Developing a machine learning based algorithm for regional time-series burned area mapping: The highly anthropized Cerrado challenge

Dhemerson E. CONCIANI[1*]; Swanni T. ALVARADO[2]; Daniel B. ALVES[3]; Thiago S. SILVA[4]

[1] Universidade Estadual Paulista (UNESP), Instituto de Biociências, Departamento de Ecologia, Avenida 24-A 1515, 13506-900, Rio Claro, Brazil. *dhemerson.conciani@unesp.br

[2] Universidade Estadual do Maranhão, Programa de Pós-Graduação em Agricultura e Ambiente, Praça Gonçalves Dias, s/n, 65800-000 Balsas, Brazil

[3] Universidade Estadual Paulista, Instituto de Biociências, Lab of Vegetation Ecology, Avenida 24-A 1515, 13506-900 Rio Claro, Brazil

[4] Biological and Environmental Sciences, Faculty of Natural Sciences, Stirling University, Stirling, FK9 4LA, UK

ABSTRACT

This study aims to develop a regional burned area (BA) algorithm for Landsat surface reflectance (SR) images by testing different machine learning (ML) algorithms. Three ML algorithms (RF, XGB, MARS) were fed and tuned by using more than 1 million of spectral signatures of BA and anthropic land-uses from a balanced dataset. As predictors, we used both SR bands and spectral indexes. Different combinations of hyperparameters were tested, being the optimal values selected by using the largest accuracy. RF overcome XGB and MARS, presenting a balanced accuracy of 98%. Validation was made by using the RF model to predict 59 scenes. RF model alone was not sufficient to generate a BA product with suitable quality (kappa= 0.53), thus, post-processing was implemented. Higher accuracy (kappa= 0.79) was obtained by combining infrastructure and terrain masks with a spatial contiguity filter. Balancing of errors prioritized a higher omission (OE= 0.16) than commission (0.09), guarantying that this product can be applied to perform regional analysis without overestimating the BA. Finally, this study launches the first Cerrado's collaborative burned area mapping platform, a simple and intuitive way to share the result with the community and take feedbacks to improve the product quality in the future.

Key- words: burned area; random forest; Landsat; landscape; land-cover; land-use.

1. Introduction

The Cerrado vegetation covers an area of c.a 2 million km², about 25% of Brazilian territory (Durigan & Ratter, 2016). The Cerrado, the largest South American savanna, evolved under natural fire regimes (Simon et al., 2009). However, contemporary fire regimes are highly affected by the established settlements, managing landscape for agriculture and livestock, thus changing the natural fire regimes according with to the local cultural and economic practices (Dias, 2006). These changes in natural fire regimes can completely alter the ecosystem's structure, composition and functionality. Increasing fire frequency is often related to the conversion of savannas into pastures for cattle grazing and opening of new croplands such as soybean or sugar cane (Daldegan et al., 2014), while decreasing fire frequency is specially observed in areas with woody encroachment (Rosan et al., 2019).

Currently, 93% of the entire Cerrado is under an anthropic matrix (Soares-Filho et al., 2014) covered mainly by agriculture, livestock and forestry (Alencar et al., 2020). In the state of São Paulo, anthropic land-use is observed in more than 95% of the Cerrado's original cover (Kronka et al., 2005). The presence of highly populated cities (Σ= 44 million inhabitants (IBGE, 2014) and dense infra-structure (eg. roads, railways, power transmission lines and telecommunication towers) are added to the Cerrado of São Paulo's landscape, increasing its complexity. For this reason, fire has been considered as a destructive force and legally banned (State Law 10 547/ 2001) to decrease fire use and mitigate its impacts on human populations. Despite the Cerrado being a fire prone ecosystem, private companies and government agencies maintain fire brigades to extinguish any fire occurrence, not mattering its cause or location (Durigan & Ratter, 2016).

When monitoring fire regime changes it is important to track where, when and how much vegetation is being affected by fire, ensuring that spatial features of the fire regime can be assessed on fire policy reviews and included on future studies of modeling, species distribution, botany, etc. Current Remote Sensing products derived from MODIS have been showing good results in evaluating natural and anthropic processes affecting land surface, allowing the global detection of active fires (Active Fire Data | MCD14DL, 1 x 1 km) and burn scars (LP DAAC - MCD64A1, 500 x 500m) since the 2000's. However, these products are not suitable to analyze regional and local patterns due to their low spatial resolution, especially in highly fragmented landscapes such as the state of São Paulo. The program “INPE Queimadas” produced burned area products based on Landsat images (AQM30, 30 x 30 m) for the entirety of Cerrado, but these products have not advanced beyond the beta phase and are limited to the 2011-2018 period, leaving a gap for the reconstruction of larger and more reliable fire regime historical series.

Traditional methods for time-series burned area classification by using moderate resolution sensors (like Landsat) are mainly based on reflectance ratios between fire sensitive bands (NIR, $SWIR_1$, $SWIR_2$) and spectral index variations (NBR, CSI, NDVI) (Bastarrika et al., 2014; Hawbaker et al., 2017; Koutsias & Karteris, 2000). These approaches present satisfactory results on homogeneous landscapes, but accuracy errors are not balanced to consider open and heterogeneous landscapes like the Cerrado. When considering highly managed landscapes exposed to constant land-use and land-cover changes (LULCC) the accuracy is highly compromised, making these products not suitable to assess ecosystem fire regime features.

Recent advancements on machine learning algorithms and open source libraries present a new opportunity to explore potential applications on automated and semi-automated burned area mapping (A. Pereira et al., 2017; Ramo & Chuvieco, 2017). Thus, we tested potential applications of different machine learning algorithms (eXtreme Gradient Boosting, Multivariate Adaptive Regression Spline, Random Forest) to reconstruct the contemporary fire regime of São Paulo’s Cerrado by using Landsat time-series data (TM, ETM+, OLI). We trained and tuned classification models by using spectral signatures of burned areas and LULCC’s from six Cerrado’s protected areas and their respective buffer zones (7 km). By using the largest accuracy value to select the optimal model, we applied it into a dense Landsat time-series and generated a standardized burned area product from 1985 to 2018 for the highly anthropized Cerrado. We performed the validation of this product by considering an independent multi-temporal burned area dataset and an adaptative post-processing routine.

Before beginning, we theorized: i) tuning the hyperparameters will affect the accuracy performance; ii) random forest and extreme gradient boosting will outperform multivariate adaptive regression spline; iii) mask of some LULCCs will be need to reach an acceptable product quality; iv) the final will be of sufficient quality to carry out environmental analysis on regional scale.

2. Methods

2.1. Study area

We focused on generating an accurate burned area product for highly anthropized Cerrado considering the São Paulo state land cover context. For this, we selected 9 WRS Landsat paths/rows covering an area of 228 776 km² (Figure 1). Parts of other states were included when sharing the same scene as our target sites. Thus, the total covered area by this study can be divided into 69% covering São Paulo state, 15% southern Minas Gerais state, 14% northern Paraná state and 2% Atlantic Ocean, the last covering small islands on São Paulo's coast.

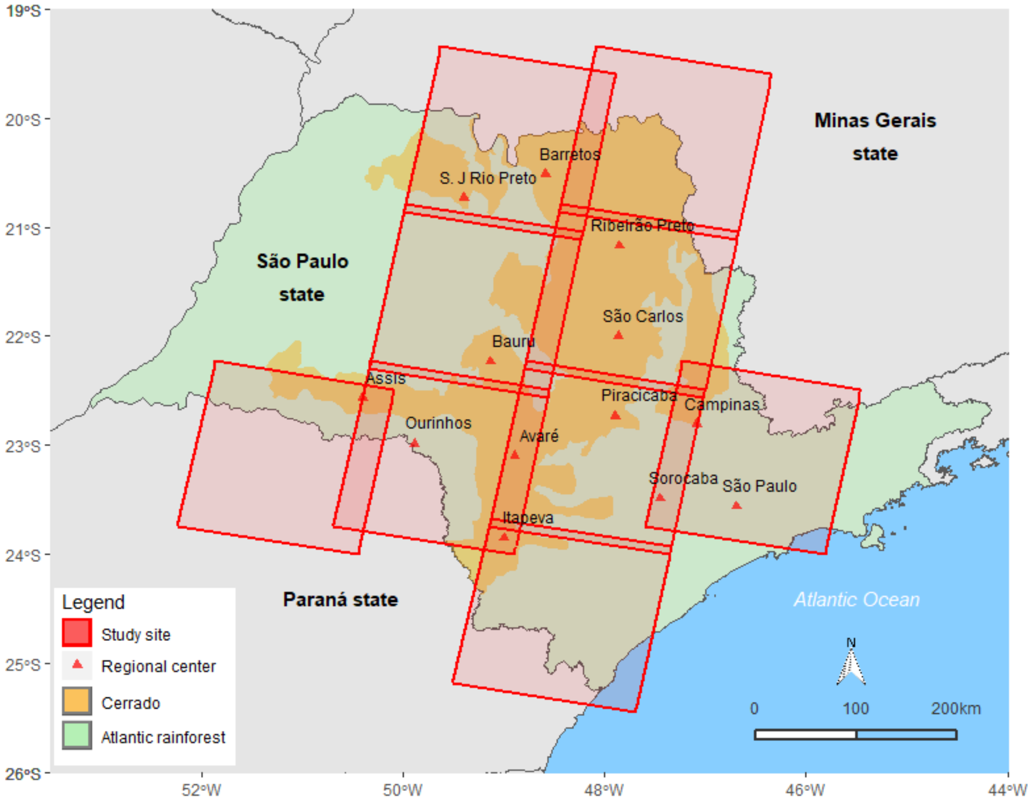


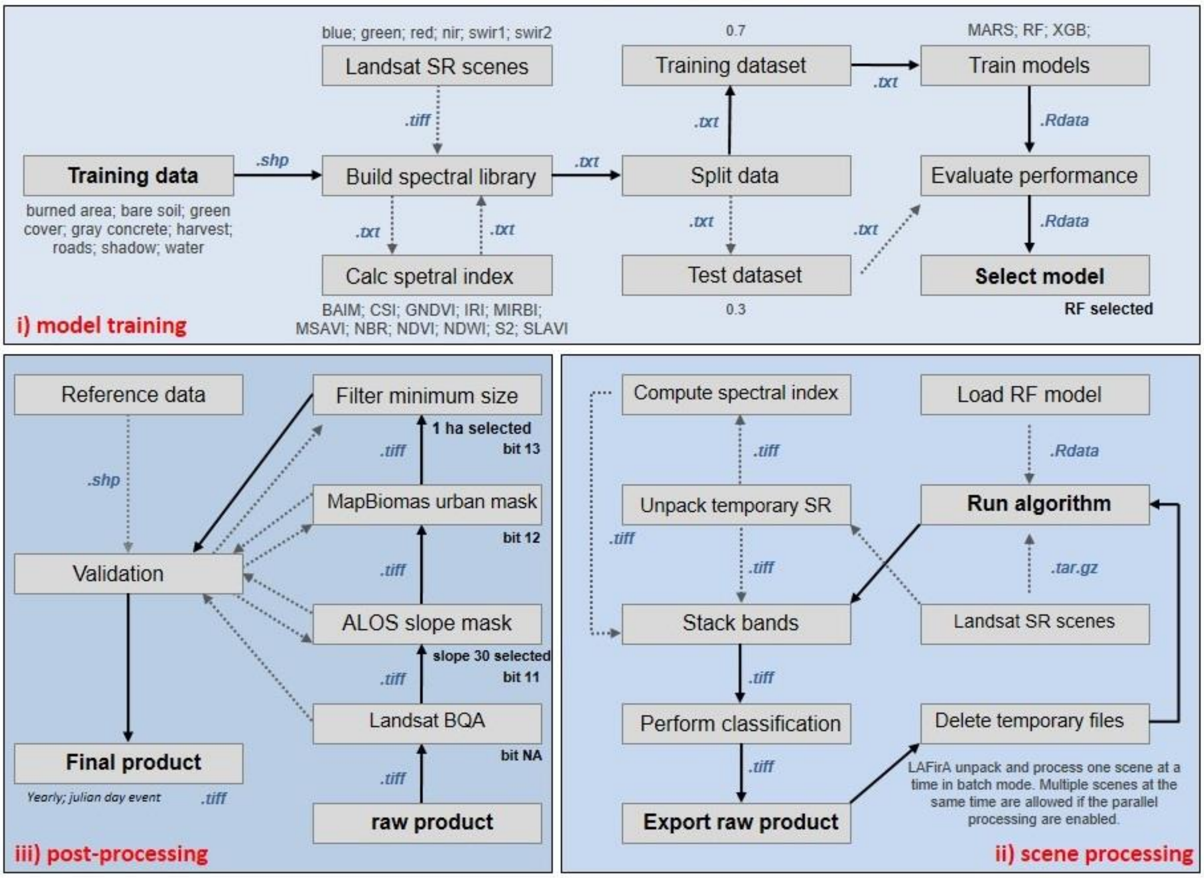
Figure 1. Study area covered by Landsat Burned Area product for the Cerrado of São Paulo state

The population for the study area was estimated around 38,3 million inhabitants in the 2010 census, distributed along 487 municipalities, representing almost 87% of São Paulo's state population and 20% of Brazil's population (IBGE, 2014). São Paulo presents the highest GDP (Gross Domestic Product) from Brazil, where the countryside is responsible by primary production (mainly sugar-cane, soybean, coffee and forestry) while the main cities rely heavily on services and heavy industries, concentrating a high population compared to neighboring cities (IBGE, 2014).

Remnants of primary native vegetation in São Paulo state are characterized by 15.7% total state area of Atlantic Rainforest and ~1% of Cerrado (Atlântica, 2017). These native remnants of both vegetation types are highly fragmented and these patches are located mainly in protected areas (Kronka et al., 2005). The Cerrado remnants are mainly dominated by forest-like formations (locally known as "cerradão"), and only very few areas of open physiognomies (locally known as "campo limpo", "campo sujo"), which are extremely rare and restricted to small patches into the anthropized matrix of the countryside (Vicente et al., 2005).

2.2. Algorithm workflow

We developed an automatic algorithm to detect burned areas in the highly anthropized Cerrado. First, we trained different classification models based on machine learning algorithms and assessed prediction performance of each one by using the balanced accuracy and kappa index. Second, the best fitted model was applied to classify a dense Landsat time-series. Finally, we balanced commission and omission errors in the burned area to ensure that the final product can be used to perform regional scale analysis. All the processing steps briefly described in this section were represented in the Figure 2 and will be detailed in the following sections.



LAFirA unpack and process one scene at a time in batch mode. Multiple scenes at the same time are allowed if the parallel processing are enabled.

Figure 2. Algorithm graphical abstract. Blue boxes show a set of one-step processes. Red labels show step titles. Gray boxes represent each individual process. Black bold labels inner gray boxes show start/end processes from a step. Black solid arrow points primary flux of processes, while gray arrow indicates secondary processes that occur in the background and feed primary processes. Gray labels offer a short description in specific boxes. Black bold labels outside gray boxes points selected parameters/setup. Blue labels points file extensions expected as input and exported as output.

All processing steps were parallelized to take advantage of multicore CPUs and made using R (RCT, 2020). Specific key-steps were accomplished by using community packages “caret”, “raster” and “rgdal” (Bivand et al., 2015; Hijmans & van Etten, 2012; Kuhn & Johnson, 2013) and jscript implementations into Google Earth Engine. Source codes are available and can be accessed in GitHub (https://github.com/musx/FireLand_SPv1). The computation infrastructure used was core i7 5820K 3.3 GHz CPU, 64GB RAM and a GTX 1060 6GB GPU.

2.3. Building the spectral library

We previously selected training sites that contain representative sample areas of native vegetation remnants and anthropic land uses. These sites correspond to six protected areas being three of full preservation (Assis, Santa Bárbara and Itirapina Ecological Stations) and three of sustainable use (Assis, Santa Bárbara and Itirapina State Forests). We also considered buffer zones of 7 kilometers around each one of these protected areas. A highly accurate and manual burn scar mapping dataset was already available for these areas from 1984 to 2016 (Conciani et al in press). This previous mapping was performed based on visual detection and manual delineation of every burn scar detected into 805 Landsat surface reflectance Level-2 scenes from *Earth Resources Observation and Science- Center Science Processing Architecture* (EROS-ESPA, https://espa.cr.usgs.gov/) for the WRS path/rows 220/75, 221/76, 222/76.

In order to create a diverse spectral library, we mapped samples over time of land uses with similar spectral signature when compared to burn scars (Table 1). Furthermore, spectral signature of generic land covers (e.g. “green cover”, “bare soil”) was also mapped in order to train a landscape classifier with ability to recognize burn scars on highly anthropized areas.

Table 1. Mapped classes to train classification models. A byte value was assigned for each class in order to identify these elements in further proceedings.

Byte	Class	Description
1	Burned area	Recently burned area, with ash presence
2	Bare soil	Soil without any type of vegetation cover
3	Green cover	Any type of green cover, forests, agriculture, pastures
4	Gray concrete	Impermeable structures, cities
5	Harvest	Recent harvest with the presence of decomposing organic matter atop the soil
6	Asphalt	Highways and paved streets
7	Shadow	Cloud and relief shadows
8	Water	Natural/ artificial water courses and water masses

We performed the spectral signature extraction from surface reflectance bands (Table 2) that matches between the mapped vectors by class and the Landsat images for each date. The data extracted in the process was compiled and exported as a database, being used to build our spectral library. A graphical summary of the spectral library is presented in the Figure 3 considering the mean reflectance value for each one of the classes.

Table 2. Surface reflectance spectral bands used to extract spectral signatures. TM = Thematic Mapper (Landsat 5); ETM+ = Enhanced Thematic Mapper Plus (Landsat 7); OLI = Operational Land Imager (Landsat 8).

Spectral band	Landsat TM and ETM+		Landsat OLI	
	Band number	Bandwidth (µm)	Band number	Bandwidth (µm)
Blue	1	0.45 - 0.52	2	0.45 - 0.51
Green	2	0.52 - 0.60	3	0.52 - 0.60
Red	3	0.63 - 0.69	4	0.63 - 0.69
NIR	4	0.77 - 0.90	5	0.77 - 0.90
SWIR$_1$	5	1.55 - 1.75	6	1.55 - 1.75
SWIR$_2$	7	2.09 - 2.35	7	2.09 - 2.35

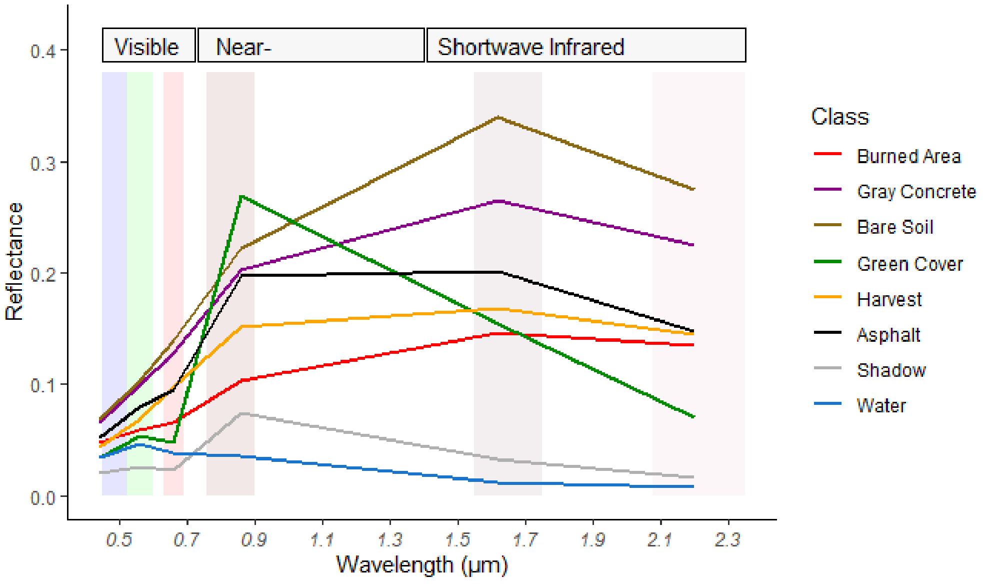


Figure 3. Mean reflectance (y-axis) over the wavelengths (x-axis) from each class present in our spectral library. Line colors represent different classes (described on legend). Background

colors represent Landsat bands ordered by wavelength (Blue, Green, Red, NIR, $SWIR_1$ and $SWIR_2$).

Finally, we used these surface reflectance database as input for the generation of several spectral indices (Table 3). We selected some of the most commonly used indexes in the literature to assess features from the burn scars, vegetation, soil and water, and included them in our spectral library.

Table 3. Spectral indexes generated to enhance our spectral library. The λ symbol represents the reflectance value of the spectral band.

Spectral Index	Reference	Formula
Burned Area Index (BAIM)	(Martín & Chuvieco, 2006)	$\frac{1}{(0.05-\lambda\,\mathrm{NIR})^2+\ (0.2-\lambda\,\mathrm{SWIR1})^2}$
Char Soil Index (CSI)	(Alistair M.S. Smith et al., 2005)	$\frac{\lambda\,\mathrm{NIR}}{\lambda\,\mathrm{SWIR1}}$
Green Normalized Difference Vegetation Index (GNDVI)	(Gitelson et al., 1996)	$\frac{\lambda\,\mathrm{NIR}-\lambda\,\mathrm{Green}}{\lambda\,\mathrm{NIR}+\lambda\,\mathrm{Green}}$
Infrared Index (IRI)	(Hardisky et al.,1983)	$\sqrt{\frac{\lambda\,\mathrm{NIR}^2+\lambda\,\mathrm{SWIR2}}{\lambda\,\mathrm{SWIR1}}}$
Mid-Infrared Bispectral Index (MIRBI)	(Trigg & Flasse, 2001)	$10\ \times\lambda\,\mathrm{SWIR1}-9.8\ \times\lambda\,\mathrm{NIR}+2$
Modified Soil Adjusted Vegetation Index (MSAVI)	(Qi et al., 1994)	$\lambda\,\mathrm{NIR}+0.5-(0.5$ $\times\ \sqrt{(2\ \times\lambda\,\mathrm{NIR}+1)^2-8\ \times\lambda\,\mathrm{NIR}-(2\ \times\lambda\,\mathrm{Red})}$
Normalized Burn Ratio (NBR)	(Key & Benson, 2006)	$\frac{\lambda\,\mathrm{NIR}-\lambda\,\mathrm{SWIR1}}{\lambda\,\mathrm{NIR}\ +\ \lambda\,\mathrm{SWIR\ 1}}$
Normalized Difference Vegetation Index (NDVI)	(Rouse et al., 1974)	$\frac{\lambda\,\mathrm{NIR}\ -\ \lambda\,\mathrm{Red}}{\lambda\,\mathrm{NIR}\ +\ \lambda\,\mathrm{Red}}$
Normalized Difference Water Index (NDWI)	(Gao, 1996)	$\frac{\lambda\,\mathrm{Green}\ -\ \lambda\,\mathrm{NIR}}{\lambda\,\mathrm{Green}\ +\ \lambda\,\mathrm{NIR}}$
Salinity Index 2 (S2)	(Douaoui et al., 2006)	$\frac{\lambda\,\mathrm{Blue}\ -\ \lambda\,\mathrm{Red}}{\lambda\,\mathrm{Blue}\ +\ \lambda\,\mathrm{Red}}$
Specific Leaf Area Vegetation Index (SLAVI)	(Lymburner et al., 2000)	$\frac{\lambda\,\mathrm{NIR}}{\lambda\,\mathrm{Red}\ +\ \lambda\,\mathrm{SWIR2}}$

A Spearman's correlation map was computed (Figure 4) to inspect the relationships between the surface reflectance bands and the generated spectral indexes.

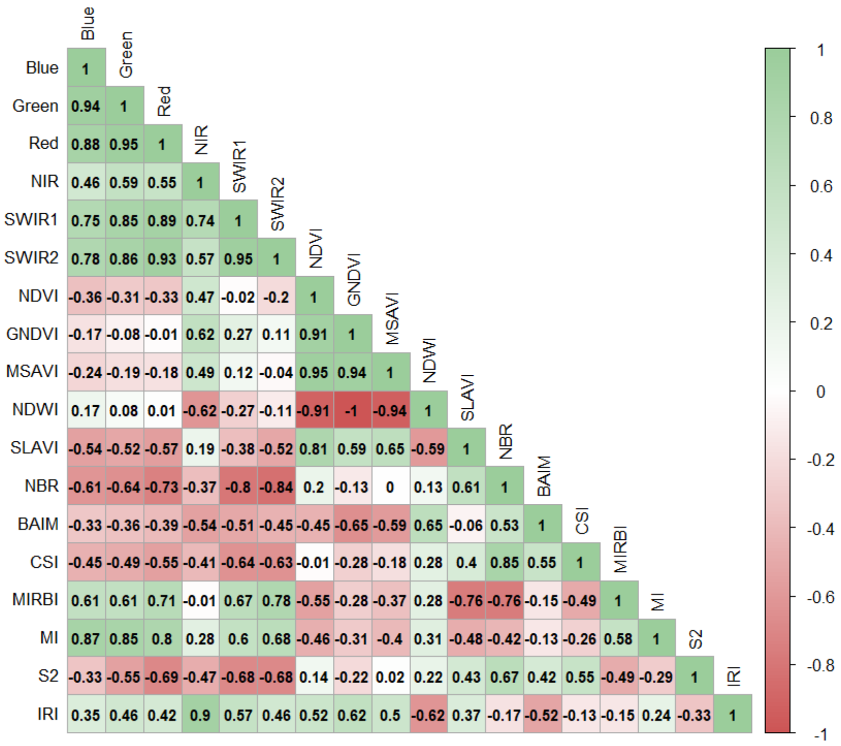


Figure 4. Correlation map between the surface reflectance bands and the spectral indexes in our spectral library. Right bar color vary between strong negative correlations (red) to strong positive correlations (green). Black number inside each square shows de Spearman's correlation (positive or negative).

2.4. Pre-processing

We computed a count of spectral signatures for each class in our spectral library and plotted them into a histogram (Figure 5A). We detected unbalanced observations per class on our database. As a remedy to prevent learning bias, we artificially balanced the frequencies by using the burned area as reference and performing a random down-sampling for classes with higher frequency than the reference (ignoring cases from the majority) and an up-sampling for classes with less frequency than the reference (replicating cases from the minority) (Figure 5B). Thus, we generated a balanced dataset containing 1,153,040 spectral signatures, being 144,130 (12.5%) of each class. This balancing strategy is described in the literature as an alternative to prevent the learning overwhelm by the majority classes (Guo et al., 2008; Provost, 2000). Furthermore, more accurate performances were reported to classifiers trained with balanced datasets when comparing them to classifiers trained with the original data (Batista et al., 2004; Jeatrakul et al., 2010; Van Hulse et al., 2007)

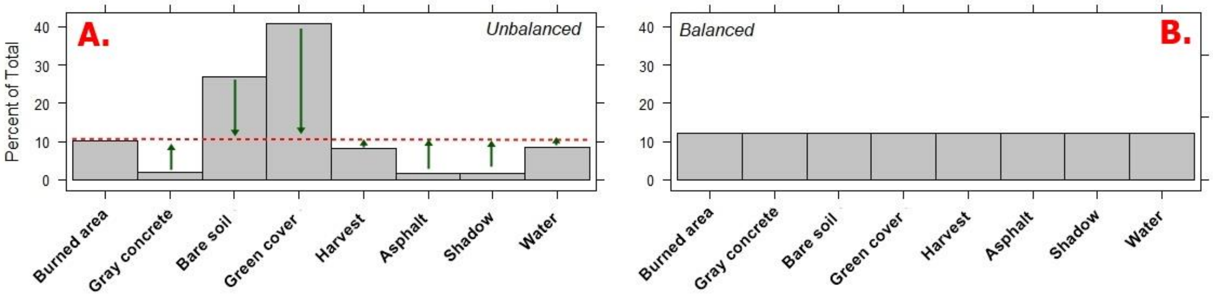


Figure 5. Histogram of frequencies by class. **A**. Original dataset. Red line indicates burned area frequency used as reference to balance other classes. Green arrows points if an up-sampling (up arrow) or down-sampling (down arrow) were performed to balance each class. **B**. Balanced dataset after up-sampling and down-sampling.

Following the most common proportions reported in the literature (Kuhn & Johnson, 2013), we divided the balanced dataset into training dataset by creating a stratified partition with 70% of the data and an test dataset by using the 30% of the remaining data. We centered and scaled the numeric data to take standard deviation one and mean of zero for all the predictors.

2.5. Model training and testing

Using the training dataset as input, we implemented machine learning algorithms considering the scope of non-parametric regressions (Multivariate Adaptive Regression Spline - MARS), decision trees (Random Forest – RF) and boosted trees (eXtreme Gradient Boosting – XGB). Each one of these algorithms have specific parameters that affects the model’s accuracy and which cannot be estimated by using the dataset (Table 4). Since there is no analytical formula available to calculate an appropriate value, these parameters are referred as *tuning parameters* or *hyperparameters*. Since these hyperparameters control the model complexity, poor choices for the inputted values can result in low accuracy or over-fitting (Kuhn & Johnson, 2013). In this way, following the adaptative search method described in Olsson & Nelson, 1975, we defined a set of candidate values for each hyperparameter (Table 4). Finally, to avoid the over-fitting, we used the k-fold cross-validation resampling technique (k= 5, repeats =3) for training and estimating the performance of the models by considering all the possible combinations between the candidate values.

Table 4. Hyperparameters description for each algorithm. Numbers following the names of the algorithms refer to the version of the R package that has been implemented. The range column represent the minimum and maximum allowed values for each hyperparameter. The candidate

values column represents the set of values that we used as input to train and evaluate different models accuracy.

Algorithm	Hyperparameter	Description	Range	Candidate values
earth 5.1.2	degree	Product degree	1 – Inf	1 – 3
	nprune	Number of terms	1 – Inf	1 – 20
RandomForest 4.6-14	ntree	Number of trees to grow	1 – Inf	1 – 750
	mtry	Number of variables randomly sampled as candidates at each split	1 – *n(β)*	2 – 8
xgboost 0.90.0.2	nrounds	Number of boosting iterations	1 – Inf	50 –150
	max_depth	Max tree depth	0 – Inf	1 – 3
	eta	Shrinkage	0 – 1	0.3 – 0.4
	gamma	Minimum loss reduction	0 – Inf	D_0
	subsample	Subsample percentage	0 – 1	0.5 – 1
	colsample_bytree	Subsample ratio of columns	0 – 1	0.6 – 0.8
	min_child_weigth	Minimum sum of instance weight	0 – Inf	D_1

Inf = Infinite; n(β) = number of predictors; $D_{0,1}$ = default hyperparameter value (0 and 1 respectively)

We computed and used the largest values of overall accuracy (ACC – eq. 1) and the Cohen’s Kappa index (Kappa – eq. 2) obtained in the training by the k-fold cross validation to select the best values for the hyperparameters as well as the optimal model trained by each algorithm. Then, we used these three finalist models (one by algorithm) to predict the test dataset and assessed the performance of each one by computing a confusion matrix comparing the predicted classes vs. reference classes. Once again, we used the largest accuracy value obtained by the test dataset classification to select the final model used in this article.

$$\mathrm{ACC} = \frac{\sum \mathrm{TP} + \sum \mathrm{TN}}{\mathrm{n}}$$

eq.1
TP= true positive; TN= true negative; n= total population

$$\mathrm{Kappa} = \frac{\left[(\sum \mathrm{TP} + \sum \mathrm{TN}) - ((\sum \mathrm{TP} + \sum \mathrm{FN}) \times (\sum \mathrm{TP} + \sum \mathrm{FP}) + (\sum \mathrm{FP} + \sum \mathrm{TN}) \times (\sum \mathrm{FN} + \sum \mathrm{TN}))/_{\mathrm{n}} \right]}{\left[\mathrm{n} - ((\sum \mathrm{TP} + \sum \mathrm{FN}) \times (\sum \mathrm{TP} + \sum \mathrm{FP}) + (\sum \mathrm{FP} + \sum \mathrm{TN}) \times (\sum \mathrm{FN} + \sum \mathrm{TN}))/_{\mathrm{n}} \right]}$$

eq.2
FN= false negative; FP = false positive

2.6. Burned area extraction into a dense time-series

Considering the study site extension, we retrieved the metadata for all the available Landsat scenes from Earth Explorer (https://earthexplorer.usgs.gov/). Assessing scene availability according cloud cover percentage (Supplementary Fig

S1), we found that a maximum of 75% cloud cover is the more suitable threshold for this study. Thus, we discarded all the scenes with more than 75% of cloud cover, preventing unnecessary processing caused by scenes with high cloud cover (NoData). Considering the metadata of processing level, we discarded scenes classified into L1GT (without precision correction) and L1GS (without terrain correction), using all remaining scenes available in the L1TP level (precision, terrain, geometric and radiometric corrections). We used this filtered list to build a request containing 4180 Level-2 scenes (surface reflectance) from 1985 to 2018 and downloaded them from Earth Resources Observation and Science- Center Science Processing Architecture (EROS-ESPA, https://espa.cr.usgs.gov/). Using the images, we calculated the same spectral indexes used to build our spectral library (Table 3) and stacked them as different bands into each one of the downloaded scenes.

For each scene, we used the final model to run per-pixel burned area and LULCC classification, being a value from 1 to 8 associated to each pixel as a result. These values from 1 to 8 correspond to the byte code of the predicted class (Table 1). Finally, using these classified scenes as input, we performed the binarization of the burned area class (1= burned area, 0= unburned – all other classes from 2 to 8). These binarized burned area data were written in new raster files containing the same metadata as the original scenes.

2.7. Burned area validation

Validation is the term used to refer to the process of assessing the accuracy of a product by comparing with an independent reference data (Roy & Boschetti, 2009). In the context of this article, we used the selection of representative places in space and a random design in time to assess the transferability of the classifier to regions outside the training data scope. In this way, we established four plots of 270 km² each (15 x 18 km) in different path/rows considering land-cover and land-use variations in São Paulo state (Figure 6, Table 5) and generated an independent validation dataset by performing the manual vectorization of burned areas over 59 cloud-free scenes across four random years (Supplementary Table S2).

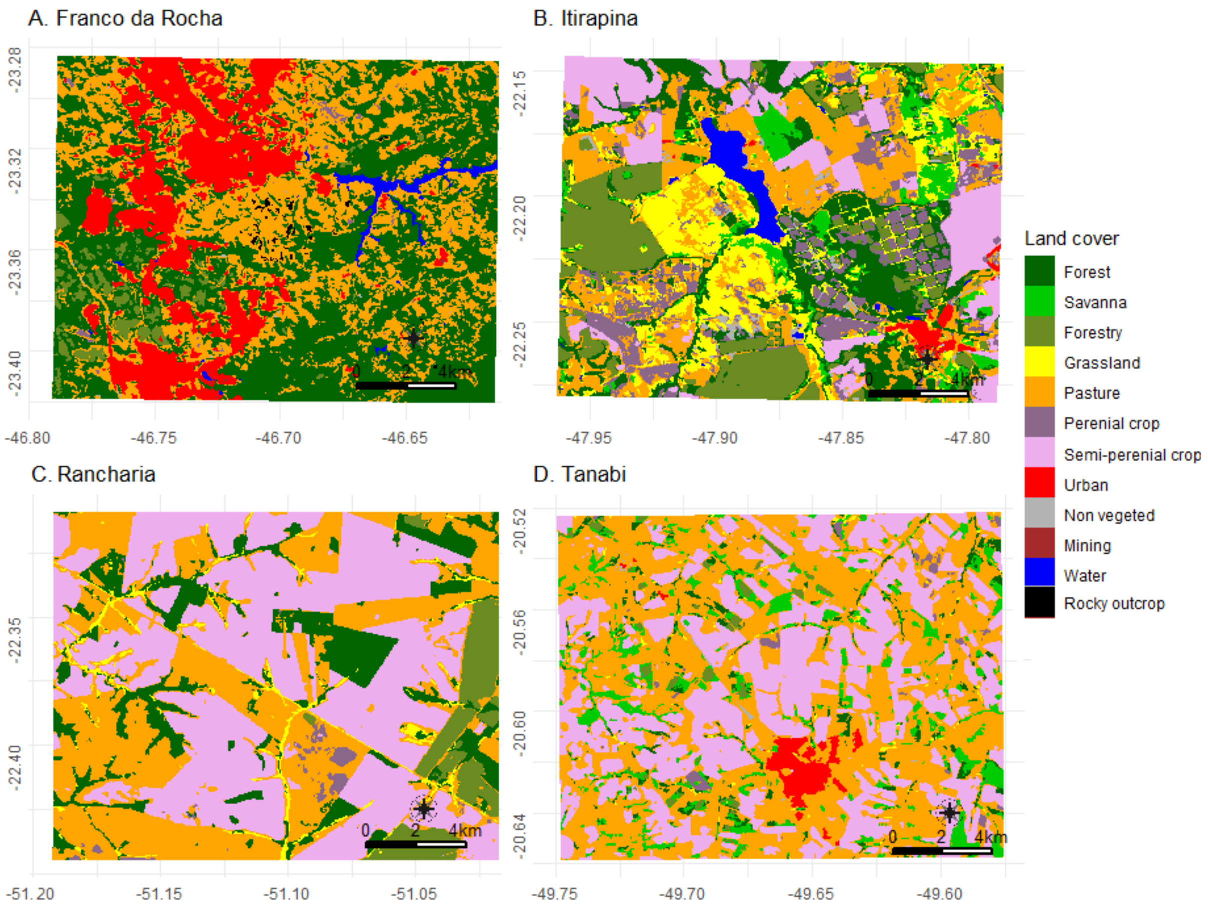


Figure 6. MapBiomas land cover for the validation plots in the year of 2018. **A.** Franco da Rocha – path 219/ row 76; **B**. Itirapina – 220/75; **C**. Rancharia – path 222/76 and; **D**. Tanabi – path 221/74.

Table 5. Landscape description and considered years for validation in each site.

Validation plot	Years	Description
Franco da Rocha	1995 2003 2017 2018	Densely populated suburban area inserted in São Paulo capital city urban zone (> 1 million inhabitants). This area presents highly rugged relief mainly covered by "Serra do Mar" Atlantic rainforest. However, few pasture areas are observed on the landscape and open Cerrado "campo sujo" fragments occur in the Juquery state park.
Itirapina	1985 1988 2015 2018	This area represents the biggest open Cerrado "campo limpo" and "campo sujo" remnants of São Paulo state, located at the Itirapina Ecological Station (~2200 ha). Outside the protected area, the landscape is dominated by cattle grazing, forestry and sugar-cane plantations. Some wetlands divide space with a rich drainage system, small urban zones (< 20 000 inhabitants), highways and railways.

Rancharia	1985 2001 2017 2018	This area corresponds to a transition between Cerrado and Atlantic rainforest. The validation plot includes a rural zone dominated by semi-perennial croplands of bean, soybean and corn. Small fragments of Atlantic rainforest remnants are maintained by farmers as a legal requirement by the National Forest Code (National Law 12 651 / 2012).
Tanabi	1995 2006 2016 2018	Regional hub in sugar and ethanol industrial production. Landscape is dominated by sugar-cane croplands with small rivers and some riparian forests. As well as the Rancharia area, here there are small fragments of “cerradão” and Atlantic rainforest maintained under legal requirement while the Cerrado area has been converted into pastures. Tanabi’s urban zone (< 25 000 inhabitants) was included in the validation plot.

We used the date metadata (yyyy-mm-dd) to match and overlap the manually mapped vectors and the binarized rasters of burned area generated by our algorithm. Then, a confusion matrix was computed to compare each one of the spatial and temporal matches; in other words, we estimated and stored in a database the kappa index (eq. 2), omission error (OE – eq. 3) and commission error (CE – eq. 4) for each one of the comparisons between references vs. predictions. We used this database to calculate the mean value of these metrics for each one of the validation plots (eq. 5) and assumed the result as a representative value of the spatio-temporal product quality. Furthermore, the quality assessment of the product provided by the validation was used to delineate the post-processing routine in order to improve the product accuracy.

$$OE = \frac{\sum FN}{n(R)}$$

eq.3

FN= false negative; n= total population; R= reference

$$CE = \frac{\sum FP}{n(P)}$$

eq.4

FP= false positive; P= predicted

$$\bar{x} = \frac{\sum_{i=1}^{n} x_{i\ (Kappa;\, OE;\, CE)}}{n(\mathrm{P} \sim VP)}$$

Pr= predicted; ~ in each; VP= validation plot

eq.5

2.8. Post-processing

As a standard procedure, we used the date metadata to match and mask (into NoData) any pixels detected as radiometric saturation, cloud, cloud shadow and water in the binarized burned area product by using the Landsat Quality Assessment Band (QA). For precaution, we applied restrictive thresholds in the QA parameters, being masked all the pixels (and also their adjacencies) that presented any of the previous described anomalies, independently of the confidence level.

Considering that our gray concrete spectral signatures is mainly composed by urban-zones and sparse buildings, since our aim is to generate a product to assess the ecosystem fire patterns, we decided to mask all the urban-zones. In this way, the MapBiomas Brasil project (collection 4.1) offers an accurate yearly classification of land-cover for the Cerrado by performing Landsat scenes classification (Alencar et al., 2020). Thus, we used the "urban-infrastructure" class from the MapBiomas products to mask our binarized burned areas. Besides that, we also used the "mining", "beach" and "rocky-outcrop" MapBiomas classes (employing our empirical knowledge that these classes don't burn) to mask our data.

Due the earth's movements (e.g. rotation, translation) the sun-earth inclination angle change across the seasons, and so the extent of the mountain shadows projected over the land surface accompany this variation (Giles, 2001). Previous studies focused in scene classifications have reported that the spectral mixture caused by the projection of mountain shadows over the surface of highly sloped areas can induce several misclassifications (Y. Chen et al., 2018; Giglio et al., 2015; Paul, 1997). In this way, we obtained the AWD3D30 v1.1 ALOS Digital Surface Model (Jaxa, 2020) from the Google Earth Engine library and derived the terrain slope for São Paulo state. We binarized slope rasters (1= slope greater than x, 0= slope less than x) by considering different slope thresholds (x= 10°, 20°, 30° and 40°) and tested how the terrain masking can improve or degrade the product quality of the burned area product in the context of this study.

Finally, to improve the consistency and the product quality, we assume that isolated pixels classified as burned area (without neighbor pixels classified as burned area) have a great chance to be misclassifications. To test this premise we implemented a "minimum spatial contiguity" filter based in the count of pixels classified as burned area that share their borders. Thus, we tested the effects in the product accuracy by masking burned area pixel aggregations less than 5 pixels (0.45 ha), 11 pixels (0.98 ha) and 16 pixels (1.44 ha) and compared them to the product without spatial contiguity filter (considering alone pixels as valid burned areas).

2.9. Final product compilation

After applying the post-processing steps and finding the best parameters for the masks and filters by balancing the omission and commission errors, we retrieved the gregorian date (yyyy-mm-dd) for each year. Final product was serialized by year ~ path/row and resulted in a library of 309 raster files in .tif format. The file names were built to store the burned area product metadata as text strings, being: i) the WRS-2 path [path] and WRS-2 row [row] as spatial descriptors inside the same separator [path+row]; ii) the gregorian year as temporal descriptor [yyyy] and; iii) the abbreviation of a short product description and the version number [jdba1], equivalent to "julian day of burned area detection, version 1". This metadata were compiled so that the final file names presented the format "pathrow_yyyy_jdba1.tif" (e.g. 22075_1985_jdba1.tif, 21976_2018_jdba1.tif, etc.).

3. Results and Discussion

3.1. Hyperparameters tuning and model selection

3.1.1. Multivariate Adaptive Regression Spline – MARS

We detected a positive effect of the hyperparameter maximum number of terms (nprune) in the models accuracy. Lower accuracies were observed by using low nprune values (ACC ranging from 0.324 to 0.402 when nprune = 2), independently of the product degree (degree). As new MARS models have been trained by increasing the number of terms, we observed strong gains in the accuracy until the nprune value = 15 (ACC ranging from 0.813 to 0.845) (Figure 7). This result points that despite the number of terms having largely contributed to the accuracy gain, this gain has a tendency to saturate since a threshold, being all the addition of complexity since this point responsible for a overwhelm of the model. This saturation pattern was statistically and empirically demonstrated by Kuhn & Johnson, 2013 and also reported as result from other studies that have tuned MARS models to make forecasts (Ferlito et al., 2017; Li et al., 2019).

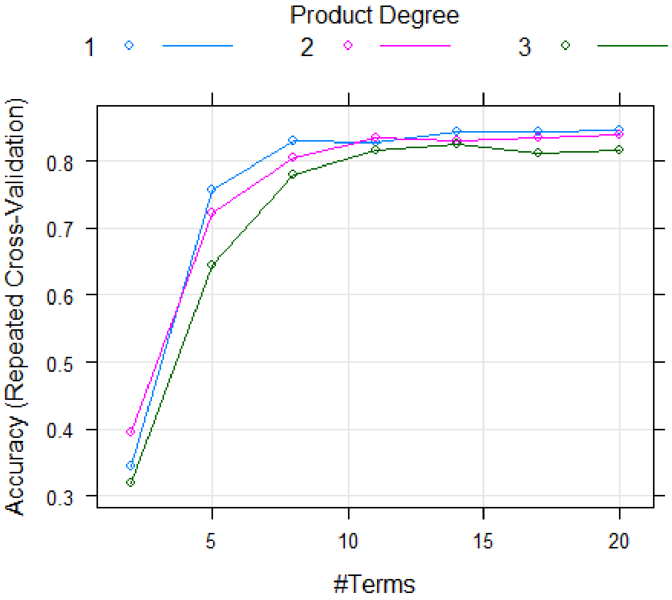


Figure 7. Multivariate Adaptive Regression Spline (MARS) hyperparameters tuning. Colored lines represent models considering different product degrees. The number of terms hyperparameter (nprune) was represented in the x-axis while the y-axis points the models accuracy.

On the other hand, although in small proportion in relation to the total number of terms, the product degree has affected the models accuracy. While the addition of terms has induced accuracy gains, the addition of product degrees has negatively affected the models by degrading the accuracy. In other words, independently of the number of terms in the model, greater values of accuracy were observed when only first order interactions between variables are allowed (degree = 1), while less accuracy was found when increasing the product degree. The combined effects of both hyperparameters can be easily observed in Figure 8, since the results showed a graphical ordered pattern. Finally, the largest accuracy value (ACC= 0.845, Kappa = 0.823) was used to select the optimal MARS model (nprune = 20 and degree = 1).

3.1.2. Random Forest – RF

We started by training an exploratory RF model using an approximation of the default value for the number of variables randomly sampled as candidates at each tree split (mtry = $\sqrt{n.of\ predictors}$ = $\sqrt{6\ SR\ bands + 11\ spectral\ index}$ = 4.12, rounded to 4). On one hand if a low number of trees are related to poor final classifiers, on the other a large number of trees are generally related to a model overwhelm and to unnecessary computation (Oshiro et al., 2012). Since no default value is available to the ntree hyperparameter, we rely on literature and set our initial candidate value as 750, an approximation of the optimal ntree value reported by Ramo & Chuvieco, 2017 to perform MODIS burned area classification.

Interestingly, a high value of accuracy (ACC= 0.976, Kappa= 0.971) was obtained from this exploratory model. Even when the model considered only a unique tree, a relatively high value of accuracy was obtained (ACC= 0.952). However, increasing in the accuracy of all the classes was observed in the window by ensembling from 5 to 100 trees (Figure 8). From this point on, an average accuracy of 0.975 was reached and a tendency of stabilization was detected for all the classes, except for the harvest class, where a small increase of accuracy (< 0.02) occurred until the 750th tree.

Since we reached a global threshold of accuracy gain (and also individual for all the classes), we assume that the input of more trees in the training is computationally unnecessary, since no more significative accuracy gains shall be observed. Thus, despite the optimal value for the number of trees can be any value since the accuracy stabilization threshold (around the 100th tree), considering the computational context of having already trained a stabilized model, we used this classifier and assumed 750 trees as our optimal ntree value in the context of this study.

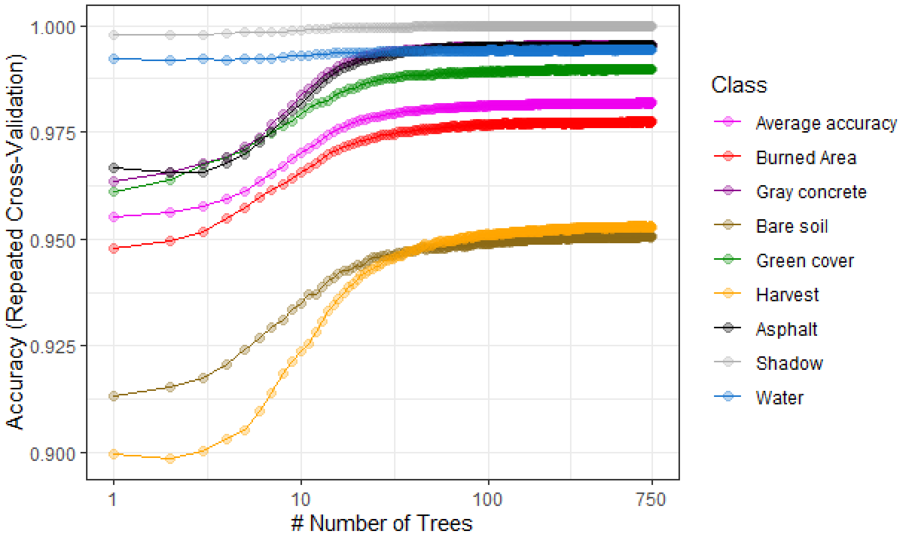


Figure 8. Random Forest (RF) number of trees tuning. Colored lines represent the accuracy error for each one of the classes. Pink link named “Average accuracy” represent the model overall accuracy by considering all the classes. The hyperparameter ntree was represented in the x-axis (log10 adjusted) while the y-axis points the accuracy value.

From this point on, we started to evaluate how the set of different values for the mtry hyperparameter affects the accuracy. For this, we tested values that corresponds to the half (2) and twice (8) of the default value (4). Considering the half value of the default, we found that the accuracy presented an insignificant drop (< 0.001), such as none accuracy gain was observed by twice the default value (ACC was held constant in 0.976). Absence of influence of the number of variables randomly sampled as candidates at each tree split goes on the opposite direction to that reported by studies that have classified optical and radar remote sensing data by using the random forest algorithm (Pal, 2005; Ramo & Chuvieco, 2017). Furthermore, other studies have related that the accuracy in the random forest was more sensitive to the mtry tuning than to the number of trees (Belgiu & Drăgu, 2016; Ghosh & Joshi, 2014; Topouzelis & Psyllos, 2012). However, Catal & Diri, 2009 have reported that the behavior of the accuracy in RF can be dependent from the dataset size and feature selection methods (based on predictors relationships). Thus, our results point that the structural particularities from our dataset can be induced by the prioritization of the ntree in relation to mtry.

3.1.3. eXtreme Gradient Boosting – XGB

We report that the subsample ratio of columns (colsample_bytree), shrinkage (eta) and subsample (subsample) did not showed determinant effects on the accuracy values, having only small variations (< 0.01) being observed by ranging the set of values proposed in the search grid (see Table 4). As reported by Joharestani et al., 2019, the number of allowed boosting iterations (nrounds) and the maximum tree depth (max_depth) appears to be the more sensitive to XGB hyperparameters. This way, we detected a positive relationship with the accuracy by increasing the values into each one of these parameters (Figure 9).

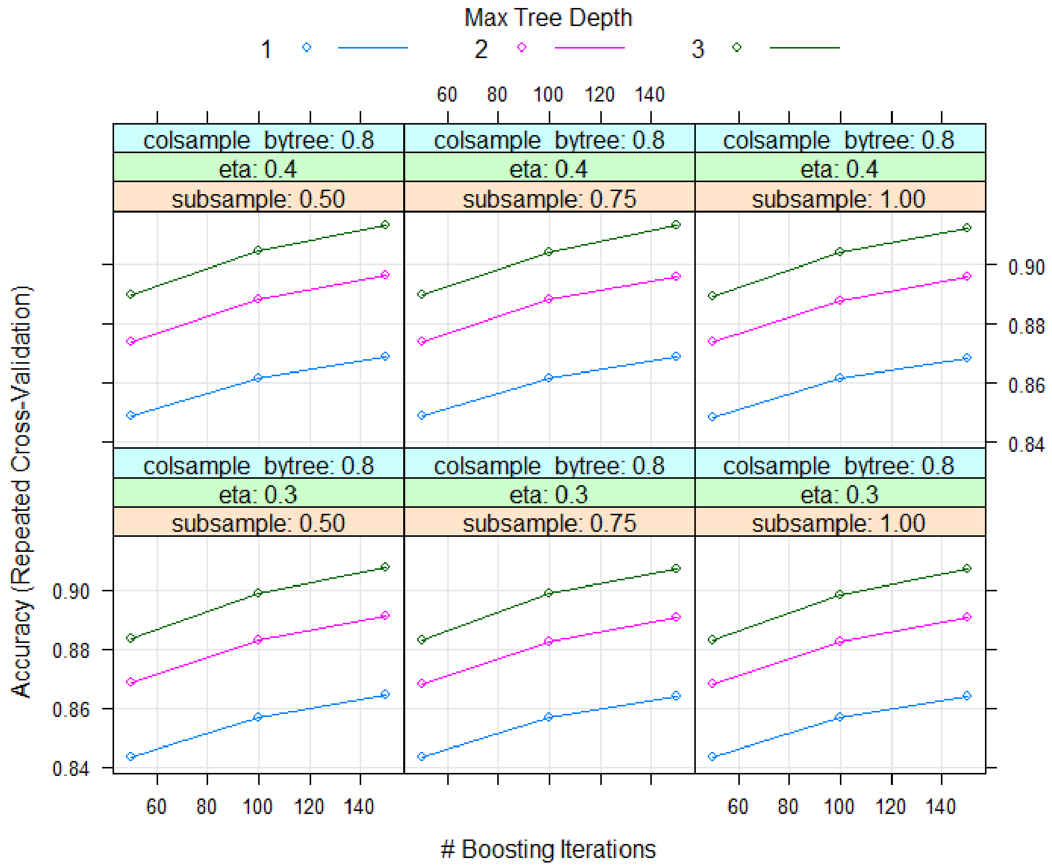


Figure 9. EXtreme Gradient Boosting (XGB) hyperparameters tuning. Colored lines represent models considering different maximum tree depth ("max_depth”). The labels inside colored upper boxes on each plot refers to the values provided to the subsample ratio of columns (colsample_bytree), shrinkage (eta) and subsample percentage (subsample). The x-axis represents the number of boosting iterations (nrounds) while the y-axis points the models accuracy.

Previous studies that have assessed the performance of classification algorithms by comparing XGB vs. RF models pointed that an optimal parameterized XGB tends to outperform RF models (Georganos et al., 2018; Naghibi et al., 2020; Joharestani et al., 2019). However, tuning the hyperparameters into the XGB algorithm is more difficult than MARS and RF algorithms for the simple reason that the first has 3.5 times more parameters to be set. Since the number of possible combinations between candidate values is a function of the number of candidate values in each parameter (NCV) raised to the total number of hyperparameters to be set (7) (NCV^7), the delineation of a detailed search grid can demand the training of thousands of models. For example: a search design that considers the input of 5 candidate values for each one of the parameters needs to train 78,125 models, making the search for the optimal hyperparameters a heuristic process that depends of computational resources availability for a long period of time.

Given that, we report that despite the optimal XGB model having reached less accuracy than the optimal RF model (- 0.063) (Table 6), this result is biased by our decision to end the XGB parameterization before the stabilization of the accuracy gain. Since a small range of eta and colsample_bytree values were tested and no saturations in the accuracy were detected by increasing the nrounds and the max_depth until the tested limits, a new set of values could have been supplied as new candidate values by following the conceptions of the Nelder-Mead method (Olsson & Nelson, 1975). However, since this heuristic search process would consume much more processing time and we had already reached an accuracy that we considered satisfactory in the context of this study by using the RF algorithm, we decided to end the tuning of hyperparameters and make a better use of the research time by processing and validating a high quality final product.

Table 6. Optimal hyperparameters and accuracy measurements for each algorithm. Size (GB) refers to the size (expressed in gigabytes) of trained models when exported as .RData files. Gray shadow points the best model obtained from the k-fold cross validation.

Algorithm	Hyperparameters	Accuracy	Kappa	Size (GB)
RF	**mtry = 4; ntree= 750**	**0.976**	**0.971**	**0.536**
XGB	nrounds= 150; max_depth= 3; eta = 0.4; gamma =0; colsample_bytree= 0.8; min_child_weigth= 1; subsample= 0.75	0.913	0.900	0.310
MARS	nprune= 20; degree= 1	0.845	0.823	19.198

3.1.4. Final model selection

Given the previous steps, we performed the classification of our test dataset (30% of the spectral library) by using the optimal model of each algorithm. Only small variations were observed when comparing the accuracy results from the training k-fold cross validation (see Table 6) and test dataset (Table 7). Thus, we confirmed our hypothesis number ii (RF and XGB > MARS) and used the largest value of accuracy (ACC= 0.982) to select the RF as the final model to be applied in the classification.

Table 7. Accuracy measurements of the test dataset classification for each algorithm. 95% C.I refers to the accuracy 95% confidence interval. Grey shadow points the selected final model by considering the largest values of accuracy and Kappa.

Algorithm	Accuracy	95% C.I	Kappa
RF	**0.982**	**0.981 – 0.983**	**0.979**
XGB	0.913	0.912 – 0.914	0.900

MARS	0.856	0.855 – 0.857	0.835

3.2. Predictors importance

Inspecting the spectral library ordination by performing a principal component analysis (PCA – Figure 10), we detected that the loadings of surface reflectance bands have largely influenced the scores from bare soil, and in a lesser extent, from the gray concrete. As expected, spectral indexes sensible to vegetation features (NDVI, GNDVI, SLAVI, MSAVI) largely influenced the ordination of green cover, such as the burned area indexes (BAIM, CSI, NBR) were largely responsible for the burned area class scores. Highly divergent classes (green cover, shadow, water) were easily separated by the PCA, however, we detected that spectral traits were shared by some of the other classes (burned area, harvest, asphalt, gray concrete, bare soil).

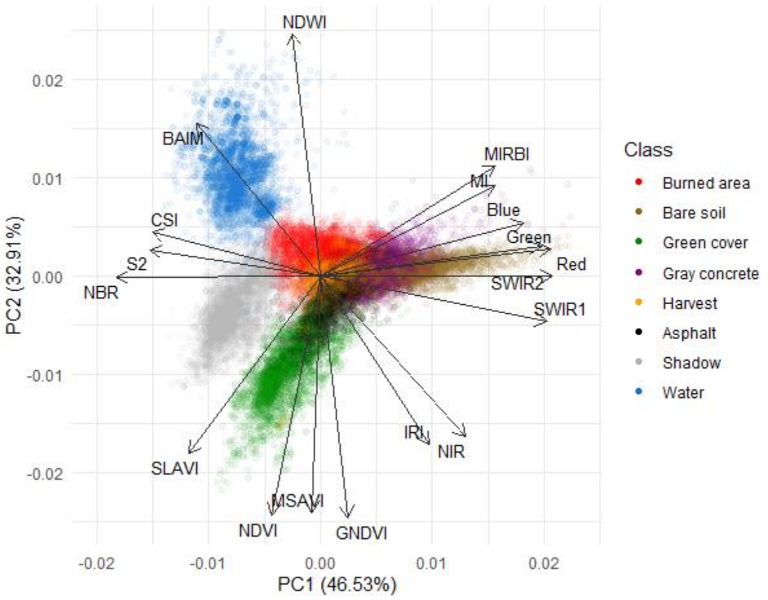


Figure 10. Principal Component Analysis (PCA) of our spectral library. Black labels represent variable names. Length of black arrows represents the loadings of each variable.

When comparing the predictors importance inside each one of the optimal models obtained from the cross validation, we detected deep differences in which and how each model has used the predictors (Figure 11). First, the MARS model used only 9 of the 17 predictors, prioritizing the use of all the surface reflectance bands (76.6% of total importance) and using only 2 of the 11 spectral indexes (NDVI and CSI). On the other hand, XGB and RF have used all the predictors, but emphasizing the importance of spectral indexes (XGB= 63.1%, RF= 64.8% of total importance). Furthermore, we detected a more homogeneous tendency of predictors importance distribution by using the RF (standard deviation ± 2.29 %) when compared to XGB (± 5.23 %) and MARS (± 7.76 %). This result highlights the ability of the RF algorithm in considering the combined effects from predictors with high collinearity by dividing the importance between them (Genuer et al., 2010), contrary to the non-parametric variable selection performed by the MARS (Doksum et al., 2008) and the feature selection performed by the XGB (T. Chen & Guestrin, 2016).

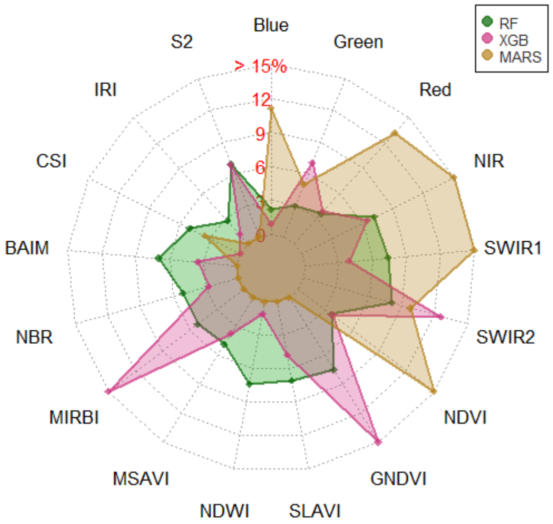


Figure 11. Predictors' importance radar plot. Black labels around the plot represent each one of the predictors used in the models training. Starting from the center (0%) and going to the outer edge (> 15%), the red labels inside the plot shows the relative importance for each predictor. Colored lines/polygons represent the implemented algorithms: Random Forest (dark green), eXtreme Gradient Boosting (magenta) and Multivariate Adaptive Regression Spline (brown).

Inspecting the confusion matrix of the test dataset by using the optimal models, we found a poor performance of MARS and XGB to make correct predictions of classes with high spectral mixture (asphalt, gray concrete and harvest), being these classes frequently misclassified among themselves or in burned area (Figure 12). Despite the spectral reflectance bands being often used to perform LULCC and burned area classification by other studies, better results were reported by adding spectral indexes (Bastarrika et al., 2011; Hayes et al., 2014). Many studies have reported that the NDVI and CSI enhances the classification of vegetation and burned area, respectively (Jia et al., 2014; Shao et al., 2016; Smith et al., 2007; Stroppiana et al., 2012). However, the non-consideration of the other spectral indexes and the emphasis in the surface reflectance bands probably have induced the several observed classification errors for the MARS algorithm.

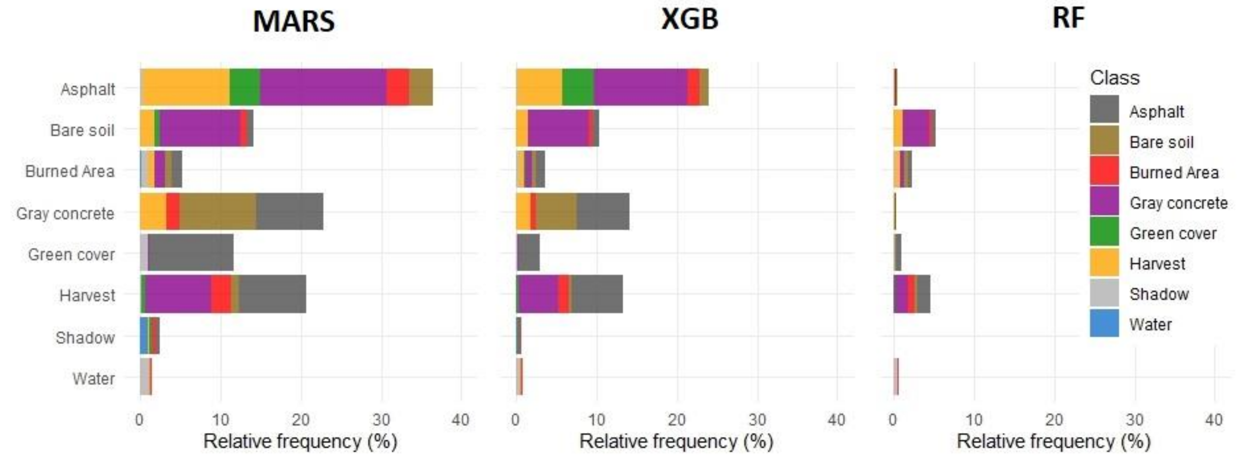


Figure 12. Distribution of total errors by predicting the test dataset. Black labels upper each plot represents the employed algorithm. The y-axis indicates the reference classes while the x-axis points the relative frequency of error per class. The error bar of each reference class is colored to represent the frequencies of wrong predicted classes

Although the XGB has attributed 63.1% of the variables importance to spectral indexes, these value is biased by the high importance lead by MIRBI (18.6%) and GNDVI (15.8%) while the average importance of all the other spectral indexes was around 3.2%. By comparing MARS and XGB, this behavior represents an importance reduction of the surface reflectance bands and an explicit replacement from CSI to MIRBI and from NDVI to GNDVI. Thus, XGB have reduced the classification errors from all the classes, but remained maintaining the same tendencies and biases that MARS.

On the other hand, the balanced use of surface reflectance bands and spectral indexes predictors in the RF practically zeroed the asphalt and gray concrete errors. Furthermore, misclassifications were largely reduced in bare soil (error= 6%), harvest (error= 5%) and burned area (error= 4%). Thus, our result suggests that the distribution of importance among highly collinear predictors have improved the machine learning capabilities in the RF, making possible that this algorithm reached a higher accuracy (98%).

3.3. *Burned area validation*

We applied the final RF model (Table 6) to perform the landscape classification of the selected validation scenes (Supplementary Table S2). We performed the burned area class binarization (1= burned, 2= unburned, all the classes from 2 to 8, see Table 1) and masked these results by applying the BQA band.

Considering the average performance of the RF model to predict burned areas into all the validation plots, first insights about the quality of this raw product showed an average kappa index of 0.53, being the average commission error (CE= 0.50) and average omission error (OE= 0.12). Despite the mean omission error being low in all the validation plots (OE - Franco da Rocha= 0.15, Itirapina= 0.12, Tanabi= 0.14, Rancharia = 0.05), high values of commission errors were observed in all the validation plots (CE – Franco da Rocha= 0.76, Itirapina= 0.50, Tanabi= 0.44, Rancharia= 0.30). We assumed that a mean omission error of 0.12 (lower= 0.05, upper= 0.15) is acceptable in the regional scope of this study and so we centered our efforts in delineating a post-processing protocol to promote the CE reduction. Inspecting our results, we found that pixels that correspond to buildings and roads were not always properly classified by our algorithm, being frequently mapped as burned area (Figure 13A). Thus, we confirmed our hypothesis iii and used this impressions as a starting point to delineate the post-processing.

First, we decided to mask the infrastructures, cities and roads due this high commission error rate. In this way, a highly accurate urban zones classification (IRS – 5 m/pixel) is available into DATAGEO (São Paulo state geospatial data repository - http://datageo.ambiente.sp.gov.br/). However, these official data refers only to a static snapshot from the time (2005). Since urban infrastructure is constantly changing, we need to include urban-zones data that considers these changes. For this, the MapBiomas Brasil project offers a multi-temporal collection of land cover and land use changes (LULCC) for the entire Cerrado in Landsat resolution (Alencar et al., 2020). We used the urban-infrastructure MapBiomas class to mask our burned area product (Figure 13B). Besides that, considering that “rocky outcrop”, “mining” and “beach” are available into MapBiomas product (and we know that these classes don’t burn), we also masked these classes in our product since much of these is frequently related to commission errors in burned area classifications (Koutsias & Karteris, 2000; Mitri & Gitas, 2004; Oechsle & Clark, 2008).

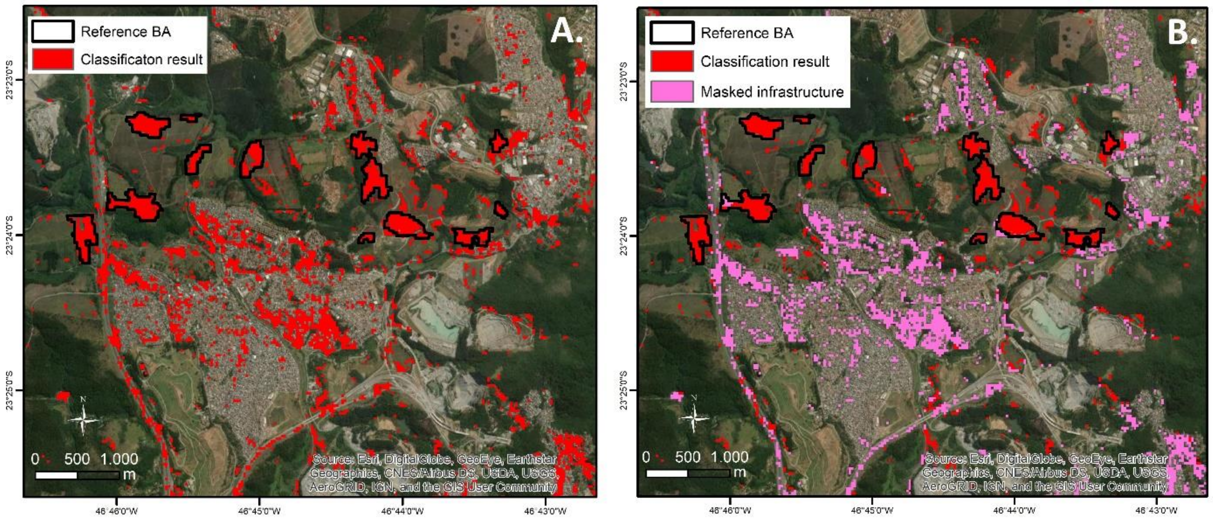

A.
Reference BA
Classificaton result
0 500 1.000
m
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
23°23'0"S
23°24'0"S
23°25'0"S
46°46'0"W
46°45'0"W
46°44'0"W
46°43'0"W
B.
Reference BA
Classification result
Masked infrastructure
0 500 1.000
m
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
23°23'0"S
23°24'0"S
23°25'0"S
46°46'0"W
46°45'0"W
46°44'0"W
46°43'0"W

Figure 13. Franco da Rocha, 2018-08-30. **A)** Raw burned area product (only BQA applied). Black polygons represent our reference burned area dataset. Red colored pixels represent pixels classified as burned area by the RF model. **B)** Burned area product masked using MapBiomas (pink). We used a high resolution scene from ESRI Imagery as background in both figures.

After applying the MapBiomas mask, we observed CE decrease in all the validation plots. Franco da Rocha as shown the biggest drop in CE (-0.22), while Itirapina (-0.03), Tanabi (-0.01) and Rancharia (-0.01) presented quasi-neutral CE reduction, probably due to absence of dense urban-zones in these validation plots. Considering OE increase as a collateral effect of product masking, we detected small increases from 0.01 to 0.02 in all the validation plots. Since CE drop and OE gain were balanced for all the validation plots, the higher CE drop in Franco da Rocha has been driven an average Kappa gain from 0.53 to 0.58.

Applying different derived slopes from ALOS AW3D30 product has not reduced CE more than 0.01 in none of the validation plots. On other hand, we detected that OE is sensitive to the restriction level from slope mask, namely, the more restrictive was the slope mask, bigger was the OE increase (Figure 14), especially in Franco da Rocha. Observing that the CE and OE was balanced (± 0.01) until intermediate degree slopes, we decided to maintain the slope mask (30º) as a post-processing step. We assume that despite the neutral influence in our validation plots, the slope mask can be useful to improve the product quality in other rough relief areas that were not considered in this validation scope but occur in the study area.

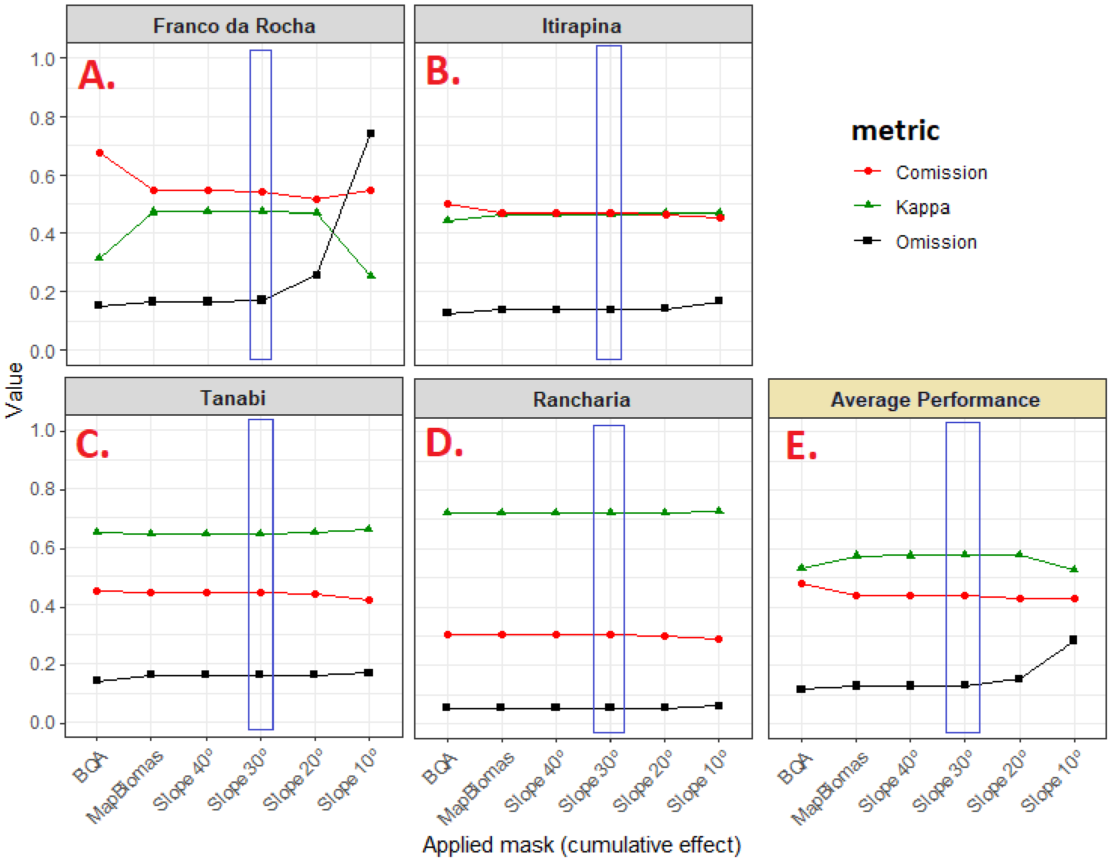


Figure 14. Cumulative performance of tested masks into each validation plot. The x-axis shows the cumulative effect of the different masks (e.g. "Slope 30º" refers to the result obtained by combining "BQA + MapBiomas + Slope 40 º + Slope 30 º"). Colored lines represent the CE (red), OE (black) and Kappa index (green). Dark blue rectangle points the selected combination of masks. Average performance (**E**) was computed by considering the mean from the results into the four validation plots.

When comparing the average results from combined masks (Landsat BQA + MapBiomas + Slope 30) with our first product (only BQA mask), we reported a CE decrease from 0.50 to 0.43 (-0.07) and a Kappa increase from 0.53 to 0.58 (+ 0.04). We observed stable CE and OE variations across masking for all the validation plots, except on Franco da Rocha, where a high CE reduction was observed from 0.76 to 0.54 (-0.22). However, higher CE than Kappa continued to be observed in Franco da Rocha (Kappa= 0.47, CE= 0.54) and Itirapina (Kappa= 0.46, CE= 0.47).

We observed that the remaining errors were distributed in a wide range of contexts. Even applying a restrictive filter to the Landsat BQA mask, we report that some small water masses, sparse clouds and their shadows were not masked. Many of these unmasked pixels corresponded to the spectral mixture between water-vegetation, water-soil, shadow-vegetation and shadow-soil, being frequently classified as false-positive burned areas. The MapBiomas mask showed a good performance to mask false-positive burned areas in dense urban zones. However, the false-positive burned areas that corresponded to sparse buildings, rural communities and small settlements were not properly masked by MapBiomas. Besides that, São Paulo state has a dense transport infrastructure (railroads and highways) being these also not masked by MapBiomas and related to false-positive burned areas.

Although the error tendencies varied between a wide ranges of contexts, we identified a pattern that joined all these contexts: the absence of spatial contiguity. In other words, while burned areas that have been classified correctly showed high spatial contiguity (higher pixels aggregation), the false-positive burned areas presented low spatial contiguity, occurring much of the times restricted to alone pixels or aggregations less than ten pixels. Thus, we decided to use the spatial contiguity as a parameter and performed a test of filters considering different numbers of minimum pixels in an aggregation to promote the false-positive burned areas masking.

Strong CE drops were detected with the lowest spatial contiguity filter (5 pixels or ~0.5 ha) (Figure 15). Considering the average performance (Figure 15E) we report a reduction in the CE from 0.43 (all previous combined masks) to 0.17 (0.5 ha filter). Contrary to the previously tested masks that have contributed to minimize CE only in specific validation plots, the spatial contiguity filter has decreased CE in all the validation plots (Figure 15A, B, C, and D), pushing up the average performance Kappa from 0.57 to 0.76. When we apply more restrictive parameters to spatial contiguity filter by increasing the minimum number of connected pixels (11 pixels = ~1 ha, 17 pixels = ~1.5 ha), we observed CE reduction in all the validation plots. However, a tendency of CE stabilization was detected since the spatial contiguity of ~ 1ha (11 pixels). On the other hand, the OE presented tendency to increase in all the validation plots when more restrictive spatial contiguity parameters were provided (Figure 15).

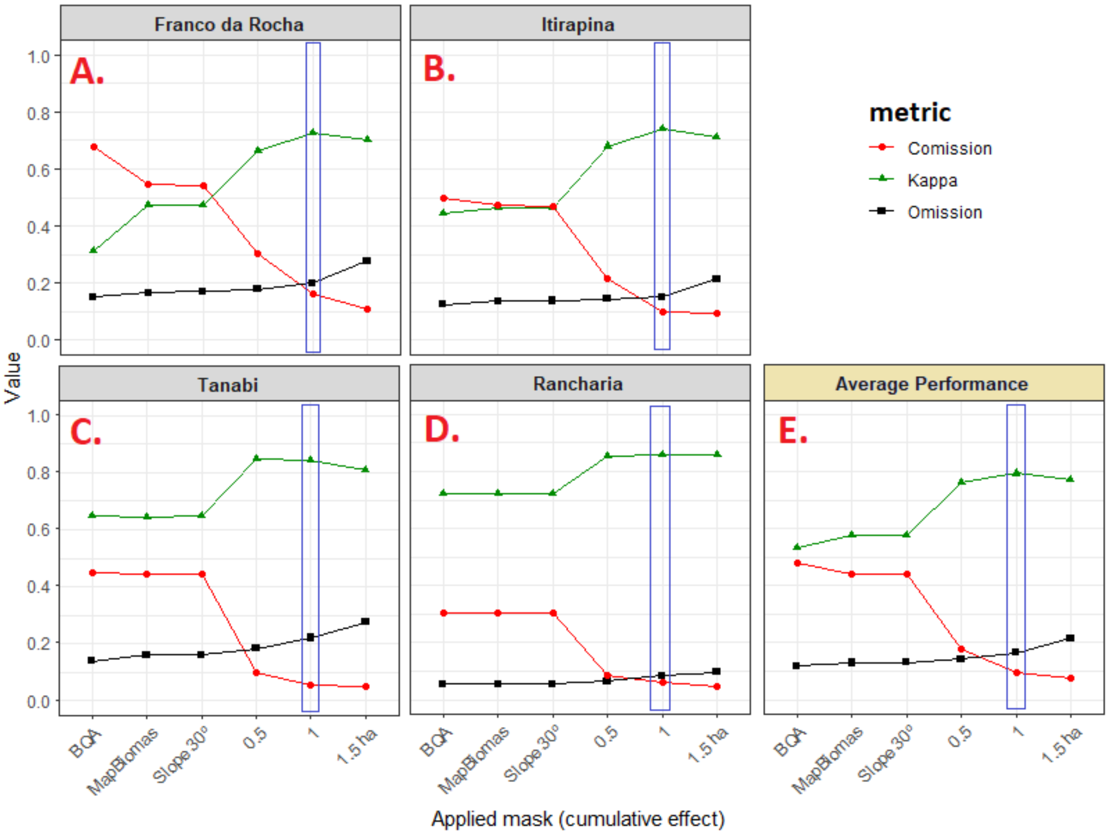


Figure 15. Performance of tested masks into each validation plots. The x-axis shows the cumulative effect of the different masks (e.g. “1 ha” refers to the result obtained by combining “BQA + MapBiomas + Slope 30 º + mask all burns less than 0.5 ha + mask all burns less than 1 ha”). Colored lines represent the CE (red), OE (black) and Kappa index (green). Dark blue rectangle point the selected combination of masks. Average performance (**E**) was computed by summarizing the mean from the results into the four validation plots.

Inspired in the MODIS burned area products strategy of error balancing (Giglio et al., 2015), we assume kappa > OE > CE as our error balancing strategy. In one hand we guarantee the maximum spatial correspondence by selecting the mask parameters with the highest kappa values. On the other hand, by prioritizing a higher OE than CE we guarantee that our burned area product is ever omitting more than committing, minimizing the risk of poor inferences in regional environmental analysis. Thus, we selected the spatial contiguity filter of ~1 ha (11 pixels of spatial contiguity using rook’s adjacency criterion – kappa = 0.79, CE= 0.09, OE= 0.16) as threshold and all fire scares below this value were excluded.

We detected that many of the previously mentioned error tendencies were corrected by the spatial contiguity filter. First, many of the sparse buildings wrongly classified as burned area by our algorithm were being masked (Figure 16A, B) as well as the spectral mixtures between water-soil and water-vegetation interfaces (Figure 16C, D).

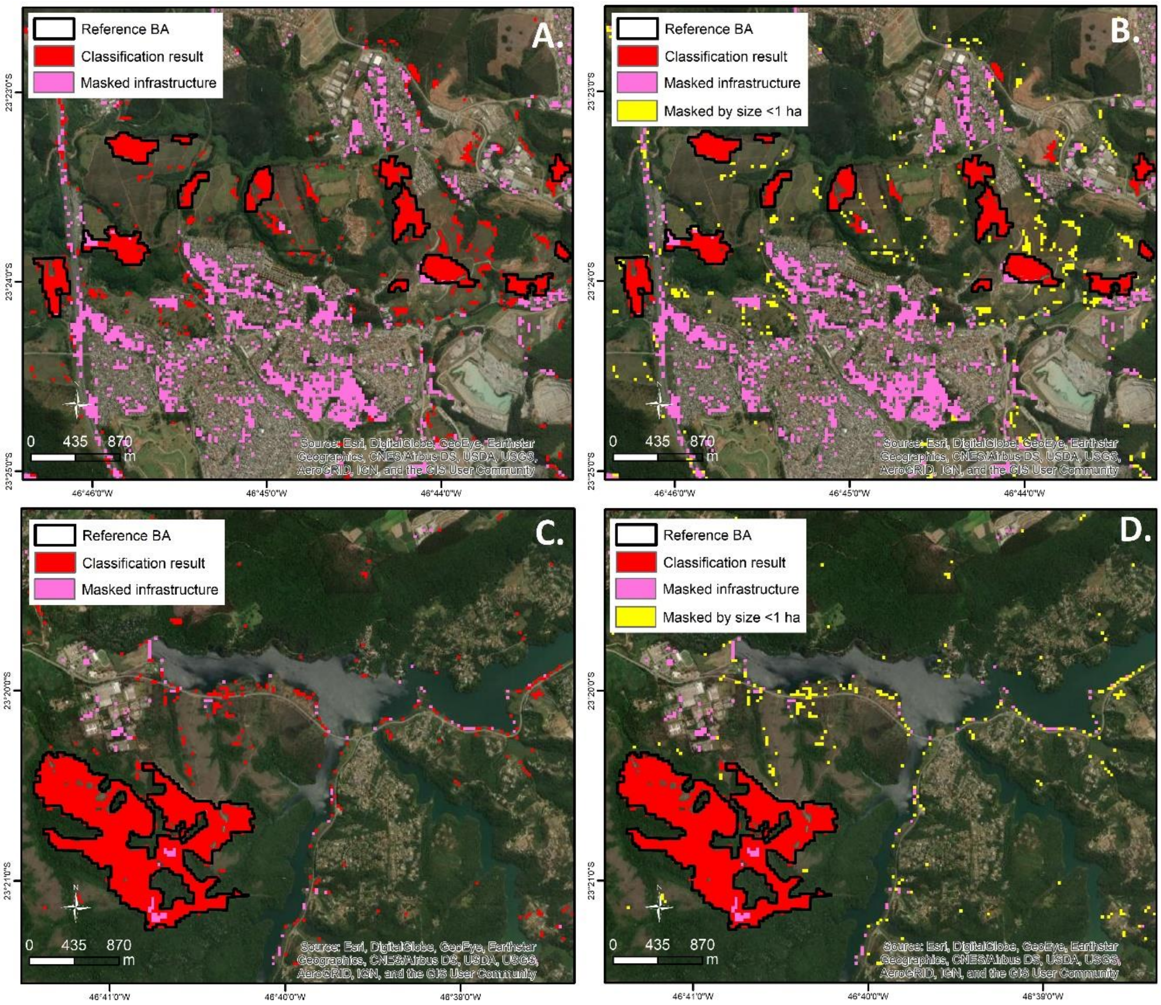

A.
Reference BA
Classification result
Masked infrastructure
0 435 870 m
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
23°23'0"S
23°24'0"S
23°25'0"S
46°46'0"W
46°45'0"W
46°44'0"W
B.
Reference BA
Classification result
Masked infrastructure
Masked by size <1 ha
0 435 870 m
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
23°23'0"S
23°24'0"S
23°25'0"S
46°46'0"W
46°45'0"W
46°44'0"W
C.
Reference BA
Classification result
Masked infrastructure
N
0 435 870 m
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
23°20'0"S
23°21'0"S
46°41'0"W
46°40'0"W
46°39'0"W
D.
Reference BA
Classification result
Masked infrastructure
Masked by size <1 ha
N
0 435 870 m
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
23°20'0"S
23°21'0"S
46°41'0"W
46°40'0"W
46°39'0"W

Figure 16. Franco da Rocha, 2018-08-30. Black polygons represent our reference burned area dataset. Red colored pixels represent pixels classified as burned area by our algorithm. Pink colored pixels represent the false-positive burned areas masked by MapBiomas. Yellow colored pixels represent the false-positive burned areas masked by minimum contiguity filter. Left boxes (**A** and **C**) show representative plots before the run of the minimum contiguity filter. Right boxes (**B** and **D**) shows the results of minimum contiguity mask by using the selected value (~ 1ha). We used a high resolution scene from ESRI Imagery as background in both figures.

Second, the native Cerrado grasslands (“campo limpo” and “campo sujo”) shows intensive phonological variations: a greenness peek into the mid wet-season and a high dehydration in the late dry season, making the spectral response of these areas in the late dry season a mixture of dry organic matter and a quartzarenic soil with high reflectance brightness. These pixels were frequently classified as false-positive burned areas (Figure 17A). However, these errors were successfully masked by the minimum spatial contiguity filter (Figure 17B). Like the sparse buildings, other small anthropic infrastructures (like roads, highways and railways) were also wrongly classified as burned areas (Figure 17C). Since these anthropic infrastructures are small-sized when compared to Landsat scale, all the false-positive errors caused by this pattern were easily removed by applying the minimum spatial contiguity filter (Figure 17D). Finally, we assume that the accuracy (kappa = 0.79) and errors (CE= 0.09, OE= 0.16) of our burned area product is balanced to make possible future regional scale environmental analysis, validating our hypothesis iv.

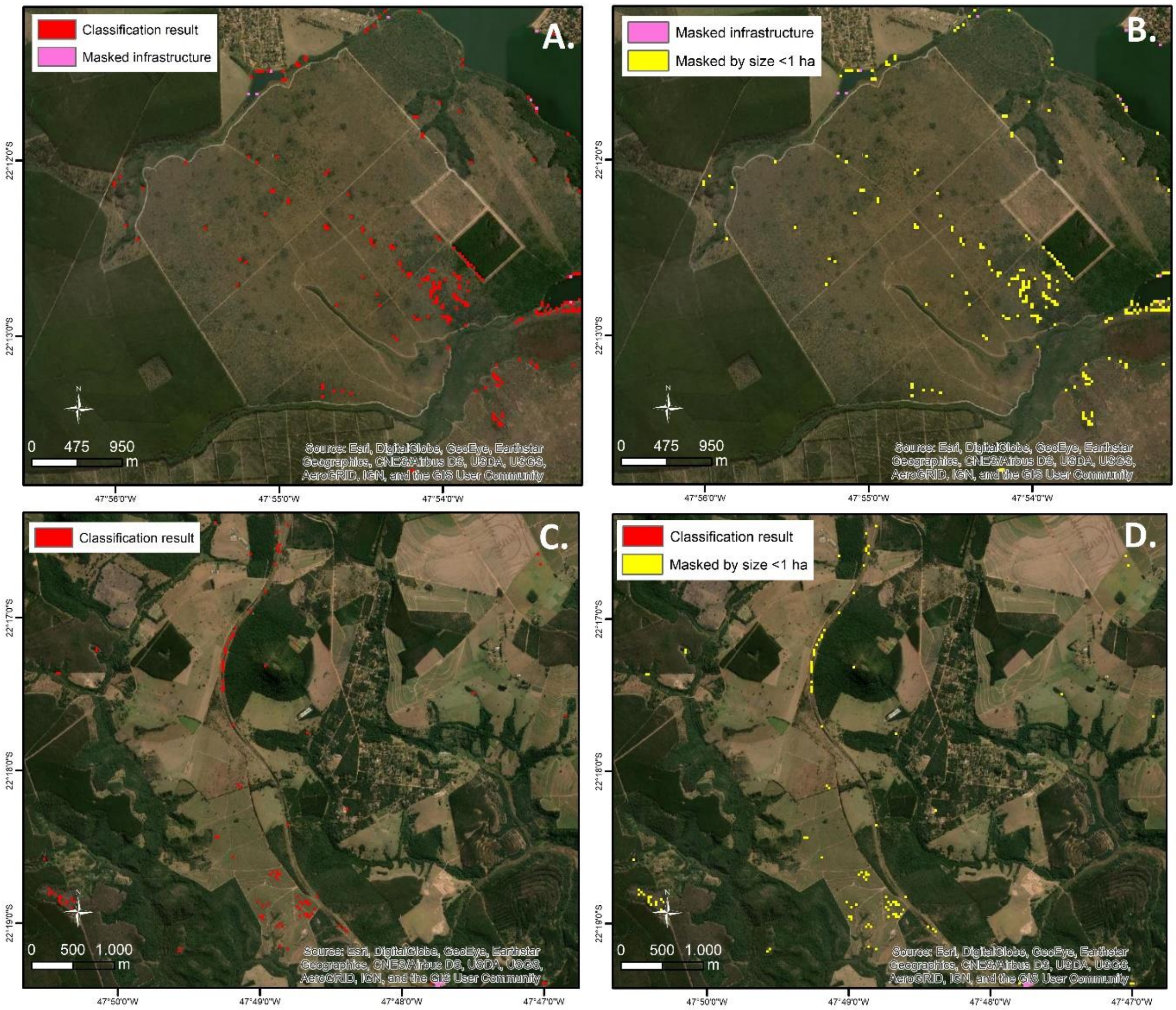
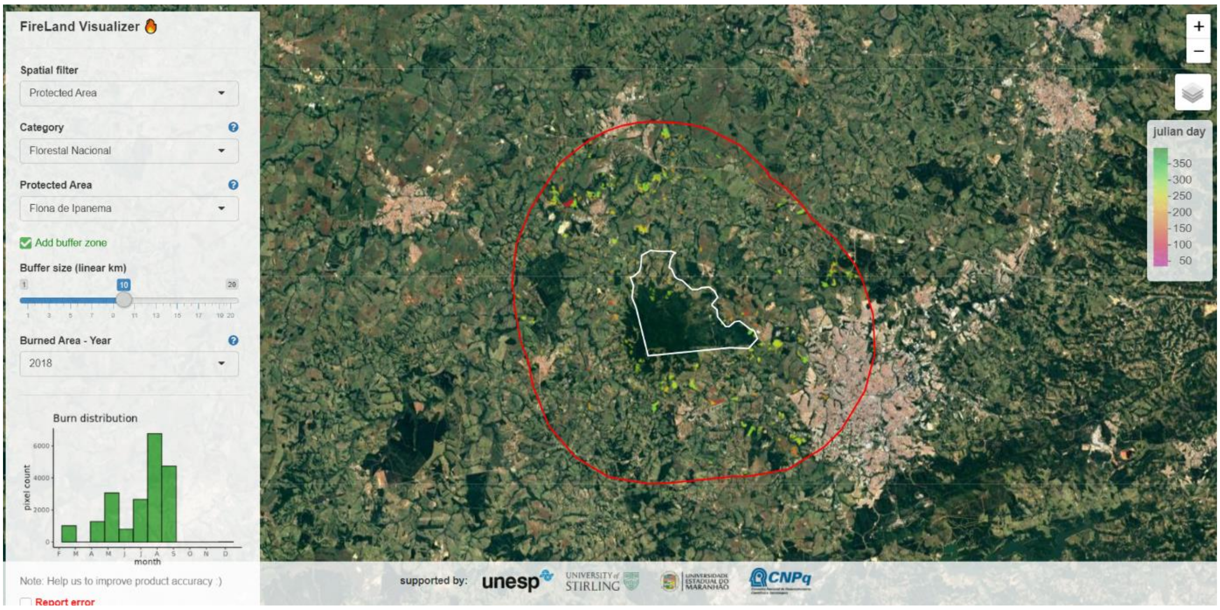


Figure 17. Itirapina, 2015-08-29. Red colored pixels represent pixels classified as burned area by our algorithm. Pink colored pixels represent the false-positive burned areas masked by MapBiomas mask. Yellow colored pixels represent the false-positive burned areas masked by minimum contiguity filter. Left boxes (**A** and **C**) show representative plots before the run of the minimum contiguity filter. Right boxes (**B** and **D**) show the results of minimum contiguity mask by using the selected value (~ 1ha). We used a high resolution scene from ESRI Imagery as background in both figures.

3.4. Final product and data access

We developed an R-Shiny web-application (Figure 18) to provide free access and navigation into our results by municipality or protected area in an interactive map (https://bit.ly/FireLandSP2). Interactive exploratory graphics were included and are recalculated every time that the end-user change spatial or temporal filters. We also implemented interactive buffer zone filters around protected areas to allow the users to inspect and assess possible human pressures near each protected area. Tools to enable the end-users to report the errors and implement their own improvements in the product are also planned and under development. Thus, by using our product as starting point, we pretend to launch the first collaborative Cerrado's burned area mapping platform.



FireLand Visualizer

Spatial filter: Protected Area

Category: Florestal Nacional

Protected Area: Flona de Ipanema

☑ Add buffer zone

Buffer size (linear km): 10

Burned Area - Year: 2018

Note: Help us to improve product accuracy :)

□ Report error

Figure 18. Graphical user-interface (UI) to access, visualize and analyze the final product.

3.5. Known issues and future development

Despite our efforts to provide a highly accurate product in regional scale, we detected some issues and biases that can affect analyses in local scale. Inspecting the final product outside our validation scope, we found that some large infrastructures (e.g. steel and petroleum industries) were not masked by the post-processing steps, being wrongly classified as burned area (Figure 19A, B). Several commission errors were also observed into managed floodplains for agriculture (e.g. rice, vegetables) (Figure 19C, D) and native floodplains locally known as “campo úmido” or “várzea” (Figure 19E, F). Thus, we report that local applications of this product by end-users need to be inspected and, if necessary, supervised by performing the necessary improvements.

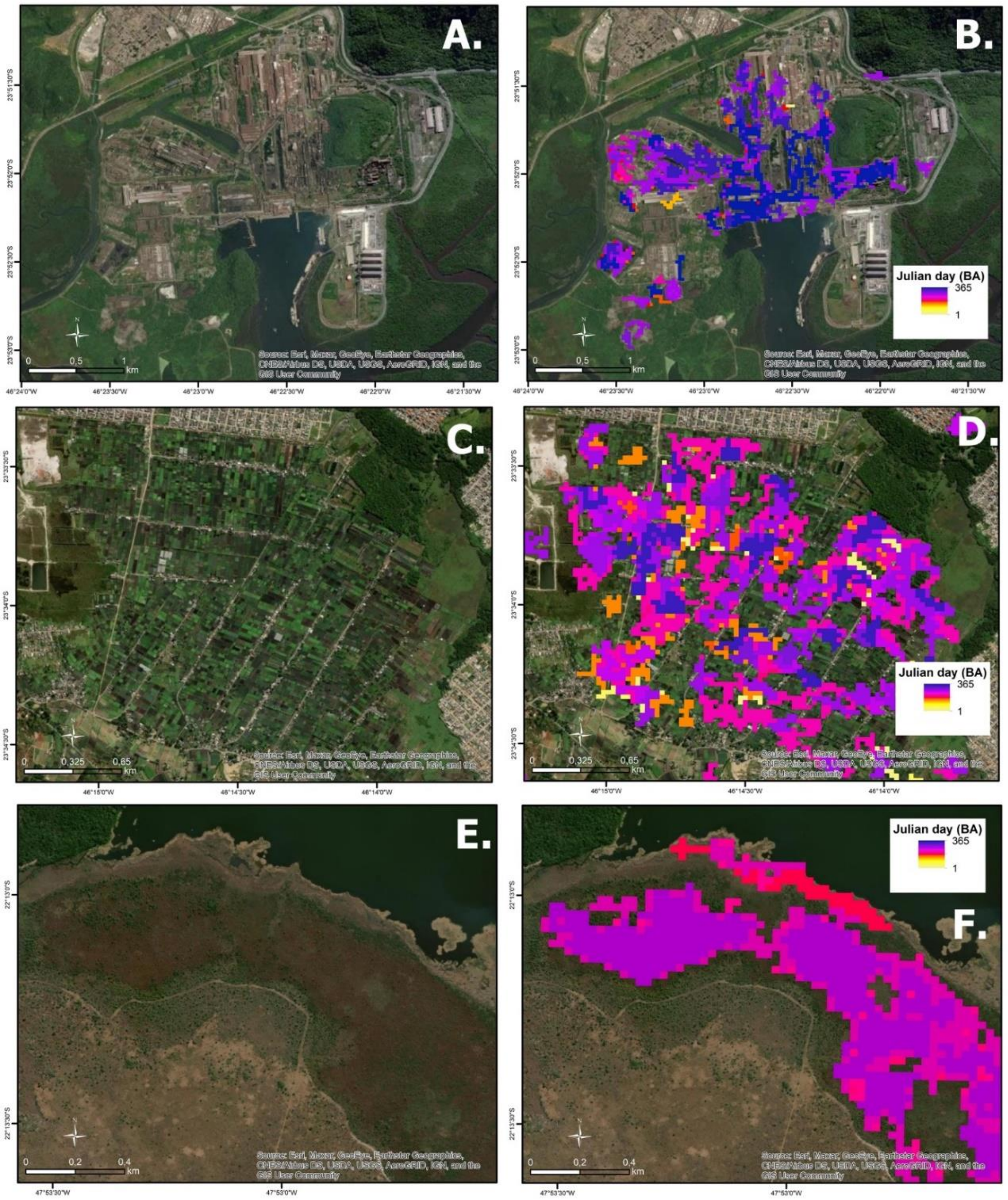


Figure 19. Known issues of the final product. All the figures refer to municipalities from São Paulo state and for burned area classifications from the year of 2018. **A/B.** Cubatão – "Siderúrgica Usiminas"; **C/D.** Mogi das Cruzes – Vegetable garden complex; **E/F.** São Carlos – Native floodplain "campo úmido de várzea". We used a high resolution scene from ESRI Imagery as background in both figures.

Another important aspect to be mentioned refers to the Landsat temporal resolution and their scenes availability. Post-fire vegetation responses over tropical savannas, like Cerrado, are quick and occur only a few months after the disturbance (Bowman et al., 2009; Coutinho, 1990). We point that despite the revisit of Landsat imagery occurs every 16 days, the availability of cloud-free scenes for the study area was rare (Supplementary Fig S1). So even if we have used scenes with until 75% of cloud-cover, it is possible that some burn scars have not been imaged before the vegetation recovery due the constant cloud-cover in repeated Landsat scenes, thus being invisible to Landsat sensors when employed alone (Alves et al., 2018; Veraverbeke et al., 2011). Furthermore, our training data of burned area considers a wide range of burn scars with different contents of ash presence. That is, if the ash spectral signal disappear from the terrain surface as a result from the rain and wind, this pixel probably would be misclassified as bare soil (Pereira, 2003; Trigg & Flasse, 2001). Thus, we report that applications of our product to analyze open Cerrado patches ("campo úmido", "campo limpo", "campo sujo") in local contexts need to be conducted with caution, if possible by using field data and empirical knowledge from the locals.

In this way, future versions of this algorithm will be developed by combining harmonized Landsat and Sentinel-2 imagery with the community contributions inside the Cerrado's collaborative burned area mapping platform. Thus, by summing efforts, we believe that the challenge of mapping burned areas in the highly anthropized cerrado can be overcome.

4. Conclusion

This study presented a reproducible methodology to generate a burned area algorithm by tuning and comparing different machine learning algorithms. In general, machine learning algorithms performed well to classify the LULCC and burned area in highly anthropized landscapes. RF was proved to be a better classifier than XGB and MARS, being used to classify and extract the burned area from a dense Landsat time-series. An adaptative post-processing have been implemented to balance omission (OE) and commission errors (CE) by using the strategy OE > CE, so that the final product showed to have quality to be employed in regional analysis (Kappa= 0.79).

This study generated the first burned area open dataset for the highly anthropized Cerrado. We recognize that this is only the first step, since some issues were reported in the known issues section and some improvements are necessary. However, we consider that this product represents many solid advancements in the fire mapping for the highly anthropized Cerrado's since no burned area data are available in regional scale for this context. Besides that, this study launched the first Cerrado's collaborative burned area mapping platform, providing free and instant access to our data and glimpsing that the challenge of mapping burned areas in complex contexts needs to be shared and overcome collectively.

5. Acknowledgments

D.E.C. is currently receiving financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) financial code 001 and received technical scholarship support from Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (grant#441968/2018-0).

6. Referencies

Active Fire Data | Earthdata. (2020). Retrieved July 29, 2020, from https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data

Alencar, A., Z. Shimbo, J., Lenti, F., Balzani Marques, C., Zimbres, B., Rosa, M., Arruda, V., Castro, I., Fernandes Márcico Ribeiro, J. P., Varela, V., Alencar, I., Piontekowski, V., Ribeiro, V., M. C. Bustamante, M., Eyji Sano, E., & Barroso, M. (2020). Mapping Three Decades of Changes in the Brazilian Savanna Native Vegetation Using Landsat Data Processed in the Google Earth Engine Platform. *Remote Sensing*, *12*(6), 924. https://doi.org/10.3390/rs12060924

ALOS Global Digital Surface Model "ALOS World 3D - 30m" (AW3D30). (2020). https://www.eorc.jaxa.jp/ALOS/en/aw3d30/index.htm

Área Queimada 30m. (n.d.). Retrieved July 29, 2020, from http://queimadas.dgi.inpe.br/queimadas/aq30m/

Atlântica, S. O. S. M. (2017). Atlas dos remanescentes florestais da Mata Atlantica período 2015-2016. *São Paulo, Brasil. Fundação SOS Mata Atlantica. Instituto Nacional Das Pesquisas Espaciais*.

Bastarrika, A., Alvarado, M., Artano, K., Martinez, M., Mesanza, A., Torre, L., Ramo, R., Chuvieco, E., Bastarrika, A., Alvarado, M., Artano, K., Martinez, M. P., Mesanza, A., Torre, L., Ramo, R., & Chuvieco, E. (2014). BAMS: A Tool for Supervised Burned Area Mapping Using Landsat Data. *Remote Sensing*, *6*(12), 12360–12380. https://doi.org/10.3390/rs61212360

Bastarrika, A., Chuvieco, E., & Martín, M. P. (2011). Mapping burned areas from landsat TM/ETM+ data with a two-phase algorithm: Balancing omission and commission errors. *Remote Sensing of Environment*, *115*(4), 1003–1012. https://doi.org/10.1016/j.rse.2010.12.005

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. In *ISPRS Journal of Photogrammetry and Remote Sensing* (Vol. 114, pp. 24–31). Elsevier B.V. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., & Hijmans, R. (2015). rgdal: Bindings for the Geospatial Data Abstraction Library 2017. *R Package Version 0.8-13*.

Borini Alves, D., Montorio Llovería, R., Pérez-Cabello, F., & Vlassova, L. (2018). Fusing Landsat and MODIS data to retrieve multispectral information from fire-affected areas over tropical savannah environments in the Brazilian Amazon. *International Journal of Remote Sensing*, *39*(22), 7919–7941. https://doi.org/10.1080/01431161.2018.1479790

Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., D'Antonio, C. M., Defries, R. S., Doyle, J. C., Harrison, S. P., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Marston, J. B., Moritz, M. A., Prentice, I. C., Roos, C. I., Scott, A. C., … Pyne, S. J. (2009). Fire in the Earth system. *Science (New York, N.Y.)*, *324*(5926), 481–484. https://doi.org/10.1126/science.1163886

Catal, C., & Diri, B. (2009). Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. *Information Sciences*, *179*(8), 1040–1058. https://doi.org/10.1016/j.ins.2008.12.001

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y., Fan, R., Yang, X., Wang, J., & Latif, A. (2018). Extraction of Urban Water Bodies from High-Resolution Remote-Sensing Imagery Using Deep Learning. *Water*, *10*(5), 585. https://doi.org/10.3390/w10050585

Coutinho, L. M. (1990). *Fire in the Ecology of the Brazilian Cerrado* (pp. 82–105). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-75395-4_6

Daldegan, G., de Carvalho, O., Guimarães, R., Gomes, R., Ribeiro, F., McManus, C., Daldegan, G. A., De Carvalho, O. A., Guimarães, R. F., Gomes, R. A. T., Ribeiro, F. D. F., & McManus, C. (2014). Spatial Patterns of Fire Recurrence Using Remote Sensing and GIS in the Brazilian Savanna: Serra do Tombador Nature Reserve, Brazil. *Remote Sensing*, *6*(10), 9873–9894. https://doi.org/10.3390/rs6109873

Dias, B. F. (2006). *Degradação ambiental: Os impactos do fogo sobre a diversidade do cerrado. In I. Garay and B. Becker (Eds.), Dimensões Humanas da Biodiversidade: O Desafio de Novas Relações Homem-Natureza no Século XXI.* Editora Vozes.

Doksum, K., Tang, S., & Tsui, K. W. (2008). Nonparametric variable selection: The EARTH algorithm. *Journal of the American Statistical Association*, *103*(484), 1609–1620. https://doi.org/10.1198/016214508000000878

Douaoui, A. E. K., Nicolas, H., & Walter, C. (2006). Detecting salinity hazards within a semiarid context by means of combining soil and remote-sensing data. *Geoderma*, *134*(1–2), 217–230. https://doi.org/10.1016/j.geoderma.2005.10.009

Durigan, G., & Ratter, J. A. (2016). The need for a consistent fire policy for Cerrado conservation. *Journal of Applied Ecology*, *53*(1), 11–15. https://doi.org/10.1111/1365-2664.12559

Ferlito, S., Adinolfi, G., & Graditi, G. (2017). Comparative analysis of data-driven methods online and offline trained to the forecasting of grid-connected photovoltaic plant production. *Applied Energy*, *205*, 116–129. https://doi.org/10.1016/j.apenergy.2017.07.124

Gao, B. C. (1996). NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, *58*(3), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., & Wolff, E. (2018). Very High Resolution Object-Based Land Use-Land Cover Urban Classification Using Extreme Gradient Boosting. *IEEE Geoscience and Remote Sensing Letters*, *15*(4), 607–611. https://doi.org/10.1109/LGRS.2018.2803259

Ghosh, A., & Joshi, P. K. (2014). A comparison of selected classification algorithms for mappingbamboo patches in lower Gangetic plains using very high resolution WorldView 2 imagery. *International Journal of Applied Earth Observation and Geoinformation*, *26*(1), 298–311. https://doi.org/10.1016/j.jag.2013.08.011

Giglio, L., Justice, C. O., Boschetti, L., & Roy, D. P. (2015). MCD64A1 MODIS. *Terra+ Aqua Burned Area Monthly L3 Global 500m SIN Grid V006 MCD64A1 (Https://Doi. Org/10.5067/MODIS/MCD64A1. 006)*.

Giles, P. T. (2001). Remote sensing and cast shadows in mountainous terrain. *Photogrammetric Engineering and Remote Sensing*, *67*(7), 833–840.

Gitelson, A. A., Merzlyak, M. N., & Lichtenthaler, H. K. (1996). Detection of red edge position and chlorophyll content by reflectance measurements near 700 nm. *Journal of Plant Physiology*, *148*(3–4), 501–508. https://doi.org/10.1016/S0176-1617(96)80285-9

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, *4*, 192–201. https://doi.org/10.1109/ICNC.2008.871

Hardisky, M. A., Klemas, V., & Smart, R. M. (1983). *The Influence of Soil Salinity, Growth Form, and Leaf Moisture on-the Spectral Radiance of ~ partina alterniflora Canopies*.

Hawbaker, T. J., Vanderhoof, M. K., Beal, Y. J., Takacs, J. D., Schmidt, G. L., Falgout, J. T., Williams, B., Fairaux, N. M., Caldwell, M. K., Picotte, J. J., Howard, S. M., Stitt, S., & Dwyer, J. L. (2017). Mapping burned areas using dense time-series of Landsat data. *Remote Sensing of Environment*, *198*, 504–522. https://doi.org/10.1016/j.rse.2017.06.027

Hayes, M. M., Miller, S. N., & Murphy, M. A. (2014). High-resolution landcover classification using random forest. *Remote Sensing Letters*, *5*(2), 112–121. https://doi.org/10.1080/2150704X.2014.882526

Hijmans, R. J., & van Etten, J. (2012). *raster: Geographic analysis and modeling with raster data. R package version 2.0–12*.

IBGE. (2014). *Cidades do Brasil*. https://cidades.ibge.gov.br/

IBGE | Censo 2010. (n.d.). Retrieved August 4, 2020, from https://censo2010.ibge.gov.br/

Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6444 LNCS*(PART 2), 152–159. https://doi.org/10.1007/978-3-642-17534-3_19

Jia, K., Liang, S., Zhang, L., Wei, X., Yao, Y., & Xie, X. (2014). Forest cover classification using Landsat ETM+ data and time series MODIS NDVI data. *International Journal of Applied Earth Observation and Geoinformation*, *33*(1), 32–38. https://doi.org/10.1016/j.jag.2014.04.015

Key, C. H., & Benson, N. C. (2006). Landscape assessment (LA). *In: Lutes, Duncan C.; Keane, Robert E.; Caratti, John F.; Key, Carl H.; Benson, Nathan C.; Sutherland, Steve; Gangi, Larry J. 2006. FIREMON: Fire Effects Monitoring and Inventory System. Gen. Tech. Rep. RMRS-GTR-164-CD. Fort Collins, CO: US*

Department Of , *164*.

Koutsias, N., & Karteris, M. (2000). Burned area mapping using logistic regression modeling of a single post-fire Landsat-5 Thematic Mapper image. *International Journal of Remote Sensing*, *21*(4), 673–687. https://doi.org/10.1080/014311600210506

Kronka, F. J. N., Nalon, M. A., Matsukuma, C. K., Kanashiro, M. M., Ywane, M. S. S., Lima, L., Guillaumon, J. R., Barradas, A. M. F., Pavão, M., & Manetti, L. A. (2005). Monitoramento da vegetação natural e do reflorestamento no Estado de São Paulo. *Simpósio Brasileiro de Sensoriamento Remoto*, *12*, 16–21.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Li, D. H. W., Chen, W., Li, S., & Lou, S. (2019). Estimation of hourly global solar radiation using Multivariate Adaptive Regression Spline (MARS) – A case study of Hong Kong. *Energy*, *186*, 115857. https://doi.org/10.1016/j.energy.2019.115857

LP DAAC - MCD64A1. (n.d.). Retrieved July 29, 2020, from https://lpdaac.usgs.gov/products/mcd64a1v006/

Lymburner, L., Beggs, P. J., & Jacobson, C. R. (2000). *Estimation of Canopy-Average Surface-Specific Leaf Area Using Landsat TM Data*.

Martín, M. P., & Chuvieco, E. (2006). Burnt Area Index (BAIM) for burned area discrimination at regional scale using MODIS data. *Article in Forest Ecology and Management*. https://doi.org/10.1016/j.foreco.2006.08.248

Mitri, G. H., & Gitas, I. Z. (2004). A semi-automated object-oriented model for burned area mapping in the Mediterranean region using Landsat-TM imagery. *International Journal of Wildland Fire*, *13*(3), 367. https://doi.org/10.1071/WF03079

Naghibi, S. A., Hashemi, H., Berndtsson, R., & Lee, S. (2020). Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. *Journal of Hydrology*, *589*, 125197. https://doi.org/10.1016/j.jhydrol.2020.125197

Oechsle, O., & Clark, A. F. (2008). Feature extraction and classification by genetic programming. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5008 LNCS*, 131–140. https://doi.org/10.1007/978-3-540-79547-6_13

Olsson, D. M., & Nelson, L. S. (1975). The nelder-mead simplex procedure for function minimization. *Technometrics*, *17*(1), 45–51. https://doi.org/10.1080/00401706.1975.10489269

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7376 LNAI*, 154–168. https://doi.org/10.1007/978-3-642-31537-4_13

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Paul, A. (1997). Rule-based classification of water in Landsat MSS images using the variance filter. *Photogrammetric Engineering & Remote Sensing*, *63*(5), 485491.

Pereira, A., Pereira, J., Libonati, R., Oom, D., Setzer, A., Morelli, F., Machado-Silva, F., & de Carvalho, L. (2017). Burned Area Mapping in the Brazilian Savanna Using a One-Class Support Vector Machine Trained by Active Fires. *Remote Sensing*, *9*(11), 1161. https://doi.org/10.3390/rs9111161

Pereira, J. M. C. (2003). Remote sensing of burned areas in tropical savannas. *International Journal of Wildland Fire*, *12*(3–4), 259–270. https://doi.org/10.1071/wf03028

Provost, F. (2000). *Machine Learning from Imbalanced Data Sets 101*. www.aaai.org

Qi, J., Chehbouni, A., Huete, A. R., Kerr, Y. H., & Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote Sensing of Environment*, *48*(2), 119–126.

https://doi.org/10.1016/0034-4257(94)90134-1

Ramo, R., & Chuvieco, E. (2017). Developing a Random Forest Algorithm for MODIS Global Burned Area Classification. *Remote Sensing*, *9*(11), 1193. https://doi.org/10.3390/rs9111193

Rosan, T. M., Aragão, L. E. O. C., Oliveras, I., Phillips, O. L., Malhi, Y., Gloor, E., & Wagner, F. H. (2019). Extensive 21st-Century Woody Encroachment in South America's Savanna. *Geophysical Research Letters*, *46*(12), 6594–6603. https://doi.org/10.1029/2019GL082327

Rouse, R. W. H., Haas, J. A. W., & Deering, D. W. (1974). Monitoring vegetation systems in the great plains with ERTS. *Third Earth Resources Technology Satellite (ERTS) Symposium*, 309–317.

Roy, D. P., & Boschetti, L. (2009). Southern Africa validation of the MODIS, L3JRC, and GlobCarbon burned-area products. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(4), 1032–1044. https://doi.org/10.1109/TGRS.2008.2009000

Shao, Y., Lunetta, R. S., Wheeler, B., Iiames, J. S., & Campbell, J. B. (2016). An evaluation of time-series smoothing algorithms for land-cover classifications using MODIS-NDVI multi-temporal data. *Remote Sensing of Environment*, *174*, 258–265. https://doi.org/10.1016/j.rse.2015.12.023

Simon, M. F., Grether, R., Queiroz, L. P. de, Skema, C., Pennington, R. T., & Hughes, C. E. (2009). Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proceedings of the National Academy of Sciences*, pnas.0903410106. https://doi.org/10.1073/PNAS.0903410106

Smith, A. M. S., Drake, N. A., Wooster, M. J., Hudak, A. T., Holden, Z. A., & Gibbons, C. J. (2007). Production of Landsat ETM+ reference imagery of burned areas within Southern African savannahs: comparison of methods and application to MODIS. *International Journal of Remote Sensing*, *28*(12), 2753–2775. https://doi.org/10.1080/01431160600954704

Smith, Alistair M.S., Wooster, M. J., Drake, N. A., Dipotso, F. M., Falkowski, M. J., & Hudak, A. T. (2005). Testing the potential of multi-spectral remote sensing for retrospectively estimating fire severity in African Savannahs. *Remote Sensing of Environment*, *97*(1), 92–115. https://doi.org/10.1016/j.rse.2005.04.014

Soares-Filho, B., Rajão, R., Macedo, M., Carneiro, A., Costa, W., Coe, M., Rodrigues, H., & Alencar, A. (2014). Cracking Brazil’s forest code. *Science*, *344*(6182), 363–364.

Stroppiana, D., Bordogna, G., Carrara, P., Boschetti, M., Boschetti, L., & Brivio, P. A. (2012). A method for extracting burned areas from Landsat TM/ETM+ images by soft aggregation of multiple Spectral Indices and a region growing algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*, *69*, 88–102. https://doi.org/10.1016/j.isprsjprs.2012.03.001

Team, R. C. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Topouzelis, K., & Psyllos, A. (2012). Oil spill feature selection and classification using decision tree forest on SAR image data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *68*(1), 135–143. https://doi.org/10.1016/j.isprsjprs.2012.01.005

Trigg, S., & Flasse, S. (2001). An evaluation of different bi-spectral spaces for discriminating burned shrub-savannah. *International Journal of Remote Sensing*, *22*(13), 2641–2647. https://doi.org/10.1080/01431160110053185

Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. *ACM International Conference Proceeding Series*, *227*, 935–942. https://doi.org/10.1145/1273496.1273614

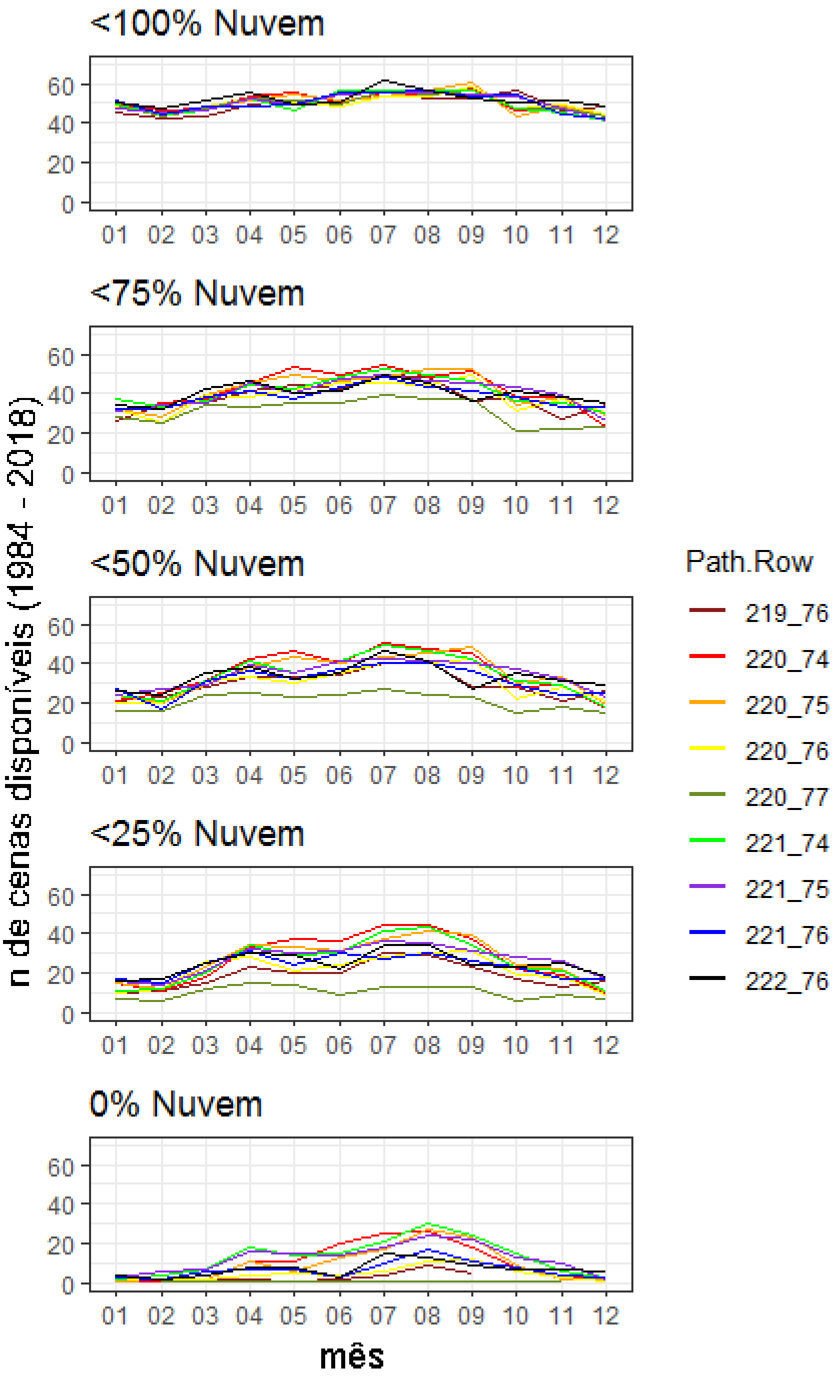
Veraverbeke, S., Lhermitte, S., Verstraeten, W., & Goossens, R. (2011). A time-integrated MODIS burn severity assessment using the multi-temporal differenced normalized burn ratio (dNBRMT). *International Journal of Applied Earth Observation and Geoinformation*, *13*(1), 52–58. https://doi.org/10.1016/j.jag.2010.06.006

Vicente, L. E., Souza Filho, C. R., & Perez Filho, A. (2005). Mapeamento de formações arenosas em fragmentos de Cerrado utilizando dados e produtos do sensor ASTER. *XII Simpósio Brasileiro de Sensoriamento Remoto. INPE, Goiânia*, 3419–3426.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, *10*(7), 373. https://doi.org/10.3390/atmos10070373

Supplementary

Suplemmentary Fig S1. Landsat scenes availability using different cloud cover thresholds.



Supplementary Table S2. List of scenes used in the product validation

Validation plot	**Scene list**
Franco da Rocha	LT05_L1TP_219076_19951002_20170106_01_T1 LT05_L1TP_219076_20030922_20161204_01_T1 LT05_L1TP_219076_20031024_20161203_01_T1 LC08_L1TP_219076_20170624_20170713_01_T1 LC08_L1TP_219076_20170726_20170810_01_T1 LC08_L1TP_219076_20170827_20170914_01_T1 LC08_L1TP_219076_20170912_20170928_01_T1 LC08_L1TP_219076_20180424_20180502_01_T1 LC08_L1TP_219076_20180510_20180517_01_T1 LC08_L1TP_219076_20180830_20180911_01_T1
Itirapina	LT05_L1TP_220075_19850319_20170219_01_T1 LT05_L1TP_220075_19850420_20170219_01_T1 LT05_L1TP_220075_19850506_20170219_01_T1 LT05_L1TP_220075_19850709_20170219_01_T1 LT05_L1TP_220075_19850810_20170219_01_T1 LT05_L1TP_220075_19850911_20170218_01_T1 LT05_L1TP_220075_19880224_20170209_01_T1 LT05_L1TP_220075_19880327_20170209_01_T1 LT05_L1TP_220075_19880701_20170208_01_T1 LT05_L1TP_220075_19880717_20170208_01_T1 LT05_L1TP_220075_19880802_20170207_01_T1 LT05_L1TP_220075_19880919_20170206_01_T1 LT05_L1TP_220075_19881106_20170205_01_T1 LT05_L1TP_220075_19881208_20170205_01_T1 LC08_L1TP_220075_20150117_20180528_01_T1 LC08_L1TP_220075_20150202_20170413_01_T1 LC08_L1TP_220075_20150423_20170409_01_T1 LC08_L1TP_220075_20150509_20170409_01_T1 LC08_L1TP_220075_20150525_20170408_01_T1 LC08_L1TP_220075_20150610_20170408_01_T1 LC08_L1TP_220075_20150728_20170406_01_T1 LC08_L1TP_220075_20150813_20170406_01_T1 LC08_L1TP_220075_20150829_20170405_01_T1 LC08_L1TP_220075_20151016_20170403_01_T1 LC08_L1TP_220075_20180314_20180320_01_T1 LC08_L1TP_220075_20180517_20180604_01_T1 LC08_L1TP_220075_20180618_20180703_01_T1
Tanabi	LT05_L1TP_221074_20060726_20161120_01_T1

	LT05_L1TP_221074_20060811_20161119_01_T1 LT05_L1TP_221074_20060912_20161118_01_T1 LC08_L1TP_221074_20160416_20170326_01_T1 LC08_L1TP_221074_20160502_20170325_01_T1 LC08_L1TP_221074_20160721_20170323_01_T1 LC08_L1TP_221074_20160923_20170321_01_T1 LC08_L1TP_221074_20161009_20170320_01_T1 LC08_L1TP_221074_20180422_20180502_01_T1 LC08_L1TP_221074_20180508_20180517_01_T1 LC08_L1TP_221074_20180625_20180704_01_T1 LC08_L1TP_221074_20180727_20180731_01_T1 LC08_L1TP_221074_20180913_20180928_01_T1
Rancharia	LT05_L1TP_222076_19850128_20170219_01_T1 LT05_L1TP_222076_19850605_20170219_01_T2 LT05_L1TP_222076_19850824_20170218_01_T1 LT05_L1TP_222076_19850909_20170218_01_T1 LT05_L1TP_222076_19851112_20170218_01_T1 LT05_L1TP_222076_20010905_20161211_01_T1 LT05_L1TP_222076_20011007_20161210_01_T1 LT05_L1TP_222076_20011226_20161210_01_T1 LC08_L1TP_222076_20170512_20170525_01_T1 LC08_L1TP_222076_20170715_20170727_01_T1 LC08_L1TP_222076_20170901_20170915_01_T1 LC08_L1TP_222076_20180515_20180604_01_T1 LC08_L1TP_222076_20180531_20180614_01_T1 LC08_L1TP_222076_20180718_20180731_01_T1