

Assessing satellite image classification for the Cerrado biome with the integration of terrain data: a comparative analysis of machine learning algorithms

Carlos Eduardo Fernandes Holanda¹ - ORCID: 0009-0004-5221-0314

Diego Rodrigues Macedo² - ORCID: 0000-0002-1178-4969

Rodrigo Affonso Albuquerque Nóbrega³ - ORCID: 0000-0001-7058-5903

¹ Universidade Federal de Minas Gerais, Instituto de Geociências, Programa de Pós-Graduação em Geografia, Belo Horizonte – Minas Gerais, Brasil.
E-mail: caduholanda@yahoo.com.br

² Universidade Federal de Minas Gerais, Instituto de Geociências, Departamento de Geografia, Belo Horizonte – Minas Gerais, Brasil.
E-mail: diegorm@ufmg.br

³ Universidade Federal de Minas Gerais, Instituto de Geociências, Departamento de Cartografia, Belo Horizonte – Minas Gerais, Brasil.
E-mail: raanobrega@ufmg.br

Received in 01st August 2024.

Accepted in 17th February 2025.

Abstract:

The present paper addresses the relevance of incorporating terrain data for analyzing satellite images in mapping land use and land cover in the Cerrado biome. Assuming that terrain influences the dynamics of landscape changes, the present investigation evaluates three machine learning algorithms: Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) in a watershed with significant topographic heterogeneity. The present study evaluated variations in image classification using Sentinel-2 satellite data. It also included a composite analysis blending Sentinel-2 data and information derived from the Shuttle Radar Topography Mission (SRTM). The results indicate that SVM exhibited the best performance, both with and without terrain data. Although DT demonstrated satisfactory results, the performance was inferior to SVM. However, DT's significantly shorter processing time presents an advantage in scenarios involving large territorial extents or computational constraints. Conversely, RF had a processing time similar to DT but recorded the lowest statistical indices among the three algorithms. Additionally, including the data cube containing elevation data and its derivatives yielded improved land use and land cover classification results for all evaluated algorithms compared to images without terrain data. This demonstrates the robustness of the process and the significant improvement in the quality of the final product.

Keywords: Land Use and Land Cover; Remote Sensing; Digital Image Processing; Nova Ponte Dam Watershed.

How to cite this article: HOLANDA CEF, MACEDO DR, NÓBREGA RAA. Assessing satellite image classification for the Cerrado biome with the integration of terrain data: a comparative analysis of machine learning algorithms. *Bulletin of Geodetic Sciences*. 31: e2025002, 2025.



This content is licensed under a Creative Commons Attribution 4.0 International License.

1. Introduction

Brazil encompasses six distinct biomes: Amazon, Atlantic Forest, Cerrado, Caatinga, Pampa, and Pantanal. The Cerrado is the second largest, representing 24% of the Brazilian territory (IBGE, 2024). The Cerrado faces economic pressures, such as grain-oriented monoculture and livestock farming, resulting in deforestation of over 45% of the original cover. Urbanization and mining also contribute to environmental degradation (Klink and Machado, 2005). Covering an area of 2,036,448 km², the Cerrado is exceptionally biodiverse, supporting endemic species, and is considered a biodiversity hotspot (Myers et al, 2000). Its deforestation threatens regional climate and water availability (Rodrigues et al, 2022). In the state of Minas Gerais, the Cerrado represents 57% of the state's original area (IBGE, 2024) and is essential for three of the major South American river basins (Paraná, São Francisco, and Tocantins-Araguaia), contributing 43% of Brazil's surface water (Strassburg et al, 2017). Anthropogenic pressures have fragmented Cerrado habitats, compromising biodiversity, and currently, around 61% of the Cerrado area has been devastated (Mapbiomas, 2024). Projections indicate that 31-34% of the remaining Cerrado areas may be lost by 2050 due to agricultural expansion and inadequate environmental protection (Strassburg et al, 2017).

Topography plays a fundamental role in determining land use and land cover dynamics, directly influencing human activities and the development of settlements (Yang et al, 2022). Mountainous terrains, for example, tend to have sparse anthropogenic uses concentrated in flat areas and valleys due to access difficulties and construction restrictions (Zhang et al, 2019). Conversely, flat and low-altitude regions tend to be more densely populated and used for agriculture and urbanization due to easy access, favorable topography, and availability of water resources (Yang et al, 2022). Additionally, slope, soil type, and altitude influence the viability of different economic activities, such as agriculture, mining, and tourism, thus shaping land use and land cover in specific areas (Steel et al, 2010). Accordingly, in recent years, the incorporation of topographic information (e.g., elevation and slope) has been increasingly explored in classification processes, either through decision tree approaches (Francisco and Almeida, 2012; Santos, Francisco and Almeida, 2015) or by integrating multiple data sources (Tian et al, 2024). In this context, topographic effects are a critical factor that directly influences classification outcomes, particularly in results based on vegetation indices (Al-Doski et al, 2022; Moreira et al, 2016; Sang et al, 2021).

Remote sensing monitoring is crucial for understanding changes in land use and land cover by estimating deforestation, overgrazing (Li et al, 2024), and surface water dynamics (Mashala et al, 2023), highlighting its essential role in environmental management (Chakraborty, 2021). The multispectral capability of satellites provides vital information on different parts of the electromagnetic spectrum, aiding in identifying objects of interest (Mather and Koch, 2011). This technology has been crucial for natural resource researchers since the launch of the Landsat-1 satellite in 1972 (USGS, 2024) to more recent systems like Sentinel-2, launched in 2015 by the European Space Agency, offering a refined spatial resolution for monitoring land cover changes over time (ESA, 2024).

Satellite images vary in spatial, spectral, temporal coverage, and viewing width, impacting data availability and product choice, which is essential for environmental applications (Hansen et al, 2008; Chakraborty, 2021). Remote sensing data obtained by Radio Detection and Ranging (RADAR) systems, such as the Shuttle Radar Topography Mission (SRTM) images, provide surface elevation models covering almost the entire planet. Using a 30-meter spatial resolution, Digital Elevation Models (DEMs) can generate georeferenced structured matrices of lines and columns representing surface elevation data per pixel (USGS, 2024).

Advances in orbital sensors and image classification algorithms increase interest in remote sensing land use and land cover data mapping (Adam et al, 2014; Mashala et al, 2023). Traditional methods, such as maximum likelihood and minimum distance, remain common, but limited capacity has led to the search for Computational Intelligence (CI) approaches (Mather and Koch, 2011; Zhong et al, 2018). In this context, machine learning, a CI subfield, develops algorithms for machines to identify patterns in data, record rules, and make decisions (Akar and Güngör, 2012). Advanced classification algorithms based on artificial intelligence have made significant progress

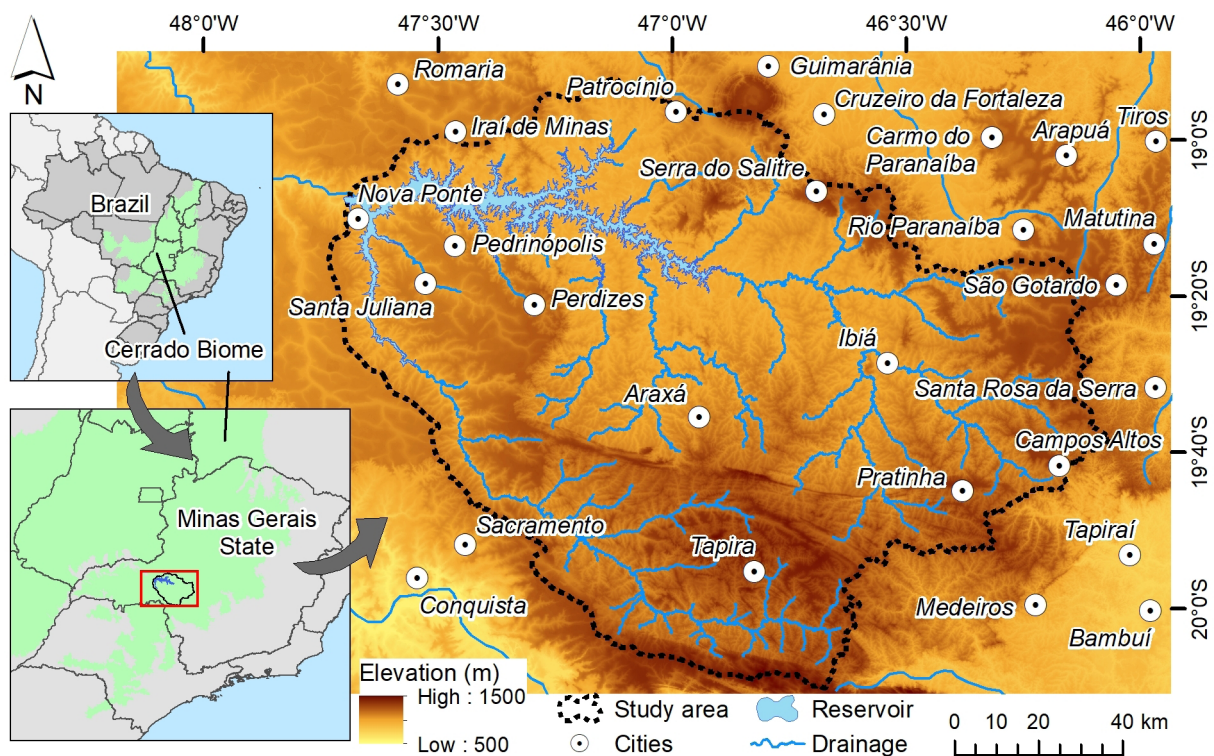
in image classification. Some of these include artificial neural networks (ANN), decision trees (DT), support vector machines (SVM), object-based algorithms, sub-pixel-based algorithms, random forests (RF), bagging, boosting, k-nearest neighbor, and contextual algorithms (Blaschke, 2010; Zhong et al, 2018). Unlike traditional pixel-based applications (Akar and Güngör, 2012), classifiers with CI approaches and learning foundations have been developed to obtain more reliable information from satellite images with higher accuracy.

The present study assumes that adding topographic information will improve the classification of Sentinel-2 satellite images, as elevation and derivatives such as terrain slope are conditioning factors for land occupation. To confirm this assumption, the central question is: Which machine learning-based image classifiers perform best in classification when using Sentinel-2 images and the SRTM terrain model? In this context, the objective of this article is to analyze and compare the performance of three machine learning-based algorithms - Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) - for land use and land cover classification in the Cerrado using an image cube composed of passive optic satellite data, active radar digital terrain model and its derived products.

2. Materials and Methods

2.1 Study Area

The study area is in the Cerrado biome and encompasses the watershed boundary of the Nova Ponte hydroelectric power dam. The total area is 15,358 km² and covers fifteen municipalities in the Alto Paranaíba region in the Minas Gerais state (Figure 1).



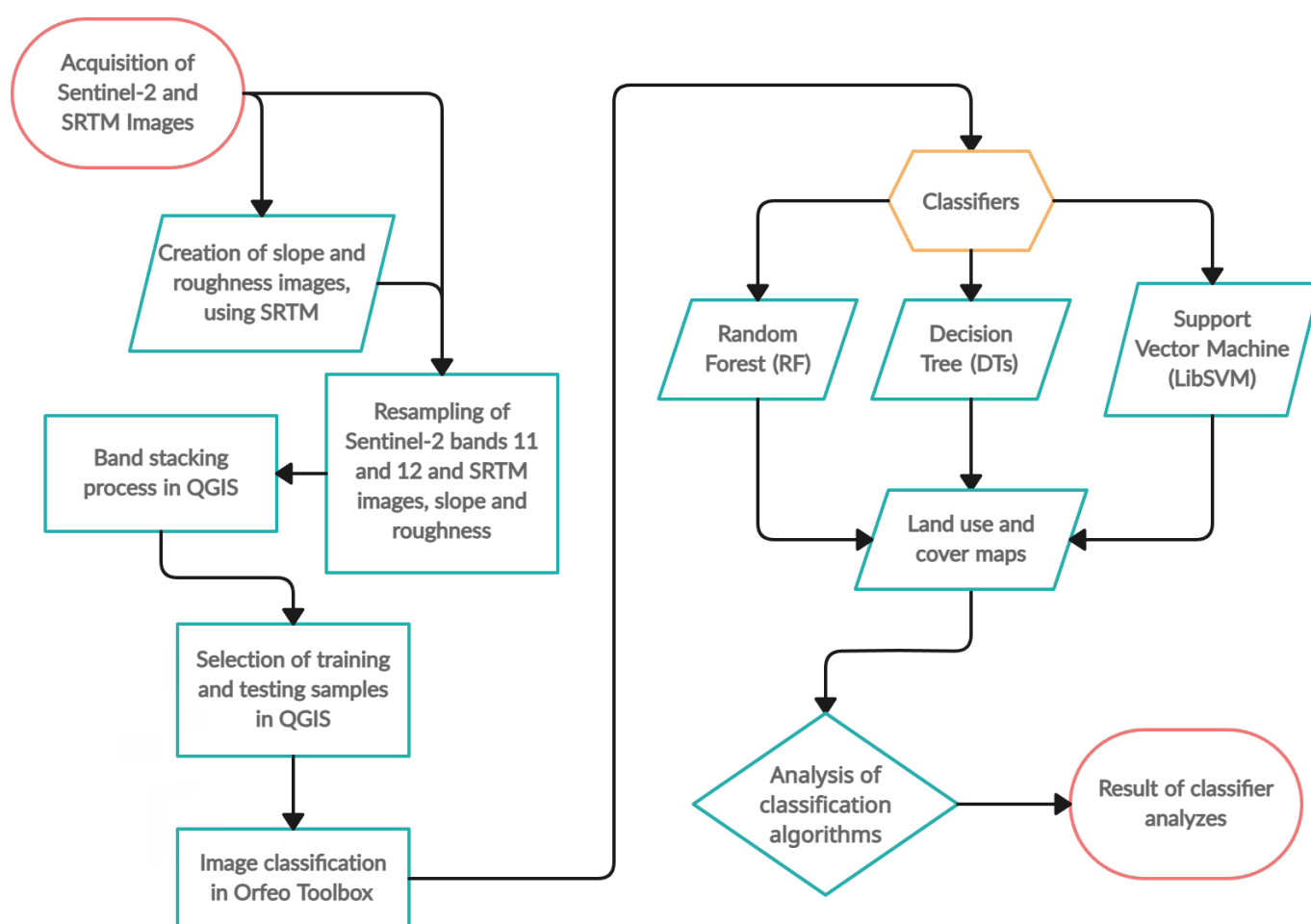
Source: Organized by the authors.

Figure 1: Location of the study area.

The region's economic base is agriculture and livestock (Macedo et al, 2014; Callisto et al, 2019). The study area includes flooded regions to the north-northeast by the Nova Ponte dam, with average altitudes of 800 m, and higher regions to the south, reaching 1,200 m (Callisto et al, 2019) (Figure 1). The Serra da Canastra National Park, located in the extreme south of São Roque de Minas municipality, contains the highest altitudes of the basin.

2.2 Methodology

Six bands of the Sentinel-2 MSI sensor (2, 3, 4, 8, 11, and 12) were analyzed and stacked with three images derived from the SRTM mission (elevation, slope, and roughness), resulting in a final image as a data cube, as shown in the flowchart (Figure 2).



Source: Organized by the authors.

Figure 2: Flowchart of activities developed in the study.

SRTM images were acquired via the Earth Explorer interface (USGS, 2024), and Sentinel-2 images were downloaded from the Copernicus Open Access Hub interface (ESA, 2024). All scenes involving the study area were acquired with radiometric calibration preprocessing, noise removal, and geometric correction. The Sentinel-2 imagery used corresponded to six scenes from 2019: 23KKV, 23KKU, 23KKT, 23KLV, and 23KLU from July 14; and 23KLT from July 4. SRTM 1 arc second data included the tiles S19W048, S19W047, S20W048, S20W047, S21W048,

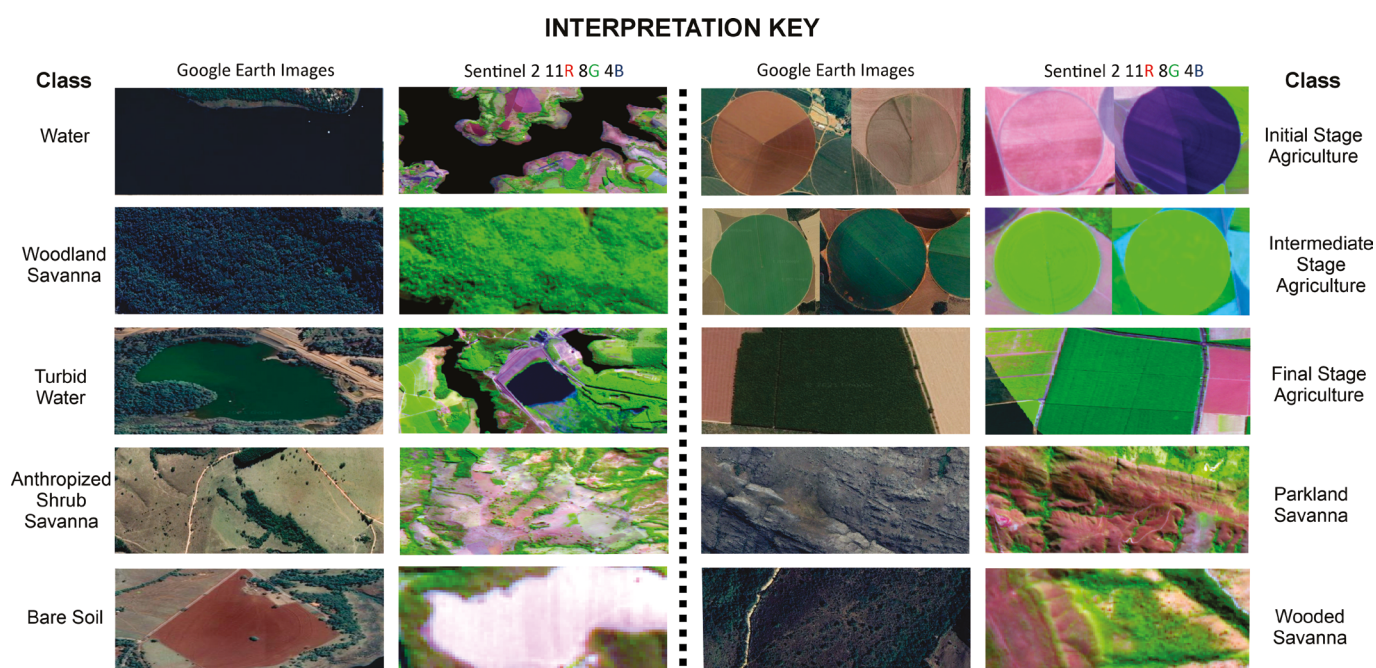
and S21W047. Slope and roughness images were generated directly from SRTM altimetry images. The methodology for generating the slope image uses a third-order finite difference equation (Horn, 1981). This technique considers a set of eight neighboring pixels relative to the central pixel, suitable for estimating the average slope on a 3x3 regular grid. The methodology for generating the roughness image employed the Terrain Ruggedness Index (TRI) proposed by Riley, DeGloria and Elliot (1999). This index is also calculated as the difference between the value of a grid cell and the average of the eight neighboring cells.

2.3 Data Cube Construction

The image cube was built using Sentinel-2 data (bands 2, 3, 4, 8, 11, and 12) and altimetry (SRTM), slope, and roughness. For Sentinel-2, bands 2, 3, and 4 correspond to the visible blue, green, and red, while the near-infrared (NIR) band corresponds to band 8. Bands 11 and 12 represent the shortwave infrared (SWIR). The 30-meter (1 arc second) SRTM image had to be resampled to 10 meters to integrate it with the visible and NIR Sentinel-2 bands using the Nearest Neighbor method. Similarly, the 20-meter resolution SWIR Sentinel-2 bands were resampled to 10 meters by the Nearest Neighbor method, using the Sentinel Application Platform (SNAP), available on the European Space Agency website (ESA).

2.4 Sample Selection for Training and Testing

A shapefile was created for training and validation samples, incorporating a new class field. Data were collected from high spatial and temporal resolution images (July 2019) via the Google Earth platform (Google, 2024). We used a random design to determine the sample distribution for training samples: 275 sample points were randomly generated for training and validation (Supplementary material 1). Using these points, polygons were delineated to represent each object and class for training and validation, established through spectral analysis in QGIS. Then, ten land use and land cover classes were defined: Water, Woodland Savanna, Turbid Water, Anthropized Shrub Savanna, Bare Soil, Initial Stage Agriculture, Intermediate Stage Agriculture, Final Stage Agriculture, Parkland Savanna, and Wooded Savanna (Figure 3). Urban centers were excluded due to pixel confusion and low kappa index performance (Congalton and Grenn, 2019). Urban centers were visually delineated using QGIS, creating a shapefile of polygons.



Source: ESA (2024); Google (2024).

Figure 3: Class Interpretation Key.

2.5 Classification

Six land use and land cover classification approaches were implemented using three machine learning algorithms applied to two datasets: [1] an ordinary multispectral Sentinel-2 stacked images, and [2] a data cube combining the multispectral Sentinel-2 images with terrain data, including elevation, slope, and roughness. The machine learning algorithms used were Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM).

Generally, a Decision Tree classifier is non-parametric, not requiring prior statistical assumptions about data distribution. The basic structure of a decision tree consists of a root node, multiple internal nodes, and a set of terminal nodes (Quinlan, 1992). The Random Forest (RF) classifier, a more sophisticated form of Bagging (Breiman, 2001), is a tree-structured classifier with added randomness. Methodologically, RF splits each node using the best among a randomly chosen subset of predictors at that node. A new training set is created from the original dataset with replacement, growing a tree from the random selection of features (Akar and Güngör, 2012). Support Vector Machines (SVM) are also non-parametric classifiers. Similarly to previous classifiers, the success of the SVM classification depends on how well it is trained. The easiest way to train this classifier is by using linearly separable classes. According to Osuana, Freund and Girosi (1997), if the training data with k samples represented as $\{X_i, Y_i\}$, where i ranges from 1 to k , X is a point in an N -dimensional space, and $Y \in \{-1, +1\}$ is a class label, these classes can be separable by a straight line if: there is a vector W orthogonal to the hyperplane (which determines the direction of the discriminant plane), and there is a scalar b that indicates the displacement of the discriminant hyperplane from the origin. Two hyperplanes can discriminate the data points in the respective classes.

Image classification was performed using the Orfeo ToolBox (Grizonnet et al, 2017) loaded in the QGIS platform, both open-source software. Before training each classifier, statistics such as global mean and standard deviation based on the combined variance of each band for both image sets were generated using OTB tools. Before training, all image features were normalized to the range $[-1, 1]$ using a linear rescaling process, which adjusts the pixel values from their original range (0 to 255) to fit within the new scale, preserving the relative differences between values.

After this step, the dataset was classified using the samples created for training and selected from analyses performed on Sentinel-2 images in false color composition. Then, the images were visually inspected using high-resolution Google Earth data.

Post-classification was also performed to standardize themes, removing isolated points classified differently from their surroundings, resulting in a classified image with reduced noise. We use Sieve Filtering to eliminate areas with less than two neighboring pixels. It is important to emphasize that the same parameters were used for both datasets, details of which are available in Supplementary Material 2. Outputs were generated for a) image and training statistics, b) image model, and c) confusion matrix, which were used to assess the overall accuracy through the kappa index, following the classification assessment method used in Manfré et al (2016).

The study aimed to assess whether incorporating a terrain model could enhance the performance of the CI classifier. It also compared the consistency of three classifiers using metrics such as the confusion matrix, kappa index, overall accuracy, and omission and commission errors. Confidence intervals ($\alpha = 0.01$) were created following the approach suggested by Olofsson et al (2014), with confusion matrices based on the pixel count for each validation polygon relative to the total mapped area. We evaluated each classifier's performance by analyzing the statistical differences in the kappa index, omission, and commission errors and whether terrain data improved or worsened the results. Additionally, we identified the best classifier for each class and evaluated the computational processing time to determine the most efficient approach overall.

3. Results

A total of 275 samples were collected for training the three algorithms, covering an area of 24.54 km² out of the total 15,358 km² of the study area, distributed among ten thematic classes (Table 1).

Table 1: Number of samples used in training the classifiers, number of pixels, and total area computed per class.

Training Samples	Sample polygons	Total area of samples (km ²)	Mean area of samples (km ²)
Water	13	5.44	0.418
Woodland Savanna	65	1.24	0.019
Turbid Water	6	0.77	0.128
Anthropized Shrub Savanna	18	3.75	0.208
Bare soil	33	1.23	0.037
Initial Stage Agriculture	44	2.21	0.05
Intermediate Stage Agriculture	32	1.52	0.048
Final Stage Agriculture	20	1.71	0.086
Parkland Savanna	26	4.53	0.174
Wooded Savanna	18	2.14	0.119
Total	275	24.54	0.089

Due to the relative similarity of the spectral responses and textures, the sampling was increased for Woodland Savanna to reduce confusion with Final Stage Agriculture. Similar problems and solutions occurred for Bare Soil and Initial Stage Agriculture. Water and Turbid Water classes needed fewer samples due to low representativeness and minimal confusion.

The total sample areas reflect the number of polygons per class. The total number of pixels per sample area for each class is evident in both datasets, whether in the Sentinel-2 image with six spectral bands or the data cube with 6 Sentinel-2 bands and three SRTM derivative ones. Notably, the number of pixels was increased by the number of bands in each cube.

3.1 Evaluating the Classifiers

The findings show how the classification using ordinary multispectral Sentinel-2 bands differs from stacking multispectral images with terrain data, which demonstrates improved results (Table 2). As expected, due to the maturity and robustness of the classifiers, the applied algorithms efficiently classified the remotely sensed digital images. The SVM achieved the best statistical indices, with a kappa of 0.91 and 0.89 for images with and without terrain data. Although DT showed inferior results, its processing time was only 40 minutes compared to SVM. RF had a time-consuming performance similar to DT but presented lower statistical indices. Using elevation, slope, and roughness data significantly improved results for all land use and land cover classification algorithms in addition to ordinary multispectral data.

Table 2: Comparison of kappa results and processing time obtained in the classifiers.

Kappa results and processing time for the classifiers	SVM	DT	RF
Sentinel 2	0.89±0.0018 / 48h	0.87±0.0019 / 40min	0.74±0.0025 / 50min
Sentinel 2 + terrain data	<i>0.91±0.0016+</i> / 48h	<i>0.89±0.0018+</i> / 40min	0.76±0.0025+ / 50min

± Indicate confidence interval to $\alpha = 0.01$; +Indicate improvement of algorithm by terrain data; Italic denotes best performance.

3.2 Comparison of Confusion Matrix Results for SVM, DT, and RF Algorithms with and without Terrain Data

The results of omission and commission errors based on the confusion matrix are presented in Tables 3 and 4. The complete confusion matrix highlighting the best and worst performances for Sentinel-2 multispectral images classified with and without SRTM-derivatives data is presented in Supplementary Material 3. The terrain model information improved algorithm performance through omission and commission errors.

Table 3: Omission errors (%) based on the confusion matrix class.

Land use and land cover Class	Omission error (%)					
	SVM	SVM+DEM	DT	DT+DEM	RF	RF+DEM
Water	0±0	0±0	0±0	0±0	0±0	0±0
Woodland Savanna	10.2±0.7	9.7±0.68	15.1±0.83	10±0.69 ⁺	4.1±0.46	1.4±0.27 ⁺
Turbid Water	0±0	0±0	0±0	0±0	0±0	0±0
Anthropized Shrub Savanna	13.3±0.45	11.9±0.43 ⁺	17±0.5	8.9±0.38 ⁺	17.3±0.5	8.2±0.36 ⁺
Bare soil	1.4±0.27	0.6±0.18 ⁺	1.16±0.25	0.7±0.19 ⁺	2±0.32	1.1±0.24 ⁺
Initial Stage Agriculture	8.8±0.49	9.9±0.52 ⁻	4.4±0.35	14.5±0.61 ⁻	65.2±0.82	51.9±0.86 ⁺
Intermediate Stage Agriculture	2.2±0.31	3.4±0.38 ⁻	3.7±0.39	0.7±0.17 ⁺	13.1±0.7	5±0.45 ⁺
Final Stage Agriculture	33.2±0.93	14±0.68 ⁺	38.8±0.96	51.5±0.98 ⁻	77.1±0.83	80.6±0.78 ⁻
Parkland Savanna	11±0.38	10.1±0.36 ⁺	15.1±0.43	7.3±0.31 ⁺	6.4±0.3	2.7±0.2 ⁺
Wooded Savanna	19.2±0.69	16.9±0.66 ⁺	19.2±0.69	13.6±0.6 ⁺	72.8±0.78	95.2±0.38 ⁻

± Indicate confidence interval to $\alpha = 0.01$; ⁺Indicate algorithm improvement by terrain data; ⁻Indicate worsened of algorithm by terrain data; Italic denotes best performance. SVM: Support Vector Machine; DT: Decision Tree; RF: Random Forest; DEM: digital elevation model derivate data.

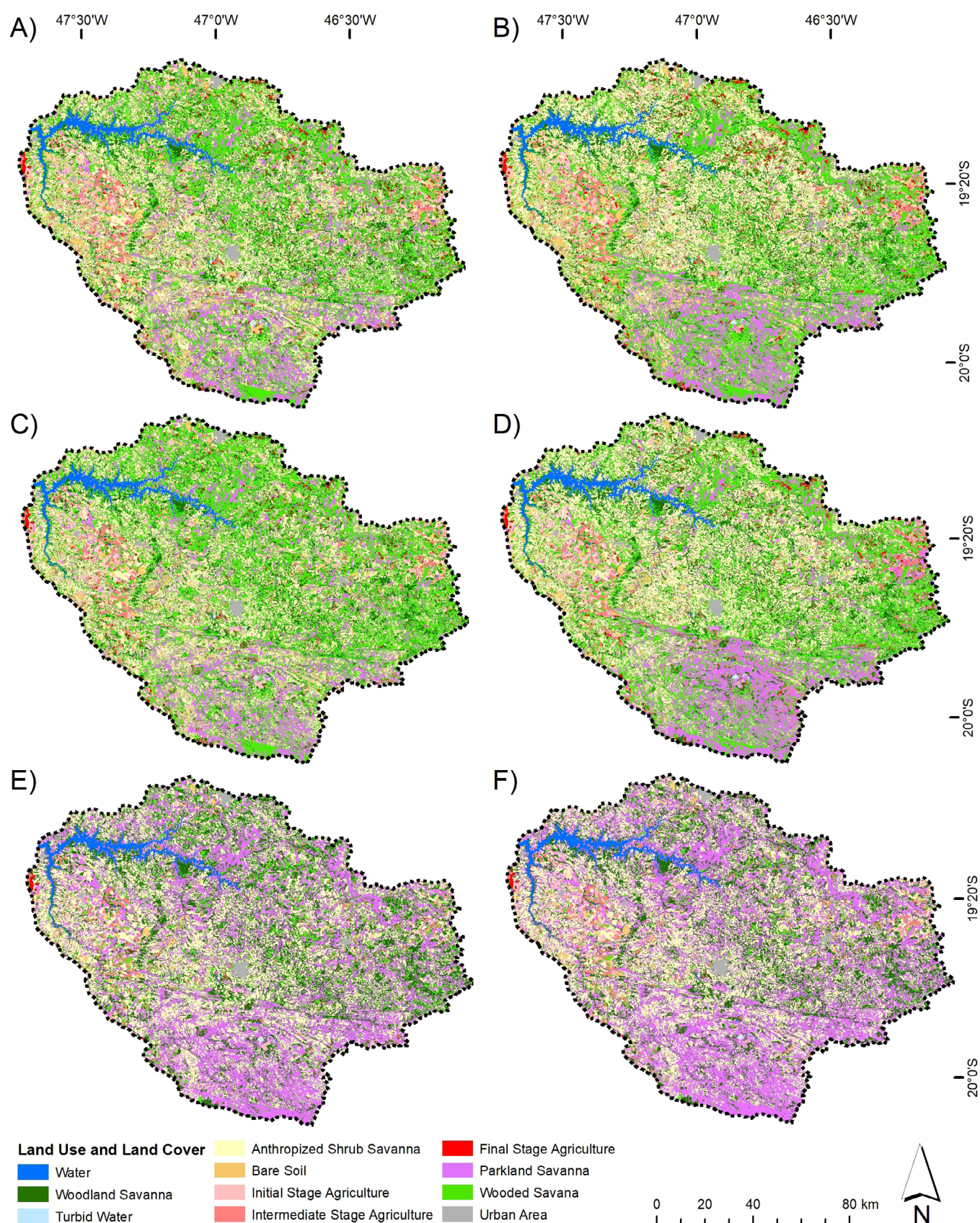
Table 4: Commission errors (%) based on the confusion matrix class.

Land use and land cover Class	Commission error (%)					
	SVM	SVM+DEM	DT	DT+DEM	RF	RF+DEM
Water	0±0	0±0	0±0	0±0	0±0	0±0
Woodland Savanna	36.5±0.93	22.6±0.89 ⁺	38.1±0.96	43.8±0.9 ⁻	58±0.75	57.7±0.75
Turbid Water	0±0	0±0	0±0	0±0	0±0	0±0
Anthropized Shrub Savanna	6.3±0.34	3.5±0.26 ⁺	6.4±0.35	1.19±0.15 ⁺	30.1±0.56	23.9±0.52 ⁺
Bare soil	18.3±0.81	19.1±0.82	15.4±0.77	11.9±0.71 ⁺	2.5±0.36	0±0 ⁺
Initial Stage Agriculture	0.5±0.13	1.1±0.19 ⁻	4.1±0.34	0.2±0.08 ⁺	16.2±0.99	1.8±0.33 ⁺
Intermediate Stage Agriculture	5±0.45	0.8±0.19 ⁺	2±0.29	0.02±0.03 ⁺	2.2±0.32	2.8±0.35
Final Stage Agriculture	3.5±0.43	2.1±0.3 ⁺	8.7±0.68	4±0.54 ⁺	0.4±0.26	0±0 ⁺
Parkland Savanna	7.8±0.33	7.2±0.32	8.4±0.35	8.8±0.34	32±0.48	33.9±0.47
Wooded Savanna	30±0.75	30.1±0.74	38.8±0.75	29.6±0.73 ⁺	37.1±1.29	24.2±3.01 ⁺

± Indicate confidence interval to $\alpha = 0.01$; ⁺Indicate algorithm improvement by terrain data; ⁻Indicate worsened of algorithm by terrain data; Italic denotes best performance. SVM: Support Vector Machine; DT: Decision Tree; RF: Random Forest; DEM: digital elevation model derivate data.

3.3 Land use and land cover maps

The mapping results are presented in Figure 4 (A, C, and E without terrain information; B, D, and F with terrain data). Overall, it is observed that models incorporating terrain data show more aggregated classes in plateau areas, particularly for the SVM (Figures 4A and 4B) and DT algorithms (Figures 4C and 4D). In contrast, the RF algorithm (Figures 4E and 4F) appears to produce a classification where some classes are scattered across the study area, consistent with the high confusion indices for natural vegetation classes observed in Table 4.



Source: Organized by the authors.

Figure 4: Land use and land cover maps: A) SVM; B) SVM+DEM; C) DT; D) DT+DEM; E) RF; F) RF+DEM. SVM: Support Vector Machine; DT: Decision Tree; RF: Random Forest; DEM: digital elevation model derivate data.

4. Discussion

As theorized, elevation, slope, and roughness data are essential for classification. These data are sources for geographic contextual information (Steinwart and Christmann, 2008), and adding them to an ordinary multispectral satellite image to compose a multisource data cube can improve the image classification model, further assisting algorithms with complementary data regarding class distribution. Analyzing and comparing image classification results using machine learning-based algorithms for the study area effectively identified the land cover and land use classes. Using the data cube with terrain derivative images such as elevation, slope, and roughness, in addition to multispectral bands and their derivatives, showed better results than the classification solely using ordinary multispectral images.

Our study area consists of spatially homogeneous landscape units (Martins et al, 2018), with steeper areas concentrated in plateaus to the south and north (near the municipalities of Tapira and Serra do Salitre, respectively) and a distinct flatter area to the west, near the municipalities of Santa Juliana and Pedrinópolis. Land use in the study area follows this pattern, with steeper areas covered by natural vegetation and flatter areas dominated by agricultural land. After incorporating terrain information, the SVM and DT algorithms highlighted this characteristic (see Figures 4B and 4D).

On the one hand, the SVM achieved the most satisfactory performance among the classifiers regarding thematic quality, while the computational cost was much higher than the other methods (see Table 2). This result can be explained by the classifier's good performance in working with a dataset with various attributes despite having few training samples (Grizonnet et al, 2017). Time-consuming of SVM is a consequence of the [1] algorithm complexity due to the quadratic optimization solver (Aryal, Sitaula and Frery, 2023), [2] training and parameter tuning (Hsu, Chang and Lin, 2016), and scalability (Steinwart and Christmann, 2008).

Al-Doski et al (2022) also found an 8.77% accuracy improvement in Landsat 8 image classification by adding NDVI and DEM data through this algorithm. In this case, the classifier that achieved equally satisfactory results but with greater computational efficiency was the DT. Thus, depending on the researcher's objective, considering processing time and the extent of the study area to be classified, this classifier is more suitable among all the evaluated ones.

Random Forest was the classifier that achieved the lowest results among the three analyzed, although satisfactory according to the statistical validation results, especially for natural cover and bare soil. Generally, comparing RF, SVM, and DT algorithms for satellite image classification reveals varied performances. RF demonstrated high accuracy in land use and land cover classification in studies using Sentinel-2 (Avci et al, 2023; Putri, 2023). In a tree species classification study, RF achieved an overall accuracy of 86% within the sample range, outperforming SVM and CART (Pantoja, Spenassato and Emmendorfer, 2023). A comparison between the Gaussian Mixture Model and RF also highlighted RF's superior efficiency in land cover monitoring (Yasaswini and Reddy, 2023). However, none of these studies incorporated altimetric data; therefore, our results introduce a sparsely investigated approach to land use and land cover classification by integrating remote sensing images with terrain-derived data, such as digital elevation models.

In this context, for all classifiers, using a composition of images that enhances their data with terrain information for generating land use and land cover classification results in better discrimination of classes and improved accuracy, as demonstrated by the employed validation data. Generally, topography influences mapping accuracy (Sang et al, 2021) and land use and land cover patterns (Zhang et al, 2019; Yang et al, 2022).

These results can also be used in future integrated works with other optical indexes in the dataset, such as a) the Normalized Difference Vegetation Index (NDVI), b) the Soil-Adjusted Vegetation Index (SAVI), and c) the Normalized Difference Water Index (NDWI), previously mentioned as alternatives to increase final classification

accuracy. Recent works have obtained satisfactory results (Zhao et al, 2022; Bueno et al, 2023), highlighting the importance of incorporating multiple layers in data cubes for further research.

It is also important to note that this study faced limitations due to the lack of control points and field data collection, primarily caused by social isolation during the COVID-19 pandemic, which partially compromised the validation of the classification results. Nevertheless, it is not uncommon to use high-resolution images to create test and validation samples in remote sensing studies, especially Google Earth images (Zhao et al, 2014) or synthetic image composites (Macedo et al, 2014). Recommendations for future work include comparing classifications in other geographic contexts to assess and compare the results. Also, the approaches can be computed and compared using graphic processing units to verify computer performance. Another suggestion for future research is to classify these algorithms on different satellite images with varying spatial resolutions (pixel dimensions), both finer and coarser, to evaluate whether the statistical validation results improve or deteriorate.

5. Conclusions

The results underscore the importance of integrating terrain data into satellite image classification for the Cerrado biome, improving the classification performance. The SVM algorithm proved the most effective, even with higher computational costs, followed by DT, offering a faster processing time alternative. The research also emphasizes the need for future studies to incorporate field data to validate results, explore different spatial resolutions of satellite images, and provide a more detailed analysis of how each terrain variable can influence the classification results. In summary, combining terrain data with machine learning algorithms promises to enhance land use and land cover classification accuracy in the Cerrado biome, contributing to environmental management and biodiversity conservation.

ACKNOWLEDGMENT

We thank Pesquisa & Desenvolvimento Agência Nacional de Energia Elétrica / Companhia Energética de Minas Gerais (P&D Aneel-Cemig GT-599), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES; finance code 001) and Conselho Nacional de Desenvolvimento Científico (CNPq; PQ-311002/2023-4 to DRM and PQ-315631/2021-0 to RAAN).

AUTHOR'S CONTRIBUTION

Carlos Eduardo Holanda: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization.

Diego Rodrigues Macedo: Conceptualization, Formal analysis, Funding acquisition, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Rodrigo Affonso Albuquerque Nóbrega: Conceptualization, Formal analysis, Supervision, Writing – review & editing.

REFERENCES

- Adam, E. et al. (2014) Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers, *International Journal of Remote Sensing*, 35(10), 3440–3458. <https://doi.org/10.1080/01431161.2014.903435>
- Akar, Ö. and Güngör, O. (2012) Classification of multispectral images using Random Forest algorithm, *Journal of Geodesy and Geoinformation*, 1(2), 105–112. <https://doi.org/10.9733/jgg.241212.1>
- Al-Doski, J. et al. (2022) Incorporation of digital elevation model, normalized difference vegetation index, and Landsat-8 data for land use land cover mapping, *Photogrammetric Engineering and Remote Sensing*, 88(8), 507–515. <https://doi.org/10.14358/PERS.21-00082R2>
- Aryal, J., Sitaula, C. and Frery, A.C. (2023) Land use and land cover (LULC) performance modeling using machine learning algorithms: a case study of the city of Melbourne, Australia, *Scientific Reports*, 13(1), 13510. <https://doi.org/10.1038/s41598-023-40564-0>
- Avci, C. et al. (2023) Comparison between random forest and support vector machine algorithms for LULC classification, *International Journal of Engineering and Geosciences*, 8(1), 1–10. <https://doi.org/10.26833/ijeg.987605>
- Blaschke, T. (2010) Object-based image analysis for remote sensing, *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1), 2–16. <https://doi.org/10.1016/j.isprsjprs.2009.06.004>
- Breiman, L. (2001) Random Forests, *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bueno, I.T. et al. (2023) Land use/land cover classification in a heterogeneous agricultural landscape using Planetscope data, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 48(M-1–2023), 49–55. <https://doi.org/10.5194/isprs-archives-XLVIII-M-1-2023-49-2023>
- Callisto, M. et al. (2019) Multi-status and multi-spatial scale assessment of landscape effects on benthic macroinvertebrates in the Neotropical Savanna, in Hughes, R.M. et al. (eds) *Advances in Understanding Landscape Influences on Freshwater Habitats and Biological Assemblages*. Bethesda, MD: American Fisheries Society Symposium 90, pp. 275–302. <https://doi.org/10.47886/9781934874561.ch14>
- Chakraborty, S. (2021) Remote Sensing and GIS in Environmental Management, in *Environmental Management: Issues and Concerns in Developing Countries*. Cham: Springer International Publishing, pp. 185–220. https://doi.org/10.1007/978-3-030-62529-0_10
- Congalton, R.G. and Grenn, K. (2019) *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. 3rd ed. Boca Raton, FL: CRC Press. <https://doi.org/10.1201/9780429052729>
- ESA - European Spatial Agency (2020) *Copernicus Sentinel Missions*. Paris: ESA. Available at: <https://sentinel.esa.int/web/sentinel/missions> [Accessed 10 November 2024].
- Francisco, C.N. and Almeida, C.M. de (2012) Avaliação de desempenho de atributos estatísticos e texturais em uma classificação de cobertura da terra baseada em objeto, *Boletim de Ciências Geodésicas*, 18(2), 302–326. <https://doi.org/10.1590/S1982-21702012000200008>
- Google (2024) *Google Earth*. Mountain View, CA: Google, Inc. Available at: <https://www.google.com.br/earth/> [Accessed 10 November 2024].
- Grizonnet, M. et al. (2017) Orfeo ToolBox: open source processing of remote sensing images, *Open Geospatial Data, Software and Standards*, 2(1), 15. <https://doi.org/10.1186/s40965-017-0031-6>
- Hansen, M.C. et al. (2008) Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data, *Proceedings of the National Academy of Sciences*, 105(27), 9439–9444. <https://doi.org/10.1073/pnas.0804042105>

- Horn, B.K.P. (1981) Hill shading and the reflectance map, *Proceedings of the IEEE*, 69(1), 14–47. <https://doi.org/10.1109/PROC.1981.11918>
- Hsu, C.W., Chang, C.C. and Lin, C.J. (2016) *A practical guide to support vector classification*. Taipei, TW: National Taiwan University, Department of Computer Science. Available at: <https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf> [Accessed 10 November 2024].
- IBGE - Instituto Brasileiro de Geografia e Estatística (2024) *Catálogo de Geosserviços, Infra Estrutura Nacional de Dados Espaciais*. Rio de Janeiro, Brazil: IBGE. Available at: <https://geoservicos.ibge.gov.br/geoserver/web/> [Accessed 10 November 2024].
- Klink, C.A. and Machado, R.B. (2005) Conservation of the Brazilian Cerrado, *Conservation Biology*, 19(3), 707–713. <https://doi.org/10.1111/j.1523-1739.2005.00702.x>
- Li, T. et al. (2024). Quantifying impacts of livestock production on ecosystem services: Insights into grazing management under vegetation restoration, *Journal of Cleaner Production*, 470, 143359. <https://doi.org/10.1016/j.jclepro.2024.143359>
- Macedo, D.R. et al. (2014) Sampling site selection, land use and cover, field reconnaissance, and sampling, in Callisto, M. et al. (eds) *Ecological conditions in hydropower basins*. Serie Peixe Vivo 3. Belo Horizonte, Brazil: Companhia Energética de Minas Gerais, pp. 61–83. <https://doi.org/10.5281/zenodo.2648039>
- Manfré, L.A., Nobrega, R.A.A. and Quintanilha, J.A. (2016) Evaluation of multiple classifier systems for landslide identification in LANDSAT Thematic Mapper (TM) Images. *ISPRS International Journal of Geo-Information*, 5(9), 164. <https://doi.org/10.3390/ijgi5090164>
- Mapbiomas (2024) *Coleções Mapbiomas: Coleção 8 (1985-2022) da Série Anual de Mapas de Uso e Cobertura da Terra do Brasil*. São Paulo, Brazil: Projeto Mapbiomas. Available at: <http://mapbiomas.org> [Accessed 10 November 2024].
- Martins, I. et al. (2018) Regionalisation is key to establishing reference conditions for neotropical savanna streams, *Marine and Freshwater Research*, 69(1), 82–94. <https://doi.org/10.1071/MF16381>
- Mashala, M.J. et al. (2023) A systematic review on advancements in remote sensing for assessing and monitoring land use and land cover changes impacts on surface water resources in semi-arid tropical environments, *Remote Sensing*, 15(16), 3926. <https://doi.org/10.3390/rs15163926>
- Mather, P.M. and Koch, M. (2011) *Computer Processing of Remotely-Sensed Images*. 4th ed. Chichester, UK: Wiley-Blackwell. <https://doi.org/10.1002/9780470666517>
- Moreira, E.P. et al. (2016) Efeito topográfico sobre índices de vegetação obtidos com dados Landsat TM: É necessário correção topográfica?, *Boletim de Ciencias Geodesicas*, 22(1), 95–107. <https://doi.org/10.1590/S1982-21702016000100006>
- Myers, N. et al. (2000) Biodiversity hotspots for conservation priorities, *Nature*, 403(6772), 853–8. <https://doi.org/10.1038/35002501>
- Olofsson, P. et al. (2014) Good practices for estimating area and assessing accuracy of land change, *Remote Sensing of Environment*, 148, 42–57. <https://doi.org/10.1016/j.rse.2014.02.015>
- Osuana, E.E., Freund, R. and Girosi, F. (1997) *Support Vector Machines: Training and Applications*. Cambridge, MA: Massachusetts Institute of Technology.
- Pantoja, D.A., Spenassato, D. and Emmendorfer, L.R. (2023) Comparison between classification algorithms: Gaussian Mixture Model - GMM and Random Forest - RF, for Landsat 8 images, *Revista de Gestão Social e Ambiental*, 16(3), e03234. <https://doi.org/10.24857/rgsa.v16n3-015>
- Putri, K.A. (2023) Analysis of Land Cover Classification Results Using ANN, SVM, and RF Methods with R Programming Language (Case Research: Surabaya, Indonesia), *IOP Conference Series: Earth and Environmental Science*, 1127(1), 012030. <https://doi.org/10.1088/1755-1315/1127/1/012030>

- Quinlan, J.R. (1992) *C4.5 Programs for machine learning*. Santa Mateo, CA: Morgan Kaufmann Publishers.
- Riley, S.J., DeGloria, S.D. and Elliot, R. (1999) A terrain ruggedness index that quantifies topographic heterogeneity, *Intermountain Journal of Sciences*, 5(1–4), 23–27.
- Rodrigues, A.A. et al. (2022) Cerrado deforestation threatens regional climate and water availability for agriculture and ecosystems, *Global Change Biology*, 28(22), 6807–6822. <https://doi.org/10.1111/gcb.16386>
- Sang, X. et al. (2021) The effect of DEM on the land use/cover classification accuracy of Landsat OLI images, *Journal of the Indian Society of Remote Sensing*, 49(7), 1507–1518. <https://doi.org/10.1007/s12524-021-01318-5>
- Santos, G.D. dos, Francisco, C.N. and Almeida, C.M. de (2015) Mineração de dados aplicada à discriminação da cobertura da terra em imagens Landsat 8 OLI, *Boletim de Ciências Geodésicas*, 21(4), 706–720. <https://doi.org/10.1590/S1982-21702015000400041>
- Steel, E.A. et al. (2010) Are we meeting the challenges of landscape-scale riverine research? A review, *Living Reviews in Landscape Research*, 4(1), 1–60. <https://doi.org/10.12942/lrlr-2010-1>
- Steinwart, I. and Christmann, A. (2008) *Support Vector Machines*. New York, NY: Springer New York. <https://doi.org/10.1007/978-0-387-77242-4>
- Strassburg, B.B.N. et al. (2017) Moment of truth for the Cerrado hotspot, *Nature Ecology & Evolution*, 1(4), 0099. <https://doi.org/10.1038/s41559-017-0099>
- Tian, T. et al. (2024) Mapping the time-series of essential urban land use categories in China: a multi-source data integration approach, *Remote Sensing*, 16(17), 3125. <https://doi.org/10.3390/rs16173125>
- USGS - United States Geological Survey (2024) *Earth Explorer*. Washington, DC: USGS. Available at: <https://earthexplorer.usgs.gov/> [Accessed 10 November 2024].
- Yang, Z. et al. (2022) The impact of topographic relief on population and economy in the Southern Anhui Mountainous area, China, *Sustainability*, 14(21), 14332. <https://doi.org/10.3390/su142114332>
- Yasaswini, C.S. and Reddy, S.N. (2023) Satellite image classification using extended Local Binary Patterns, SVM and CNN, *International Journal of Scientific Research in Science and Technology*, 10(3), 775–784. <https://doi.org/10.32628/IJSRST523103138>
- Zhang, J. et al. (2019) Topographical relief characteristics and its impact on population and economy: A case study of the mountainous area in western Henan, China, *Journal of Geographical Sciences*, 29(4), 598–612. <https://doi.org/10.1007/s11442-019-1617-y>
- Zhao, J. et al. (2022) A land cover classification method for high-resolution remote sensing images based on NDVI deep learning fusion network, *Remote Sensing*, 14(21). <https://doi.org/10.3390/rs14215455>
- Zhao, Yuanyuan et al. (2014) Towards a common validation sample set for global land-cover mapping, *International Journal of Remote Sensing*, 35(13), 4795–4814. <https://doi.org/10.1080/01431161.2014.930202>
- Zhong, Y. et al. (2018) Computational intelligence in optical remote sensing image processing, *Applied Soft Computing*, 64, 75–93. <https://doi.org/10.1016/j.asoc.2017.11.045>

Supplementary Material 1

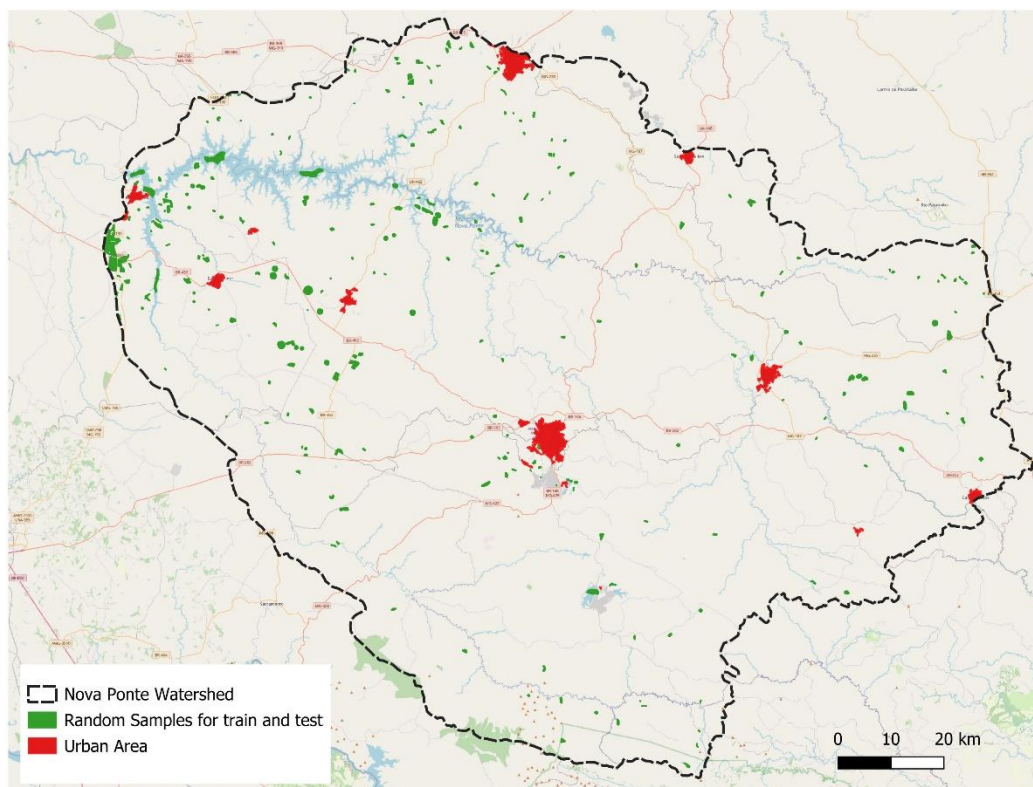


Figure S1.1: Randomly sampled areas to train and test the algorithms.

Supplementary Material 2

Table S2.1: Parameters for training images.

Parameters for training images	Maximum size of training samples	Maximum size of validation samples per class	Number of samples bound by the minimum	Training and validation sample rate
Sentinel 2	-1	-1	1	0.5
Sentinel 2 + Terrain data	-1	-1		0.5

Where:

- Maximum training sample size per class default value: 1000. Maximum size per class (in pixels) of the training sample list (default = 1000) (no limit = -1). If equal to -1, then the maximum size of the training sample list available per class will equal the surface area of the smallest class multiplied by the aspect ratio of the training sample.
- Maximum validation sample size per class default value: 1000. Maximum size per class (in pixels) of the validation sample list (default = 1000) (no limit = -1). If equal to -1, then the maximum size of the available validation sample list per class will equal the surface area of the smallest class multiplied by the aspect ratio of the validation sample.
- Number of samples limited by minimum default value: 1. Limit the number of samples for each class by the number of samples available for the smallest class. The ratios between training and validation are respected. The default is true (= 1).
- Training and validation sample ratio default value: 0.5. Ratio between training and validation samples (0.0 = all training, 1.0 = all validation) (default = 0.5).

Table S2.2: Parameters for post-classification of images.

Parameters for Post-classification of images	Radius of structuring element (in pixels)	Label for the class with no data	Label for the undecided class	Threshold for isolated pixels default value
Sentinel 2	4	0	0	2
Sentinel 2 + Terrain data	4	0	0	2

Where:

- Radius of the structuring element (in pixels) default value: 1. The radius of the ball-shaped structuring element (in pixels).
- Label for the class with no data default value: 0. Label for the class without data. These input pixels keep their label without data in the output image.
- Label for the class Undecided default value: 0. Label for the class Undecided.
- Threshold for isolated pixels default value: 1. Maximum number of neighbors with the same label as the central pixel to consider that it is an isolated pixel.

Table S2.3: Parameters for the Random Forest (RF) algorithm.

Parameters for Random Forest (RF)	Maximum depth of the tree	Minimum number of classes at each node	Termination criteria for regression tree	Group possible values of a variable category into $k \leq \text{cat}$	Size of randomly selected feature subset at each node of the tree	Maximum number of trees in the forest	Sufficient accuracy (OOB error)
Sentinel 2	5	10	0	10	0	100	0.01
Sentinel 2 + Terrain data	5	10	0	10	0	100	0.01

Where:

- Maximum depth of tree default value: 5. The depth of the tree. A low value is likely to be underfitting; conversely, a high value is likely to be overfitting. The optimal value can be obtained using cross-validation or other suitable methods.
- Minimum number of samples at each node default value: 10. If the number of samples at a node is less than this parameter, the node is not split. A reasonable value is a small percentage of the total data, for example, 1 percent.
- Termination criteria for regression tree default value: 0. If all the absolute differences between an estimated node value and the training samples' values are less than this regression accuracy parameter, the node is not split.
- Group the possible values of a categorical variable into $K \leq \text{cat}$ clusters to find a suboptimal split default value: 10. Group the possible values into $K \leq \text{cat}$ clusters to find a suboptimal split.
- • Size of the randomly selected feature subset at each node in the tree default value: 0. The randomly selected feature subset size at each node in the tree is used to find the best splits. If you set this to 0, the size will be set to the square root of the total number of features.
- Maximum number of trees in the forest default value: 100. The maximum number of trees in the forest. Typically, the more trees you have, the better the accuracy. However, the improvement in accuracy usually slows down and reaches an asymptote at a certain number of trees. Also to keep in mind, increasing the number of trees increases the prediction time linearly.
- Sufficient accuracy (OOB error) default value: 0.01. Sufficient accuracy (OOB error).

Table S2.4: Parameters for the Decision Tree (DT) algorithm.

Parameters for Decision Tree (DT)	Maximum depth of the tree	Minimum number of samples in each node	Termination criteria for regression tree	Group possible values of a variable category into $k \leq \text{cat}$
Sentinel 2	10	10	0.01	10
Sentinel 2 + Terrain data	0	10	0.01	10

Where:

- Maximum depth of tree default value: 10. The training algorithm attempts to split each node while its depth is less than the maximum possible depth of the tree. The actual depth may be less if the other

termination criteria are met and/or if the tree is pruned.

- Minimum number of samples in each node default value: 10. If the number of samples in a node is less than this parameter, then this node will not be split.
- Termination criteria for regression tree default value: 0.01. Suppose all the absolute differences between an estimated node value and the training samples' values at this node are less than this regression accuracy parameter. In that case, the node will not be split any further.
- Group the possible values of a categorical variable into $K \leq \text{cat}$ clusters to find a suboptimal split default value: 10. Group the possible values into $K \leq \text{cat}$ clusters to find a suboptimal split.

Table S2.5: Parameters for the Support Vector Machine (LibSVM) algorithm.

Parameters for Support Vector Machines (LibSVM)	Kernel type for SVM	SVM model type	Cost parameter C	Cost parameter Nu
Sentinel 2	Linear	Classification of support vectors C	1	0.5
Sentinel 2 + Terrain data	Linear	Classification of support vectors C	1	0.5

Where:

- Kernel type for SVM: Linear Kernel, no mapping is done; this is the fastest option.
- SVM model type: This formulation allows imperfect class separation. The penalty is set via the cost parameter C.
- Cost parameter C Default value: 1. SVM models have a cost parameter C (1 by default) to control the tradeoff between training errors and forcing hard margins.
- Cost parameter Nu: Default value: 0.5. Cost parameter Nu, in the range 0..1, the higher the value, the smoother the decision.

Supplementary Material 3

Table S3.1: Confusion matrix (pixels), omission and commission error, overall accuracy and kappa index for Support Vector Machine (SVM) using Sentinel image.

	Water	Woodland Savanna	Turbid Water	Anthropized Shrub Savanna	Bare soil	Initial Stage Agriculture	Intermediate Stage Agriculture	Final Stage Agriculture	Parkland Savanna	Wooded Savanna	Total
Water	54480	0	0	0	0	0	0	0	0	0	54480
Woodland Savanna	0	11184	0	1	0	0	188	407	1	683	12464
Turbid Water	0	0	7728	0	0	0	0	0	0	0	7728
Anthropized Shrub Savanna	0	21	0	32510	832	76	119	3	577	3392	37530
Bare soil	0	0	0	173	12151	0	0	0	0	0	12324
Initial Stage Agriculture	0	0	0	4	1955	20213	0	0	2	0	22174
Intermediate Stage Agriculture	0	329	0	0	0	0	14911	1	0	12	15253
Final Stage Agriculture	0	5180	0	0	0	0	58	11440	0	462	17140
Parkland Savanna	0	3	0	1676	0	34	429	0	40374	2874	45390
Wooded Savanna	0	898	0	366	0	0	1	16	2848	17281	21410
Total	54480	17615	7728	34730	14938	20323	15706	11867	43802	24704	245893
Comission error	0	0.365	0	0.063	0.183	0.005	0.05	0.035	0.078	0.3	
Omission error	0	0.102	0	0.133	0.014	0.088	0.022	0.332	0.11	0.192	
Global exaction (Po)	0.9										
Kappa index	0.89										

Table S3.2: Confusion matrix (pixels), omission and commission error, overall accuracy and kappa index for Support Vector Machine (SVM) using Sentinel image and terrain data.

	Water	Woodland Savanna	Turbid Water	Anthropized Shrub Savanna	Bare soil	Initial Stage Agriculture	Intermediate Stage Agriculture	Final Stage Agriculture	Parkland Savanna	Wooded Savanna	Total
Water	54480	0	0	0	0	0	0	0	0	0	54480
Woodland Savanna	0	11252	0	0	0	0	110	326	0	776	12464
Turbid Water	0	0	7728	0	0	0	0	0	0	0	7728
Anthropized Shrub Savanna	0	21	0	33029	714	38	15	0	380	3333	37530
Bare soil	0	0	0	85	12239	0	0	0	0	0	12324
Initial Stage Agriculture	0	0	0	1	2181	19964	0	0	28	0	22174
Intermediate Stage Agriculture	0	318	0	0	0	0	14933	2	0	0	15253
Final Stage Agriculture	0	2182	0	0	0	0	0	14733	0	225	17140
Parkland Savanna	0	0	0	1051	0	199	0	0	40796	3344	45390
Wooded Savanna	0	774	0	82	0	0	0	1	2762	17791	21410
Total	54480	14547	7728	34248	15134	20201	15058	15062	43966	25469	245893
Comission error	0	0.226	0	0.035	0.191	0.011	0.008	0.021	0.072	0.301	
Omission error	0	0.097	0	0.119	0.006	0.099	0.034	0.14	0.101	0.169	
Global exaction (Po)	0.92										
Kappa index	0.91										

Table S3.3: Confusion matrix (pixels), omission and commission error, overall accuracy and kappa index for Decision Tree (DT) using Sentinel image.

	Water	Woodland Savanna	Turbid Water	Anthropized Shrub Savanna	Bare soil	Initial Stage Agriculture	Intermediate Stage Agriculture	Final Stage Agriculture	Parkland Savanna	Wooded Savanna	Total
Water	54480	0	0	0	0	0	0	0	0	0	54480
Woodland Savanna	0	10581	0	1	0	0	140	886	1	855	12464
Turbid Water	0	0	7728	0	0	0	0	0	0	0	7728
Anthropized Shrub Savanna	0	38	0	31113	1082	184	98	13	611	4391	37530
Bare soil	0	0	0	144	12180	0	0	0	0	0	12324
Initial Stage Agriculture	0	0	0	76	901	21194	0	0	0	3	22174
Intermediate Stage Agriculture	0	363	0	19	0	0	14688	1	0	182	15253
Final Stage Agriculture	0	5405	0	0	0	0	72	10487	0	1176	17140
Parkland Savanna	0	3	0	1552	236	724	0	0	38504	4371	45390
Wooded Savanna	0	722	0	363	7	3	5	108	2920	17282	21410
Total	54480	17112	7728	33268	14406	22105	15003	11495	42036	28260	245893
Comission error	0	0.381	0	0.064	0.154	0.041	0.02	0.087	0.084	0.388	
Omission error	0	0.151	0	0.17	0.0116	0.044	0.037	0.388	0.151	0.192	
Global exaction (Po)	0.89										
Kappa index	0.87										

Table S3.4: Confusion matrix (pixels), omission and commission error, overall accuracy and kappa index for Decision Tree (DT) using Sentinel image and terrain data.

	Water	Woodland Savanna	Turbid Water	Anthropized Shrub Savanna	Bare soil	Initial Stage Agriculture	Intermediate Stage Agriculture	Final Stage Agriculture	Parkland Savanna	Wooded Savanna	Total
Water	54480	0	0	0	0	0	0	0	0	0	54480
Woodland Savanna	0	11212	0	10	0	0	0	347	0	895	12464
Turbid Water	0	0	7728	0	0	0	0	0	0	0	7728
Anthropized Shrub Savanna	0	0	0	34178	0	23	0	0	0	3329	37530
Bare soil	0	0	0	94	12230	0	0	0	0	0	12324
Initial Stage Agriculture	0	0	0	110	1666	18949	0	0	1449	0	22174
Intermediate Stage Agriculture	0	95	0	0	0	0	15135	0	0	23	15253
Final Stage Agriculture	0	8366	0	0	0	0	4	8307	0	463	17140
Parkland Savanna	0	0	0	199	0	30	0	1	42071	3089	45390
Wooded Savanna	0	305	0	0	0	0	0	0	2617	18488	21410
Total	54480	19978	7728	34591	13896	19002	15139	8655	46137	26287	245893
Comission error	0	0.438	0	0.0119	0.119	0.002	0.0002	0.04	0.088	0.296	
Omission error	0	0.1	0	0.089	0.007	0.145	0.007	0.515	0.073	0.136	
Global exaction (Po)	0.91										
Kappa index	0.89										

Table S3.5: Confusion matrix (pixels), omission and commission error, overall accuracy and kappa index for Random Forest (RF) using Sentinel image.

	Water	Woodland Savanna	Turbid Water	Anthropized Shrub Savanna	Bare soil	Initial Stage Agriculture	Intermediate Stage Agriculture	Final Stage Agriculture	Parkland Savanna	Wooded Savanna	Total
Water	54479	1	0	0	0	0	0	0	0	0	54480
Woodland Savanna	0	11952	0	0	0	0	276	18	23	195	12464
Turbid Water	0	0	7728	0	0	0	0	0	0	0	7728
Anthropized Shrub Savanna	0	101	0	31018	91	233	12	0	4642	1433	37530
Bare soil	0	0	0	254	12070	0	0	0	0	0	12324
Initial Stage Agriculture	0	448	0	10876	220	7707	1	0	2922	0	22174
Intermediate Stage Agriculture	0	223	0	4	0	0	13252	0	0	1774	15253
Final Stage Agriculture	0	13207	0	0	0	0	0	3924	1	8	17140
Parkland Savanna	0	50	0	1629	0	1232	0	0	42449	30	45390
Wooded Savanna	0	2514	0	617	0	29	16	0	12411	5823	21410
Total	54479	28496	7728	44398	12381	9201	13557	3942	62448	9263	245893
Comission error	0	0.58	0	0.301	0.025	0.162	0.022	0.004	0.32	0.371	
Omission error	0	0.041	0	0.173	0.02	0.652	0.131	0.771	0.064	0.728	
Global exaction (Po)	0.77										
Kappa index	0.74										

Table S3.6: Confusion matrix (pixels), omission and commission error, overall accuracy and kappa index for Random Forest (RF) using Sentinel image and terrain data.

	Water	Woodland Savanna	Turbid Water	Anthropized Shrub Savanna	Bare soil	Initial Stage Agriculture	Intermediate Stage Agriculture	Final Stage Agriculture	Parkland Savanna	Wooded Savanna	Total
Water	54480	0	0	0	0	0	0	0	0	0	54480
Woodland Savanna	0	12278	0	22	0	0	0	0	135	29	12464
Turbid Water	0	0	7728	0	0	0	0	0	0	0	7728
Anthropized Shrub Savanna	0	0	0	34416	0	50	0	0	2992	72	37530
Bare soil	0	0	0	142	12182	0	0	0	0	0	12324
Initial Stage Agriculture	0	0	0	9548	9	10646	428	0	1543	0	22174
Intermediate Stage Agriculture	0	552	0	0	0	0	14478	0	0	223	15253
Final Stage Agriculture	0	13817	0	3	0	0	0	3317	1	2	17140
Parkland Savanna	0	0	0	1113	0	156	0	0	44121	0	45390
Wooded Savanna	0	2419	0	0	0	0	0	0	17974	1017	21410
Total	54480	29066	7728	45244	12191	10852	14906	3317	66766	1343	245893
Comission error	0	0.577	0	0.239	0	0.018	0.028	0	0.339	0.242	
Omission error	0	0.014	0	0.082	0.011	0.519	0.05	0.806	0.027	0.952	
Global exaction (Po)	0.79										
Kappa index	0.76										

Table S3.7: Comparison of the results of the confusion matrices of the SVM, DT and RF algorithms with and without terrain data (DEM). The algorithm that presented the best result based on omission errors for each class is marked in green, while the one that presented the greatest confusion based on commission errors between the classes is in orange.

	Water	Woodland savanna	Turbid water	Anthropized shrub Savanna	Bare soil	Beginning agriculture	Intermediate agriculture	Advanced agriculture	Parkland Savanna	Wooded Savanna
Water	SVM without/with DEM; DT without/with DEM; RF without DEM	RF without DEM								
Woodland savanna		RF with DEM		RF with DEM			RF without DEM	DT without DEM	RF with DEM	DT with DEM
Turbid water			SVM without/with DEM; DT without/with DEM; RF without DEM							
Anthropized shrub Savanna		RF without DEM		RF with DEM	DT without DEM	RF without DEM	SVM without DEM	DT without DEM	RF without DEM	DT without DEM
Bare soil				RF without DEM	RF with DEM					
Beginning agriculture		RF without DEM		RF without DEM	SVM with DEM	DT without DEM	RF with DEM		RF without DEM	DT without DEM
Intermediate agriculture		RF with DEM		DT without DEM			DT with DEM	SVM with DEM		RF without DEM
Advanced agriculture		RF with DEM		RF with DEM			DT without DEM	SVM with DEM	RF without/with DEM	DT without DEM
Parkland Savanna		RF without DEM		SVM without DEM	DT without DEM	RF without DEM	SVM without DEM	DT with DEM	RF with DEM	DT without DEM
Wooded Savanna		RF without DEM		RF with DEM	DT without DEM	RF with DEM	RF with DEM	DT without DEM	RF with DEM	DT with DEM