西 南 交 通 大 学 学 报

第 55 卷 第 2 期
2020 年 4 月

JOURNAL OF SOUTHWEST JIAOTONG UNIVERSITY

Vol. 55 No. 2
Apr. 2020

**ISSN: 0258-2724** **DOI：10.35741/issn.0258-2724.55.2.61**

Research article

Earth Sciences

# LAND COVER CLASSIFICATION BASED ON CLOUD COMPUTING PLATFORM

# 基于云计算平台的土地覆被分类

**Nghia Viet Nguyen [a,] *, Thu Hoai Thi Trinh [b], Hoa Thi Pham [b], Trang Thu Thi Tran [b], Lan Thi Pham [a], Cuc Thi Nguyen [a]**

[a] Hanoi University of Mining and Geology
No.18 Vien St., Duc Thang ward, Bac Tu Liem district, Hanoi, Vietnam, nguyenvietnghia@humg.edu.vn
[b] Hanoi University of Natural Resources and Environment
No 41 A Phu Dien road, Phu Dien precinct, North-Tu Liem district, Hanoi, Vietnam

***Received: February 01, 2020 ▪ Review: April 2, 2020 ▪ Accepted: April 25, 2020***



**Abstract**

Land cover is a critical factor for climate change and hydrological models. The extraction of land cover data from remote sensing images has been carried out by specialized commercial software. However, the limitations of computer hardware and algorithms of the commercial software are costly and make it take a lot of time, patience, and skills to do the classification. The cloud computing platform Google Earth Engine brought a breakthrough in 2010 for analyzing and processing spatial data. This study applied Object-based Random Forest classification in the Google Earth Engine platform to produce land cover data in 2010 in the Vu Gia - Thu Bon river basin. The classification results showed 7 categories of land cover consisting of plantation forest, natural forest, paddy field, urban residence, rural residence, bare land, and water surface, with an overall accuracy of 73.9% and kappa of 0.70.

**Keywords:** Google Earth Engine, Classification, Land Cover, Vu Gia-Thu Bon River Basin

**摘要** 土地覆盖是影响气候变化和水文模型的关键因素。从遥感图像中提取土地覆盖数据已经通过专门的商业软件进行。但是，计算机硬件和商业软件算法的局限性是昂贵的，并且需要很多时间，耐心和技巧来进行分类。云计算平台谷歌地球引擎在 2010 年为分析和处理空间数据带来了突破。本研究在 GGE 平台中应用了基于对象的随机森林分类，以产生 2010 年 Vu Gia-Thu Bon 流域的土地覆盖数据。分类结果显示，人工林，天然林，水田，城市居住地，农村居住地，光秃秃的土地和水面共 7 种土地覆被，总准确度为 73.9%，卡帕为 0.70。

**关键词:** 谷歌地球引擎，分类，土地覆盖，Vu Gia-Thu Bon 流域

## I. INTRODUCTION

Building land cover and land use data from remote sensing images has often been done using algorithms extracted from commercial software such as ERDAS, LPS, ENVI, GEOMATICA, and ECOGNITION. However, with large areas, the use of commercial software is limited by computer hardware and algorithms from commercial software, leading to slow processing speeds and higher prices for mapping. Since its appearance, Google Earth Engine (GEE) has made great progress in significantly enhancing its computing power and becoming completely free to users. GEE is a cloud computing platform designed to store and process huge data sets (at petabyte-scale), which currently stores an extensive catalog of earth observation data including satellite image data and vector data [1], [6], [8].

The computing approach of GEE is an application programming interface (API) that integrates JavaScript and Python, allowing for the easy development of parallel algorithms suitable for large data analysis. On the one hand, GEE is accessible through a web-based integrated development environment (IDE) using the JavaScript API. The web platform (IDE) allows users to visualize images and analyze results, tables, and charts easily. On the other hand, the Python API provides the same set of methods for making requests to the tool and accessing the catalog but does not allow the visualization of the web IDE [6]. The GEE Python library processes requests to GEE and receives results. The information returned to JavaScript is displayed in the browser. Spatial information is displayed with the Google Maps application programming interface and graphical data is visualized with the Google API.

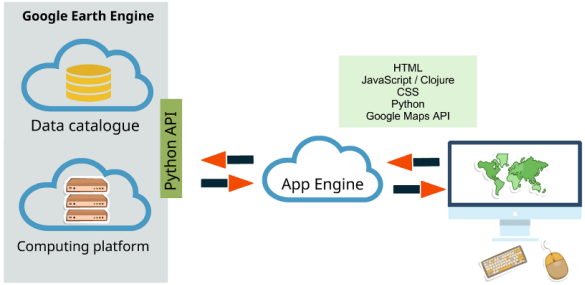


Figure 1. The infrastructure for developing spatial applications provided by Google [1] (cloud computing platform – data catalog – Google Earth Engine)

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 (superpixels). 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 [2]. That is why currently the object-based approach is preferred because this approach takes contextual information into account and can eliminate spot contamination in classification results.

Classification algorithms are often categorized into small groups, such as tested and non-tested classifications, parametric and non-parametric classifications, or hard and soft classifications. However, these classification algorithms are often affected by parameters such as the selection of classification samples, the uniformity of the study areas, the sensors, and the number of classification classes. Therefore, new classification algorithms with higher accuracy classification are constantly being developed. Among them, machine learning classification (MLC) is one of the most reliable and increasingly used classification methods in remote sensing [3], [4], [7].

For the above-mentioned reasons, the authors of this paper have developed land cover data for an area in the Vu Gia-Thu Bon river basin based on GEE. An object-based classification approach and random forest (RF) classification (an algorithm of MLC classification) were used for Landsat thematic mapper (TM) images from 2010.

## II. STUDY AREA

The selected study area was a section of the Vu Gia-Thu Bon river basin (Figure 2). The basin is located in Central Vietnam and covers over 10,000 $km^2$. The Vu Gia-Thu Bon river basin is one of the nine largest basins in Vietnam, stretching from $14^0$57'10" to $16^0$03'50" North latitude and from $107^0$12'50" to $108^0$44'20" East longitude. It includes a small part of Kon Tum Province, the whole of Quang Nam Province, and the city of Da Nang.

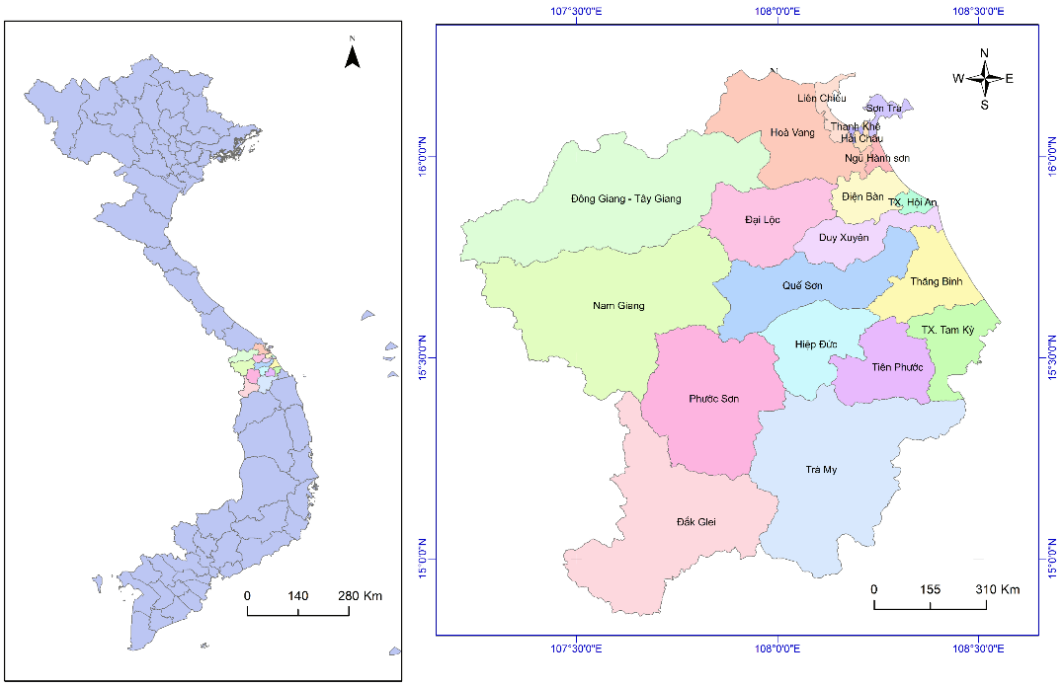


Figure 2. Vu Gia-Thu Bon river basin

The study area was over 5,000 km$^2$ in size. Its terrain is quite complex, strongly divided, and tends to gradually tilt from west to east. This area has a variety of terrain types like high mountains in the west, midlands in the center, and narrow plains and coastal sand dunes to the east. While the mountainous terrain has an average altitude of 700–800 m with the highest over 1500 m, the hilly terrain has an average altitude of 100–200 m. The coastal plains are relatively flat with altitudes below 30 m.

## III. Land Cover Mapping

Figure 3 shows the steps for land cover mapping of the study area in the Vu Gia-Thu Bon basin based on the object-based approach scale and RF classification on the GEE cloud-computing platform.

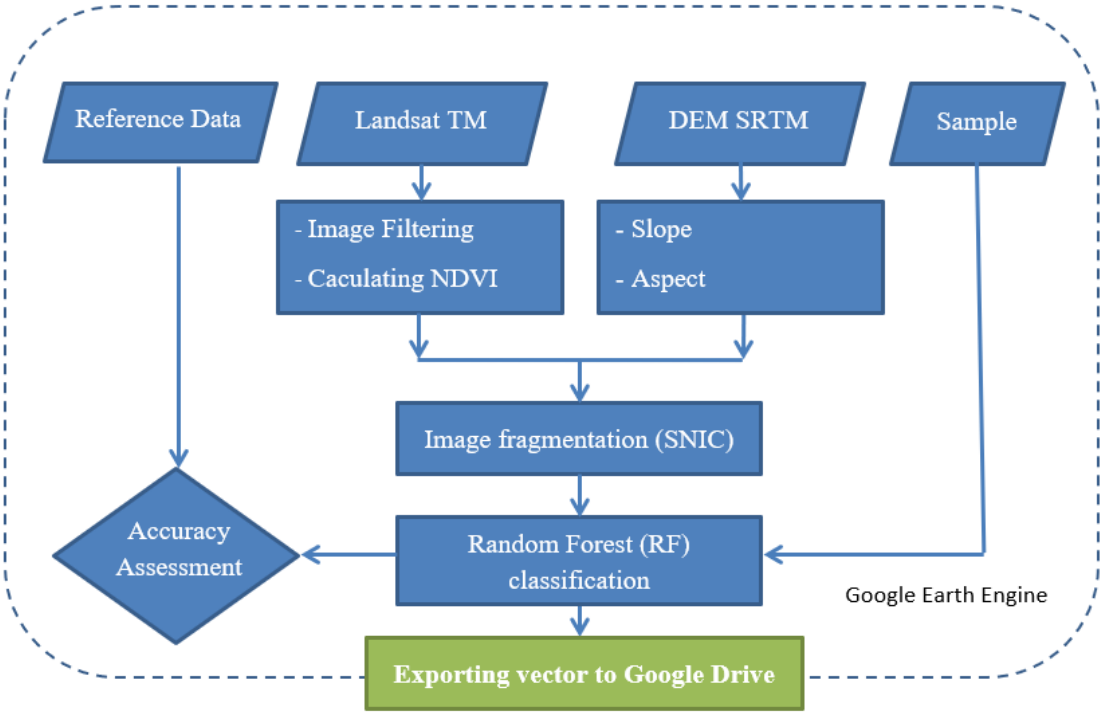


Figure 3. Map of steps for land cover mapping of the study area in the Vu Gia-Thu Bon basin on Google Earth Engine

First, the input including the Landsat 5 image database and Shuttle Radar Topography Mission DEM data at 30 m resolution was prepared. Next, study area images for March 24, 2010 from the

Landsat 5 image database were selected. Then, the NDVI plant index was calculated according to the following formula and the slope was mapped:

$$NDVI = (NIIR - R)/(NIR + R) \quad (1)$$

where *NIR* is the spectral reflectance value in the near infrared channel (channel 4), and *R* is the reflectance value of the red channel (channel 3).

Second, image fragmentation, the process of aggregating single pixels into an object considering the contextual information of the adjacent data areas, was carried out. The image fragmentation creates areas or objects based on specific parameters such as geometry, scale, and uniformity. In this study, the simple non-iterative clustering (SNIC) algorithm was applied for image fragmentation.

The algorithm starts with initializing the central pixels by the pixels selected in the image plane. The relationship between these pixels and the central pixels is measured by the distance in the five-dimensional space (color space and spatial position) according to the following formula [9]:

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

where spatial position $x = [x \quad y]^T$, the CIELAB color space $c = [l \quad a \quad b]^T$, and *s* and *m* are the normalized factors for spatial and color distance.

For images with *N* pixels, each obtained pixel cluster *K* will have *N/K* pixels. Assuming the image cluster is square, the value *s* in Equation (2) will be $s = \sqrt{N/K}$. *m* is the tightness factor which is selected and supplied by the user.

Starting from the central pixel, the SNIC algorithm selects the next pixel to add to the cluster. The selected pixel is the pixel with the smallest distance to the central pixel among the 4–8 pixels near the central pixel.

The data to be fragmented included 6 image channels of Landsat 5 data (3 channels in the visible range, 1 near-infrared channel, and 2 medium-infrared channels), NDVI data, Shuttle Radar Topography Mission data, slope data, and aspect data. The parameters in this study were selected to suit the image fragmentation of the study area. These parameters included a fragmentation size of 30, density of 7, adjacent data of 4, and adjacent size of 256.

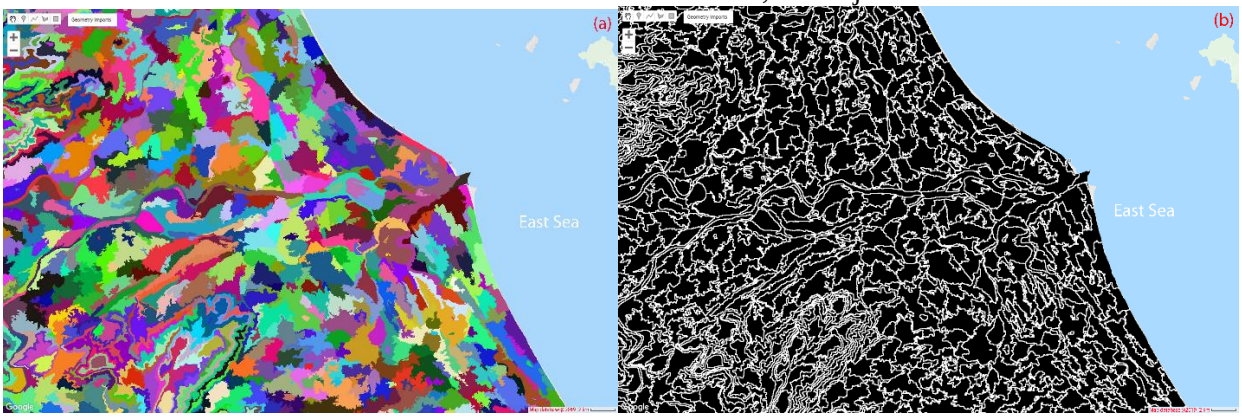


Figure 4. Results of image fragmentation in the study area

The third step identified the land covers and created sample data. Sample data and reference data were used for classification, and their accuracy were interpreted from the 2010 Google Earth images. Sample data were selected randomly with a sample size of 976 polygons for all 7 types of land cover. 07 types of land cover are classified: artificial forest (RTN), natural forest (RTN), paddy area (LUA), urban area (DDT), rural area (ONT), bare land (DT), and water system (TH).

The fourth step tested the classification performed on GEE with fragmented images. A sample value was taken on the image object with information from the 6 Landsat image channels and the complementary information NDVI, STRM. Random Forest classification method, a member of the decision tree classification algorithm, was selected.

The RF method builds a collection of decision trees from the data set, to ensure that decision trees are independent. RF then randomly selects the observation sample, and randomly selects the attribute. Each decision tree predicts a result, and the final result with the highest accuracy is the one predicted by many decision trees [5], [10]. The classification parameters and the code used to perform classification on the image object of the study area are illustrated in Figure 5.

```
//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');
```

Figure 5. The code to perform object-based classification

The fifth step assessed the accuracy of classification products. Assessing the accuracy of classification data is an important part of extracting data from remote sensing images, in order to determine the quality of the classification results. The reference data used in this study are random sample points, 100 sample points for each class. The samples were interpreted and identified on Google Earth images from 2010, then entered into the classification system on the Google Earth engine.

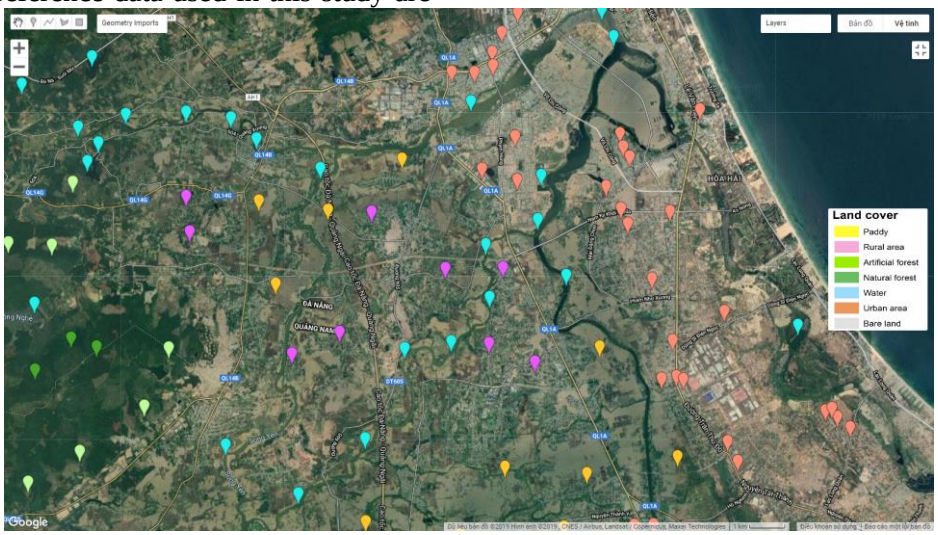


Figure 6. Sample distribution for accuracy assessment

The quality of classification data is based on evaluation criteria determined from the error matrix. Accuracy evaluation criteria include: Overall error (OA) determined by the total number of correct classification points divided by the total number of points, omission error (OE) production accuracy (PA), confusion error (CE, use accuracy (UA), accuracy of each land cover (F) and finally coefficient K which is a measure of uniformity accuracy of classification data with reference data corresponding to the following formulas:

$$OA = \left( \sum_{i=1}^{k} x_{ii} / N \right) \times 100\% \quad (3)$$

$$OE = \left[ \left( \sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i+1} \right) / \left( \sum_{i=1}^{k} x_{i+1} \right) \right] \times 100\% \quad (4)$$

$$PA = \left( x_{ii} / \sum_{i=1}^{k} x_{1+i} \right) \times 100\% \quad (5)$$

$$CE = \left[ \left( \sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{1+i} \right) / \left( \sum_{i=1}^{k} x_{1+i} \right) \right] \times 100\% \quad (6)$$

$$F = 2 \times (PA \times UA) / (PA + UA) \quad (7)$$

$$K_{hat} = \left( N \times \sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i+1} \times x_{1+i} \right) / \left( N^2 - \sum_{i=1}^{k} x_{i+1} \times x_{1+i} \right) \quad (8)$$

where N is the total number of samples, t, xii are the components on the main diagonal, $x_{i+1}$ and

$x_{1+i}$ are the totals by row and column in the error matrix, k is the number of land covers.

## IV. Results and Discussions

The accuracy of classification results are evaluated through reference data which are independent points sampled on 2010 Google Earth images randomly and evenly spread across the study area. The error matrix is a symmetric spatial matrix which allows the comparison of similarities and differences in properties of classification products. The reference data and the error matrix are shown in Table 1.

Table 1.
The error matrix

| | Paddy area | Rural area | Artificial forest | Natural forest | Bare land | Water system | Urban land | UA | OE | F |
|---|---|---|---|---|---|---|---|---|---|---|
| Paddy area | 66 | 16 | 4 | 3 | 0 | 0 | 4 | 71 | 290 | 68 |
| Rural area | 12 | 67 | 5 | 1 | 6 | 0 | 9 | 670 | 330 | 67 |
| Artificial forest | 8 | 2 | 75 | 12 | 0 | 2 | 0 | 758 | 24 | 754 |
| Natural forest | 5 | 5 | 16 | 84 | 0 | 0 | 0 | 764 | 23 | 80 |
| Bare land | 4 | 4 | 0 | 0 | 63 | 0 | 23 | 670 | 33 | 64 |
| Water system | 1 | 0 | 0 | 0 | 0 | 98 | 0 | 990 | 1 | 985 |
| Urban land | 4 | 6 | 0 | 0 | 31 | 0 | 64 | 610 | 390 | 62 |
| PA | 66.0 | 67.0 | 75.0 | 84.0 | 63.0 | 98.0 | 64.0 | | | |
| CE | 34.0 | 33.0 | 25.0 | 16.0 | 37.0 | 2.0 | 36.0 | | | |

According to the formula (3) and (8), from the error matrix, the overall error is 73.9% and the Kappa coefficient is 0.70. As shown in Table 1, water system land cover is extracted with the highest accuracy of 98.5% corresponding to the production accuracy of 98.0% and the use accuracy of 99.0%. The land cover of bare urban area has the lowest accuracy of 62.4% with the omission error of 36.0% and the confusion error up to 39.0%. Most of the confusion is between urban area and bare land.

The area and spatial distribution of seven types of land cover including artificial forest, natural forest, paddy area, urban area, rural area, bare land, and water system are shown in Table 2 and Figure 7. The total overall area is 5471.258 km2 (Table 1).

Natural forest is the largest type of land cover, comprising 3579.968 km2 and accounting for 65.4% of the total study area; this is shown in dark green in Figure 7. This type of land cover is concentrated in the southwest with high mountainous terrain. The area of paddy land cover ranks second, corresponding to 716.324 km2 and comprising 13.1% of the total study area; it is shown in yellow. The bare land cover has the lowest area at 15.651 km2 and accounts for 0.3% of the total study area.

Table 2.
The land cover area in the study area

| No. | Land cover | Area (km2) |
|---|---|---|
| 1 | Paddy area | 716.324 |
| 2 | Rural area | 511.235 |
| 3 | Artificial forest | 228.335 |
| 4 | Natural forest | 3579.968 |
| 5 | Bare land | 15.651 |
| 6 | Water system | 136.625 |
| 7 | Urban area | 283.121 |
| | Total | 5471.258 |

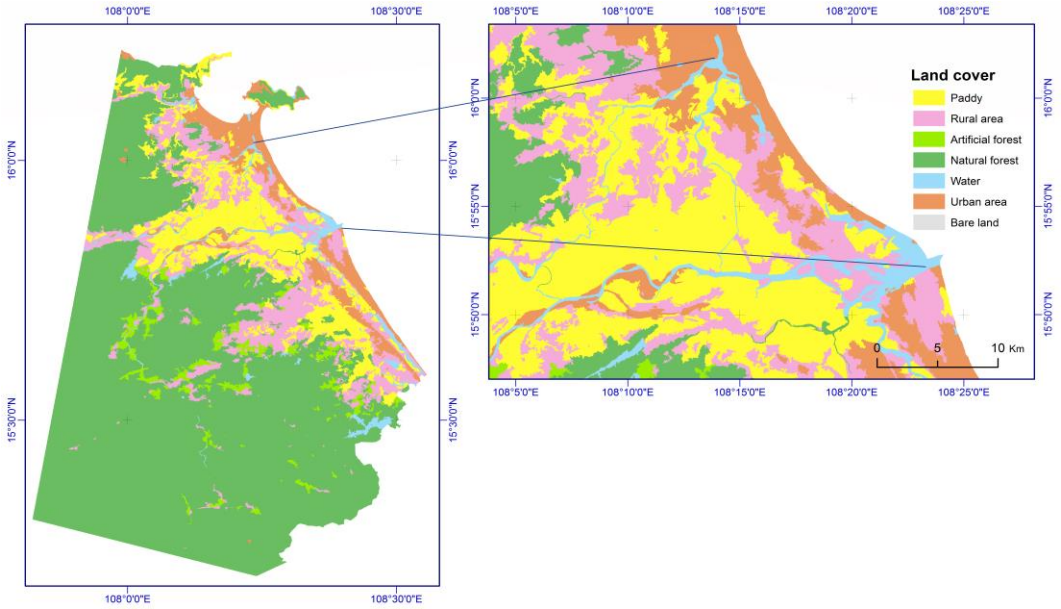


Figure 7. Object-based classification results of the study area on Google Earth Engine. Classification results for the whole study area (left image), classification results enlarged on Google Earth Engine for a part of the study area (right image)

## V. CONCLUSION

This study demonstrates an effective approach for building land-cover data with Landsat 5 remote-sensing imagery using the object-based RF classification method on the cloud computing platform. Compared to commercial software, the classification based on cloud computing with the Google Earth Engine produces rapid classification results; users can build algorithms and parameters suitable for different study areas. However, this method still depends on the speed of the internet.

The classification results showed seven categories of land cover including artificial forest, natural forest, paddy area, urban area, rural area, bare land, and water system with an overall accuracy of 73.9% and kappa of 0.70. This level of accuracy ensures the use of land cover data for subsequent steps in the analysis.

## ACKNOWLEDGMENT

This research was funded by Project "Researching on the application of remote sensing and GIS technology to determine the quantity, demand and allocation of surface water resources for the planning of inter-provincial river water resources, case study in Vu Gia - Thu Bon river basin", No. TNMT.2017.02.05 by the University of Natural Resources and Environment.

## REFERENCES

[1] POORTINGA, A., CLINTON, N., SAAH, D., CUTTER, P., CHISHTIE, F., MARKERT, K.N., ANDERSON, E.R., TROY, A., FENN, M., TRAN, L.H., BEAN, B., NGUYEN, Q., BHANDARI, B., JOHNSON, G., and TOWASHIRAPORN, P. (2018) An Operational Before-After-Control-Impact (BACI) Designed Platform for Vegetation Monitoring at Planetary Scale. *Remote Sensing*, 10 (5), 760.

[2] BLASCHKE, T. (2010) Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, pp. 2-16.

[3] CAMPOS-TABERNER, M., GARCÍA-HARO, F.J., CAMPS-VALLS, G., GRAU-MUEDRA, G., NUTINI, F., CREMA, A., and BOSCHETTI, M. (2016) Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring. *Remote Sensing of Environment*, 187, pp. 102-118.

[4] GARCÍA-HARO, F.J., CAMPOS-TABERNER, M., MUÑOZ-MARÍ, J., LAPARRA, V., CAMACHO, F., SÁNCHEZ-ZAPERO, J., and CAMPS-VALLS, G. (2018) Derivation of global vegetation biophysical parameters from EUMETSAT Polar System. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139, pp. 57-74.

[5] GISLASON, P.O., BENEDIKTSSON, J.A., and SVEINSSON, J.R. (2006) Random Forests for Land Cover Classification.

*Pattern Recognition Letters*, 27 (4), pp. 294-300.

[6] MATEO-GARCÍA, G., GÓMEZ-CHOVA, L., AMORÓS-LÓPEZ, J., MUÑOZ-MARÍ, J., and CAMPS-VALLS, G. (2018) Multitemporal Cloud Masking in the Google Earth Engine. *Remote Sensing*, 10 (7), 1079.

[7] MAHDIANPARI, M., SALEHI, B., MOHAMMADIMANESH, F., HOMAYOUNI, S., and 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.

[8] MUTANGA, O. and KUMAR, L. (2019) Google Earth Engine Applications. *Remote Sensing*, 11 (5), 591.

[9] ACHANTA, R. and S¨USSTRUNK, S. (2017) Superpixels and Polygons Using Simple Non-Iterative Clustering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, July 2017*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 4651- 4660.

[10] WASKE, B. and BRAUN, M. (2009) Classifier ensembles for land cover mapping using multitemporal SAR imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64 (5), pp. 450-457.

## 参考文:

[1] POORTINGA，A.，CLINTON，N.，SAAH，D.，CUTTER，P.，CHISHTIE，F.，MARKERT，K.N.，ANDERSON，E.R.，TROY，A.，FENN，M.，TRAN，L.H.，BEAN，B.，NGUYEN，Q.BHANDARI，B.，JOHNSON 和 G.TOWASHIRAPORN，P。（2018）一个可操作的控制后影响（工商银行）设计的平台，用于行星尺度的植被监测。遥感，10（5），760。

[2] BLASCHKE，T.（2010）基于对象的遥感影像分析。ISPRS 摄影测量与遥感杂志，65，第 2-16 页。

[3] CAMPOS-TABERNER，M.，加西亚·哈罗，F.J.，CAMPS-VALLS，G.，GRAU-MUEDRA，G.，NUTINI，F.，CREMA，A.，和 BOSCHETTI，M.（2016）多时相和多分辨率叶面积指数检索，用于本地水稻作物的可操作监测。环境遥感，187，第 102-118 页。

[4] GARCÍA-HARO，F.J.，CAMPOS-TABERNER，M.，MUÑOZ-MARÍ，J.，LAPARRA，V.，CAMACHO，F.，SÁNCHEZ-ZAPERO，J。和 CAMPS-VALLS，G。（2018）从欧盟气象卫星极地系统推导全球植被生物物理参数。ISPRS 摄影测量与遥感杂志，139，第 57-74 页。

[5] GISLASON，P.O.，BENEDKTSSON，J.A.，和 SVEINSSONSON，J.R.（2006）随机森林分类。模式识别快报，27（4），第 294-300 页。

[6] G. MATEO-GARCÍA，L。GÓMEZ-CHOVA，J.，AMORÓS-LÓPEZ，J.，MU.OZ-MARÍ，J。和 CAMPS-VALLS，G。（2018）谷歌地球引擎中的多时态云掩蔽。遥感，10（7），1079。

[7] MAHDIANPARI，M.，SALEHI，B.，MOHAMMADIMANESH，F.，HOMAYOUNI，S.，和 GILL，E.（2019）使用前哨 1 号和前哨 1 号在空间分辨率为 10 米的纽芬兰首次湿地清单地图。谷歌地球引擎云计算平台上的前哨 2 号数据。遥感，11（1），43。

[8] MUTANGA，O. 和 KUMAR，L.（2019）谷歌地球引擎应用程序。遥感，11（5），591。

[9] ACHANTA，R. 和 S¨USSTRUNK，S.（2017）使用简单的非迭代聚类的超像素和多边形。于：2017 年 7 月，夏威夷，檀香山，电气工程师学会计算机视觉与模式识别会议论文集。新泽西州皮斯卡塔维：电气与电子工程师协会，第 4651 至 4660 页。

[10] WASKE，B. 和 BRAUN，M.（2009）使用多时态 SAR 影像进行土地覆盖制图的分类器集合。ISPRS 摄影测量与遥感杂志，64（5），第 450-457 页。