

APPLYING RANDOM FOREST CLASSIFICATION TO MAP LAND USE/LAND COVER USING LANDSAT 8 OLI

Huong Thi Thanh Nguyen^{*1}, Trung Minh Doan¹, Volker Radeloff²

¹Department of Forest resource & Environment management (Frem), Faculty of Agriculture and Forestry, Tay Nguyen University, Le Duan Str. 567, Buon Ma Thuot City, Daklak Province, Vietnam huongthanh.frem@gmail.com

^{*}huongthanh.frem@gmail.com; ¹doanminhtrungmt12@gmail.com; radeloff@wisc.edu

² Department of Forest and Wildlife Ecology 120 Russell Laboratories, 1630 Linden Drive Madison WI 53706-1598 USA

KEYWORDS: Classification; Landsat 8 OLI; Land use Land cover; Random Forest; Decision Tree

ABSTRACT:

This study used the Random Forest classifier (RF) running in R environment to map Land use/Land cover (LULC) of Dak Lak province in Vietnam based on the Landsat 8 OLI. The values of two RF parameters of ntree (number of tree) and mtry (the number of variables used to split at each node) were tested and compared. In current study the best results indicate the number of suitable decision trees involved in the classification process is 300 (ntree), and the suitable number of variables used to split at each node is 4 variables (mtry). These parameters were used to classify 7 bands multi-spectral resolution from 1-7 of Landsat 8 into ten classes of LULC including natural broad-leaved evergreen, semi-evergreen, dipterocarp deciduous forest, plantation forest, rubber, coffee land, crop land, barren land, residential area and water surface. The overall accuracy of 90.32% with Kappa coefficient of 0.8434 was found in this case.

1. INTRODUCTION

A proper LULC map including forest map is important for general planning as well as development strategies. In global context, the necessity for the development of repeatable, efficient, and accurate monitoring of land cover change is paramount to successful management of our planet's natural resources (Campbell et al. 2015). Therefore using remotely sensed data have become indispensable in producing such maps due to their multi-temporal and reasonable cost. Since the first generation of Landsat was launched in 1972 along with the rapid development of information technology, the development of the LULC maps have changed dramatically by the high technological approaches. Various types of satellite imagery and methods have been explored with an expectation of providing a reliable and up-to-date information system of LULC in low cost. Accordingly, many methods have been tested to improve accuracy of classification using satellite images. The conventional commonly used method is the pixel-based classification, which provides classification results based on supervised classification such as Maximum likelihood or support vector machine (SVM), or unsupervised such as ISODATA or k-mean. While the pixel-based classification maximum likelihood classification, each pixel is tested for all possible classes and the pixel is assigned to the class with the highest posterior probability (Lowe and Kulkani, 2015), the object-based classification is not based on individual pixels, but rather on the entire object that can be identified in the image, such as the shape and texture of the object, and the relationship between the objects as well (e.g. Dharamvir, 2013; Blaschke, et al., 2008; Machala and Zejdová, 2014; Zhou and Troy, 2008). Fuzzy sets (e.g. Mora et al. 2017; Sowmya and Sheelarani, 2011; Krihnapuram & Keller 1996) and neural network (e.g. Apte and Patravali, 2015; Mora et al., 2017) based methods have also been used for image classification.

In recent years, the development of algorithms including classification running in open source software e.g R has received a strong interest from the scientific community around the world for reasons that it is the cheapest, the broadest, and the most professional statistical computing environment. It allows data processing automation, import/export to various platforms, extension of functionality and open exchange of

scripts/packages, and also allows handling and generation of maps (Hengl, 2007). Random Forest is one of such packages. The Random Forest algorithm has been used in many data mining applications, however, its potential is not fully explored for analyzing remotely sensed images (Lowe and Kulkarni, 2015). Some studies proved RF is a potential method to map land cover comparing to conventional methods (Lowe and Kulkarni, 2015; Pelletier et al, 2016; Basten, 2016). As other methods, number of classes distinguished is one of reasons influencing the overall accuracy of classification. Aslan and San (2015) [8] have used RF to classify 13 different types of LULC based on image 8 with an overall accuracy (OA) of 79.90%, Lowe and Kulkarni (2015) used Landsat 8 to map LULC of soil, water, forest and vegetation with OA of 96.25%. Basten (2016) used RF to classify the three types of water, land, and buildings with 84% in OA, but OA reduced to 61% when six different types of LULC were classified. Normally, purpose of LULC map produced will decide how many classes should be interest, simultaneously, the LULC maps should be done in reasonable cost and allowable reliability. Hence, the attentions are always paid to find out suitable solutions to meet such purpose.

In this study, the RF classifier was used to map LULC from Landsat 8 OLI image and the RF classification algorithm. The study evaluated the RF parameters and combined spectral bands to obtain the best choice in this case.

2. MATERIALS AND METHOD

2.1 Study area

The study was conducted in Daklak province. This province is located in Central Highlands of Vietnam which located between in 12°09'45" to 13°25'06" latitude North and 107°28'57" to 108°59'37" longitude East. The size of study area is about 22,500 square km (150 x 155km). The average elevation is 400 - 800 meters above sea level. This area covered by different kinds of land use/land cover. In recent years a large of naturally forested areas converted to other land especially rubber and coffee land due to economic purpose. This leads to an unsustainable development which is one of reasons of climate change. The remaining natural forest types are dominated evergreen broadleaved, mixed pine, bamboo and broadleaved,

dipterocarp deciduous, and semi deciduous forest. Many of area have been disturbed by human over time at different levels. Many of valuable species trees have been selectively logged. The Figure 1 shows site study (in pink colour area).

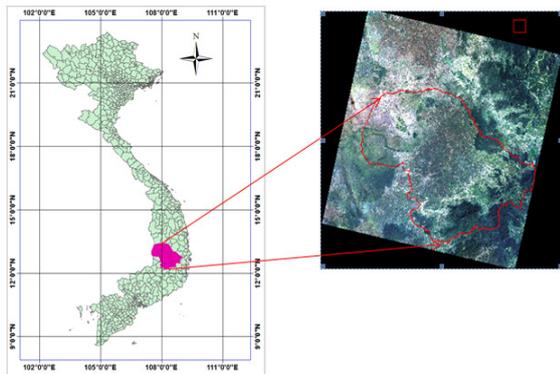


Figure 1. The area of study site

2.2 Data

The data used in this study is one scene of Landsat 8 Operational Land Imager (OLI) which was free downloaded from the website <http://earthexplorer.usgs.gov>. The image was acquired on February 7, 2017. The spectral bands for OLI are shown in Table 1. Our tests considered bands 1 through 7.

Table 1. Landsat 8 OLI Band Descriptions

	Bands	Wavelength (micrometers)	Resolution (meters)
Landsat 8 OLI	Band 1- Coastal aerosol	0.43-0.45	30
	Band 2- Blue	0.45-0.51	30
	Band 3- Green	0.53-0.59	30
	Band 4- Red	0.64-0.67	30
	Band 5- Near Infrared (NIR)	0.85-0.88	30
	Band 6- SWIR 1	1.57-1.65	30
	Band 7- SWIR 2	2.11-2.29	30
	Band 8- Panchromatic	0.50-0.68	15
	Band 9- Cirrus	1.36-1.38	30

2.3 Method

Random Forest classifier suggested by Breiman in 2001 is a supervised algorithm based on decision trees and improved bagging and bootstrap techniques. The RF contains a large number of trees. Each tree in the forest is grown from training pixels which is randomly selected to train RF classification. The remaining samples are considered to be out-of-bag cases (OOB) and these samples are used to estimate the classification error as trees are added to the forest (each bootstrap sample produces a tree and n_{tree} are grown from $n_{bootstrap}$ samples). The OOB is also used to measure input variable importance. Two parameters that need to be defined in this classification algorithm are n_{tree} (number of trees to grow) and m_{try} (the number of variables to split at each node). After the RF model is created, each result of the bootstraps votes for the most common class and the output is a classification result. The model is created depending on the most-voted classify of each decision tree (n_{tree}).

This number of variables used to split a RF node (denoted by m) corresponds to the square root of the number of input variables (Liaw&Wiener, 2002). By limiting the number of variables used for a split, the computational complexity of the algorithm is

reduced, and the correlation between the trees also decreases. The split process of RF decorrelates the different trees, there by the classification results less variable and more reliable.

The out-of-bag (OOB) error estimate: Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the tree. Put each case left out in the construction of the k_{th} tree down the k_{th} tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case n was OOB. The proportion of times that j is not equal to the true class of n averaged over all cases is the OOB error estimate. This has proven to be unbiased in many tests (Breiman, 2001a).

The estimation of variable importance was considered by two indicators: i) Mean Decrease Accuracy (MDA) and ii) Mean Decrease Gini (MDG). MDA is the accuracy of each independent variable classified based on the error rate (OOB). Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability of an item with label being chosen times the probability of a mistake in categorizing that item.

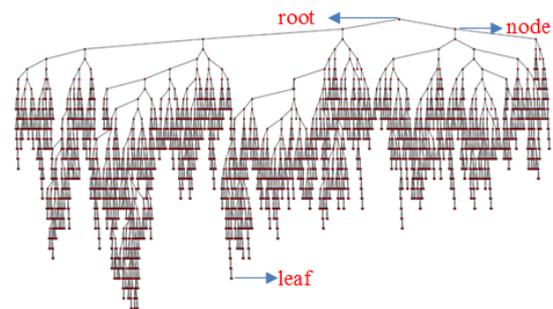


Figure 2. Decision Tree (Source: Basten, 2016)

This study used 7 bands of Landsat 8 OLI, from bands 1 to 7 to perform RF classification. The training areas for each land use/land cover type were selected based on Google Earth, field data and prior knowledge as well as available data. Training areas were distributed through the class to ensure the adequate representation of all the classes. There are 10 classes visually defined in the image, they are (1) evergreen broadleaved forest, (2) semi-evergreen forest (3) dipterocarp deciduous forest, (4) plantation forest, (5) rubber, (6) coffee field, (7) crop land, (8) residential area, (9) water surface and (10) barren land. The models were used as supervised classifiers to classify pixels based on their spectral signatures. Each pixel was represented by a vector of 7 grey values of 7 bands. The training set of 484 polygons corresponding to 27,918 pixels of the 10 classes as mentioned above was used to train classification. Additionally, an independent data set of 255 polygons with 12,338 pixel was collected to assess the classification result. Matrix confusion was produced statistically, which helps evaluating the accuracy of the classification result based on overall accuracy, producer's and user's accuracy as introduced by Congalton and Green, 1999.

The study used the Random Forest package running in R (CRAN) suggested by Liaw and Wiener. The n_{tree} and m_{try} evaluated to select the best one. The best m_{try} and n_{tree} is obtained based on the lowest OOB error.

3. RESULTS AND DISCUSSION

3.1 Random Forest parameters

The two parameters of *ntree* and *mtry* were evaluated their influence on classification performances using input spectral data bands within sample polygons. The *ntree* and *mtry* parameters by default of 500 trees and *mtry* of 2 respectively were put in the function `plot(modelRF)` to find the most suitable tree (*ntree*) based on OOB error rate as shown in Figure 3.

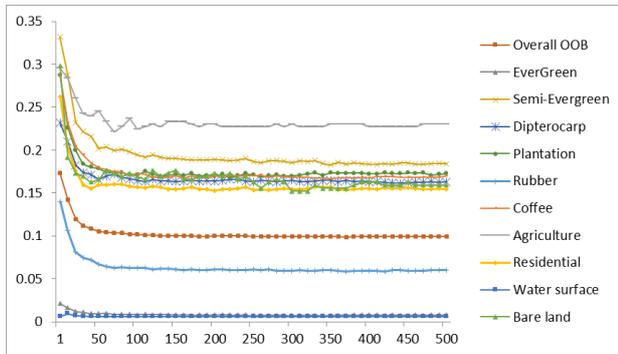


Figure 3. The variance of the classifications based on the number of trees *ntree*

The OOB errors varied from 0.007 to 0.332. The OOB errors of all classes were quite high in the range of 0 to 100 trees. the OOB errors were stable at 300 trees onwards. Comparing to others, the lowest OOB errors were indicated by water surface and evergreen broadleaved forest, dipterocarp deciduous was followed. The highest OOB errors were found by crop land and semi-evergreen forest.

The selection of appropriate *ntree* for the classification process depends on the degree of stability and saturation of the error as well as the least time to reach the lowest error. The smaller the number of decides trees, the lower the computational time and vice versa. Figure 4 shows that at *ntree* = 300, the level of error is almost saturated and very little fluctuation, hence, in this case we used *ntree* of 300 for the model.

The number of variables used for splitting at each node in the process (*mtry*) based on the number of trees (*ntree*) identified above was calculated through the `tuneRF()` function, which is shown in Figure 4a.

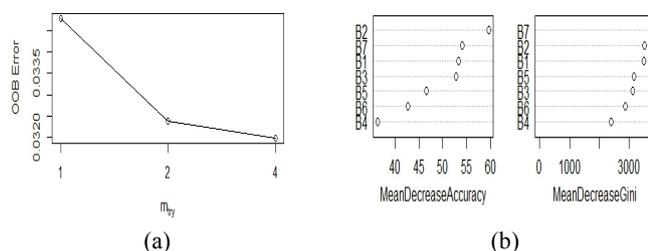


Figure 4. (a) Level of OOB error of *mtry*; (b) The variable importance of classification

The results showed that the highest OOB error was found with 3.48% when only one variable (*mtry* = 1 (band)) was used. The following error was of 3.24% corresponding *mtry* = 2. Meanwhile the lowest OOB error was indicated with *mtry* = 4 (OOB error of 3.20%). However not significant difference was found among these scenarios. The difference between the worse and the best was 0.28%.

Importance() function and `varImpPlot()` function were used to consider the MDA and MDG values and `varUsed()` function was employed to select the potential variables which are actually used in RF. The Table 2 described the use frequencies of variables.

Table 2. Frequencies that variables are used

Spectral band	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7
Frequencies	44,554	44,705	44,250	41,507	43,240	39,830	44,088

Figure 4b and Table 2 show comparing to others, the variable of 4 (band 4) and 6 (band 6) had the lowest impact to the model with MDA by 36.46 and 42.75 and the frequencies of 41,507 and 39,830 respectively. However the difference was not significant.

The all results presented above indicated that though there was different among the parameters in the model, the discrimination was not much significant. Therefore we tested different scenarios using different the three parameters of *ntree*, *mtry* and the spectral bands to obtain the best solution. These scenarios were presented below:

- RF1 classification model: using default parameters *ntree* by 500 trees; *mtry* = 2 and all bands involved in the classification process.
- RF2 classification model: using the parameters of *ntree* 300 and *mtry* = 4, all bands involved bands in the classification process.
- RF3 classification model: *ntree* by 500 and *mtry* = 4 t with all bands involved in the classification.
- RF4 classification model: using the parameters were detected as *ntree* by 300 and *mtry* = 4, with band 1,2,3,5,7 involved in the classification process (except for band 4 and 6) to consider whether these bands actually affect the results of the model.

The results were assessed using OOB estimate of error rate (%), overall accuracy and Kappa as identified in Table 3.

Table 3. Assessment of Random Forest classification models

Model	RF1	RF2	RF3	RF4
Criteria				

OOB estimate of error rate (%)	3.24	3.28	3.23	3.75
Overall Accuracy (%)	90.18	90.32	90.30	89.80
Kappa	0.8396	0.8434	0.8427	0.8344

Table 3 shows the low error OOB (~ 3%) was presented in all the predictions. The overall accuracy and Kappa coefficient are quite high with OA > 90% and K > 84%. It is observed there is a

little difference on the classification accuracy, about 0.5% OA difference between the worst and the best parameter. Comparing to others, the FR4 shows the lowest accuracy. This indicates in this case the exclusion of bands which presented as the worse ones could cause the lower accuracy. As expected, there is no significant difference in accuracy among RF1, RF2, and RF3. Overall, the RF2 yielded comparable results to the RF1 and RF3, while the processing time was significantly lower. Thus the RF2 was considered as the best choice with ntree of 300, mtry of 4 with 7 bands involved in the model.

3.2 Mapping LULC of Daklak province, Vietnam

The RF2 was used to classify LULC from Landsat 8 OLI for Daklak province of Vietnam, and the independent data was used for accuracy assessment the classification as presented in Table 4.

Table 4. Accuracy of the RF2 classification model

LULC	UA (%)	PA (%)	LULC	UA (%)	PA (%)
Evergreen	94,23	99,13	Coffee	50,25	36,53
Semi-evergreen	61,36	75,96	Crop land	92,47	91,69
Dipterocarp forest	88,46	54,12	Residential area	90,34	95,41
Plantation forest	38,46	70,95	Water surface	99,34	99,91
Rubber	88,12	71,84	Barren land	89,39	56,19
Overall accuracy: 90,32 % Kappa coefficient: 0,8434					

(Where:UA : User Accuracy; PA : Producer Accuracy)

The overall accuracy under the study obtained 90.32%. From the error matrix, the computed kappa coefficient was 0.84 representing an almost agreement (Landis and Koch, 1977).

Within the forest categories, the best result came from Evergreen broadleaved forest with about 94% and 99% for user's and producer's accuracy, respectively. By contrast, the lowest accuracies were given by coffee and plantation forest with 50% and 38 %; 70% and 36% for user's and producer's accuracy, respectively. Except for semi-evergreen forest the remaining lands obtained UA higher 80%. Most of LULC gained an PA higher 70% except for barren land.

Based on the classification model, using predict() function to estimate the whole study area. The classification image was made from the plot() function and then exported to the GEOTiff format for editing into the LULC map using ArcGIS 10.1 software.

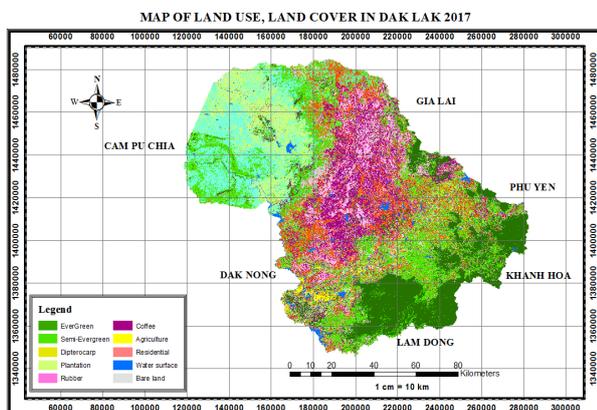


Figure 5. Map of land use/land cover in Dak Lak 2017

The area and percentage of each class of LULC in Dak Lak is shown in Figure 6 below.

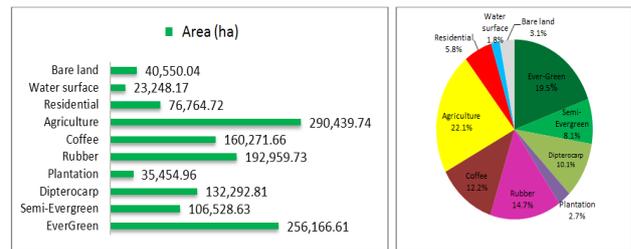


Figure 6. The area and percentage of LULC of Daklak, Vietnam

The study results show that the total area of LULC types in Dak Lak is 1,314,677.07ha, of which the area covered by forest cover is 40.35%, equivalent to 530,443.01ha and non-forest land area is 59.65%, equivalent to 784,234.06ha. Evergreen forest accounts for 19.50% of total area, and Semi-evergreen forest, Dipterocarp forest and plantation forests accounting for 8.10%, 10.06% and 2.70% of total area. In the non-forest LULC group, crop land occupies the highest area of 22.09% of the total area, equivalent to 290,439.74 ha; rubber, coffee, residential area, barren land and water surface accounted for 14.7%, 12.2%, 5.8%, 3.10% and 1.8% of the total area, respectively.

4. CONCLUSIONS

The obtained results indicated that the Random Forest was the potential method to map LULC from the satellite image. In this study, Random Forest Classifier was used for mapping 10 types of LULC from Landsat 8 OLI imagery. The approaches were based on the out-of-bag (OOB) estimate of error to optimize the RF parameters, simultaneously, the effect of spectral band of Landsat 8 (from 1-7) was evaluated. The results reveal that, the best choice were obtained when using all of the bands (1 through 7) and the suitable parameters of ntree = 300 and mtry = 4 for the prediction. The overall accuracy of 90.32% was gained in this case.

ACKNOWLEDGMENT

This work is initial part of the research project under the PEER program (Partnerships for Enhanced Engagement in Research), a U.S. government program to fund scientific research in developing countries. This is a competition program sponsored by USAID in partnership with several other US Government agencies and administered by the US National Academy of Sciences (NAS). The Landsat 8 OLI imagery was obtained from USGS website, all rights reserved. The authors would like to thank all of the people involved in collecting field data for classification and validation and the PEER Program who funded for this work.

REFERENCES

- Apte, K. S., and Patravali, D. S., 2015a. Development of Back Propagation Neural Network Model for Extracting the Feature from a Satellite Image using Curvelet Transform. *International Journal of Engineering Research and General Science*, ISSN 2091-2730, 3(1), pp.226-236.
- Aslan, N., and San, D. K., 2015. The Usage Of Combined Landsat 8 Imagery And Additional Bands For Random Forest Classification Improvement. *Asian Association on Remote Sensing*. Access Feb. 2018. <http://a-a-r->

[s.org/acrs/index.php/acrs/acrs-overview/proceedings-1?view=publication&task=show&id=2068](https://www.sciencedirect.com/journal/ISPRS-International-Archives-of-Photogrammetry-and-Remote-Sensing).

Basten K., 2016. Classifying Landsat Terrain Images via Random Forests. Bachelor thesis Computer Science in Radboud University, Netherlands.

Blaschke, T., Lang, S., Hay, G.J. (Eds). 2008. Object based image analysis. Springer, Heidelberg, Berlin, New York, 817p.

Breiman L., 2001a. Random Forests. *Machine Learning Journal Paper*, 45(1), pp. 5-32.

Breiman L., and Cutler A., 2007. Random Forests. <https://www.stat.berkeley.edu/~breiman/RandomForest/>.

Campbell, M., Congalton, R.G., Hartter, J., Ducey, M., 2015a. Optimal land cover mapping and change analysis in northeastern Oregon using Landsat imagery. *Photogrammetric Engineering and Remote Sensing*, 81 (1), pp. 37-47. DOI: 10.14358/PERS.81.1.37.

Congalton, R. G., and Green, K., 1999. Assessing the accuracy of remotely sensed data: Principles and practices, Boca Raton, Lewis Publishers.

Dharamvir., 2013. Object Oriented Model Classification of Satellite Image. *CDQM*, 16(1), pp. 46-54.

Efron B., 1979. Bootstrap Methods: Another look at the Jackknife. *The Annals of statistics*, 7(1), pp.1-26.

Hengl, T., 2007. A practical guide geostatistics mapping of environment variable. *JRC European commission*. 165p.

Kennedy, R.E., et al., 2015a. Attribution of disturbance change agent from Landsat time-series in support of habitat monitoring in the Puget Sound region, USA. *Remote Sensing of Environment*. <http://dx.doi.org/10.1016/j.rse.2015.05.005>

Krishnapuram, R., Keller, J. 1996. The possibilistic c-means

algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems* 4, pp. 385-393. DOI: 10.1109/ICDCSyst.2012.6188738.

Landis, R.J. and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159-174.

Liaw, A., and Wiener, M., 2002. Classification and Regression by randomForest. *R News*, 2 (3), pp. 18–22.

Lowe B. and Kulkarni A., 2015a. Multispectral Image Analysis Using Random Forest. *International Journal on Soft Computing (IJSC)*.6 (2), pp. 1-14.

Machala, M., and Zejdová, L., 2014a. Forest Mapping Through Object-based Image Analysis of Multispectral and LiDAR Aerial Data. *European Journal of Remote Sensing*, 47(1), pp.117-131, DOI: 10.5721/EuJRS20144708.

Mora, A., Santos, M. A. T., Lukasik, S., Silva, M. N. J., Falcão, J. A., Fonseca, M, J. and Ribeiro A.R., 2017. Land Cover Classification from Multispectral Data Using Computational Intelligence Tools: A Comparative Study. *Information* 2017, 8(147); doi:10.3390/info8040147 www.mdpi.com/journal/information.

Pelletiera, C., Valeroa, S., Inglada, J., Championb, N., Dedieu, G. 2016. Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*. Volume 187, 15 December 2016, pp.156-168.

Sowmya, B., and Sheelarani, B., 2011a. Land cover classification using reformed fuzzy C-means. *Sadhana*, 36(2), pp. 153–165.

Zhou, W., and Troy, A., 2008a. An Object-Oriented Approach for Analyzing and Characterizing Urban Landscape at the Parcel Level. *International Journal of Remote Sensing*. 29(11), pp. 3119-3135, DOI: 10.1080/01431160701469065.