Final Report

Analysis of spatio-temporal data - Deforestation in Brazil 2013 - 2020

Submitted by:
Lukas Bäcker
Ilka Pleiser

Lecturer:
Prof. Edzer Pebesma

Institute for Geoinformatics

March 24, 2021

Contents

List of Figures

List of Tables

1 Motivation

The Amazon rainforest in South America comprises more than 50 % of the remaining tropical forest in the whole world. Around 60 % of this forest area can be located in Brazil. It has the greatest biodiversity of all tropical forests in the world. The Amazon rainforest is home to a good 2.5 million insect species, tens of thousands of plant species and 2,000 species of mammals and birds. Among them are animal and plant species that are native only to the Amazon region, such as the jaguar, the sloth or the black caiman (*Amazon rainforest*, 2021).
In addition, the Amazon region offers excellent conditions for growing various cereals and fruits. For example, soybeans, avocados, coffee and cocoa can be grown here under very good climatic conditions. There are also various precious metals such as gold in the soil of the Amazon rainforest (*Zerstörung im Schatten der Corona-Krise*, 2020). Cultivation or mining require large, open areas, for which the Amazon rainforest has been systematically cleared and branded in recent decades. Even for the simple trade of tropical timber huge areas are cleared in the Amazon. This is often done illegally and is not always directly and fully visible from the ground. Satellite images provide information about the size of the cleared areas. For example, the Brazilian environmental agency INPE was able to determine that 563 square kilometres of Amazon rainforest were destroyed by raw logging in November 2019 alone (*Regenwald-Abholzung in Brasilien hat sich verdoppelt*, 2019). Also the Brazilian Institute of Environment and Renewable Natural Resources (IBAMA) uses the analysis of satellite images to fight against illegal deforestation in Brasil sending out armed special forces. This project aims at the understanding of analysing spatio-temporal data in the context of deforestation in Brasil's Amazon rainforest from the years 2013 to 2019. With help of the NDVI value (Normalized Density Vegetation Index) we analyse a sample area and validate its results.

2 Research Question

Our research questions are:

1. How meaningful is the identification of deforestation areas based on the NDVI to determine whether forest recession is of natural or human-influenced origin?

2. Does it make a significant difference using annual or semi-annual data for the analysis of question 1?

3 Methodology

This section will describe the data that was used and the workflow we constructed in this project.

3.1 Data

The data basis of this project were Landsat-8 images from 2013 to 2020. Spatially we used four tiles and temporally between 109 and 115 time steps per tile. The test area is located in the Amazon region of Brazil and has an area of 1,200,000 ha. It is an area known for deforestation. The data was given and not chosen by us except the Landsat image for 2020 which we downloaded at `https://earthexplorer.usgs.gov/`.

3.2 Image Preparation

The first step of the image preparation was to process the downloaded Landsat-8 satellite data for subsequent calculations including the tailoring of each individual image to the test area. Furthermore, the NDVI had to be calculated so that vegetation can be easily distinguished from non-vegetation. To do so, Landsat-8 band 4, which displays the red color and the 5th band including the near infrared light were needed resulting in the formula for the NDVI as following:

$$(\text{Band 5} - \text{Band 4}) \;/\; (\text{Band 5} + \text{Band 4})$$

(*Landsat Normalized Difference Vegetation Index*, n.d.)

The image data has already been made available to us in cropped form except the images from 2020 which we downloaded from the Earth Explorer. Further steps, which belong to the image preparation stage, were on the one hand the cleaning of clouds and on the other hand the calculation of the NDVI. The code for these steps was also made available to us for the images from 2013 to 2019. We calculated an average image per year or half-year, which contains no clouds, is cropped, has a pixel size of 25 x 25 metres and shows the NDVI value. The downloaded images from 2020 contained some clouds that we masked the layer to ignore clouds and cloud shadows so that the classification is not distorted by these. We used the workflow of page `https://www.earthdatascience.org/courses/earth-analytics/multispectral-remote-sensing-modis/intro-spectral-data-r/` as basic concept using Band 1 of the Landsat-8 data. Band 1 senses deep blues and violets with whom one can detect water in the air and therefore clouds and other aerosols. After a few intermediate steps, the layer of the cloud mask could be displayed (figure 1). The final result of this step, the RGB image without clouds, can be seen in figure 2.

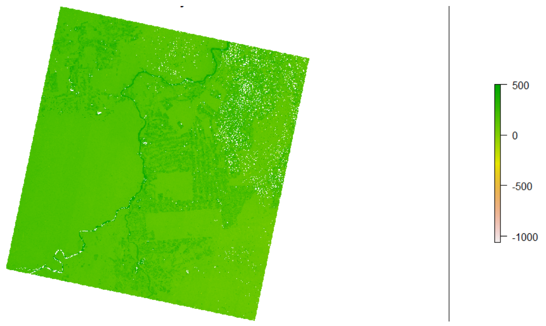

Figure 1: Cloud mask image 2020. Green area: no clouds, white spots: clouds (Source: Own figure)

Figure 2: Completed, cloudless RGB-image (Source: Own figure)

3.3 Annual vs. Semi-annual

Since we worked with both annual and semi-annual image data, two stacks were needed that are set up with the package "gdalcubes" (*gdalcubes: Earth Observation Data Cubes from Satellite Image Collections*, n.d.). The package brings a collection of functions for dealing with multispectral, multitemporal raster data cubes and made it easy to process the satellite images to data with predefined time dimension and spatial extension. The images of both cases have a temporal extension from the years 2013 to 2019. Depending on the size of time step, one or two images are processed per year displaying the NDVI of each cell of the chosen research area.

3.4 Analysis

3.4.1 Raster Calculation

Since not all pixels that belong to forest according to the calculation of the NDVI, but may belong to cultivations, parks or gardens, for example, it makes sense to exclude pixels that do not belong to forest before calculating the forest return. We have set an NDVI threshold of 0.8 here, as this is defined as the threshold value for dense vegetation (Londhe, 2019).

Already in this step we could get a rough picture of the deforestation development. For this purpose, we have had the percentage and absolute shares of tropical forest pixels output as a histogram. The results are shown in the following figures (figure 3 and figure 4).

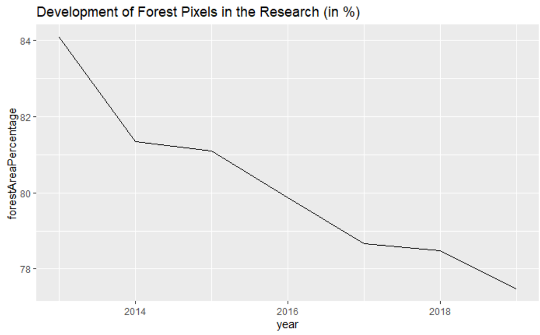

Figure 3: Development of forest pixels share from 2013 - 2019 in percent (Source: Own figure)

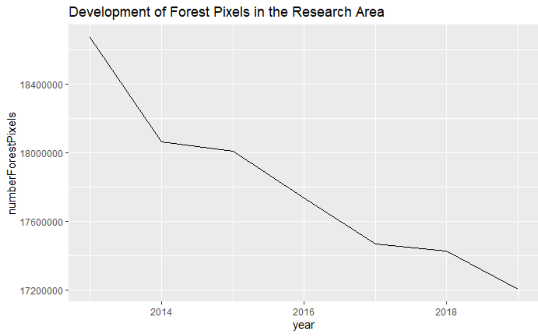

Figure 4: Development of forest pixels share from 2013 - 2019 in ha (Source: Own figure)

To calculate the deforested areas, we have inverted the visualisation so that the pixels that do not belong to forest are highlighted in colour. We have carried out this visualisation for all years and half-years (see figure 5).

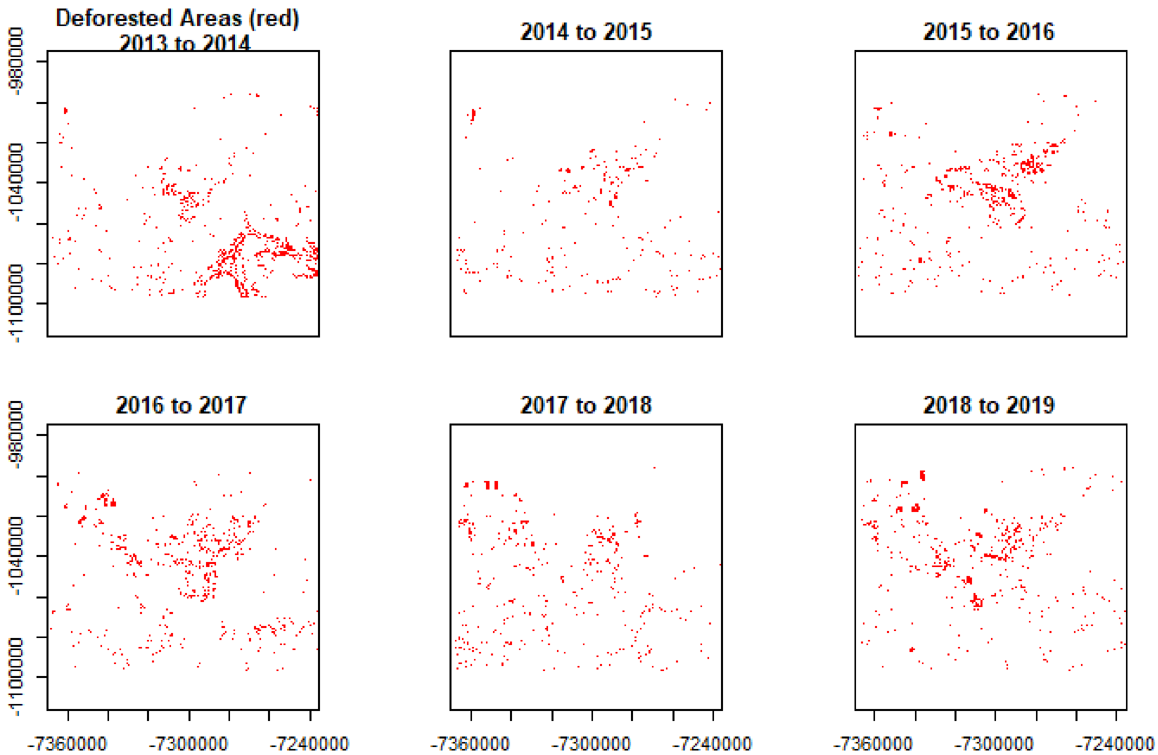

Figure 5: Deforested areas from 2013 - 2019 annually (Source: Own figure)

We now knew which pixels were deforested in which year. Now we wanted to determine the sum of all deforested pixels in this period. We therefore added all layers together resulting in 6.

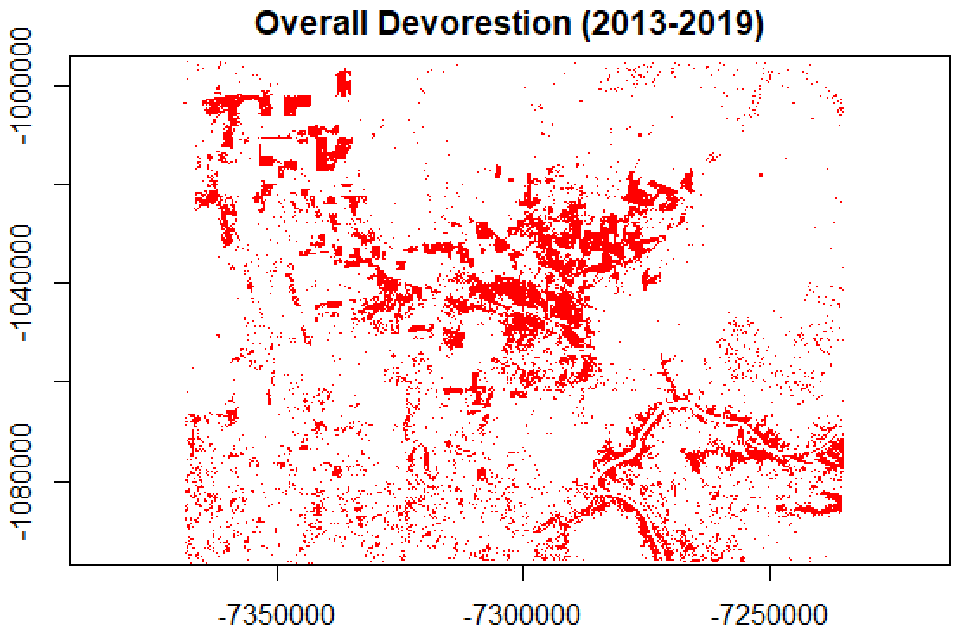

Figure 6: Deforested areas from 2013 - 2019 annually (Source: Own figure)

3.4.2 Natural Dieback Versus Man-Made Deforestation

Not all forest decline is due to human intervention. Natural forest decline is also favoured by various factors. In order to distinguish human-cleared forest areas from naturally deforested areas, we have applied the focal function two times. This calculates "focal ("moving window") values for the neighbourhood of focal cells using a matrix of weights, perhaps in combination with a function" (*focal: Focal values*, n.d.).

The first focal function is applied to a 3 x 3 matrix. It first checks whether the cell under consideration has itself been declared as "deforested" or "not deforested". If the cell has been identified as "deforested", the function checks whether more than 3 neighbouring cells (from the Moore neighbourhood) are "deforested". If this is the case, the cell under consideration is still considered "deforested". Otherwise it is overwritten as "not deforested". This removes solitary pixels (pixels without neighbours). The second focal function is applied to a 5 x 5 matrix. It also first checks whether the cell under consideration has itself been declared as "deforested" or "not deforested". If the cell has been identified as "deforested", the function checks whether more than 18 neighbouring cells (also from the Moore neighbourhood) are "deforested". If this is the case, the cell under consideration is still considered "deforested". Otherwise it is overwritten as "not deforested". This gives us a larger filtering. The reason for this step is that man-made (systematic) raw deforestation is often applied over a large area. The result image of these two steps is shown in figure 7.

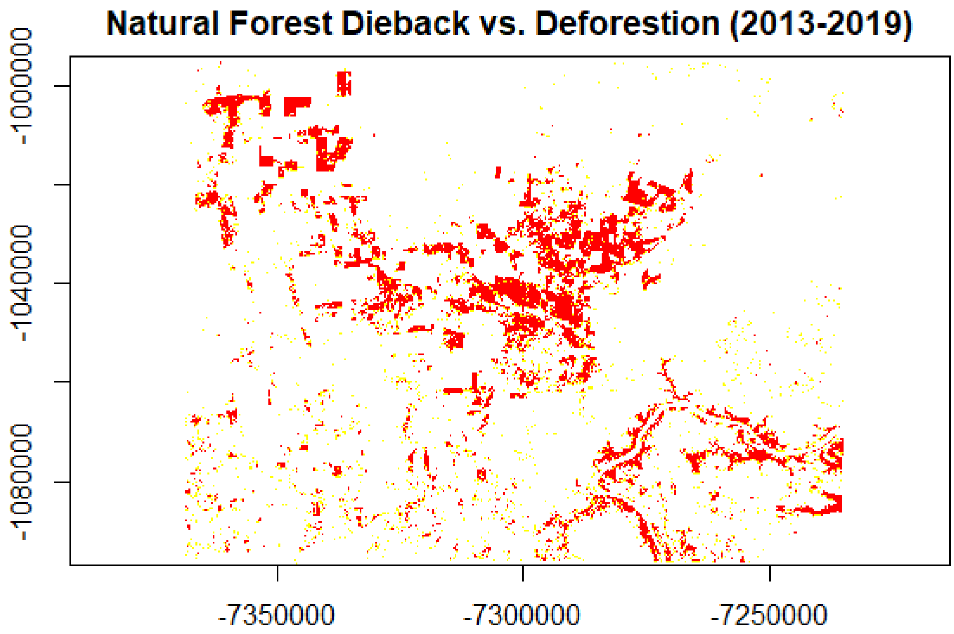

Figure 7: Natural Dieback vs. Deforestation 2013 - 2019. Red pixels show the deforestation, yellow pixels show the natural dieback (Source: Own figure)

Looking at this calculation more closely, with a smaller time step (here 2019 to 2020) and on a smaller area, the accuracy can be better assessed.

Figure 8: Excerpt aerial view 2019 (Source: Own figure)

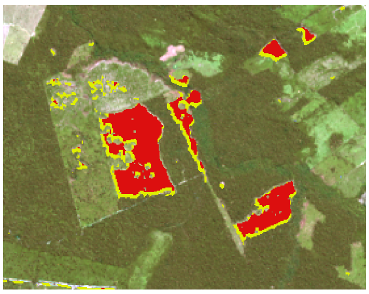

Figure 9: Added deforested areas after one year (Source: Own figure)

Figure 10: Excerpt aerial view 2020 (Source: Own figure)

3.4.3 Supervised Classification

After deciding to validate our outcomes with another classification we choose a random-forest supervised classification. For this we first collected polygons as training-data using QGIS and giving each polygon a class of those three: "StayDeforest", "StayForest" and

”NewDeforest“, which means that we wanted to differ areas that were already deforested in 2019, areas that stay forest in 2020 and those which were deforested from 2019 to 2020. As base for data-collection we used a 2019 and the 2020 aerial photograph. The test data set contained a total of 21 entries. With these data we then ran the random-forest classification on layer-stacks of NDVI layers. For the annual data we set up a stack with the calculated NDVI of 2019 and 2020 and for the semi-annual data we used the NDVI raster of the first half-year of 2019 due to the fact, that the second half-year of 2019 was strongly cloud-covered. For both time intervals we had then a classification based on only the NDVI from a previous image and the current image to check for changes.

3.5 Validation

We checked the results of the classifications of the annual and semi-annual data with cross-validations. The results of the cross-validations are shown in the following.

Annual

Table 1: Result of the cross-validation annual

	newDeforest	stayDeforest	stayForest	Sum
newDeforest	**735**	3	0	738
stayDeforest	1	**1613**	199	1813
newForest	2	232	**5389**	5623
Sum	738	1848	5588	**8174**

We calculate the $\text{accuracy}_{\text{annual}}$:

$\text{accuracy}_{\text{annual}} = (735 + 1613 + 5389) \div 8174$

This results in an accuracy of about 0.95 or 95%.
We have chosen the Kappa index as the validation measure for the classification. The kappa index can be calculated with the following formula:

$$K = \frac{p_0 - p_{\text{e}}}{1 - p_{\text{e}}}$$

Where

$$p_0 = \frac{\sum_{i=1}^{k} n_{\text{ii}}}{N}$$

and

$$p_{\text{e}} = \frac{\sum_{i=1}^{k} n_{\text{i}} \cdot n_{\text{i}}}{N^2}$$

It therefore follows:

$$K_{\text{annual}} = \frac{\frac{1}{8174}(735 + 1613 + 5389) - \frac{1}{8174^2}(738 \cdot 738 + 1848 \cdot 1813 + 5588 \cdot 5623)}{1 - \frac{1}{8174^2}(738 \cdot 738 + 1848 \cdot 1813 + 5588 \cdot 5623)}$$

Accordingly

$$K_{\text{annual}} = 0.8865943503$$

Semi-annual

Table 2: Result of the cross-validation semi-annual

	newDeforest	stayDeforest	stayForest	Sum
newDeforest	**736**	2	0	738
stayDeforest	1	**1616**	196	1813
newForest	2	226	**5395**	5623
Sum	739	1844	5591	**8174**

We calculate the accuracy$_{\text{semi-annual}}$:

accuracy$_{\text{semi-annual}}$ = (736 + 1616 + 5395) ÷8174

This results in an accuracy of about 0.95 or 95%.
Here, too, we have calculated the Kappa index.

$$K_{\text{semi-annual}} = \frac{\frac{1}{8174}(736 + 1616 + 5395) - \frac{1}{8174^2}(739 \cdot 738 + 1844 \cdot 1813 + 5591 \cdot 5623)}{1 - \frac{1}{8174^2}(739 \cdot 738 + 1844 \cdot 1813 + 5591 \cdot 5623)}$$

Accordingly

$$K_{\text{semi-annual}} = 0.889153004$$

4 Results

First, the two line graphs from figure 3 and 4 show a steady decline in forest pixels from year 2013 to 2019. In 2013, about 84% (more than 1,000,000 ha) of the pixels in our test area were still forest pixels. In 2019, the proportion has decreased to just under 78% (about 940,000 ha). This corresponds to a decrease of 6% (or 72000 ha) in 6 years.

The monitored classification of the annual data has an accuracy of around 95% and a kappa index of around 0.89. The monitored classification of the semi-annual data has an accuracy of around 95% and a kappa index of around 0.89. Both correspond to an almost perfect match.

5 Discussion

Regarding research question 1:
The NDVI proves to be extremely meaningful. This can be said with certainty, as the supervised classifications (annual and semi-annual) were calculated solely on the NDVI layers. The kappa indices prove this.

Regarding research question 2:
We first applied the window(focal)-function to our data (annual and semi-annual). Since no significant differences could be found here, we performed the supervised classifications with subsequent validation to check the results of the window-function. It could not be established that the classification on year-round means gave better or worse results than on biannual means. This is confirmed by the kappa indices (both around 0.95). However, it should be noted that the last image of 2019 was not used in the semi-annual calculation because it contained too many clouds. In the full-year picture, this data, which is sometimes more up-to-date, would still be included. With the semi-annual we had to fall back on data that is at least half a year old. Otherwise we would have had a very patchy result.
Although the supervised classification confirms the results of the window function, the supervised classification is preferable to the window function. The reason for this is that a threshold must be defined below which pixels are removed. This threshold can be chosen relatively untrue, so pixels that could have been useful for further calculation are removed. In figure 11 you can see that there areas in our research area that were detected by the supervised classification but not by the raster calculations after running the window functions.

Figure 11: Comparison of areas classified as newly deforested between window-calculation and supervised classification. Red areas: window-function, blue areas: supervised classification (Source: Own figure)

6 Future work

Future work should first address whether the yellow areas identified in our work as "natural decline" at the edge of cleared areas (shown in red and purple) were caused by man-made rough grazing. Another interesting question is whether cleared (supposedly naturally deforested) lines (presumably roads) suggest future raw earthworks. This could be done by looking at whether deforested areas appear in a certain area around these lines, within a certain time. If this were to be confirmed, the surrounding areas of these lines could be specifically monitored, if the possibility of illegal clearing exists. In our work, we completely disregarded this analysis, as we wanted to concentrate on large areas first.

References

Amazon rainforest. (2021). Retrieved from https://en.wikipedia.org/wiki/Amazon_rainforest

focal: Focal values. (n.d.). Retrieved from https://www.rdocumentation.org/packages/raster/versions/3.4-5/topics/focal

gdalcubes: Earth Observation Data Cubes from Satellite Image Collections. (n.d.). Retrieved from https://cran.r-project.org/web/packages/gdalcubes/index.html

Landsat normalized difference vegetation index. (n.d.). Retrieved from https://www.usgs.gov/core-science-systems/nli/landsat/landsat-normalized-difference-vegetation-index?qt-science_support_page_related_con=0#qt-science_support_page_related_con

Londhe, D. (2019, 05). *Arbitrary cut-off values for forest using ndvi?*

Regenwald-Abholzung in Brasilien hat sich verdoppelt. (2019). Retrieved from https://www.dw.com/de/regenwald-abholzung-in-brasilien-hat-sich-verdoppelt/a-51679932

Zerstörung im Schatten der Corona-Krise. (2020). Retrieved from https://www.tagesschau.de/ausland/brasilien-coronavirus-amazonas-101.html