Object-Based land cover classification of the Vu Gia – Thu Bon river basin on the cloud computing platform

View the article online for updates and enhancements.

Object-Based land cover classification of the Vu Gia – Thu Bon river basin on the cloud computing platform

N V Nghia[1*], T TH Thu[2], B TH Tham Hong Thi Bui[2], D T Huong[3], N T Cuc[3]

[1] Department of Mine Surveying, Hanoi University of Mining and Geology, Hanoi, Vietnam.
[2] Department of Elementary Surveying, Hanoi University of Natural Resources and Environment, Hanoi, Vietnam
[3] Faculty of Basic Sciences, Hanoi University of Mining and Geology, Hanoi, Vietnam.

E-mail: nguyenvietnghia@humg.edu.vn; nguyenvietnghia@gmail.com

Abstract. The arrival of cloud computing platform Google Earth Engine (GEE) in 2010 has brought a breakthrough for analysing and processing spatial data. Applying algorithms on this platform has overcome the limitations of commercial software while processing data in building thematic databases, including land cover data. These data are a critical factor for climate change and hydrological models. This study applied Object-based Random Forest (RF) classification in the Google Earth Engine platform to produce land cover data from Landsat 8 data of the Vu Gia - Thu Bon river basin. The classification results showed 7 categories of land cover consisting of artificial forest, natural forest, paddy area, urban area, rural area, bare land, and body water, with an overall accuracy Kappa of 0.70.

1. Introduction

Remote sensing is known as a science and an art of Earth surface information acquisition without any physical direct contact [11]. The most fundamental factor of remote sensing is its incoming electromagnetic radiation which is reflected, transmitted or absorbed by the surface when incident upon the surface.

There are different methods for classifying and extracting object information from remote sensing images, such as unsupervised classification, supervised classification, threshold classification, fuzzy logic-based method,… [13]. Based on these methods, specialised software (or commercial software) has been established (like ERDAS, LPS, ENVI, GEOMATICA,...) to automate image processing.

One of the noticeable advantages of commercial software is that information classification and extraction on remote sensing images can be quickly carried out with high accuracy and reliability. However, for large areas, due to its hardware and software limitations, the software did not promote superiority, its speed of processing was slow linking and co-operating objects between images, areas,... encountered many difficulties; leading to increased time and cost of project implementation [12]. However, Google Earth Engine (GEE) has overcome these disadvantages. GEE is an application built on cloud computing platform. GEE consists of two platforms: a graphical user interface called Explorer (https://explorer.earthengine.google.com/#workspace) and Javascript's application programming interface (API) called the Code Editor (https://code.earthengine.google.com/) [6]. GEE can be accessed with a web-based integrated development environment (IDE) through the JavaScript API. While the

web platform (IDE) enables users to easily visualise images, tables and charts; the Python API allows sending requests to the tool and access to the catalog. However, API cannot visualise the wed IDE platform [15]. The GEE Python Library processes requests to GEE and then receives the results. The information sent back in JavaScript will be presented in the browser. While graphical data is visualised with the Google API, spatial information is shown with the Google Maps application programming interface. On this interface, users can write and run scripts to share and repeat processes of geospatial data analysis and processing. GEE enables environmental geographic data analysis on a global scale with a storage capacity of pentabytes of image data [12]. GEE data is aggregated from various sources of satellite image data from NASA, NOAA, ESA and others. GEE takes advantage of a computational system optimized for parallel geospatial data processing. According to [12], GEE is widely applied in many fields from agroforestry, ecosystems to economics and epidemics, among which the most numerous is studies using GEE for forests and Plant cover, followed by land use, ecosystems, moist soil and hydrology.

Since GEE is a very powerful application, users not only can solve large-scale problems, take advantage of the huge free photo resources but also can flexibly devise object classification and extraction programs suitable for research and application purposes.

Land cover/land use data from remote sensing images can be built based on approach scales and classification algorithms. In general, the approach scales can be divided into three categories: approach by pixels, approach by sub-pixels and approach by objects (super pixels). The pixel-based approach scale usually relies on the spectral value of each pixel; as a consequence, the obtained land cover data after classification are often spotted with other land cover types, especially when the image has a high resolution [4]. That is why currently the object-based approach is preferred because this approach takes contextual information into account and can eliminate spots contamination in classification results.

As mentioned above, there are many different methods of classification. Classification methods are often grouped into categories such as inspected, unqualified, or parametric and non-parametric, or hard and soft classifications. Classification algorithms are affected by many parameters such as the selection of classification samples, the uniformity of the study area, the sensor, the number of classification classes... Therefore, the choice of algorithms and calculation parameters (classification samples, the uniformity of the study area, sensors, the number of classification classes...) will determine the accuracy and reliability of land cover map and land use in the study area. Currently, the machine learning classification method (MLC) is the most reliable classification method and is increasingly used by many remote sensing [5,8,14].

Vu Gia - Thu Bon river basin is a large river system, most of which flows in the territory of Quang Nam and Da Nang cities, while the upper part partly lies on Kon Tum and Quang Ngai lands in Central Vietnam. The Vu Gia - Thu Bon river basin plays an important role in the socio-economic development of these localities like hydroelectricity development and mineral exploitation.

In this study, GEE application is used to classify and extract object information on Landsat 8 images of Vu Gia - Thu Bon river basin. The approach to classify objects in the study area is based on the Random Forest (RF) classification method, currently the most reliable classification algorithm of Machine learning classification algorithms (MLC).

2. Basis of research method

2.1. Random Forest Algorithm
In fact, the land cover classification on remote sensing images is a form of image clustering. The Random Forest algorithm in this study is a machine learning algorithm used to solve both classification and regression problems.

Random trees are trees which are randomly generated from a set of trees with K random characteristics at each node. They can be efficiently constructed and the combination of big sets of random trees can result in high prediction accuracy [2]. Besides, Random Forest is not only efficient, interpretable but also non-parametric for different kinds of datasets. Because of its model interpretability

and prediction accuracy, Random Forest becomes more special than other machine learning methods [2].

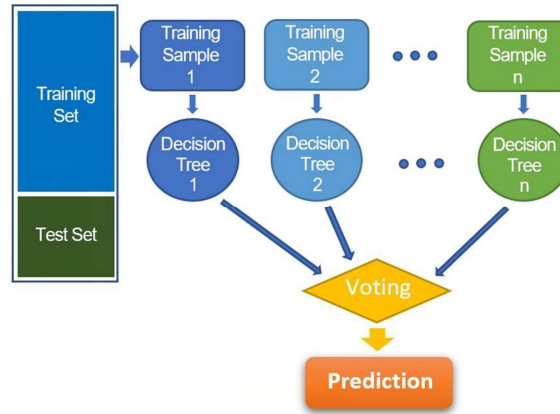The operation of the Random Forest algorithm is presented in Figure 1. [9,20]:

Figure 1. Random Forest algorithm diagram

As shown in the Figure 1, the following steps shall be done to complete the operation:

- Step 1: Select random samples from a given data set.
- Step 2: Set up a decision tree for each sample and get predictable results from each tree decision.
- Step 3: Vote for each predicted result.
- Step 4: Choose the most predicted result as the final prediction.

The Random Forest algorithm is an accurate and powerful method because of the number of trees that participate in the process. This algorithm does not suffer from overfitting problems because it takes the average of all predictions, which cancels biases. Besides, Random Forest algorithm can also handle missing values. Accordingly, two ways to handle these values are using the mean to replace continuous variables and calculating the average of the missing values. Then, the relatively important features can be obtained, from which the features that contribute the most to classifiers are chosen.

2.2. Some basic calculations in the study

Basic calculations in the study are mentioned as follows:

The first calculation is NDVI (normalized difference vegetation index). The NDVI is measured by the following formula (1) [18,19].

$$NDVI = (NIIR - R)/(NIR + R) \tag{1}$$

where: NIR is the spectral reflectance value in the near infrared channel;

R is the reflectance value of the red channel.

The second is SNIC clustering algorithm to segment images. The SNIC algorithm starts with initialising the central pixels (pixel centroid) with the pixels selected in the image plane. The relationship of pixels with the center point is measured by the distance in 5-dimensional space (color space and spatial position) according to the formula (2) [1].

$$d_{j,k} = \sqrt{\frac{\|x_j - x_k\|_2^2}{s} + \frac{\|c_j - c_k\|_2^2}{m}} \tag{2}$$

where: $x = [x \quad y]^T$ is the spatial position;;

$c = [l \quad a \quad b]^T$ is the CIELAB color space;

s and m are normalisation factors for spatial and color distances.

For images with N pixels, each resulting cluster of K images will contain N / K pixels. Assuming the cluster is square, the value s in equation (2) will be:

$$s = \sqrt{N / K} \tag{3}$$

where: m is the tightness factor, selected and provided by the user.

The next calculation is the quality of classified data, which is based on evaluation criteria determined from the error matrix. Accuracy evaluation criteria include the following factors:

- The Overall accuracy (OA), determined by the total number of correct classification points divided by the total number of points [3,6].

$$OA = \left(\sum_{i=1}^{k} x_{ii} / N \right) \times 100\% \tag{4}$$

- Omission error (OE): $OE = \left[\left(\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i+1} \right) \Big/ \left(\sum_{i=1}^{k} x_{i+1} \right) \right] \times 100\%$ (5)

- Producer's accuracy (PA): $PA = \left(x_{ii} / \sum_{i=1}^{k} x_{1+i} \right) \times 100\%$ (6)

- Comission error (CE): $CE = \left[\left(\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{1+i} \right) \Big/ \left(\sum_{i=1}^{k} x_{1+i} \right) \right] \times 100\%$ (7)

- The user's accuracy (UA): $UA = \left(x_{ii} / \sum_{i=1}^{k} x_{i+1} \right) \times 100\%$ (8)

- Accuracy of each cover (F): $F = 2 \times (PA \times UA) / (PA + UA)$ (9)

- Kappa coefficient, a measure of synchronous accuracy of classification data with reference data:

$$Kappa = \left(N \times \sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i+1} \times x_{1+i} \right) \Big/ \left(N^2 - \sum_{i=1}^{k} x_{i+1} \times x_{1+i} \right) \tag{10}$$

where: N is the total number of samples, x_{ii} is the components on the main diagonal, x_{i+1} and x_{1+i} are the total components in rows and total in columns in the error matrix, k is the number of overlays.

3. The study area

The Vu Gia - Thu Bon River basin is a large river system in the Central Coast region of Vietnam. This system originates from Kon Tum province, flows through Quang Nam province, Da Nang city and then runs into the East Sea at Cua Dai and Cua Han. Da Nang city and Quang Nam province are two provinces located in the central economic region. In addition, the Vu Gia - Thu Bon River basin is one of the nine biggest river basins in Vietnam, stretching from $14^{0}57'10''$ to $16^{0}03'50''$ of North latitude, $107^{0}12'50''$ to $108^{0}44'20''$ of East longitude. The case study area is situated in Central Vietnam, covering an area of over 10,000 km^2.

This area has a complex and strongly divided topography. The terrain tends to tilt from the West to the East, with all kinds of landscape like high mountains, midlands, coastal plains and sand dunes. The mountainous terrain has an average altitude of 700÷800 m with the highest point of over 2000, the hilly terrain has an average altitude of 100÷200 m with wavy and bowl-formed shape, the coastal plains are relatively flat, with the elevation of below 30 m, including narrow eastern plains and coastal sand dunes/beaches [17].
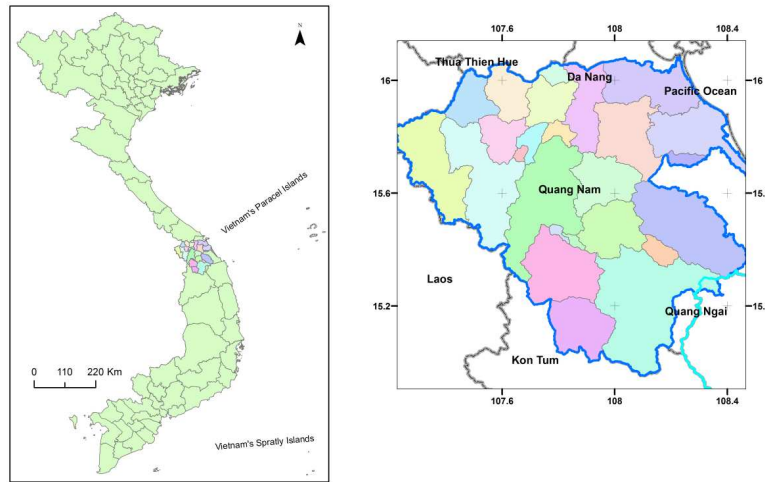
Figure 2. The Vu Gia - Thu Bon river basin

3.1. Materials

In this study, two multispectral cloud – free Landsat 8 OLI - TIRS images (path 126, row 46) with a spatial resolution of 30×30 meters taken from 01 May to 31, May 2014 during the dry season with cloud cover below 10% were used. These Landsat standard terrain correction products (L1T) were obtained from United States Geological Survey (USGS – http://glovis.usgs.gov) website.

It is known that the Landsat 8 satellite was launched on February 11, 2013 by NASA and operated in a sun-synchronous orbit with a 16 - day repeat cycle. The Landsat 8 OLI_TIRS sensors acquired 11 spectral bands along a 185 km orbital swath.

In addition, the study also used SRTM global elevation model. It is reported that SRTM which is managed by NASA's Jet Propulsion Laboratory is the joint project between NASA, German and Italian space agencies, and National Geospatial-Intelligence Agency (https://www.jpl.nasa.gov/news/news.php?release=2014-321). Moreover, 980 field samples were taken to evaluate the accuracy of classification results.

3.2. Methodology

The land cover mapping of the Vu Gia - Thu Bon river basin according to the object-based RF classification on the GEE cloud computing platform is modeled in the following diagram:
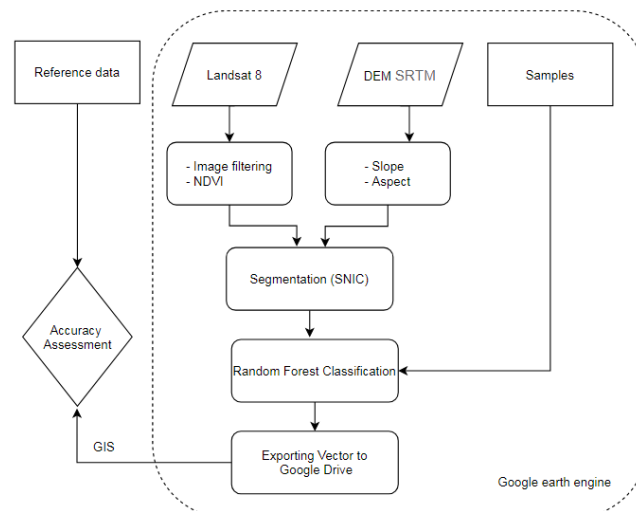

Figure 3. Steps for land cover mapping of the Vu Gia - Thu Bon river basin on Google earth engine

As presented in the above figure, the mapping process includes the following steps:

- Step 1: Declare the input data.

The input data includes Landsat 8 satellite imagery, STRM height data model (DEM) of the experimental area, catchment boundary, and control sample data.

- Step 2: Calculate NDVI according to the formula (1).

From the elevation numerical model, a slope map (Slope) is created.

- Step 3: Segment images by SNIC clustering algorithm.

This is the process of aggregating single pixels into an object while regarding the contextual information of the adjacent data area. Image segmentation creates areas or objects based on specific parameters such as geometry, scale, uniformity.

The SNIC clustering algorithm (presented in section 2.2.) was used to segment images in this study. Starting from the central pixel, the SNIC algorithm selected the next pixel to add to the cluster; the selected pixel was the pixel with the smallest distance to the central pixel in 4÷8 pixels nearby the central pixel. The segmentation data consisted of 6 image channels of Landsat 8 data (3 channels in the visible range, 01 near-infrared channel, and 02 medium-infrared channels), NDVI data, STRM data, Slope data and Aspect. In this study, the selected parameters to suit the image segmentation of the study area included the ratio size of 30, density of 8, adjacent data of 8, neighborhood size of 256.
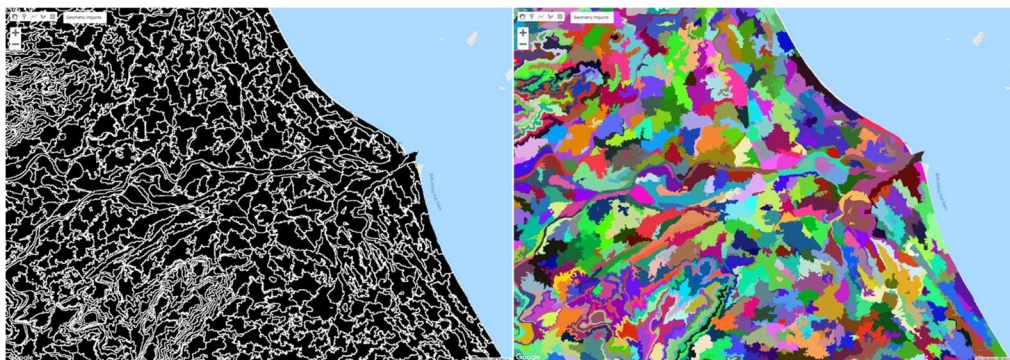

Figure 4. Results of image segmentation in the study area

- Step 4: Classify land cover and select sample data.

The sample data were randomly selected with the sample size of 764 polygon for all 7 types of land cover including paddy area, rural area, urban area, natural forest, artificial forest, bare land, and body water.

Table 1. Description of land cover of the study area

No.	Type of land cover	Description	Number of samples
1	Paddy area	Paddy areas, from May 1, 2014 to May 31, 2014, when paddy had already been harvested	180
2	Rural area	Areas of low construction density land mixed with vegetation	80
3	Artificial	Forests planted on non-forest land; or reclaimed med from natural	79
4	Natural	Forests often in high places and difficult to access	120
5	Bare land	Areas of sandy, barren land	39
6	Body water	Rivers, lakes and water entities	176
7	Urban area	Residential areas with a high construction density	88

- Step 5: Verify land cover classification on GEE.

GEE verified classification was applied to segmented images. The sample values were taken on the image information from 6 Landsat 8 image channels and NDVI, STRM. As the result, the program code

classified by object-based RF method is shown in Table 2.

Table 2. Program code classified by Random forest method

```
//training data
var mm=lua1.merge(ont1).merge(rsx1).merge(rtn1).merge(dt).merge(n1).merge(odt1) ;
print(mm,'mm');
// select training data
var training = objectPropertiesImage.addBands(addndvi).select(['b1','b2','b3','b4','b5','b6','nd','slope','aspect','b1_
.updateMask(seeds)
.sampleRegions({
  collection: mm,
  properties: ['landcover'],
  scale: 30
});
print(training,'training');
 //Classification method randomForest
var classifier = ee.Classifier.randomForest(100).train({
  features: training,
  classProperty: 'landcover',
  inputProperties: ['b1','b2','b3','b4','b5','b6','nd','b1_1']
});
print(classifier,'classifier');
//Clasified
var classified = objectPropertiesImage.classify(classifier);
print(classified,'classified');
Map.addLayer(classified, {min: 0, max: 8, palette: palette}, 'Land cover Classification');
```

   - Step 6: Evaluate classification accuracy.

The criteria for evaluating classification accuracy were programmed on the basis of the formulas (4), (5), (6), (7), (8), (9) and (10). The accuracy of classified products were evaluated by reference data. The reference data were independent points randomly sampled on 2014 Google Earth photos and evenly spreading across the study area. The error matrix (in Table 3) is a symmetric spatial matrix to compare similarities and differences of classified products with reference data on ArcGIS 10.3.

Table 3. Error matrix

		GEE classification										
		0	1	2	3	4	5	6	sum	PA	CE	F
	0	40.0	5.0	10.0	12.0	0.0	0.0	0.0	67.0	59.7	40.3	69.0
	1	2.0	33.0	9.0	3.0	0.0	0.0	1.0	48.0	68.8	31.3	66.7
	2	1.0	5.0	172.0	51.0	0.0	0.0	0.0	229.0	75.1	24.9	72.1
GEE reference	3	1.0	2.0	52.0	538.0	0.0	0.0	0.0	593.0	90.7	9.3	89.8
	4	3.0	1.0	3.0	0.0	8.0	0.0	0.0	15.0	53.3	46.7	69.6
	5	2.0	4.0	2.0	1.0	0.0	14.0	0.0	23.0	60.9	39.1	75.7
	6	0.0	1.0	0.0	0.0	0.0	0.0	4.0	5.0	80.0	100.0	80.0
Total		49.0	51.0	248.0	605.0	8.0	14.0	5.0	980.0	OA	Kappa	
UA		81.6	64.7	69.4	88.9	100.0	100.0	80.0		0.8	0.7	
OE		18.4	35.3	30.6	11.1	0.0	0.0	20.0				

3.3. Experimental results

The object-based classification results on google earth engine of the study area are presented in Figure 5 with the whole classification in the left image and the partial ones zoomed in on the Google earth engine in the right image.

The area and spatial distribution of 7 types of land cover including paddy area, rural area, urban area, natural forest, artificial forest, bare land, and body water are shown in Table 4 and Figure 5. The total area of the study area is 10068.082 km^2. The natural forest covers the largest area of 5989.667 km^2,

accounting for 59.49% of the total study area, which is shown in dark green in Figure 6. This type of land cover is concentrated in the Southwest with typically high hills. Artificial forest is the second largest land cover area of 2656.417 km^2, making up 26.38% of the total study area. Urban area has the smallest area of 41.052 km^2, contributing only 0.41% to the total study area.
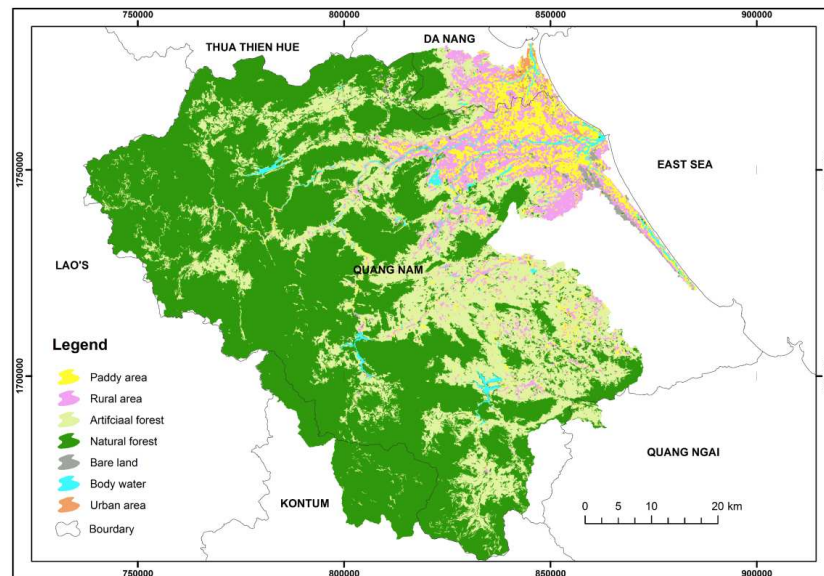

Figure 5. Object-based classification results of the Vu Gia - Thu Bon river basin.

Table 4. The area of 7 land cover types

No.	Land cover	Area (km^2)
1	Paddy area	537.263
2	Rural area	674.761
3	Artificial forest	2656.417
4	Natural forest	5989.667
5	Bare land	49.775
6	Body water	119.147
7	Urban area	41.052
	Total	**10068.082**

4. Conclusion

On the basis of Google earth engine cloud computing technology, a map of the Vu Gia - Thu Bon river basin in the Central region of Vietnam was established with 7 types of land cover, namely paddy area, rural area, urban area, natural forest, artificial forest, bare land, and body water. The accuracy of land cover classification was assessed according to different types of indexes, showing that the classification accuracy reflected not only the overall error of the classification results but also the accuracy of the individual land covers through the neutral F coefficient between UA and PA. The land cover classification results of the Vu Gia - Thu Bon river basin on the GEE platform had the Kappa of 0.7.

For each study area, users need to get a sufficiently large number of samples for classification. In addition, a survey to get a set of parameters (such as segmentation size, density, adjacent data, neighborhood size) that is suitable for the study area should also be conducted. However, since GEE is an application that uses an internet connection, the processing speed of the program depends on the speed of the network connection. Therefore, a stable, high-speed network connection is needed when establishing a classification map, especially for large areas.

Acknowledgments

References

[1] Achanta R, Susstrunk S 2017 Superpixels and polygons using simple non-iterative clustering *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

[2] Ali J, Khan R, Ahmad N, Maqsood I 2012 Random forests and decision trees *International Journal of Computer Science Issues (IJCSI)* **9(5)** 272

[3] Ao Z, Su Y, Li W, Guo Q, Zhang J 2017 One-class classification of airborne LiDAR data in urban areas using a presence and background learning algorithm *Remote Sensing* **9(10)** 1001

[4] Blaschke T 2010 Object based image analysis for remote sensing *ISPRS J. Photogramm. Remote Sensing* **65** 2–16

[5] Campos-Taberner M, García-Haro F J, Camps-Valls G, Grau-Muedra G, Nutini F, Crema A, Boschetti M 2016 Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring *Remote Sensing of Environment* **187** 102-118

[6] Congalton R G, Green K 2019 Assessing the accuracy of remotely sensed data: principles and practices *CRC press*

[7] Dean J, Ghemawat S 2004 *MapReduce: Simplified data processing on large clusters*

[8] García-Haro F J, Campos-Taberner M, Muñoz-Marí J, Laparra V, Camacho F, Sánchez-Zapero J, Camps-Valls G 2018 Derivation of global vegetation biophysical parameters from EUMETSAT Polar System *ISPRS journal of photogrammetry and remote sensing* **139** 57-74

[9] Gislason P O, Benediktsson J A, Sveinsson J R 2006 Random forests for land cover classification. *Pattern Recognition Letters* **27(4)** 294-300

[10] *U.S. Releases Enhanced Shuttle Land Elevation Data* Available at: https://www.jpl.nasa.gov/news/news.php?release=2014-321

[11] Jensen J 2006 *Remote sensing of the environment an earth resource perspective: low price edition Pearson Education* (New Delhi, India)

[12] Kumar L, Mutanga O 2018 Google Earth Engine applications since inception: Usage, trends, and potential *Remote Sensing* **10(10)** 1509

[13] Li M, Zang S, Zhang B, Li S, Wu C 2014 A review of remote sensing image classification techniques: The role of spatio-contextual information *European Journal of Remote Sensing* **47(1)** 389-411

[14] Mahdianpari M, Salehi B, Mohammadimanesh F, Homayouni S, Gill E 2019 The first wetland inventory map of newfoundland at a spatial resolution of 10 m using sentinel-1 and sentinel-2 data on the google earth engine cloud computing platform *Remote Sensing* **11(1)** 43

[15] Mateo-García G, Gómez-Chova L, Amorós-López J, Muñoz-Marí J, Camps-Valls G 2018 Multitemporal cloud masking in the Google Earth Engine *Remote Sensing* **10(7)** 1079

[16] Mutanga O, Kumar L 2019 *Google Earth Engine Applications: Multidisciplinary Digital Publishing Institute*

[17] Tran V T 2013 *Building flood maps in the downstream areas of the Vu Gia – Thu Bon river basin* (Vietnamese)

[18] Tucker C, Miller L D, Pearson R 1973 Measurement of the combined effect of green biomass, chlorophyll, and leaf water on canopy spectroreflectance of the shortgrass prairie. *Remote Sensing of Earth Resources*

[19] Tucker C J, Miller L D, Pearson R L 1975 *Shortgrass prairie spectral measurements* [for terrain analysis and photomapping]

[20] Waske B, Braun M 2009 Classifier ensembles for land cover mapping using multitemporal SAR imagery *ISPRS journal of photogrammetry and remote sensing* **64(5)** 450-457