of the calculated AGB from the field data

4.2 Relation Between Radar Backscatter and AGB

In this study, the above-ground biomass estimation tested with the relation between RADAR backscatter and the above-ground biomass filed data. When analyzing the relations, non-linear model which has the highest accuracy in terms of correlation was also plotted in the same plot. When analyzing radar data, backscatter noise can cause lack of correlation between AGB and backscatter. In that case, acquiring multitemporal images and obtaining an image, which has mean radar backscatter can reduce the effect of radiometric noises as well as the RADAR backscatter changes due to the climatic variations(Nashrrullah *et al.*, 2012).



(a) (b)

Figure 4: Best fit curves for the scatter diagrams between (a)AGB and HH backscatter and (b)AGB and HV backscatter

According to these 208 plotted points and the R^2 values, it is shown that non-linear regression curves are not perfectly fit the data. When comparing with the HV and HH polarization, HV polarized backscatter gives the best estimate as it has a higher R^2 value (0.275). Above-ground biomass saturation with L band is clearly visible after 100 t/ha level. These polarization qualities can be further described as texture parameters for random forest modeling.

4.3 Random Forest Model Training and Validation

Random forest modeling results were taken separately for radar model, optical model, and combination of radar and optical model. When considering the accuracy of these three models it is clearly shown that the Optical model gives an accurate estimation (R^2 = 0.443 ,MAE = 78.53) of the above-ground biomass. The radar data model has the lowest performance in terms of accuracy (R^2 = 0.227).

Table 3: Results obtain from RF models

AGB Model	Training Dataset			Validation Dataset		
	R^2	RMSE	MAE	R^2	RMSE	MAE
RADAR	0.916	55.125	44.203	0.227	113.787	89.333
RADAR+Optical	0.925	51.78	39.94	0.391	103.523	80.774
Optical	0.905	56.61	43.387	0.443	101.459	78.526

When comparing the non-linear biomass model approach (which is based on the polarized radar backscatter with the AGB values) and the Radar data based RF model, the non-linear model has a good performance. But both approaches have a low correlation value. The combination of radar data and optical data can increase the accuracy of the biomass estimation model (R^2 = 0.391) than it only based on Radar parameters.

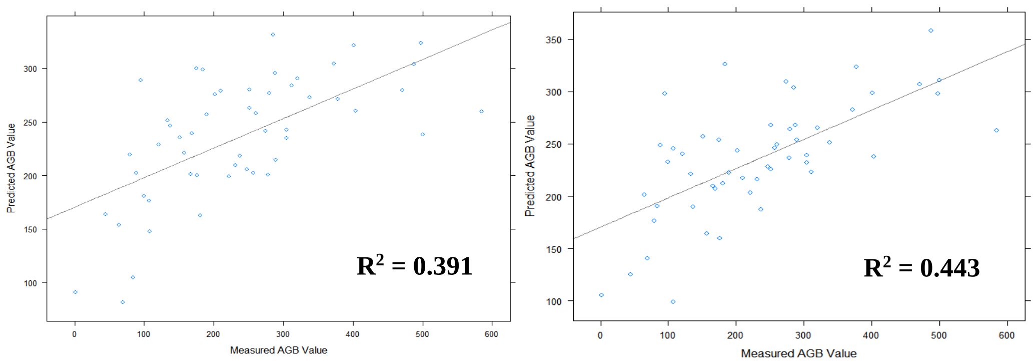


Figure 5: Predicted vs Measured Biomass Graph (a) fusion model (b) optical model

The above graphs show the linear relation between predicted and measured biomass of optical and fusion models. When considering the accuracy of these models, we can use these models to map the biomass distribution despite the complex topography and the complex structure of the forest which affect the accuracies of these models. Multitemporal image generation to reduce the climatic effects on the radar backscatter and the effectiveness of the topographic correction used for the radar data may not be sufficient for particular purposes. Insufficient AGB field data for the random forest model generation can also be a reason for the low performances of these models. Geometric registration errors of the AGB field plots due to plot location establishment from low accurate hand-held GPS instruments can make uncertainties in the biomass estimation process.

Table 4: Variable importance obtain from RF models

Variable	RF Model	
importance	Optical	Optical+RADAR
1	PCA1	HH_GLCM Variance
2	Brightness	HH_GLCM Mean
3	PCA2	HV_GLCM Variance
4	NDVI	HH_GLCM Correlation
5	Greenness	HV
6	EVI	HH
7	RVI	Greenness
8	SAVI	PCA1
9	Wetness	HV_GLCM Correlation
10	PCA3	Wetness
11	NRI	HH_Homogenity
12	Elevation	HV_Contrast

When considering the variable importance gain from the random forest models, there are specific AGB predictors which come across in both optical and fusion model. Principle component1, Greenness, and Wetness from optical data and radar texture parameters were ranked higher in terms of increase in mean square error of predictions.

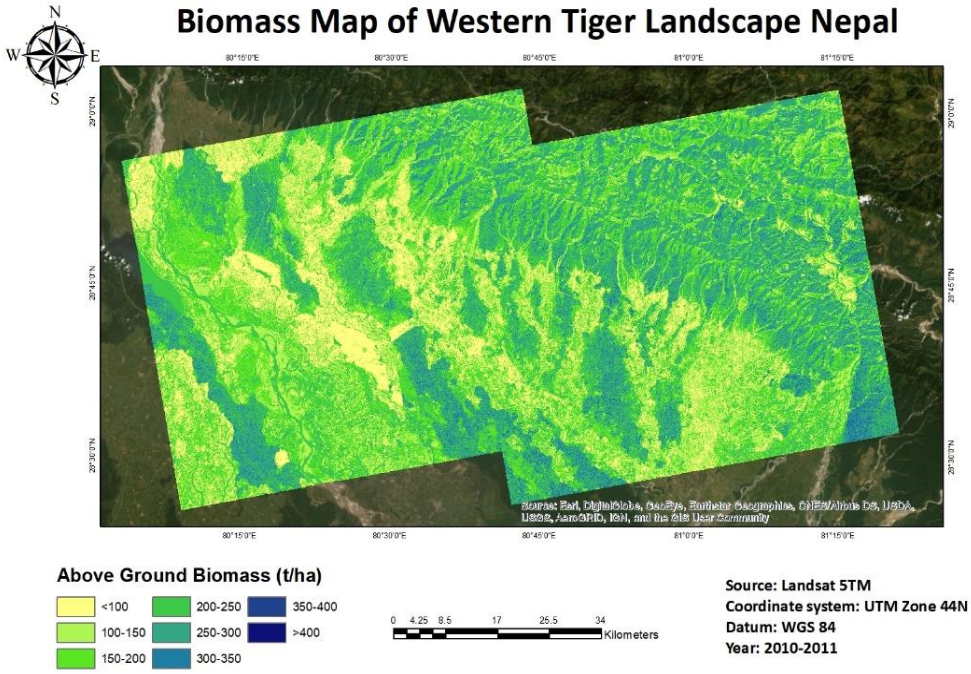


Figure 6: Biomass map derived from the optical data model

5. CONCLUSIONS AND RECOMMENDATIONS

This study investigated the potential of data integration of ALOS PALSAR and Landsat 5 data for estimating the forest biomass of the Western Tiger Landscape of Nepal using random forest method. The performance of the RF Models based on optical data, radar data and a combination of radar and optical data is also tested. Based on the finding in this research, the following conclusions are drawn:

- The Landsat 5 data could be used to estimate the forest AGB with moderate accuracy, while the ALOSPALSAR data alone is not enough for estimating the forest AGB.
- When combining ALOS PALSAR data with Landsat 5 optical data, the RF model gives best results than AGB estimation only based on Radar backscatter.
- Principle components, tasseled cap components, and radar texture parameters play an essential role when predicting the results.

In complex terrain and forest stand structures, it is better to combine the radar and optical data to utilize top of canopy detection from optical data and radar penetration capabilities from radar data. Even though optical radar fusion makes better estimations, correcting the topography of the radar data is quite challenging. More care should be taken to remove the terrain effects with an appropriate model-based slope correction.

SAR data has the limitation of signal saturation when sensing high biomass values, and this study also clearly shows the L band biomass saturation at 100-150 ton/ha. Using longer wavelengths like P band for the biomass estimation studies can extend the saturation limit. Further studies on biomass estimation can be done combining Lidar derived data and Radar polarimetric parameters along with the optical Data.
Artificial neural networks and support vector machine are two other efficient machine learning approaches that can be used in the biomass modeling process.

The optical random forest biomass model successfully produced a biomass map that indicates the spatial distribution of forest biomass on a large scale, but the accuracy was relatively low when comparing it with other biomass studies. Since the traditional filed AGB estimation techniques are time consuming and expensive, remote sensing methods still can provide the spatial distribution of the biomass on a large scale effectively.

6. ACKNOWLEDGMENT

The ALOS/PALSAR data and the AGB ground truth data was obtained from a JAXA Mini Project 2011-2012 carried out for estimating above ground forest biomass of the Western Tiger Landscape of Nepal in 2011

7. REFERENCES

References from journal:

Deus, D. (2016) 'Integration of ALOS PALSAR and Landsat data for land cover and forest mapping in northern Tanzania,' Land, 5(4). doi: 10.3390/land5040043.

Duchelle, A. E. et al. (2018) 'What is REDD+ achieving on the ground?', Current Opinion in Environmental Sustainability. Elsevier B.V., 32, pp. 134–140.

Nashrrullah, S. et al. (2012) 'Estimation of above ground forest biomass in a tiger Habitat of the Western Nepal using alos data and field inventory', 33rd Asian Conference on Remote Sensing 2012, ACRS 2012, 2(January), pp. 1156–1163.

Vaglio, G. L. et al. (2017) 'Potential of ALOS2 and NDVI to estimate forest above-ground biomass, and comparison with lidar-derived estimates', Remote Sensing, 9(1). doi: 10.3390/rs9010018.

Wang, L. et al. (2016) 'Estimation of biomass in wheat using random forest regression algorithm and remote sensing data', Crop Journal. Elsevier B.V., 4(3), pp. 212–219.