*proceedings*

MDPI

*Proceedings*

# Comparative of Machine Learning Algorithms and Datasets to Classify Natural Coverage in the Cajas National Park (Ecuador) Based on GEOBIA Approach †

**Diego Pacheco Prado [1,2,*] and Luis Ángel Ruiz [1]**

[1] Geo-Environmental Cartography and Remote Sensing Group (CGAT), Universitat Politècnica de València, 46022 Valencia, Spain
[2] Universidad del Azuay, Cuenca, Ecuador
* Correspondence: diepacpr@upv.es
† Presented at the II Congress in Geomatics Engineering, Madrid, Spain, 26–27 June 2019.

Published: 16 July 2019

**Abstract:** GEOBIA is an alternative to create and update land cover maps. In this work we assessed the combination of geographic datasets of the Cajas National Park (Ecuador) to detect which is the appropriate dataset-algorithm combination for the classification tasks in the Ecuadorian Andean region. The datasets included high resolution data as photogrammetric orthomosaic, DEM and derivated slope. These data were compared with free Sentinel imagery to classify natural land covers. We evaluated two aspects of the classification problem: the appropriate algorithm and the dataset combination. We evaluated SMO, C4.5 and Random Forest algorithms for the selection of attributes and classification of objects. The best results of kappa in the comparison of algorithms of classification were obtained with SMO (0.8182) and Random Forest (0.8117). In the evaluation of datasets the kappa values of the photogrammetry orthomosaic and the combination of Sentinel 1 and 2 have similar values using the C4.5 algorithm.

**Keywords:** Cajas National Park; Geobia; classification; machine learning; Sentinel

## 1. Introduction

Information on land cover is crucial not only for mapping vegetation cover and land use changes, but also for the role they play in natural hazard processes e.g. land sliding, erosion, flash floods, etc. [1]. In tropical regions, such as Cajas National Park—CNP (located in the Ecuadorian Andes), the land cover data is not updated frequently because of the unavoidable cloud cover. Therefore, it is crucial to further explore classification methods and feature selection taking into account data that is not affected by cloud coverage.

Past works used a combination of terrain data and satellite images as Sentinel 1 (S1) and Sentinel 2 (S2), obtained from the European Spatial Agency (ESA), to resolve land cover classification problems in sites such as Wayang-Windu and Patuha [1], Wuhan, south-central China [2]. Other case is presented by Clerici et al. in the Colombian Andean [3]. The S2 with 13 spectral bands provide data from visible, Near-infrared (NIR) and Short-wave infrared (SWIR) regions [4] and the red edge bands represent a significant spectral enrichment with respect to other commonly used sensors in Land Use Land Cover (LULC) mapping exercises [3].

For the creation of land cover maps some works use the Geographic Object-Based Image Analysis (GEOBIA) devoted to partitioning remote sensing imagery into meaningful image-objects [5].

The aim of this research is to compare the performance of three algorithms of classification as C4.5 decision tree, Sequential Minimal Optimization (SMO) and Random Forest (RF) with different combination of datasets to determinate the appropriate combination dataset-algorithm to classify land cover in CNP using the GEOBIA approach.

## 2. Materials and Methods

CNP is located in the Ecuadorian inter-Andean region (Figure 1), being an important protected area for the country. It has approximate 28,585 ha and elevations ranging 3160–4450 m. The park contains 235 lakes and two main vegetation types: high-elevation Andean forest and paramo.

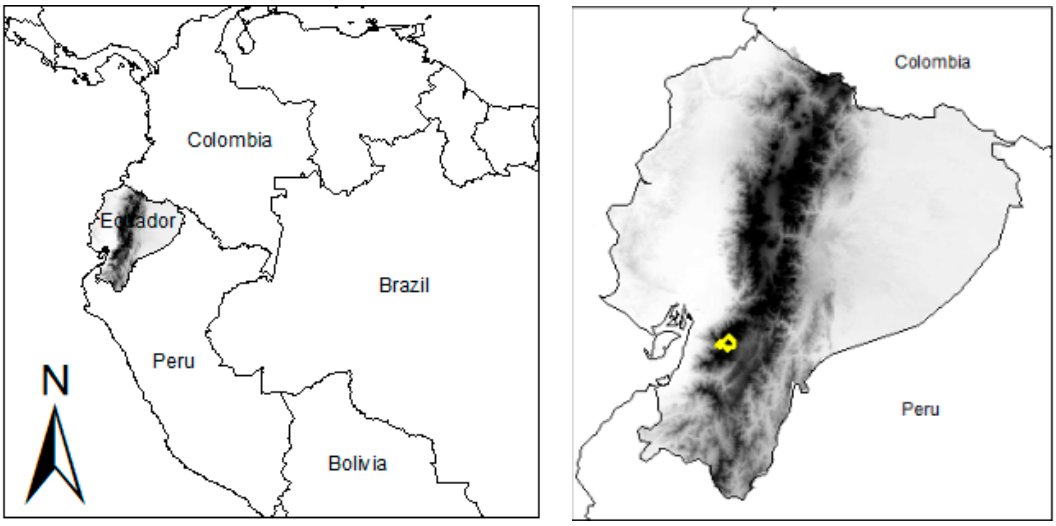


**Figure 1.** Location of study region. (**a**) Ecuador within South America. (**b**) Cajas National Park in Ecuador.

The objects for GEOBIA modeling were provided by the LULC map of Azuay province scale 1:5000 [6]. Using these objects we extracted the features from datasets as Digital Elevation Model-DEM and slope derivated from DEM, photogrammetric orthomosaic (PO), S1 and S2 satellite imagery of 12 April 2018 using Fetex 2 [7]. S1 data was used to generate surface roughness [3] and was downloaded with the Google Earth Engine (GEE) platform. The details of the datasets are available in https://bit.ly/2DyJG43. The complete workflow is detailed in Figure 2.

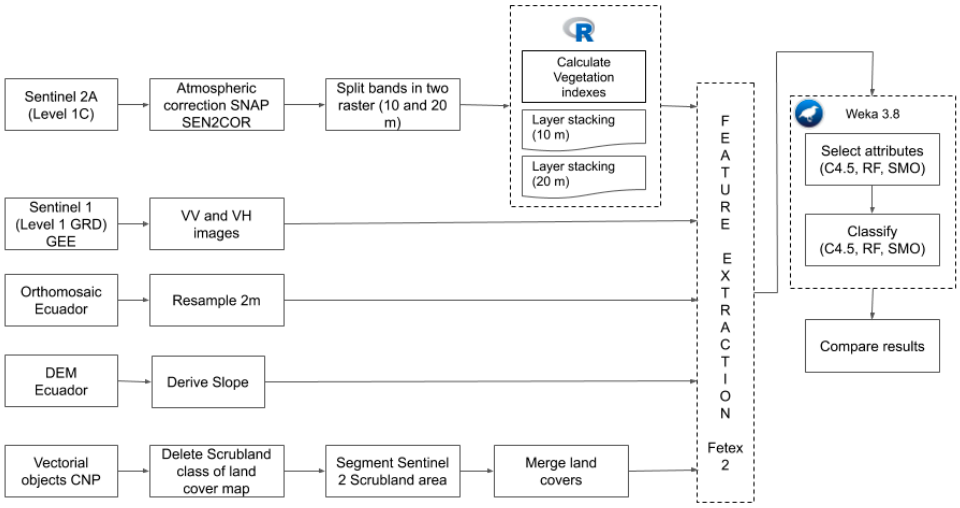


**Figure 2.** Workflow Characterization process.

For the selection of attributes and classifications of objects we evaluated three machine learning algorithms: C4.5 used to generate a decision trees [1], RF [8] that combines a lot of individual decision trees witch is tested with aleatory data of equal distribution, and the SMO (optimization algorithm used to train Support Vector Machine), that uses heuristics to partition the training problem into smaller problems that can be solved analytically.

We combined all attributes (unweighted) in the one file. The process of selection of attributes with the three algorithms reduce the number of attributes for the classification process. In all cases we set the attribute selection mode to "Use full training set" and the "GreedyStepwise" algorithm. Additionally, to determine the appropriate dataset combination we evaluated only the C4.5 algorithm with each dataset and latter combining them, for the ease of transforming the decision trees in to rules for the construction of maps.

## 3. Results

The 9371 objects of CNP were divided in different land covers such as rock outcrops (3095.47 ha), cushion or wetlands (1295.16 ha), native forest (250.53 ha), lakes or lagoons (1179.28 ha), scrubland (21,007.04 ha), shrub vegetation—chaparro (1459.32 ha) and bare soils (40.53 ha). Classes with lower number of objects and objects not related with natural land covers were discarded. After the feature extraction we obtained 8848 objects (~28,576 ha) with the PO, 2413 objects (~27,137 ha) with S2 bands of 10 m and 1030 (~24,990 ha) with S2 20 m bands. The objects without attributes are due to digitalization (very small objects) and the raster resolution.

The attributes merging in one single data file generated 363 features of 1030 objects to analyze with machine learning techniques. With the process of selection of attributes, we reduced the number of features and obtained a list of attributes for each algorithm. In the case of C4.5 the selected features were the mean of bands 10, 12 and 14 of S2 (20 m), the mean of bands 2 and 5 of S2 (10 m), the uniformity and entropy of the DEM, and the uniformity of slope and majority of PO. With RF features selected are the mean of bands 9 and 13 of S2 (20 m), the mean of band 5 of S2 (10 m), DEM contrast, the Gray-Level Co-Occurrence Matrix (GLCM) inverse difference moment of slope, the majority of bands 1 and 4, the standard deviation of edgeness factor and the ratio between semi variance values at second and first lag of PO. For SMO and other algorithms the attribute selection results are detailed in the link https://bit.ly/2vnxEpK. The C4.5 was the algorithm that less considered attributes of the PO dataset, thus it was used to evaluate the datasets individually and combined.

The larger kappa values for the classification were obtained for SMO and RF (Table 1). We must to highlight that SMO is product of S2, S1 and PO datasets, considering that the PO dataset cannot be generated very frequently because of cost factor. Additionally, Table 1 presents the evaluation for each dataset and test two combinations with C4.5. The best kappa value is obtained with only PO dataset and with the combination of freely imagery S1 and S2 with 10 m of pixel size. The detailed table with the results of classification is located in https://bit.ly/2J0fRgj.

**Table 1.** True Positive percentage and kappa values of classification process with C4.5, RF and SMO algorithms.

| | **Algorithms Evaluated Folds: 10** | | | **Datasets Individual and Combining Evaluated with C4.5 Folds: 4** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **C4.5** | **SMO** | **RF** | **S2 (10 m)** | **S2 (20 m)** | **S1 (10m )** | **DEM** | **SLOPE** | **PO** | **S2_10M +S2_20M** | **S2_10m +S1_10m** |
| Correctly Classified Instances (%) | 86.68 | 89.5 | 89.4 | 83.77 | 81.63 | 69.97 | 71.5 | 71.43 | 85.4 | 83.38 | 85.52 |
| Kappa statistic | 0.77 | 0.82 | 0.81 | 0.72 | 0.68 | 0.4 | 0.44 | 0.44 | 0.75 | 0.71 | 0.75 |

## 4. Conclusions

We evaluated a combination of different datasets to determinate the characteristics that were selected when the attributes were unweighted. With the machine learning algorithms (C4.5, SMO and RF) we opted for the use of S2 imagery followed by DEM and slope features. From the evaluation

of the datasets the combination of S1 and S2 imagery with pixel size of 10 meters present similar values of kappa comparing with PO. Thus, they must be considered for LULC updates with free imagery.

Regarding the land cover classes, the only algorithm that could classify native forests was Random Forest. In other algorithms the confusion matrix shows that the Native forest were confused with shrub vegetation—chaparro. The different confusion matrix that show each classes is located in http://bit.ly/2VxYBWi.

**Funding:** This research was funded by UNIVERSIDAD DEL AZUAY in the context of investigation project 2018-77 denominated Multispectral analysis of natural coverage of the Cajas National Park through drones and satellite images.

**Acknowledgments:** This research was supported by the Instituto de Estudios de Régimen Seccional del Ecuador (IERSE) of Universidad del Azuay (Ecuador) and the Geo-Environmental Cartography and Remote Sensing Group (Cartografia GeoAmbiental y Teledeteccion—CGAT) of the Polytechnic University of Valencia (Spain).

## References

1. Shrestha, D.P.; Saepuloh, A.; Van der Meer, F. Land cover classification in the tropics, solving the problem of cloud covered areas using topographic parameters. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *77*, 84–93.
2. Liu, Y.; Gong, W.; Hu, X.; Gong, J. Forest type identification with random forest using Sentinel-1A, Sentinel-2A, multi-temporal Landsat-8 and DEM data. *Remote Sens.* **2018**, *10*, 1–25.
3. Clerici, N.; Valbuena Calderón, C.A.; Posada, J.M. Fusion of Sentinel-1a and Sentinel-2A data for land cover mapping: A case study in the lower Magdalena region, Colombia. *J. Maps* **2017**, *13*, 718–726.
4. Borràs, J.; Delegido, J.; Pezzola, A.; Pereira, M.; Morassi, G.; Camps-Valls, G. Clasificación de usos del suelo a partir de imágenes sentinel-2. *Rev. Teledetec.* **2017**, *2017*, 55–66.
5. Hay, G.J.; Castilla, G. Object-based image analysis: Strengths, weaknesses, opportunities and threats (SWOT). In Proceedings of the The International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences, Salzburg University, Salzburg, Austria, 4–5 July 2006; pp. 4–5.
6. Tenesaca, C.; Quindi, T.; Delgado, G.; Toledo, E.; Delgado, O. Generación del mapa de cobertura y uso del suelo de la provincia del Azuay. *Universidad Verdad* **2017**, 23–37.
7. Ruiz, L.A.; Recio, J.A.; Fernández-Sarría, A.; Hermosilla, T. A feature extraction software tool for agricultural object-based image analysis. *Comput. Electron. Agric.* **2011**, *76*, 284–296.
8. Navarro, J.A.; Algeet, N.; Fernández-Landa, A.; Esteban, J.; Rodríguez-Noriega, P.; Guillén-Climent, M.L. Integration of UAV, Sentinel-1, and Sentinel-2 Data for Mangrove Plantation Aboveground Biomass Monitoring in Senegal. *Remote Sens.* **2019**, *11*, 77.