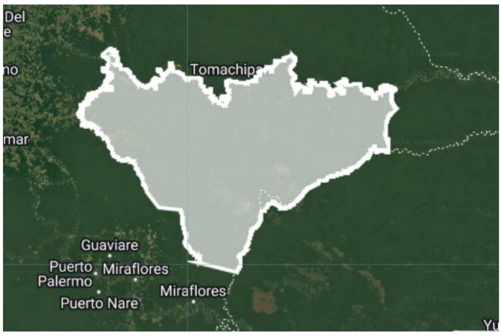
Assessing the land use change within Nukak protected area

Abstract



Nukak is the ID19992 colombian region of the WDPA database of protected areas[1]. The aim of this work is to assess the land use change inside that protected area comparing the years 2016 and 2020.
The idea is to compare satellite images of those years with the following properties:

1. I will look for low cloud-covered images, which is hard in this region due to the local climate and hydrology;
2. I will use NDVI to mask which regions have *loss* or *gain* in “photosynthetic biomass”, not relating it directly to a specific land use change;
3. I will try to compare images from the same period of the year not to misunderstand a land use change instead of a natural change of NDVI due to the annual phenology of the plant;
4. I will classify land use in both gain and loss regions using a supervised classification over PCA bands made from optical bands. The information classes are **forest**, **agriculture** and **deforested**;
5. I will then classify land use change separating *likely manned* (**forest to deforested**, **forest to agriculture**, **agriculture to deforested**), *probably unmanned reasons* (**agriculture to forest** and **deforested to forest**) and *probably manned reasons* (**deforested to agriculture**). I will come back later on this point;
6. I will then make an accuracy assessment over the point 4. I could not make accuracy assessment on point 5 because I did not find any ground truth data about land use change between those years.

Images selection

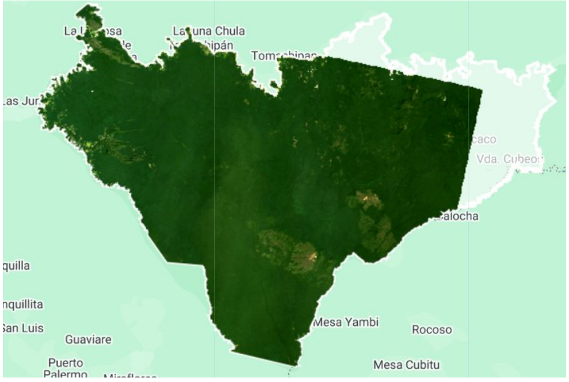
Thanks to this, the required properties of the image are:

- *spectral resolution*: multispectral optical bands containing at least RGB and NIR bands;
- *time resolution*: there is no particular bound over time resolution but the greater it is the better I could find a day where there is a desired cloud cover;
- *space resolution*: I am trying to identify a manned land use change that in this region is usually done for big agriculture or cattle purposes so I ask for a space resolution under 50m which means a pixel over a field will hardly contain the forest around. In this resolution I also consider the extent of the *tiled image*. Due to the difficulty in acquiring not cloudy images I would prefer a single not cloudy image rather than more images that will be time displayed by cloud cover and can lead to misalignments in NDVI in different regions;
- *radiometric resolution*: I am trying to make a coarse classification of forest/agriculture/bare fields so I have not particular bounds on the radiometric resolution.

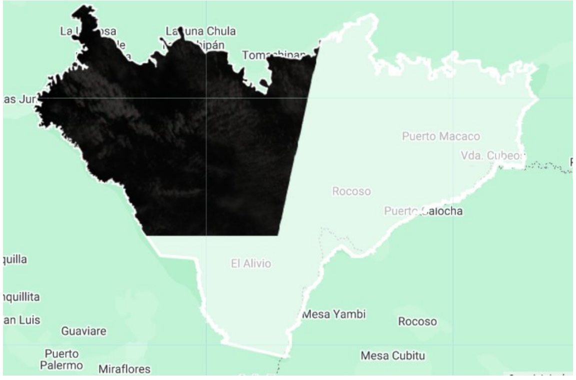
[1] https://www.protectedplanet.net/19992

The candidates are *Sentinel-2-MSI* (S2) and *Landsat-8-OLI/TIRS* (LC8) and due to tiles of S2 images that taken as single have a partial cover of the area, I chose the LC8 satellite. I add that Sentinel-2B was not already launched at that epoch so the time resolution of the region would be about 10 days where LC8 satellite has a revisit time of 8 days.

LC8 tile



S2 tile



The chosen image collection is *USGS Landsat 8 Level 2, Collection 2, Tier 1*, for the following reasons:

- *Landsat 8*: it is the optical Landsat satellite from 2013;
- *Level 2A*: to compute NDVI I am interested in surface reflectance so, not having any skill in deriving atmospheric corrections, I have chosen to take the geometric and atmospheric corrected images;
- *Collection 2*: since 2021 USGS has been migrating into Collection 2. Collection 1 was the first major archive elaboration of the LC8 images but it is now deprecated. The algorithms and corrections in Collection 2 are more reliable[2];
- *Tier 1*: are the data that meets the NASA quality requirements about atmospheric and geometric corrections. Taking the Tier 2 means that I have more images but the atmospheric and geometric corrections are not at the highest standard.

Laguna Chula
Puerto Macaco
Vda. Cubeo
Rocoso
El Alivio
Mesa Yambi
Rocoso
Guaviare
Puerto Palermo
Mesa Cubitu

The chosen period of the year is climatic winter (from January to March) that should be the period in the local climate that has lowest cloud coverage[3]. In particular the least cloud period in that years (2016 and 2020) was found to be January for both years. The two images of comparison are found from GEE to be:

- 2016-01-25 15:01:15
- 2020-01-04 15:01:23

The maximum cloud coverage over the tile (not over the Nukak region) was chosen to be 10% and I took the least covered images. I then applied scale factors over the LC8 images[4], because for archive purposes there are scales and offsets on the images bands. Then I made a *cloud* (and *shadow*) *mask* over the image, lightblue regions in the image, even if the cloud coverage filter is very strict. This is because, as said before, the cloud cover percentage property of the

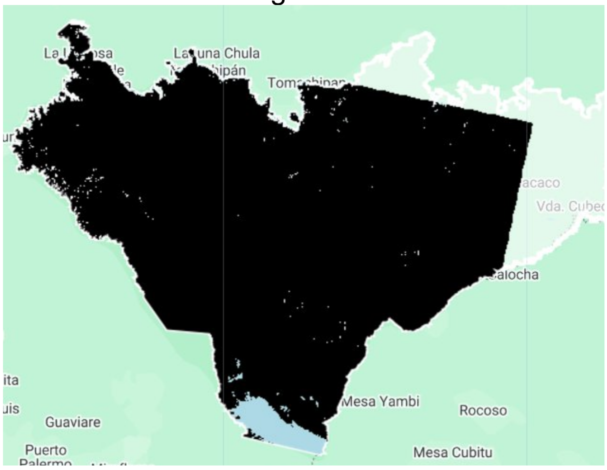
[2] https://www.usgs.gov/landsat-missions/landsat-collections
[3] https://weatherspark.com/y/25940/Average-Weather-in-Miraflores-Colombia-Year-Round
[4] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1_L2

image is over the whole tile but my region can be affected by a different cloud cover depending on its distance from clouds. This pixel mask is then applied to the images, keeping in my images only high quality pixels. The masked region is considerable, but by photo-interpretation of the image in different periods, there shouldn't be a great land use change so for a first study of land use change this is not a problem.

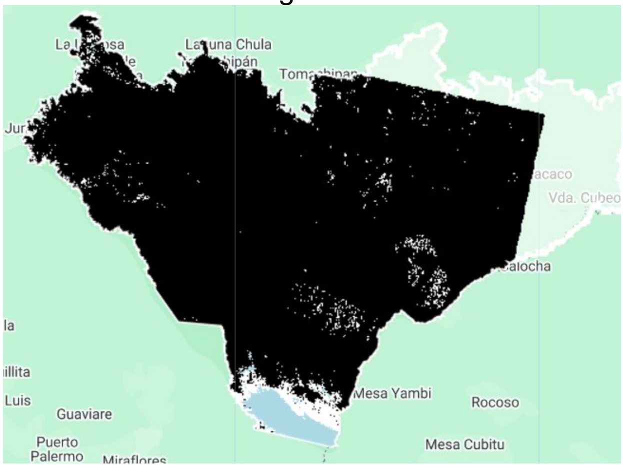
Loss and gain regions

Then I computed *Normalized Difference Vegetation Index*[5] (NDVI) for both images and the difference between those indices. Most of the image has changed little and I want to highlight great changes in NDVI, that will be (indirectly) related to land use change. For this reason I computed the mean and the standard deviation of the NDVI difference over the whole Nukak area. *Loss regions* will be identified as the regions where the NDVI difference is at least 2σ below the mean NDVI difference while *gain regions* are the regions where the NDVI difference is at least 2σ over the mean NDVI difference.
This could be a questionable if the land use change in that region were a very diffuse phenomenon, but considering that this is a protected area this should not be the case.

Loss regions mask



Gain regions mask



Classification of loss and gain regions

Now I want to classify loss and regions regions over both images (for a total of 4 combinations of images). To do this, I want to use the widest spectral information of the satellite image so I use the bands SR_B2 (*blue*), SR_B3 (*green*), SR_B4 (*red*), SR_B5 (*NIR*), SR_B6 (*SWIR1*) and SR_B7 (*SWIR2*). I did not use the band SR_B1 (*coastal blue*) because I am not sure it is reliable in vegetation classification. I could have used all these 6 bands for the classification but for a coarse classification it is usually convenient to choose just the lowest number of bands needed to separate the spectral classes, in order to reduce omission and commission errors, thus improving in accuracy and extrapolation. To do this, I used the *Principal Component Analysis* (PCA), that is an algebraic way to assess variance distributions among bands through creating linear combinations of bands with a defined variance[6]. I did PCA over both images and identified the new "synthetic" bands distribution of variance. Looking at the eigenvalues I could assess that the greatest amount of variance was in three PCA bands (that I called pc1, pc2, pc3), while the others are only a negligible correction to them.

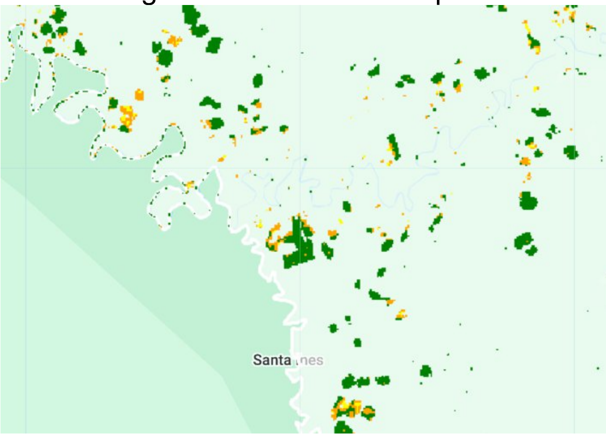
[5] https://www.indexdatabase.de/db/i-single.php?id=401
[6] Strictly speaking PCA is used as a feature reduction technique to keep only informative (combination of) bands. Instead of making a supervised classification over PCA bands, a *Canonical Discriminant Analysis* (CDA) could have been carried out using all the original bands and using a training set to select the bands which make better clusters (greater between-class distance, smaller within-class distance).

Supervised classification requires a training set for the algorithm to be applied. The best choice would be ground truth, then another (satellite) RGB image of the region, for example with higher resolution. I did not have neither of them but using photo-interpretation of the same image in RGB, I could distinguish **forest**, **agriculture** and **deforested** regions. Using a widely accepted rule of thumb, I chose a number of training pixels between:

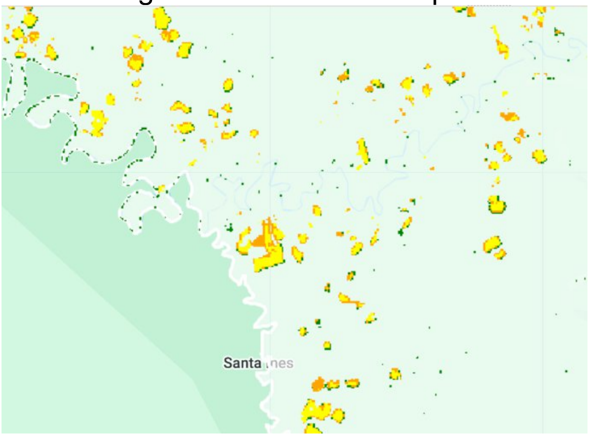
10 * Number of class * Number of bands and 100 * Number of class * Number of bands

I had 3 classes and 6 true bands, but I am actually classifying using three PCA bands so the right number is among 90 and 900. I have chosen 150, that is 50 pixels per class. These pixels are chosen mostly on the left hand corner of the image where deforestation is visually most prominent. The classifier used is smileCART[7] and the result of land cover classification is[8]:

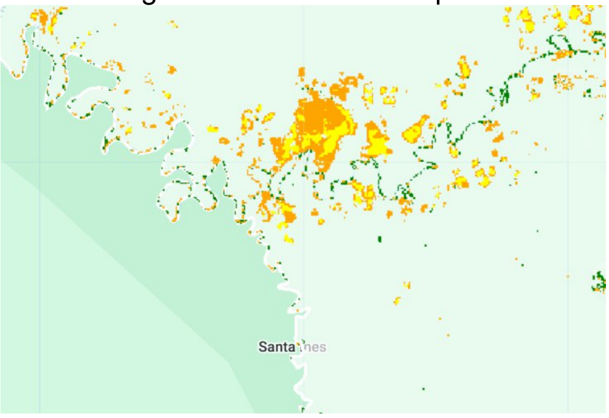
Loss regions classification map in 2016



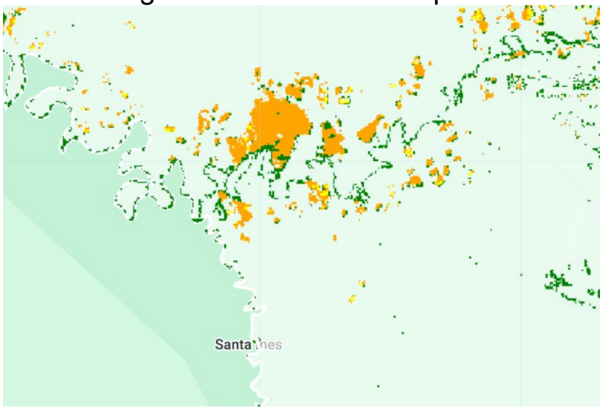
Loss regions classification map in 2020



Gain regions classification map in 2016



Gain regions classification map in 2020



There are some remarks to be made at this point:

- there are regions where the classification has not changed: I am not interested in these regions and in the following I will exclude them;
- in the "gain" row I see that the NDVI is increasing, that is a 2016 region classified as bare has become in 2020 either agriculture or forest or a region 2016 classified as agriculture has become in 2020 forest. With similar but opposite consideration I see what happens in the "loss" row. This is a self-consistence check of the algorithm made;
- see that among "gain" regions there are regions near the river (a curvy line in the "gain" row). This can be correct since the growing rate should be greater near a water source;

[7] "Classification and Regression Trees,". L. Breiman, J. Friedman, R. Olshen, C. Stone. Chapman and Hall, 1984.

[8] I zoomed around Santa Ines region to simplify reader view. The colours of the map are **forest**, **agriculture** and **deforested**.

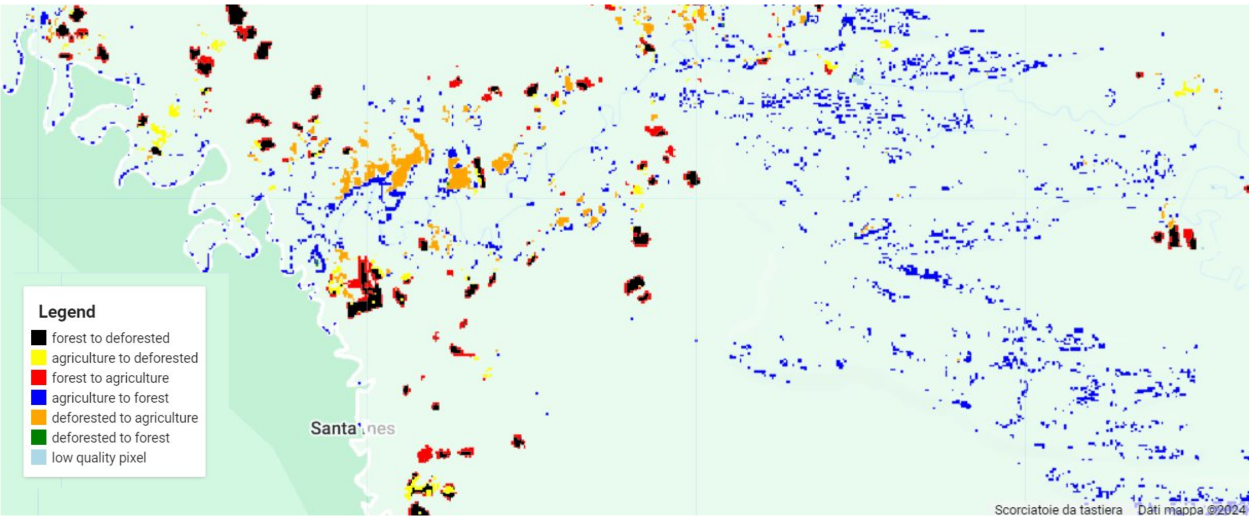
- see that among “loss” regions there are square-like contiguous regions that is a clear sign of manned origin;
- see that there is a vast area of bare/deforested region that from 2016 to 2020 has become agriculture. This can be a newly deforested area in 2016 that from that on has become an agriculture field.

Land use change classification

Finally, I classified the land use change this way:

- loss regions can be further distinguished into **forest to deforested**, **forest to agriculture**, **agriculture to deforested**. As an example a pixel will be classified forest to deforested if it was classified as forest in 2016 and as deforested/bare in 2020;
- gain regions can be further distinguished into **agriculture to forest**, **deforested to forest** and **deforested to agriculture**.

The result of this classification is shown for the Santa Ines region:



Accuracy assessment

For the accuracy assessment, I computed the confusion matrix using an independent set of testing pixels, in the same number of the training pixels. I ended with an overall accuracy of 98.7%, consumers’ accuracies of 100% for forest, 98.04% for bare/deforested, 98% for agriculture and a kappa coefficient of 98%.

Final remarks

- **agriculture to forest**: in this class I can misidentify fruit trees, so the land use has not really changed. This also applied to **forest to deforested** and **forest to agriculture**. A better investigation will combine also information about NDVI as a synthetic layer to separate the forest class between true forest and fruit trees, because simply speaking NDVI of true forest should be higher than fruit trees NDVI. In this class can also contribute an expanding forest border, where little forest trees grow. After dealing with those problems, **agriculture to forest** should be the information class of *abandoned croplands*;
- **deforested to forest**: as the **agriculture to forest**, this class can be misidentified with fruit trees, that should go to **deforested to agriculture** class. One way to exclude it that is valid for **deforested to forest** but not for **agriculture to forest** is to filter by the

growing year of fruit trees. If 4 years are few to a fruit tree to grow from the seed, this class should not be affected by that problem. After dealing with those problems, **deforested to forest** should be the information class of *abandoned croplands*;

- **deforested to agriculture**: one problem of this class is on the whole deforested classification where I can misidentify a deforested region that is instead an agriculture region but either with small crop plants or fallow fields or freshly mowed croplands. This means that I have to separate the deforested class into a soil class and a small crop class so that the land use of some regions has not really changed. The same applies to agriculture to deforested. After dealing with those problems, **deforested to agriculture** should be the information class of *agriculture use of newly deforested region*;
- **forest to deforested**, **forest to agriculture**, agriculture to deforested are likely to be manned because natural reduction of tree cover can be due to:
    - spontaneous (for example due to a lightning) fires. These are unlikely to happen in equatorial forest due to high amount of both soil and air moisture and trees are generically healthy in this region. Furthermore, the evidence of no fires is the shape of the deforested regions;
    - storms have usually a greater scale and an homogeneous pattern and this is not the case for loss regions;
    - illnesses or pests are usually contiguous. They appear to be local only when some geo-climate conditions are local but in this equatorial forest the conditions are very homogeneous.

  The above considerations about fruit trees and fallow fields applied also here. For these reasons, after dealing with them, I am confident to say that:

    - **forest to deforested** will be *newly deforested region* where newly means closer to 2020 than 2016;
    - **forest to agriculture** will be *old deforested region* where old means closer to 2016 than 2020;
    - agriculture to deforested will be *region in a period between two different crop species*.
- I assumed in the study that the land use change is a restricted phenomenon. This can be seen to be true for this protected area but cannot always be true for all others so extrapolation of this algorithm should be made very carefully;
- I used a single Landsat 8 image even if it didn’t cover all the Nukak protected area. The reason is that as explained above I wanted to compare images possibly in the same period of the year. Taking an image of the excluded region but with low cloud coverage could mean to choose an image of a different period of the year, invalidating the considerations I made over those images;
- low quality pixels, for example cloud pixels, have not been studied. This could be a problem if a complete report of the protected area is needed. A possible solution is to analyse all the images of this period to see if that cloud masked region are visible in other images, that may have many clouds but in other parts of the image;
- NDVI may not be the best index for land use change, because it depends also on soil variety and soil moisture when soil is not covered by canopy. Other indices (SAVI, MSAVI, EVI, ...) can be used instead. An NBR index can be used for assessing deforestation through fire;
- rivers: in the information classes above I did not mention any water class. This means that pixels of the rivers are in other classes.