COMBINING ENVIRONMENTAL AND LANDSAT ANALYSIS READY DATA FOR VEGETATION MAPPING: A CASE STUDY IN THE BRAZILIAN SAVANNA BIOME

H. N. Bendini[1],*, L. M. G. Fonseca[1], M. Schwieder[2], P. Rufin[2], T. S. Korting[1], A. Koumrouyan[1], P. Hostert[2,3]

[1] INPE, National Institute for Space Research, Sao Jose dos Campos, Brazil - (hugo.bendini, leila.fonseca, thales.korting)@inpe.br
[2] Geography Department, Humboldt-Universität zu Berlin, Berlin, Germany -
(marcel.schwieder, philippe.rufin, patrick.hostert)@geo.hu-berlin.de
[3] IRI THESys, Integrative Research Institute on Transformations of Human-Environment Systems, Berlin, Germany

KEY WORDS: Vegetation mapping, Cerrado, Phenology, Data mining, Random Forest

ABSTRACT:

The Cerrado biome in Brazil covers approximately 24% of the country. It is one of the richest and most diverse savannas in the world, with 23 vegetation types (physiognomies) consisting mostly of tropical savannas, grasslands, forests and dry forests. It is considered as one of the global hotspots of biodiversity because of the high level of endemism and rapid loss of its original habitat. This work aims to analyze the potential of Landsat Analysis Ready Data (ARD) in combination with different environmental data to classify the vegetation in the Cerrado in two different hierarchical levels. Here we present results of a pixel-based modelling exercise, in which field data were combined with a set of input variables using a Random Forest classification approach. On the first hierarchical level, with the three classes savanna, grasslands and forest, our model results reached f1-scores of 0.86, 0.87 and 0.85 leading to an overall accuracy of 0.86. In the second hierarchical level we differentiated a total of 12 vegetation physiognomies with an overall accuracy of 0.77.

1. INTRODUCTION

The Brazilian Savanna, also known as Cerrado, is the second largest biome in Brazil (MMA, 2015), considered as a biodiversity hotspot and providing environmental services of global importance. Despite that, the Cerrado has lost around 88 Mha (46%) of its native vegetation with a projection that 31-34% of the remaining biome is likely to be cleared by 2050 (Strassburg et al., 2017). The rate of conversion of native Cerrado vegetation is up to two times the conversion observed in the Amazon in the past five years (Rocha et al., 2011). Most of the native vegetation conversion tends to occur in areas with dense vegetation that have favorable climate and soil conditions and in flat terrains that are suitable for mechanized farming (Alencar et al., 2020). The conversion of natural vegetation into agricultural land is leading to major carbon emissions (Noojipady et al., 2017) and biodiversity loss (Ratter et al., 1997), stressing the importance of frequent mapping approaches that enable to monitor and assess ongoing change processes.

The Brazil Investment Plan (BIP) under the Forest Investment Program (FIP) seeks to promote sustainable land use and forest management improvement in the Cerrado Biome in order to reduce pressure on remaining forests, reduce greenhouse gas (GHG) emissions and increase carbon dioxide sequestration (Tuchschneider, 2013). As part of BIP, the project "Development of systems to prevent forest fires and monitor vegetation cover in the Brazilian Cerrado" aims to improve Brazil's capacity to monitor deforestation, prevent the risk of forest fires and improve models for estimating greenhouse gas (GHG) emissions, making tools and data available to environmental agencies [1].

The project will provide the basis for improving the management of water, forest and soil resources in the Brazilian Cerrado, which, together with other projects financed by the FIP in Brazil, should promote the sustainable management of forests. In the context of this project, one of the activities it to modify the existing land cover classification system for the Cerrado developed by IBGE (Brazilian Institute of Geography and Statistics) on the basis of the Food and Agriculture Organization of the United Nations (FAO) Land Cover Classification System framework. Therefore, it will be possible to discriminate forest from non-forest vegetation taking into account the spectrum of structural vegetation complexity in the Cerrado. However, mapping heterogeneous tropical areas, such as the Cerrado, is challenging due to the natural, climatic and topographic factors and the peculiarities of the characteristic physiognomies.

To make mapping approaches comparable Ribeiro and Walter (2008) divided the major Cerrado formations into a dominant herbaceous stratum (Grasslands), shrublands (savannas) and a woody-dominated stratum (Forests).

It has been shown that remote sensing based approaches have the potential for mapping these classes, but the strong seasonality of natural vegetation and the spectral ambiguities between some physiognomies makes it hard to differentiate them (Sano et al., 2010, Grecchi et al., 2013).

While the Brazilian National Institute for Space Research's official deforestation and land use maps still rely mostly on visual image interpretation (INPE, 2019b), some (semi-) automatic approaches have been developed by Ferreira et al. (2007) , Neves et al. (2019) , Girolamo-Neto et al. (2017) , Schwieder et al. (2016) . Even though these approaches are useful in specific case studies, they were usually restricted to comparably small extents and do not account for variations in environmental characteristics present across the entire Cerrado extent.

Previous studies highlighted the benefits of dense remote sensing time series, derived land surface phenological metrics (LSP)

* Corresponding author
[1] More information in: http://fip.mma.gov.br/projeto-fm/

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-953-2020 | © Authors 2020. CC BY 4.0 License.

and analyzed their relationship to the grassland-savanna-forest gradient of the Cerrado. They suggested the importance of integrating other environmental variables, such as terrain (Sano et al., 2019, Bendini et al., 2019b, Schwieder et al., 2018).

Recent advances in remote sensing technologies offer great opportunities for mapping land use and land cover over large extents. Several operational sensor systems are currently in orbit and the development of infrastructure for remotely sensed data storage and dissemination enables to consistently derive ARD (Potapov et al., 2020, Frantz, 2019). This development is catalyzed by the accessibility of cloud computing platforms (e.g. Amazon Web Services (AWS), Google Earth Engine (GEE) and Microsoft Azure), and the rapid evolution of machine learning approaches in the field of remote sensing. Together, these developments enable to explore the full potential of integrated data analyses, in which metrics derived from time series of ARD (e.g., phenological or spectral-temporal metrics) are combined with environmental data that are meaningful for a specific domain.

Recently, Alencar et al. (2020), as part of the MapBiomas project (https://mapbiomas.org), which aims to generate annual land use and land cover classification of Brazil, proposed a methodology based on GEE, in which they combined mosaics of spectral metrics, including sub-pixel fractions, indices, individual spectral bands of all Landsat-7 ETM+ (Enhanced Thematic Mapper Plus) and Landsat-8 OLI (Operational Land Imager) imageries between 1985 to 2017, and slope data from ALOS (Advanced Land Observing Satellite) global digital surface model with 30 meters (m) spatial resolution. This approach enabled them to classify the vegetation of the entire Cerrado on an annual basis for the considered period. They achieved a very promising overall accuracy of 0.71 considering the 33 years. Although this is a great effort and contributes to the analysis of land changes over time, a significant confusion between grasslands and savannas is apparent in these data. Besides, they do not account for all vegetation types present in the Brazilian Cerrado. These limitations are critical in the context of public policy decisions (e.g., based on the forest code) and, therefore, classification errors may affect the implementation of environmental conservation efforts in this highly threatened biome.

Large-scale mapping of the Cerrado vegetation using remote sensing technologies is still a challenge due to the high spatial variability and spectral similarity among its vegetation types. To the best of our knowledge, there is no research published that differentiated the natural vegetation for the entire Cerrado, with a high level of thematic detail in terms of vegetation physiognomies.

Therefore, the objective of this study is to analyze the potential of Landsat ARD combined with various environmental data to classify the different physiognomies in the Cerrado based on two hierarchical levels of thematic detail.

2. MATERIALS AND METHODS

2.1 Fieldwork and reference data

Reference data were collected during roadside surveys across five thousand kilometers on the most important remaining areas of natural vegetation in the biome Cerrado during March and July, 2019, in which a group of specialists on vegetation visually identified the classes and registered them on a Global Positioning System (GPS)-enabled tablet. The stopping points were selected randomly.

A mosaic composed of the most recent available Landsat-8 OLI images was used to navigate along the regions by using a GPS device connected to a tablet. High resolution images available in Google Earth were used as auxiliary data. We used the physiognomy definitions described in Table 1, based on the first and second hierarchical level of Ribeiro and Walter (2008) classification system. These physiognomies were defined by Ribeiro and Walter (2008) and consist in a hierarchical classification structure. The first hierarchical level (referred as level-1) consists on three classes: grassland, savanna and forest; which are further split in a total of 12 sub classes in level-2.

We compiled the reference data for both thematic levels from several field campaigns and data collection efforts as follows. We collected 652 ground-truth data points during the roadside surveys: savanna (479), forest (92), grassland (81), from level-1, and for level-2: Campo limpo (14), Campo rupestre (58), Campo sujo (9), Cerradao (9), Cerrado rupestre (10), Cerrado sensu stricto (381), Mata de Galeria and Ciliar (61), Mata de galeria (32), Mata seca (22), Palmeiral (14) and Vereda (74).

These datasets were completed with the official vegetation map on a scale of 1:250,000 produced by IBGE in 2012 and updated in 2013. This map was produced by the RADAMBRASIL project [2] on a scale of 1:1,000,000 and later adaptations were carried out to the one used in this study. The legend of the vegetation in the Cerrado utilized on the map refers to the one published in the technical manual of Brazilian vegetation (IBGE, 2012).

We utilized data kindly provided by the State Environmental Departments of Brazil, and also included ground samples from a public repository (Câmara et al., 2019) . Manually collected samples, based on Google Earth image visual inspection, were also included.

We prepared annual vegetation cover layers for year 2014 utilizing PRODES data (INPE, 2019a) [3], which includes forest and deforested polygons. We produced a mask based on PRODES data to remove samples that were within deforestation areas. We proceeded a quality screen on samples by visual inspection based on Google Earth imagery.

Finally, our reference database comprised a total of 2828 samples, distributed in the following way. Level-1: savannas (1250), grasslands (805) and forests (773). For the level-2 classes, the database contained samples for Campo limpo (276), Campo rupestre (210), Campo sujo (319), Cerradão (160), Cerrado rupestre (162), Cerrado sensu stricto (580), Ipuca (91), Mata riparia (446), Mata seca (76), Palmeiral (135), Parque de Cerrado (246), Vereda (127). For this work, as we used data from different sources, we proceeded harmonization between both IBGE legend (IBGE, 2012) and Ribeiro and Walter (2008). Figure 1 shows the spatial distribution of the samples.

[2] The Amazon Radar Project (RADAMBRASIL Project, was created in July, 1975, to collect data on mineral resources, soils, vegetation, land use and cartography of the national territory, with the aim of an integrated mapping of the natural resources

[3] Available from http://terrabrasilis.dpi.inpe.br/download/dataset/cerrado-prodes/vector/prodes_cerrado_2000_2018_v20190405.zip (download date: 17-Dec-19)

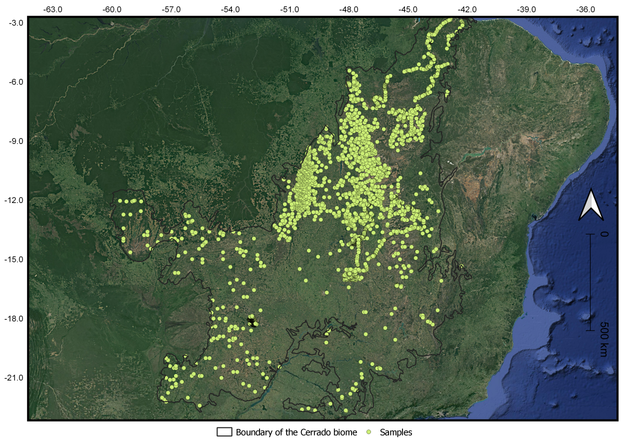

Figure 1. Distribution of the reference samples (n = 2828) across Cerrado biome.

The description of the physiognomies of the Cerrado on the first and second hierarchical level of (Ribeiro et al., 2008) classification system is shown in the Table 1.

2.2 Predictor variables

In this section we will describe the predictor variables. We selected informative variables that condition the different forms of life that characterize the physiognomies described by Ribeiro and Walter (2008) .

We derived Land Surface Phenological metrics (LSP) from a dense Landsat enhanced vegetation index (EVI) time series of the phenological season 2013 - 2014. Time series data gaps, due to sensor errors or cloud cover, were interpolated using a radial basis convolution filter (Schwieder et al., 2016, Bendini et al., 2019a). A total of 11 LSP were derived using TIMESAT (Jönsson, Eklundh, 2004), extracted for the seasonal cycle observed in the EVI time series. These metrics relate for example to the start and end of the season, the maximum EVI value of the season or the amplitude and are explained in detail in Jönsson, Eklundh (2004) .

Spectral temporal metrics (STM) (Griffiths et al., 2013, Rufin et al., 2015) were calculated for the dry season of the year 2014, consisting of the surface reflectance (SR) median values of the cloud free pixels of all spectral bands of Landsat images acquired between June and August 2014.

To account for variations in less dynamic environmental variations we included soil property maps from the Global Gridded Soil Information (SoilGrid) database (Hengl et al., 2017). This database is based on global compilation of soil profile data and environmental layers. The outputs of SoilGrids are global soil property maps at six standard depth intervals at a spatial resolution of 250 meters. We included variables related to sand, clay, silt content, and organic carbon concentration as well as the pH of the soil in 30 centimeters (cm) depth.

Table 1. Description of physiognomies on the 1st and 2nd hierarchical level of Ribeiro and Walter (2008) legend.

Level-1	Level-2	Description
Grassland	Campo limpo	Predominantly herbaceous, with rare shrubs and complete absence of trees. It is found more frequently on slopes, plateaus, water sources, surrounding the Veredas and on the edge of gallery forests.
	Campo rupestre	Predominantly herbaceous and shrubby, with occasional presence of trees up to 2m and occupying stretches of rocky outcrops. It usually occurs at altitudes above 900m.
	Campo sujo	Exclusively herbaceous, with sparse shrubs and sub-shrubs often made up of less developed individuals of Cerrado tree species.
Savanna	Cerrado sensu stricto	Low and tortuous trees, with irregular and twisted branches, usually with evidence of burning. In the rainy season the sub-shrub and herbaceous strata become exuberant due to their rapid growth.
	Cerrado rupestre	A subtype of Cerrado sensu stricto that occurs in rocky environments.
	Parque de Cerrado	Characterized by the presence of trees grouped in small elevations of land, known as ”murundus”. The trees have an average height of 3 to 6m and form a tree cover of 5% to 20%.
Forest	Cerradao	Forest with similarities to savannas due to species composition. Tree cover can range from 50 to 90%. The average height of the trees varies from 8 to 15m. Although it may be perennial, many species show deciduousness.
	Mata de Galeria and Ciliar	Mata de Galeria follows small rivers forming corridors. Perennial and with sudden transition to savannas. The average height varies between 20 and 30m, with an overlap of crowns providing tree coverage of 70 to 95%. Mata Ciliar is alongside medium and large rivers, where arboreal vegetation does not form galleries.
	Mata Seca	Characterized by different levels of deciduous trees, depending on conditions and mainly on the depth of the soil. The average height of the trees varies between 15 and 25m. Tree cover of 50 to 95%.
Savanna	Palmeiral	Characterized by the presence of a single species of arboreal palm. In general, they are found on well-drained land, although they also occur on poorly drained land, where galleries can follow the drainage lines.
	Vereda	Physiognomy with the predominance of *Mauritia Flexuosa* arboreal palm (Buritis), amid more or less dense clusters of shrub-herbaceous species. The Veredas are surrounded by Campo Limpo, which is generally flooded.
Forest	Ipuca	Ipucas are fragments of alluvial seasonal semi deciduous forests that flood every six months and occur in the region of the Araguaia plain due to the meeting of the Cerrado, Amazonian and Pantanal biomes.

https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-953-2020 |

Height Above the Nearest Drainage (HAND) is a quantitative topographic algorithm based on SRTM (Shuttle Radar Topography Mission) data (Rennó et al., 2008). It uses topographic data to obtain the vertical distance of each pixel in the watershed in relation to the drainage. In this way, it is strongly related to the water conditions in the soil and, therefore, to the natural vegetation present in the land cover. In this work, the HAND data for the Cerrado were generated from the SRTM with spatial resolution of 1 arc-second in which faults were corrected by the USGS (United States Geological Survey) based on the 3-arc-second SRTM. The processes for generating HAND are as follows: refilling of sinks, calculation of flow direction and accumulation area, generation of drainage channels. The processing task was carried out in TerraHidro (Abreu et al., 2012).

Terrain data such as elevation, slope, vertical and horizontal curve were derived from both Topodata database (de Morisson Valeriano, de Fátima Rossetti, 2012) as well as Tandem-X DEM (Gruber et al., 2012).

Table 2 shows all the selected variables, their brief description and its units of measurement.

Predictor	Description [units]
TANDEM_X_90	Altitude [m]
Topodata_HN	Horizontal curvature [m]
Topodata_VN	Profile curvature [m]
Topodata_SN	Slope[%]
BDTICM	Absolute depth to bedrock [mc]
BLDFIE	Bulk density [kg/m^3]
CECSOL	Cation Exchange Capacity [cmolc/kg]
CLYPPT	Clay particles [%]
OCSTHA	Soil organic carbon stock [ton/ha]
ORCDRC	Soil organic carbon content [permille]
PHIHOX	pH index measured in water [pH]
SLTPPT	Silt particles [%]
HAND_1K	Vertical Distance [m]
LSP_Amp	Amplitude [EVI]
LSP_BV	Base value [EVI]
LSP_EoS	End of Season [day of the year - DOI]
LSP_EoSVal	End of Season value [EVI]
LSP_LDer	Left Derivative [EVI]
LSP_LoS	Length of Season [DOI]
LSP_Mfit	Peak [EVI]
LSP_MoS	Middle of Season [DOI]
LSP_RDer	Right Derivative [EVI]
LSP$_{soS}$	Start of Season [DOI]
LSP_SoSVal	Start of Season [EVI]
STM_blue_med	median SR at Blue band [%]
STM_green_med	median SR at Red band [%]
STM_nir_med	median SR at NIR band [%]
STM_swirI_med	median SR at SWIR I [%]
STM_red_med	median SR at Red band [%]
STM_swirII_med	median SR at SWIR II band [%]

Table 2. Selected variables, their brief description and its units of measurement.

We re-sampled all data to 30 m x 30 m and tiled it into 60 x 60 km tiles for mass processing using FORCE (Frantz, 2019).

2.3 Classification and Validation

The predictor variables were used together with the reference data to train a Random Forest classifier (RF). RF is a non-parametric machine learning algorithm that is based on decision trees. As individual decision trees are prone to errors, RF uses an ensemble of many decision trees that were independently trained with random subsets of the input data to overcome this limitation (Breiman, 2001). The algorithm implementation in R (R Core Development Team, 2019) further allows to assess the variable importance of each input variable based on the Gini coefficient (Liaw et al., 2002).

The classification accuracy was assessed using Monte Carlo simulation, in which 1000 simulations were carried out by randomly selecting 70% of the samples to train the RF classification model, while the remaining 30% were used for validation. In each iteration, a confusion matrix was calculated, and the average confusion matrix was used to derive the overall accuracy and the class-wise f1-scores.

Additionally, the model was validated using independent ground truth samples, which were randomly drawn from a land cover reference map of the Brasilia National Park. This map was generated by the Image Processing and GIS Laboratory of Federal University of Goiás (LAPIG/UFG) for the year 2010, based on high resolution imagery, and the map created by Manuel et al. (2003). For a visual comparison, we used the final model to map the spatial patterns of the observed vegetation in the Brasilia National Park.

3. RESULTS AND DISCUSSIONS

3.1 Results

On the hierarchical level-1 our model results reached an overall accuracy of 0.86, ranging from 0.82 to 0.90 after 1000 iterations, with class-wise f1-scores of 0.86, 0.87 and 0.85 for the classes savanna, grasslands and forest. The producers accuracies for these classes were respectively 0.85, 0.90 and 0.87. The related confusion matrix is shown in Figure 2.

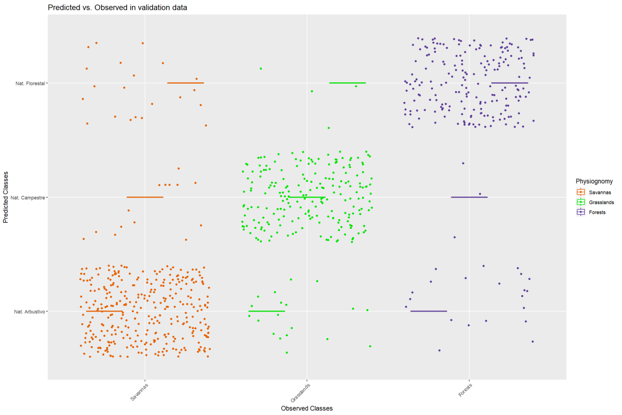

Figure 2. Confusion matrix for the classification model of Level-1.

The level-2 classification with 12 classes was assessed with a mean overall accuracy of 0.77, ranging from 0.72 to 0.81 Figure 3 shows the confusion matrix for the model for classifying level-1.

The five most important variables for the level-1 classification model were the Tandem-X elevation (TANDEM_X_90), soil fractions of clay (CLYPPT), the median of the surface reflectance of the red band

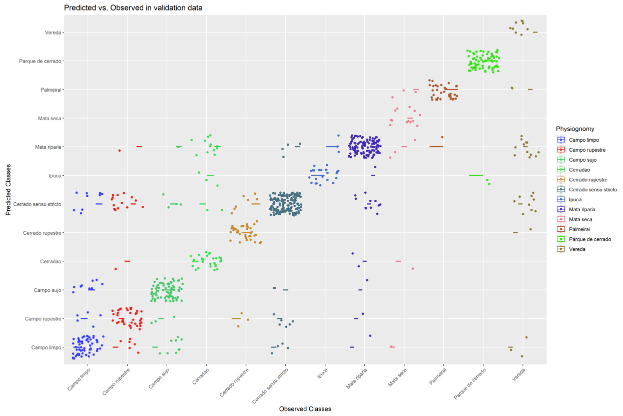

Figure 3. Confusion matrix for the classification model of Level-1.

(STM_2014_red_med), the EVI value during the start of the season (LSP_Metrics_2014_SoSVal) and the minimum EVI value during the season (LSP_Metrics_2014_BV). Figure 4 shows 3D-scatterplot of three most important variables for the Level-1 classification model based on the Gini importance index.

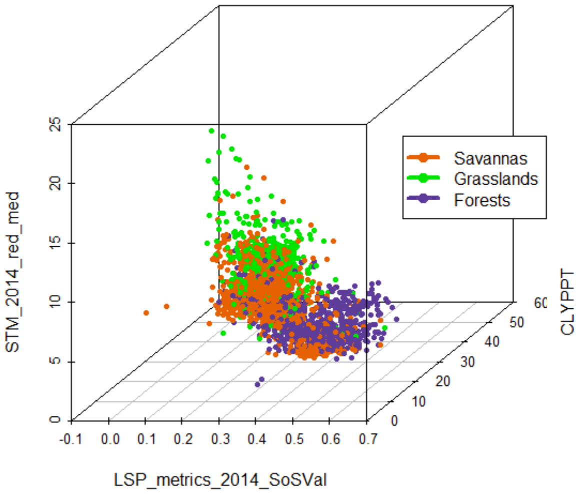

Figure 4. 3D-scatterplot of three of the most important variables for the level-1 classification model based on the Gini importance index.

We can observe higher SR values on the red band during the dry season for grasslands, showing an increasing gradient from forests to grasslands. This was also observed by other authors that worked with spectral characterization of the Cerrado vegetation (Ferreira, Huete, 2004, Jacon et al., 2017).

We can also observe in Figure 4 that higher values of clay content are present in the grassland class. We checked the samples with the 1:250,000 scale Brazilian official soil Map (IBGE, 2017). The majority of the grasslands samples are within latosols, which are soils with a large clay content. On the other hand, the other classes are most distributed on entisols and plinthosols, which could have favored the low clay concentration in these classes. More studies are needed to understand these patterns, and also to analyze if there is a bias on the samples distribution.

The five most important variables for the level-2 classification model were the Tandem-X elevation (TANDEM_X_90), the EVI value during the peak of the season (LSP_Metrics_2014_Mfit), the Cation Exchange Capacity of soil (CECSOL), the EVI value during the start of the season (LSP_Metrics_2014_SoSVal) and the minimum EVI value during the season (LSP_Metrics_2014_BV). Figure 5 shows 3D-scatterplot of three most important variables for the level-2 classification model based on the Gini importance index.

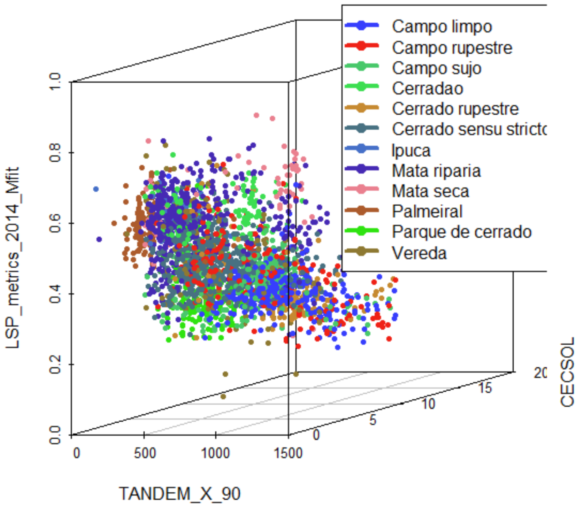

Figure 5. 3D-scatterplot of the three most important variables for the level-2 classification model based on the Gini importance index.

Grasslands classes as Campo Limpo, Campo Sujo and Campo Rupestre are more rare than the others, so they mostly remain in environmental protection areas, mainly plateaus, which have higher elevations. Campo rupestre class usually occurs in regions with elevation higher than 900 meters (Ribeiro et al., 2008). Also, as we can observe, these grasslands are commonly above latosols, which are soils with higher content of clay and cation capacity exchange.

The Palmeiral class is characterized by the massive presence of palm trees, that usually occurs in lowlands, in the plateau of Maranhão-Piauí states (Floresta de Cocais and Costeiro ecoregions). They are associated to ground surface water. The same for Parque de Cerrado class, that is mostly concentrated on the Araguaia river floodplain (Bananal ecoregion) (Ribeiro et al., 2008). Furthermore, these ecoregions have the lowest elevations in the Cerrado (Sano et al., 2019).

Additionally, we validated the level-1 model using completely independent ground truth data in the Brasilia National Park (Ferreira, 2003) with an overall accuracy of 0.71, and f1-class scores of 0.64, 0.76 and 0.73, for savanna, grasslands and forest. This shows the robustness of a global model that enables to classify the remaining natural vegetation of the entire Cerrado with a high level of thematic detail.

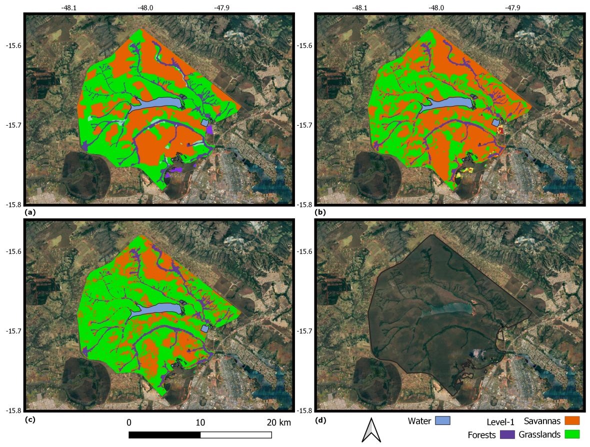

Figure 6. Map comparisons: a) Ferreira et al., 2003; b) Mapbiomas Collection 4.0; c) Our map; d) Google Satellite (April 7, 2019).

3.2 Discussion

The use of ARD data derived from Landsat time series and a selection of environmental data, together with a well distributed set of reference data, enabled us to differentiate the main vegetation physiognomies of the Cerrado on two hierarchical levels.

Even though our classifications could be assessed with high overall accuracies we observed confusion between savannas and grasslands. Alencar et al. (2020) also showed some areas where the native vegetation appeared as a mosaic of grasslands, savannas, and forests, indicating a limitation of Landsat data to discriminate these three vegetation types properly, which might be due to the moderate spatial resolution (30 m).

The producers accuracies of our classes for level-1 are higher than reported by Alencar et al. (2020) (0.61, 0.54, 0.75 for savanna, grassland and forest, respectively).

Our results suggest that combining various important environmental data with ARD is a promising approach for differentiating these vegetation physiognomies. This finding is in line with other studies that showed the correlation between the different vegetation gradients of the Cerrado to the EVI time series can be used as a proxy of the phenological variation (Schwieder et al., 2018), suggested the importance of environmental variables for the delimitation of different types of vegetation (Sano et al., 2019) and highlighted the relation between terrain proxies and the phenological information (Bendini et al., 2019b).

Visually we can see that the maps are consistent when comparing to other maps. In case of Ferreira et al. (2003) , the visual comparison needs to be done carefully, as the map is based on spatial segments. Additionally, it needs to be considered that fire events, which are frequent in the Cerrado, might have changed the vegetation structure. These can be reasons for a underestimating of the savannas in both our approach and Mapbiomas approach.

By assessing the confusion matrix from Figure 3, confusion between the grassland physiognomies and Cerrado Sensu stricto becomes apparent. This can be associated to the fact that we merged the different sub classes of Cerrado Sensu stricto (Cerrado Ralo, Típico and Denso) into one class. And even in the field, it can be hard to differentiate Cerrado Ralo and Campo Sujo. Ribeiro and Walter (2008) claim that several Cerrado coverages do not have evident transition areas. So, despite the high accuracies for the level-1 model, we assume that our approach is still underestimating the savannas in that region.

For future investigations we aim to analyze the samples and try to improve the number of ground truth points by conducting additional fieldwork. The use of data from different sources can lead into semantic problems on the class labeling. We also point to the opportunity of using Sentinel-1 SAR data to improve the separation between the sub classes of Cerrado Sensu Stricto, once they are mostly associated to biomass density. Berger et al. (2019) explored the potential of multi-temporal Sentinel-1 data for herbaceous biomass mapping in savanna ecosystem over Kruger National Park (KNP), South Africa. They achieved good results indicating that Sentinel-1 time series can be successfully employed for this type of mapping application.

In case of Veredas, which achieved the lowest accuracies, we believe that in particular for this class, the spatial context is determinant. As we can see from Table 1, the Vereda physiognomy is characterized by the predominance of Buritis, with more or less dense clusters of shrub-herbaceous species and surrounded by Campo Limpo, being generally flooded. We can see this by analyzing the confusion matrix from Figure 3. The most part of the misclassifications for Vereda class was observed with Campo Limpo, Mata Riparia and Cerrado Sensu Stricto. Perhaps this specific pattern can not be detected from a Landsat-30m pixel. We point out to opportunities of using Deep Learning algorithms, specially the Convolutional Neural Networks, that take into account the spatial context, to detect this physiognomy.

And as far as we know, it is the first time to report a (semi-) automatic methodology that can achieve such good accuracies for level-2 classification, considering samples from the different regions of the whole Cerrado. The results are promising and we aim to improve the classification models presented here, and run the methodology for all over the biome.

Numerous projects have been developed to produce Earth Observation (EO) Data Cubes such as the Australian Geoscience Data Cube (AGDC) (Dhu et al., 2017, Lewis et al., 2017) and the Framework for Operational Radiometric Correction for Environmental monitoring (FORCE) (Frantz, 2019), that enable to pre-process remote sensing data and organize and store them ARD for immediate analysis. The Brazil Data Cube project[4], currently developed by INPE provides multidimensional ARD data cubes from medium-resolution EO images, including Landsat, CBERS and Sentinel satellites images.

We believe that once these maps can be generated on a pixel-level and using ARD combined with other environmental data that can be integrated on a EO Datacube structure, using cloud processing platforms, i.e. Google Earth Engine, Amazon Web Services or Azure, there is a possibility of deriving important proxies related to ecosystem structure and to biodiversity consequently, being potentially considered for developing a system for monitoring biodiversity.

[4] More information can be seen at http://brazildatacube.org/.

https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-953-2020 |

ACKNOWLEDGEMENTS

The project "Development of systems to prevent forest fires and monitor vegetation cover in the Brazilian Cerrado" (World Bank Project #P143185) – Forest Investment Program (FIP).

REFERENCES

Abreu, E. S., Rosim, S., Renno, C. D., de Freitas Oliveira, J. R., Jardim, A. C., de Oliveira Ortiz, J., Dutra, L. V., 2012. Terrahidro &#x2014 a distributed hydrological system to delimit large basins. *2012 IEEE International Geoscience and Remote Sensing Symposium*, IEEE.

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. et al., 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.

Bendini et al., 2019a. Detailed agricultural land classification in the Brazilian cerrado based on phenological information from dense satellite image time series. *IJAEOG*, 82, 101872.

Bendini, H. N. et al., 2019b. Assessing Satellite-Derived Phenological Metrics and Terrain Data As a Proxy for Vegetation Dynamics Along the Brazilian Savanna Corridor. *PECORA - International Symposium on Remote Sensing of Environment*, 38. http://pecora.asprs.org/wp-content/uploads/2019/05/Preliminary-Program-52919.pdf.

Berger, C., Werner, S., Wigley-Coetsee, C., Smit, I., Schmullius, C., 2019. Multi-temporal sentinel-1 data for wall-to-wall herbaceous biomass mapping in kruger national park, south africa — first results. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), 5–32.

Câmara, G. et al., n.d. Land cover change maps for mato grosso state in brazil: 2001-2017 (version 3). PANGAEA - Data Publisher for Earth Environmental Science.

de Morisson Valeriano, M., de Fátima Rossetti, D., 2012. Topodata: Brazilian full coverage refinement of SRTM data. *Applied Geography*, 32(2), 300–309. https://doi.org/10.1016/j.apgeog.2011.05.004.

Dhu et al., 2017. Digital Earth Australia – Unlocking New Value from Earth Observation Data. *Big Earth Data*, 1(1-2), 64–74.

Ferreira et al., 2007. Spectral linear mixture modelling approaches for land cover mapping of tropical savanna areas in Brazil. *International Journal of Remote Sensing*, 28(2), 413–429.

Ferreira, L. G., Huete, A. R., 2004. Assessing the seasonal dynamics of the Brazilian Cerrado vegetation through the use of spectral vegetation indices. *International Journal of Remote Sensing*, 25(10), 1837–1860. https://doi.org/10.1080/0143116031000101530.

Ferreira, M. E., 2003. Análise do modelo linear de mistura espectral na discriminação de fitofisionomias do parque nacional de brasília (bioma cerrado).

Frantz, D., 2019. FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. *Remote Sensing*, 11(9), 1124.

Girolamo-Neto et al., 2017. Assessment of texture features for Brazilian savanna classification: A case study in Brasília national park. *Braz. J. Cartogr*, 69, 891–901.

Grecchi, R. C., Gwyn, Q. H. J., Bénié, G. B., Formaggio, A. R., 2013. Assessing the spatio-temporal rates and patterns of land-use and land-cover changes in the Cerrados of southeastern Mato Grosso, Brazil. *International journal of remote sensing*, 34(15), 5369–5392.

Griffiths, P., van der Linden, S., Kuemmerle, T., Hostert, P., 2013. A pixel-based Landsat compositing algorithm for large area land cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5), 2088–2101.

Gruber, A., Wessel, B., Huber, M., Roth, A., 2012. Operational TanDEM-X DEM calibration and first validation results. *ISPRS Journal of Photogrammetry and Remote Sensing*, 73, 39–49. https://doi.org/10.1016/j.isprsjprs.2012.06.002.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2), e0169748. https://doi.org/10.1371/journal.pone.0169748.

IBGE, 2012. Manual técnico da vegetação brasileira.

IBGE, 2017. Pedologia.

INPE, 2019a. PRODES Annual increment of deforested areas in the Brazilian Cerrado. National Institute for Space Research.

INPE, 2019b. Projeto terraclass cerrado-mapeamento do uso e cobertura vegetal do cerrado. http://www.dpi.INPE.br/tccerrado/.

Jacon, A. D., Galvão, L. S., dos Santos, J. R., Sano, E. E., 2017. Seasonal characterization and discrimination of savannah physiognomies in Brazil using hyperspectral metrics from Hyperion/EO-1. *International Journal of Remote Sensing*, 38(15), 4494–4516. https://doi.org/10.1080/01431161.2017.1320443.

Jönsson, P., Eklundh, L., 2004. TIMESAT—a program for analyzing time-series of satellite sensor data. *Computers & geosciences*, 30(8), 833–845.

Lewis et al., 2017. The Australian Geoscience Data Cube — Foundations and Lessons Learned. *RSE*, 202, 276–292.

Liaw, A., Wiener, M. et al., 2002. Classification and regression by randomForest. *R news*, 2(3), 18–22.

MMA, M. o. E., 2015. Biomas.

Neves et al., 2019. Hierarchical classification of brazilian savanna physiognomies using very high spatial resolution image, superpixel and geobia. *IGARSS 2019*, IEEE, 3716–3719.

Noojipady, P., Morton, C. D., Macedo, N. M., Victoria, C. D., Huang, C., Gibbs, K. H., Bolfe, L. E., 2017. Forest carbon emissions from cropland expansion in the Brazilian Cerrado biome. *Environmental Research Letters*, 12(2), 025004. https://doi.org/10.1088/1748-9326/aa5986.

Potapov et al., 2020. Landsat Analysis Ready Data for Global Land Cover and Land Cover Change Mapping. *Remote Sensing*, 12(3), 426.

Ratter, J. A., Ribeiro, J. F., Bridgewater, S., 1997. The Brazilian cerrado vegetation and threats to its biodiversity. *Annals of botany*, 80(3), 223–230.

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., Waterloo, M. J., 2008. HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. *Remote Sensing of Environment*, 112(9), 3469–3481. https://doi.org/10.1016/j.rse.2008.03.018.

Ribeiro et al., 2008. As principais fitofisionomias do bioma Cerrado. *Cerrado: ecologia e flora*, 1, 151–212.

Rocha, G. F., Ferreira, L. G., Ferreira, N. C., Ferreira, M. E., 2011. Detecção de desmatamentos no bioma Cerrado entre 2002 e 2009: padrões, tendências e impactos. *Revista Brasileira de Cartografia*, 63(3).

Rufin, P., Müller, H., Pflugmacher, D., Hostert, P., 2015. Land use intensity trajectories on Amazonian pastures derived from Landsat time series. *International Journal of Applied Earth Observation and Geoinformation*, 41, 1–10. https://doi.org/10.1016/j.jag.2015.04.010.

Sano, E. E., Rodrigues, A. A., Martins, E. S., Bettiol, G. M., Bustamante, M. M., Bezerra, A. S., Couto, A. F., Vasconcelos, V., Schüler, J., Bolfe, E. L., 2019. Cerrado ecoregions: A spatial framework to assess and prioritize Brazilian savanna environmental diversity for conservation. *Journal of Environmental Management*, 232, 818–828. https://doi.org/10.1016/j.jenvman.2018.11.108.

Sano, E. E., Rosa, R., Brito, J. L., Ferreira, L. G., 2010. Land cover mapping of the tropical savanna region in Brazil. *Environmental monitoring and assessment*, 166(1-4), 113–124.

Schwieder et al., 2016. Mapping Brazilian savanna vegetation gradients with Landsat time series. *IJAEOG*, 52, 361–370.

Schwieder, M., Leitão, P., Pinto, J., Teixeira, A., Pedroni, F., Sanchez, M., Bustamante, M., Hostert, P., 2018. Landsat phenological metrics and their relation to aboveground carbon in the Brazilian Savanna. *Carbon balance and management*, 13(1), 7.

Strassburg, B. B. N., Brooks, T., Feltran-Barbieri, R., Iribarrem, A., Crouzeilles, R., Loyola, R., Latawiec, A. E., Filho, F. J. B. O., de M. Scaramuzza, C. A., Scarano, F. R., Soares-Filho, B., Balmford, A., 2017. Moment of truth for the Cerrado hotspot. *Nature Ecology & Evolution*, 1(4). https://doi.org/10.1038/s41559-017-0099.

Team, R. C. D., 2019. R: A language and environment for statistical computing (version 3.5. 2, r foundation for statistical computing, vienna, austria, 2018).

Tuchschneider, D., 2013. Brazil - development of systems to prevent forest fires and monitor vegetation project (english). Technical report, World Bank Group, Washington, D.C.