Hierarchical mapping of Brazilian Savanna (Cerrado) physiognomies based on deep learning

Alana K. Neves[a,*] **Thales S. Körting**[a] **Leila M. G. Fonseca**[a] **Anderson R. Soares**[a] **Cesare D. Girolamo-Neto**[b] **and Christian Heipke**[c]

[a]INPE - National Institute for Space Research, Division of Earth Observation and Geoinformatics (DIOTG), São José dos Campos – SP, Brazil
[b]Vale Institute of Technology, Belém, Brazil
[c]Leibniz Universität Hannover, Institute of Photogrammetry and GeoInformation, Hannover, Germany

Abstract. The Brazilian Savanna, also known as Cerrado, is considered a global hotspot for biodiversity conservation. The detailed mapping of vegetation types, called physiognomies, is still a challenge due to their high spectral similarity and spatial variability. There are three major ecosystem groups (forest, savanna, and grassland), which can be hierarchically subdivided into 25 detailed physiognomies, according to a well-known classification system. We used an adapted U-net architecture to process a WorldView-2 image with 2-m spatial resolution to hierarchically classify the physiognomies of a Cerrado protected area based on deep learning techniques. Several spectral channels were tested as input datasets to classify the three major ecosystem groups (first level of classification). The dataset composed of RGB bands plus 2-band enhanced vegetation index (EVI2) achieved the best performance and was used to perform the hierarchical classification. In the first level of classification, the overall accuracy was 92.8%. On the other hand, for the savanna and grassland detailed physiognomies (second level of classification), 86.1% and 85.0% were reached, respectively. As the first work that intended to classify Cerrado physiognomies in this level of detail using deep learning, our accuracy rates outperformed others that applied traditional machine learning algorithms for this task. [DOI: 10.1117/1.JRS.15.044504]

Keywords: Savanna; Cerrado; physiognomy; semantic segmentation; spectral channels; protected area.

Paper 210442 received Jul. 9, 2021; accepted for publication Oct. 1, 2021; published online Oct. 15, 2021.

1 Introduction

The ecosystems of savanna cover ~20% of the Earth's terrestrial area. Especially in tropical regions, they are rich in biodiversity[1] and water resources[2] and play an important role in carbon stock due to their content of above and below-ground biomass.[3] The Brazilian Savanna, also known as Cerrado, is considered a global hotspot for biodiversity conservation, containing around 4800 endemic species.[1,4] In addition, its water resources feed the three largest watersheds in South America: the Amazon, Prata, and São Francisco. The Cerrado comprises 24% of the Brazilian territory, being the second largest biome in the country. However, despite its ecological relevance, only 8.6% of its native vegetation is in protected areas, i.e., specific regions established to protect biodiversity, water bodies, and other environmental resources.[5] Approximately 47% of the Cerrado native vegetation have already been converted to other land uses, especially pasture (29%) and annual agriculture (9%).[6] Moreover, from 2008 to 2019, the deforestation rates in the Cerrado have been higher than in the Amazon biome in 9 of these 12 years interval.[7] This heavy loss of native vegetation has severe environmental consequences, such as vegetation fragmentation, habitat loss,[8] and reduction of water yield and carbon stocks.[9,10]

*Address all correspondence to Alana K. Neves, alana.neves@inpe.br

In this scenario, accurate mapping of Cerrado vegetation is essential to support policies against deforestation and, consequently, to maintain the provision of ecosystem services, since these maps are crucial to assess biodiversity, improve carbon stock estimation within the biome, and guide conservation policies.

The use of remote sensing imagery to perform Cerrado vegetation mapping is still a challenge due to the high spatial variability and spectral similarity among its vegetation types, also known as physiognomies. According to the well-known classification system proposed by Ribeiro and Walter,[11] there are 25 physiognomies, which vary in vegetation structure, tree density, edaphic conditions, and floristic composition. These physiognomies can be grouped into three major ecosystem groups: grasslands, savannas, and forests. Therefore, this classification system has an intrinsic hierarchical structure, where the previous identification of the three major ecosystem groups in a first level of classification may improve the identification of more detailed physiognomies in subsequent levels. Nevertheless, only few studies have been exploring the hierarchical aspect of the classification system to map Cerrado vegetation.[12,13] Neves et al.[12] used geographic object-based image analysis (GEOBIA) techniques and the random forest (RF) algorithm to compare hierarchical and nonhierarchical approaches to classify seven Cerrado physiognomies. Although the authors have observed the superiority of the hierarchical approach over the nonhierarchical one, the accuracy for some physiognomies was still low. More recently, Ribeiro et al.[13] used GEOBIA and RF and also included a spatial contextual ruleset to represent other environmental factors (e.g., soil type, slope, and elevation) to improve the Cerrado vegetation classification. They classified 13 classes, including 11 vegetation types, with an overall accuracy (OA) of 87.6%. However, their semiautomatic methodology still relied on some subjective tasks, such as the selection of image segmentation parameters.

Considering nonhierarchical approaches, some works have been mapping the Cerrado natural vegetation using remote sensing techniques.[6,14–18] The TerraClass Cerrado project[6] utilized Landsat images and region growing image segmentation followed by visual interpretation to map land use and land cover in the entire Cerrado biome in 2013. The forest and nonforest (grasslands and savannas) classes from TerraClass map obtained accuracies between 60% and 65%. The MapBiomas Project[14,19] classified the entire Cerrado from 1985 to 2017 using Landsat images (30 m) and RF algorithm to identify forest, savanna, and grassland. Encompassing smaller study sites, other works identified more detailed vegetation classes.[15–18] Jacon et al.[16] used a hyperspectral image (30-m spatial resolution) to perform automatic classification of seven physiognomies based on time series and multiple discriminant analysis. Ferreira et al.[17] and Schwieder et al.[18] used Landsat images to classify five and seven vegetation classes, respectively. Ferreira et al.[17] employed spectral linear mixture models and Mahalanobis distance, whereas Schwieder et al.[18] utilized the support vector machine (SVM) method and phenological metrics extracted from dense time series.

Regardless of the classification methods, studies have shown that detailed Cerrado vegetation mapping based on Landsat-like imagery is a challenging task since the Cerrado biome is composed by heterogeneous and seasonal natural vegetation types. In addition, some works[12,15,20] performed a semiautomatic classification using high spatial resolution image, such as WorldView-2 (2-m spatial resolution)[12,15] and RapidEye imagery (5-m spatial resolution).[13,20] Despite the use of a more detailed spatial resolution, the misclassification in transition areas remained an issue. Each physiognomy has a unique biodiversity and is responsible for a specific amount of carbon stocked above and below the ground.[10,11] For this reason, improving the detailed physiognomies mapping is crucial, since so far the mapping initiatives handled well the identification of the three major groups of ecosystems but reached low accuracies for individual physiognomies.[12,15–17]

To improve the physiognomies discrimination to generate a detailed vegetation map, a large variety of machine learning techniques have been employed. Convolutional neural networks (CNNs) are able to perform end-to-end classification, learning features from an input dataset, and presenting increasing complexity through the layers of the network.[21] The results achieved with such methods often outperform those obtained with traditional machine learning algorithms, such as RF or SVM.[22] For savanna vegetation, some efforts have already been made with deep learning to delineate tree crowns.[23,24] Nogueira et al.[25] were the first to employ a deep learning-based method to identify vegetation patterns, which include different tree heights,

tree cover and shrub, and herbaceous vegetation. Using RapidEye imagery, entire regular image patches were designated as only one class (forest, savanna, or grassland), resulting in a considerable mixture of classes in a single patch. A semantic segmentation (also known as pixelwise classification) of the three major ecosystem groups was performed by Neves et al.,[26] using a modified U-net architecture and eight spectral bands of the WorldView-2 satellite image. Compared with the classification approach performed by Nogueira et al.,[25] the semantic segmentation results in a better class delineation.

The U-net[27,28] belongs to the group of fully convolutional neural networks (FCNNs[29]). Compared with more traditional CNNs such as LeNet[30] and AlexNet[31] that predict a single class for each image patch, FCNNs are tailored to the task of semantic segmentation. In particular, they take an image patch with an arbitrary number of channels as input and predict a label-map usually of the same size as the input. The U-net is composed of a multilayer convolutional encoder that successively reduces the spatial resolution and increases the number of filters per kernel and a multilayer convolutional decoder, which upscales the features to the original spatial resolution. They further use skip-connections between encoder and decoder layers, of the same spatial resolution, to preserve low-level details, required for the precise prediction of object boundaries.

In deep learning methods, originally developed in the computer vision field, the analysis of the contribution of different spectral bands to improve the network accuracy is not yet well explored. The spectral behavior of the physiognomies and their respective major groups rely on the information contained in different wavelengths, represented here by satellite spectral bands. However, the majority of works with deep learning approaches used only red, green, and blue channels[22,32] or included the near-infrared (NIR) one.[25,33] In addition, very few initiatives have applied some hierarchical behavior in classification tasks.[34,35]

Therefore, the objective of this work is threefold: (1) to hierarchically classify the Brazilian Savanna physiognomies based on deep learning techniques according to the classification system proposed by Ribeiro and Walter;[11] (2) to evaluate different combinations of spectral bands taken as input dataset in the deep learning classification; and (3) to evaluate the deep learning classification performance in relation to the different training samples selection methods. To the best of our knowledge, this is the first work that produced a Brazilian Savanna map in this level of detail based on deep learning techniques.

2 Materials and Methods

This study was performed according to the flowchart presented in Fig. 1. In phase A of the methodology [Fig. 1(a)], we investigated several combinations of spectral bands as input in the processing to perform a semantic segmentation of three major Cerrado ecosystem groups (grassland, savanna, and forest). Figure 1(b) shows the steps required to perform the semantic segmentation approach using an adapted U-net CNN architecture.[28] Following, we applied the best input dataset resulting from the first part to perform the hierarchical Cerrado vegetation mapping [Fig. 1(c)]. All processing procedures are described in detail in the following sections.

2.1 *Study Site*

As study site, a Brazilian protected area was chosen to ensure the native Cerrado vegetation analysis. The Brasília National Park (BNP) (Fig. 2), located in the Federal District, Brazil, comprises ~423 km^2 of native Cerrado vegetation, which encompasses the major physiognomies found in the Cerrado biome.[17,36] It contains several endangered species[37] and a dam that is responsible for 25% of the Federal District's water supply. This protected area was also used as study site in several other works,[12,15,17,18,36] which attests its representativeness and facilitates a comparison among the results.

According to the existing physiognomies in the study site and the classification system proposed by Ribeiro and Walter,[11] we differentiated two hierarchical levels of classes. In the first level, three major ecosystem groups (also known as formations) were classified: forest, savanna, and grassland. In the forest formations, there is a predominance of arboreal species, forming continuous or discontinuous canopy. In the second level, forest was maintained as gallery forest (*Mata de Galeria*), since it is the only forest physiognomy with significant presence in this area.

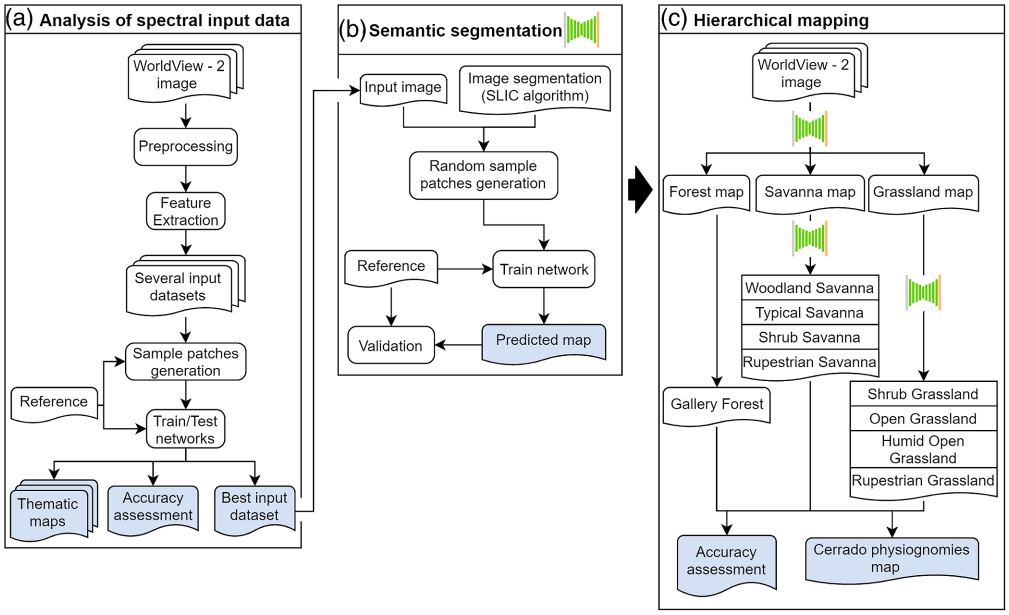

Fig. 1 Methodological flowchart presenting: (a) the analysis of spectral input data, which generates the most accurate input dataset that will be used in the next part. (b) The semantic segmentation approach. (c) The hierarchical mapping methodology, where the semantic segmentation icon is used to indicate every time the semantic segmentation is performed.

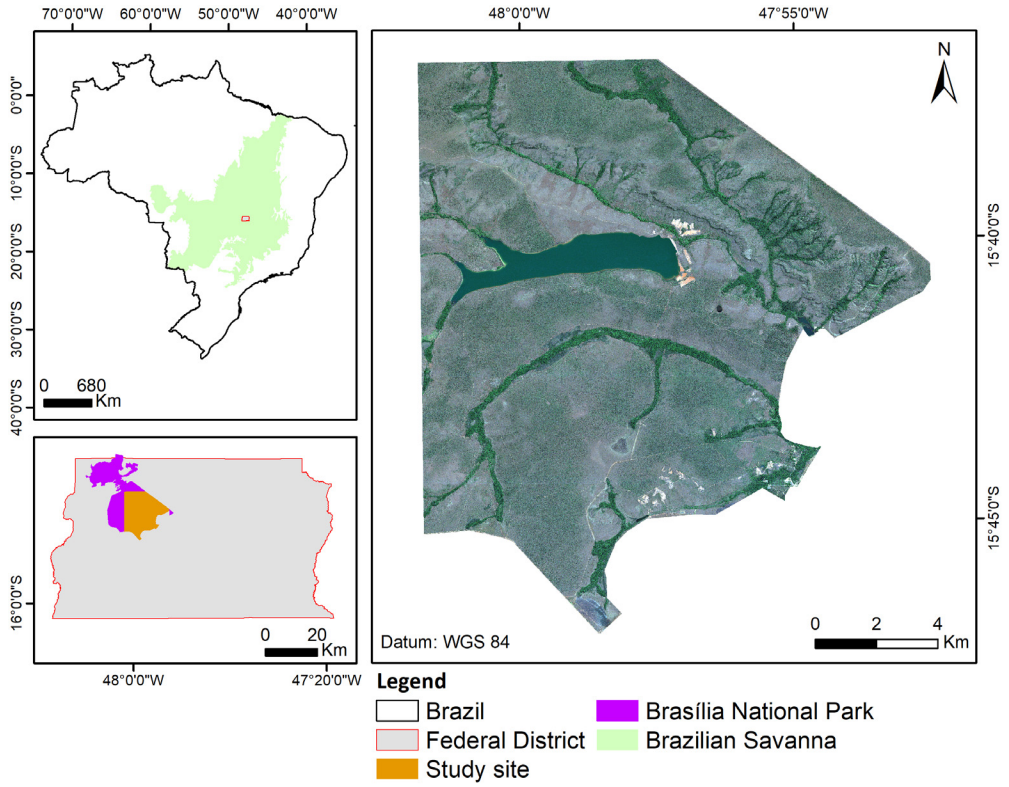

Fig. 2 Location of the BNP image (true color composite) in the Brazilian savanna.

In savanna formations, the presence of continuous canopy is uncommon, and there are trees and shrubs scattered over grasses. The areas identified as savannas in the first level were subdivided into woodland savanna (*Cerrado Denso*), typical savanna (*Cerrado Típico*), Rupestrian savanna (*Cerrado Rupestre*), shrub savanna (*Cerrado Ralo*), and *Vereda* in the second level.

Table 1 Detailed description of the physiognomies, adapted from Ribeiro and Walter classification system.[11] Formation and physiognomy columns represent the first and second level of classes in this study, respectively.

Formation	Physiognomy	Vegetation	Tree cover	Tree height	Other aspects
Forest	Gallery forest (*Mata de galeria*)	Riparian forest, closed canopy, composed predominantly by evergreen (deciduous) trees	From 70% to 95%	20 to 30 m	Follows small size rivers and streams, vegetation forms galleries over water bodies
Savanna	Woodland savanna (*Cerrado Denso*)	Mostly arboreal, presence of shrub and herbaceous layer (less dense due to the tree shades)	From 50% to 70%	5 to 8 m	Presence of tortuous trees with twisted branches. Usually presents evidence of forest fires
	Typical savanna (*Cerrado Típico*)	Mostly arboreal-shrubby, without continuous canopy. Intermediate savanna type between woodland and shrub savanna	From 20% to 50%	3 to 6 m	Presence of tortuous trees with twisted branches. Usually presents evidence of forest fires
	Shrub savanna (*Cerrado Ralo*)	Mostly arboreal-shrubby. Presents a significant shrub-herbaceous layer	From 5% to 20%	2 to 3 m	Represents the lowest and least dense of the *Sensu Stricto* savannas
	Rupestrian savanna (*Cerrado Rupestre*)	Arboreal-shrubby, presents shrub-herbaceous layer and occurs in Rupestrian environments	From 5% to 20%	2 to 4 m	Presence of rocky outcrops. Usually occurs in mosaics, which includes other types of vegetation
	Vereda	Presents only one species of palmtree, the Buriti (*Mauritua flexuosa*), in mainly flooded terrain	From 5% to 10%	12 to 15 m	Does not form a closed canopy and is usually surrounded by shrub-herbaceous vegetation
Grassland	Shrub grassland (*Campo Sujo*)	Shrub-herbaceous. Evident presence of shrubs, subshrubs, and isolated low trees	No canopy formation	—	Presents subdivisions according to topographic and edaphic factors
	Open grassland (*Campo Limpo*)	Mostly herbaceous, the presence of shrubs and subshrubs are insignificant and there are no trees	No canopy formation	—	Presents subdivisions according to topographic and edaphic factors
	Humid open grassland (*Campo Limpo úmido*)	Predominantly herbaceous and seasonally flooded (marsh)	No canopy formation	—	Consists of a subdivision of open grassland. Occurs in terrains with high water table and usually surrounds *Veredas* and riparian forests
	Rupestrian grassland (*Campo Rupestre*)	Shrub-herbaceous, possible presence of low trees	No canopy formation	up to 2 m	Usually occurs at altitudes above 900 m and in areas with rocky outcrops

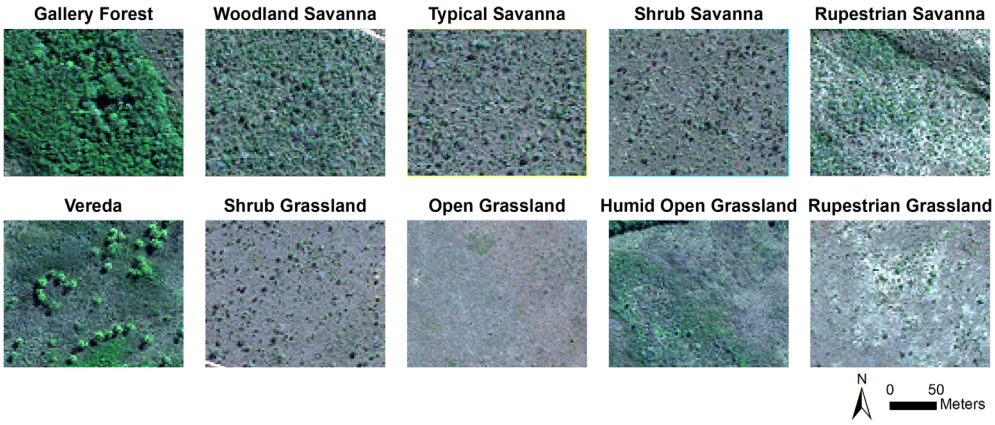

Fig. 3 Patterns of the physiognomies in the WorldView-2 image (true color composite).

In grasslands, there are predominantly herbaceous species and some shrubs. Four subclasses were differentiated in the second level: shrub grassland (*Campo Sujo*), open grassland (*Campo Limpo*), Rupestrian grassland (*Campo Rupestre*), and humid open grassland (*Campo Limpo Úmido*). The humid open grassland is a subtype of open grassland, but it was considered an independent class due to its significant presence in the study site. The Cerrado physiognomies hierarchy and their individual definitions and characteristics are presented in Table 1. Besides, their patterns in a WorldView-2 image true color composite can be observed in Fig. 3.

2.2 *Remote Sensing Data, Preprocessing, and Feature Extraction*

The WorldView-2 image (tile ID 103001003373A600), with spatial resolution of 2 m and acquired on July 22, 2014, was utilized in this work. Although the Cerrado vegetation is strongly influenced by the seasonality, the chosen image is from the dry season. According to Jacon et al.,[16] the spectral separability of the physiognomies increases during the dry season. This image has eight spectral bands: coastal (400 to 450 nm), blue (450 to 510 nm), green (510 to 580 nm), yellow (585 to 625 nm), red (630 to 690 nm), red-edge (705 to 745 nm), NIR 1 770 to 895 nm), and NIR2 (860 to 1040 nm). Initially represented in digital numbers, the image was converted to surface reflectance using the fast line-of-sight atmospheric analysis of hypercubes (FLAASH) algorithm[38] available in the ENVI 5.2 software.

Besides the spectral bands, other features were included as input data in the experiments: two-band enhanced vegetation index[39] and three components of LSMM[40] (vegetation, soil, and shade components). The EVI2 is given as

$$\mathrm{EVI2} = 2.5\frac{\rho\mathrm{NIR} - \rho R}{\rho\mathrm{NIR} + 2.4\rho R + 1}, \tag{1}$$

where ρNIR is the reflectance in the NIR band and ρR is the reflectance in the red band.

To create the LSMM, 10 endmembers (pure pixels) were selected for each component, which is calculated as

$$r_i = \sum_{j=1}^{N}(a_{ij} \times x_j) + e_i, \tag{2}$$

where r_i is the spectral reflectance mean for the i'th band of a pixel with N components; j is the number of components; i is the number of bands; a_{ij} is the spectral reflectance of the j'th component of a pixel for the i'th band; x_j is the proportional value of the j'th component of the pixel; and e_i is the error for the i'th band.

As we did not have a detailed Cerrado vegetation map and FCNNs require classified patches (rather than points) for training, the entire WorldView-2 image was classified through visual interpretation and, then, used as reference data. This was performed by some specialists in remote

sensing, who have previous experience in mapping Brazilian Savanna vegetation. Considering our intention to differentiate natural vegetation types (Fig. 3), the other minority areas were masked in the reference data (built-up areas, water bodies, bare soil, and burned areas).

2.3 *Network Architecture*

In this work, a variation of the U-net architecture,[27] proposed by Kumar,[28] was used in all tasks of pixelwise classification (semantic segmentation). The architecture[28] mainly follows the design-choices of Ronneberger et al.[27] However, the architecture was modified to use zero-padding, instead of unpadded convolution, to preserve the spatial size along the network. As a further modification, the upsampling is based on transposed convolutions with a stride of two along both spatial dimensions.

Network parameters such as the number of layers and number of filters per layer are shown in Fig. 4. The input layer (gray color) represents the size of the sample patches (160×160). The depth (N) is the number of bands. The output layer has the same size of the input layer, but the depth is represented by the number of classes C. Every other layer is represented according to the legend; a 2×2 max-pooling layer, for instance, is illustrated in pink. The numbers in brackets are the image sizes in each layer followed by the number of filters. As in the original U-net, skip-connections are used to concatenate information of high spatial resolution (but low complexity) with information of low spatial resolution (but high complexity).

While the last network layer for semantic segmentation is usually modeled by softmax function, here we chose the sigmoid function because it presented higher OAs in the preliminary tests. This allows the model to predict independent probabilities per class and per pixel. Final class predictions are obtained by choosing the respective classes with highest probabilities. The network was implemented in a Python environment, using Keras[41] with TensorFlow[42] as backend. The NVIDIA GeForce RTX 2070 super (8 GB) GPU was used.

2.4 *Analysis of Spectral Input Data*

To test the network performance according to the spectral data used as input, eight datasets were created. The first one is composed by the red, green, and blue bands, and it was considered the baseline dataset, since it is the simplest one and presents the spectral bands commonly used in deep learning approaches. Then, we added one type of information (e.g., NIR spectral bands or EVI2) at each of the following datasets to evaluate how it could improve the model performance. An additional dataset using only the LSMM components was also used. Since the LSMM vegetation component is highly correlated to the EVI2, we did not include a dataset using them together. The datasets are summarized below:

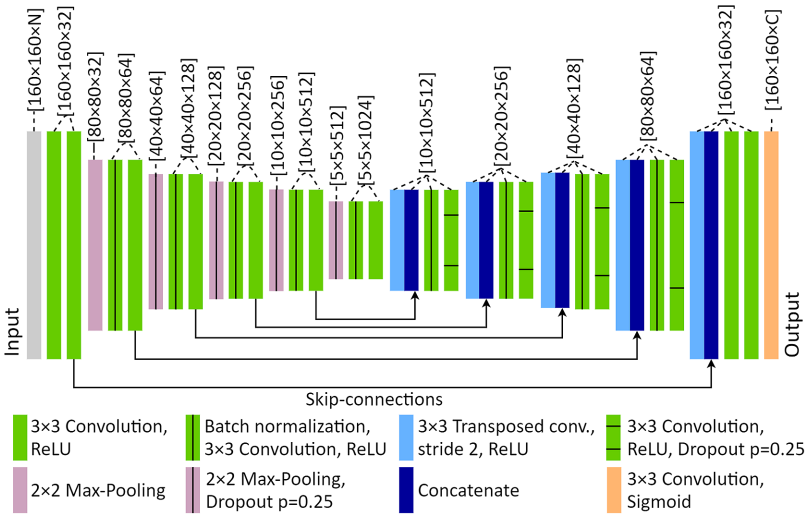

Fig. 4 Modified U-net architecture (adapted from Ref. 28). The N, in the input size, is the number of bands, while the C, in the output size, corresponds to the number of classes.

- RGB (red, green, and blue bands)
- RGB + EVI2
- RGB + LSMM
- RGB + rededge
- RGB + NIR1 + NIR2
- RGB + NIR1 + NIR2 + rededge
- 8 WorldView-2 bands
- LSMM (three LSMM components: vegetation, soil, and shade).

All datasets were divided into regions A, B, and C (Fig. 5), which contain roughly similar distributions of the three classes of the first level of classification. Thereafter, the datasets and the reference data were cropped into nonoverlapping and adjacent tiles of 160×160 pixels to be used as samples. Tiles with any no-data value were excluded from further processing (i.e., pixels originally covering built-up areas, water bodies, bare soil, and burned areas). To increase the amount of samples, six data augmentation techniques were employed: transposition, horizontal, and vertical flips and three rotations: 90 deg, 180 deg, and 270 deg (clockwise direction).

The samples selected from regions A, B, and C were combined as follows: 70% of the samples from two regions (e.g., A and B) were randomly selected for training, and 30% for validation. The resulting network was then tested in the remaining region (e.g., C). This experiment was repeated three times, i.e., until the three regions had a semantic segmentation resulting from the cross-validation approach. Table 2 shows the number of samples used in each experiment. For the training and validation sets, those numbers include samples generated through a data augmentation procedure.

During training, the early stopping criterion (also known as "patience" in Keras) was set to 50, i.e., if after 50 epochs the validation accuracy did not increase, the training process was halted

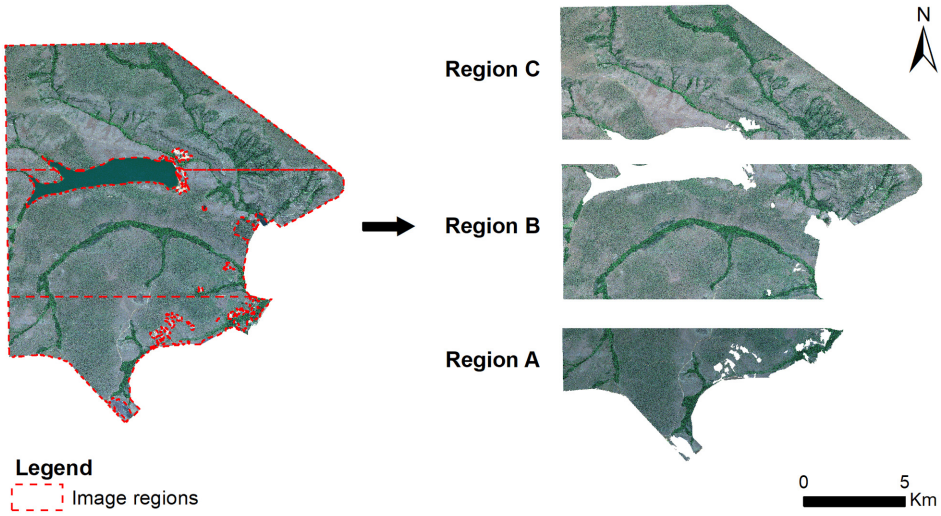

Fig. 5 Regions A, B, and C used to generate sample patches in the spectral input data analysis, according to Table 2.

Table 2 Regions and number of samples used for training, validation, and testing procedures in each cross-validation experiment.

Experiment	Training regions	Validation regions	Testing region
1	A + B: 70% (5439 samples)	A + B: 30% (2331 samples)	C (645 samples)
2	B + C: 70% (6951 samples)	B + C: 30% (2982 samples)	A (336 samples)
3	A + C: 70% (4802 samples)	A + C: 30% (2065 samples)	B (774 samples)

to avoid overfitting. To reduce the misclassification in the tile borders, the resulting image was created through a sliding window approach with steps of 20 pixels. All procedures [Fig. 1(a)] were executed for each one of the eight datasets to classify the three first level classes (forest, savanna, and grassland). The outputs of this phase are some thematic maps, accuracy measures as well as the best set of input data to be utilized in the next processing phase to achieve the hierarchical classification.

2.5 *Hierarchical Semantic Segmentation*

Using the input dataset, which yielded the best performance in the previous processing phase, we carried out the semantic segmentation approach to hierarchically classify the Cerrado physiognomies. Differently from the first level, the classes of the second level are unbalanced, i.e., they do not present similar distributions across the three image regions (Fig. 5). As the unbalanced class distribution can create artifacts, a different approach for sample generation was employed. Initially, the WorldView-2 image was partitioned into so-called superpixels[43] using the simple linear iterative clustering (SLIC), implemented in the Scikit-Learn Python package.[44] SLIC is an adaptation of the *k*-means algorithm,[45] which computes the weighted distance measure through a combination of color (in the CIELAB color space) and spatial proximity. As input for SLIC, we use the red, green, and blue bands, since the CIELAB color space is defined by the lightness (color brightness). It is also possible to control the superpixels compactness. If its value is large, spatial proximity is more relevant, therefore superpixels are more compact (close to a square in shape). However, when the compactness value is small, they adhere more to image boundaries and have less regular size and shape.[46] In this study, a compactness value equals to 400 was used. This value depends on the data values range used and was chosen empirically to create the superpixels that adhere well to the interest patterns.

Figure 6 shows how sample patches are generated. For each superpixel, a class was assigned based on the majority of corresponding pixel classes in the reference image. Thereafter, superpixels centroids were calculated and used as the center point for each sample of 160×160 pixels. Each sample corresponds to the pair composed of one patch for the reference image and one patch for the WorldView-2 image. For each class of interest, 1000 centroids were randomly selected to generate sample patches. The sample patches may contain transition areas between physiognomies, which is a positive aspect, because it enables the network to learn the context in which each

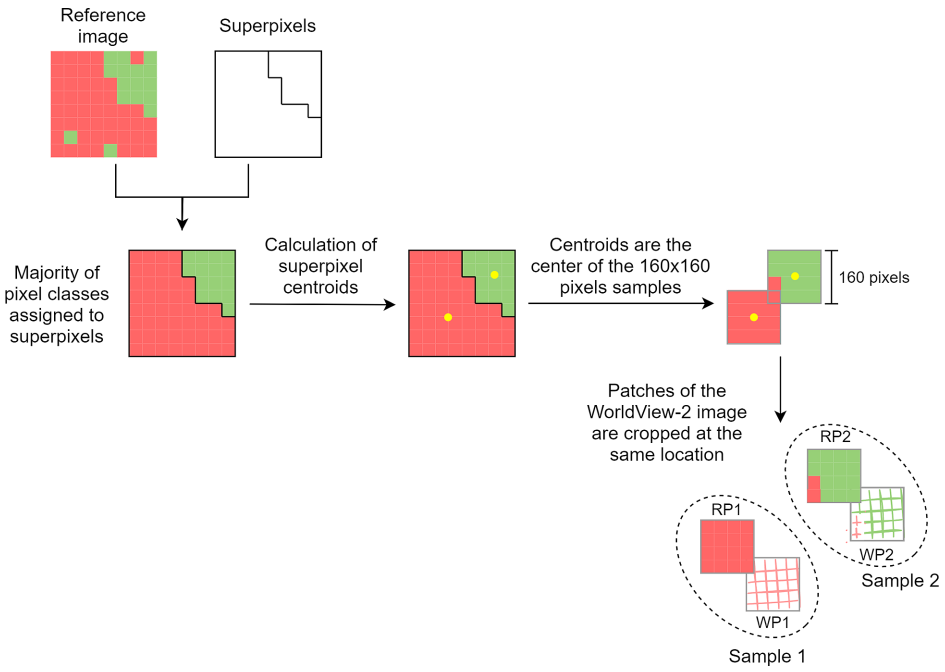

Fig. 6 Generation of sample patches. RP1 and WP1 are the reference and WorldView-2 patches of sample 1, respectively, and RP2 and WP2 are the reference and WorldView-2 patches of sample 2.

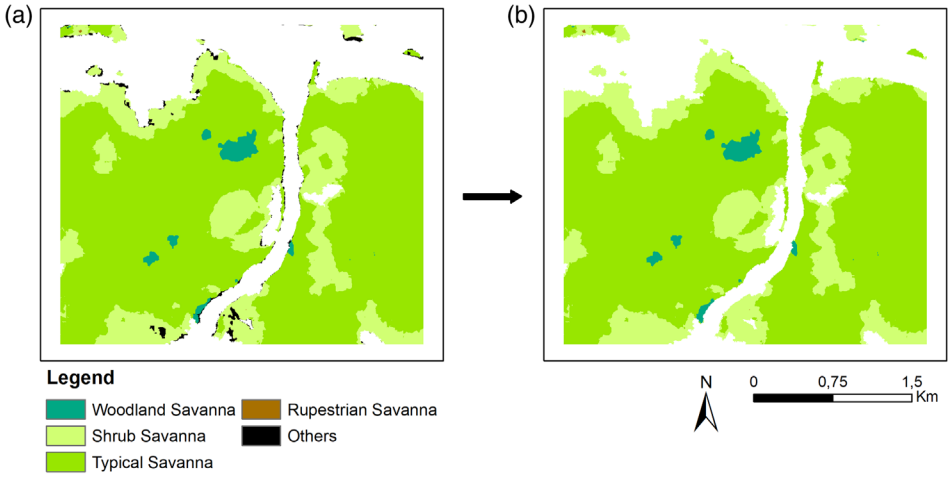

Fig. 7 Example of savanna classification (second level of classification) demonstrating: (a) minority edges predicted as others (in black); and (b) replacement of the others class by the second highest probability of the network output.

physiognomy occurs. Similar to the previous experiment (Sec. 2.4), all samples that contain any no-data value were excluded. Thus, the six data augmentation techniques mentioned before were applied for the remaining training and validation samples.

The complete samples set was randomly split: 70% and 30% were assigned for training and validation, respectively. The sliding window approach was also employed to create the results. To use the same sample generation approach in the entire hierarchical process, the semantic segmentation was repeated for the first level (forest, grassland, and savanna). Subsequently, the semantic segmentation approach was employed for each resulting savanna and grassland maps. The final Cerrado physiognomies map (and the respective accuracy metrics) is composed of the forest map (gallery forest), the savanna map (shrub savanna, typical savanna, woodland savanna, Rupestrian savanna, and *Vereda*), and the grassland map (open grassland, shrub grassland, Rupestrian grassland, and humid open grassland). These last two were generated in the second level of classification. These methodological steps are represented in Figs. 1(b) and 1(c).

The *Vereda* physiognomy has a minority presence in the BNP, so its area is not large enough to be included into deep learning classification. Therefore, this physiognomy was included in the first level of classification (as part of the overall savanna physiognomy), but it was manually identified in the second level. Consequently, it is present in the final map but not considered in the confusion matrices and accuracy metrics. Another relevant detail concerns the generation of samples for the savanna and grassland physiognomies in the second level of classification. When generating the samples of the four grassland types, for instance, any other class present in the sample patch (e.g., forest or savanna) was considered as others, a temporary class. If the pixels corresponding to others had simply been excluded, the network would be unable to understand patterns of transitions between grassland physiognomies and other classes of forest and savanna, for example. As the output of the network is a probability for each pixel and each class, when the network tried to classify a pixel as others, the second highest probability was considered, according to the example of Fig. 7.

2.6 *Accuracy Assessment*

The obtained semantic segmentation maps were then compared with the respective reference data, and confusion matrices were generated. Since we are performing a hierarchical classification, misclassifications on the first level will directly influence the results of the second level, i.e., if a pixel of shrub savanna was classified as grassland in the first level, it will still be considered as an error in the confusion matrix of the second level. Based on each confusion matrix, the following metrics were computed: OA, recall, precision, and $F1$-score. The OA corresponds to the percentage of pixels with the respective labels assigned correctly. Precision, also known

as user's accuracy, is the proportion of pixels predicted for a class and actually belonging to that class; it is the complement of the commission error. Recall, also known as producer's accuracy, is the proportion of pixels of a particular class successfully identified; it is the complement of omission error. The $F1$-score is the harmonic mean of precision and recall, computed for each class.

3 Results

3.1 *Assessment of the Spectral Input Information*

Table 3 presents the accuracies of the training and validation steps for the assessment of the spectral input information, as well as the number of epochs needed to stabilize the network for each experiment. It took about 5 h to train each network. Due to the patience parameter of 50, all

Table 3 Training and validation accuracies for all datasets in the three experiments (see description in Table 2). The highest values in training and validation for the three experiments are given in bold.

Input data	Exp.	Epochs	Training acc. (%)	Validation acc. (%)
RGB + LSMM	1	23	88.5	89.8
	2	52	91.7	88.2
	3	16	89.1	87.9
Eight band	1	47	91.3	88.6
	2	34	90.6	89.0
	3	77	92.1	89.4
LSMM	1	25	89.9	89.8
	2	31	89.8	89.4
	3	89	93.5	89.3
RGB + NIR1 + NIR2	1	5	87.7	88.9
	2	35	89.4	88.6
	3	93	93.0	89.5
RGB + NIR1 + NIR2 + RedEdge	1	55	**91.7**	89.9
	2	29	88.7	89.2
	3	87	**93.7**	**90.4**
RGB+RedEdge	1	10	88.6	89.4
	2	28	90.5	89.2
	3	53	91.4	89.8
RGB (532)	1	47	89.2	89.1
	2	70	92.5	88.7
	3	28	90.1	89.2
RGB + EVI2	1	47	89.7	**90.4**
	2	58	**92.7**	**89.7**
	3	32	89.4	88.6

trainings stopped before 100 epochs, and stabilization occurred between epochs 5 and 93. In general, all accuracy values were higher than 87.5%. The highest accuracies in training were reached using the RGB + NIR1 + NIR2 + RedEdge dataset in experiments 1 and 3, whereas, in experiment 2, the highest accuracy was obtained for RGB + EVI2. In the validation step, the highest accuracies for experiments 1 and 2 were reached using RGB + EVI2, whereas the highest accuracy in experiment 3 was observed with the RGB + NIR1 + NIR2 + RedEdge dataset.

For the test step, the OAs and $F1$-score per class are presented in Table 4. The OA varied from 87.4%, using RGB + LSMM, to 89.3% with the RGB + EVI2 dataset. It could be expected that the eight band dataset would achieve the highest performance, since it contains more bands and, consequently, most of spectral information. However, it obtained the second worst OA value of 87.6%. Despite presenting the lowest OA, the RGB + LSMM dataset had the highest $F1$-score (0.91) for the forest class. This is also reflected in the class delineation in the mapping result. For savanna and grassland, the highest $F1$-scores (0.92 and 0.84, respectively) were achieved with the same dataset with best OA, RGB + EVI2.

To analyze in more detail the results of Table 4, Fig. 8 shows selected patches of the WorldView-2 image, the reference, and the thematic maps using the RGB + EVI2 and the RGB + LSMM datasets. In this scale, the misclassified areas between grassland and savanna (G × S), grassland and forest (G × F), and savanna and forest (S × F) are highlighted in different colors. Despite the small difference between the best (RGB + EVI2) and the worst (RGB + LSMM) datasets of 1.9% points, the resulting maps show significant dissimilarities.

In all maps, the major areas of misclassification occur between grassland and savanna (G × S), followed by savanna and forest (S × F). There are only few areas of confusion between grassland and forest (G × F). This behavior is expected, since the confusions of classification occur mainly in transition areas, considering an increasing scale of vegetation density in the existing physiognomies (i.e., G × S and S × F). In addition, the higher forest $F1$-score with RGB + LSMM, already observed in Table 4, is also reflected in the maps. In Fig. 8, we notice a better delineation of forest areas when using this dataset, even better than in RGB + EVI2 dataset results.

3.2 *Detailed Physiognomies Mapping*

For the hierarchical classification, we used the input dataset composed by RGB + EVI2 bands, since it achieved the best performance in the assessment of the spectral input information for the first level, especially for savanna and grassland. For the first level of classification, the accuracy during training was of 97.9%, achieved after 147 epochs. The confusion matrix for the validation step is presented in Table 5. The matrix is presented in terms of number of pixels, and the OA was of 92.8%. Forest obtained the highest $F1$-score of 0.95, and the other two classes achieved

Table 4 OAs (%) and classes $F1$-score for all input datasets. The highest values are given in bold.

		Classes $F1$-score		
Input dataset	OA (%)	Grassland	Savanna	Forest
RGB + LSMM	87.4	0.81	0.90	**0.91**
Eight band	87.6	0.81	0.90	0.89
LSMM	87.8	0.82	0.90	**0.91**
RGB + NIR1 + NIR2	87.9	0.82	0.90	0.89
RGB + NIR1 + NIR2 + RedEdge	88.3	0.83	0.91	0.90
RGB + RedEdge	88.4	0.82	0.91	0.89
RGB (532)	88.6	0.83	0.91	0.89
RGB + EVI2	**89.3**	**0.84**	**0.92**	**0.91**

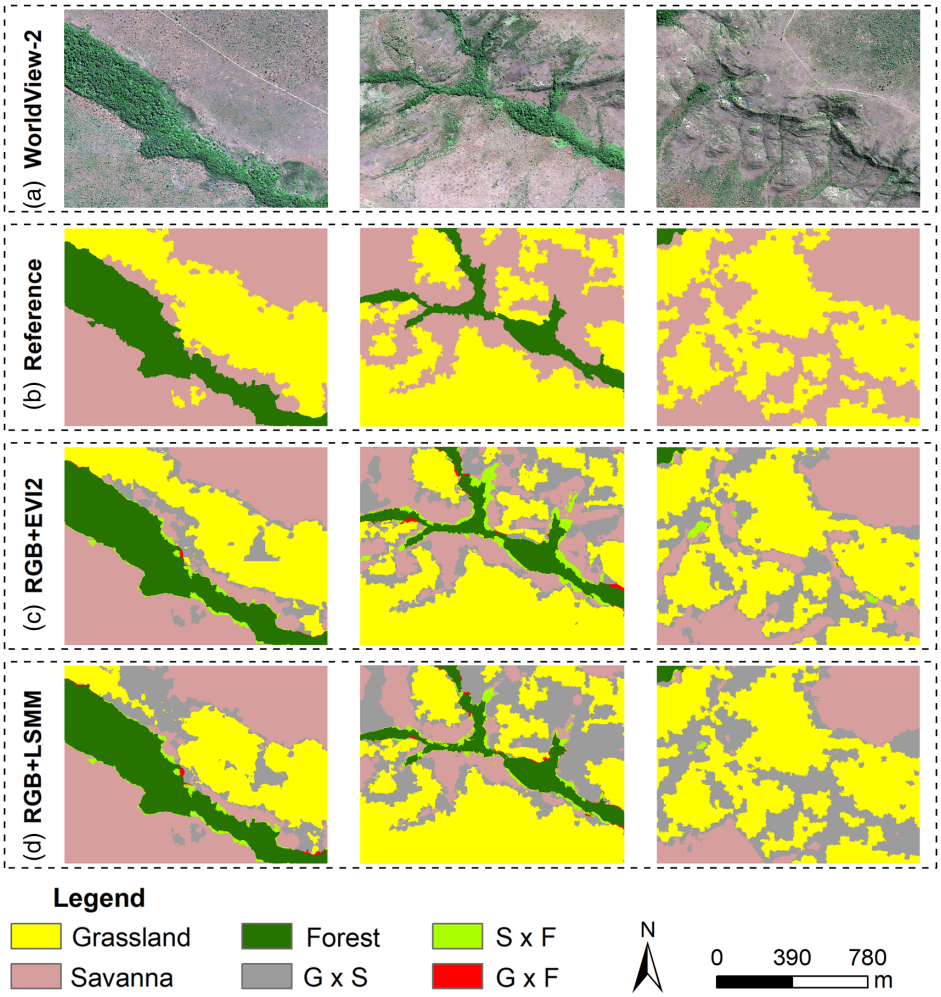

Fig. 8 Patches of: (a) the WorldView-2 image; (b) the reference data; (c) resulting thematic map using RGB + EVI2 dataset; and (d) resulting thematic map using RGB + LSMM datasets. G × S are the misclassified areas between grassland and savanna; S × F, between savanna and forest; and G × F, between grassland and forest.

Table 5 Confusion matrix (in number of pixels), precision, recall, and *F*1-score (highlighted in bold) for the first level of classification, using the RGB + EVI2 dataset. OA = 92.8%.

	Reference				
Predicted	Grassland	Savanna	Forest	Total	Precision
Grassland	5,906,317	315,673	8,084	6,230,074	**0.95**
Savanna	697,394	7,788,065	93,307	8,578,766	**0.91**
Forest	36,929	96,423	2,441,111	2,574,463	**0.95**
Total	6,640,640	8,200,161	2,542,502	—	—
Recall	**0.89**	**0.95**	**0.96**	—	—
*F*1-score	**0.92**	**0.93**	**0.95**	—	—

$F1$-scores higher than 0.91. The grassland recall (0.89) was the only metric lower than 0.90, as 10.5% of the grassland pixels were misclassified as savanna. In general, our deep learning approach yielded a very accurate classification of the three classes. The confusion between grassland and savannas, which presented the highest error rate, occurs mainly along the class borders, where it is indeed difficult to define when one class becomes another, even in field campaigns.

In a hierarchical classification, results of the first level affect directly the performance of the second level. Considering the detailed reference of savanna physiognomies and the resulting classification of the first level, it is shown that 92.7% of the woodland savanna was correctly included in savanna class, while 6.4% was misclassified as forest (Table 6). The typical savanna had the highest percentage of area classified as savanna (98.2%). While the Rupestrian savanna showed the highest percentage of misclassification, 88.3% of the area were classified as savanna, whereas 10.5% was classified as grassland. Pixels of savanna classified as forest or grassland were included as errors in the recall of the second level of classification for savanna physiognomies (Table 7).

Similar to Table 6, typical savanna showed the best performance in the classification of the detailed physiognomies, with $F1$-score of 0.91 (Table 7). The other three savanna physiognomies achieved $F1$-scores from 0.84 (shrub savanna) to 0.88 (Rupestrian savanna). Most of the misclassified pixels of woodland and shrub savannas were labeled as typical savanna. That was an expected behavior, since shrub, typical, and woodland savanna (in this order) compose an increasing scale of vegetation density and biomass. Regarding Rupestrian savanna, the network was able to identify its pattern among the savanna physiognomies, resulting in a precision of 0.91. However, the confusion of 10.5% of this physiognomy with grassland in the previous level of classification leads to a recall of only 0.84. During the training step, the accuracy for savanna physiognomies was of 96.6% when 190 epochs were executed. In the validation step, the OA was

Table 6 Analysis of the result of the second level of classification for savanna physiognomies regarding the first level resulting map (%).

	Reference (savanna physiognomies)			
Predicted (first level)	Woodland savanna	Typical savanna	Shrub savanna	Rupestrian savanna
Savanna (correct)	92.7	98.2	91.0	88.3
Classif. as grassland	0.9	0.7	8.4	10.5
Classif. as forest	6.4	1.1	0.6	1.2

Table 7 Confusion matrix (in number of pixels) for the savanna physiognomies, in the second level of classification (OA = 86.1%). Precision, Recall and $F1$-score are highlighted in bold.

	Reference					
Predicted	Woodland savanna	Typical savanna	Shrub savanna	Rupestrian savanna	Total	Precision
Woodland savanna	286,275	40,812	2,348	888	330,323	**0.87**
Typical savanna	23,794	4,107,171	295,190	6,227	4,432,382	**0.93**
Shrub savanna	2,258	325,652	2,402,537	4,019	2,734,466	**0.88**
Rupestrian savanna	232	10,219	11,411	231,146	253,008	**0.91**
Total	337,128	4,566,258	2,980,008	274,599	—	—
Recall	**0.85**	**0.90**	**0.81**	**0.84**	—	—
$F1$-score	**0.86**	**0.91**	**0.84**	**0.88**	—	—

of 95.6%. However, when errors from the first level of classification are propagated to the second level, the OA becomes 86.1%.

Considering the reference of the detailed grassland physiognomies, almost all (99.4%) open grassland were correctly classified as grassland in the first level (Table 8). Rupestrian, shrub, and humid open grassland had 94.1%, 85.5%, and 79.4% of their areas classified as grassland, respectively. For the shrub and humid open grassland, 14.2% and 18.0%, respectively, were misclassified as savanna in the first level. In the hierarchical classification system proposed by Ribeiro and Walter,[11] humid open grassland is a subtype of open grassland. However, it was considered an independent class in this work as it presents a pattern very different from the traditional open grassland. Besides that, preliminary experiments showed that separating these two classes increased the open grassland's OA by more than 2% points.

During the training step, the OA of the classification of the detailed grassland physiognomies was 96.3% with 171 epochs. In the validation, the network achieved an OA of 95.6%. After the inclusion of the errors from the first level of the classification, the grassland physiognomies OA decreased to 85.0%. The resulting confusion matrix is presented in Table 9. The $F1$-scores varied from 0.86 (humid open grassland) to 0.94 (open grassland). Shrub and Rupestrian grassland had $F1$-scores of 0.89 and 0.93, respectively. The largest amount of misclassified pixels of shrub grassland was classified as open grassland. For open, Rupestrian, and humid open grassland, the largest amount of misclassified pixels was classified as shrub grassland, the majority grassland physiognomy in the BNP.

The results of savanna and grassland physiognomies are presented in Fig. 9. This figure also includes gallery forest, generated in the first level of classification. Therefore, a detailed mapping of all physiognomies in the BNP is created. In addition to the reference and the predicted images,

Table 8 Analysis of the result of the second level of classification for grassland physiognomies regarding the first level resulting map (%).

	Reference (grassland physiognomies)			
Predicted (first level)	Open grassland	Shrub grassland	Rupestrian grassland	Humid open grassland
Grassland (correct)	99.4	85.5	94.1	79.4
Classif. as savanna	0.5	14.2	5.5	18.0
Classif. as forest	0.1	0.3	0.4	2.6

Table 9 Confusion matrix (in number of pixels) for the grassland physiognomies, in the second level of classification (OA = 85.0%). Precision, Recall and $F1$-score are highlighted in bold.

	Reference					
Predicted	Open grassland	Shrub grassland	Rupestrian grassland	Humid open grassland	Total	Precision
Open grassland	1,662,547	78,198	4,319	3,426	1,748,490	**0.95**
Shrub grassland	110,249	2,858,558	9,961	16,827	2,995,595	**0.95**
Rupestrian grassland	3,898	12,359	379,596	3,283	399,136	**0.95**
Humid open grassland	4,170	12,286	938	740,197	757,591	**0.98**
Total	1,791,208	3,461,580	419,588	962,759	—	—
Recall	**0.93**	**0.83**	**0.90**	**0.77**	—	—
$F1$-score	**0.94**	**0.89**	**0.93**	**0.86**	—	—

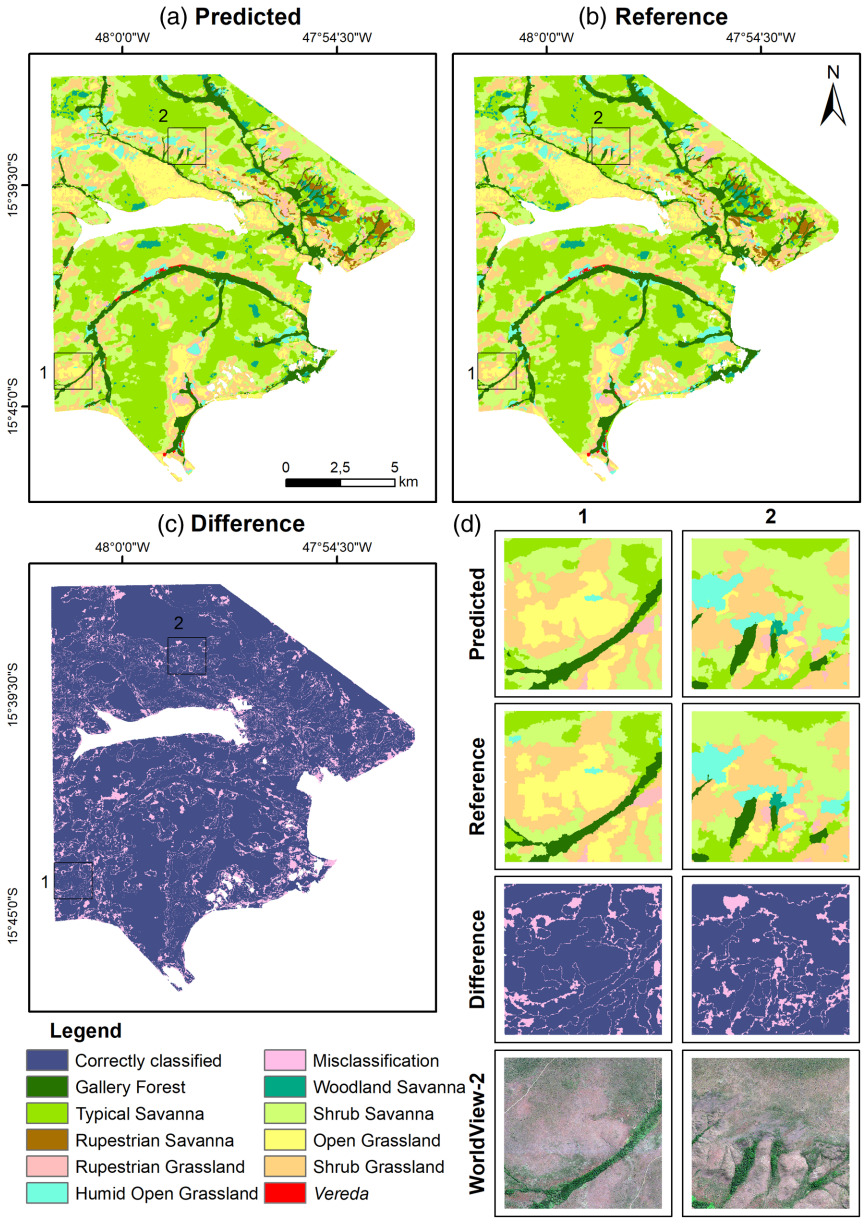

Fig. 9 Result of the detailed physiognomies mapping, showing: (a) predicted classified image; (b) reference image; (c) difference between predicted and reference images; and (d) zoom of two regions to show the result in more detail.

the image representing the differences between both is also presented to clearly show right and wrong results. To better visualize the development of the hierarchical classification, two zoomed regions are also presented. Just like in Tables 7 and 9, the errors from the first level of classification are carried over to the second level, so no misclassified area from the first level can become correctly classified [represented in dark blue, Fig. 9(c)] in the second level.

4 Discussion

4.1 *Assessment of Spectral Input Information and Samples Generation*

The majority of research in optical RS that applies deep learning techniques use images with the three most common channels (spectral bands), the RGB (red, green, and blue).[22,32] Others also

include an NIR channel.[25,33] These works rarely use all available satellite bands, sometimes because of the increase in processing time, or because the network architecture and the algorithms are prepared to use only three input layers. In the RS field, several extraspectral information, such as yellow or red-edge bands (available in WorldView-2), may be used. Zhang et al.[47] tested the efficiency of using datasets containing four and eight bands of the WorldView-2 and WorldView-3 images. The authors achieved better accuracies with the eight-band dataset, although they classified only different urban targets (e.g., buildings and roads). While more bands in general improve the results, in this study, it was observed that the use of some bands, such as the yellow band (present in the 8 band dataset), did not improve the classification of vegetation patterns. Thus, the increase in the number of bands is not necessarily directly related to an increase of OA.

Deep learning networks learn features from the training data to identify the desired classes. For this reason, it is believed that it is not necessary to give the network supplementary information, commonly called handcrafted features (e.g., vegetation indices, fractions of the LSMM), in addition to the satellite spectral bands.[48] However, our results showed the opposite: by combining a handcrafted feature (vegetation index) and original data, we obtained the best OA. The EVI2 enhanced the information regarding vegetation biomass in a way that does not occur when using only the red and the NIR bands. Thus, the OA of the RGB + EVI2 dataset was higher than the OA of other datasets containing red and NIR bands (eight band, RGB + NIR1 + NIR2, and RGB + NIR1 + NIR2 + RedEdge). A better performance when including vegetation indices was also observed by Kerkech et al.,[49] although the authors used different indices and tested them in a different domain (crop disease detection).

On the other hand, the inclusion of the three fractions (vegetation, soil, and shade) of the LSMM as input did not increase the OA. However, the vegetation and shade LSMM fractions highlighted and represented well the conditions of the dense and high vegetation as well as the presence of shades (due to differences in tree heights), improving the forest delineation. This resulted in the highest $F1$-score for this class using RGB+LSMM dataset. The extraction of LSMM fractions from high spatial resolution images is still useful, since the mixture of vegetation targets is still present in a 2×2 m pixels.[50,51] In a pixel of woodland savanna, for instance, the proportions of vegetation and shade fractions are mainly higher than in a pixel of shrub savanna. These fractions are widely used to classify vegetation in Cerrado[14,15,17,52,53] with traditional machine learning techniques. In the deep learning field, new methodologies have been tested to generate the LSMM fractions as results of some CNN,[54] but they were never tested as input layers in semantic segmentation using deep learning approach.

Regarding the generation of samples for the network training and validation, we employed two procedures. The first, used in the analysis of spectral input data, splits the image in three parts, two of which generated the training and validation samples and the last one was used to test the network. In the second procedure, employed in hierarchical classification, the samples were generated based on superpixel centroids. The first procedure is commonly used, since it is able to detect patterns of the same class in different regions of the image.[55,56] However, in a case with several classes with different occurrences across the study site, the second procedure may be more appropriate to ensure the classification represents all classes and contexts. In this study, the OA with the first procedure was of 89.3%, raising to 92.8% in the second one. Both of them for the first level of classification and using the RGB+EVI2 dataset.

Fu et al.[57] used an approach similar to the second procedure. They generated the samples using object centroids but with a multiresolution segmentation algorithm to classify general classes (e.g., water, road, and building). The accuracies of this approach with a CNN were 5% to 10% higher than the accuracies achieved with GEOBIA. Superpixel segmentation algorithms, such as SLIC, create more uniform objects, whereas traditional segmentation algorithms (e.g., multiresolution segmentation) generate objects with different sizes and shapes. Using this last case to create the network samples could potentially generate patches with irregular proportions of the classes due to the centroid positions:[58] in segments with irregular shapes, the centroids are not always located inside the patch. A supposed elongated and curved segment of gallery forest, for example, could generate a centroid outside that class and fail to represent that pattern.

4.2 *Hierarchical Classification*

The high OAs achieved in our results are assumed to be mainly related to two aspects: the deep learning methodology and the high spatial resolution of the WorldView-2 imagery. Using coarser spatial resolution and the same three classes of the first level, previous works[6,14,59] achieved lower OAs using traditional machine learning techniques. Although these works identified only three classes, it is worth noting that each class has a high intraclass variability. Therefore, the grassland class, for instance, contains the pattern (i.e., spectral behavior) of many different types of grasslands (four, in the BNP). Still using coarser spatial resolution (30 m), Ferreira et al.[17] and Schwieder et al.[18] improved the detail of vegetation classes. Schwieder et al.[18] had recalls of 86% and 64% for gallery forest and shrub grassland, respectively, whereas we achieved 95% and 83%.

In Ref. 17, the classes shrub Cerrado and woodland Cerrado are equivalent to our shrub grassland and shrub savanna classes, respectively. While they correctly classified 75% of both, we presented recalls of 83% and 81% for shrub grassland and shrub savanna. Thus, probably only the handcrafted features, used by the traditional machine learning algorithms, are not enough to acquire all the information needed about the vegetation classes. Improving the spatial resolution of the input imagery is also not enough to classifying the Cerrado vegetation. Keeping the RF algorithm and switching to the same spatial resolution used in this work (2 m), Girolamo Neto[15] and Neves et al.[12] obtained 88.9% and 88.2%, respectively, when differentiating the classes forest, savanna, and grassland. When dealing with the physiognomies, they achieved recalls of 39.15% and 39.25% for shrub savanna and 63.32% and 72.51% for shrub grassland, respectively, whereas we achieved 81% and 83% for these two classes.

Each detailed physiognomy has its own floristic composition, vegetation density, and edaphic factors. The woodland savanna, due to its dense vegetation, is often confused with forests, whereas the shrub savanna, with a more sparse vegetation, is confused with shrub grassland, a grassland physiognomy.[12,15,16] The confusion between forest and grasslands is rather rare. Although such a confusion is surprising, it may be related to the presence of humid open grassland areas in the BNP. This physiognomy is located predominantly close to the gallery forests[11] and, consequently, the misclassified areas occur at the boundaries between these two classes.

The physiognomies, according to the classification system used in this work,[11] present an increasing scale of density and, consequently, biomass. Under these circumstances, the most common errors occur in transition areas between the physiognomies. Although this error is more intense when coarser images are used, this happen to be an issue in every mapping of savanna physiognomies.[16–18] The majority of works that classified detailed physiognomies using traditional machine learning techniques performed the validation using independent random points.[12,15,16,52] As we performed a semantic segmentation approach, the validation samples (as well as the training samples) were independent patches (160×160 pixels) entirely classified. Thus, our approach generates a more robust evaluation of the delineation of the physiognomies and, consequently, the misclassification in transition areas.

The use of a hierarchical classification approach intended to minimize the confusion between savanna and grassland physiognomies or between any of them with forest patterns. Despite accounting for the misclassifications of the first level, this approach was efficient to map the detailed physiognomies, since it achieved higher accuracy rates than other works that intended to perform a similar task without using hierarchy.[15–18] Compared with the other few works that also used hierarchical approaches,[12,13] ours presented superior accuracy rates. Thus, we demonstrated the potential of applying deep learning techniques to open problems in the RS field.

It is important to note, though, that in many applications, and specially in RS, one of the main limitations of deep learning is the requirement of a large amount of training samples to achieve acceptable results.[60] The amount of samples used in this work (see Secs. 2.4 and 2.5) was satisfactory to differentiate the physiognomies in the BNP. Despite covering a proportionally small area of the Cerrado biome, the BNP is a preserved area representative of the Cerrado ecosystems and contains the major physiognomies of the biome.[17] However, due to the physiognomy heterogeneity across the Cerrado biome,[61] the application of our methodology in the entire biome would require the inclusion of more samples during the training phase.

5 Conclusions

This study proposed and evaluated a new methodology based on deep learning techniques to hierarchically classify the Cerrado physiognomies, according to the Ribeiro and Walter[11] classification system. The use of deep learning techniques enabled the creation of maps with higher detail and accuracy than other techniques such as GEOBIA and general machine learning. Although it does not completely prevent misclassifications, the use of a hierarchical approach reduced the confusion between detailed classes. Testing several datasets as input in the networks showed that the best dataset was composed by RGB bands plus EVI2.

Each Cerrado physiognomy has a different amount of biomass above and below ground and a unique biodiversity. Consequently, the proper identification and delineation of the Cerrado physiognomies are fundamental to truly understand the role of savanna biomes in the global carbon cycle. In addition, Cerrado vegetation maps can be used as a proxy in biodiversity studies, since a high rate of endemism can be found in its physiognomies. The greatest limitation when mapping the physiognomies of the Cerrado, the second largest biome of a continental-size country such as Brazil, is the lack of reference data.

The difficulty in differentiating the physiognomy patterns associated to the biome extension results in few options of vegetation maps, available only for three classes (usually forest, savanna, and grassland) and with a spatial resolution of around 30 m. This limitation is even more problematic when dealing with a semantic segmentation approach, since reference points are not enough and entirely classified patches are required as samples for network training. In addition, we should point out that using a high spatial resolution image also has some limitations to provide a ground truth regarding the Cerrado vegetation types. Due to the huge spatial variability of this vegetation, even in field campaigns it is hard to determine where a physiognomy ends and another one starts.

Under these circumstances, we performed the hierarchical classification methodology using deep learning in a relevant protected area, the BNP, and was quite efficient to classify the three major groups of physiognomies, in a first level, and 10 detailed physiognomies, in the second level. As the Cerrado contains several particularities across the biome, the reproduction of the method for the entire biome would require the availability of high spatial resolution images and reference data in the same spatial resolution to generate more samples for training and validation.

For future work, we suggest the inclusion of additional satellite data to consider other aspects of the physiognomies (e.g., satellite image time series to include the physiognomies seasonality in the analysis) in the mapping. As investigations of time series (without considering the spatial context) with well-known techniques have been carried out already and may not be enough, it is also suggested to keep the high spatial resolution and apply deep learning architectures that are appropriate for time series data, such as the long short-term memory networks. The inclusion of active RS data, such as radar and LiDAR (light detection and ranging) data, can also provide additional information, especially of the vegetation structure, to differentiate the physiognomies patterns.

Acknowledgments

The authors thank the DigitalGlobe Foundation for supplying the image for this work and cooperating with INPE. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and the Brazilian National Research Council (CNPq) (Grant Nos. 140372/2017-2 and 303360/2019-4). Part of this study was carried out while the first author was working at IPI (Institut für Photogrammetrie und GeoInformation), Leibniz Universität Hannover, Germany, with a scholarship from the Deutscher Akademischer Austauschdienst (DAAD). The authors declare no conflict of interest.

References

1. B. B. Strassburg et al., "Moment of truth for the cerrado hotspot," *Nat. Ecol. Evol.* **1**(4), 0099 (2017).

2. L. G. Ferreira et al., "Equivalent water thickness in savanna ecosystems: MODIS estimates based on ground and EO-1 hyperion data," *Int. J. Remote Sens.* **32**(22), 7423–7440 (2011).
3. S. C. Ribeiro et al., "Above- and belowground biomass in a Brazilian cerrado," *For. Ecol. Manage.* **262**(3), 491–499 (2011).
4. R. A. Mittermeier et al., "Global biodiversity conservation: the critical role of hotspots," in *Biodiversity Hotspots*, F. E. Zachos and J. C. Habel, Eds., pp. 3–22, Springer, Berlin, Heidelberg (2011).
5. MMA – Ministério do Meio Ambiente, "Cadastro Nacional de Unidades de Conservação," 2000, https://www.mma.gov.br/areas-protegidas/cadastro-nacional-de-ucs.
6. INPE – National Institute for Space Research, "Projeto TerraClass Cerrado – mapeamento do uso e cobertura vegetal do Cerrado," 2015, http://www.dpi.inpe.br/tccerrado/Metodologia_TCCerrado_2013.pdf.
7. INPE - National Institute for Space Research, "Annual deforestation in Brazilian Savannah," 2020, http://www.obt.inpe.br/cerrado.
8. R. D. Françoso et al., "Habitat loss and the effectiveness of protected areas in the cerrado biodiversity hotspot," *Nat. Conserv.* **13**(1), 35–40 (2015).
9. F. M. Resende et al., "Consequences of delaying actions for safeguarding ecosystem services in the Brazilian Cerrado," *Biol. Conserv.* **234**, 90–99 (2019).
10. J. Grace et al., "Productivity and carbon fluxes of tropical savannas," *J. Biogeogr.* **33**(3), 387–400 (2006).
11. J. F. Ribeiro and B. M. T. Walter, "As principais fitofisionomias do bioma cerrado," *Cerrado* **1**, 151–212 (2008).
12. A. K. Neves et al., "Hierarchical classification of Brazilian savanna physiognomies using very high spatial resolution image, superpixel and Geobia," in *IEEE Int. Geosci. and Remote Sens. Symp.*, IEEE, pp. 3716–3719 (2019).
13. F. F. Ribeiro et al., "Geographic object-based image analysis framework for mapping vegetation physiognomic types at fine scales in neotropical savannas," *Remote Sens.* **12**(11), 1721 (2020).
14. A. Alencar et al., "Mapping three decades of changes in the Brazilian savanna native vegetation using Landsat data processed in the Google earth engine platform," *Remote Sens.* **12**(6), 924 (2020).
15. C. D. G. Neto, "Identificação de fitofisionomias de Cerrado no Parque Nacional de Brasília utilizando random forest aplicado a imagens de alta e média resoluções espaciais," PhD Thesis, National Institute for Space Research (INPE), p. 186 (2018).
16. A. D. Jacon et al., "Seasonal characterization and discrimination of savannah physiognomies in brazil using hyperspectral metrics from Hyperion/EO-1," *Int. J. Remote Sens.* **38**(15), 4494–4516 (2017).
17. M. Ferreira et al., "Spectral linear mixture modelling approaches for land cover mapping of tropical savanna areas in Brazil," *Int. J. Remote Sens.* **28**(2), 413–429 (2007).
18. M. Schwieder et al., "Mapping Brazilian savanna vegetation gradients with landsat time series," *Int. J. Appl. Earth Obs. Geoinf.* **52**, 361–370 (2016).
19. C. M. Souza et al., "Reconstructing three decades of land use and land cover changes in Brazilian biomes with Landsat archive and earth engine," *Remote Sens.* **12**(17), 2735 (2020).
20. J. C. O. Filho, "Avaliação do uso da abordagem orientada-objeto com imagens de alta resolução rapideye na classificação das fitofisionomias do cerrado," Master's Thesis, University of Brasília – UNB, Brasília, DF – Brazil, p. 44 (2017).
21. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
22. E. Guirado et al., "Deep-learning versus Obia for scattered shrub detection with Google earth imagery: *Ziziphus lotus* as case study," *Remote Sens.* **9**(12), 1220 (2017).
23. M. Brandt et al., "An unexpectedly large count of trees in the west African Sahara and Sahel," *Nature* **587**(7832), 78–82 (2020).
24. D. Torres et al., "Semantic segmentation of endangered tree species in Brazilian savanna using deeplabv3+ variants," in *IEEE Latin Am. GRSS & ISPRS Remote Sens. Conf.*, IEEE, pp. 515–520 (2020).
25. K. Nogueira et al., "Towards vegetation species discrimination by using data-driven descriptors," *in 9th IAPR Workshop Pattern Recognit. Remote Sens.*, IEEE, pp. 1–6 (2016).

26. A. Neves et al., "Semantic segmentation of brazilian savanna vegetation using high spatial resolution satellite data and U-net," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **V-3-2020**, 505–511 (2020).
27. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
28. S. Kumar, "Deep U-net for satellite image segmentation," 2018, https://github.com/reachsumit/deep-unet-for-satellite-image-segmentation/.
29. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3431–3440 (2015).
30. Y. LeCun et al., "Handwritten digit recognition with a back-propagation network," in *Adv. Neural Inf. Process. Syst.*, pp. 396–404 (1990).
31. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105 (2012).
32. T. Kattenborn, J. Eichel, and F. E. Fassnacht, "Convolutional neural networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery," *Sci. Rep.* **9**(1), 17656 (2019).
33. S. E. Jozdani, B. A. Johnson, and D. Chen, "Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification," *Remote Sens.* **11**(14), 1713 (2019).
34. C. Yang, F. Rottensteiner, and C. Heipke, "A hierarchical deep learning framework for the consistent classification of land use objects in geospatial databases," *ISPRS J. Photogramm. Remote Sens.* **177**, 38–56 (2021).
35. Y. Guo et al., "CNN-RNN: a large-scale hierarchical image classification framework," *Multimedia Tools Appl.* **77**(8), 10251–10271 (2018).
36. L. Ferreira et al., "On the use of the EOS–MODIS vegetation indices for monitoring the Cerrado region, Brazil: insights and perspectives," Anais X SBSR, Foz do Iguacu, pp. 20–26 (2001).
37. ICMBIO - Instituto Chico Mendes de Conservação da Biodiversidade, "Plano de Manejo do Parque Nacional de Brasília," 1998, http://www.icmbio.gov.br/portal/images/stories/imgs-unidades-coservacao/PARNA%20Brasilia.pdf.
38. T. Perkins et al., "Retrieval of atmospheric properties from hyper and multispectral imagery with the FLAASH atmospheric correction algorithm," *Proc. SPIE* **5979**, 59790E (2005).
39. Z. Jiang et al., "Development of a two-band enhanced vegetation index without a blue band," *Remote Sens. Environ.* **112**(10), 3833–3845 (2008).
40. Y. E. Shimabukuro and J. A. Smith, "The least-squares mixing models to generate fraction images derived from remote sensing multispectral data," *IEEE Trans. Geosci. Remote Sens.* **29**(1), 16–20 (1991).
41. F. Chollet et al., "keras," 2015, https://keras.io.
42. M. Abadi et al., "Tensorflow: large-scale machine learning on heterogeneous systems" (2015).
43. C. Çila and A. A. Alatan, "Efficient graph-based image segmentation via speeded-up turbo pixels," in *IEEE Int. Conf. Image Process.*, IEEE, pp. 3013–3016 (2010).
44. F. Pedregosa et al., "Scikit-learn: machine learning in Python," *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. J. MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symp. Math. Stat. and Probab.*, Oakland, California, Vol. 14, pp. 281–297 (1967).
46. R. Achanta et al., "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(11), 2274–2282 (2012).
47. P. Zhang et al., "Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery," *Sensors* **18**(11), 3717 (2018).
48. X. X. Zhu et al., "Deep learning in remote sensing: a comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.* **5**(4), 8–36 (2017).
49. M. Kerkech, A. Hafiane, and R. Canals, "Deep leaning approach with colorimetric spaces and vegetation indices for vine diseases detection in UAV images," *Comput. Electron. Agric.* **155**, 237–243 (2018).

50. J. Nichol and M. S. Wong, "Remote sensing of urban vegetation life form by spectral mixture analysis of high-resolution ikonos satellite images," *Int. J. Remote Sens.* **28**(5), 985–1000 (2007).
51. X. Sun et al., "Vegetation abundance and health mapping over southwestern antarctica based on worldview-2 data and a modified spectral mixture analysis," *Remote Sens.* **13**(2), 166 (2021).
52. T. Miura et al., "Discrimination and biophysical characterization of Cerrado physiognomies with EO-1 hyperspectral hyperion," *Simpósio Brasileiro Sens. Remoto* **11**, 1077–1082 (2003).
53. C. H. Amaral et al., "Mapping invasive species and spectral mixture relationships with neotropical woody formations in southeastern brazil," *ISPRS J. Photogramm. Remote Sens.* **108**, 80–93 (2015).
54. X. Zhang et al., "Hyperspectral unmixing via deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.* **15**(11), 1755–1759 (2018).
55. Z. Pan et al., "Deep learning segmentation and classification for urban village using a worldview satellite image based on U-net," *Remote Sens.* **12**(10), 1574 (2020).
56. R. Andrade et al., "Evaluation of semantic segmentation methods for deforestation detection in the Amazon," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **XLIII-B3-2020**, 1497–1505 (2020).
57. T. Fu et al., "Using convolutional neural network to identify irregular segmentation objects from very high-resolution remote sensing imagery," *J. Appl. Remote Sens.* **12**(2), 025010 (2018).
58. Y. Chen, D. Ming, and X. Lv, "Superpixel based land cover classification of vhr satellite image combining multi-scale CNN and scale parameter estimation," *Earth Sci. Inf.* **12**(3), 341–363 (2019).
59. H. Bendini et al., "Combining environmental and landsat analysis ready data for vegetation mapping: a case study in the Brazilian savanna biome," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **XLIII-B3-2020**, 953–960 (2020).
60. L. Ma et al., "Deep learning in remote sensing applications: a meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.* **152**, 166–177 (2019).
61. E. E. Sano et al., "Cerrado ecoregions: a spatial framework to assess and prioritize brazilian savanna environmental diversity for conservation," *J. Environ. Manage.* **232**, 818–828 (2019).

Alana K. Neves is an associate researcher of Brazil Data Cube project at National Institute for Space Research (INPE). She is an environmental engineer and received her master's and PhD degrees in remote sensing from INPE in 2017 and 2021, respectively. Her main research topics are digital image processing, GEOBIA, time series, machine learning, and deep learning applied to the classification and monitoring of native vegetation, deforestation, regeneration, and land use and land cover.

Thales S. Körting received his PhD in remote sensing, with an MS degree in applied computing, both titles obtained from Brazil's National Institute for Space Research – INPE. He is also a computer engineer at FURG. Nowadays, he is a researcher at INPE. His research areas include remote sensing image segmentation, multitemporal analysis, image classification, data mining algorithms, and artificial intelligence.

Leila M. G. Fonseca is graduated in electrical engineering, master of electrical engineering and computer science, and PhD in applied computing. Since 1985, she has been working at the National Institute for Space Research (INPE), Brazil. She has been the head of Image Processing Division (2011 to 2014) and the director of Earth Observation Coordination at INPE (2014 to 2018). Currently, she works on environmental monitoring projects and the field of image processing and pattern recognition applied to remote sensing.

Anderson R. Soares has an advanced degree in geoprocessing from the Federal Institute of Paraíba, Brazil, his master's degree in geodesic science and geoinformation technology from the University of Pernambuco, Brazil, and his PhD in remote sensing from Brazilian Institute

for Space Research (INPE), Brazil, in 2019. He is currently a data scientist at Cognizant Technology Solutions/Bayer Crop Science. His main research interests are geographic object-based image analysis (GEOBIA), time series analysis, and spatiotemporal segmentation.

Cesare D. Girolamo-Neto is a postdoctoral research fellow at Vale Institute of Technology, Brazil, and holds a PhD in remote sensing from the National Institute for Space Research (INPE), Brazil. He has experience in image processing and machine learning applied to several remote sensing applications, such as vegetation, agriculture, deforestation, and forest fires. He has a strong knowledge of Brazilian biomes, especially in the Amazon and savannah.

Christian Heipke is a professor of photogrammetry and remote sensing, Leibniz Universität Hannover, where he leads a group of about 25 researchers. His professional interests comprise all aspects of photogrammetry, remote sensing, image understanding, and their connection to computer vision and GIS. He has authored or coauthored more than 300 scientific papers, more than 70 of which appeared in peer-reviewed international journals. From 2012 to 2016, he was ISPRS secretary general; currently he serves as ISPRS president.