Burned Area Mapping in Rondonia using tools in R

Interuniversity Master in Mathematical, Statistical and Computational Modeling

PUBLIC UNIVERSITY OF NAVARRE

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Rubén Andueza San Román

Supervisors: María Dolores Ugarte and Manuel Montesino-San Martín

Abstract

The Amazon rainforest is one of the most important biodiversity reservoirs of the Earth and suffers environmental pressure due to the exploitation of its resources and the conversion of forest areas into farming areas. In recent years, the climate crisis has become a political weapon that makes these protected areas vulnerable to government changes and their new environmental policies. Its monitoring can help to preserve the biome and control the international standards and agreements compliance. To carry out such surveillance, satellite images are one of the best available resources. They allow computing indexes to study burned and/or deforested areas. The present work aims to determine the burned area in the Brazilian state of Rondonia during August from 2011 to 2019 using a classification algorithm and some statistical techniques with R programming language. The results obtained are compared to those provided by the National Institute for Space Research (INPE) from Brazil and NASA's burned area product.

Resumen

El Amazonas es uno de los reservorios de biodiversidad más importantes a nivel global y sufre una gran presión medioambiental debido a la explotación de sus recursos y a la conversión de espacios forestales en zonas agrícolas. En los últimos años la crisis climática se ha convertido en un arma política que hace que estas áreas protegidas sean muy vulnerables a cambios en los gobiernos y en sus nuevas políticas medioambientales. Su monitorización puede ayudar a preservar el bioma y controlar el cumplimiento de normativas y acuerdos internacionales. Para llevar a cabo dicha vigilancia, las imágenes satelitales son uno de los mejores recursos disponibles. Permiten el cálculo de índices para el estudio de zonas incendiadas o deforestadas. El presente trabajo pretende determinar el área quemada en el estado brasileño de Rondonia en agosto desde 2011 hasta 2019 mediante un algoritmo de clasificación y diferentes técnicas estadísticas utilizando R. Los resultados obtenidos se comparan con aquellos facilitados por el Instituto Nacional de Pesquisas Espaciales (INPE) de Brasil y con los datos disponibles en la base de datos de la NASA.

1 Introduction

1.1 Wildfires and fire management policies

Fire is a natural disaster that plays different roles in Earth's ecosystems. On the one hand, it contributes to the global climate system emitting greenhouse gases and trace gases like carbon dioxide or nitrous oxide, which change the atmospheric chemistry. On the other hand, fires promote diversity and natural regeneration. Some species require wildfires for survival and others are sensitive to them (Kelly and Brotons, 2017).

But the human factor breaks the balance of the fire effect in Earth's ecosystems. Fires are one of the most common tools used for land management, and if appropriately used, they can be an important tool to improve and control vegetation and to prevent deforestation and big wildfires. The aim of controlled fires for land management goes from cleaning unwanted vegetation to reduce fuel loads until promoting seed germination and vegetation regeneration. But fires are also used for deforestation and clearing land for farming. These uncontrolled applications carry damages on critical biomes for the Earth ecosystem.

In some developed countries government policies tend to suppress all kinds of wildfires. A good example is the European Union. The "FOREST FIRES: Sparking firesmart policies in the EU" report (Directorate-General for Research and Innovation of the European Commission, 2018) gives scientific-based recommendations to legislate, in order to improve forest fire risk management and to reduce their catastrophic consequences. This report presents a good approach for a permanent dialog between science, management, and policy makers. The fire management cycle and the recommendations in each phase are summarized in Fig. 1.

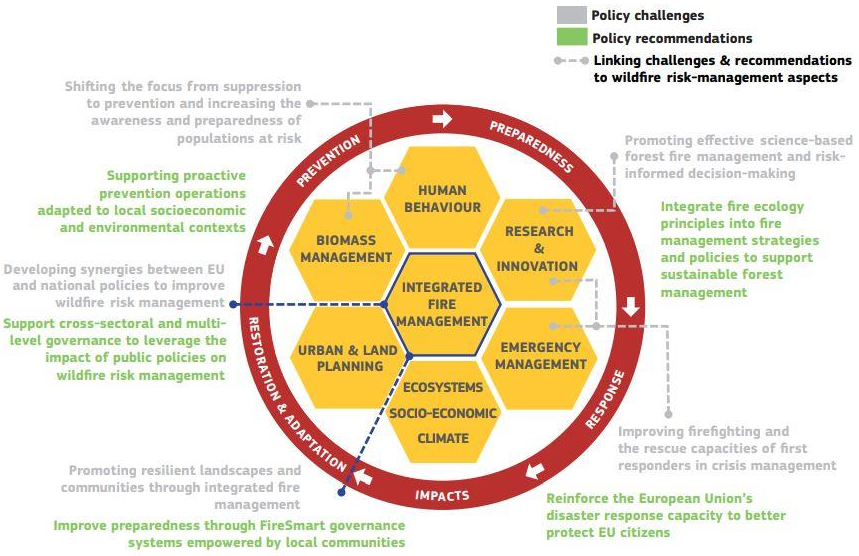

Figure 1: The fire management cycle (Directorate-General for Research and Innovation of the European Commission, 2018)

But other countries promote uncontrolled fires, directly or indirectly, for land management. Some underdeveloped countries or regions, like Central Africa, have less restrictive laws in this aspect, and they see fires as a tool for deforestation and economic opportunities.

Somewhere in between is Brasil that has implemented strong policies like the monitorization of land change carried out by the Brazilian National Institute for Space Research (INPE). Also, the Action Plan and Control of Legal Amazon Deforestation (PPCDAM) presented in 2004 has helped to control deforestation, which is related to wildfires (one of the most common tools used for clearing forest vegetation). But the political changes in 2019 have resulted in an unprecedented fire activity during August since 2010 (the period 2011-2018 the fire activity and the burned area were almost stable), although some studies indicate this increase around the average of the full-time series (Lizundia-Loiola, J et al., 2020).

Many studies about wildfire modeling have been developed in the last 20 years. The aim of these studies is to prevent wildfires in protected ecosystems with different approaches. Some of them develop models that predict the behavior of fires (Rodriguez-Aseretto et al., 2013). Other studies are centered on mapping burned areas to control environmental law enforcement, and once the burned areas are mapped some of them determine the severity of the fires (Escuin et al., 2008) to establish forest damages and CO_2 emissions (Hoelzemann et al., 2004). The present study focuses on mapping burned areas through satellite images and compare the results with other sources to check if there is an increase of burned areas in certain regions in August 2019.

In Sections 1.2 and 1.3 the region of interest and the method used to detect burned areas are presented. Section 2 consists of the physical base of the algorithm (Section 2.1), the data used (Section 2.2), and the explanation of the classification algorithm chosen to detect burned areas (from Section 2.3 to Section 2.11). The new approach developed in this work is presented in Section 3. Exploratory data analysis and results are detailed in Section 4. Some concluding remarks are summarized in Section 6.

1.2 Region of interest

The Amazon rainforest has an extension of 7 million km^2, the largest in the Earth. It is crucial to the world's oxygen and carbon cycles. It generates 6% of the planet's oxygen and also absorbs a big amount of CO_2. But in recent years the wildfires in this kind of rainforest have increased to alarming rates that suggest some of the tropical rainforests will be emitters of CO_2 in the next decades if the fire activity continues increasing (Hubau et al., 2020).

The Amazon is a tropical rainforest divided by the equator into two parts. As a common tropical rainforest, the climate is hot and humid, with a high amount of precipitation. Although it rains considerably throughout the year, two weather seasons can be defined, the rainy one and the dry one. In the first half of the year, the Amazon region located in the northern hemisphere is in the dry season and the

region located in the southern hemisphere is in the rainy season. The second part of the year is the opposite.

As exposed above, INPE reported an increase in the number of active fires during August 2019 based on information gathered by the Moderate Resolution Image Sprectroradiameter (MODIS), an instrument on board NASA's satellites. Historically, August and September have been the most affected months by wildfires in the Brazilian Amazon biome (see Fig. 2).

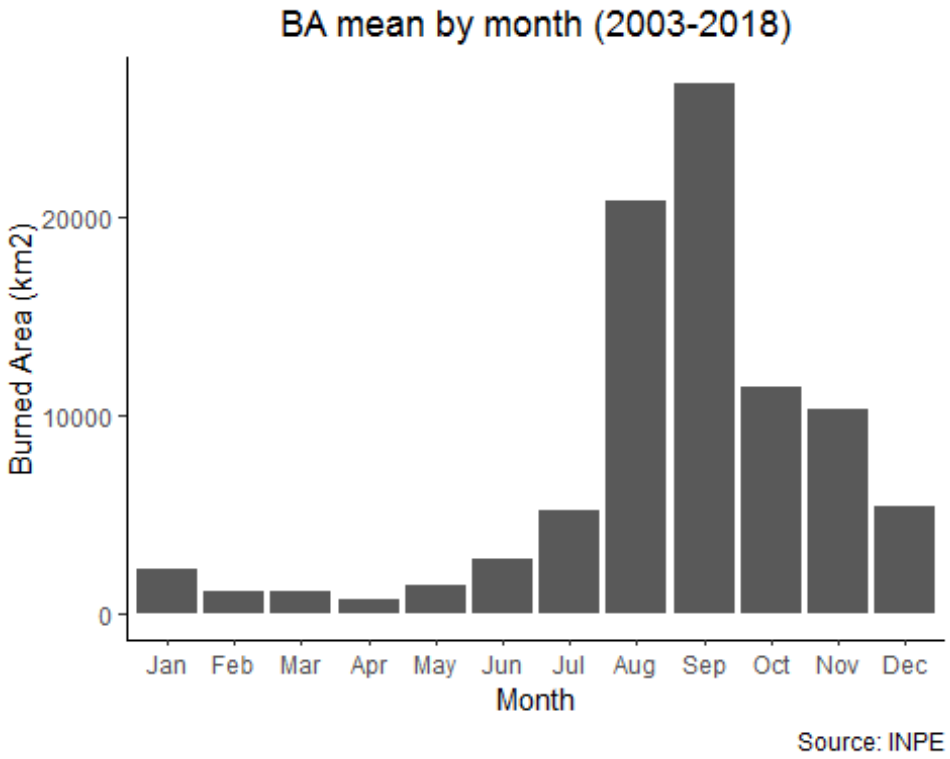

Figure 2: Burned Area (BA) detected in Brazilian Amazon biome (on average). The data have been provided by INPE under request (AQ1km product).

Regions in the rainy season are not affected by wildfires and also have lots of clouds that generate missing data. In consequence, we are looking for a region in the southern hemisphere, which is strongly affected by wildfires in August (the critical month in this Amazon crisis in 2019). In Fig. 3 we can see a two-dimensional density map of active fires detected in August 2019 in the Brazilian Amazon biome.

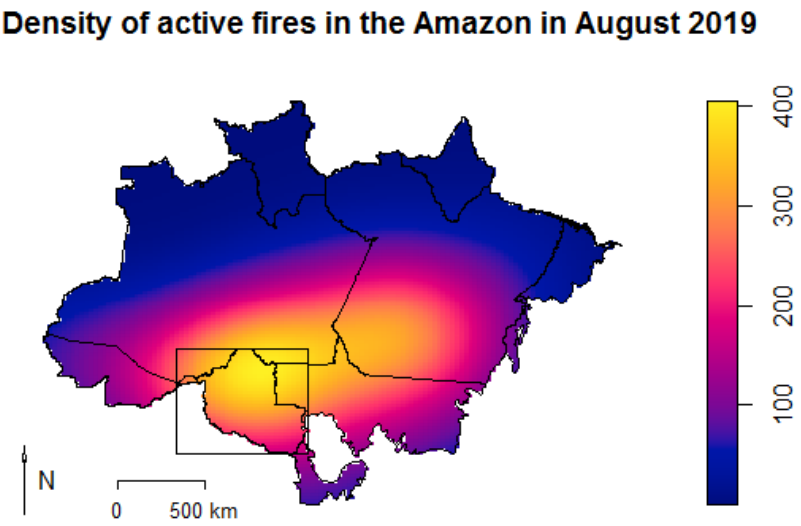

Figure 3: Brazilian Amazon biome active fires density map. Source: MODIS active fire product.

This density map is based on MODIS active fire product generated by the algorithm developed by Giglio et al., (2016). The method used to generate the map of Fig. 3 is a point pattern analysis. Using the `spatstats` package (Baddeley et al., 2015) a function based on the distance from each pixel to the active fire points is computed to calculate the number of active fires detected per unit area. More details about this computation are given in the package's reference manual (page 339).

Some regions of the states of Rondonia, Amazonia, Para, and Mato Grosso in Brasil are highly affected by fire activity according to NASA's active fire data in August 2019. This study is focused in Rondonia (the squared region in Fig. 3), a medium-size state that has the highest active fire density. The size of this state (237.576 km^2) is adequate to the scope of this study. Other big states would imply a high computational cost and in the cases of Amazonia and Para, which are near the Equator, the weather seasons are not clearly marked and there are usually many missing data due to the presence of clouds.

1.3 Methods for mapping burned areas

First studies about the burned area (BA) mapping collected information from ground estimations of fire management teams, but the different methodologies of data collection made it difficult to process this information. In 1972 the first Landsat satellite (ERTS-1) was launched, and two years later the first BA mapping algorithm was presented (Hitchcock and Hoffer, 1974).

Since then, different methods have been developed taking data from various sources. A great review of historical approaches of BA mapping is presented in Chuvieco et al., (2019). In this paper, machine learning algorithms are slightly mentioned, but it is centered around physically-based algorithms due to their

versatility and adaptability. They can be classified by the spectral regions of the electromagnetic spectrum used (Solar domain, Middle infrared and thermal domain or microwave domain) or the method used to extract data (optical sensors, radar, lidar or synergetic approaches). They also can be classified as global or regional algorithms, depending on the size of the region it is being applied to (regional products are developed to be applied to specific areas and so that they consider the peculiarities of each region).

1.4 Approach of this study

In our case, we map BA somehow reproducing the NASA's algorithm introduced by Giglio et al. (2018). It is a physically-based algorithm that uses the solar domain data acquired by optical sensors (satellite images from the MODIS sensors) and generates a global product identified as MCD64A1. We apply this algorithm to a specific region in the Brazilian Amazon (Rondonia) and we compare NASA's results with ours's which have a spatial coherence approach (this will be detailed in subsequent steps).

INPE developed a regional algorithm for the South America region (Libonati et al., 2015) but it is in a lower stage of maturity than the NASA's algorithm and also it is more difficult to reproduce. Therefore, we select NASA's algorithm rather than INPE's to conduct our analysis.

2 The algorithm

In this section, NASA's algorithm for Burned Area mapping using satellite data is presented. The basic physical background is introduced to support the empirical knowledge applied in some sections. However, the mathematical background is the primary focus of the section. All thresholds established in the algorithm are also based on more than 20 years of research and evolve with each version of the algorithm.

Section 2.1. explains the basic physical background about the electromagnetic spectrum and its role to differentiate between objects or phenomena such as fires from images. Then, in Section 2.2., we will cover the available sources and data products of satellite imagery. We end up explaining the algorithm scheme in Section 2.3, where the subsequent sections are summarized.

2.1 Physical basis

Many aspects influence on BA mapping through satellite images: the type of fire, the vegetation chemical and physical characteristics, and the temporal difference between fire extinction and image acquisition are critical.

We will use data from the MODIS sensor onboard the Terra and Aqua satellites. These sensors measure electromagnetic radiation (ER) reflected from the Earth. The ER is a form of energy that flows through the vacuum or material mediums as electrical and magnetical waves. It can be expressed in terms of energy, wavelength, or frequency. MODIS data are available in terms of wavelength and expressed in meters.

The electromagnetic spectrum (ES) is the range of all types of ER. MODIS sensors measure ER from 400 nm to 14.400 nm divided into 36 bands, which correspond to a small part of the ES, as shown in Fig. 4.

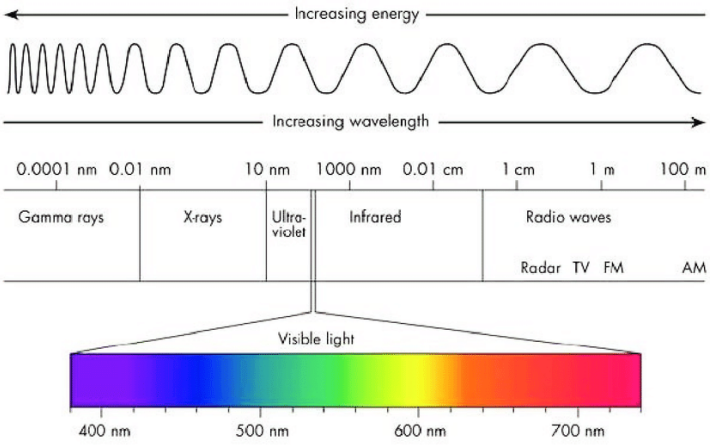

Figure 4: Diagram of the light's electromagnetic spectrum (Baier, 2012).

The solar domain will be the ES range used for BA mapping. This range includes visible light (400-700 nm), near-infrared (800-2500 nm), and short-wave infrared (1400-2200 nm).

Sun light traveling through the atmosphere, reflected by the Earth's surface, and captured by satellites suffers from several sources of error. The measurements collected by sensors go through several corrections levels to compensate for these disturbances. The outcome from the last level is known as surface reflectance, i.e. the reflectance of light as it would be measured near the surface. Reflectance refers to the portion of light reflected by the surface of the Earth, with atmospheric correction (it corrects for the effects of atmospheric gases and aerosols). Reflectance is a unitless ratio of surface radiance to surface irradiance with a value that falls between 0 and 1. But the application of the atmospheric correction results in values between -0.1 and 1.6 (Vermote, 2015).

The reflectance across the spectrum is known as spectral signatures, and every objects or surface states have their own. Fig. 5 represents four spectral signatures of different land states: unburned, surface fire (only the surface vegetation and other kinds of fuel present in the ground burn), crown fire (it is the most intense, it burns trees from the ground to the crown), and light charcoal (the situation after a crown fire occurred). The spectral signatures show different behaviors in each situation. A combination of certain ranges of the solar domain will be a useful tool to detect burned areas. Also, the type of fire affects the algorithm capacity of detecting areas where a surface fire occurred because the spectral signature of burned areas is similar to that of unburned areas.

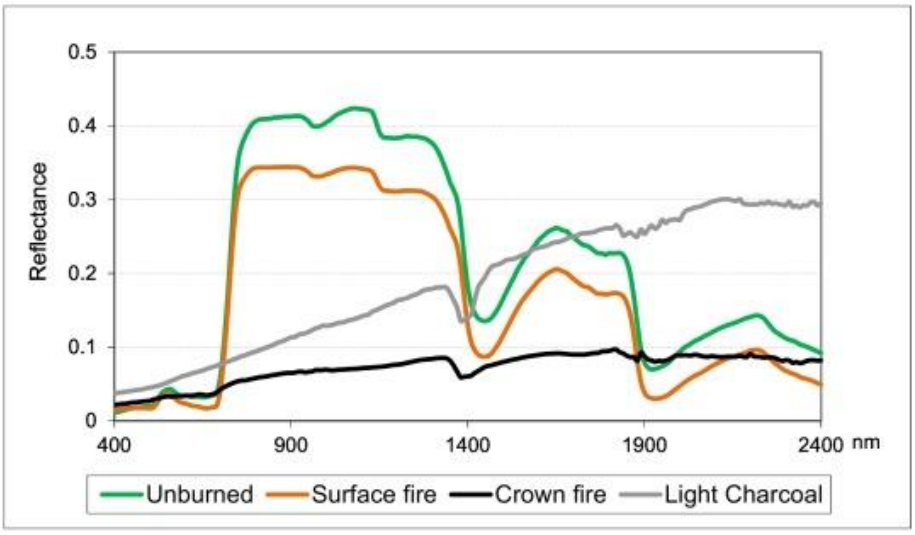

Figure 5: Solar domain behavior. Chuvieco et al., (2019)

2.2 Types of data and sources

Spatial data are usually referred as information associated with coordinates based on a reference system (Bivand, 2013). The **coordinate reference system** (CRS) is how a map is projected on a plain. In particular, MODIS land products are produced in three different projections: Sinusoidal, Lambert Azimuthal Equal-Area, and Geographic.

According to the technique used to store them, we can differentiate two types of data. The first type is raster data which is composed of grid cells identified by row and column. Satellite images are geographic areas divided into a grid of pixels. The size of the pixels is called the **resolution** of the image.

The second is vector data (points, lines, and polygons), which is used to represent entities as exactly as possible. We will use this kind of data to represent our region of interest (the state of Rondonia), and `SpatialPoints` (vector data with points as their geometric feature) to represent active fires. Both data' types allow more information than just geographical reference and this information is called **attribute**, which can be a single value for each pixel/point/line/polygon or a vector with different features.

In this work, we will also use data generated by other algorithms. The data have been preprocesed, which means that their accuracy have been tested. MODIS distinguishes three validation levels (or maturity levels) depending on how extensively data have been tested against ground truth. Stage 3 represents robust data with well-known uncertainties, Stage 1 is data evaluated against few observations and Stage 2 is an intermediate state. Most datasets used in our analysis belong to Stage 3 with the exception of the burned area from INPE (Table 1).

In MODIS, images are systematically divided following the grid system represented in Fig. 6. Each of the grid cells is called a tile. The tiles covering the region of interest for the time period of interest (August, 2011-2019) were downloaded and uncompressed using `RGISTools` (Perez-Goya et al., 2020). The following processing steps were carried out using the `raster` package and its functionalities (Hijmans, 2019).

Table 1: *All data products used in the study. In the product name it also is identified the source of the data (NASA or INPE)*

Product	Reference	Description	Temporal resolution	Data type	CRS	Spatial resolution	Stage of maturity
Surface reflectance (MOD09GA/ MYD09GA) (NASA)	Vermote, E. F., & Vermeulen, A., 1999	It estimates the value of the surface spectral reflectance as it would be measured at ground level, suppressing atmospheric gases and aerosol influence of the radiation data.	Daily	Raster	Sinusoidal projection	500 m / 1000 m	3
Land cover (MCD12Q1) (NASA)	Cover, M. L., & Change, L. C., 1999.	It identifies 17 different categories of land cover following the International Geosphere-Biosphere Program (IGBP) scheme.	Annually	Raster	Sinusoidal projection	500 m	3
Burned area (MCD64A1) (NASA)	Giglio et al., 2018	It is produced by the algorithm explained in this work.	Monthly	Raster	Sinusoidal projection	500 m	3
Active fires (NASA)	Giglio et al., 2016	It identifies points with high changes in surface temperature (thermal anomalies) and strong emissions of mid-infrared radiation produced by fires.	Daily	SpatialPoints	Longlat projection (WGS84)	-	Near-real-time standard quality
Burned area (BA1km) (INPE)	Libonati et al., 2015	It detects burned area in South America region.	Monthly	SpatialPolygons / Raster	Longlat projection (WGS84)	1000 m	2

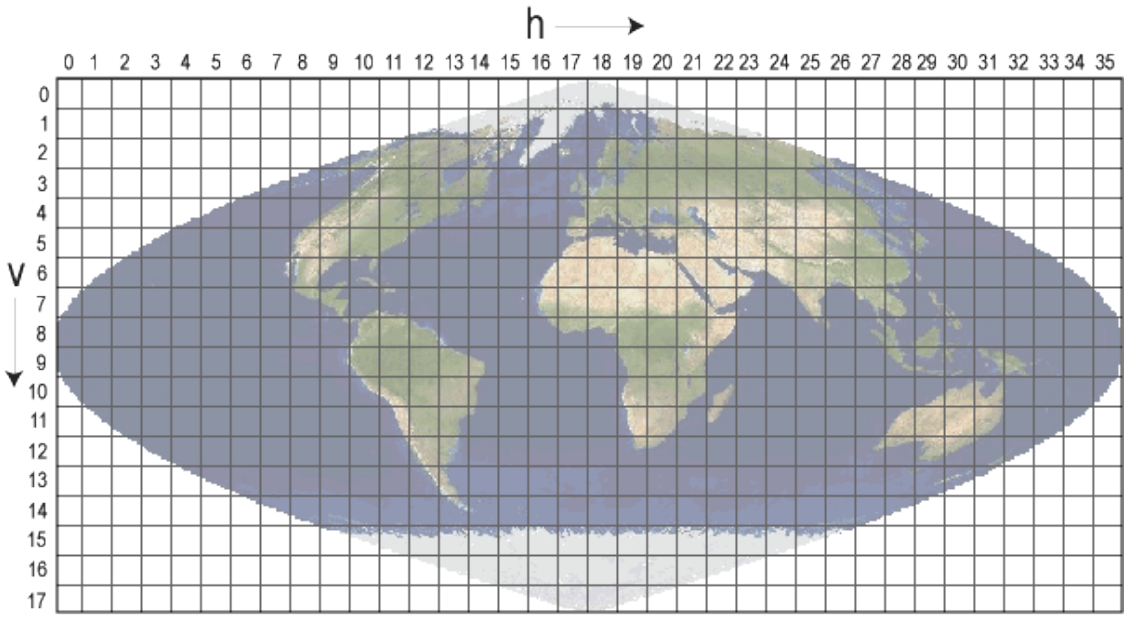

Figure 6: MODIS Tile Grid

Modis active fire product (MODIS Collection 6 NRT Hotspot / Active Fire Detections MCD14DL. Available on-line [https://earthdata.nasa.gov/firms]) is available under request in shapefile format. Also INPE burned area product (AQ1km) is available under request in shapefile format (INPE's product was requested by email, it is not available on its website). Both are loaded using `rgdal` (Bivand et al., 2019) package and the data is processed using functions of `raster` and `sp` (Bivand, 2005) packages.

2.3 Scheme and summary of the algorithm

The scheme of the algorithm divided in 3 main parts is presented in Fig. 7:

The first part (1) of the flowchart corresponds to the input data of the algorithm. The data have been downloaded in different compressed formats and extracted to the appropriate format as explained in Section 2.2. In Appendix 2, the structure of the database generated by the algorithm is presented. It is important to be careful when storing the data downloaded and generated by the algorithm.

The second part (2) corresponds to the preprocessing of the data to get the physical and temporal information needed. Initially, the vegetation index is computed for the specified period (Section 2.4) using surface reflectance as input data. Then the algorithm looks for a trend change in the index time series that may indicate a fire occurred in a pixel (Section 2.5). It ends up saving temporal and physical information of the day whenever a trend change is detected (composite images generated in Sections 2.5 and 2.6).

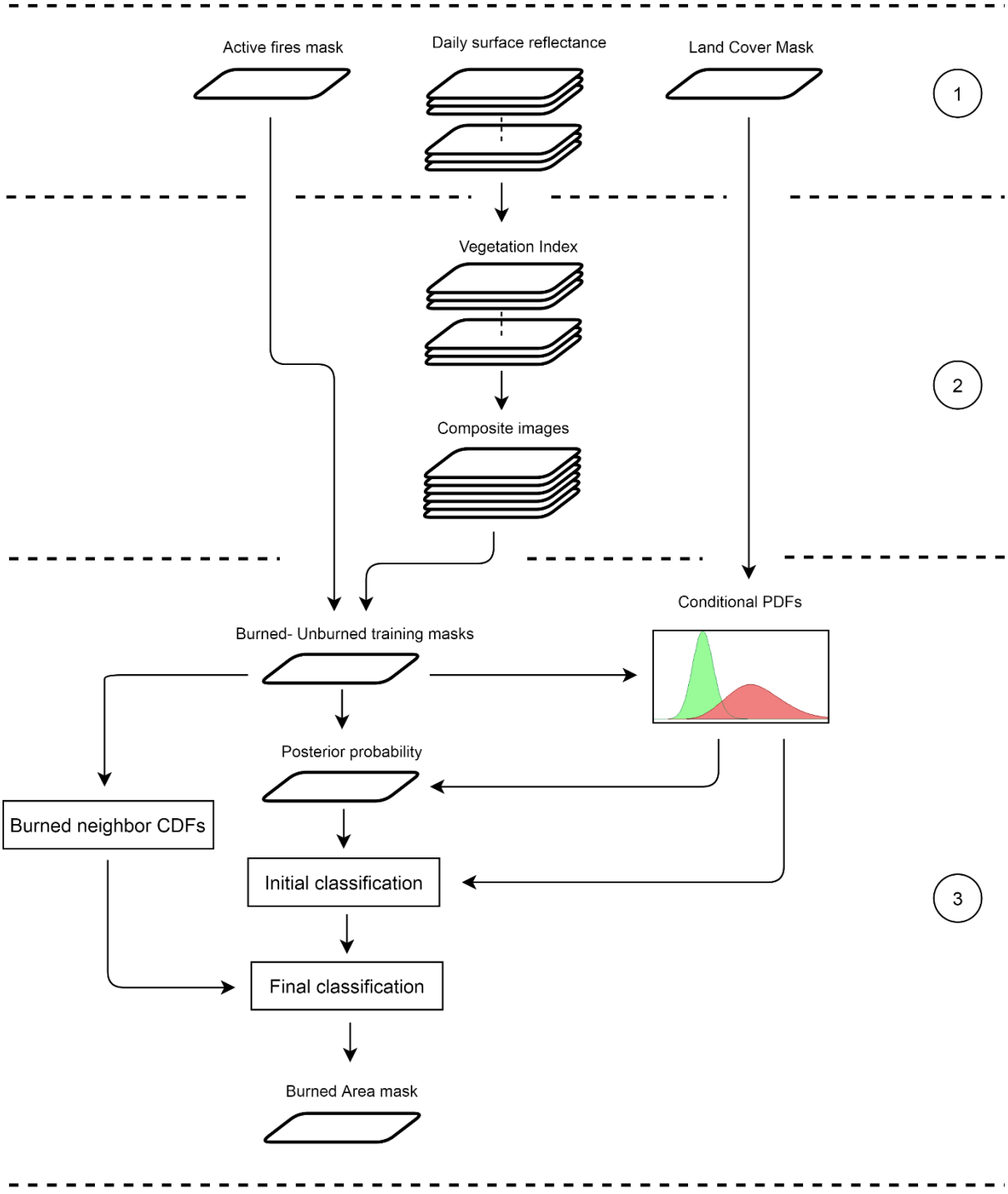

Figure 7: Algorithm flowchart

In the last part (3), the classification algorithm uses as input data the composite images generated in the previous step, the active fires mask of the month being processed, and the land cover of the region of interest to generate burned and unburned training masks (Section 2.7).

A classification algorithm uses input features to categorize observations into a distinct number of classes. But, before the classification step, the algorithm needs to be trained with data whose final class is known (called training data).

The difference between this algorithm and a common classification one is that we need to generate the training data every time we apply the algorithm. We will establish burned and unburned training pixels based on the active fires detected and some temporal and spatial behavior of different features.

Usually, the training data are randomly selected to avoid overfitting of the classification algorithm. In our case, the algorithm is trained every time it is applied, and the adjustment of the classification algorithm to each training data is desired. But the original spatial application ignores the land cover particularities of the regions affected by wildfires. In this work, we present a new approach to deal with this issue, i.e., to take into account the land cover particularities (Section 3).

Finally, the classification step is divided into two phases. The first one (Section 2.10) is based on the physical behavior of the data (see Sections 2.8 and 2.9). The last one is based on the spatial coherence to relabel pixels depending on their neighbor's class (Section 2.11). The output of this step is the burned area of the region being studied (the final result).

2.4 The Vegetation Index

The burn-sensitive Vegetation Index (VI) is a combination of two spectral bands of the electromagnetic spectrum (Roy and Landmann, 2005). It is used to detect a change of trend in a time series when a fire occurred. The VI index is computed as shows Equation (1).

$$VI_i = \frac{\rho_{5i} - \rho_{7i}}{\rho_{5i} + \rho_{7i}} \qquad (1)$$

where $\rho_5$and ρ_7 correspond to the bands 5 and 7 (1240 and 2130 nm, respectively) of the MODIS reflectance product (MOD09GA/MYD09GA) and i denotes the days in the time series with valid data for each pixel. Additionally the product offers a quality band (QA) that, among other things, provides information about cloud coverage. The computation of the VI and the interpretation of the QA band was carried out with `RGISTools`.

The index takes values in the range [-1,1]. Unburned pixels usually take values between 0.4 and 0.9, depending on the characteristic of the land cover. Burned pixels usually take values between -0.2 and 0.2, depending on the features of the fire (surface fire, crown fire, the proportion of the pixel burned…).

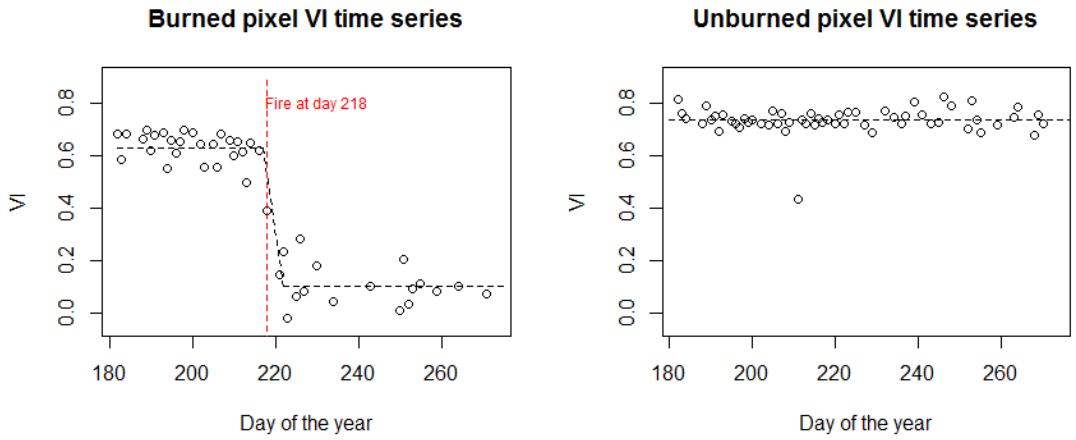

Figure 8: Difference in VI time series in Rondonia in August 2019

Fig. 8 shows the ideal behaviour of the time series of VI for a burned and unburned pixels (left and right, respectively). But, in other cases, the time series does not show this prominent change of trend on the burned pixels, and this deviation from ideal behavior can be fixed applying some spatial coherence computations.

The VI index uses valid reflectance data. As explained in Section 2.1, MODIS reflectance product values are in the range [-0.1,1.6] because the application of the atmospheric correction algorithm produces values out of the initial range [0,1]. The burned area mapping algorithm understands reflectance as the proportion of light reflected by the Earth and pixels with values out of the range [0,1] are removed.

Also, pixels covered by clouds are removed using the QA band of the surface reflectance product. This band is extracted and processed to obtain cloud information using the `modCloudMask` function of the `RGISTools` package. All pixels that were covered by clouds are removed. It is a critical step of the algorithm because it generates almost 40% of lost values on the entire dataset.

MODIS sensors are integrated on Terra and Aqua satellites and both generate an image of the Earth every one or two days. Two vegetation indexes (VI_{Terra} and VI_{Aqua}) are generated using data from both satellites and and they are combined to create one single image (VI) per day by getting the maximum value of each pixel. The nominal equatorial crossing time is different for each satellite and this allows us to avoid the presence of clouds in some pixels. Differences between the Terra and Aqua surface reflectance are assumed to be negligible after NASA's atmospheric corrections and due to the narrow time lapse.

In the following we calculate the VI of the region of interest explained in Section 3 at July 21, 2018. This date is selected, as every step mentioned in the process occurs on this day. Reflectance bands are examined looking for values out of the range [0,1] (Fig. 9).

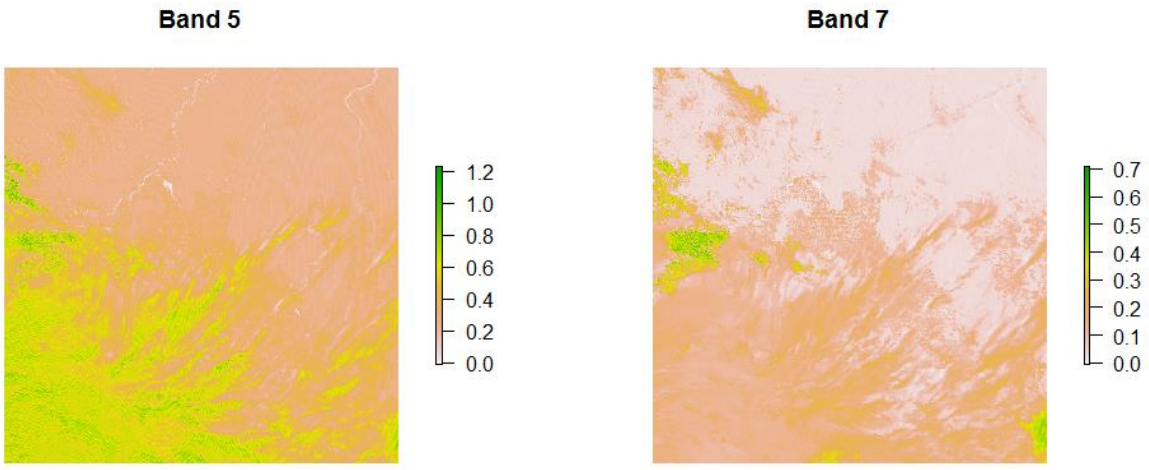

Figure 9: Bands 5 and 7 of the surface reflectance. Band 5 have invalid data out of the range [0,1]

Band 5 has values in the range [-0.0076,1.2305], and band 7 in the range [-0.0038,0.7093]. The invalid reflectance values are removed obtaining the images of Fig. 10.

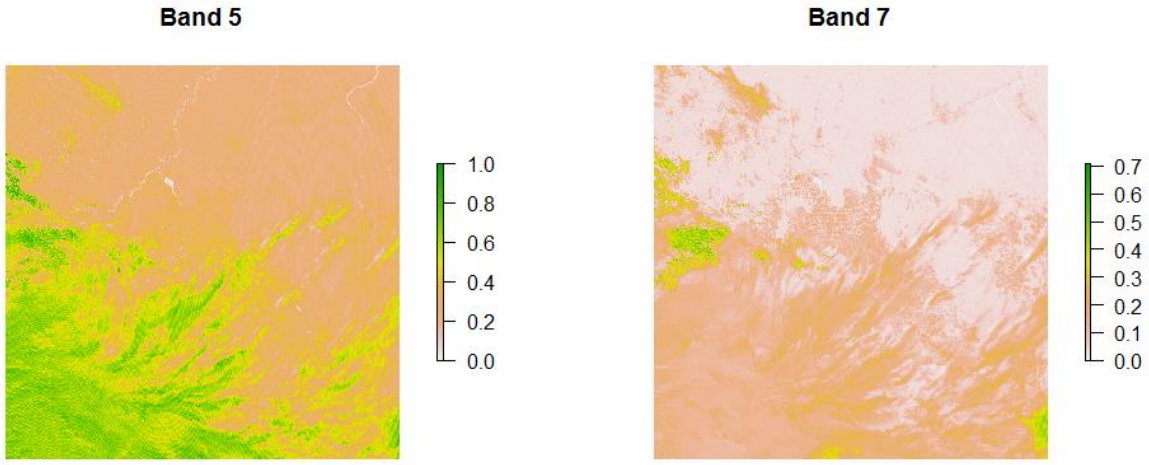

Figure 10: Bands 5 and 7 of the surface reflectance only with valid data.

Then, the VI index is computed using Equation (1) (Fig. 11 on the left). The QA band of the reflectance product is processed to obtain the cloud information, and pixels that are covered by clouds are removed (Fig. 11 on the right).

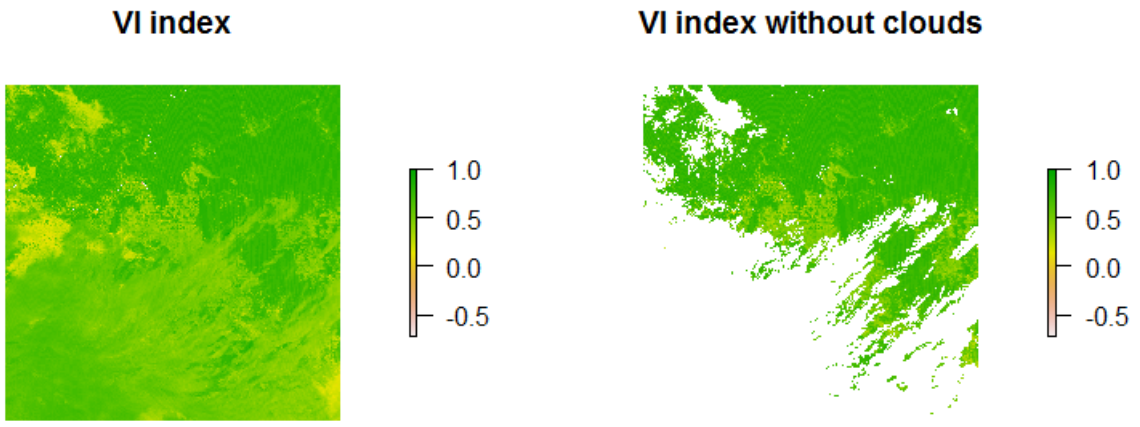

Figure 11: VI index (left graph) and VI index cloud-free (right graph).

The same process is applied to data obtained by the Aqua satellite. Fig. 12 shows both indexes only with valid data.

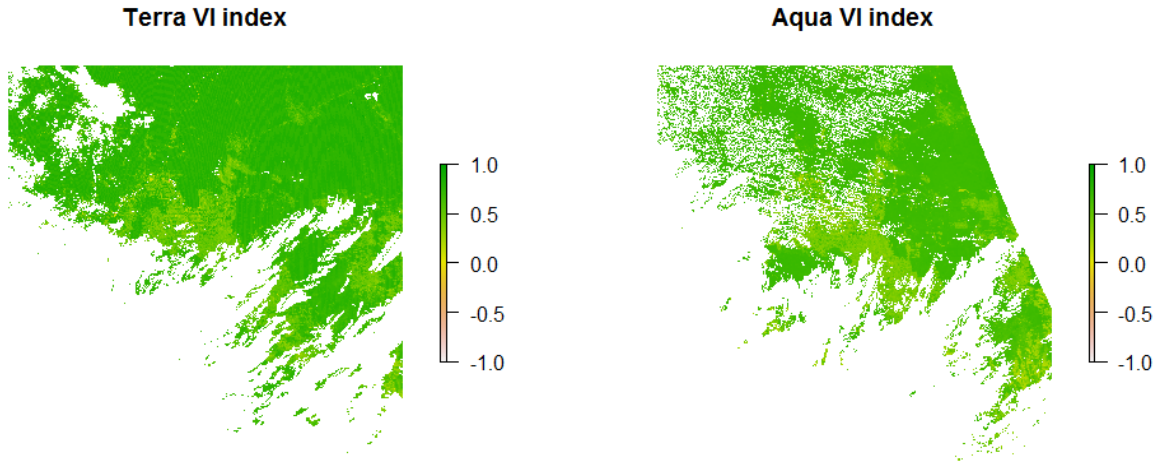

Figure 12: Terra and Aqua VI indexes. Each one is computed with data obtained by the correspondent satellite (Terra and Aqua)

The Terra VI index has 54% of missing values that correspond to pixels with invalid reflectance values and pixels covered by clouds. Also, the Aqua VI index has 66% of missing values. As explained in this section, to deal with the amount of missing data the both Terra and Aqua VI indexes are combined getting the maximum value of each pixel to create a single image that has 44% of missing values (Fig. 13).

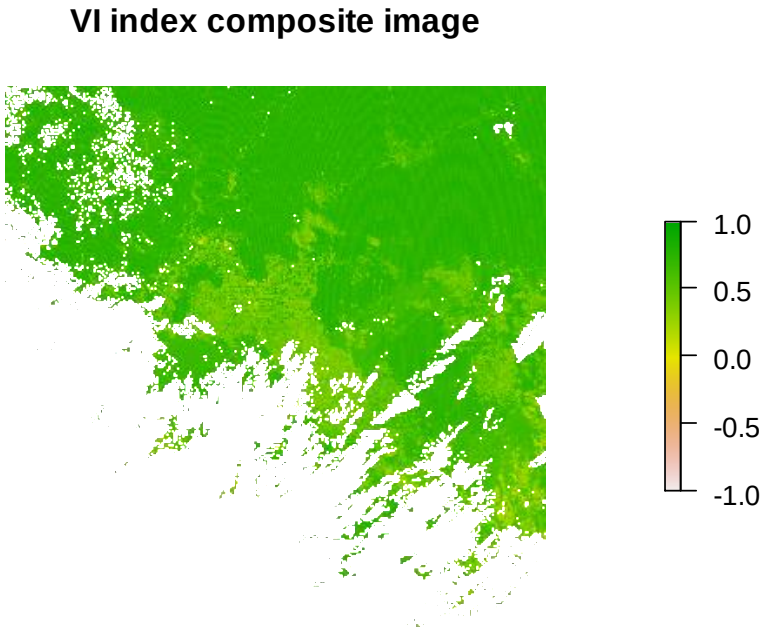

Figure 13: Composite VI index generated using Terra and Aqua data.

This process is repeated for each day of the period specified. The resulting time series is going to be studied in the next section to detect trend changes.

2.5 Trend changes detection

The algorithm needs data from the month that is being processed and previous and later months (3 months altogether). Then, the VI index is computed for every day of each pixel of the time series and this generates a 3-dimensional array VI(x,y,z) , where x and y correspond to the coordinates of the pixel and z corresponds to the day in the time series.

For a time series of VI from a single pixel, we consider a moving window with two legs (pre-wind and post-wind) (see, Fig. 14). Each leg has a length of 8 days (w = 8). The 8 days may not be consecutive as there might be some missing values between measurements. Data gaps lead to uncertainty which will be treated in following steps.

For every pixel (x, y) and considering a time window starting at k, we calculate the following statistics for the pre-wind leg:

$$VI_{pre}(x,y,k) = mean(VI_i(x,y),p) \qquad i \in (k,\ldots,k+w-1) \qquad (2)$$

$$\sigma_{pre}(x,y,k) = stdev(VI_i(x,y),p) \qquad (3)$$

Equation (2) and (3) are calculated for a pixel at location (x,y) and temporal index i. $p = 10$ indicates observations within the 10-90 percentile interval.

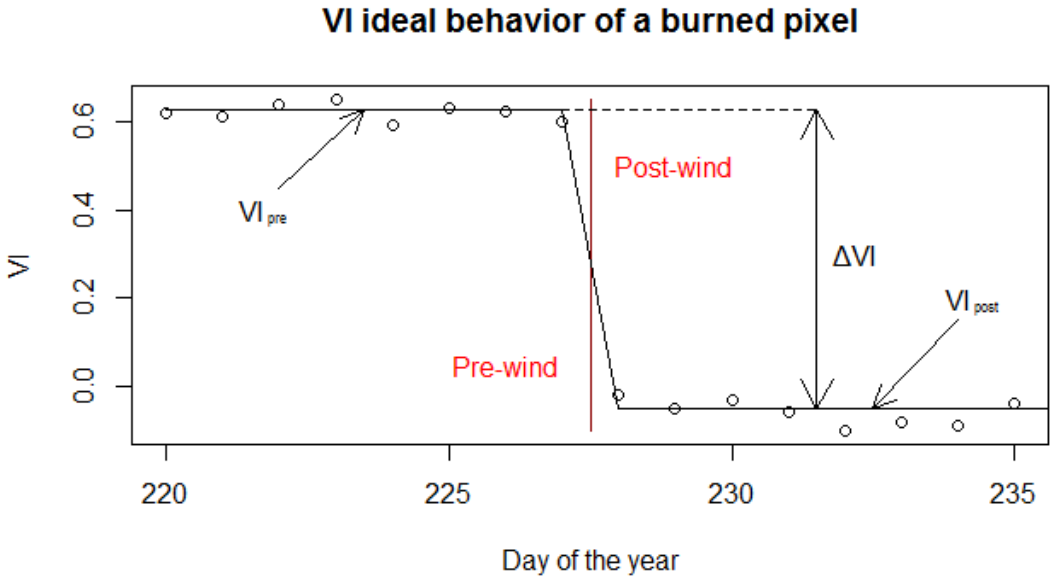

Figure 14: Scheme of a pixel pre-wind and post-wind.

The same statistics are computed to the post-burn window in Equations (4) and (5):

$$VI_{pos}(x, y, k) = mean(VI_i(x, y), p) \qquad i \in (k + w, \ldots, k + 2w - 1) \qquad (4)$$

$$\sigma_{pos}(x, y, k) = stdev(VI_i(x, y), p) \qquad (5)$$

Every pixel takes candidates for the pre-burn window no earlier than 30 days and selects the $w = 8$ days with valid data that are closest to k. If $w \leq 8$, in any of the two windows, this pixel will be labeled as unclassified and ignored in the next steps (equivalent method for post-burned window).

With these statistics, a measure of the difference between both windows in a pixel is computed in Equation (6), and it is called Separability:

$$S(x, y, k) = \frac{\Delta VI(x, y, k)}{[\sigma_{pre}(x, y, k) + \sigma_{pos}(x, y, k)]/2} \qquad (6)$$

where $\Delta VI(x,y,k) = VI_{pre}(x,y,k) - VI_{pos}(x,y,k)$ is evaluated for every k day. A time-series of Separability values can follow 3 different patterns. Whenever a fire takes place, the VI (Eq. 1) rapidly decreases (Fig. 14 on the left) and Eq. 6 results in a high value of separability (Fig. 15 on the left). Pixels where a fire does not occur, the VI index remains almost stable and we get separability values $S \approx 0$ (Fig. 15 on the right). When the VI (Eq. 1) rapidly increase, Eq. 6 results in large negative value of separability. It is not common to get negative values of separability but it could happen during a rapid recovery phase after a burn.

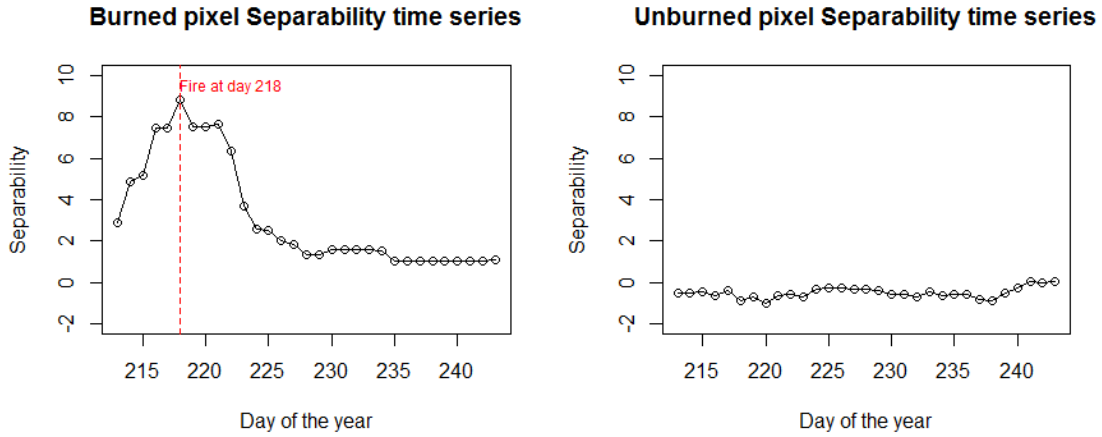

Figure 15: Separability behavior of burned and unburned pixels in Rondonia (August 2019)

Separability is a way to find a change of trend in a time series. There are a lot of more efficient methods (Militino, Moradi and Ugarte, 2020) to check trend changes. However, this method has proven effective, and the statistical metrics from Eq. 2-6 are useful for gaining further insights in subsequent steps of the analysis.

Also, the maximum of every time series is located and recorded (S^*), together with its temporal index in the time series (k^*). This value indicates the day when a pixel has burned (in case a fire occurred). Some composite images are saved to use them in subsequent phases.

The maximum separability day of each pixel is:

$$t^*(x,y) = (t_{k^*+w-1} + t_{k^*+w})/2 \qquad (7)$$

The result of Equation (7) represents the middle point between the last observation of the pre-burn window (t_{k^*+w-1}) and the first observation of the post-burn window (t_{k^*+w}) of the day of maximum separability for each pixel. The $t^*(x,y)$ value

corresponds to the day of the year, and the algorithm assigns that day as the one that the fire was detected.

A time interval is computed to measure the distance between the last and the first points of both windows in Equation (8). In satellite images, there are a lot of missing values (clouds, invalid data…), and a measure of uncertainty is needed.

$$\Delta t^*(x,y) = t_{k^*+w-1} - t_{k^*+w} \qquad (8)$$

The $\Delta t^*(x,y)$ value is used in equation (11) to deal with the influence of lost values in the computation of the $t^*(x,y)$ and the details of this equation will be explained in Sections 2.7.1 and 2.7.2.

The following composite images of the day of maximum separability (k^*) are saved: separability value ($S^*(x,y)$), the variation of the mean between the two windows ($\Delta VI^*(x,y)$), and the mean of the post-burn window ($VI^*_{pos}(x,y)$). All of them with Equations (7) and (8) will be used in some steps of the algorithm. The $S^*(x,y)$ can be better understood graphically.

Figure 16: Image of the value of maximum separability in the region of interest explained in Section 3 in August 2019

In Fig. 16, almost 71% of the pixels have a maximum separability value smaller than 2. A small value of separability means there is no significant trend change in VI time series, and hence, these pixels will be labeled as unburned. The criterion used to discriminate pixels with $S^*(x,y)$ value is detailed in Sections 2.7.1 and 2.7.2.

2.6 Spatial coherence measure

In this section, the concept of spatial coherence is introduced. The spatio-temporal evolution of the wildfires follows different patterns depending on multiple ecological and environmental factors, such as vegetation or precipitations. Also, the spatial growth of the wildfires is influenced by their neighbors. But depending on the projection of data and the location, we have to select a different number of neighbors.

The sinusoidal projection has a different impact on the Earth. Fig. 17 shows how the projection affects to 5 different locations.

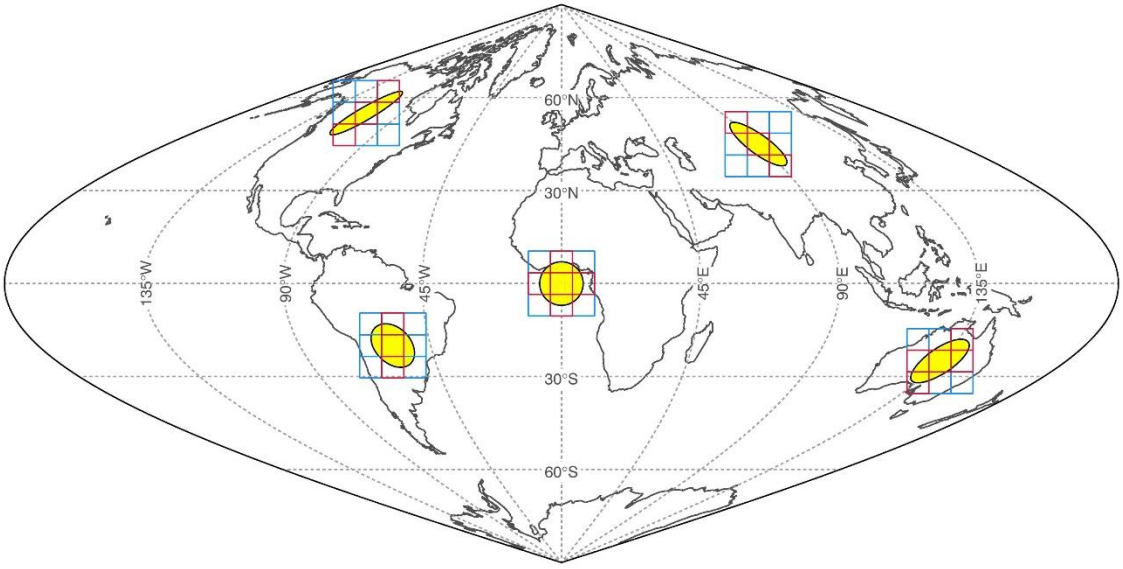

Figure 17: The equi-distant spatial kernels. Giglio et al., (2018)

Each region is distorted differently according to their location. The central Africa region situated at latitude and longitude (0,0) is not distorted. When latitude and longitude increases and/or decreases the distortion is remarkable (the cases of Australia and North America). The tile system used by NASA showed at Fig. 6 does not consider the particular distortions in each location caused by the projection of the coordinates.

The sinusoidal projection corresponds to the pseudocylindric projection showed in Fig 18, but the tile system interprets the map as if there is no distortion (similar to the cylindric projection which is the less distorted one). Therefore, it must be corrected in spatial coherence computations to consider only neighbors that represent close land.

The 5 locations presented in Fig. 17 can be interpreted as follows. The 3X3 kernel represents pixels in the tile system in each location (not drawn to scale), that does not consider the sinusoidal projection. The distortion is represented by the yellow

ellipse that have variable orientation and eccentricity depending on the location. Finally, the red squares of each 3X3 kernel correspond to neighbors that contain land near the pixel. The blue pixels represent neighbors in the tile system that, out of sinusoidal projection, represent land far away from the central pixel of the kernel. The use of physical and temporal information associated with the blue pixels will adversely affect the accuracy of the results.

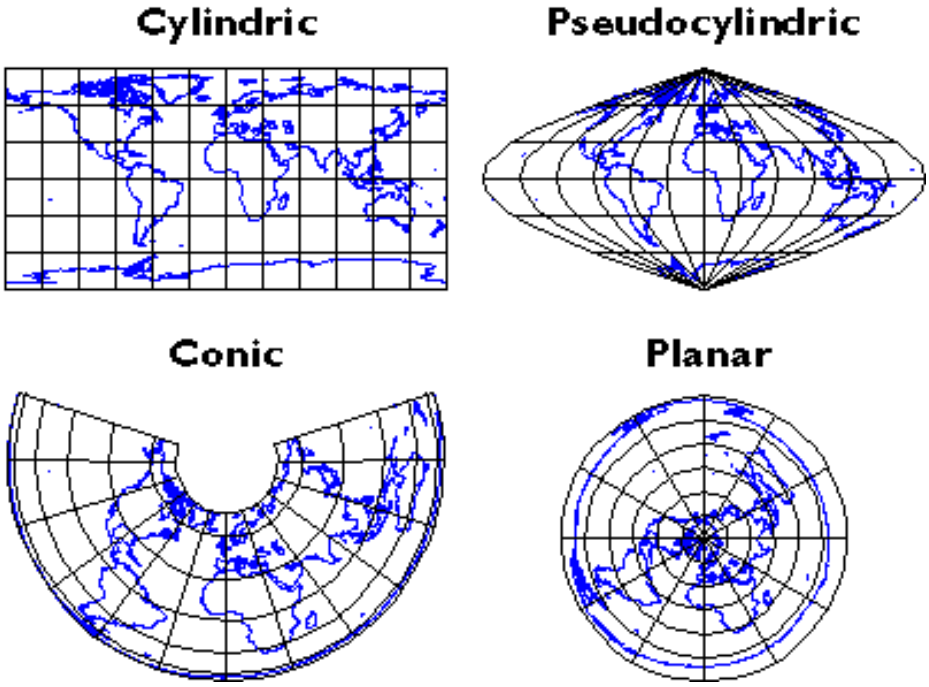

Figure 18: Four categories of map projections. (Penn State University, 2012)

We will use the kernel located on top of South America in several computations (Rondonia is under this kernel). In previous versions of the algorithm the kernel used for all regions was the “rook’s case” (kernel on top of central Africa), and it did not consider the sinusoidal projection effect. The methodology used to establish the kernel used in each region is not explained by NASA’s algorithm, which make it more difficult to standardize the process developed in this study to be applied in any region of the Earth.

In this step, the algorithm looks for areas that have high spatial coherence. It is based on the premise that if a fire occurred, neighboring pixels have similar values of the day of maximum separability. We take a composite image of $t^*(x.y)$ (Equation (7)), which is the estimation of the day where a burn occurred in a pixel (if a burn really happened).

We will choose for our region of interest a kernel with the central pixel, and its upper and lower neighbors. For each pixel, we compute the standard deviation of the selected kernel values. Then, an order ranked filter is applied to the image obtained,

which is a smoothing filter that assigns the 25th percentile of the kernel values to the central pixel. This technique is applied to reduce the uncertainty caused by lost values and it has been tested empirically by the NASA. We obtain an image that is denoted temporal texture, σ_t^* (Fig. 19).

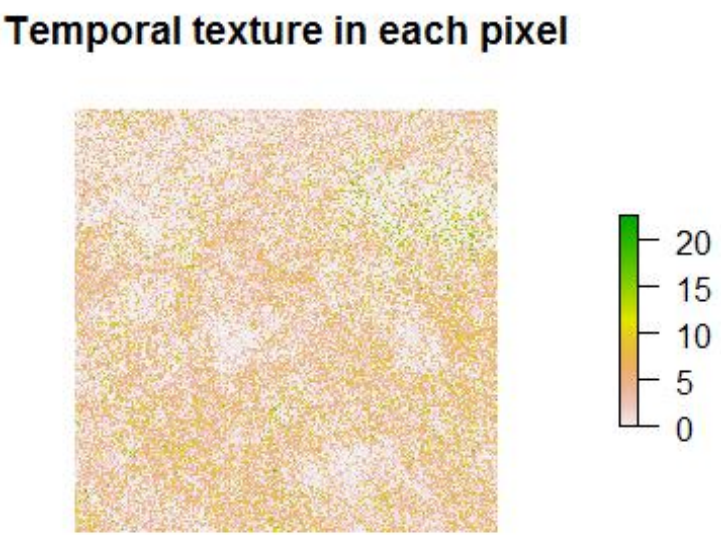

Figure 19: Image of the temporal texture in the region of interest explained at section 3 in August 2019.

The values of the temporal texture in August 2019 are in the range of [0,26] (measured in days). Low values indicate regions with high spatial coherence on the day a fire occurred. It is more likely to detect a fire in a region with a small temporal texture than in regions with high temporal texture. Figure 19 shows that most of the temporal texture values in this region after having applied the smoothing filter are compatible with fire occurrence. The 99 % of the pixels has temporal texture values lower than 8. The criterion to discriminate into burned-unburned labels using the temporal texture is detailed in the next section.

2.7 Selection of the training pixels.

In this step, the methodology to extract the data needed to train the classification algorithm is presented. The concept of training data was presented in Section 2.3. The burned and unburned training pixels are selected based on the following criterion:

2.7.1 Unburned pixels.

The algorithm looks for pixels that have big value of temporal coherence with their neighbors (Equation (10)), and the change of trend in the VI time series is small

(Equation (9)). All pixels that satisfy either of the following conditions are a priori very unlikely to experience a wildfire and therefore they are labeled as unburned pixels:

$$S^*(x,y) < 2 \qquad (9)$$

$$\sigma_t^*(x,y) > 8 \qquad (10)$$

2.7.2 Burned training pixels

Active fires are points where a thermal anomaly is detected (Giglio et al. 2016). The active fire algorithm uses surface temperature and surface reflectance data to detect anomalies that are associated with points (not pixels) where a fire occurred on a specific date. But an active fire is not always turning into a wildfire, it can burn only a small percentage of a pixel. When most of the pixel has not burned and the value of the VI does not change, the algorithm does not detect a trend change (small value of separability) and the pixel is labeled as unburned.

But the MODIS active fire product is a great source to identify burned pixels. Initially, the algorithm generates a mask of pixels where one or more active fires were detected. Then, this mask is filtered with a morphological erosion (Clayden, 2019). Using the same kernel defined in the Section 2.6, only pixels with neighbors where an active fire was detected are selected as burned training pixels. If active fires are detected in three consecutive pixels, the probability of detecting the central pixel as burned with the burned area mapping algorithm is greater than in other cases.

The following threshold tests are applied to the eroded burned training mask:

$$|t_f(x,y) - t^*(x,y)| \leq \Delta t^*(x,y) + 5 \quad days \qquad (11)$$

Equation (11) checks the difference between the day an active fire was detected ($t_f(x,y)$) and the estimated date where a pixel has burned ($t^*(x,y)$). The result is compared with the temporal uncertainty measure computed in Equation (8), and also five days of margin are added due to the amount of missing data.

Also, the following thresholds are computed:

$$S^*(x,y) \geq 2 \qquad (12)$$

$$\sigma_t^*(x,y) \leq 8 \qquad (13)$$

Equation (12) looks for the big change of trends in the VI time series and Equation (13) looks for areas with small values of temporal texture.

Pixels identified as burned in the morphological erosion that also satisfies Equation (11), (12), and (13) form the initial burned classification mask. We will denote these initial burned pixels as the first seeds of the algorithm. They are called seeds because they have a high probability of being burned and they are at a location where more active fires were detected in other neighbors. So these seeds are likely to be located at a big wildfire. The location of each seed forms an individual cluster (j) that will grow constantly adding more seeds if their neighbors satisfy some thresholds. All pixels labeled as not burned will not be candidates in this step.

For each cluster j, $\Delta VI^*\{j\}$ and $VI^*_{pos}\{j\}$ represent the ΔVI^* and VI^*_{pos} values of the pixels that belong to each cluster j. The cluster grows spatially if the neighboring pixels meet the following criteria:

$$\Delta VI^*(\mathrm{x,y}) \; > \; 25\text{th percentile of } \Delta VI^*\{j\} \qquad (14)$$

$$VI^*_{pos}(x,y) < 75\text{th percentile of } VI^*_{pos}\{j\} \qquad (15)$$

$$\sigma_t{}^*(\mathrm{x,y}) \; \leq \; 3 \text{ days} \qquad (16)$$

$$d_l(x,y) \leq R_g \qquad (17)$$

where $d_l(x,y)$ is the distance of the candidate pixel at (x,y) to the initial seed, and $R_g = 10$ km is established empirically (big wildfires do not have normally a radius bigger than 10 km). All pixels that meet Equations (14), (15), and (16) are included as seeds in the corresponding cluster and $\Delta VI^*\{j\}$ and $VI^*_{pos}\{j\}$ are updated with the values of the new seeds. Then, an iteration process starts adding more seeds to the clusters until no candidate is added or Equation (17) is not met. All the pixels that belong to a cluster at the end of this process form the initial burned training mask.

2.7.3 Unburned training pixels

In this section, the unburned training mask is computed again using information from the initial burned training mask, and all pixels that satisfy one of the following conditions are labeled as unburned:

1) If the pixel was labeled as unburned in the initial classification (Section 2.7.1).

2) In the second condition, we use the information from the burned training mask. All pixels that were not labeled in either of both masks and the distance ($d_B(x,y)$) to a burned pixel is bigger than $R_d = 5km$ ($d_B(x,y) \geq R_d$) are labeled as unburned.

As exposed at the beginning, those thresholds are based on empirical knowledge.

2.8 Conditional probability density functions

From now on, every step until the final classification will be conditioned on the pixels land cover class. As exposed in the introduction, the vegetation chemical and physical characteristics have a big influence on the value of the reflectance and in consequence, in the value of the index and the other calculations derived from it. The land cover product allows dealing with these different features, identifying clusters of pixels with similar physical and chemical characteristics, which means similar changes in reflectance values after a burn.

So far, the algorithm has generated two masks, one mask with burned pixels and another with unburned pixels. For each class, we derive the conditional probability density functions (PDF), $P(\Delta VI^*|B)$ and $P(\Delta VI^*|U)$, where B and U stand for burned and unburned pixels.

The assumption made here is that if the data are divided by land cover class, similar values of ΔVI^* for burned and unburned pixels will be obtained. Hence, these data are expected to have Gaussian distributions and a Gaussian kernel density estimator (KDE) is used to generate both distributions.

A KDE is a non-parametric method used to estimate the probability density function (PDF) of a random variable. The kernel is a symmetric and non-negative function that integrates to 1. It generates one distribution for each point of the sample and then sums all of them generating the final distribution (Fig. 20).

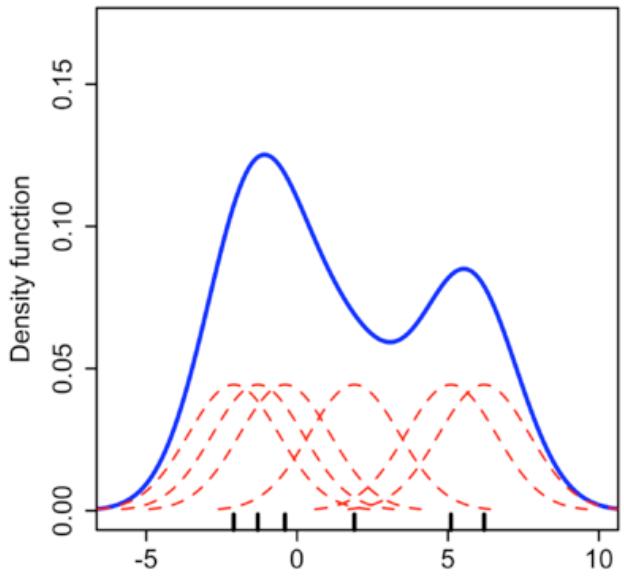

Figure 20: Gaussian kernel estimator example.

The following Gaussian KDE is used to generate the corresponding distributions:

$$P(u) = C \sum_{i} exp\left(-\frac{[u-u_i]^2}{2\sigma_k^2}\right) \qquad (18)$$

where C in Equation (18) is the normalization factor that ensures the integral of P overall u is 1. The parameter (σ_k^2) corresponds to the bandwidth which is the way to model the smoothness of the distribution. It controls the amount of data involved in each distribution of the kernel function. In this case the value of the bandwidth is $\sigma_k = 0.02$ (empirically computed).

Once we have generated burned and unburned distributions, we decide for each land cover if burned and unburned pixel distributions are separated enough to discriminate between them knowing the ΔVI^*. For each land cover, we compute the difference between the median ($Q_l^{(50)}$) of the burned and unburned ΔVI^* training samples (Eq. 19).

$$\Delta Q_l^{(50)} = Q_l^{(50)}(\Delta VI^*|B) - Q_l^{(50)}(\Delta VI^*|U) \qquad (19)$$

The pixels belonging to the class *l* are considered unburned if the difference in the median is below -0.05 or the difference in the median is below 0 and the size of the sample of the burned pixels for the class is small (<100).

These thresholds have been established knowing that the value of $\Delta Q_l^{(50)}$ is expected to be greater than 0 for normal burns, but a -0.05 margin value is allowed when burned pixels of a specific class are more than 100, to offset the cropland fires. Cropland fires (intentioned most times) are not included in the actual analysis and so that all pixels belonging to this class are automatically labeled as unburned.

2.9 Prior and posterior burned probability

In this step, the Bayes' rule is used to extract a probability mask based on the PDF distributions of burned and unburned pixels. The burned-unburmed labels are established in Section 2.7. The computation of the prior and posterior probability only consider pixels of land classes that passed the separability test. These classes have no overlap on the burned and unburned distributions of the ΔVI^* (Section 2.8).

First, a prior probability ($P_B(x,y)$) is computed considering the distance to the nearest pixels labeled as burned ($d_B(x,y)$). If the pixel was labeled as unburned, $P_B(x,y) = 0$, otherwise the prior probability is computed using the formula $exp(\frac{x^2}{2})$ as follows:

$$P_B(x,y) = (P_{max} - P_{min})exp\left(-\frac{d_B(x,y)^2}{2\sigma_p^2}\right) + P_{min} \qquad (20)$$

with $P_{max} = 0.5$, $P_{min} = 0.01$, and $\sigma_p = 2$ km. Equation (20) is an arbitrary method used in Giglio et al. (2018) based on the idea that pixels were more likely burned if they are close to a burned pixel. The value $\sigma_p = 2$ km was established empirically, knowing that it is not common to have burned pixels more separated than 2 km each other after having identified the burned training mask where most of the pixels that belong to big wildfires have been labeled as seeds in each cluster.

The probability of a pixel being burned decays when the distance to a burned pixel increases as shows Fig. 21:

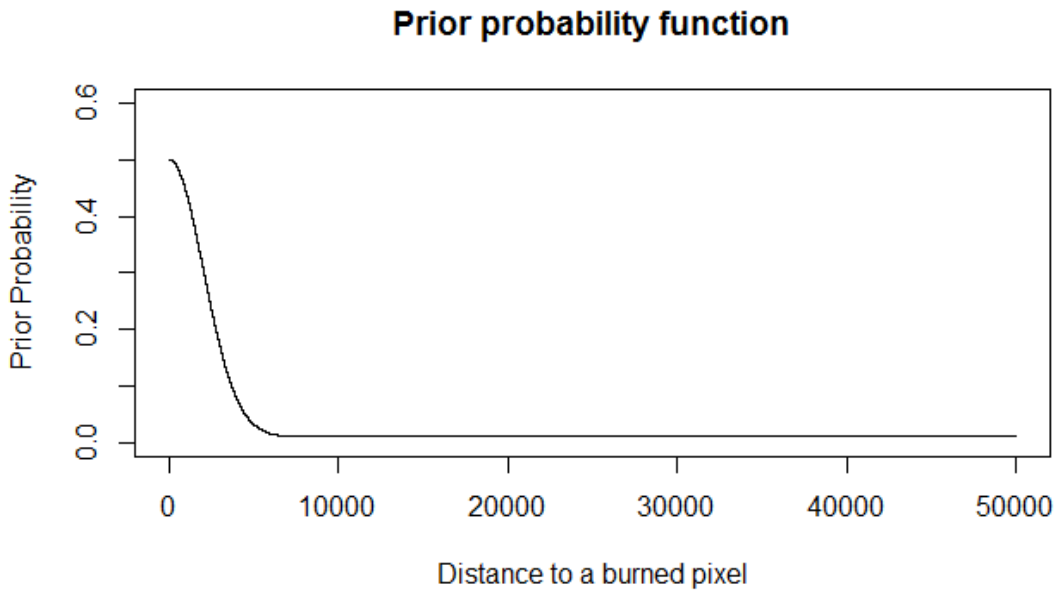

Figure 21: Prior probability function.

The prior probability values are in the range $P_{min} \leq P_B \leq P_{max}$. The probability P_B when the distance $d_B(x, y)$ increases converge to $P_{min} = 0.01$ value, and those which are near a burned pixel have $P_B \approx 0.5$. When we have the P_B value, the probability that a pixel is not burned is $P_U = 1 - P_B$.

After obtaining the prior probability and given the observed value of change in VI, the posterior burned probability is computed using the Bayes' rule in Equation (21):

$$P(B|\Delta VI^*(x,y)) = \frac{P_l(\Delta VI^*|B)P_B(x,y)}{P_l(\Delta VI^*|B)P_B(x,y) + P_l(\Delta VI^*|U)P_U(x,y)} \qquad (21)$$

This value represents the probability of an event having occurred, knowing the value of another event. In this case, the event (a specific pixel burned) has occurred, so we try to determine how strong is our belief about our initial prediction.

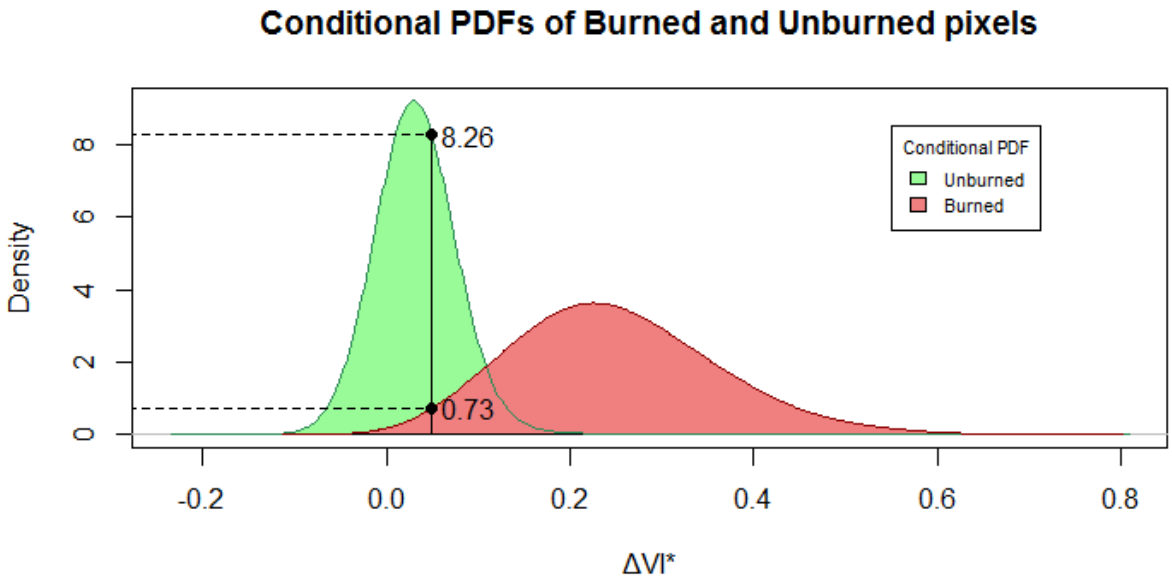

Figure 22: Distributions of burned and unburned pixels of class 2 in Rondonia in August 2019. Example about how are taken the two values of both distributions to apply the Bayes' Rule.

Pixels close to a burned pixel have a high prior probability $P_B \approx 0.5$ and that gives the weight to the ΔVI^* value when computing the posterior probability ($P_B \approx P_U \approx 0.5$). As shown in Fig. 22, if the burned and unburned prior probabilities are similar, the computation of the posterior probability is only about proportions (50-50) of the values of both PDFs depending on the ΔVI^* of the pixel.

Otherwise, in those which are far away from a burned pixel the prior probability function converges to 0.01 ($P_U \approx 0.99$), and the value of ΔVI^* will not have much influence on the posterior probability computation. At the end of this step, the algorithm generates a probability mask with the posterior probability ($P(B|\Delta VI^*(x,y))$) of each pixel (see Fig. 23).

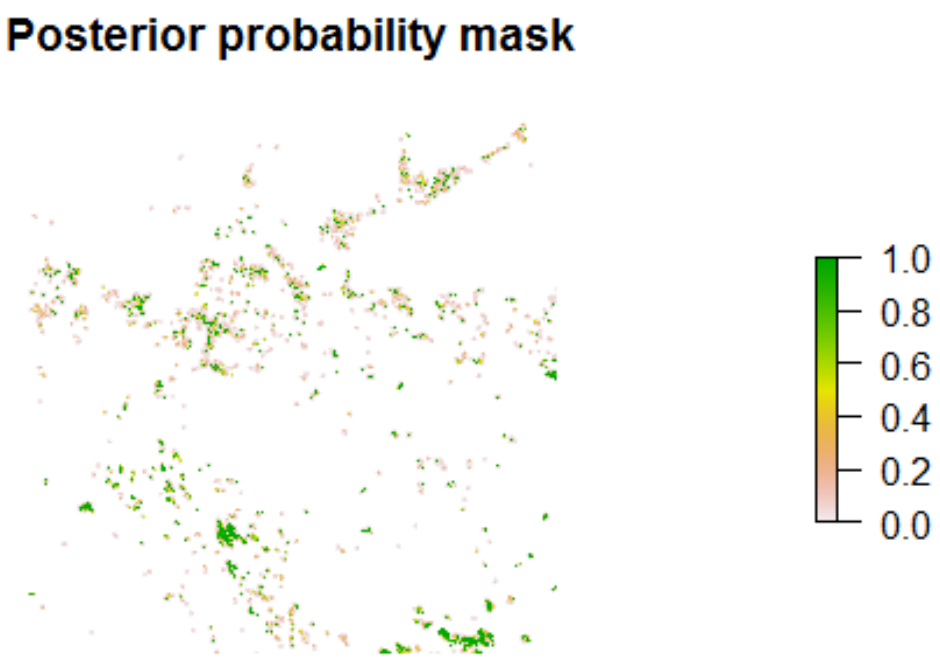

Figure 23: Posterior probability mask

2.10 Initial classification

This first classification is focused on physical characteristics. The burned training mask finds pixels where an active fire was detected and has neighbors in the same situation. Then, it looks for pixels near each other with a big change of trend in the VI time series in the same period. Finally, the pixels are divided by land cover class to deal with physical and chemical characteristics and a probability to be burned is assigned to each one depending on ΔVI^* and the distance to the nearest burned pixel.

Now, a new mask of burned - unburned pixels is generated based on this previous knowledge. Initially, all pixels classified as unburned in the training mask are labeled as unburned. Then, the following computations are conditioned to the land cover of the remaining pixels.

All land cover classes that did not pass the separability test were identified as unburned in previous steps and are equally labeled because in this initial classification all thresholds will be conditioned to the land cover class, and in those classes, there was no evidence that the burned and unburned conditional PDF were different enough.

The remaining pixels are labeled as burned if they satisfy the following thresholds:

1) If the posterior probability $P(B|\Delta VI^*) \geq 0.5$
2) If the $VI^*_{pos}(x,y) \leq Q_l^{(98)}(VI^*_{pos}|B)$
3) If $\sigma_l^*(x,y) \leq Q_l^{(98)}(\sigma_l^*|B)$

After this, a new mask is generated with burned and unburned pixels forming the initial classification

2.11 Final classification

This final classification looks for spatial coherence in burned areas. Two-pixel categories have been defined: burned (n_B) and unburned (n_U). In this step, another category is defined as tentative burned pixel n_{CB}. We take every burned pixel and its neighbors with the kernel introduced in Section 2.6 and look at their days of maximum separability (t^*). The number of neighbors which days of maximum separability are in a window of 21 days centered in the t^* of the central grid ($|t(x,y) - t(x',y')| \leq 10$) are counted and recorded (this condition relaxes the criterion of equation (13)).

Then, a cumulative density function of burned neighbors of burned pixels is computed, which means the number of burned neighbors that a burned pixel has. It is mathematically written as follows:

$$F(n_B|B) = P(N_b \leq n_B|B) = \sum_{N=0}^{n_B} P\,(N_b = N|B) \qquad (22)$$

In the final classification burned pixels will be relabeled as unburned if they have more unburned than burned neighbors ($n_B < n_U$) and if the cumulative probability is $F(n_B|B) < 0.1$.

On the other hand, unburned pixels are relabeled as burned if the number of burned neighbors is greater than the number of unburned neigboods ($n_B < n_U$) and $n_{CB} \geq 1$.

This will be the final result of the whole algorithm, which is a mask of burned-unburned pixels of the region of interest. In this study, the region of interest is Rondonia, and the burned area mapped in August 2019 is represented in Fig. 24.

Finally, we calculate the burned area for the month, multiplying the area of each pixel by the number of pixels. The area of each pixel of the images of this study is colloquially referred as “500 m”, but the real pixel size is 463.3127 m (Giglio et al, 2015). Thus, the area of each pixel is 0.2147 km^2.

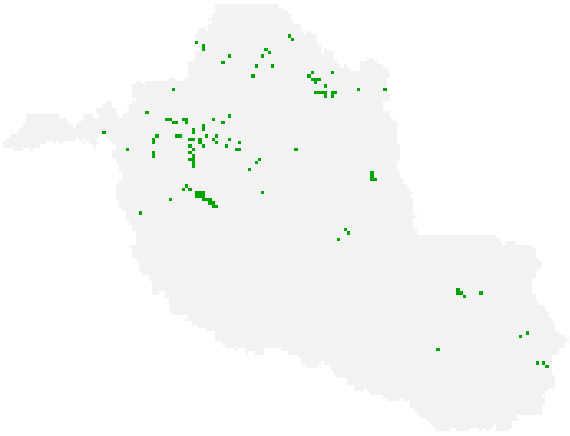

Figure 24: Burned area in Rondonia in August 2019. The resolution of the image misrepresents the real burned area.

3 Spatial coherence approach.

As many steps are conditioned to the land cover due to the influence of the chemical and physical characteristics on detecting BA, this new approach is oriented to compute the burned area of specific regions. Small areas are usually covered by more than one tile. That is the case of Rondonia, the Brazilian state that belongs to the Amazon biome.

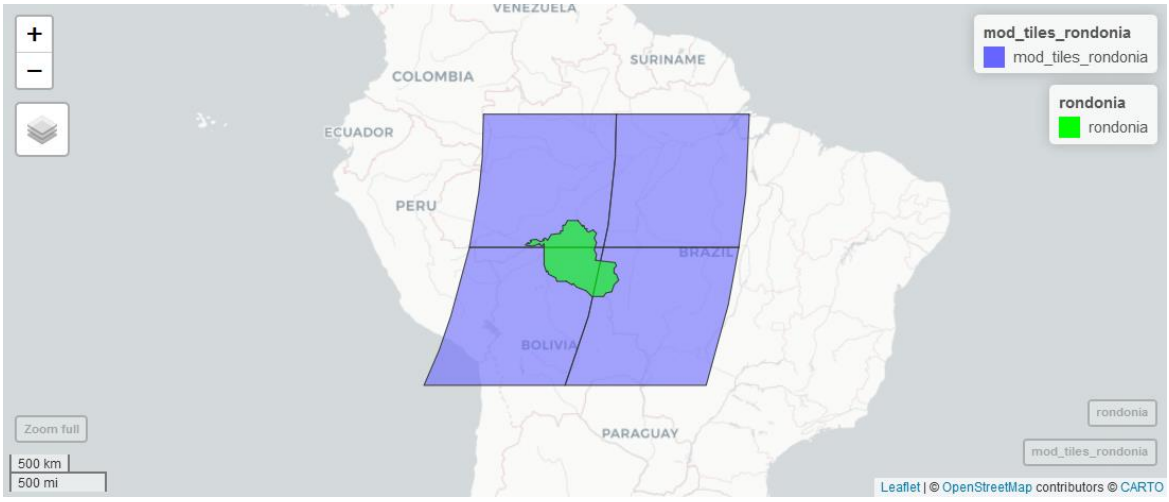

Figure 25: Tiles needed to generate a rectangular shape that contains Rondonia. The top right corresponds to h:11 v:09 tile, the top left to h:12 v:09 tile, the bottom left to h:11 v:10 tile and the bottom right to h:12 v:10 tile.

Rondonia is covered by 3 tiles, but we work with rectangular images and the four tiles of the Fig. 25 are needed to create a rectangular shape that contains this region.

The aim of this work is not to explain the physical and chemical basis of the land cover conditioning, and then, we do not intend to detail all land cover classes. In Table 2 (Appendix 3) a short explanation of each class is presented. It has been extracted from the user guide of the land cover product (MCD12Q1) (Friedl et al. 2015).

We are just looking for the proportion of each class in each region or tile. At this point, we only need to know that same land cover means similar reflectance values. Fig. 26 shows the proportion of pixels in each land cover class for each tile shown in the Fig. 25 (considering the same tile order).

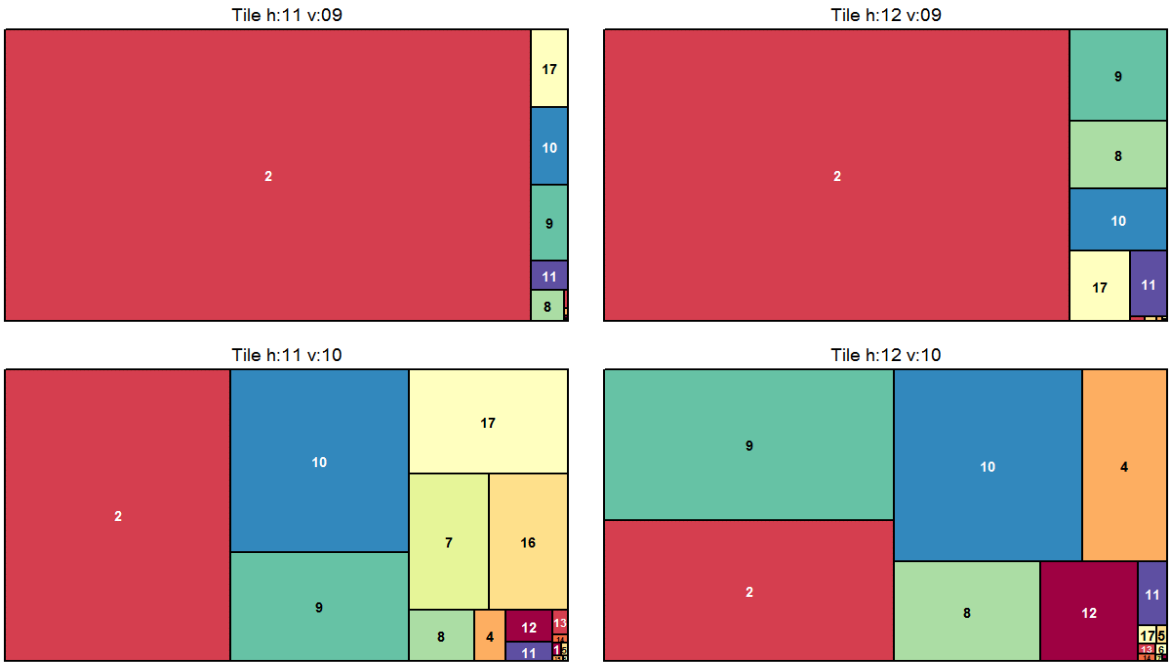

Figure 26: Land cover of the four tiles that cover Rondonia. Class cover ID numbers are those in the MCD12Q1 product and are specified in Appendix 3.

Many of the classification steps work with conditional probability densities. Those are estimated by KDE as introduced in Section 2.8, which are influenced by the amount of data available to generate each class probability density. The proportion of pixels of each class varies in the four tiles. This is why some classes with a small proportion of pixels will generate poor distributions and will be labeled as unburned by the separability test.

But if we apply the BA mapping algorithm to the rectangular shape that contains Rondonia we are underestimating the burned area because the algorithm is supposed to be applied to a tile size image (2400x2400 pixels) and Rondonia's rectangular shape has only 58% of data and that generates poor distributions as we have less training data. For getting more accurate results we will expand this region from the centroid of Rondonia in the four directions until we have a tile size image (2400x2400 pixels). The advantage of this method is the similarity in the proportion of pixels of each class in Rondonia and in the new region called "Rondonia tile size extended".

Fig. 27 demonstrates that Rondonia and Rondonia extended have a similar proportion of pixels of each class and the new image will generate good burned-unburned distributions of the most important classes. In this particular case, the Rondonia state is in the border of the Brazilian Amazon Biome (next to Cerrado's biome), which also gives more bias to the results as there is an important difference in the characteristics of the vegetation in the four tiles, as Fig. 26 shows. This

particularity will be discussed in the conclusions once the results have been exposed.

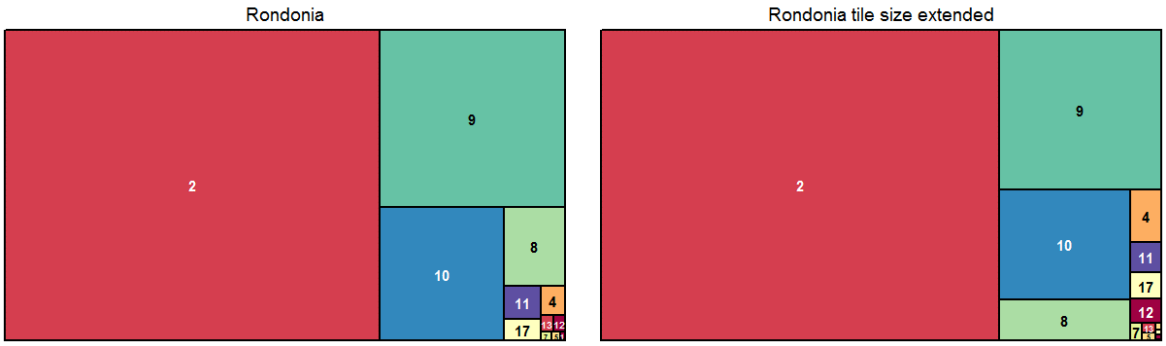

Figure 27: Land cover of Rondonia shape and Rondonia extended to tile size shape

4 Results

4.1 Temporal scope of the study

In Section 1.2 August and September were identified as the most affected fire months historically. Also, in August 2019 INPE and NASA reported an important increase in active fires (see Fig. 31), so this month is selected as the one to be studied.

But the performance of an algorithm cannot be tested only in a single month, so we are going to study every August in the period 2011-2019.

4.2 Exploratory data analysis.

4.2.1 Vegetation index.

For every year, we have the vegetation index for each day from July 1st to September 31st generated using both Aqua and Terra reflectance data. Initially we have 184 images that correspond to the VI_{Aqua} (92 images) and VI_{Terra} (92 images). Then, we combine both as explained in Section 2.4 to obtain the 92 final VI images. These data correspond to the region “Rondonia tile size extended” (Section 3) and can be represented as a three-dimensional array (images ordered by day) Fig. 28:

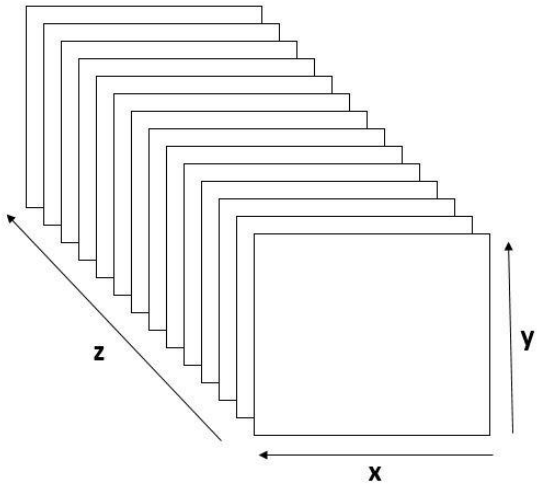

Figure 28: Satellite image time series scheme

Here, x and y correspond to the location of the pixels and z corresponds to the day in the time series. In the actual case, the resolution of an image is 2400 x 2400, which means that every image has 5760000 pixels. Thus, every year we are going to study trend changes in 5760000 time series. The 3-month time window condition means that each time series has 92 days. But the biggest drawback is the missing values due to the presence of clouds in the satellite images and this is the reason

why we use 3 months data to compute the BA for only one month, as explained below.

In all data analysis preprocess, lost values are one of the main issues. In this case, the presence of clouds causes almost 40% of lost values in the dataset. As explained above, we are looking for trend change in the VI index in a time series of 92 days for each pixel. If there is no fire in a pixel, the value of the VI index will remain almost constant. If there is a fire, the VI value will decrease and later increase during a recovery phase. The speed of recovery depends on the land cover, the strength of the fire and other characteristics.

We are not going to replace lost values, but instead, we will use only real valid data and we will model the gap of this valid data at the date the fire is detected. To validate the use of Equations (7) and (8) in different steps, we define two random variables. The first one is the discrete random variable X: *Number of valid data in a time series of 92 days(July-September) in "Rondonia extended to tile size" region*. It uses data from the period 2011-2019. The probability mass function of X can be seen in Fig. 29.

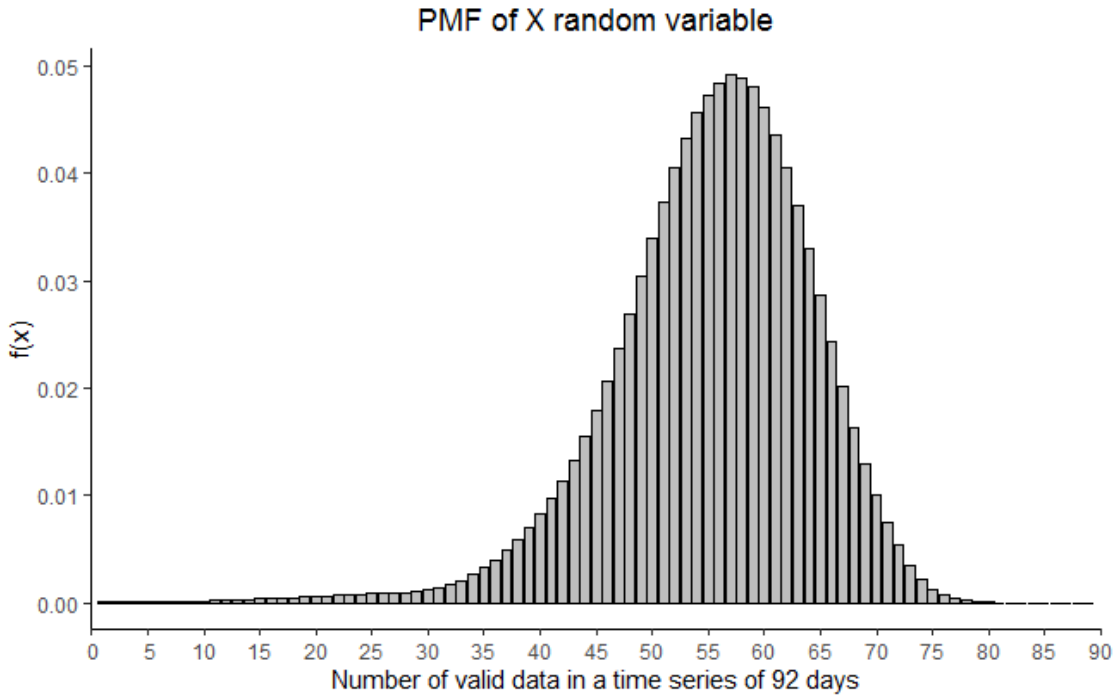

Figure 29: Probability mass function of the random variable X.

This random variable has the following features:

- It is discrete and it only can take values in the interval [0,92].

- It counts the number of times an event (presence of valid data) occurs in a period (time-series of 92 days).
- It has no upper bound, (example: if on the day 80 the number of valid pixels is 50, in the day 81 it can be 51).
- The expected value is E[X] = 55.27.

With the PMF and these features, we can assume that X follows a Poisson distribution with parameter $\lambda = E[X] = 55.27$, which is the expected number of valid data in this region and period.

But, the distribution in time of this valid data is also important. Therefore, we define the second random variable Y: * the maximum number of consecutive missing data in a time series of 92 days(July-September) in “Rondonia extended to tile size” region*. Probability mass function of Y is shown in Fig. 30.

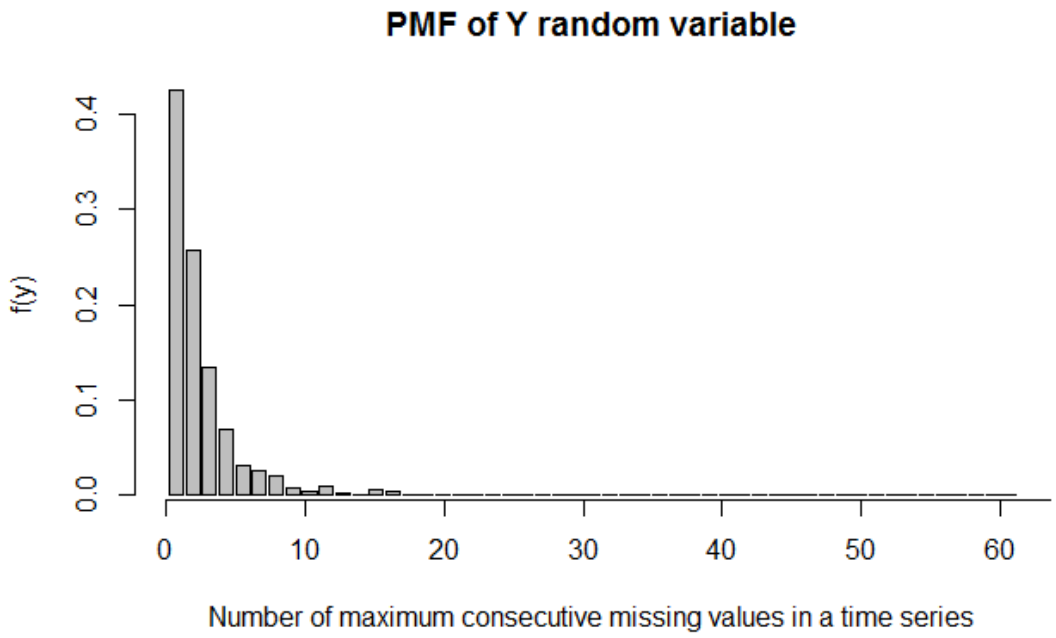

Figure 30: Probability mass function of the random variable Y.

The expected value of this random variable is $\lambda = E[x] = 2.47$ days. . This means that if there is a gap in the time series between two pixels with valid values it is expected to be equal or smaller than 2.47 days.

With these two random variables, we can validate Equations (7) and (8), and their use throughout the algorithm. The expected number of valid pixels in the time series is 55.27, and the expected maximum separation between them is 2.47, which means

the valid data are evenly distributed throughout the time series, and so that the uncertainty of a 2-3 pixels gap can be easily modeled.

4.2.2 Active fires

Active fire data allow us to set up the first seeds of the initial burned classification mask. At this point, a logical relationship can be established: the more active fire points are detected, the more burned area is mapped. It is an indicator of the increase or the decrease of the burned area that will take more time to be estimated as the algorithm needs a 3-month time window.

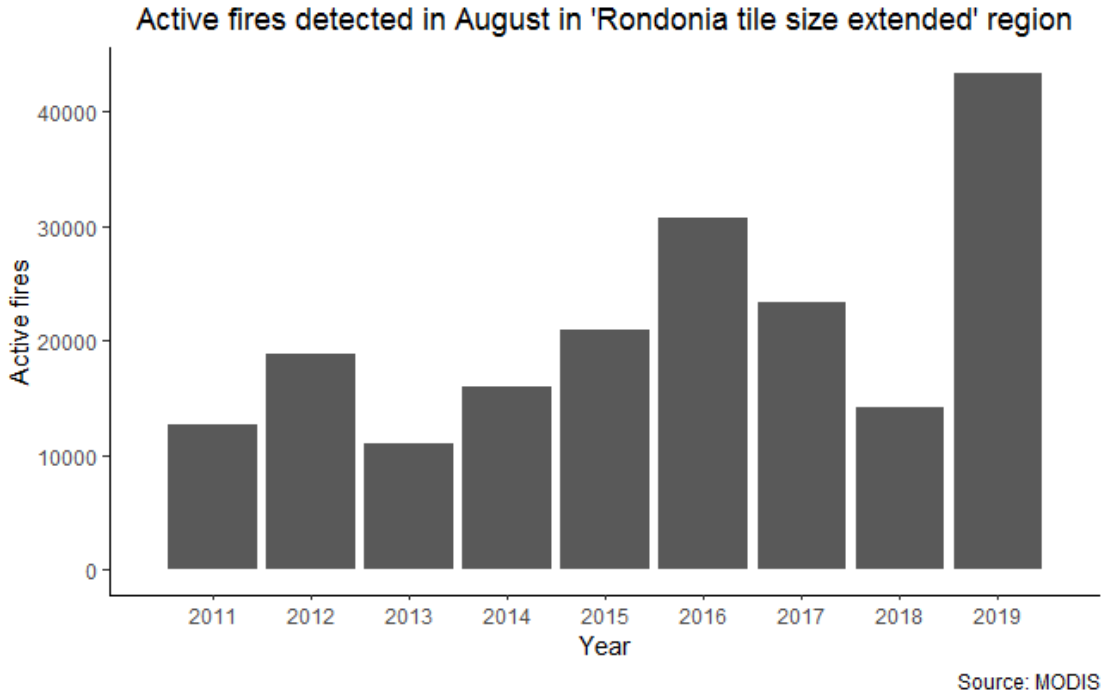

Figure 31: Active fires detected in the region of interest in August.

The increase in active fires detected in August 2019 (Fig. 31) is remarkable compared to previous years, which a priori will involve more burned area mapped.

In Section 1.2 Rondonia was selected as the region of interest because it was the state with the highest active fire density of the Amazon biome. This decision along with the spatial coherence approach will test the performance of the algorithm in regions with a high active fire density that are divided (by 3 in this case) in the MODIS tile system.

4.2.3 Land cover

The main idea about land cover data has been exposed in Section 3. Every year the land cover algorithm generates a new mask based on the changes in reflectance associated with different physical and chemical characteristics.

The time-period of this study goes from 2011 to 2019. The amount of pixels relabeled to other classes in this period is not relevant in this kind of analysis because the percentage is very small, and it does not adversely affect the accuracy of the results.

4.3 Comparation with NASA's and INPE's burned area products

In this section, the results obtained by applying the algorithm are exposed. A comparison of these results with those obtained by NASA and INPE is also presented.

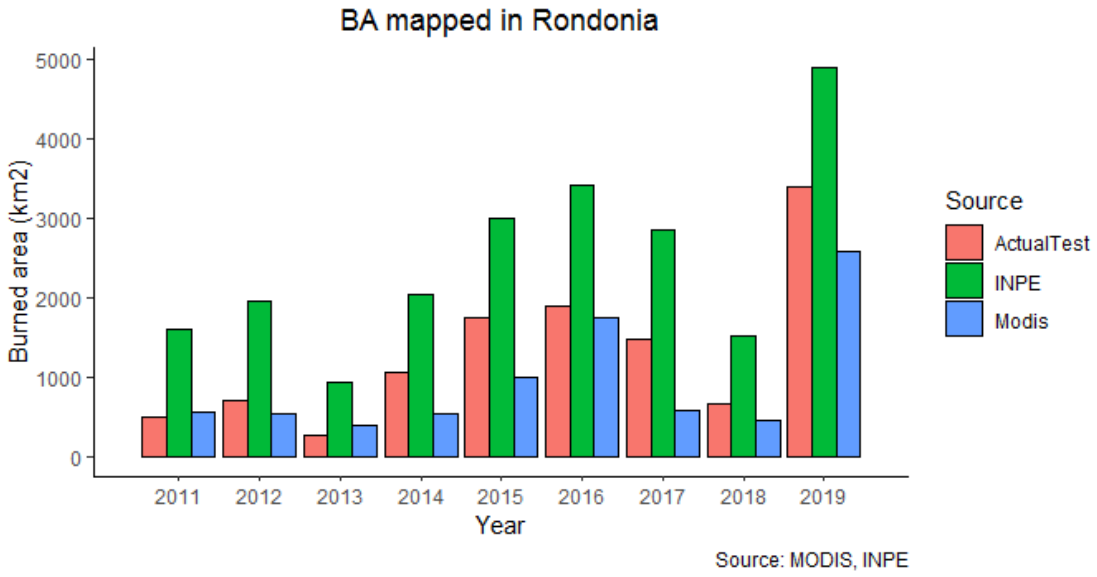

Figura 32: Burned Area detected by 3 algorithms

As Fig. 32 shows, INPE's algorithm detected 3.06 times more burned area (on average) than the MODIS one in Rondonia between 2011 and 2019. It uses other physical features (the red light and near-infrared light of the solar domain) and other classification methods based on tables developed by the authors (Libonati et al., 2015). Also, it is in stage 2 of maturity in contrast to MODIS which is at a higher level. For these reasons, we will focus on comparing only both MODIS approaches.

The idea of spatial coherence was presented to improve the burned-unburned distributions of the principal land covers, and in consequence, to generate more

accurate results of the burned area mapped. Also, as Fig. 3 shows, Rondonia is one of the most affected regions by active fires in August 2019 in the Amazon biome. We are going to test the performance of the MODIS algorithm when we have a region that is divided into 3 different tiles in the original approach (Fig. 25) and it is highly affected by the fires (Fig. 3).

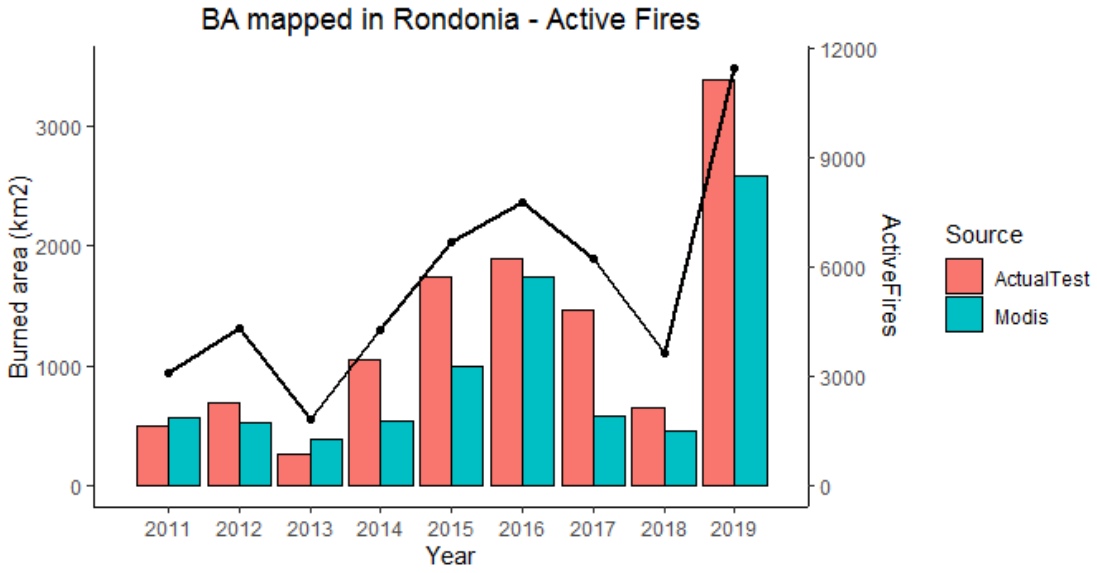

Figure 33: Burned Area compared with MODIS data (and active fires detected). The right axis is referred to the number of the active fires detected and the black line is the active fires detected in August of each year by MODIS.

In Fig. 33 we can appreciate the assumption made above that the more active fires are detected, the more burned area is mapped. Both approaches draw a similar shape (also similar to the active fire line), but with some particularities.

The first particularity can be seen in years with a small impact on active fires. This idea was presented in Giglio et al. (2018), where they remark that active fires in tropical and subtropical latitudes are under-represented due to the overpass frequency of the Terra and Aqua sensors. The solution proposed in this paper to deal with this problem is to extend the region being studied to obtain more active fires detected and have more training data. They add to the region of interest (a tile) more pixels that belong to other tile neighbors to obtain more training data.

In this study the extension of the region of interest to obtain more training data in years with less fire activity is not applied. Therefore, in years with less fire activity (2011 and 2013), the burned area is underestimated compared to the MODIS burned area product (MCD64A1).

Giglio et al. (2018) paper does not detail what is the procedure used to deal with this issue. Also, the extension of the region of interest means more data to be processed,

which derives on an increase of the computational cost. As can be seen in Appendix 1, the computational cost without adding that modification is too big and we cannot assume a new increase of the data to be processed with the available resources.

Nevertheless, these results also show that the BA detected using this new approach is 1.43 times bigger than the MODIS tile method when the region of interest is highly affected by active fires and also the BA mapping of this region is involved in different tiles. Fig. 31 indicates NASA's algorithm underestimates the burned area in years with a medium-high incidence of active fires (2012, 2014, 2015, 2017, 2018, and 2019) using the tile system.

This increase shows the misuse of the tile system, a rigid spatial application of the algorithm. Active fires are the most important input data. They are used to generate the burned training mask in Section 2.7.2. The more burned training pixels are detected in section 2.7.2, the more separated burned and unburned distributions are generated in section 2.8. But active fires occur at different locations each year. Before the spatial application, the algorithm should study which regions have high active fires density (Fig. 3) and look at those that are divided into more than two tiles (Fig. 25). These regions are at a disadvantage and they should not be considered in the regular computation (tile system). They should be computed as exposed in this study. This correction factor should be used as a complement to the tile system that shows poor behavior in regions that are at a disadvantage because of their location and their fire activity is not considered.

5 Conclusions

In this study, the NASA's algorithm developed by Giglio et al. (2018) has been reproduced in R code as faithfully as possible. Initially Rondonia has been selected as the region of interest based on the active fires density in August 2019. The active fire density and the land cover coherence (Section 3) of the region being studied are key in the approach followed in our study.

The first step of the algorithm is the computation of the Vegetation index (VI) (Section 2.4). VI is used to identify a reflectance change in time that can be associated with a fire. Then the burned and unburned training data are selected based on active fires (Giglio et al. 2016) and the information extracted from the VI index (Sections 2.4, 2.5, 2.6 and 2.7).

The training data are used to generate burned and unburned distributions (Section 2.8). But the reflectance values are different depending on the land cover of the Earth's surface. Therefore, burned and unburned distributions are generated separately for each land cover. The separability test verifies if the distributions are adequate to discriminate burned and unburned pixels. Then, the burned and unburned distributions are used to generate the so called posterior probability mask (Section 2.9). The value of each pixel of the posterior probability mask represents the probability of having burned based on all the previous physical and temporal knowledge.

Finally, the algorithm has two classification phases. The first one is based on the posterior probability mask (physical and temporal knowledge) (Section 2.10). The last one is based on the spatial coherence (Section 2.11). Pixels are relabeled (or not) to the opposite class (burned-unburned) depending on the label of their neighbors.

The most important change introduced in this study is related to the spatial application of the algorithm, which is focused in a region with a high active fires density that is divided into three parts in the MODIS tile system (Rondonia).

The results obtained with this new approach indicate that the burned area detected in this region is underestimated in years with high active fires density. In our study, the burned area is 1.43 times (on average) the area mapped by the MODIS product. This underestimation highlights the limitations of the tile system, because it is an algorithm based on active fires and it does not consider the density of the active fire for its spatial application. Also, MODIS product MCD64A1 is not a reliable source of information for region/states similar to Rondonia that are clearly at a disadvantage due to their location. It is on the border of the biome and divided in three parts by the tile system in the original spatial application of the algorithm.

In Section 4.3 we propose the application of a correction factor to fix this underestimation. The correction factor is based on the modification of the spatial application of the algorithm in regions with high fire activity density (Fig. 3) and divided into more than 2 parts in the tile system (Fig. 25). As Giglio et al. (2018) is a probabilistic algorithm, the regions computed using this new approach should be

removed in the original approach and replaced by the results obtained with the new approach presented in this study.

But some non-detailed steps in the original literature lead to assumptions like the kernel selected at Section 2.6. The methodology used to generate this kernel in each region is not explained and we assume the use of the kernel located above South America. Also, in tropical and subtropical regions, active fires are under-represented, and therefore, the burned area is underestimated. This issue is not rectified, but it is taken into account when explaining the final results. Finally, it is a physics-based algorithm that relies on empirical knowledge. The results and conclusions obtained in this work rely on them.

Despite the lack of information explained in the previous paragraph and the limitations of the code to be used to calculate burned areas in other regions, R is a great source to solve problems such as the computation of burned areas. There are many available spatial data packages to download and process satellite data (for example (`RGISTools`) and to analyze and model all kinds of spatial data (`raster`, `sp`, `spatstats`...). This makes R an adequate open-source tool to be used for spatial data analysis.

This Master's project deals with topics related to statistical modeling not covered in the master's subjects (like the analysis of a huge spatial data set of terabytes magnitude and 40 % of missing values), laying aside machine learning algorithms and combining probabilistic, statistical and physical knowledge to determine an event that has already occurred.

6 Appendix 0: Quick review of scripts/functions developed in R code

This appendix introduces the functions and scripts personally developed for the analysis. The definitions are intended to help following the R code related to this thesis.

ImagesDownload: It uses functions from the `RGIStools` package to download, extract, and mosaic the images of the MODIS products (see section 2.2 for more details about the products).

Main: It contains the core of the algorithm that uses auxiliary functions to obtain the BA.

The auxiliary functions:

activeFires and **getMonth**: The first one loads the shapefile of the active fires into a `SpatialPointsDataFrame`, and the second one selects only the fires that occurred in the month been studied.

cover_VI: It creates a composite image of the index VI using VI_{Aqua} and VI_{Terra}, and saves it to a file.

FinalClassification (c): It generates the final BA map of the region of interest (the result of the algorithm).

FinalTrainingMask: It grows the clusters of the seeds that form the initial burned training mask.

getMaxDay (c): It returns the value of the specified position in a time series.

incertidumbre (c): It returns two images with the values of the equations (7) and (8).

maxSeparabilityDay (c): It detects the maximum in the separability index time series.

meanSdQuantile (c): It computes the mean and the standard deviation of a vector after having removed values smaller than the 10th percentile and bigger than the 90th percentile.

preVIcals: It generates two images of each day of the month been studied, one with the mean of VI index of the pre-burn window of each pixel and other with the standard deviation (equations (2) and (3)).

posVIcals: It generates two images of each day of the month been studied, one with the mean of VI index of the post-burn window of each pixel and other with the standard deviation (equations (4) and (5)).

probMask_generate: It generates the probability mask used for the initial classification applying the Bayes' rule and the PDFs of burned and unburned masks generated by the KDEs.

reflectanceCleaner: It removes invalid data from the reflectance bands 5 and 7.

separability: This function identifies a change of trend in a time series of the VI index (equation (6)).

TempConsistent: It gets the number of neighbors that are temporally consistent with the criteria explained at the final classification.

varVI: This function computes the VI index and saves it to a file.

The functions followed by (c) have been developed as an argument of the calc function of the raster package.

7 Appendix 1: Computational cost

The analysis was conducted with an Intel(R) Core(TM) i7-6500U CPU @2.5GHz and 8GB of RAM memory. The computational cost of the BA mapping process divided by section and the most expensive functions applied in each one is exposed in this appendix:

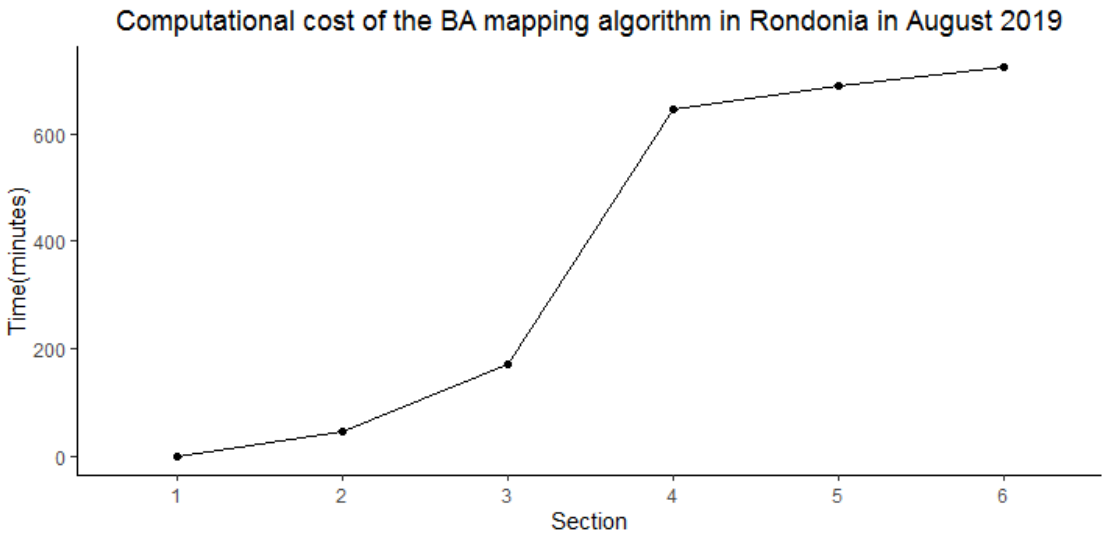

Figure 35: Cumulative computational cost

1) MODIS images download:

The computational cost of this part is not included because it depends on the internet connection speed. The size of the files in the HDF compressed format of all products used in this work is 415 GB.

2) Satellite images (extraction and mosaic):

The extraction of bands 5, 7, and QA of the images and tiles downloaded and mosaic the images getting only the region of interest. It represents 6.34 % of the total cost of the algorithm.

3) VI index:

This section is related to the computation of the indexes VI_{Terra} and VI_{Aqua} and the combination of both to generate the final composites of VI. It represents 17.36 % of the total cost of the algorithm.

The function `projectRaster` of the raster package used to change the resolution of the cloud mask from 1000 m to 500 m (the resolution of the VI index) is the most expensive one. It is applied 2 times, one to each cloud mask of the VI indexes (VI_{Terra} and VI_{Aqua}). However, using the function `beginCluster` the computations

are sent to 3 different cores (in this particular case with the available resources) and the computational cost of this section decreases 46 % (getting the percentage presented before (17.36 %)).

4) Separability:

It is the one with the highest computational cost of the entire algorithm (calculation of equations (2), (3), (4), (5), and (6)). The compute of the 10th and 90th percentiles of 345.600.000 windows of length 8 are the reason why this section represents 65.43 % of the total cost of the algorithm.

5) Composite images derived from the day of maximum separability:

It computes the equations (7), (8), the temporal texture σ_t^* and the composite images $S^*(x,y)$, $\Delta VI^*(x,y)$, and $VI_{pos}^*(x,y)$ of the day of maximum separability for each pixel. It represents 5.79 % of the total cost of the algorithm.

The most expensive computation is the *temporal texture*, σ_t^*, using the `focal` function of the `raster` package that represents 33.33 % of this section.

6) Burned-Unburned training masks and classification:

The resting computes of the algorithm uses only single images (not time series as the previous sections), and that means less computational cost. This section represents 5.10 % of the total cost of the algorithm, but it is the only one that shows variations over the years, depending on the number of active fires detected the month being processed.

In this final section, the number of training pixels (active fires) influence the running time of the algorithm, and particularly in two functions: the `distance` function of the `raster` package that calculates the distance of the *NA* pixels to the nearest one that has valid data and the own developed function called `FinalTrainingMask` which grows the cluster of each seed in the initial burned training mask.

8 Appendix 2: Structure of the database generated

For each year the algorithm generates a folder with the name of the year that has the following structure:

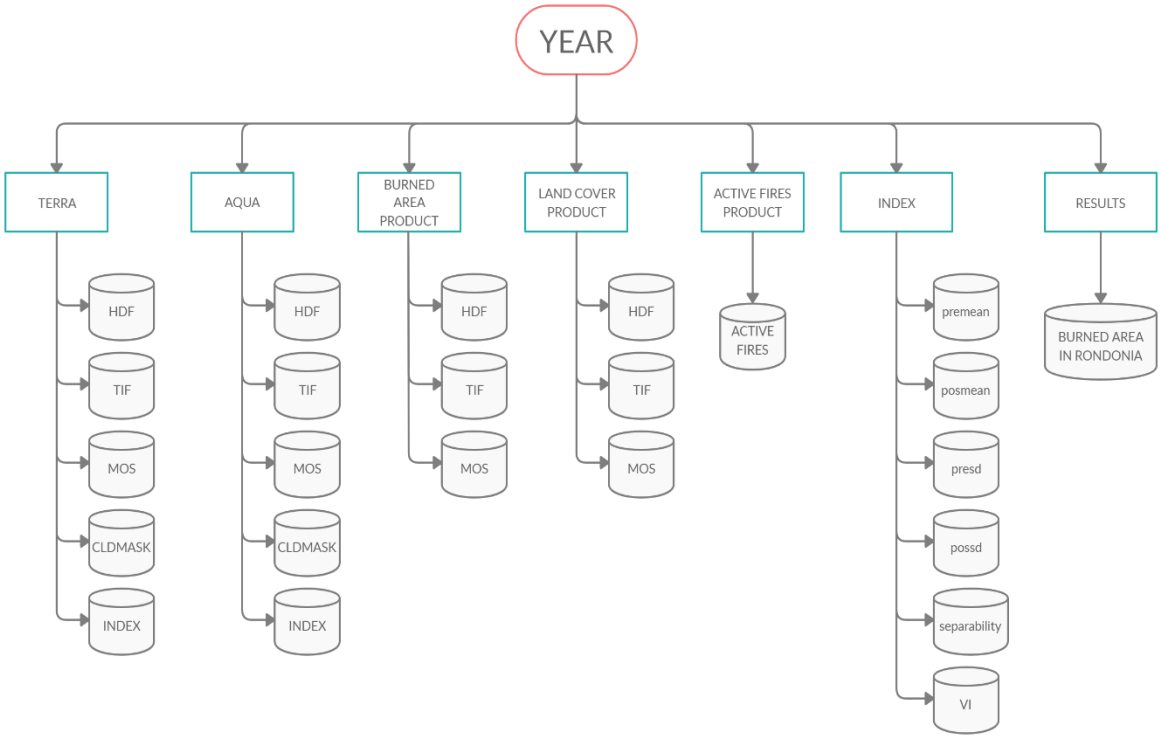

Figure 36: Database structure

0) The file format of MODIS' products explanation:

MODIS products are available on NASA's website and can be downloaded in the HDF compressed format. Using `RGIStools`, the tiles covering the region of interest are downloaded for the established period. Then, the satellite images are extracted to *.tif* format and finally, the mosaic of the region of interest is generated and form the database of this product that will be used by the algorithm.

In Fig. 36, all subfolders named HDF contain the compressed satellite images of a certain MODIS product, TIF subfolders own the *.tif* images of the tiles extracted from the corresponding *.hdf* files and MOS subfolders own the mosaicked images of the region of interest in *.tif* format.

SUBFOLDERS STRUCTURE

1) TERRA and AQUA:

Both folders have the same structure with different data obtained from Terra or Aqua satellites.

- HDF, TIF, and MOS: the surface reflectance MODIS product is downloaded, and bands 5, 7, and QA (the quality band with cloud presence information) are extracted and processed until obtaining the *.tif* image of the region of interest of the specified time series.
- CLDMASK: QA band information is processed and saved to delete pixels covered by clouds.
- VI: the VI index is computed and saved cloud-free for each satellite data.

2) Burned Area Product:

The burned area MODIS product is downloaded and processed as explained in the initial section of this appendix.

3) Land Cover Product:

The land cover MODIS product is downloaded and processed as explained in the initial section of this appendix.

4) Active fires:

This subfolder contains the shapefile of the active fires near-real-time MODIS product of the specified time series.

5) Index:

The VI subfolder contains the composites created from the TERRA and AQUA VI indexes.

The following computations are saved because of their computational cost in this subfolders:

- premean: results of the equation (2).
- presd: results of the equation (3).
- posmean: results of the equation (4).
- possd: results of the equation (5).
- separability: results of the equation (6).

6) Results:

The results obtained at the end of the algorithm explained in this work are saved in this subfolder in *.tif* format.

9 Appendix 3: Land cover classes description

In this appendix, the land cover classes' explanation extracted from the MCD12Q1 product user guide is presented (Table 2).

Table 2: Land cover classes extracted from MCD12Q1 product user guide

Value	Description
1	*Dominated by evergreen conifer trees (canopy >2m). Tree cover >60%.*
2	*Dominated by evergreen broadleaf and palmate trees (canopy >2m). Tree cover >60%.*
3	*Dominated by deciduous needleleaf (larch) trees (canopy >2m). Tree cover >60%.*
4	*Dominated by deciduous broadleaf trees (canopy >2m). Tree cover >60%.*
5	*Dominated by neither deciduous nor evergreen (40-60% of each) tree type (canopy >2m). Tree cover >60%.*
6	*Dominated by woody perennials (1-2m height) >60% cover*
7	*Dominated by woody perennials (1-2m height) 10-60% cover.*
8	*Tree cover 30-60% (canopy >2m).*
9	*Tree cover 10-30% (canopy >2m).*
10	*Dominated by herbaceous annuals.*
11	*Permanently inundated lands with 30-60% water cover and >10% vegetated cover.*
12	*At least 60% of area is cultivated cropland.*
13	*At least 30% impervious surface area including building materials, asphalt, and vehicles.*
14	*Mosaics of small-scale cultivation 40-60% with natural tree, shrub, or herbaceous vegetation.*
15	*At least 60% of area is covered by snow and ice for at least 10 months of the year.*
16	*At least 60% of area is non-vegetated barren (sand, rock, soil) areas with less than 10% vegetation.*
17	*At least 60% of area is covered by permanent water bodies.*
255	*Has not received a map label because of missing inputs.*

10 References

Baddeley A., Rubak E., and Turner R., (2015). Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press. URL http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/

Baier, S., (2012). What is Full Spectrum Lighting? Last visit: June 1, 2020. https://lumenistics.com/what-is-full-spectrum-lighting/

Bivand R., Keitt T., and Rowlingson B., (2019). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.4-8. https://CRAN.R-project.org/package=rgdal

Clayden J., (2019). mmand: Mathematical Morphology in Any Number of Dimensions. R package version 1.5.4. https://CRAN.R-project.org/package=mmand

Chuvieco, E., et al., (2019). Historical background and current developments for mapping burned area from satellite Earth observation. Remote Sens. Environ., 225, pp. 45-64; doi:10.1016/j.rse.2019.02.013

Cover, M. L., and Change, L. C., (1999). MODIS land cover product algorithm theoretical basis document. ATBD version, 5, 1-72.

Directorate-General for Research and Innovation of the European Commission, (2018). Forest Fires: Sparking firesmart policies in the EU. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/info/publications/forest-fires-sparking-firesmart-policies-eu_en

Escuin, S., Navarro Cerrillo, R., and Fernández, P.. (2008). Fire severity assessment by using NBR (Normalised Burn ratio) and NDVI (Normalised Difference Vegetation Index) derived from LANDSAT TM/ETM images. International Journal of Remote Sensing - INT J REMOTE SENS. 29. 1053-1073; doi:10.1080/01431160701281072.

Friedl, M., Gray, J., and Sulla-Menashe, D. (2015). MCD12Q2 MODIS/Terra+Aqua Land Cover Dynamics Yearly L3 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC. (User guide) https://doi.org/10.5067/MODIS/MCD12Q2.006

Giglio, L. , Boschetti, L., Roy, D.P., Humber, M.L., and Justice, C.O., (2018), The collection 6 MODIS burned area mapping algorithm and product. Remote Sens. Environ., 217, pp. 72-85; doi:10.1016/j.rse.2018.08.005

Giglio, L., Schroeder, W., and Justice, C. O., (2016). The collection 6 MODIS active fire detection algorithm and fire products. Remote Sensing of Environment, 178, 31-41; doi:10.1016/j.rse.2016.02.054

Giglio, L., Justice, C., Boschetti, L., and Roy, D. (2015). MCD64A1 MODIS/Terra+Aqua Burned Area Monthly L3 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC. (User guide) doi:10.5067/MODIS/MCD64A1.006

Wicham, H., Francois, R., Henry, L. and Müller, K. (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

Hitchcock, H., and Hoffer, R., (1974). Mapping a recent forest fire with ERTS-1 MSS data. Remote Sens. Earth Resour. 3, 449–461

Hoelzemann, J. J., Schultz, M. G., Brasseur, G. P., Granier, C., and Simon, M. (2004). Global Wildland Fire Emission Model (GWEM): Evaluating the use of global area burnt satellite data. Journal of Geophysical Research: Atmospheres, 109(D14).

Hubau, W., (2020). Asynchronous carbon sink saturation in African and Amazonian tropical forests. Nature 579, 80–87; doi:10.1038/s41586-020-2035-0

Kelly, L., and Brotons, L., (2017). Using fire to promote biodiversity. Science 355, 1264–1265; doi:10.1126/science.aam7672

Libonati, R., DaCamara, C.C., Setzer, A.W., Morelli, F., and Melchiori, A.E., (2015). An Algorithm for Burned Area Detection in the Brazilian Cerrado Using 4 μm MODIS Imagery. Remote Sens. 7, 15782-15803; doi:10.3390/rs71115782

Lizundia-Loiola, J., Pettinari, M.L., and Chuvieco, E., (2020). Temporal Anomalies in Burned Area Trends: Satellite Estimations of the Amazonian 2019 Fire Crisis. Remote Sens. 12, 151, doi:10.3390/rs12010151

Militino, A.F., Moradi, M., Ugarte, M.D. (2020). On the Performances of Trend and Change-Point Detection Methods for Remote Sensing Data, Remote Sens., 12, 1008; doi:10.3390/rs12061008

NASA, (2003). What is the Product Maturity Status? MODLAND Science Team, Last visit: March 15, 2020, https://landweb.modaps.eosdis.nasa.gov/cgi-bin/QA_WWW/newPage.cgi?fileName=maturity

Pebesma, E.J., and Bivand R.S., (2005). Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

Rodriguez-Aseretto D., de Rigo D., Leo M.D., Corts A., and San-Miguel-Ayanz J., (2013). A data-driven model for large wildfire behaviour prediction in Europe. Procedia Computer Science, vol. 18, no. 0, pp. 1861-1870; doi:10.1016/j.procs.2013.05.355.

Roger S. Bivand, Pebesma E., and Gomez-Rubio V., (2013). Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/

Roy, D.P. and Landmann, T., (2005). Characterizing the surface heterogeneity of fire effects using multi-temporal reflective wavelength data, International Journal of Remote Sensing, 26, 4197-4218; doi:10.1080/01431160500112783

Penn State University, (2012). What are Map Projections? Last visit: June 5, 2020. https://www.e-education.psu.edu/geog160/node/1918

Pérez-Goya U., Montesino-SanMartin M., Militino A.F, and Ugarte M.D., (2020). RGISTools: Handling Multiplatform Satellite Images. R package version 1.0.0. https://CRAN.R-project.org/package=RGISTools.

Vermote, E. F., & Vermeulen, A., (1999). Atmospheric correction algorithm: spectral reflectances (MOD09). ATBD version, 4, 1-107.

Vermote, E., (2015). MOD09A1 MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC. (User Guide); doi:10.5067/MODIS/MOD09A1.006

Wickham, H., (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. URL https://ggplot2.tidyverse.org