Journal of South American Earth Sciences 153 (2025) 105366

Contents lists available at ScienceDirect

Journal of South American Earth Sciences

journal homepage: www.elsevier.com/locate/jsames

Prediction of forest fire susceptibility using machine learning tools in the Triunfo do Xingu Environmental Protection Area, Amazon, Brazil

Kemuel Maciel Freitas [a,*], Ronie Silva Juvanhol [b], Christiano Jorge Gomes Pinheiro [c], Anderson Alvarenga de Moura Meneses [d]

[a] Federal University of Western Pará/UFOPA, Institute of Geosciences and Engineering, Graduate Program in Amazon Natural Resources, Vera Paz St., Salé, s/n, 68000-000, Santarém, PA, Brazil
[b] Federal University of Piaui/UFPI, BR 135, Km 03, Planalto Horizonte, 64900-000, Bom Jesus, PI, Brazil
[c] Federal University of Espírito Santo/UFES, Department of Rural Engineering, Alto Universitário, s/n, 29500-000, Alegre, ES, Brazil
[d] Federal University of Western Pará, Institute of Geosciences and Engineering, Laboratory of Computational Intelligence, Vera Paz St., Salé, 68000-000, Santarém, PA, Brazil

ARTICLE INFO

Keywords:
Remote sensing
Amazon
Forest fire
Prediction
Random forest
XGBoost

ABSTRACT

Machine learning tools have demonstrated promising results for fire prediction, which have included the generation of models that have been developed across a large range of contexts and locations. This research aims to map areas susceptible to forest fires within the Triunfo do Xingu Environmental Protection Area, employing machine learning algorithms to ascertain the influence of environmental, topographic and socioeconomic factors on fire occurrence. For this purpose, the Random Forest and Extreme Gradient Boosting regression models were used to predict kernel density values calculated over 15,291 confirmed burn points between 2010 and 2020, using 11 predictor factors, including Altitude, Slope, Aspect, Topographic Wetness Index, Precipitation, Temperature, Proximity to Inhabited areas, Proximity to Roads, Land Use and Cover, Vegetation Continuous Fields, and the Normalized Difference Vegetation Index. To evaluate the performance of the algorithms, the metrics used were Mean Absolute Error, Root Mean Square Error, and the Coefficient of Determination. The test results showed that the models had similar performance, and both the Random Forest (RMSE = 36.26, MAE = 17.45, and $R^2 = 0.99$) and the Extreme Gradient Boosting (RMSE = 35.73, MAE = 18.74, and $R^2 = 0.99$) demonstrated good predictive capacity. The elaborated map presented areas of high and very high susceptibility occupying 39% of the total area of the conservation unit, mainly located in the central-east and central-west regions. The variables with the greatest importance and contribution to the final model were environmental and socioeconomic variables, notably precipitation, distance from inhabited areas, and land use type.

1. Introduction

Throughout history, Amazonian communities have used fire as a tool for management of soil, pastures, and agricultural plantations, among other uses (Oliveira et al., 2020). The use of controlled fire, as well as its natural occurrence in specific environments and situations, has benefitted human activities, especially in ecosystems wherein fire is a naturally occurring phenomena, such as the savanna, where plants have morphological and physiological adaptations to the periodic presence of fire (Simon et al., 2009).

However, as a function of increasing pressure from agricultural activities and deforestation in the Amazon region, fire sometimes escapes from the control of those who purposely set it, thus becoming a destructive forest fire which can rapidly propagate itself and cause irreparable damage to the forest, the surrounding environment, and the human communities that inhabit nearby areas (Fonseca-Morello et al., 2017). Considering this, it is important that strategies are adopted that can protect environmental resources and the ecosystem services provided by the forest.

The management of an area as vast as the Legal Amazon, approximately 5 million km^2, or 58.9% of Brazilian territory (IBGE – Brazilian Institute of Geography and Statistics, 2020), would require large-scale

* Corresponding author.
E-mail addresses: engkemuel@gmail.com (K.M. Freitas), roniejuvanhol@ufpi.edu.br (R.S. Juvanhol), christiano.pinheiro@ufes.br (C.J.G. Pinheiro), anderson.menesesv@ufopa.edu.br (A.A.M. Meneses).

https://doi.org/10.1016/j.jsames.2025.105366
Received 25 November 2024; Accepted 16 January 2025
Available online 19 January 2025

optimization of available resources such as labor and money to reduce vulnerability to fire and improve decision making processes related to prevention, monitoring, and combat of environmental risks (Almeida et al., 2020; Ozenen Kavlak et al., 2021).

In this context, new technologies have shown promise in the search for solutions to problems caused by environmental risks. For example, the use of artificial intelligence tools has been shown to be effective in the prediction and monitoring for forest fires (Abid, 2021). The capacity to analyze large volumes of data in real time enables early detection of areas that are more susceptible to forest fires, thus permitting a more rapid and efficient response to their combat.

A large range of models that use machine learning and deep learning methods have been developed to predict the occurrence of forest fires in different contexts and situations. These models have been created using software such as Neural Networks, Algorithms form Classification and Regression Trees, Random Forest, Support Vector Machine, Deep Neural Network, XGBoost, among others, in Asia (Pourghasemi et al., 2020a; Pang et al., 2022), in Europe (Tonini et al., 2020; Michael et al., 2021), in North America (Phelps and Woolford, 2021; Abid, 2021) and in Brazil, principally in the southeast and central-west regions (Juvanhol et al., 2023; Rubí et al., 2023). However, there is still a great need for studies conducted in the Amazon, especially in protected areas that are subject to intense pressure from deforestation and recurring forest fires, such as is the case in the Environmental Protection Area (EPA) Triunfo do Xingu, in the state of Pará. This EPA was shown by INPE (National Institute for Space Research), through the PRODES (Program for the Calculation by Satellite of Deforestation in the Legal Amazon) project to be the most deforested conservation unit in the Legal Amazon and which has the highest number of fire hotspots (INPE – National Institute for Space Research, 2023).

In this context, the objective of this study was to predict and map areas that are susceptible to forest fires in the Triunfo do Xingu-PA

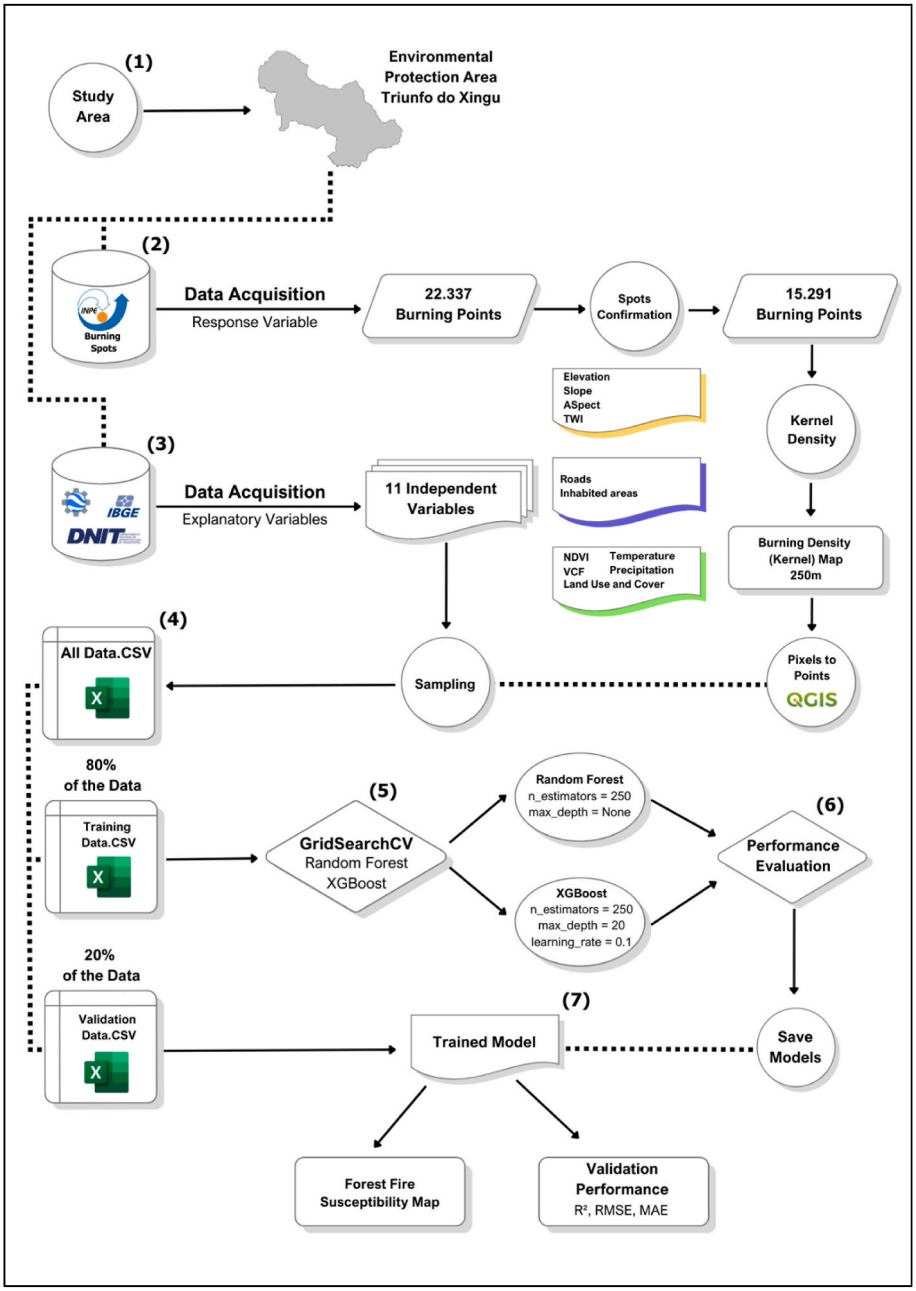

Fig. 1. Flowchart with the visual representation of main procedures and tools involved in this research.

Environmental Protection Area, using machine learning algorithms to verify the influence of environmental, topographical, and socioeconomic factors on fire occurrence.

Therefore, the principal contributions of this research are, (i) the application of machine learning tools to predict forest fire susceptibility in a protected area that is exposed to risk factors; (ii) identification and discussion of the aspects that most contribute to forest fire susceptibility, and (iii) compare the Random Forest and Extreme Gradient Boosting prediction models according to the cross-validation statistical method.

2. Material and methods

The flowchart for the procedures and tools used in this study is presented in Fig. 1.

2.1. Study area

The study area was the Environmental Protection Area (EPA) Triunfo do Xingu, a sustainable use conservation unit in the state of Pará, created by State Decree nº 2.612 on December 04, 2006. The EPA has a total area of 1,679,280.52 ha distributed between the municipalities of Altamira and São Félix do Xingu (Fig. 2). The National System of Conservation Units (SNUC), created by the Federal Law nº 9.985 of July 18, 2000, establishes that sustainable use units have as principal objective the harmonization of conservation with the sustainable use of resources.

The EPA Triunfo do Xingu is located in the Southeast mesoregion of the state of Pará. The climate is characterized as humid tropical and is classified as Ami in the Köppen classification system, with two well-defined seasons, one that is relatively dry between May and October, and the other rainy, between November and April. The average annual precipitation in the region is 2000 mm.year^{-1}, with average annual temperature of 27 °C and relative humidity of 80% (Carvalho et al., 2022).

The EPA Triunfo do Xingu is part of a mosaic of protected areas called the Terra do Meio, which, including this EPA, is composed of 10 conservation units (CUs). Besides the 10 conservation units, the mosaic of the Terra do Meio also contains indigenous territories (IT). However, the Triunfo do Xingu EPA has been indicated by several studies as a critical area with respect to deforestation and forest fires Araújo et al. (2017).

2.2. Burn (kernel) density – response variable

Active Fire were used as the target or response variable, and this variable was represented by the measurement of the density of fires in an area. Active Fire data were downloaded from the Wildfire Program (Programa Queimadas (BDQueimadas)) from the National Institute for Space Research (INPE), from the site https://terrabrasilis.dpi.inpe.br/queimadas/portal/. Each active fire spot indicates fire occurrence in a pixel that varies in size depending on the spatial resolution of the satellite.

The reference satellite used by INPE is the AQUA_M-T, which has a spatial resolution of 1 km and can detect fire fronts with a minimum of 30 m in length and 1 m in width. Therefore, one single pixel could have one or several fire fronts, which would generate just one active fire spot when there is actually more than one fire front. In contrast, one very

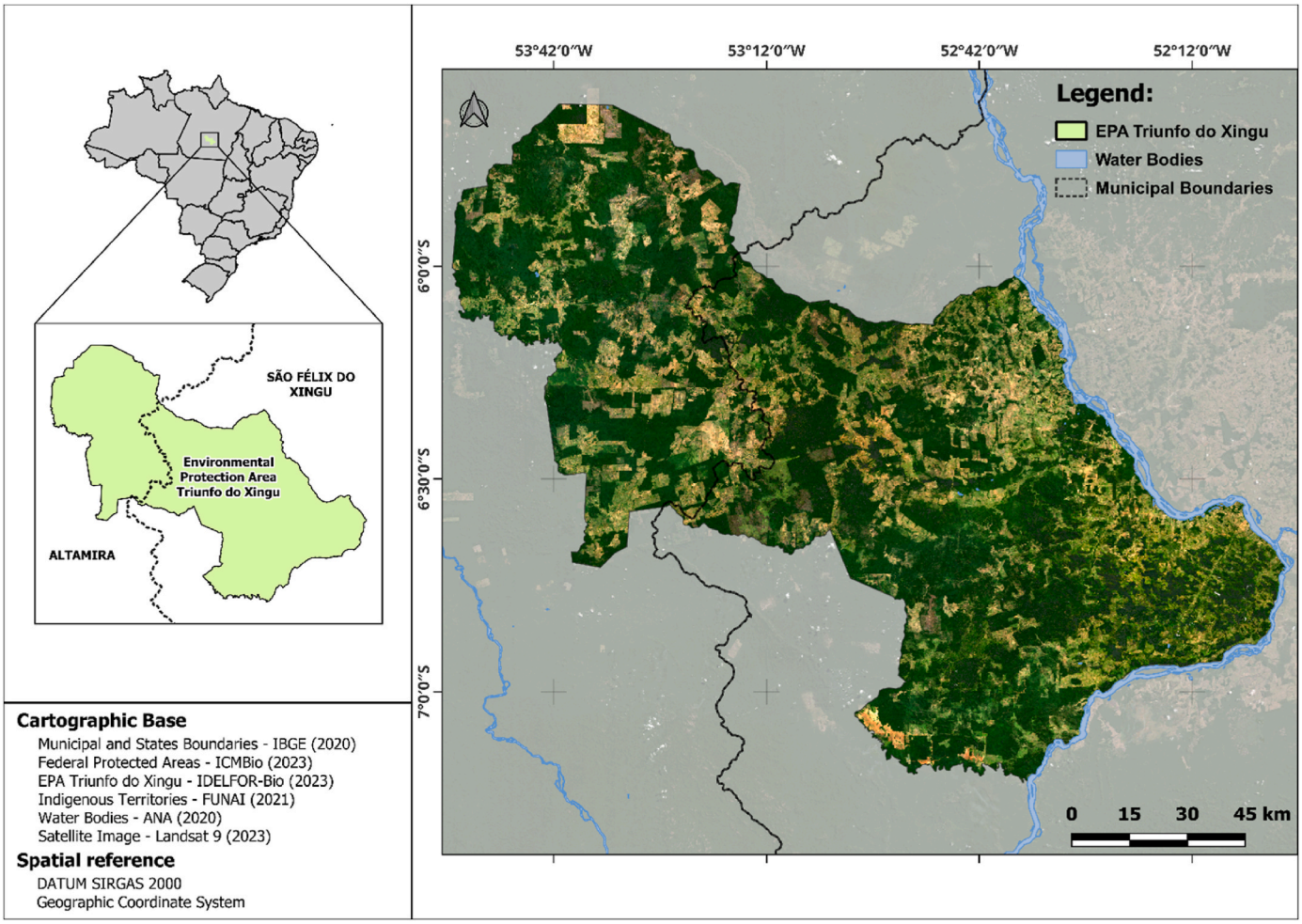

Fig. 2. Location Map of Environmental Protection Area (EPA) Triunfo do Xingu, Pará, Brazil.

large extension of fire could be detected across different pixels, which would then be registered as several active fire spots for one single event (INPE – National Institute for Space Research, 2023).

In the current study, the data from the internal area of the EPA Triunfo do Xingu were filtered to include the period 2010 to 2020, which yielded a total of 22,337 active fire spots detected by the reference satellite. However, since these active fire spots did not necessarily represent a single burning event, an additional step was taken to confirm these active fires. To this end, the product from the Burned Area program from INPE (Libonati et al., 2015; INPE – National Institute for Space Research, 2023) was used, which was validated by Rodrigues et al. (2019). Monthly data from the period 2010 to 2020 was downloaded and compared to the reference satellite data so that only points that were contained within burned areas were maintained in the final analysis. This procedure thus yielded a total of 15,291 validated active fire in the EPA Triunfo do Xingu, as shown in Fig. 3.

The kernel density technique was used to spatially organize the data throughout the entire study area with the objective of inferring the spatial variation of validated active fire and verify tendencies and patterns in individual fire events (Bertolla et al., 2014; Juvanhol et al., 2023).

The method used for the Kernel estimate is described in Rizzatti et al. (2020), wherein two principal factors, the radius of influence (R), and the estimation function (k). The radius of influence determines the area centered around the point being estimated. All points that are included in the radius are considered in the estimate calculation, while the estimation function is related to properties used to smooth the kernel.

To calculate Kernel density, vector data for each year were imported into the software QGIS 3.28.15, reprojected into the plane coordinate system SIRGAS 2000, zone UTM 22S, and combined into a single layer. In this way, the radius of influence (R) was calculated using the following equation:

$$R = \overline{X} \pm \overline{X_\sigma} \quad \text{(Equation 1)}$$

Where, $\overline{X}$ represents the average of the average distances between validated active fires, and $\overline{X_\sigma}$ represents the average of the standard deviations.

To calculate the overall average of the average distances and the standard deviations the tool “distance matrix” was used, the output of which yields average distance, standard deviation, and the maximum and minimum distance for each point, from which averages can then be calculated. In this way, the radius that best represents the distribution of the points throughout the EPA is done through subtraction of or the sum of the averages; for this study, subtraction of averages best represented the spatial distribution of the points. The selected estimation function was the quartic since it attributed greater weight to points that were closer to each other and less to those that were further away and displayed a gradual reduction as distance increased (Porter and Reich, 2012). Furthermore, the resolution of the raster output layer was established at 250 m using the size of the pixels.

To determine the dataset that represented the entire area, the conversion tool Raster Pixels to Points was used, which transformed each pixel in the image into central points that served as a base to extract information from each of the independent variables.

2.3. Explanatory features

The predictive variables used were Altitude, Slope, Aspect, Topographical Wetness Index (TWI), Precipitation, Temperature, Proximity to Urban Centers, Proximity to Roads, Land Use and Cover, Vegetation Continuous Fields, and the Normalized Difference Vegetation Index (NDVI) (Table 1).

The following subsections describe the methods used to acquire the datasets for each of three groups of variables, namely, topographic, socioeconomic, and environmental.

2.3.1. Topographic variables

The subset of topographic variables including Altitude, Slope, Aspect, and TWI (Fig. 4), were obtained using the tool *Google Earth Engine* – GEE, derived from digital elevation models – DEM elaborated using data from the SRTM (*Shuttle Radar Topographic Mission*) available from the USGS (*United States Geological Survey*), with a resolution of 30 m.

The *Topographic Wetness Index* (TWI) reflects the tendency for water accumulation at a specific point in an area, such that the lower the value the less is the accumulation of water, and therefore the greater the probability of fire occurrence (Pourtaghi et al., 2015; Nóbrega et al., 2018). The TWI is calculated using equation (2) below, initially proposed by Beven and Kirkby (1979):

$$TWI = \ln\left(\frac{\alpha}{\tan\beta}\right) \quad \text{(Equation 2)}$$

Where, α represents the contribution area and β is the slope in radians.

2.3.2. Socioeconomic variables

The variables associated with human presence and socioeconomic

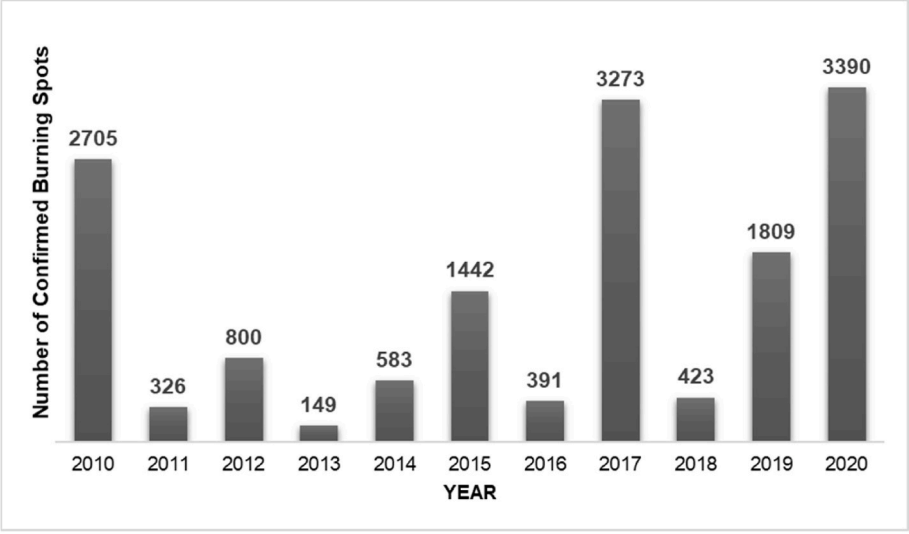

Fig. 3. Number of validated active fire registered by the National Institute of Space Research (INPE) in the EPA Triunfo do Xingu between 2010 and 2020.

Table 1
Explanatory Features considered in the analysis of the Fire Susceptibility Prediction Models.

FEATURE		SHORT NAME	RANGE
Topographic	Elevation	ELEVATION	167 m → 644 m
	Slope	SLOPE	0° → 60°
	Aspect	ASPECT	0° → 360°
	Topographic Wetness Index	TWI	7,3 → 17,06
Socioeconomics	Distance from roads	DIST_ESTRA	1 km → 15 km
	Distance from Inhabited areas	DIST_URB	1 km → 100 km
Environmental	Vegetation Continuous Fields	VCF	8% → 81%
	Temperature	TEMP	27 °C → 34 °C
	Precipitation	PRECIPITATION	2000 → 2300 mm.year^{-1}
	Normalized Difference Vegetation Index	NDVI	−1 → 1
	Land Use and Cover	USO_SOLO	Forest Savanna Wetland Grassland Pasture Temporary Crop Soybean Mining River, Lake

activities, represented by the proximity to communities and roads (Fig. 5), were obtained using the tool *Multi Ring Buffer* available in the software QGIS 3.28.7. This software creates a series of buffers at specific distances around urban areas and roads, and the buffers are then mixed together and dissolved to avoid superposition of values. The original data for communities and urban centers were sourced from the Brazilian Institute for Geography and Statistics – IBGE (2010), while that for roads was obtained from manual correction of vector data in shapefiles for federal and state highways, available from the DNIT – National Department of Transport Infrastructure (2022).

2.3.3. Environmental variables

To represent the subset of variables associated with the environment, the following characteristics were used: classes of Land Use and Cover, Vegetation Continuous Fields, and the Normalized Difference Vegetation Index – NDVI (Fig. 6).

The classes of Land Use and Cover are products taken from collection 8 of the Annual Series of Maps of classes of Land Use and Cover from the MapBiomas project, which spans the years 1965–2022 (MapBiomas Project, 2023; de Souza et al., 2020). The classes were downloaded with the aid of Google Earth Engine – GEE, in the GeoTIFF format at a scale of 1:100.000 and a spatial resolution of 30 m, using the year 2020 as a reference.

The Normalized Difference Vegetation Index (NDVI) was calculated using Landsat images from the interval 2010 to 2020 with a spatial resolution of 30 m. During image pre-processing, radiometric corrections were made to correct for the different generations of Landsat

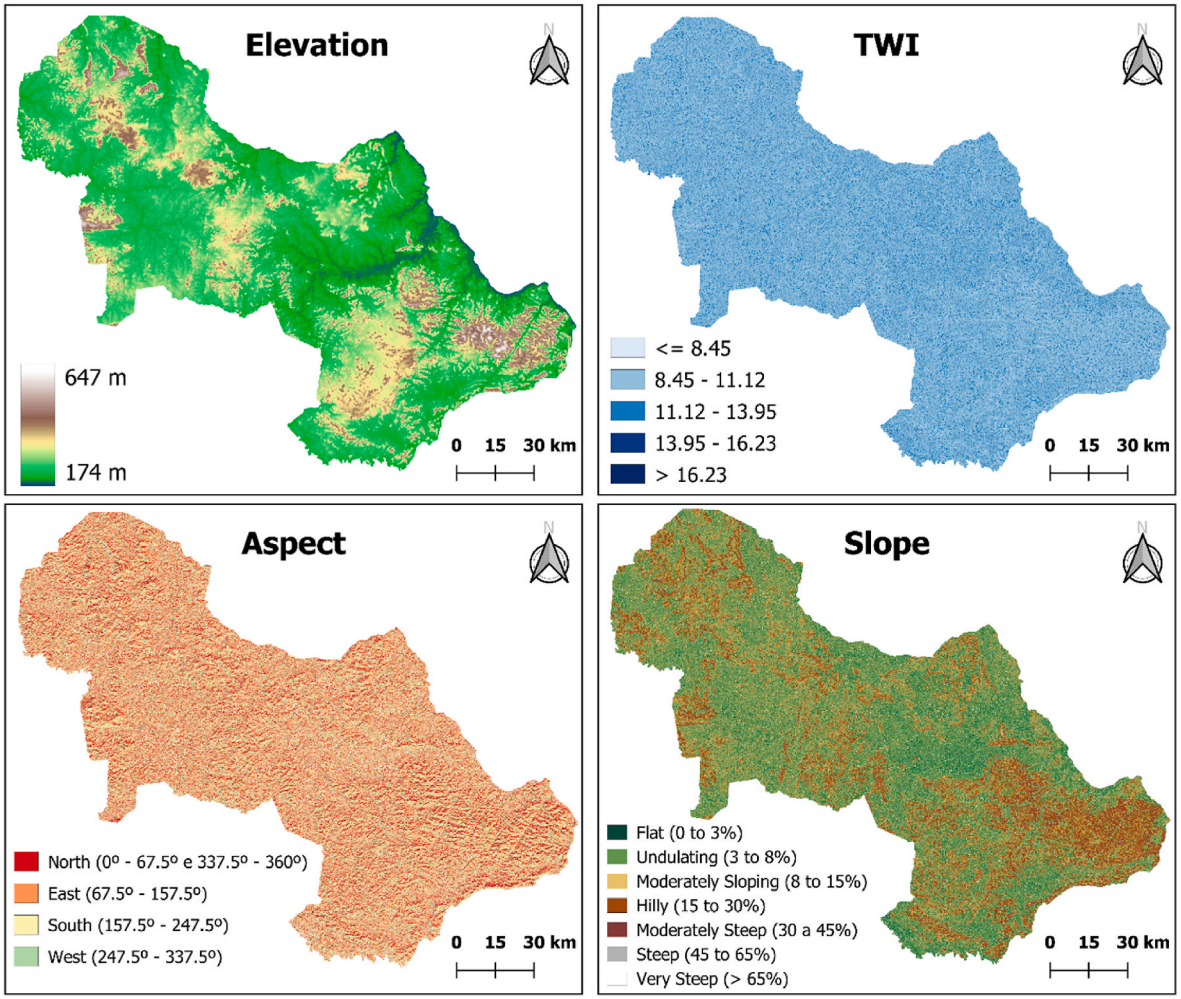

Fig. 4. Elevation, Aspect, Topographic Wetness Index (TWI) and Slope characteristics of the EPA Triunfo do Xingu, Para, Brazil.

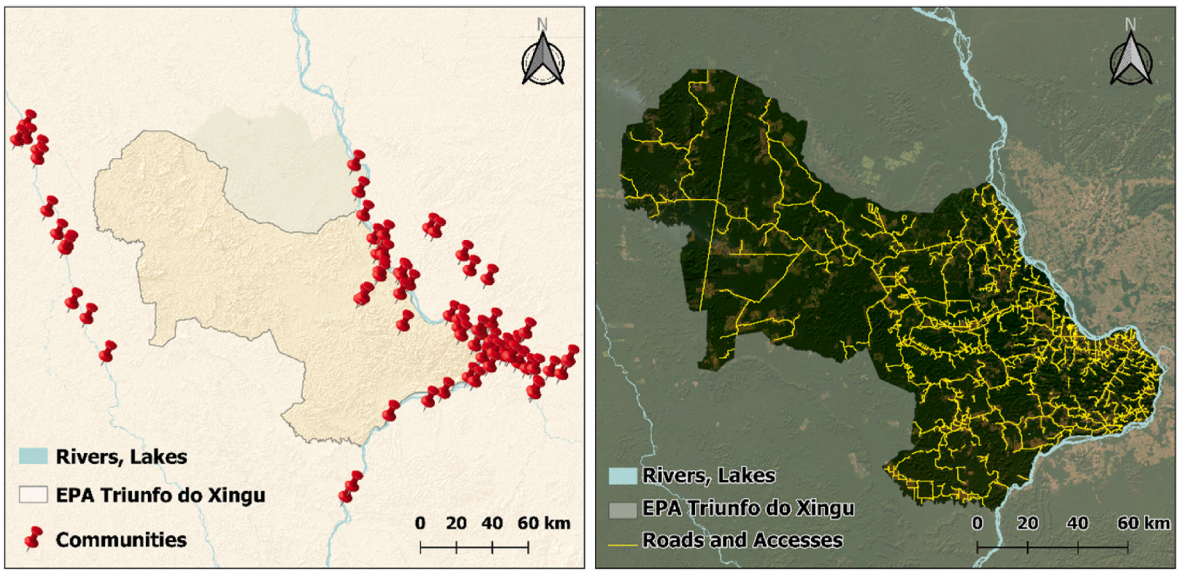

Fig. 5. Communities (Inhabited areas) and Roads Located in EPA Triunfo do Xingu and its surroundings, Para, Brazil.

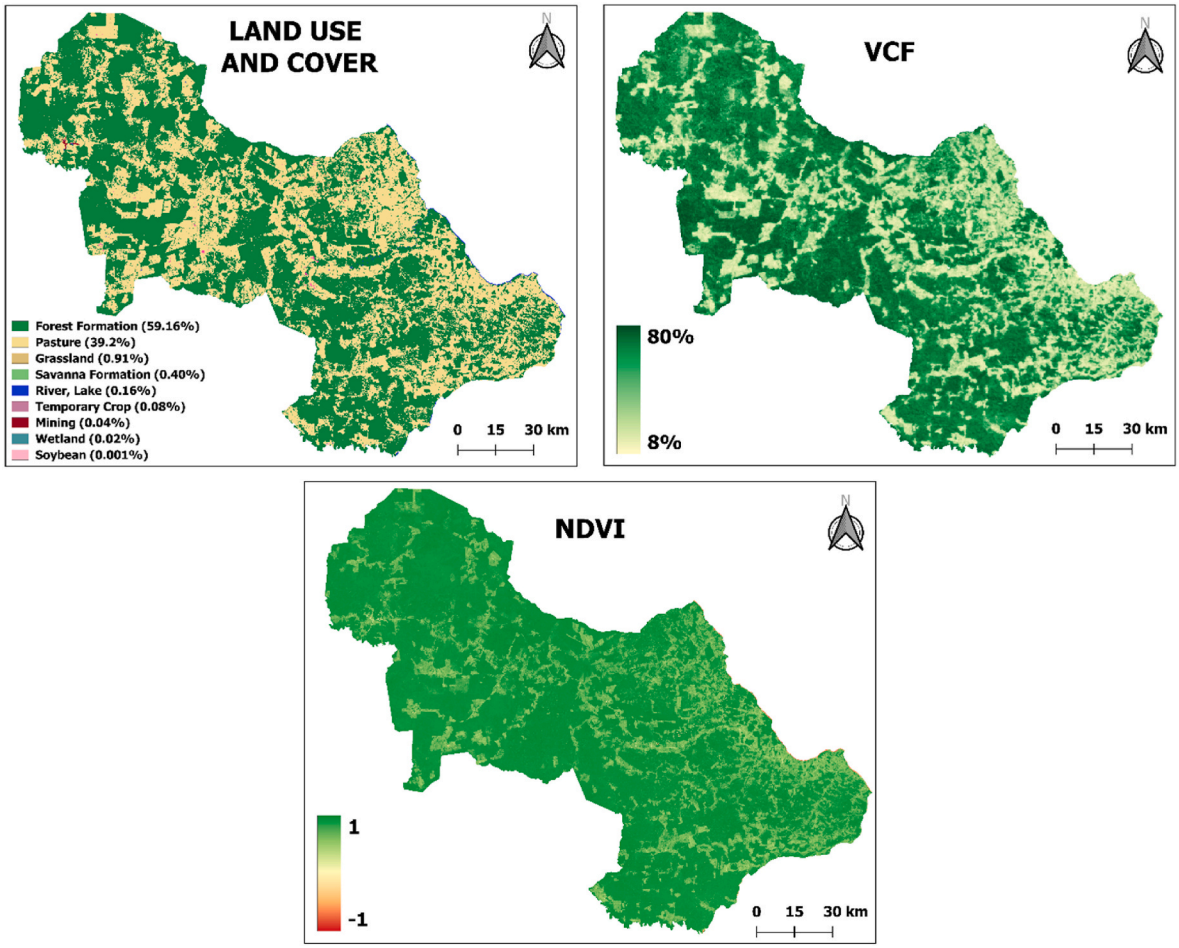

Fig. 6. Land Use and Cover, Vegetation Continuous Fields (VCF) and Normalized Difference Vegetation Index (NDVI) characteristics of the EPA Triunfo do Xingu, Para, Brazil.

sensors that created these images (Landsat 7 and 8), and interfering agents were removed (clouds and shade, considered up to 10%). The annual averages for the NDVI values were calculated for each of the analyzed years, and then the median of these values was extracted, and a single image was generated that represented the entire period. The NDVI was calculated using equation (3):

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad \text{(Equation 3)}$$

Where NIR represents the reflectance values in the near infrared band, and RED represents the reflectance values in the red band.

The Vegetation Continuous Fields (VCF) variable is represented by the percent tree cover of the MODIS annual product Vegetation Continuous Fields (MOD44B), with a spatial resolution of 250 m. The images were obtained using Google Earth Engine for each year between 2005 and 2020, and then the median was extracted to represent the entire period.

The climatic factors temperature and precipitation (Fig. 7), directly influence the occurrence, frequency, and intensity of fires by enabling or preventing the ignition of fuel (Vadrevu et al., 2010; Moreira et al., 2020). Precipitation data were obtained from GEE, using the dataset *Climate Hazards Group Infrared Precipitation with Station data* (CHIRPS), which has daily data taken at 3-h intervals at a spatial resolution of approximately 0.05°. In this study, the average annual precipitation was calculated for the years 2010–2020.

Land Surface Temperature (LST) was measured using the method proposed by De Jesus and Santana (2017) for the calculation of temperature from Landsat images. The average temperature data were from the years 2010–2020, and temperature (°C) for each year was calculated using equation (4):

$$LST = \frac{T_\beta}{1 + \left(\lambda \times \frac{T_\beta}{1.438} \times \ln \varepsilon\right)} - 273.15 \quad \text{(Equation 4)}$$

Where, T_β represents the value of the *at-satellite* temperature, which was calculated by converting the thermal band into a number using the scale factors and offset, λ represents the wavelength of the thermal band, and ε represents emissivity, as calculated by equation (5):

$$\varepsilon = 0.004 \times fv + 0.986 \quad \text{(Equation 5)}$$

Where *fv* represents the fraction of vegetation, which was obtained using equation (6):

$$fv = \left(\frac{NDVI - NDVI_{min}}{NDVI_{max} - NDVI_{min}}\right)^2 \quad \text{(Equation 6)}$$

Where, NDVI represents the Normalized Difference Vegetation Index, obtained using equation (3), $NDVI_{min}$ represents the lowest NDVI measured, and $NDVI_{max}$, the largest NDVI value.

2.4. Processing data

2.4.1. Pre-processing

The data were pre-processed in a programming environment using the Python language (https://www.python.org/) with the aid of the pandas (https://pandas.pydata.org/) library, which enable data analysis and manipulation. Missing data, which can occur principally along the edges of an image, were verified and substituted, and in the case of qualitative data, substitution was made using the value with the highest frequency, and for missing quantitative data the average was used (Batista and Monard, 2003).

In this regression scenario the input values must be numeric. In this way, categorical input values were substituted by quantitative values using the Integer Encoding method, wherein each categorical variable receives a single whole number value so that each category will be represented by a unique number, thus avoiding the creation of new columns of information, which reduces the input into the models.

2.4.2. Cross validation – training and validation data

The data were divided into training/testing and validation subsets (Fig. 8), with 80% of the data used for training and testing the models, and the other 20% were used to validate the trained models (Rácz et al., 2021). The objective of the division of the data into training/testing and validation subsets was to test the capacity for generalization of the models and to evaluate the model's learning process with the goal of application to data that were not used in the training steps (Dangeti, 2017).

Comparison of the tested models was done using the cross-validation statistical technique, which divides the dataset into k subsets (or *folds*) and then training the model k times using k-1 subsets for training, and the remaining subset for testing, with results being returned for each round of training and testing.

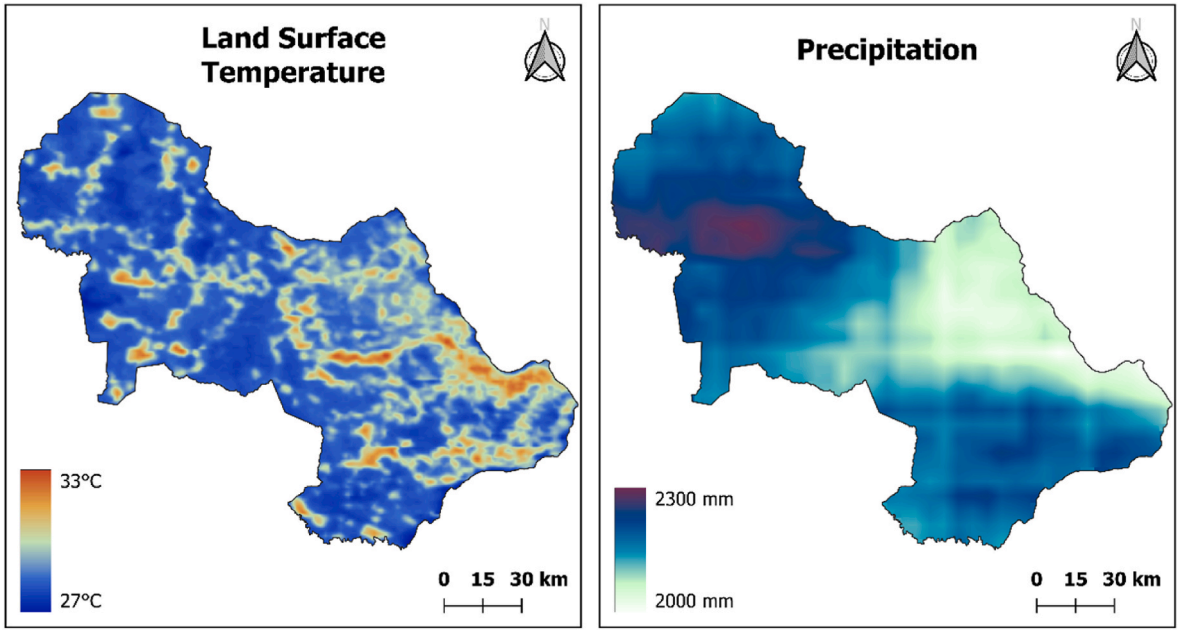

Fig. 7. Land Surface Temperature (LST) and Average Annual Precipitation characteristics of the EPA Triunfo do Xingu, Para, Brazil.

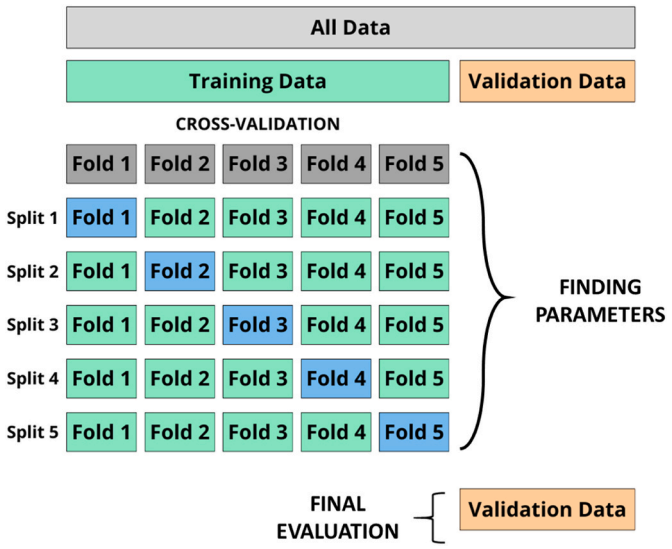

Fig. 8. Division of the input data into training/test data to perform cross-validation and validation data to test model generalization capacity.

2.5. Machine learning configuration

2.5.1. Algorithms

The *Random Forest* (RF) algorithm is an offshoot of the Decision Trees algorithm, and it represents an improved version of Decision Trees by eliminating some of its limitations. RF was introduced by Ho (1995) and then further developed by Breiman (2001). RF belongs to the ensemble category of algorithms, which combine predictions from multiple models to improve robustness and the precision of the predictions (González et al., 2020; Da Silva et al., 2022).

The functioning of RF is based on the creation of a collection of Decision Trees, each one constructed from different subsets of training data and random variables. The RF is a method based on *bagging* (Breiman, 1996), wherein the final model makes its predictions based on the predictions of several independently-trained trees that are combined through voting (in the case of classification) or by using the averages (in the case of regression) (Rokach, 2010).

As is the case for Random Forest, XGBoost is also an ensemble method, and is also a combination of multiple decision trees. However, in contrast to RF, XGBoost is based on the concept of boosting (Schapire and Freund, 2013), wherein models are sequentially constructed with the objective of correcting the errors in the previous model. At the end of the training, each prediction sample will have a corresponding score that reflects the contribution of each tree to the final prediction, and the corresponding scores are summed to obtain the final prediction value (Guo et al., 2020).

2.5.2. Model configuration

To evaluate the performance of the algorithms, the best architectures were defined using the function GridSearchCV from the scikit-learn library in Python 3.10, which analyzes different combinations of predefined parameters, analyzing the performance of the algorithm tested in each combination, and then lastly, choosing the parameters that had the best results using the cross-validation. To compare two different models, in this case RF and XGBoost, it is necessary to conduct two GridSearchCV functions since each one is configured with the specific parameters from each model and applied to the training and test data to find the best combination.

For the RF model, the class RandomForestRegressor from the ensemble module was used, which is part of the scikit-learn library. The hyperparameters that were explored were the maximum number of trees in the forest (parameter *n_estimators*) varying between 100, 150, 200 and 250 trees, and the maximum depth of each decision tree in the forest (parameter *max_depth*), which varied between 10, 20, 30 and without limits. For the XGBoost model, the class XGBRegressor from the library xgboost was used, and the identical parameters from the RF model were tested, along with the variation in the learning rate (parameter *learning_rate*) for 0.001, 0.01 and 0.1.

In this way, the best configuration found for the Random Forest algorithm was for 200 trees (n_estimators = 200) and without a maximum depth (max_depth = None). For the XGBoost algorithm, the configuration that presented the best performance was for 250 trees (n_estimators = 250), with a maximum depth of 20 (max_depth = 20) and a learning rate of 0.1 (learning_rate = 0.1). The results for each one of the tested variations for the RF and XGBoost algorithms are in the supplementary material section.

2.5.3. Hardware and software configuration

The computer used for the training and validation steps had an Intel® Core™ i5 – 7400 processer and 16 GB RAM memory. Regarding software configurations, Python version 3.10 was used, with Spyder version 5.4.3 as the Integrated Development Environment – IDE, and for creating maps, QGIS version 3.28.15 software was used.

2.6. Performance evaluation

To verify the performance of each model in predicting the response variables, three metrics were evaluated, the *Mean Absolute Error* – MAE, the *Root Mean Squared Error* – RMSE, and the *R-Squared* – R^2, described by equations (7)–(9):

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{(Equation 7)}$$

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{(Equation 8)}$$

$$R^2(y,\hat{y}) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{(Equation 9)}$$

Where, y represents the actual observed values, $\hat{y}$ represents the predicted values, and $\bar{y}$ represents the average values of y. The results for R^2 range from 0 to 1, with results closer to 1 conferring greater explanatory power in relation to the predicted data. In contrast to R^2, the values of MAE do not have a maximum limit and measure the difference between the real and predicted values, while the values of RMSE, despite using the same scale as the MAE values, penalize for outliers (Harrison, 2020).

The metrics are calculated in the cross-validation using the function 'cross validate' from the scikit-learn library, which allows for the determination of the number of folds (in this case, k = 10) and the metrics to be evaluated (MAE, RMSE and R^2). Subsequently, Mann-Whitney non-parametric test was used to compare the two algorithm models and to determine the best performance.

Finally, the model was validated with the previously separated validation data, testing its performance using data that were omitted during the training/testing phase.

2.7. Forest Fire Susceptibility Prediction Map Elaboration

Based on the choice of the best model using the described tests, the model was saved and applied to the entire dataset, thus generating the prediction values. These values were uploaded into a GIS system and classified into five classes of fire susceptibility, which were very low, low, medium, high, and very high. The statistical method used for the classification was the Jenks Natural Breaks algorithm, which adjusts the breaks between the interval classes to minimize the internal variation

within each class and maximize the variation between them (Jenks and Caspall, 1971).

2.7.1. Feature importance and contribution

The RF and XGBoost models provide data that aid in the comprehension of the importance of each predictor variable. The importance of each one of the variables used in the training step was obtained using the function feature_importances_ for both models. Another characteristic that was explored in this analysis was the contribution of each variable to the predictions. In this case, the SHAP values were used (Shapley Additive Explanations), which attributed a specific contribution of each variable to each model prediction. In contrast to feature importance, which identifies the variables that are the most influential in the global predictive capacity of the model, a feature contribution permits comprehension of how each variable contributes to the predictions of the model at an individual level, by using an average contribution made over the series of all the predictions as a whole (Marcílio and Eler, 2020).

The values of the feature contribution describe the influence between each variable and the predicted results, and this influence can be either positive or negative. This means that, as a variable increases, the result also tends to increase, and the contribution will be positive; in contrast, as a variable increases, the result tends to decrease, and the contribution will be negative.

3. Results and discussion

3.1. Machine learning models performance

The results from the ten split iterations of cross-validation of the Random Forest and XGBoost algorithms evaluated by the RMSE, MAE and R^2 metrics from the training data are shown in Table 2 and displayed in Fig. 9.

The Mann-Whitney non-parametric test for the RMSE and R^2 metrics showed no statistically significant differences between the two algorithms at a 95% confidence level; however, there was a significant difference for MAE.

These results demonstrate that the performance of both models was satisfactory, and due to lower computational cost as a function of faster execution time (Table 2), the XGBoost algorithm was chosen over Random Forest. Additionally, XGBoost offers greater scalability to be able to increase the quantity of input data, among other advantages (Chen and Guestrin, 2016). Furthermore, due to the similarity of performance between the algorithms, there were no apparent differences in the analysis of the map of prediction of fire susceptibility.

The use of ensemble algorithms has been widely used in the prediction of forest fires, such as in studies that used the regression method (Xie and Peng, 2019; Shmuel and Heifetz, 2023), as well as in classification studies (Michael et al., 2021; Bilucan et al., 2024; Seddouki et al., 2023; Agrawal et al., 2023), wherein the Random Forest and XGBoost ensemble algorithms had superior performance compared to the other tested models. These studies corroborate the similarity of the performance between the two algorithms, however, the XGBoost algorithm had better performance with respect to the unit cost value of the tests used in this analysis.

3.2. Validation data results

The capacity for generalization of the XGBoost model was tested using the model generated in the training/testing step using a previously unused set of data. The results of the validation of this model (Fig. 10) yielded average values RMSE (33.9902), MAE (18.1331), and R^2 (0.9934), demonstrating that the algorithm was able to make reliable predictions since it yielded validation results that agreed with those from the training/testing phase (Hastie, 2009; Arlot and Celisse, 2010).

3.3. Forest Fire Susceptibility Prediction Map

Using the validated model, the forest fire susceptibility map was generated from the classification of the values of validated active fire density into very low, low, average, high, and very high classes (Fig. 11). The areas that received a very low classification with respect to forest fire susceptibility are in dark green in Fig. 9 and represent just 0.7% of the total area of the EPA. These areas are located in the extreme south, central-west, and north regions of the EPA. In contrast, the areas classified as being of low- (light green) and average susceptibility (yellow), represent 27.36% and 32.94% of the total area, respectively, and represent the general condition of the EPA, with the areas radiating from the edges towards the center of the EPA. The areas of high susceptibility are concentrated in the central-east and central-west regions, and also in the southeast portion of the EPA, representing 25.95% of the total area of the EPA. The areas that require the highest level of monitoring, with a forest fire susceptibility classification of very high, occupy 13.05% of the total area, and, as for the medium and high classes, are also concentrated in the central-east and central-west regions.

Areas with very high susceptibility can be compared to areas with low and very low susceptibility (Table 3) in relation to each of the predictor variables. For example, for distance to roads, regions of lower susceptibility are an average distance of 3 km from a road, while regions

Table 2
Runtime and Cross-Validation for 10 splits (k-10 folds), Mean, Standard Deviation, Median, Minimum and Maximum values of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2) for Random Forest and XGBoost algorithms using the training data.

RANDOM FOREST					**XGBOOST**				
Runtime: 3947 s					**Runtime:** 1141 s				
		RMSE	**MAE**	**R^2**			**RMSE**	**MAE**	**R^2**
CROSS-VALIDATION (K = 10)	**SPLIT 1**	36.8718	17.4682	0.9925	**CROSS-VALIDATION (K = 10)**	**SPLIT 1**	35.8865	18.8054	0.9929
	SPLIT 2	34.8124	17.2537	0.9933		**SPLIT 2**	33.5265	18.6657	0.9936
	SPLIT 3	34.7051	17.2384	0.9932		**SPLIT 3**	35.0228	18.3015	0.9931
	SPLIT 4	36.2188	17.4631	0.9926		**SPLIT 4**	34.7056	18.7247	0.9932
	SPLIT 5	36.6295	17.5724	0.9926		**SPLIT 5**	38.4170	19.2998	0.9918
	SPLIT 6	36.6142	17.5465	0.9925		**SPLIT 6**	37.1396	18.8565	0.9923
	SPLIT 7	37.6799	17.5411	0.9918		**SPLIT 7**	36.0321	18.7801	0.9927
	SPLIT 8	36.3186	17.4276	0.9925		**SPLIT 8**	36.9110	18.8405	0.9922
	SPLIT 9	37.0203	17.7360	0.9924		**SPLIT 9**	35.9579	18.8380	0.9928
	SPLIT 10	35.0535	17.3349	0.9931		**SPLIT 10**	33.7524	18.3432	0.9936
MEAN		36.2616	17.4582	0.9926	**MEAN**		35.7351	18.7455	0.9929
STD. DV.		1.0313	0.1453	0.0004	**STD. DV.**		1.4520	0.2660	0.0006
MEDIAN		36.4990	17.4656	0.9926	**MEDIAN**		35.9222	18.7927	0.9929
MIN.		34.5729	17.2384	0.9918	**MIN.**		33.5265	18.3015	0.9918
MAX.		38.0857	17.7360	0.9933	**MAX.**		38.4170	19.2998	0.9936

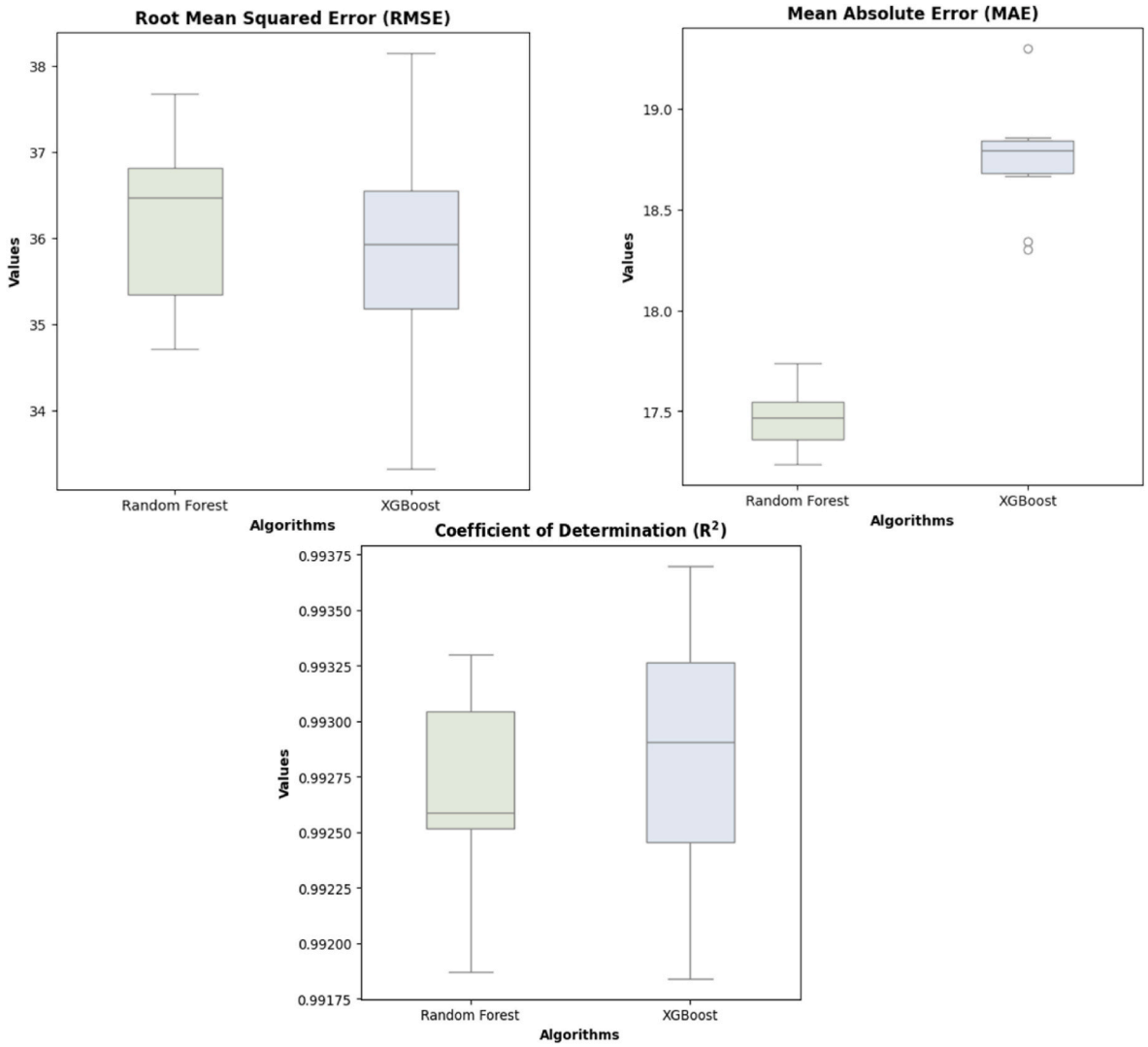

Fig. 9. Boxplot of Cross-Validation values of Root Mean Squared Error, Mean Absolute Error and Coefficient of Determination for Random Forest and XGBoost using the training data.

of greater susceptibility are just 2 km from a road. For inhabited areas, regions of very low and low susceptibility are slightly more distant (average of 40 km) compared to areas with greater susceptibility (average of 38 km).

With respect to topographic characteristics, very low and low susceptibility areas, and areas with very high susceptibility have a similar profile between each other and with the other susceptibility classes in the EPA. These regions have low elevation, about 260 m above sea level, and slopes that range between 5 and 7%, with little accumulation of water on the soil surface and a slight predominance of a northern aspect of the hillsides. This demonstrates that áreas with low and high risk have vegetation that is susceptible to fire, explained by the higher level of incident radiation due to the predominance of northern aspect of the landscape. This aspect receives intense solar radiation during the hottest part of the day, which could have a direct effect on drying of fuel on the forest floor.

In contrast, when analyzing the environmental variables, there are several notable differences, principally in relation to average annual temperature and precipitation, land use types, and vegetation cover. Areas with low and very low fire susceptibility had an average temperature that was 1.5 °C lower compared to areas with greater susceptibility, 32 °C and 33.5 °C, respectively. Average annual precipitation showed the same trend, with areas of greater fire susceptibility having an average of 2000 mm.year^{-1}, while areas with lower fire susceptibility had an average of 2200 mm.year^{-1}.

The predominant land use type in each fire susceptibility class showed that there was a predominance of forest in areas with lower susceptibility, while in areas with very high susceptibility pasture was the dominant land use type. With respect to vegetation, in areas of low and high susceptibility to fire, there was healthy vegetation, showing an average NDVI of 0.8, however, when analyzing the VCF values, areas with lower fire susceptibility had a greater percentage of vegetation cover, with an average of 56% compared to areas with greater fire susceptibility (47%).

The predominance of pasture in areas that have greater fire susceptibility is a result of the process of transformation of natural ecosystems into productive agroecosystems where fire is used to clear forests and manage pastures (de Vasconcelos et al., 2013). This process has been indicated by the Amazonian Institute of Man and the Environment (IMAZON) as being related to deforestation that is a result of timber extraction and the spread of livestock pastures (Souza et al., 2018). Dos Santos et al. (2020, 2023), working in the region of the EPA Triunfo do Xingu and in the municipality of São Félix do Xingu, related that in the last 10 years, 12,000 km^2 of forest were substituted by pasture, and that the management used in these pastures has caused fire to spread into forested areas.

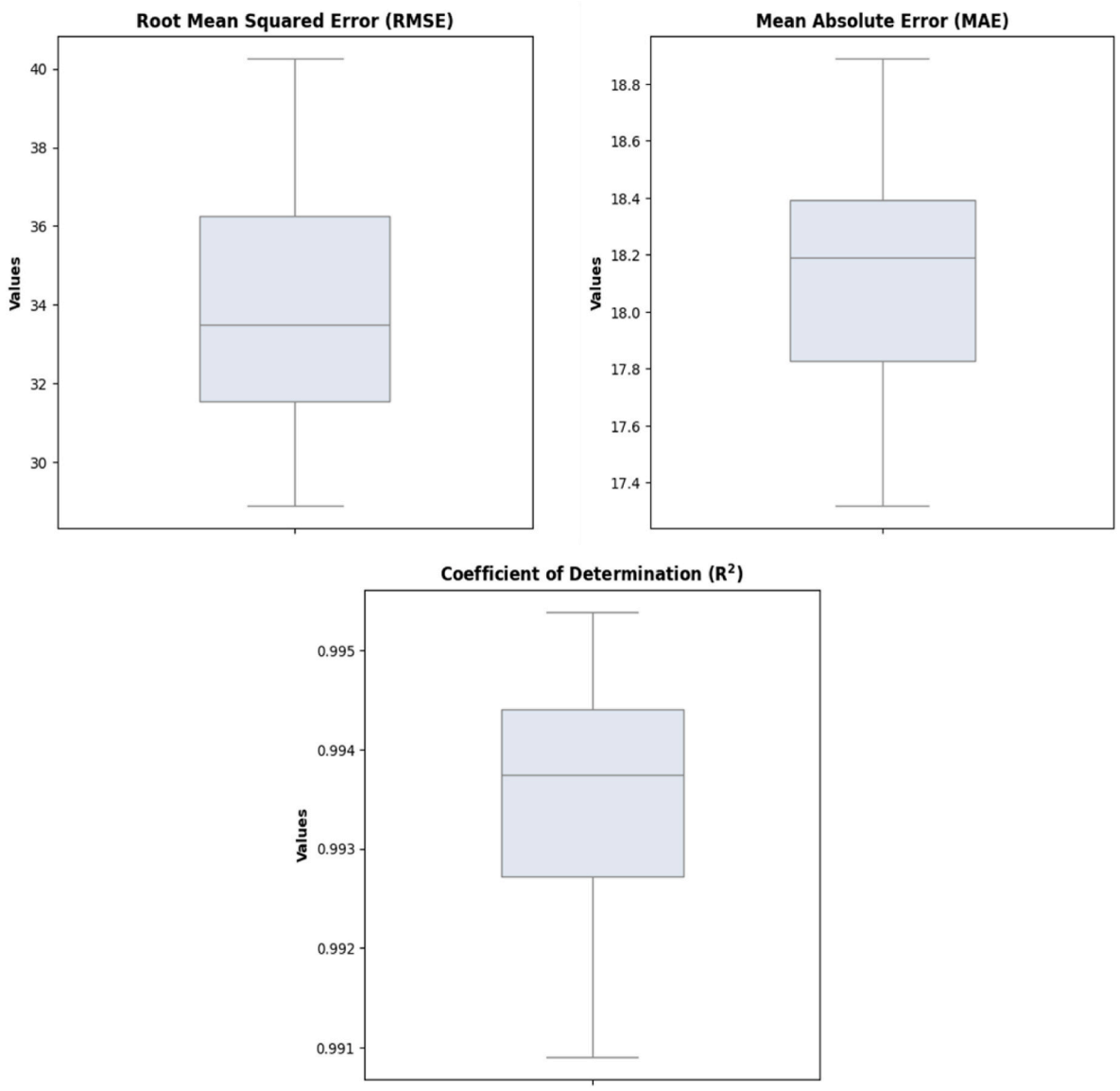

Fig. 10. Boxplot of Cross-Validation values of Root Mean Squared Error, Mean Absolute Error and Coefficient of Determination for XGBoost using validation data.

The lack of variation in the NDVI values between the areas of lower and higher susceptibility reflects the weak influence of factors related to vegetation on forest fires. The NDVI values for each area, about 0.8, indicate that these areas have healthy and dense vegetation. Furthermore, areas that are less susceptible to fire were located only in areas with greater continuity of forest vegetation (56%), which forms a dense natural barrier against the propagation of fires (Loudermilk et al., 2022; Tian et al., 2022). On the other hand, even though the NDVI values in more susceptible areas were high, the predominance of pastures suggests that it is a vegetation type that is more susceptible to fire due to its simple and uniform structure (Neary and Leonard, 2020; Yan and Liu, 2021), besides having a more fragmented structure, as shown by the VCF values, which facilitates fire ignition and propagation.

3.4. *Feature influence and feature contribution*

Table 4 lists the percentage importance of each independent variable to the explanation of variance in the response variable, in this case the predicted values of validated active fire density. The greater the importance value, the larger the impact that the variable had on the predictive capacity of the model (Casalicchio et al., 2019).

The Precipitation variable had greater relative percent importance (70.45%) for the model, suggesting that average annual precipitation is a significant factor in the susceptibility of fire occurrence. The next highest importance value was for distance to Urban Areas (15.02%) and then Land Use (7.87%), demonstrating that these were important variables for the model. Other variables such as Elevation, Distance to Roads, and Temperature had lower percentages but still had an important role in fire susceptibility, although to a smaller extent than the aforementioned variables. The influence of the remainder of the variables was very low, suggesting that they had minimum or no impact on the model, although it is important to analyze the overall context and not simply examine each variable separately, because even a variable with a very low importance variable might add important information to the model. In this sense, even though a specific variable appears not to be important for the reduction of uncertainty in the model's predictions, it still might be relevant to fire occurrence since causal or linear relationships between variables is not measured by the feature importance parameter.

In contrast to feature importance, Fig. 12 shows the average contribution of each independent variable to the prediction of the response variable. The variable with the greatest absolute average contribution

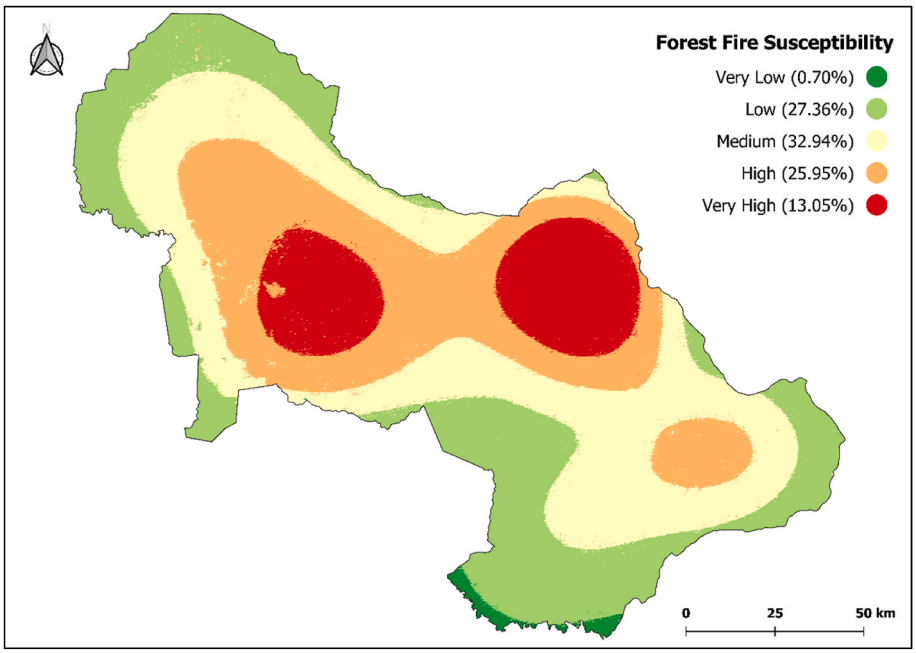

Fig. 11. Forest Fire Susceptibility Prediction Map for EPA Triunfo do Xingu, Para, Brazil.

Table 3
Characterization of areas with Very Low and Low Forest Fire Susceptibility in relation to Very High Forest Fire Susceptibility Areas considering the explanatory features studied.

FEATURE		VERY LOW AND LOW SUSCEPTIBILITY AREAS	VERY HIGH SUSCEPTIBILITY AREAS
Socioeconomics	Distance from roads (km)	Mean = 3	Mean = 2
	Distance from inhabited areas (km)	Mean = 38	Mean = 40
Topographic	Elevation (m)	Mean = 269	Mean = 253
	Slope	Mean = 7.3%	Mean = 5.2%
	TWI	Mean = 10.45	Mean = 10.6
	Aspect	North (26.52%)	North (25.71%)
		East (25.01%)	East (25.83%)
		South (24.41%)	South (24.11%)
		West (24.04%)	West (24.33%)
Environmental	Temperature	Mean = 32 °C	Mean = 33.5 °C
	Precipitation	Mean = 2200 mm. year^{-1}	Mean = 2000 mm. year^{-1}
	Land Use and Cover	Forest Areas (67.57%)	Pasture (53.58%)
		Pasture (31.24%)	Forest Areas (43.41%)
		Others (1.2%)	Others (3%)
	NDVI	Mean = 0.8	Mean = 0.8
	VCF	Mean = 56%	Mean = 47 %

Table 4
Feature Importance of each explanatory feature in the final XGBoost model.

FEATURE	IMPORTANCE
Precipitation	70.45%
Distance from Inhabited areas	15.02%
Land Use and Cover	7.87%
Elevation	2.29%
Temperature	1.79%
Distance from roads	1.38%
VCF	0.62%
NDVI	0.41%
Slope	0.07%
TWI	0.03%
Aspect	0.02%

was Distance to Urban Areas (+34.66), followed by precipitation (−20.06), Distance to a Road (−3.82), Land Use (−3.46), Elevation (−2.60), Temperature (−2.26) and NDVI (−1.48), while Aspect, Slope, TWI and VCF did not make significant contributions.

The strong influence of precipitation, allied with its negative contribution, indicates that annual average precipitation plays an important rle in the prediction of forest fire susceptibility by decreasing the probability of fire occurrence in areas with greater precipitation. Pourghasemi et al. (2020b), Mohajane et al. (2021), Celis et al. (2023), Juvanhol et al. (2023), and Hang et al. (2024), studying the prediction and occurrence of fires using machine learning, also reported that precipitation was one of the most important predictor variables, highlighting the influence of rainfall on making fuel on the forest floor less flammable during rainy periods.

In the studies mentioned above, especially that of Pourghasemi et al. (2020b) and Mohajane et al. (2021), conducted in Iran and Morocco, respectively, the smallest volume of annual precipitation was 1200 and 635 mm.year^{-1}, respectively. Such low volumes of annual rainfall could stimulate forest fire occurrence since vegetation becomes increasingly sensitive to fire with lower volumes of water availability. In the Amazon region, high annual rainfall causes the opposite relationship, as described by Dos Santos et al. (2023). These authors reported that high volumes of rainfall in the region, about 2000 mm.year^{-1}, provides water to the vegetation year-round. In this way, water serves to limit the potential for ignition of fire and mitigates its effects.

The Distance to Urban Areas variable, despite having a lower importance value than precipitation (Table 4), has a high absolute average contribution (Fig. 10), which demonstrates its relevance in this study of prediction of forest fires. Previous studies, such as Bui et al. (2017), Kalantar et al. (2020) and Pham et al. (2020), used different machine learning algorithms to predict fire susceptibility, and reported that the distance to inhabited areas was among the most important variables for fire prediction. With respect to the variables that indicate proximity to human activities, such as distance to roads, despite having lower importance and contribution to the model, play a fundamental

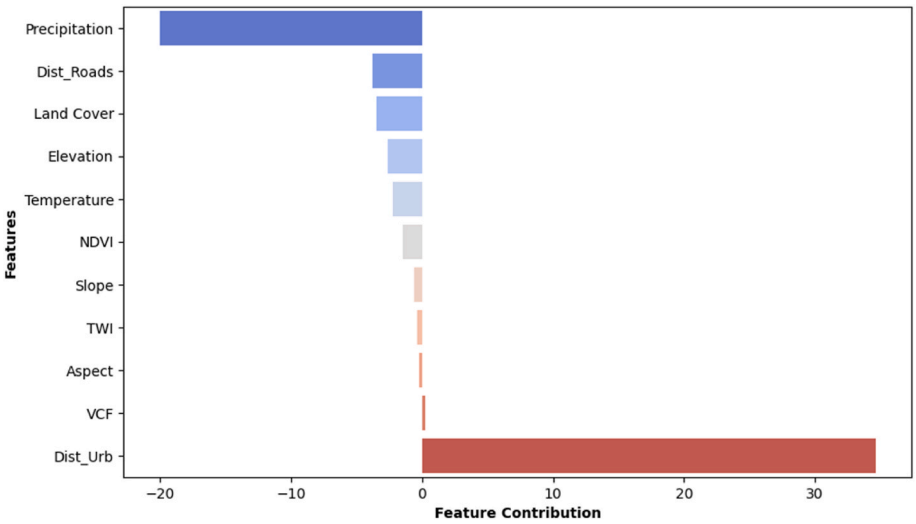

Fig. 12. Average Contribution of each explanatory variable to the final Forest Fire Susceptibility prediction model.

role in the analysis of anthropogenic factors, which have been shown to be one of the principal factors that cause forest fires (Nami et al., 2018; Pang et al., 2022; Celis et al., 2023).

The minimal influence shown in the regression of distance to roads (Table 4) could be explained by the incompatibility between the two variables (distance to roads and kernel density). The distance to roads loses its importance when correlated with the kernel density output instead of with the localization of the validated active fire (Amatulli et al., 2006). This variable could be better analyzed, for example, by considering the absolute distance from the validated active fire, or as a function of the density of the road network (Chuvieco, 2012). These alternative approaches highlight the importance of distance to roads as a risk factor for forest fires, with due to its role in pre-fire suppression operational activities.

Greater proximity to inhabited areas and roads shown in areas predicted to be more susceptible to fires is clear evidence of the impacts of urbanization in these areas. The greater concentration of human activities in these areas cause socioeconomic impacts that are reflected not only in an increase in the occurrence of fires, but also in landslides, floods, erosion, contamination of groundwater, and deforestation (Pasqualotto and Sena, 2017; Gama et al., 2019). In studies on the dynamics of active fire in the EPA Triunfo do Xingu, Rosan et al. (2017) and Dos Santos et al. (2023) reported that greater concentrations of active fire in the EPA occurred in areas of greater population concentration along new occupation fronts and areas of economic expansion. Sousa et al. (2016), highlighted that the EPA has a large diversity of areas that have been cleared along roads, with forest areas and agricultural fields in dispute between family-based and large-scale farmers, which increases deforestation pressure on these areas and is directly related to an increase in fires, as shown by Dos Santos et al. (2023).

With respect to topographic variables, these are generally less predictive of fire occurrence compared to other factors. However, among topographic variables, altitude was the most relevant (Table 4, Fig. 10), and the contribution of slope is explained through its importance in models of fire behavior (Jain et al., 2020; Chen et al., 2024).

Land use and vegetation cover is another factor that had considerable influence and an important contribution in this study. Since it is a reflection of the interaction between human activity and the natural environment, its influence and contribution could have been masked by other elements but were still able to offer important information to the prediction model. The importance of land use has been discussed by Pourtaghi et al. (2016), Pourghasemi et al. (2020b) and Pham et al. (2020) as one of the most important variables in the process of predicting and mapping fire susceptibility.

4. Conclusion

This study demonstrated the capacity of two ensemble algorithms, Random Forest and XGB, to predict, with different levels, areas that are susceptible to the occurrence of forest fires in the EPA Triunfo do Xingu. The similar performance of the models indicates that both can be used for prediction mapping in these areas. The map based on the results from the XGB model showed areas of greater susceptibility in the central-east and central-west portions of the EPA. The environmental and socioeconomic variables had greater importance and contribution to the model, showing that areas that are more susceptible to fire are nearer to human activity, receive a lower annual volume of rainfall, and have pastures as the predominant land use cover.

With respect to the tested models, even though there were no differences in predictive capacity, the XGB model displayed superior computational performance compared to the RF and is preferred for use with factors such as execution time and scalability.

To our knowledge, this is the first study of the application of machine learning techniques in the EPA Triunfo do Xingu, which makes an important contribution to better understanding of the dynamics of fire occurrence in this protected area. Nevertheless, the use of data with better spatial and temporal resolution, when available, could improve our understanding and help to combat forest fires in conservation areas, thus avoiding the loss of forest resources and the ecosystem services they provide.

In this context, to meet the challenge of protecting forests and the communities that depend on them, it is important to conduct studies that take new approaches, integrate new sources of data, and that combine new artificial intelligence techniques, machine learning, and deep learning with human expertise. In this way, we can leverage artificial intelligence and pave the way to a more secure and sustainable future where technology becomes a vital ally in the preservation of natural ecosystems.

CRediT authorship contribution statement

Kemuel Maciel Freitas: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ronie Silva Juvanhol:** Data curation, Formal analysis, Investigation, Methodology, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. **Christiano Jorge Gomes Pinheiro:** Conceptualization, Investigation, Supervision, Visualization, Writing – review & editing. **Anderson Alvarenga de Moura**

Meneses: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing.

Ethical statement

Hereby, I Kemuel Maciel Freitas consciously assure that for the manuscript "Prediction of Forest Fire Susceptibility using Machine Learning Tools in the Triunfo do Xingu Environmental Protection Area, Amazon, Brazil" the following is fulfilled.

1) This material is the authors' own original work, which has not been previously published elsewhere.
2) The paper is not currently being considered for publication elsewhere.
3) The paper reflects the authors' own research and analysis in a truthful and complete manner.
4) The paper properly credits the meaningful contributions of co-authors and co-researchers.
5) The results are appropriately placed in the context of prior and existing research.
6) All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference.
7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.
8) If accepted, the article will not be published elsewhere in the same form, in English or in any other language, including electronically, without the written consent of the copyright-holder.

I agree with the above statements and declare that this submission follows the policies outlined in the Guide for Authors and in the Ethical Statement.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was financed by the Coordination of Superior Level Staff Improvement – Brasil (CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) – Finance Code 001. The authors would like to thank Dr. Troy Beldini for help with language and proof reading the article. We would also like to thank A. Folguera from the Journal of South American Earth Sciences for the editorial handling.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsames.2025.105366.

Data availability

Data will be made available on request.

References

Abid, F., 2021. A survey of machine learning algorithms based forest fires prediction and detection systems. Fire Technol. 57 (2), 559–590. https://doi.org/10.1007/s10694-020-01056-z.

Agrawal, N., Nelson, P.V., Low, R.D., 2023. A novel approach for predicting large wildfires using machine learning towards environmental justice via environmental remote sensing and atmospheric reanalysis data across the United States. Rem. Sens. 15 (23), 5501. https://doi.org/10.3390/rs15235501.

Almeida, T.E.G., Flores, M.D.S.A., Sobrinho, M.V., 2020. Mapeamento de Risco de Desastre por Incêndio Florestal na Amazônia: Uma Abordagem Multifatorial no Município de Moju (PA). InterEspaço: Revista de Geografia e Interdisciplinaridade 6 (19), e202009. https://doi.org/10.18764/2446-6549.202009.

Amatulli, G., Rodrigues, M.J., Trombetti, M., Lovreglio, R., 2006. Assessing long-term fire risk at local scale by means of decision tree technique. J. Geophys. Res.: Biogeosciences 111 (G4). https://doi.org/10.1029/2005JG000133.

Araújo, E., Barreto, P., Baima, S., Gomes, M., 2017, pp. 24–28. Belém: Imazon.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statist. Surv. 4, 40–79. https://doi.org/10.1214/09-SS054.

Batista, G.E., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17 (5–6), 519–533. https://doi.org/10.1080/713827181.

Bertolla, J.M., Kawamoto, M.T., Falcão, J.G., Tandel, M.D.C.F.F., Govone, J.S., 2014. Processos pontuais aplicados ao estudo da distribuição espacial de enfermidades na área urbana da cidade de Rio Claro, SP. Revista da Estatística da Universidade Federal de Ouro Preto 3 (3), 684–688.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrol. Sci. J. 24 (1), 43–69. https://doi.org/10.1080/02626667909491834.

Bilucan, F., Teke, A., Kavzoglu, T., 2024. Susceptibility mapping of wildfires using XGBoost, random forest and AdaBoost: a case study of mediterranean ecosystem. In: Bezzeghoud, M., et al. (Eds.), Recent Research on Geotechnical Engineering, Remote Sensing, Geophysics and Earthquake Seismology. MedGU 2022. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-031-48715-6_22.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Bui, D.T., Bui, Q.T., Nguyen, Q.P., Pradhan, B., Nampak, H., Trinh, P.T., 2017. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. Agric. For. Meteorol. 233, 32–44. https://doi.org/10.1016/j.agrformet.2016.11.002.

Carvalho, A.B., Moreira, R.P., Herrera, J.A., 2022. Aspectos da Dinâmica Climática de Altamira-PA. Rev. Percurso 14 (2), 23–34.

Casalicchio, G., Molnar, C., Bischl, B., 2019. Visualizing the feature importance for black box models. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (Eds.), Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018, Lecture Notes in Computer Science(.), vol. 11051. Springer, Cham. https://doi.org/10.1007/978-3-030-10925-7_40.

Celis, N., Casallas, A., Lopez-Barrera, E.A., Felician, M., De Marchi, M., Pappalardo, S.E., 2023. Climate change, forest fires, and territorial dynamics in the Amazon rainforest: an integrated analysis for mitigation strategies. ISPRS Int. J. Geo-Inf. 12 (10), 436. https://doi.org/10.3390/ijgi12100436.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Chen, R., He, B., Li, Y., Fan, C., Yin, J., Zhang, H., Zhang, Y., 2024. Estimation of potential wildfire behavior characteristics to assess wildfire danger in southwest China using deep learning schemes. J. Environ. Manag. 351, 120005. https://doi.org/10.1016/j.jenvman.2023.120005.

Chuvieco, E. (Ed.), 2012. Remote Sensing of Large Wildfires: in the European Mediterranean Basin. Springer Science & Business Media. https://doi.org/10.1007/978-3-642-60164-4.

Da Silva, D.G., Geller, M.T.B., Dos Santos Moura, M.S., De Moura Meneses, A.A., 2022. Performance evaluation of LSTM neural networks for consumption prediction. e-Prime-Advances in Electrical Engineering, Electronics and Energy 2, 100030. https://doi.org/10.1016/j.prime.2022.100030.

Dangeti, P., 2017. Statistics for Machine Learning. Packt Publishing Ltd.

De Jesus, J.B., Santana, I.D., 2017. Estimation of land surface temperature in caatinga area using Landsat 8 data. Journal of Hyperspectral Remote Sensing 7 (3), 150–157. https://doi.org/10.29150/jhrs.v7.3.p150-157.

de Souza, A.A., Oviedo, A., dos Santos, T.M., 2020. Impactos na qualidade do ar e saúde humana relacionados ao desmatamento e queimadas na Amazônia Legal brasileira, vol. 21. Instituto Socioambiental, São Paulo, SP, Brazil.

de Vasconcelos, S.S., Fearnside, P.M., de Alencastro Graça, P.M.L., Dias, D.V., Correia, F.W.S., 2013. Variability of vegetation fires with rain and deforestation in Brazil's state of Amazonas. Remote Sensing of Environment 136, 199–209. https://doi.org/10.1016/j.rse.2013.05.005.

DNIT – National Department of Transport Infrastructure, 2022. DNITGeo data viewer. Available at: https://servicos.dnit.gov.br/vgeo/. (Accessed 6 July 2023).

Dos Santos, M.G., Neris, J.P.F., De Freitas, T.P.M., 2020. Uso de geotecnologias na análise espacial dos focos de calor no município de São Félix do Xingu, Pará. Geografia: Publicações Avulsas 2 (1), 395–419.

Dos Santos, G.G., Neris, J.P.F., Coelho, R.D.F.R., 2023. Dinâmica dos Focos de Calor na Área de Proteção Ambiental Triunfo do Xingu, Amazônia Paraense. Revista GeoAmazônia 11 (22), 23–45. https://doi.org/10.18542/geo.v11i22.13770.

Fonseca-Morello, T., Ramos, R., Steil, L., Parry, L., Barlow, J., Markusson, N., Ferreira, A., 2017. Queimadas e Incêndios Florestais na Amazônia Brasileira: Porque as Políticas Públicas Têm Efeito Limitado? 1. Ambiente Sociedade 20, 19–38. https://doi.org/10.1590/1809-4422asoc0232r1v2042017.

Gama, L.H., Oliveira, M.J., Silva, T.C., Neves, S., Dias, G., 2019. Dinâmica de Uso do Solo e sua Relação com os Focos de Calor na Área de Preservação Ambiental Triunfo Do Xingu-PA. Enciclopedia Biosfera 16 (29).
González, S., García, S., Del Ser, J., Rokach, L., Herrera, F., 2020. A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities. Inf. Fusion 64, 205–237. https://doi.org/10.1016/j.inffus.2020.07.007.
Hang, H.T., Mallick, J., Alqadhi, S., Bindajam, A.A., Abdo, H.G., 2024. Exploring forest fire susceptibility and management strategies in Western Himalaya: integrating ensemble machine learning and explainable AI for accurate prediction and comprehensive analysis. Environmental Technology & Innovation 35, 103655. https://doi.org/10.1016/j.eti.2024.103655.
Harrison, M., 2020. Machine Learning–Guia de referência rápida: trabalhando com dados estruturados em Python. Novatec Editora.
Hastie, T., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York. https://doi.org/10.1007/978-0-387-21606-5.
Ho, T.K., 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1. IEEE, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994, 278-282.
IBGE – Brazilian Institute of Geography and Statistics, 2020. Legal Amazon. Available at: https://www.ibge.gov.br/geociencias/cartas-e-mapas/mapas-regionais/15819-amazonia-legal.html?=&t=o-que-e. (Accessed 6 July 2023).
INPE – National Institute for Space Research, 2023. Monitoring system – wildfire database. Available at: http://queimadas.dgi.inpe.br/queimadas/bdqueimadas. (Accessed 6 July 2023).
Jain, P., Coogan, S.C., Subramanian, S.G., Crowley, M., Taylor, S., Flannigan, M.D., 2020. A review of machine learning applications in wildfire science and management. Environ. Rev. 28 (4), 478–505. https://doi.org/10.1139/er-2020-0019.
Jenks, G.F., Caspall, F.C., 1971. Error on choroplethic maps: definition, measurement, reduction. Ann. Assoc. Am. Geogr. 61 (2), 217–244. https://doi.org/10.1111/j.1467-8306.1971.tb00779.x.
Juvanhol, R.S., Fiedler, N.C., Santos, A.R.D., Peluzio, T.M., Silva, W.B.D., Pinheiro, C.J.G., Sousa, H.C.P.D., 2023. Use of machine learning as a tool for determining fire management units in the brazilian atlantic forest. An Acad. Bras Ciências 95 (2), e20201039. https://doi.org/10.1590/0001-3765202320201039.
Kalantar, B., Ueda, N., Idrees, M.O., Janizadeh, S., Ahmadi, K., Shabani, F., 2020. Forest fire susceptibility prediction based on machine learning models with resampling algorithms on remote sensing data. Rem. Sens. 12 (22), 3682. https://doi.org/10.3390/rs12223682.
Libonati, R., DaCamara, C.C., Setzer, A.W., Morelli, F., Melchiori, A.E., 2015. An algorithm for burned area detection in the Brazilian Cerrado using 4 μm MODIS imagery. Rem. Sens. 7 (11), 15782–15803. https://doi.org/10.3390/rs71115782.
Loudermilk, E.L., O'Brien, J.J., Goodrick, S.L., Linn, R.R., Skowronski, N.S., Hiers, J.K., 2022. Vegetation's influence on fire behavior goes beyond just being fuel. Fire Ecology 18 (1), 9. https://doi.org/10.1186/s42408-022-00132-9.
MAPBIOMAS PROJECT, 2023. Collection 8 of the annual land cover and land use maps of Brazil (1985-2022). MapBiomas Data 1. https://doi.org/10.58053/MapBiomas/VJIJCL.
Marcílio, W.E., Eler, D.M., 2020. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, pp. 340–347. https://doi.org/10.1109/SIBGRAPI51738.2020.00053.
Michael, Y., Helman, D., Glickman, O., Gabay, D., Brenner, S., Lensky, I.M., 2021. Forecasting fire risk with machine learning and dynamic information derived from satellite vegetation index time-series. Sci. Total Environ. 764, 142844. https://doi.org/10.1016/j.scitotenv.2020.142844.
Mohajane, M., Costache, R., Karimi, F., Pham, Q.B., Essahlaoui, A., Nguyen, H., Laneve, G., Oudija, F., 2021. Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area. Ecol. Indicat. 129, 107869. https://doi.org/10.1016/j.ecolind.2021.107869.
Moreira, P.A.G., Mendes, T.A., Santos, D.F.D., 2020. Avaliação de locais potenciais para instalação de torres de observação para prevenção de risco de incêndios florestais. Ciência Florest. 30, 1266–1282. https://doi.org/10.5902/1980509839686.
Nami, M.H., Jaafari, A., Fallah, M., Nabiuni, S., 2018. Spatial prediction of wildfire probability in the Hyrcanian ecoregion using evidential belief function model and GIS. Int. J. Environ. Sci. Technol. 15, 373–384. https://doi.org/10.1007/s13762-017-1371-6.
Neary, G.D., Leonard, M.J., 2020. Effects of fire on grassland soils and water: a review. Grasses and grassland aspects 1–22. https://doi.org/10.5772/intechopen.90747.
Nóbrega, L.O., Lazzarini, G.M.J., Viola, M.R., Batista, A.C., De Carvalho, E.V., Giongo, M., 2018. Forest Fire susceptibility index for assessing the history of fire occurrences in the indigenous land of Kraholândia, Brazil. Advances in Forestry Science 5 (2), 325–332. https://doi.org/10.34062/afs.v5i2.5841.
Oliveira, V.F.R., Silva, E.D.S., Vick, E., Silva, B., 2020. Geoprocessamento aplicado ao mapeamento de risco a incêndios. Revista Brasileira de Geografia Física 13 (3), 1194–1212. https://doi.org/10.26848/rbgf.v13.3.p1194-1212.
Ozenen Kavlak, M., Cabuk, S.N., Cetin, M., 2021. Development of forest fire risk map using geographical information systems and remote sensing capabilities: Ören case. Environ. Sci. Pollut. Control Ser. 28 (25), 33265–33291. https://doi.org/10.1007/s11356-021-13080-9.
Pang, Y., Li, Y., Feng, Z., Feng, Z., Zhao, Z., Chen, S., Zhang, H., 2022. Forest fire occurrence prediction in China based on machine learning methods. Rem. Sens. 14 (21), 5546. https://doi.org/10.3390/rs14215546.
Pasqualotto, N., Sena, M.M., 2017. Impactos ambientais urbanos no Brasil e os caminhos para cidades sustentáveis. Revista Educação Ambiental em Ação 61.
Pham, B.T., Jaafari, A., Avand, M., Al-Ansari, N., Dinh Du, T., Yen, H.P.H., Phong, T.V., Nguyen, D.H., Van Le, H., Mafi-Gholami, D., Prakash, I., Thuy, H.T., Tuyen, T.T., 2020. Performance evaluation of machine learning methods for forest fire modeling and prediction. Symmetry 12 (6), 1022. https://doi.org/10.3390/sym12061022.
Phelps, N., Woolford, D.G., 2021. Comparing calibrated statistical and machine learning methods for wildland fire occurrence prediction: a case study of human-caused fires in Lac La Biche, Alberta, Canada. Int. J. Wildland Fire 30 (11), 850–870. https://doi.org/10.1071/WF20139.
Porter, M.D., Reich, B.J., 2012. Evaluating temporally weighted kernel density methods for predicting the next event location in a series. Spatial Sci. 18 (3), 225–240. https://doi.org/10.1080/19475683.2012.691904.
Pourghasemi, H.R., Kariminejad, N., Amiri, M., Edalat, M., Zarafshar, M., Blaschke, T., Cerda, A., 2020a. Assessing and mapping multi-hazard risk susceptibility using a machine learning technique. Sci. Rep. 10 (1), 3203. https://doi.org/10.1038/s41598-020-60191-3.
Pourghasemi, H.R., Gayen, A., Lasaponara, R., Tiefenbacher, J.P., 2020b. Application of learning vector quantization and different machine learning techniques to assessing forest fire influence factors and spatial modelling. Environ. Res. 184, 109321. https://doi.org/10.1016/j.envres.2020.109321.
Pourtaghi, Z.S., Pourghasemi, H.R., Rossi, M., 2015. Forest fire susceptibility mapping in the Minudasht forests, Golestan province, Iran. Environ. Earth Sci. 73 (4), 1515–1533. https://doi.org/10.1007/s12665-014-3502-4.
Pourtaghi, Z.S., Pourghasemi, H.R., Aretano, R., Semeraro, T., 2016. Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. Ecol. Indicat. 64, 72–84. https://doi.org/10.1016/j.ecolind.2015.12.030.
Rácz, A., Bajusz, D., Héberger, K., 2021. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. Molecules 26 (4), 1111. https://doi.org/10.3390/molecules26041111.
Rizzatti, M., Lampert Batista, N., Cezar Spode, P.L., Bouvier Erthal, D., Mauro de Faria, R., Volpato Scotti, A.A., Trentin, R., Petsch, C., Turba Costa, I., Quoos, J.H., 2020. Mapeamento da COVID-19 por meio da densidade de Kernel. Metodologias E Aprendizado 3, 44–53. https://doi.org/10.21166/metapre.v3i0.1312.
Rodrigues, J.A., Libonati, R., Pereira, A.A., Nogueira, J.M., Santos, F.L., Peres, L.F., Setzer, A.W., 2019. How well do global burned area products represent fire patterns in the Brazilian Savannas biome? An accuracy assessment of the MCD64 collections. Int. J. Appl. Earth Obs. Geoinf. 78, 318–331. https://doi.org/10.1016/j.jag.2019.02.010.
Rokach, L., 2010. Pattern classification using ensemble methods. World scientific 75. https://doi.org/10.1142/11325.
Rosan, T.M., Anderson, L.O., Vedovato, L., 2017. Avaliação da origem de focos de calor em anos de extremos climáticos na Amazônia brasileira. Rev. Bras. Cartogr. 69 (4), 731–741. https://doi.org/10.14393/rbcv69n4-44331.
Rubí, J.N., de Carvalho, P.H., Gondim, P.R., 2023. Application of machine learning models in the behavioral study of forest fires in the Brazilian Federal District region. Eng. Appl. Artif. Intell. 118, 105649. https://doi.org/10.1016/j.engappai.2022.105649.
Seddouki, M., Benayad, M., Aamir, Z., Tahiri, M., Maanan, M., Rhinane, H., 2023. Using machine learning coupled with remote sensing for forest fire susceptibility mapping. Case study Tetouan province, Northern Morocco. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 48, 333–342. https://doi.org/10.5194/isprs-archives-XLVIII-4-W6-2022-333-2023.
Shmuel, A., Heifetz, E., 2023. A machine-learning approach to predicting daily wildfire expansion rate. Fire 6 (8), 319. https://doi.org/10.3390/fire6080319.
Simon, M.F., Grether, R., de Queiroz, L.P., Skema, C., Pennington, R.T., Hughes, C.E., 2009. Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. Proc. Natl. Acad. Sci. USA 106 (48), 20359–20364. https://doi.org/10.1073/pnas.090341010.
Sousa, R.D.P., Silva, R.C.D., Miranda, K.A.F.D.N., Amaral Neto, M.A., 2016. Governança socioambiental na Amazônia: Agricultura familiar e os desafios para a sustentabilidade em São Félix do Xingu–Pará. Instituto Internacional de Educação do Brasil.
Souza Jr., C., Fonseca, A., Nunes, S., Salomão, R., Ribeiro, J., Martins, H., 2018. Desmatamento em áreas protegidas. O estado das áreas protegidas. Imazon 2–12.
Tian, Y., Wu, Z., Li, M., Wang, B., Zhang, X., 2022. Forest fire spread monitoring and vegetation dynamics detection based on multi-source remote sensing images. Rem. Sens. 14 (18), 4431. https://doi.org/10.3390/rs14184431.
Tonini, M., D'Andrea, M., Biondi, G., Degli Esposti, S., Trucchia, A., Fiorucci, P., 2020. A machine learning-based approach for wildfire susceptibility mapping. The case study of the Liguria region in Italy. Geosciences 10 (3), 105. https://doi.org/10.3390/geosciences10030105.
Vadrevu, K.P., Eaturu, A., Badarinath, K., 2010. Fire risk evaluation using multicriteria analysis—a case study. Environ. Monit. Assess. 166, 223–239. https://doi.org/10.1007/s10661-009-0997-3.
Xie, Y., Peng, M., 2019. Forest fire forecasting using ensemble learning approaches. Neural Comput. Appl. 31 (9), 4541–4550. https://doi.org/10.1007/s00521-018-3515-0.
Yan, H., Liu, G., 2021. Fire's effects on grassland restoration and biodiversity conservation. Sustainability 13 (21), 12016. https://doi.org/10.3390/su132112016.