

Forest over classification from multi-temporal MODIS images in Southeast Asia using decision tree

Sijie Wu¹, Jianxi Huang², Xingquan Liu¹, Guannan Ma²

¹ School of Geosciences and Info-Physics, Central South University, 410083, Changsha, China

² College of Information and Electrical Engineering, China Agricultural University, 100083, Beijing, China

{vr_hjx@163.com, dx01wusijie@163.com}

Abstract: MODIS data is of significant for the classification of regional forest cover due to its high temporal resolution and high spectral resolution. Forest cover is an important parameter for forest ecosystem. The objective of this preliminary study is to mapping forest cover from mutli-temporal MODIS data with decision tree. The classification forest samples were selected from four global land cover datasets with specific rules. The selected samples were used to generate rules of the decision tree for the classification of forest cover. The study results show that mutli-temporal remote sensing data with decision tree method have great potential to improve the regional forest cover mapping.

Keywords: Multi-temporal; forest cover; decision tree

Introduction

Identification of types of forest has significant for forestry source monitoring and management. Because of the capability of acquiring regional surface information, remote sensing has become a reliable tool for identifying types of forest in regional and global scales. Currently, most of the applications of remote sensing classification are the traditional statistical pattern recognition method, such as minimum distance, parallelepiped, maximum likelihood, and mixed-distance method, cyclic cluster method and other supervision or unsupervised classification method. Because of the existence of spatial resolution of remote sensing image itself and "same object with different spectrum", "different objects with same image" phenomenon, misclassification and leakage of points occur more frequently. These factors lead to the low classification accuracy. New methods of pattern classification are as follows: fuzzy classification, classification based on texture description of Markov random field model, classification of wavelet analysis, fractal texture method, neural network and expert system classification, etc[1-3]. Currently, remote sensing information composite technique is widely used [4, 5]. In recent years, researchers primarily utilize satellite remote sensing with vegetation surface temperature, terrain elements and other non-remote sensing of forest vegetation on the ground to identify sub-categories [6].

In the past decades, applied research results on the large area forest cover mainly

using AVHRR data has achieved great success[9,10]. Due to limitations AVHRR data for land cover mapping applications, there are still many uncertainties [11]. Thus, with the launch of Terra satellite, the use of MODIS data in regional scale studies of forest cover has been developed. Using MODIS global supervised classification model, Muchoney [12] classified the vegetation and land cover in central United States. On this basis, with IGBP classification system and the STEP global plots database and MLCCA (MODIS land cover classification algorithm method. Friedl [13] carried out a global land cover classification with a total of five months of MODIS data in 2000. Based on pattern decomposition method (PDM), Cen[14] conducted a study of land cover classification using MODIS data of the Kii Peninsula of Japan in 2001. Using MODIS 8 day composite reflectivity products, Carrão [15] evaluated the efficiency of MODIS hyper spectral data and relative land cover classification for a long time. On the basis of studying and comparing the abroad classification algorithm that having good application effects, Wu [16] achieved quantitative judgments on continuous coverage of MODIS data and generated the land use status classification. Liu [17] proposed a classification method using MODIS data to select and extract classification feature and do large area land use/cover classification combined with the multi-temporal characteristics. Classification test was conducted in Shandong Province in China.

As for classification data, MODIS data can provide more data products. Visible data and near infrared data of the MODIS can response the growth characteristics of different vegetation types in different periods well, it is perfectly suited for forest classification. Taking into account of close contact of the distribution of vegetation with the climate and soil, coupled with influence of the climate by the altitude, slope and other terrain factors, it is necessary to add the soil and terrain data in the classification of forest types.

2 Study area and data

There are 11 countries in Southeast Asia: Vietnam, Laos, Cambodia, Thailand, Myanmar, Malaysia, Singapore, Indonesia, Brunei, and the Philippines, Timor-Leste. The special geographical location makes the Southeast Asia having hot and humid climate and lush tropical forests. Wet equatorial climate and a tropical monsoon climate are two types of the Southeast Asia. The main natural vegetation here is the tropical rain forest and tropical monsoon forest. Figure1 show the study area of Southeast Asia.

It can be divided into two sub-areas:

1. Indochina area: The climate here is tropical continental monsoon climate. The climate of Malay Peninsula is wet equatorial climate. Annual rainfall of the Malay Peninsula and the rainy coast of Indochina are tropical rain forest landscape. Indochina with dry and wet season is Tropical monsoon forest landscape. Less rainfall Interior plains and valley are savannah landscape. Indochina base is mountain mixed forest. Coast of North Bay and the Gulf of Siam is filled with mangroves.

2. Southeast Asian island district is also called the Malay Archipelago area. It belongs to maritime equatorial rainy climate. The Philippine islands belong to maritime tropical monsoon climate, mainly for tropical rain forest landscape. Southeast Asia with the Indonesian forest area for first is the world's second-largest

rain forest (after Brazil). Papua new several second is in the western Pacific after Indonesia's second-largest forest resources garden.

In Southeast Asia, land surface is mainly composed by woodland, grassland, residential areas, etc., among which forest cover is the most typical one. In order to improve classification accuracy in the region, a lot of land surface sample data is needed. Because the large scope of the study area and too many type and the far distance, it is difficult to achieve the field sampling of vegetation types. Through data analysis on the existing global surface types, we found the land cover of the typical areas changes less, and many classification products can separate the type of surface area of typical areas very well. Therefore, in areas where vegetation integrated we reference map and interactively select the sample data, and using ROI to define the surface type.



Fig. 1 Study area (Southeast Asia)

In the establishment of classification algorithms, including supervised and unsupervised classification are in common use. In order to improve the accuracy of classification, decision-making regression tree algorithm of supervised classification was used in this research. It is characterized by the better use of sample data and extracting decision to create classification algorithm according to the sample data. The user can understand the decision-making rules well.

Decision tree is a tree structure similar with binary tree or multi-branches tree. The decision tree algorithm is one of the inductive learning algorithms. Induction is getting the regular conclusion by contrast, analysis, summary and summarization from characteristics and attributes of examples that known and

seemingly with no specific order and rules. This conclusion was used to generate decision tree as the rule of interpretation, then classification was carried out on cases out of samples. Early famous decision tree algorithm was ID3 posed by Quinlan [19] in 1993. This is one of the most basic decision tree algorithms. C4.5 algorithm was posed by Quinlan [20] in 1993 as Subsequence of ID3 and became basis of later decision tree algorithms. SLIQ (supervised Learning in Quest) algorithm is a fast scalable decision tree algorithm posed by IBM researchers that suitable for processing the data of large scale. CART decision tree, also called classification regression tree, was a data survey and prediction algorithm posed by Breiman, Friedman, Olshen and Stone in 1984.

QUEST is a new decision tree construction algorithm is formulated by Loh and Shih in 1997. The algorithm separates the variable selection and split point selection, not only suitable for continuous variables, but also for discrete variables, and it overcome the other decision tree algorithms that prefer to choose those predictor variables with more potential split point for classification. So basically it is unbiased on variable selection. And through hyper plane constituted by more than one predictor variable, this decision tree can distinguish between multiple members and non-class members of the class in the feature space. Studies have shown that QUEST decision tree classification computation's speed and accuracy are better than other decision tree construction method.

3 Material and Method

3.1 Data

MODIS standard products which produced by this project were using to do forest classification. Taking into account the classification of forests and other land cover classification using vegetation status, MODIS 13Q products were downloaded, including visible light reflectivity. Considering the region's distribution of the forest influenced by the soil type and altitude, we collected global 10km soil data and STRM-DEM data and extracted the slope information. In order to maintain the consistency of the above data, the data project coordinate system and projection were normalized. Although the MODIS data was standardization, there are still clouds of pollution. Using the SG remove cloud algorithms, NDVI-SG Index data was generated. As the greenness of forests changed little during the growth in the 16 days, the index data was synthesized, that is, two 16-day month synthetic data was made into the index data. The 12th monthly indices data was instead by 23rd layer data. By integrating the above data, the input data for classification software was ready. When carrying out the classification of the data by adding soil analysis, marked polygons were found which classification results were not the same size as same as grid soil data. After analysis, it was found that this is due to their resolution (the MODIS 1/10) was inconsistent with the image data. In the later experiments, the soil data were excluded. Taking into account the amount of data and storage limits, in order to support the Bit types of data files, when storing data, using altitude / 20 which altitude below 5,000 m in the region. Degree

of slope data used (Degree) units. NDVI data was set to 0 when less than 0; Greater than 0, the draw ratio is 250 times.

3.2 Forest Classification Method

Based on four global land cover datasets, the samples with pure pixels were extracted. MERIS with 300 meters was used for validate the accuracy of percentage of each sample. And then the overall classification accuracy was discussed.

3.3 The choice of sample set of forest types

Five global Land Cover datasets, including UMD, IGBP-DIS Cover, MODIS, GLC2000 and MERIS 2005 Global Land Cover data were collected in our study. 1 km land cover map of the first four class of Southeast Asia was encrypted cut for classification of 500 meters. Combined with GLOBCOVER, a standard reference data for classification in this region was formed. Because these data in time is quite different and system of classification is different, classification system of forest mapping in Southeast Asia which was provided by Chinese Academy of Forestry was using for combining land cover types, encoding conversion and forming a comprehensive map of forest types. When producing the integrated forest type's graphs, the types with relatively large probability of appearance were on behalf of the forest types in the region, after the adoption of uniform codes. In order to analyze differences in forest types, using red color to represent the type inconsistent, the white for the same type. Types can be seen change little from on higher ground area; types change more in the flat terrain and more human activities places. Using ENVI-ROI function, the sample data were collected in each typical area. Based on the sample analysis and the decision-making algorithm, regression analysis of data for decision-making was conducted to form the classification decision in the region. Figure2, figure3, figure4 and figure5 and figure 6 represent IGBP-DIS, MODIS, GLC 2000, UMD and MERIS forest cover map respectively. And figure7 show the integrated forest cover results.

Table1 A brief description of 5 global land cover datasets

No	Data name	Data source	Data explanation
1	UMD Global land cover data	University of Maryland (http://www.geog.umd.edu/landcover/1km-map.html)	spatial resolution 1km the 1990s
2	IGBP-DISCover Global land cover data	International Geosphere Biosphere Program (http://edcsns17.cr.usgs.gov/glcc/globe_int.html)	spatial resolution: 1 km the 1990s
3	MODIS Global land	Boston iversity (http://duckwater.bu.edu/lc/datasets.html)	spatial resolution: 1 km The early 21st century

	cover data		
4	GLC 2000 Global land cover data	European Commission, Joint Research Centre (http://www-tem.jrc.it/glc2000/)	spatial resolution: 1 km The early 21st century
5	MERIS 2005 Global land cover data	CNES, CNRS, IRD, Météo-France, and INRA (http://postel.mediasfrance.org/en/DOWNLOAD/)	spatial resolution: 300m 2005-2006

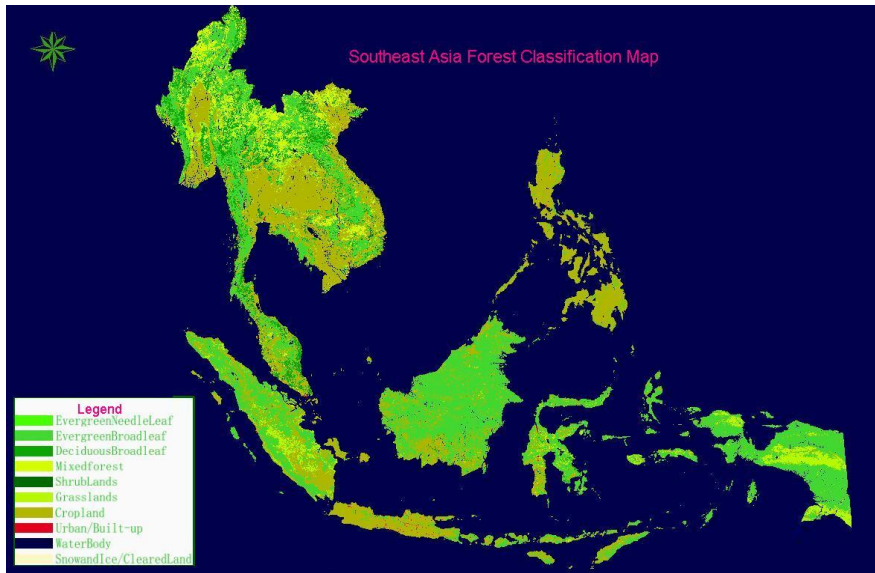


Fig.2 IGBP-DIS Cover Global land coverdata

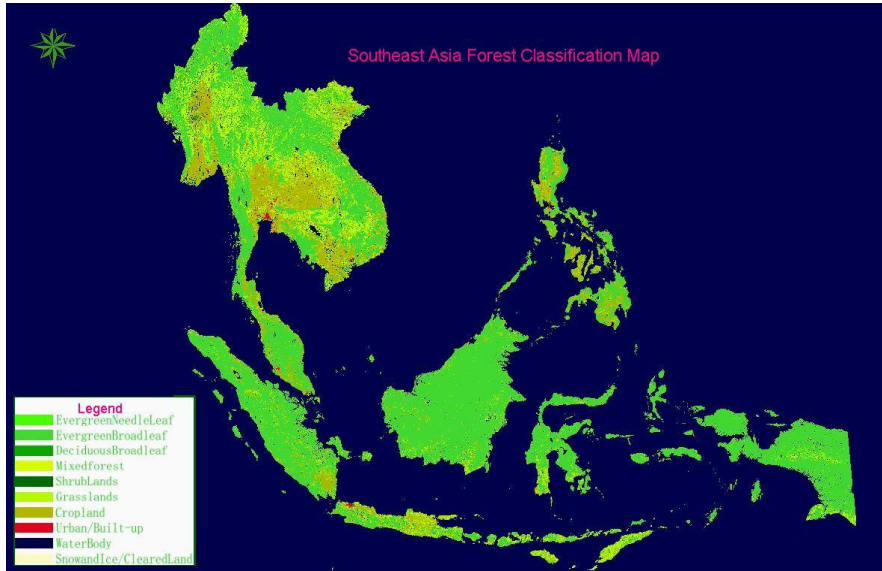


Fig.3 MODIS Global land cover data

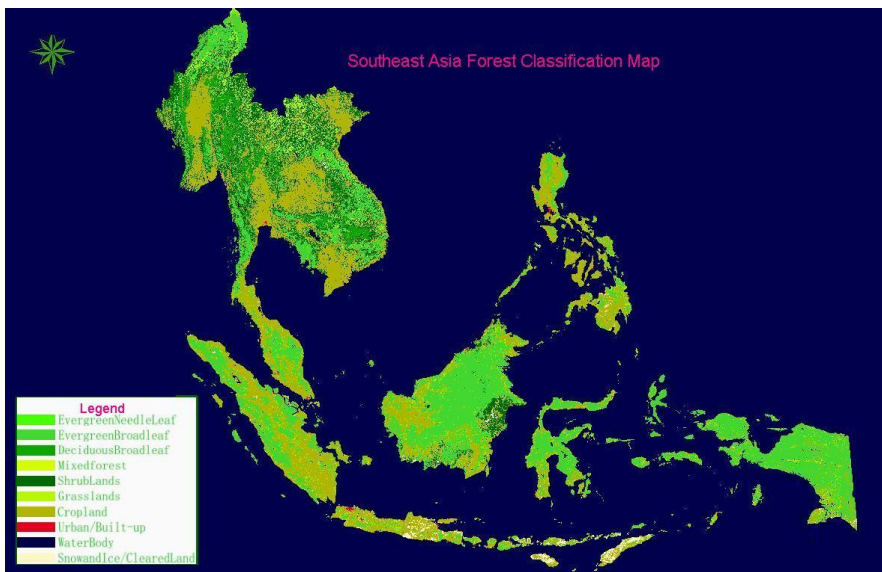


Fig.4 GLC 2000 Global land cover data

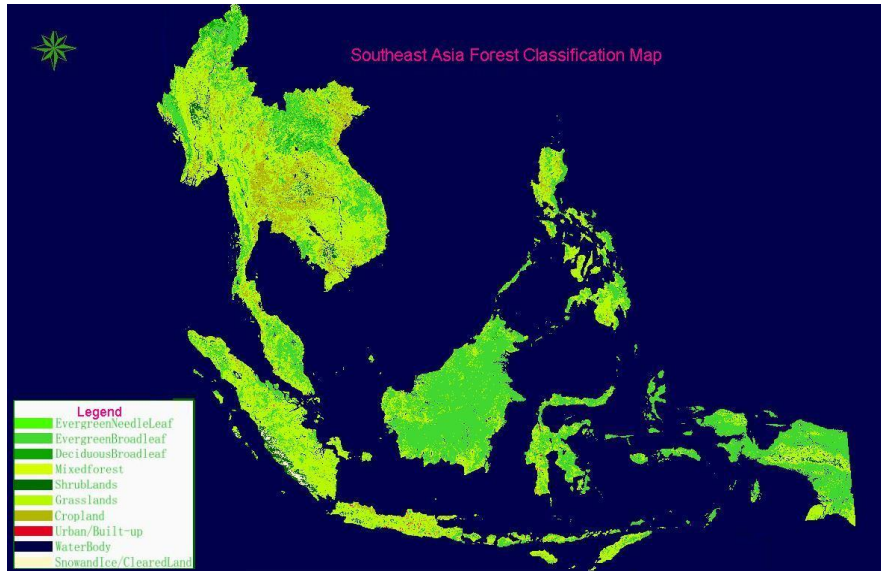


Fig.5 UMD Global land cover data

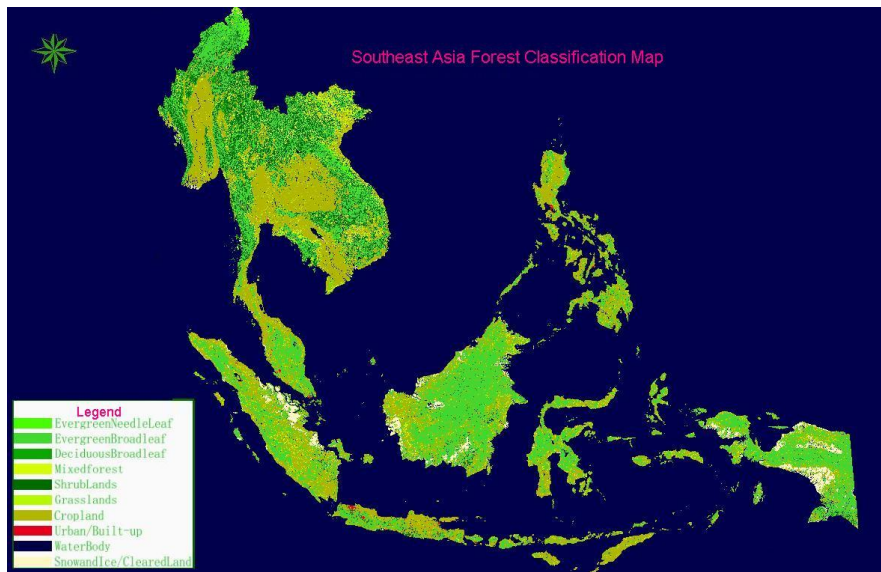


Fig.6 MERIS 2005 GLOBCOVER

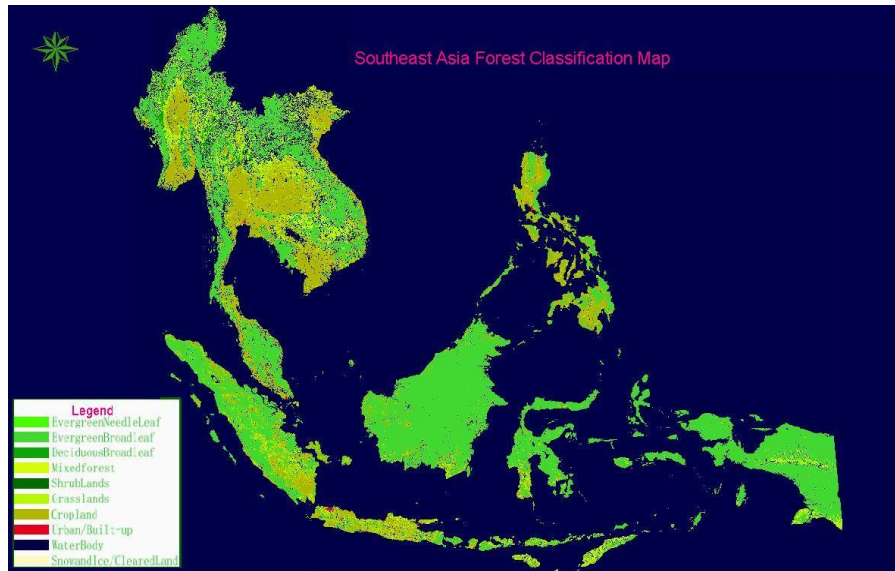


Fig.7 integrated results

3.4 Decision Tree Classification Algorithm

The key of the decision tree algorithm is to establishment decision rules through the sample points. There are several decision-making algorithms to establish the rules, but algorithms for massive remote sensing data classification have deficiencies. Especially there are few algorithms suitable for application in forest cover classification. Through the relevant literature and data on foreign inquiries, Quest algorithm has been selected to generate the decision tree of classification. The basic principle is as follows:

1) First, predict variables are selected. Correlation between all predictor variables X and target variable Y is analyzed in turn. If X is discrete, the association strength between X and Y will be calculated using the chi-square test respectively, and find the P value; If X is ordered or continuous variables, P value will be calculated using analysis of variance.

2) All the P values are compared to the value preset boundary (the default is 0.05). If they are less than the critical value, the minimum value of the variable P is selected as a branch variable; If they are all larger than the cutoff value, when X is continuous or ordered variables, P value will be calculated using Levene homogeneity test of variances, and when P value is less than the cutoff value, the minimum value of the variable P will be chosen as a branch of variables. If P values of homogeneity of variance test are large than critical value, the minimum value of the variable P in the first step will be the direct selection as the branch variable.

3) If the elected Branch variables are of discrete categorical variables, through the transformation, the goal variable Y will be maximized when X values are different. And largest discriminate coordinate will be calculated.

4) If Y is more classified categories, the mean number of X will be calculated for each type of Y value. These categories will eventually merge into two categories using the cluster analysis algorithm. This will simplify the multiclass discrimination problem to two types of discrimination.

5) Split point position is finally decided using quadratic discriminate analysis. And the selected original value of predictor variable X is got for building classification rules.

Based on the QUEST algorithm, the system combines the selection of the sample, decision analysis and classification. Firstly, using the sample data, samples were selected based on ROI selection tools in ENVI software. Then sampled were changed into data formats supported by QUEST, and QUEST algorithm interface was called by automated programs developed by IDL for establishing the decision tree automatically (suitable for established rules for ENVI decision data). Finally, classification was carried out using ENVI decision tree analysis tools to achieve judgments of the types of classification of forests in Southeast Asia. Classification accuracy and speed are greatly improved.

3.5 Sample precision analysis

Table 2 accuracy analysis of selected sample points

Types	Evergreen needle leaf forest	cities and towns	Water body	Snow and ice + Bare land	Evergreen broadleaf forest	Decidu ous needle leaf forest	Mixed forest	Shrub lands	Grass- lands	Crop -land
Evergreen needle leaf forest	59	0	0	0	0	0	0	2	5	4
Urban and built-up	1	75	2	0	0	0	0	0	1	2
Water body	0	0	2 8	0	0	0	0	0	0	0
Snow and ice + Bare land	0	2	0	48	0	0	0	3	0	0
Evergreen broadleaf forest	3	1	0	0	118	1	0	2	2	4
Deciduous needle leaf forest	0	0	0	0	0	62	0	0	2	0
Mixed forest	1	0	0	0	0	0	43	0	0	0

Shrub lands	0	0	0	6	3	0	1	80	0	1
Grasslands	7	1	0	0	1	6	0	2	11 9	4
Cropland	6	3	0	0	6	1	0	1	6	1 60
Classification accuracy	76.62%	91.46 %	93.33 %	88.89%	92.19%	88.57 %	97.73 %	88.89 %	88.15 %	91.4 3 %
Overall Classification accuracy	83.02 %									

Using QUEST algorithm, regression accuracy analysis of sample points established by decision tree is showed as follows. The accuracy is better than 75%, and the average accuracy reaches 83.02%. The above results only show that the regression accuracy of samples is high, but cannot represent the actual accuracy. The classification results are needed to be analyzed.

4 Results

In order to fully use the data features in the data processing, four schemes have been taken into account: scheme 1 implement the SG cloud-remove processing, and month-synthesis used maximum NDVI values (SGMaxLANDTYPE); scheme 2 carry out the SG cloud-remove processing, and month-synthesis used average NDVI values (MeanLANDTYPE); scheme 3 without implementing SG cloud-remove processing, and month-synthesis used maximum NDVI synthesis (SGMaxLANDTYPE); scheme 4 without implementing SG cloud-remove processing, and month-synthesis used average NDVI values (MeanLANDTYPE). Data samples were collected in the same position of sampling point on the above to construct different trees, and carry out the forest cover classification with quest decision tree algorithm. Classification results are shown in figure 8 to figure 11:

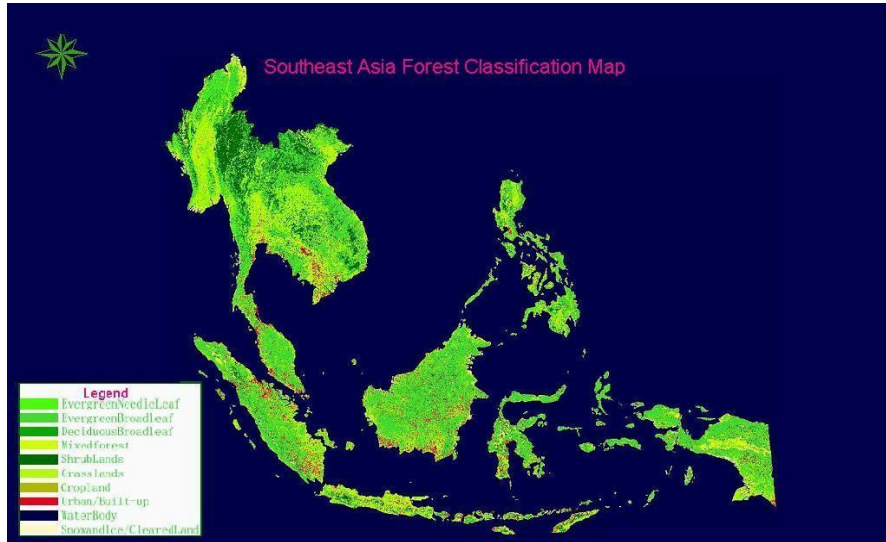


Fig.8 Forest classification result with SG cloud-remove processing and months synthesis using maximum NDVI values

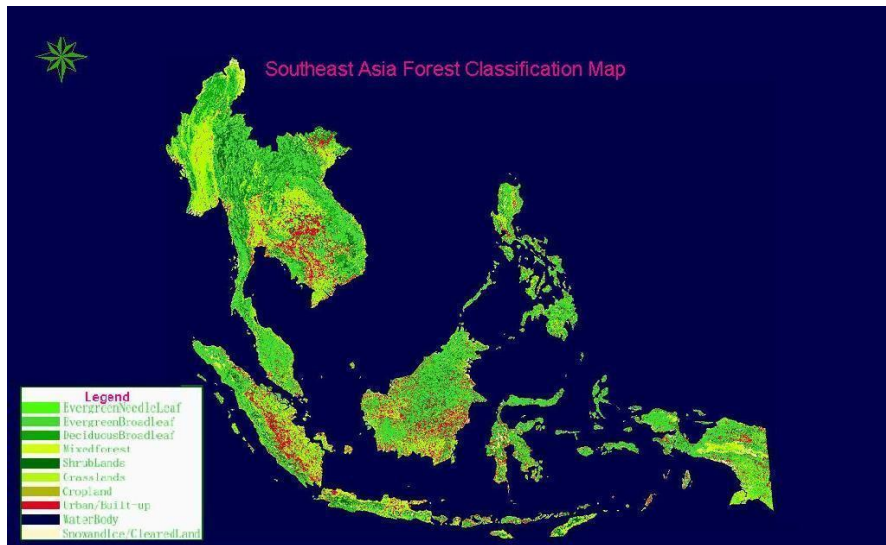


Fig.9 Forest classification result with SG cloud-remove processing and months synthesis using average NDVI values

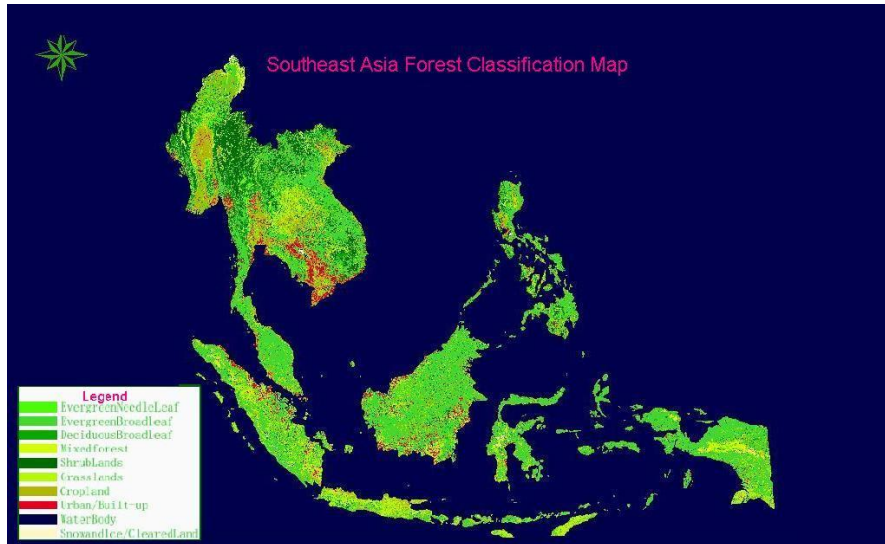


Fig.10 Forest classification result with month synthesis using maximum NDVI values

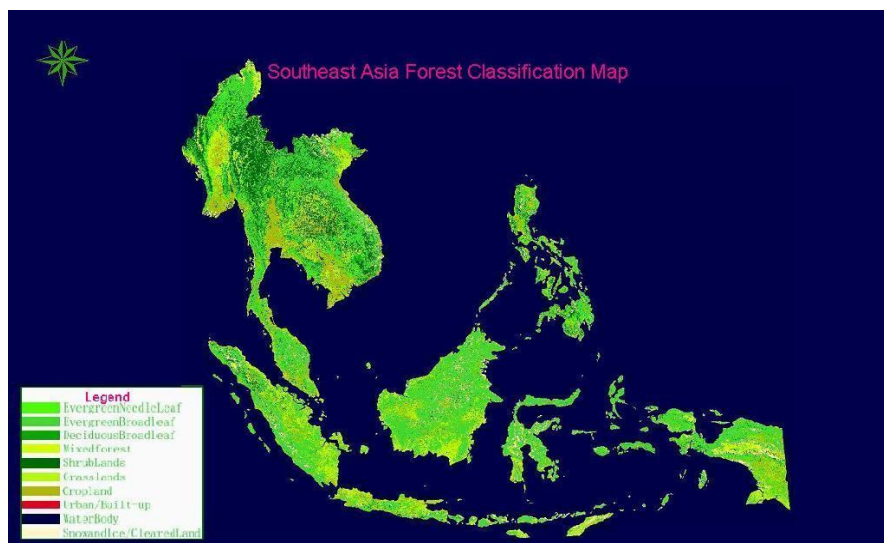


Fig.11 Forest classification result with month synthesis using average NDVI values

Since 2005 global land classification data (GLOBCover2005) is the update data and its spatial resolution is better than 500m, it can be used as reference data for the accuracy verification of classification results of this project. In order to evaluate the classification accuracy, the results of 4 categories of data were compared with the GLOB Cover 2005. The accuracy of MaxLANDTYPE, MeanLANDTYP,

SGMaxLANDTYPE ,SGMeanLANDTYPE are 65.05%,66.19%,76.66%, 78.14% respectively. It can be seen that high precision can be reached using SG cloud-remove processing and month-synthesis using average NDVI values (more than 78%). Meanwhile, the proportion of each category and the proportion of surface vegetation in each country were statistically. Table 3 indicates the results. With analysis on the proportion of different vegetation in different countries, the results showed SGMean LANDTYPE classification has achieved the highest accuracy.

Table 3 the distribution table of the proportion of different vegetation in Southeast Asia

No	Types	SGMaxLA NDTYPE	SGMeanLA NDTYPE	MaxLAN DTYPE	MeanLA NDTYPE	GLOBCo ver2005
1	Evergreen needle leaf forest	4.10	7.44	8.89	5.37	0.97
2	Evergreen broadleaf forest	40.63	38.61	35.73	37.29	35.00
3	Deciduous broadleaf forest	4.95	3.13	2.75	4.87	1.80
4	Mixed forest	1.91	1.53	1.96	1.79	6.41
5	Shrub lands	12.40	14.00	12.07	11.28	10.69
6	Grasslands	14.42	14.43	30.34	24.76	0.33
7	Cropland	16.22	15.14	0.00	0.00	39.56
8	Urban and built-up	4.46	0.00	5.11	11.31	0.15
9	Water body	0.31	0.31	0.20	0.21	2.06
10	Snow and ice + Bare land	0.58	5.39	2.96	3.12	3.02

5. Discussions

MODIS provide the multi-temporal forest cover information. Decision tree has advantage of understandable in structure, interpreted by its rules, calculating and precise in getting result, which lead to its high development in the field of forest cover classification. Multi-temporal MODIS-NDVI data with SG filtering have been used to the forest cover classification. We compare forest cover classification accuracy with the four schemes for time series MODIS-NDVI processing using SG cloud remove algorithm using Quest decision tree method. The results show that SG Mean LANDTYPE classification has achieved the highest accuracy. It can be seen that the decision tree algorithm with multi-temporal MODIS has great potential in regional forest cover mapping.

Acknowledgements

This work is supported by the National Science Foundation of China (NSFC) project (NO.40901161), and Chinese Universities Scientific Fund (Project No. 2011JS142) .

References

1. Ning Huang, Xiaojun Liu, Minhui Zhu, Shou rong Zhang: Remotely Sensed Image Classification Technology Review. *Journal of test and measurement technology*(in Chinese). 15(2): 86-92(2001)
2. Shuang Li , Shengyan Ding, Shuming Xu: Study of Remotely Sensed Image Classification Technology. *Journal of Henan University(Natural Science)*(in Chinese).32(2): 70-73(2002)
3. Jianping Wu, Xingwei Yang: Purification of Training Samples in Supervised Classification of Remote Sensing Data. *Remote Sensing for Land & Resources* (in Chinese), 26(1): 36-41(1996)
4. Shupeng Chen, Yingshi Zhao: *Remote Sensing Analyst*. Surveying and Mapping Press, Beijing, (1990)
5. Yingshi Zhao: *Principles and methods of analysis of remote sensing applications* .Science Press, Beijing, 2003
6. Xusheng Liu, Xiaoli Zhang: Research Advances and Countermeasures of Remote Sensing Classification of Forest Vegetation. *Forest Resources Management* (In Chinese) 1: 61-64, (2004)
7. Hansen M., Dubayah, R.&Defries R.: Classification trees: an alternative to traditional land cover classifiers. *International Journal of Remote Sensing*, 17(5):1075-1081(1996)
8. R.S.Defries, M.Hansen, J.R.G.Townshend, et al.: Global land cover classifications at 8 km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers. *International Journal of Remote Sensing*, 19(16):3141-3168(1998)
9. Hansen,M., DeFries,R., Townshend, et al. :Global land cover classification at 1km resolution using a decision tree classifier. *International Journal of Remote Sensing*, 21: 1331-1365(2000)
10. Kressler F P, Steinnocher K T: Detecting land cover changes from NOAA 2AVHRR data by using spectral mixture analysis. *International Journal of Applied Earth Observation and Geo information*, 1 (1): 21-26(1999)
11. Loveland.T.R., Zhu.Z., Ohlen,D.O.,et al: An analysis of the IGBP global land cover characterization process. *Photogrammetric Engineering and Remote Sensing*, 65(9):1021-1031(1999)
12. D.Muchoney, J.Borak, H.Chi, et al. : Application of the MODIS global supervised classification model to vegetation and land cover mapping of Central America. *Remote Sensing*, 21(6&7):1115—1138(2000)
13. M.A.Friedl, D.K.McIver, J.C.F.Hodges, et al.: Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, 2002, 83: 287-302.
14. Cen Yi, Liangpei Zhang, Shinobu Furumi, et al.: Land cover classification in Kii Peninsula of Japan using Terra/MODIS data.Third International Symposium on Multi-spectral Image Processing and Pattern Recognition. *Proceedings of the SPIE*, 5286:692-695(2003).

15. Carrão et al.: Contribution of multispectral and multi-temporal information from MODIS images to land cover classification, *Remote Sensing of Environment*, 112 (2008): 986–997.
16. Fei Wu: MODIS Data Land Use Classification. China Academy of Forest, (2005)
17. Yonghong, Liu, Zheng Niu: Regional Land Cover Image Classification and Accuracy Evaluation Using MODIS Data. *Remote Sensing Technology and Application*, 19(4):217-224(2004)
18. Quinlan J R: Induction of decision trees, *Machine Learning*, 1986, (1):81-106.
19. J. R. Quinlan: C4.5 Programs for Machine Learning. Morgan Kaufmann, San Mateo, Calif (1993)
20. Manish Mehta, Rakesh Agrawal, Jorma Rissanen: SLIQ: A fast scalable classifier for data mining, In *EDBT96*, Avignon, France, March (1996).