Initial Results on Landuse/Landcover Classification Using Pixel-Based Random Forest Algorithm on Sentinel-2 Imagery over Enrekang Region

J S Nurfadila¹, S Baja², R Neswati², D Rukmana³, and Z Zylshal⁴

¹Post Graduate Student of Agriculture Science, Hasanuddin University, Indonesia
²Department of Soil Science, Hasanuddin University, Indonesia
³Department of Agriculture Socio-Economics, Hasanuddin University, Indonesia
⁴Remote Sensing Data and Technology Center, National Institute of Aeronautics and Space (LAPAN), Indonesia

(Correspondence author email: dilajs01@gmail.com)

Abstract: Land use classification is the basis for making further policy in many fields including agriculture. Effective methods in landuse/landcover (LULC) classification are essential for later application in policy making. The development of remote sensing technology has been increasing rapidly. The use of Earth Observing (EO) Sentinel-2 imagery can greatly help LULC mapping over large area. As the basic input on the assessment of land availability and suitability, it is important to perform LULC in such way that it is objective, replicable, and accurate. This study aim to performed state-of-the-art Random Forest algorithm on multitemporal Sentinel-2 imagery on LULC extraction over Enrekang Region. With its 10 m spatial resolution as well as multitemporal information, acquired on December as a representation of the rainy season and in July as a representation of the dry season, it is expected to produce a more optimal LULC maps. Confusion matrix were then performed using visually interpreted Pan-sharpened and orthorectified SPOT-6/7 imagery to calculate the accuracy. The output of LULC classification based were expected to reach 95% overall accuracy.

1. Introduction

Land use information is essential in many ecology management and plannings. Land use and landcover are the most influential factors in the ecosystem[1][2]. Currently, Land use information is increasingly important because high population growth rates make human land use in large and scattered areas truly very complex. Increased land recruitment and competition in land use both in the needs of agricultural production, plantations, industry, services, and settlements encourage the birth of ideas about how to make the most profitable land use decisions from limited resources. The accuracy and updating of land cover maps are essential, because as a foundation of to understand the dynamics of the environment, including climate, as well as interactions in the environment. Land cover classification based on remote sensing technology has become a research that continues to be developed until now, and has proven to be the most useful and effective approach in many aspects of earth observation applications [3]. The Development of Earth Observation satellites, such as Sentinel 2, make it possible to extract more detailed information about the condition of the earth and surface in the large area, short time and with
relatively cheaper costs [4][5]. Sentinel 2 is a program of the European Space Agency (ESA) with the good resolution, so it is very good to be used for monitoring [6][7][8]. Improved image resolution both spectral and temporal, effective in optimizing estimates of land use (Galiano et al., 2011). The results of the research by Vuolo et al (2018) sentinel 2 accuracy can reach 92-95% overall accuracy. Sentinel 2 has 13 bands, of which 4 bands have a resolution of 10 m, 6 bands with a resolution of 20 meters and 3 bands with a resolution of 60 meters. The Sentinel 2 sensor covers an area of 290 Km, with a 5-day temporal resolution. With good spatial resolution, temporal resolution and coverage, Sentinel 2 satellite data are used as a good monitoring data. In this study, visual classification used the help of SPOT 6 satellite imagery. The high resolution of Spot 6 satellite helps in clarifying objects that are less clearly visible on Sentinel 2. SPOT 6 have the spatial resolution of 6 m and spectral band Blue (60.04 nm), Green (57.33 nm), Red (37.47 nm ) and NIR (28.25 nm). The characteristics of SPOT and Sentinel 2 can be seen from table 1.

| Satellite | Band | Spectral Range | Spacial Resolution (M) |
|-----------|------|----------------|------------------------|
| SPOT 6    | Pankromatik | 0,450 - 0,745 μm | 1.5                   |
|           | Blue   | 0,450 - 0,520μm  | 6                     |
|           | Green  | 0,530-0,590μm    | 6                     |
|           | Red    | 0,625- 0,695μm   | 6                     |
|           | NIR    | 0.760-0.890 μm   | 6                     |
| Sentinel 2| B1     | 430–457 μm       | 60                    |
|           | B2     | 448–546 μm       | 10                    |
|           | B3     | 538–583 μm       | 10                    |
|           | B4     | 646–684 μm       | 10                    |
|           | B5     | 694–713 μm       | 20                    |
|           | B6     | 731–749 μm       | 20                    |
|           | B7     | 769–797 μm       | 20                    |
|           | B8     | 763–908 μm       | 10                    |
|           | B8A    | 848–881 μm       | 20                    |
|           | B9     | 932–958 μm       | 60                    |
|           | B10    | 1336–1411 μm     | 60                    |
|           | B11    | 1542–1685 μm     | 20                    |
|           | B12    | 2081–2323 μm     | 20                    |

Various methods of land use classification by using remote sensing imagery such as maximum likelihood classifiers (MLCs), neural networks, support vector machines (SVM), Random Forest (Breiman, 2001) with fairly good results accuracy. There have been many studies showing the success of using learning methods such as the Random Forest, in optimising the accuracy of land cover classifications (Mellor et al., 2015). Random Forest (RF) (Breiman, 2001) is a learning method that has been widely applied in remote sensing [9][10]. Many studies have shown that the results of classification with random forest show accurate results and with fast processing [9], and can be relied on areas that have complex landscapes [11]. Random Forest does an ensemble method consist of several decision trees as classifiers. The class that results from this classification process is taken from the most classes produced by the decision trees in the Random Forest. Voting the available decision trees makes the accuracy of the Random Forest increase. Although the researchers used two Sentinels as data analyzed with 10 and 20 meters spectral resolution, a visual classification was still very difficult. The purpose of the research on land use/cover classification using sentinel 2 and the random forest method is to investigate, implement the use of Sentinel 2 satellite data using the Random Forest method to produce land cover estimates that achieve 95% overall accuracy.
2. Material and Methods
The aim of the land use/land cover mapping in this study is to provide thematic information on different land use categories, used as a basis for assessing land availability, and as basic information for the land suitability for further research. Data sources used in this classification are Sentinel 2 satellite data with 10 m spatial resolution, also using spot satellite data 6 as material for validation in addition to direct field observations. procedures in this land use classification can be seen from Figure 1.

![Figure 1. Workflow LULC Classification.](image)

2.1. Study Area
Enrekang Regency is located about ± 240 km north of Makassar City, or geographically located between 3°14'36"-3°50'0" south latitude with 119°40'53" - 120°6'33" east longitude. The total area of Enrekang Regency is 1,786.01 km². The topography of the Enrekang Regency consists of highlands located on the West, East, South, and North, including Alla, Curio, Anggeraja and Malua Districts. While the lowlands lie in the middle, which includes Enrekang District and part of Maiwa Subdistrict.
2.2. Data Collection and Pre-Processing

The classification of land use in Enrekang Regency used Sentinel 2 satellite data obtained from two recordings, April as the representation of rain month and September as a representation of the dry month, as well as SPOT 6 satellite data used to assist validation.

Land use classification results are influenced by Pre-processing data [12]. Pre-processing data includes atmospheric correction, geometric correction and cropping area. Because the Sentinel-2 used in this study on two recordings, it is important to do atmospheric correction [13]. For that reason, Semi Automatic Plungin Classification is needed [14]. Atmospheric correction use the Dark Object Substraction (DOS1) method by reducing the value of reflection with the dark pixel value. This method for correcting diffusion effects [14]. Dos are the simplest method in atmospheric corrections [12], by reducing the smallest digital number value with all the digital number values present in the image [15][12].

Sentinel 2 and Spot have different geometric orbits and displays, so require normalization to the same appearance and angle. Sentinel-2 Level 1 products must be corrected Geometrically [6]. Geometric correction is done by Image Retrification Workflow. Geometric correction needs to be done by comparing reference data that has clearer location data [15]. Cropping is done by limiting the research area based on enrekang district administrative boundaries.

2.3. Determination of Land Use Classes, Sample and Sample Training

The land use class used in this study refers to the medium-scale Indonesian National Standard 2014 rule, the class of land use/cover to be mapped is adjusted to the dominance of land cover in the study area which can be seen from the table 2. The strategy of determining the number of samples is very important because it will relate to the accuracy of the results and the confidence in the final map. The selected sample is representative of each class that has been determined. This study will use a stratified random sampling pattern, the sample will be stratified into 15 classes of land use/cover. Allocation of sample
size depends on the parameters estimated, but if the overall accuracy of the map, optimal allocation is recommended [16]. Until now there has been still no clear and strict rules regarding the number of samples needed to assess accuracy (McCoy, 2005). According to Conngalton and Green (1999), a minimum sample of 50,000 is needed for an area of more than 1000000, whereas for more than 12 classes, 75-100 samples are needed. In this study using a formula developed by Fitzpatrick-Lins (1981) namely

\[ N = \frac{Z^2 (p) (q)}{E^2} \]

N is the number of samples. Z is (standard for 95% confidence level), p is the expected accuracy, q is 100 - p, and E is the allowable error. So that the total number of samples used in this study is 759 whiles for each class requires a sample of 51. Polygon will be made to identify all classes of land cover that are set. The sample is taken randomly where 66% is for training and 34% for data validation.

Table 2. Land Use Class

| Land Use Class       | Description                                                                 |
|----------------------|-----------------------------------------------------------------------------|
| River                | Body of water flowing through the channel at the surface of the ground      |
| Scrub                | Clumps of plants with short, creeping stems, several cm to about 1.5 m high.|
| Dry Rice Fields      | Rice Fields in dry lands                                                    |
| Rice fields          | Rice Field in wet lands                                                     |
| Plantation           | Plants are rather tall with regular patterns and spacing                    |
| Setlement            | Houses, offices, mosques and other buildings                                |
| Dry land agriculture | agriculture on dry land such as onions, carrots and so on with a regular pattern |
| Forest               | trees, trees remain year-round green, tall plants with irregular patterns    |
| Cloud                | Clouds                                                                       |
| Cloud shadow         | shadows formed by clouds                                                    |

2.4. Random Forest Method

It is a development of Classification and Regression Tree (CART) method by applying bootstrap aggregating (bagging) and random feature selection methods (Breiman, 2001). The classification results with random forest have higher accuracy compared to the Neural Network classifier, Classification and regression tree (CART) classifiers method [9][17]. The Random forest method has several advantages such as more efficient results for large data, can handle thousands of variables without reducing it, gives an estimate of variables that are important in classification, can produce unbiased estimates of generalisation errors, computationally lighter than other tree ensemble methods [10][18]. Random forest is a class allocation prediction method based on a decision tree [19][20]. Polygon will be created to identify all designated land cover classes. The sample is taken randomly where 66% is for training and 34% for data validation. Random exclusion of 33% of samples is effective for reducing sample size [20]. Data training uses spectral variables, textures, and topography.

2.5. Accuracy Assessment

Accuracy testing is very important by remote sensing researchers. Data validation is used to assess the accuracy of land cover classification by applying the random forest method. the use of error matrix method is the most common for the test results of land cover mapping. The accuracy of the classification results is measured using the confusion matrix method or also known as error matrix or matching matrix, namely by comparing the actual classes with the classification
classes. Through the confusion matrix method, indicators of accuracy and error can be obtained in the classification results.

3. Result And Discussion

Land Cover Classification in this study we can see in table 3.

Table 3. Land Use Classification

| DN | Reference                  | Map / Prediction         |
|----|----------------------------|--------------------------|
| 0  | unclassified               | unclassified             |
| 1  | cloud                      | Cloud                    |
| 2  | cloud shadow               | cloud shadow             |
| 3  | Forest                     | Forest                   |
| 4  | Dry land agriculture       | Dry land agriculture     |
| 5  | Setlement                  | Setlement                |
| 6  | Plantation                 | Plantation               |
| 7  | Rice fields                | Rice fields              |
| 8  | Dry Rice Fields            | Dry Rice Fields          |
| 9  | Scrub                      | Scrub                    |
| 10 | River                      | River                    |
In testing processing methods and models in classification, a test is needed accuracy of the classification results. The quality of the results of the classification of closing land use from remote sensing data measured by reference to the data test which is some image information in the target classification image area. Test data is made through digitized images based on references from field survey information and / or from interpretation using remote sensing data with more spatial resolution high, where the type of land cover class in the data test is the same as target classification class. Making data test is similar to making training data but the area is wider, and with the area requirements used the data test is not may be the same as the training data area.

Figure 3. Land Use Class in Study Area.
In addition to measuring overall accuracy, errors can also be calculated measured classifications for each class with commission errors and errors that should enter certain classes or omission errors (omission error).

Table 5. Error Omission Each Class

| Map class         | User's Accuracy [%] | Estimate | Standard Error | 95 % Interval | Producer's Accuracy [%] | Estimate | Standard Error | 95 % Interval |
|-------------------|---------------------|----------|----------------|---------------|-------------------------|----------|----------------|---------------|
| (1) Cloud         | 99.99               | 0.00     | 99.99          | 99.99         | 0.00                    | 99.99    | 0.00           | 100.00        |
| (2) Cloud shadow  | 99.63               | 0.01     | 99.61          | 99.65         | 0.01                    | 99.92    | 0.99           | 99.98         |
| (3) Forest        | 96.62               | 0.03     | 96.56          | 96.67         | 0.05                    | 97.46    | 97.67          |               |
| (4) Dry land agriculture | 89.99 | 0.05     | 89.90          | 90.08         | 0.23                    | 55.86    | 56.74          |               |
| (5) Setlement     | 92.80               | 0.04     | 92.72          | 92.88         | 0.21                    | 86.48    | 87.28          |               |
| (6) Plantation    | 72.16               | 0.07     | 72.02          | 72.30         | 0.10                    | 90.99    | 91.37          |               |
| (7) Rice fields   | 77.56               | 0.07     | 77.43          | 77.69         | 0.22                    | 65.60    | 66.44          |               |
| (8) Dry Rice Fields | 93.48 | 0.04     | 93.40          | 93.56         | 0.17                    | 96.31    | 96.98          |               |
| (9) Scrub         | 96.71               | 0.03     | 96.65          | 96.76         | 0.17                    | 93.09    | 93.75          |               |
| (10) Sungai       | 99.26               | 0.01     | 99.23          | 99.29         | 0.35                    | 97.25    | 98.60          |               |

The accuracy of the classification results is measured using the confusion matrix method or also known as an error matrix or matching matrix, which is a table comparison of the actual classes of test data with the classification classes. Through the confusion matrix method can be obtained indicators of accuracy and errors in classification results. Table 4 illustrates accuracy measurements classification results with confusion matrix.
Table 6. Confussion Error matriks

| Reference Class | (1) Cloud | (2) Cloud shadow | (3) Forest | (4) Dry land | (5) Settlement | (6) Plantation | (7) Rice fields | (8) Dry Rice Fields | (9) Scrub | (10) Sungai |
|-----------------|----------|-----------------|-----------|-------------|--------------|--------------|-------------|-------------------|-----------|------------|
| Map class       | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -8 | -9 | -10 | Sum |
| (1) Cloud       | 289092 | 0 | 0 | 0 | 28 | 0 | 0 | 1 | 1 | 0 | 289122 |
| (2) Cloud shadow| 0 | 31437 | 0 | 0 | 42 | 0 | 59 | 0 | 0 | 17 | 31555 |
| (3) Forest      | 0 | 16 | 27744 | 3 | 84 | 456 | 410 | 1 | 2 | 0 | 28716 |
| (4) Dry land    | 0 | 0 | 0 | 2167 | 186 | 5 | 0 | 50 | 0 | 0 | 2408 |
| agriculture     | 0 | 0 | 308 | 6033 | 107 | 28 | 22 | 0 | 1 | 6501 |
| (5) Settlement  | 2 | 0 | 0 | 308 | 6033 | 107 | 28 | 22 | 0 | 1 | 6501 |
| (6) Plantation  | 0 | 0 | 680 | 14 | 194 | 7917 | 1383 | 287 | 496 | 0 | 10971 |
| (7) Rice fields | 5 | 0 | 2 | 679 | 249 | 52 | 3732 | 3 | 0 | 90 | 4812 |
| (8) Dry Rice Fields | 6 | 0 | 0 | 659 | 68 | 0 | 0 | 10512 | 0 | 0 | 11245 |
| (9) Scrub       | 2 | 0 | 10 | 1 | 40 | 146 | 41 | 1 | 7082 | 0 | 7323 |
| (10) Sungai     | 0 | 0 | 0 | 18 | 20 | 0 | 0 | 0 | 0 | 0 | 5104 |
| Sum             | 289107 | 31453 | 28436 | 3849 | 6944 | 8683 | 5653 | 10877 | 7581 | 5212 | 397795 |

4. References

[1] K. Cegielska et al., “Land use and land cover changes in post-socialist countries: Some observations from Hungary and Poland,” Land use policy, vol. 78, no. April, pp. 1–18, 2018.

[2] P. H. Verburg, J. van de Steeg, A. Veldkamp, and L. Willemen, “From land cover change to land function dynamics: A major challenge to improve land characterization,” J. Environ. Manage., vol. 90, no. 3, pp. 1327–1335, 2009.

[3] P. Gong et al., “Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data,” Int. J. Remote Sens., vol. 34, no. 7, pp. 2607–2654, 2013.

[4] C. Zhang et al., “An object-based convolutional neural network (OCNN) for urban land use classification,” Remote Sens. Environ., vol. 216, no. April, pp. 57–70, 2018.

[5] E. Gumus and P. Kirci, Selection of spectral features for land cover type classification, vol. 102. Elsevier Ltd, 2018.

[6] M. Drusch et al., “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services,” Remote Sens. Environ., vol. 120, pp. 25–36, 2012.

[7] K. Segl, L. Guanter, F. Gascon, T. Kuester, C. Rogass, and C. Mielke, “S2eteS: An End-to-End Modeling Tool for the Simulation of Sentinel-2 Image Products,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 10, pp. 5560–5571, 2015.

[8] Q. Wang and P. M. Atkinson, “Spatio-temporal fusion for daily Sentinel-2 images,” Remote Sens. Environ., vol. 204, no. April, pp. 31–42, 2018.

[9] M. Belgiu and L. Drăgu, “Random forest in remote sensing: A review of applications and future directions,” ISPRS J. Photogramm. Remote Sens., vol. 114, pp. 24–31, 2016.

[10] L. Collins, P. Griffioen, G. Newell, and A. Mellor, “The utility of Random Forests for wildfire severity mapping,” Remote Sens. Environ., vol. 216, no. December 2017, pp. 374–384, 2018.

[11] A. Mellor, S. Boukir, A. Haywood, and S. Jones, “Using ensemble margin to explore issues of training data imbalance and mislabeling on large area land cover classification,” 2014 IEEE Int. Conf. Image Process. ICIP 2014, vol. 105, pp. 5067–5071, 2014.
[12] D. Phiri, J. Morgenroth, C. Xu, and T. Hermosilla, “Effects of pre-processing methods on Landsat OLI-8 land cover classification using OBIA and random forests classifier,” Int. J. Appl. Earth Obs. Geoinf., vol. 73, no. April, pp. 170–178, 2018.

[13] D. Colson, G. P. Petropoulos, and K. P. Ferentinos, “Exploring the Potential of Sentinels-1 & 2 of the Copernicus Mission in Support of Rapid and Cost-effective Wildfire Assessment,” Int. J. Appl. Earth Obs. Geoinf., vol. 73, no. July, pp. 262–276, 2018.

[14] L. Leroux, L. Congedo, B. Bellón, R. Gaetano, and A. Bégué, “Land Cover Mapping Using Sentinel-2 Images and the Semi-Automatic Classification Plugin: A Northern Burkina Faso Case Study,” QGIS Appl. Agric. For., vol. 2, pp. 119–151, 2018.

[15] L. Yan, D. P. Roy, Z. Li, H. K. Zhang, and H. Huang, “Sentinel-2A multi-temporal misregistration characterization and an orbit-based sub-pixel registration methodology,” Remote Sens. Environ., vol. 215, no. April, pp. 495–506, 2018.

[16] M. J. Steinhausen, P. D. Wagner, B. Narasimhan, and B. Waske, “Combining Sentinel-1 and Sentinel-2 data for improved land use and land cover mapping of monsoon regions,” Int. J. Appl. Earth Obs. Geoinf., vol. 73, no. April, pp. 595–604, 2018.

[17] Z. Chunhui, G. Bing, Z. Lejun, and W. Xiaoqing, “Classification of Hyperspectral Imagery based on spectral gradient, SVM and spatial random forest,” Infrared Phys. Technol., vol. 95, pp. 61–69, 2018.

[18] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, “An assessment of the effectiveness of a random forest classifier for land-cover classification,” ISPRSJ. Photogramm. Remote Sens., vol. 67, no. 1, pp. 93–104, 2012.

[19] D. C. Duro, S. E. Franklin, and M. G. Dubé, “Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests,” Int. J. Remote Sens., vol. 33, no. 14, pp. 4502–4526, 2012.

[20] B. Melville, A. Lucieer, and J. Aryal, “Object-based random forest classification of Landsat ETM+ and WorldView-2 satellite imagery for mapping lowland native grassland communities in Tasmania, Australia,” Int. J. Appl. Earth Obs. Geoinf., vol. 66, no. November 2017, pp. 46–55, 2018.

[20] B. Melville, A. Lucieer, and J. Aryal, “Object-based random forest classification of Landsat ETM+ and WorldView-2 satellite imagery for mapping lowland native grassland communities in Tasmania, Australia,” Int. J. Appl. Earth Obs. Geoinf., vol. 66, no. November 2017, pp. 46–55, 2018.