RANDOM FOREST CLASSIFICATION OF JAMBI AND SOUTH SUMATERA USING ALOS PALSAR DATA

Mulia Inda Rahayu[1*] and Katmoko Ari Sambodo[1]
[1]Remote Sensing Technology and Data Center, LAPAN, Jakarta
[*]E-mail: iinsudhie@yahoo.com

Abstract. Recently, Synthetic Aperture Radar (SAR) satellite imaging has become an increasing popular data source especially for land cover mapping because its sensor can penetrate clouds, haze, and smoke which a serious problem for optical satellite sensor observations in the tropical areas. The objective of this study was to determine an alternative method for land cover classification of ALOS-PALSAR data using Random Forest (RF) classifier. RF is a combination (ensemble) of tree predictors that each tree predictor depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In this paper, the performance of the RF classifier for land cover classification of a complex area was explored using ALOS PALSAR data (25m mosaic, dual polarization) in the area of Jambi and South Sumatra, Indonesia. Overall accuracy of this method was 88.93%, with producer's accuracies for forest, rubber, mangrove & shrubs with trees, cropland, and water classes were greater than 92%.

Keywords: *Land cover, ALOS-PALSAR, random forest (RF), classification, remote sensing*

1 INTRODUCTION

Land cover mapping and monitoring is one of the major applications on Earth observing satellite sensor data and is essential for the estimation of land cover change (Rodriguez-Galiano *et al*., 2012). Increased numbers of satellite sensor images have made easier to establish land cover monitoring programs for large area mapping over regular time intervals (Friedl *et al*., 1999).

Optical sensors such as Landsat TM/ETM+ and SPOT have proven an efficient tool for various applications like land cover mapping, change detection and disaster control (Huang *et al*., 2007). These optical sensor have limitations in acquiring cloud free imagery on a regular basis and difficulties in performing spectral classification for certain types of land features.

In addition to optical sensors, microwave sensors has become an increasing popular data source especially for land cover mapping. Compared to optical sensors, active microwave sensors can provide their own illumination. The longer wavelengths enable penetration of atmospheric condition such as rain, sleet, fog, haze, smoke, precipitation, and clouds (Haack *et al*., 2000). The advantage of active microwave sensors such as Synthetic Aperture Radar (SAR) is their ability to obtain images under various weather conditions during both day and night time.

Land cover map can be generated by digital image classification of remote sensing data. A variety of classification methods, from traditional per-pixel based parametric algorithm such as maximum likelihood, to advanced non parametric algorithm such as neural network, support vector machine, and decision tree have been used to map land cover using remote sensing data. Non-parametric classifier have increasingly become important approaches for multisource data classification (Lu and Weng, 2007). Machine learning algorithms have become more accurate and efficient alternatives to conventional parametric algorithm, when faced with large dimensional and complex data spaces and have been used for large area mapping (Huang *et al*., 2002).

An ensemble learning technique called Random Forest (RF) is known to be one of the most efficient classification methods. Ensemble learning techniques have higher accuracy than other machine learning algorithms because the group of classifiers performs more accurately than any single classifier (Ghimire *et al.*, 2010; Akar and Güngör, 2012). Akar and Güngör (2012) reported that for IKONOS image over urban area, RF algorithm gives 10% higher classification accuracy than Support Vector Machine algorithm, whereas Gentle and Boost algorithm has the lowest classification accuracy (14% lower than RF). The aim of this paper was to explore the use of Random Forest algorithm for land cover mapping using ALOS-PALSAR 25m dual polarization mosaic data in part of Jambi and South Sumatera Province, Indonesia for the year of 2010.

2 MATERIALS AND METHOD

2.1 Data

The SAR data used in this study is shown in Figure 1. The SAR data were mosaic ALOS-PALSAR data, 25 m resolution, L-band, dual polarization (HH-HV). The research area was part of Jambi and South Sumatera Province, Indonesia. These data were acquired in year 2010 by Advanced Land Observing Satellite (ALOS) and pre-processed (orthorectification and slope correction) by JAXA-EORC (Japan Aerospace Exploration Agency – Earth Observation Research Center).

Figure 1 showed a set of ground survey information and the PALSAR image. There were some different types of land cover: forest, swamp forest, acacia, rubber, mangrove, shrubs, oil palm, coconut, cropland, bare soil, settlement, and water area. By analyzing these data, a set of regions of interest (ROI) was defined. The entire ROI datasets would be divided into two datasets. Approximately 60% of ROI datasets were used for training and remaining 40% of ROI data were used for testing the RF classifier. From the testing dataset, the classification accuracy based on analysis of the confusion matrix can be estimated.

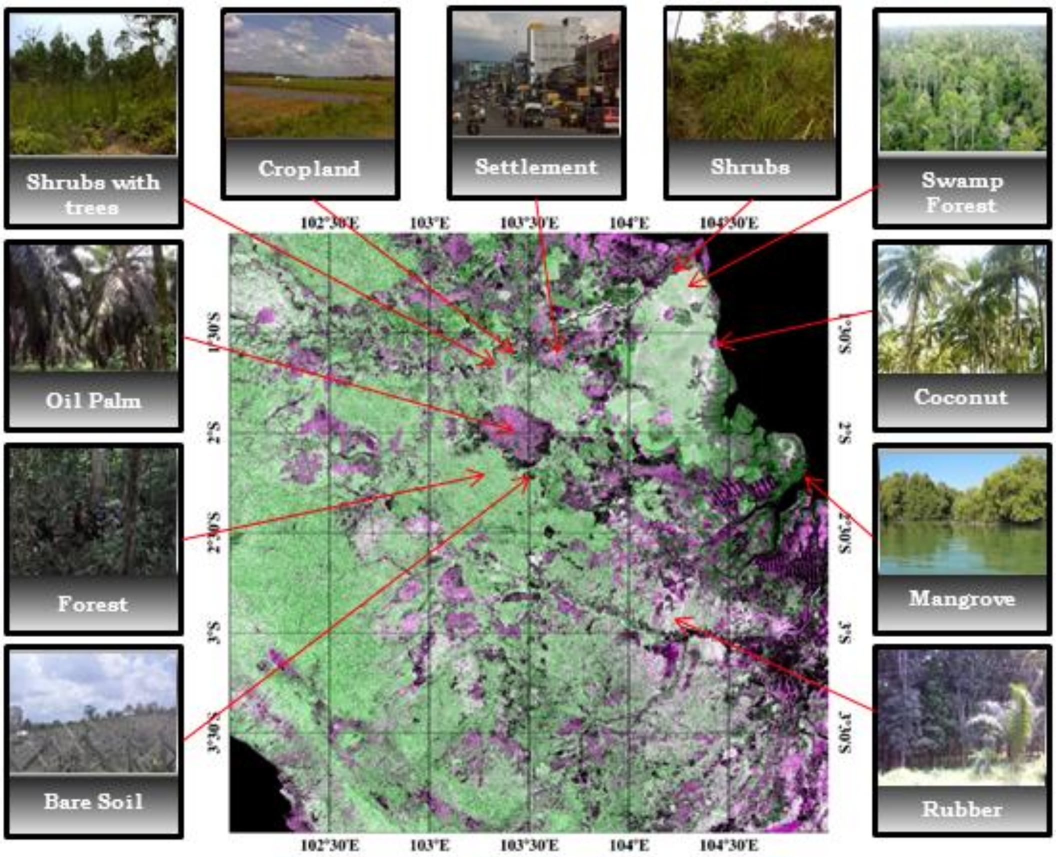


Figure 1. ALOS PALSAR 25m mosaic data and ground survey information in part of Jambi and South Sumatera Province.

2.2 Data Analyses

Figure 2 showed the flowchart of the ALOS-PALSAR classification used in this study. It started with the conversion of Digital Number (DN) of ALOS-PALSAR data to Gamma Naught γ^0 in decibel unit, which was defined as radar backscatter per unit area of the incident wavefront (perpendicular to slant range) (Motohka, 2012):

$$\gamma^0 = 10 * \log_{10}\langle DN^2 \rangle + CF \quad [\text{dB}] \qquad (1)$$

Where, the calibration factor, $CF = -83.0[\text{dB}]$, and $\langle ... \rangle$ represent averaging over 3x3 window size.

Based on the ground survey information, at least 10 ROIs were selected for each class. The statistics (mean and variance-covarian) were calculated and plotted in HV-HH feature space as an ellipse. Each ROI should be in small ellipse shape which was indicate that the selected samples were quite homogeneous. This selection and evaluation process should be done iteratively. When two or more classes were highly overlapping, these classes would be grouped into a single class. It was better to obtain high classification accuracy with less number of classes, rather than used the entire class information but with low accuracy. Approximately 60% of ROI datasets were used for training and remaining 40% of ROI data were used for testing the RF classifier. Once the training samples for each class have been generated, the RF classification was then performed. Random Forest classification method will be described in the following sub section.

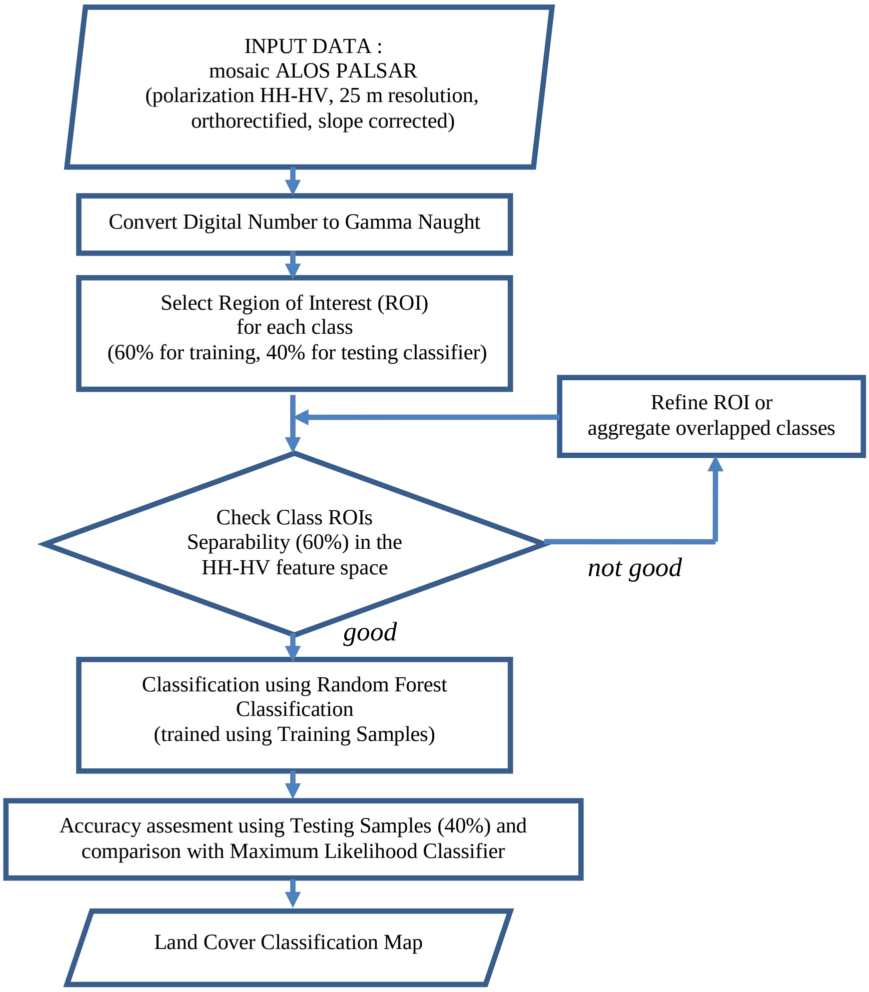

INPUT DATA :
mosaic ALOS PALSAR
(polarization HH-HV, 25 m resolution,
orthorectified, slope corrected)
Convert Digital Number to Gamma Naught
Select Region of Interest (ROI)
for each class
(60% for training, 40% for testing classifier)
Refine ROI or
aggregate overlapped classes
Check Class ROIs
Separability (60%) in the
HH-HV feature space
not good
good
Classification using Random Forest
Classification
(trained using Training Samples)
Accuracy assesment using Testing Samples (40%) and
comparison with Maximum Likelihood Classifier
Land Cover Classification Map

Figure 2. Flowchart of the ALOS-PALSAR classification method.

2.3 Random forest classification

RF is a combination (ensemble) of tree predictors, which each tree predictor depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). A random forest classifier consists of a collection of tree-based classifiers as follows:

$$\{h(x, \Theta k), k = 1,...\}$$

where x is the input vector and Θk is independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x (Breiman, 2001).

The classification worked as follows: the random trees classifier took the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". During the training, a different subset of training data were selected with replacement to train each tree, while remaining training data were used to estimate error and variable importance.

Random forests used bootstrap samples with replacement to grow a large collection of classification trees, which assigned each pixel to a class based on the maximum number of votes that a class receives from the collection of trees. Random forests did not overfit and it was very fast, so it was possible to run as many trees as user want (Breiman and Cutler, 2005).

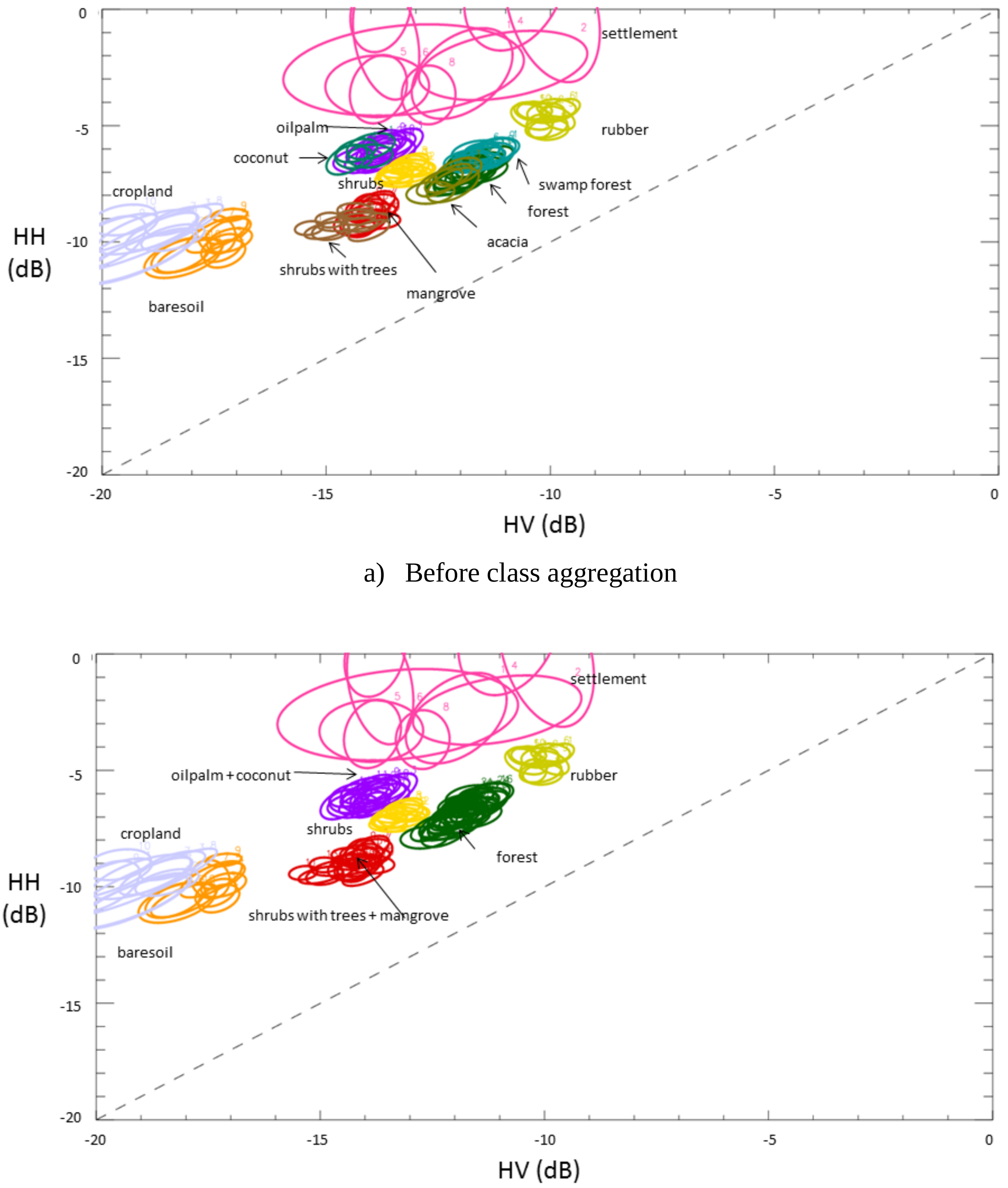
After the RF classification process was completed, the classification accuracy was then estimated using confusion matrix. The RF classification result was compared with the result obtained from Maximum Likelihood algorithm.

3 RESULTS AND DISCUSSION

Before the classification process, we evaluated the training sample. As a preliminary step, based on field survey information, there were 13 land cover classes that can be found, namely forests, swamp forests, acacia, rubber, mangrove, shrubs, shrubs with trees, oil palm, coconut, cropland, bare soil, settlement, and water area. From HV-HH feature space plot (Figure 3a), it was noted a lot of overlapping classes. For example, the forest class was overlapped with acacia and swamp forest class. The oil palm plantation class was overlapped with coconut plantation class. The mangrove class was overlapped with "shrubs with trees" class. These overlapped classes complicated the classifier in determining the optimum class boundaries and decreased the classification accuracy. Therefore, the overlapped classes were grouped into one class (Figure 3b). Forest, swamp forest, and acacia were grouped into "forest" class. Similarly, mangrove and shrubs with trees were grouped into "mangrove + shrubs with trees" class. Oil palm and coconut were also grouped into "oil palm + coconut" class. As the result of class aggregation, there were nine land cover classes namely forest, rubber, mangrove+shrubs with trees, oil palm+ coconut, shrubs, cropland, bare soil, settlement, and water class.

In the HV-HH feature space, water class was not visible because the position of the water class was on the bottom left in the feature space and was far from the other classes (under -20 dB) (Figure 3a, 3b).

The classification results using RF classifier is shown in Figure 4, and the corresponding confusion matrix is presented in Table 1. Water, rubber, cropland, and mixed mangrove+shrubs with trees can be separated. The forest could be separated with other classes, but some misclassification between forest, rubber, mangrove & shrubs with trees were also occurred, mainly due to their similar radar backscattering characteristics. The shrubs and baresoil couldnot be well identified by RF classifier. For shrubs class, there was 294 pixels identified as forest, and 182 pixels indentified as oilpalm. This could be due to some mix pixels in those objects. As well as shrubs, for baresoil class there was 172 pixels identified as cropland.



a) Before class aggregation

b) After class aggregation

Figure 3 The HV-HH feature space plot of the class sample ROIs (Region of Interest).

To evaluate the performance of RF algorihm, we try to compare this result with the result obtained from Maximum Likelihood algorithm (Table 2). The classification result using Maximum Likelihood is shown in Figure 5. The RF method produced better overall classification accuracy (88,93%). From the RF result, producer's accuracy of forest (92.17%), mangrove+shrubs with tree (93.91%), oilpalm+coconut (87.28%), cropland (94.13%), and settlement (83.87%) class were higher than the Maximum Likelihood result. From the RF result, the oilpalm+coconut could be well separated from other classes. There were 1908 pixels identified as oilpalm+coconut, 182 pixels identified as shrubs, but none of the pixel identified as rubber, cropland, baresoil and water. While from the maximum likelihood result, there were 1727 pixels identified as oilpalm+coconut, 287 pixels identified as shrubs, 79 pixels identified as baresoil, and none pixel identified as rubber and water.

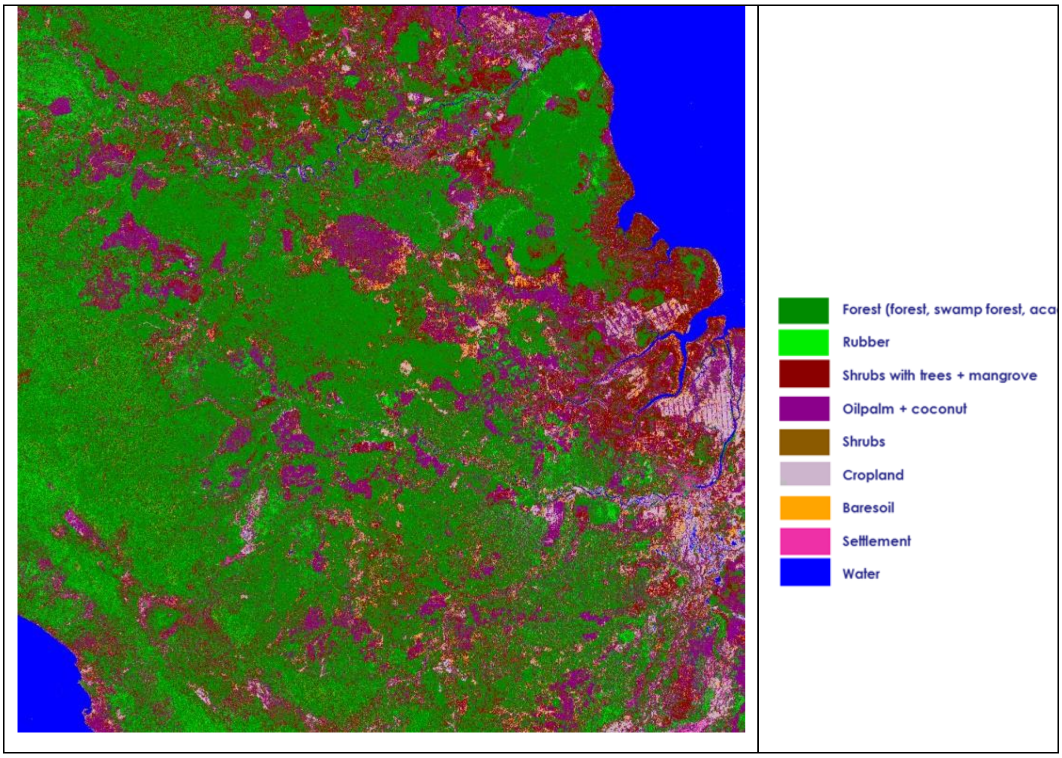

Forest (forest, swamp forest, aca
Rubber
Shrubs with trees + mangrove
Oilpalm + coconut
Shrubs
Cropland
Baresoil
Settlement
Water

Figure 4. Classification result using Random Forest classifier.

Table 1. Confusion matrix of Random Forest classifier.

Reference Data Classified Data	Forest	Rubber	Mangrove + shrubs with trees	Oilpalm + coconut	Shrubs	Cropland	Baresoil	Settlement	Water	User's accuracy
Forest	4428	49	47	24	294	0	0	35	0	90.79
Rubber	87	1314	0	0	0	0	0	43	0	91.00
Mangrove + shrubs with trees	108	0	2259	7	21	2	5	0	0	94.05
Oilpalm + coconut	20	0	39	1908	182	11	0	116	0	83.83
Shrubs	154	0	71	182	689	0	0	1	0	62.81
Cropland	0	0	0	0	0	1234	172	0	0	87.77
Baresoil	0	0	5	0	0	64	341	0	0	83.17
Settlement	7	26	0	65	1	0	0	1014	0	91.11
Water	0	0	0	0	0	0	0	0	1583	100.00
Producer's accuracy	92.17	94.60	93.91	87.28	58.05	94.13	65.83	83.87	100.00	

Overall accuracy: 88.93%   Kappa Coefficient = 0.868

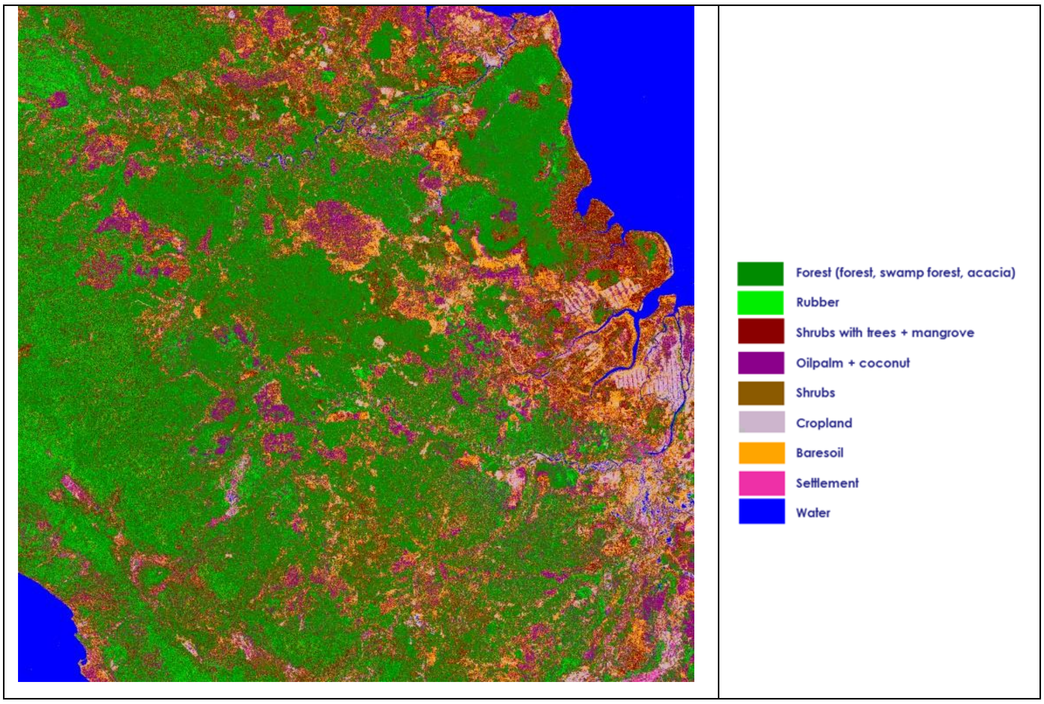

Forest (forest, swamp forest, acacia)
Rubber
Shrubs with trees + mangrove
Oilpalm + coconut
Shrubs
Cropland
Baresoil
Settlement
Water

Fig. 5. Classification result using maximum likelihood classifier.

Table 2. Confusion matrix of Maximum Likelihood classifier

Reference Data Classified Data	Forest	Rubber	Mangrove + shrubs with trees	Oilpalm + coconut	Shrubs	Cropland	Baresoil	Settlement	Water	User's accuracy
Forest	4083	62	25	17	145	0	0	41	0	93.97
Rubber	138	1322	0	0	0	0	0	124	0	83.46
Mangrove + shrubs with trees	161	0	2079	1	16	0	0	0	0	92.11
Oilpalm + coconut	15	5	6	1727	124	1	0	181	0	83.88
Shrubs	403	0	71	287	860	0	0	3	0	52.96
Cropland	0	0	0	1	0	1232	237	0	0	83.81
Baresoil	4	0	240	79	42	78	281	23	0	37.62
Settlement	0	0	0	74	0	0	0	837	0	91.88
Water	0	0	0	0	0	0	0	0	1583	100.00
Producer's accuracy	84.99	95.18	85.87	79.00	72.45	93.97	54.25	69.23	100.00	

Overall accuracy: 84.32% Kappa Coefficient = 0.816

4 CONCLUSION

Using Random Forest algorithm to generate land cover classification from ALOS PALSAR 25m mosaic data, we produced nine different classes i.e., forest, rubber, mangrove & shrubs with trees, oilpalm & coconut, shrubs, cropland, bare soil, settlement, and water. The results of RF algorithm were compared with the results of maximum likelihood algorithm. The result showed that oilpalm+coconut class can be well separated from other class. The RF produced better performance with 88.93% overall accuracy (Kappa value = 0.868) than maximum likelihood, while producer's accuracies for forest, mangrove + shrubs with trees, oilpalm + coconut, cropland, and settlement classes were higher than maximum likelihood result.

ACKNOWLEDGMENT

The authors would like to thank JAXA for providing the ALOS PALSAR 25m mosaic data within the framework of the JAXA Kyoto & Carbon Initiative.

REFERENCES

Akar and Güngör, 2012, Classification of multispectral images using Random Forest algorithm, *Journal of Geodesy and Geoinformation,* 1(2):105–112.

Breiman, 2001, Random forests. *Machine Learning,* 45:5–32.

Breiman, L and A. Cutler, 2005, Random forests, http://www. stst. berkeley. edu/ ~breiman/RandomForests/cc_home.htm [accessed on 11 July 2013].

Friedl, M.A., C.E. Brodley, and A. Strahler, 1999, Maximizing land cover classification accuracies produced by decision tree at continental to global scales, *IEEE Transactions on Geoscience and Remote Sensing,* 37(2):969–977.

Rodriguez-Galiano, V.F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J.P Rigol-Sanchez, 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.

Ghimire, B., J. Rogan, and J. Miller, 2010, Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic, *Remote Sensing Letters,*1:45–54.

Haack, B.N., N.D. Herold, & M.A. Bechdol, 2000, Radar and optical data integration for landuse/ land-cover mapping, *Photogrammetric Engineering & Remote Sensing,* 66(6):709–716.

Huang, C., L.S. Davis, and J.R.G. Townshend, 2002, An assessment of support vector machines for land cover classification, *International Journal of Remote Sensing,* 23(4):725–749.

Huang, H., J. Legarsky, M. Othman, 2007, Land-cover classification using Radarsat and Landsat Imagery for St.Louis, Missouri. *Photogrammetric Engineering & Remote Sensing,* 73(1):037-043.

Lu and Weng, 2007, A survey of image classification methods and techniques for improving classification performance, *International Journal of Remote Sensing,* 28(5):823–870.

Motohka, T., 2012, Introduction on forest change mapping using PALSAR gamma-naught change, *International Workshop and Training on Pi-SAR-L2 Data Analysis for Forest Carbon Monitoring, Ship Detection, Disaster Monitoring, Geometric Evaluation, and Crop Monitoring (JAXA Training Materials).*