Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project

Evaluating the Changes in Temporal Dynamics of Climate Variables and Wildfires in Pantanal of Tropical South America

by

Zihui Ge

zihui.ge20@imperial.ac.uk
GitHub Login: acse-zg120

Supervisors:

Dr. Minerva Singh
Dr. Adriana Paluszny

27th August 2021

Abstract

The wildfire dynamics in Tropical South America are critically important to understand changes in ecosystems, and for helping fire prevention, control, and resource allocation. In this study, local regression, second-order regression, and the Mann-Kendall Trend test were applied to analyse the trend of fire occurrences and precipitation in the Pantanal biome in South America. The result confirmed the existence of a negative correlation between fire occurrences and precipitation. On the other hand, two multi-class classification methods, decision tree and random forest algorithms in machine learning were applied to predict and classify fires into five categories: category 1 ($<$100MW), category 2 (100-500MW), category 3 (500-1000MW), category 4 (1000-1500MW) and category 5 ($\geq$ 1500MW). Fire information in Pantanal was collected using Moderate Resolution Imaging Spectroradiometer (MODIS) MCD14ML from 2000 to 2018. The three related climate factors were used as dependent variables for modelling, which are precipitation, solar radiation, and temperature. The model using the decision tree algorithm had an accuracy of 88.54% while the model had an accuracy of 88.28% using the random forest algorithm, which indicated the performances of both models were excellent. Also, the most important factor found in this study was the temperature in the three climate variables.

Keywords: fire occurrences, regression, Mann-Kendall Trend test, multi-class classification, machine learning, decision tree, random forest

1 Introduction

Fire has affected ecosystems and interfered with the evolution of species on a global scale (Moreira de Araújo, Ferreira & Arantes 2012). Moreover, many blazing fires have ravaged forest resources as well as threatening public safety and biodiversity (Martell 2007). Burned land has increased by more than 420 Mha every year, mainly occurred in savannas and grasslands at a global level (Giglio, Boschetti, Roy, Humber & Justice 2018). Meanwhile, each year billions of dollars are used for fire control and management to mitigate the damage of wildfires (Jain, Coogan, Subramanian, Crowley, Taylor & Flannigan 2020).

While being the country with largest land mess in South America, Brazil also has the greatest amount of biodiversity dominated by tropical rain forests and grasslands (Bond, Woodward & Midgley 2005). However, due to agricultural activities, climate change and human activities, the number of fires has increased in Brazil in the last twenty years (de Oliveira-Junior, Teodoro, da Silva Junior, Baio, Gava, Capristo-Silva, de Gois, Correia Filho, Lima, de Barros Santiago et al. 2020) (Caúla, de Oliveira-Júnior, de Gois, Delgado, Pimentel & Teodoro 2017). Furthermore, the frequency of destructive fire occurrences has risen significantly in the past few years in some biomes like Pantanal with no long history of fires in Brazil (de Oliveira-Junior et al. 2020). To understand the association of fires, it is essential to investigate the interaction of factors leading to fire occurrences. Also, knowing the trends of fire occurrences in spatial and temporal terms allows us to predict fires dynamics for fire management and damage control.

There are two aspects of determining fires in spatial and temporal terms: trends analysis and fire prediction. A traditional way to analyse fire dynamics and investigate the correlations between climate variables and fire occurrences is to use statistical methods such as linear regression and logistic regression. In contrast to statistical methodologies, many ongoing algorithms use machine learning algorithms, which provide more accurate predictions on fire trends and modelling (Chuvieco, Aguado, Yebra, Nieto, Salas, Martín, Vilar, Martínez, Martín, Ibarra et al. 2010). Machine learning algorithms build models based on learning algorithms, and make predictions with independent and dependent variables (Barreto & Armenteras 2020) from extracted data resources. Wei and Wang et al. built a model using a random forest algorithm in machine learning in order to predict the probability of fire occurrences with related factor information in topography, human activities, vegetation, and climate variables (Barreto & Armenteras 2020). The model has excellent performance with an accuracy of 94%, and it can be used in fire management and control as base information for identifying geolocations in which attention needs to prioritise (Barreto & Armenteras 2020).

In many direct or potential factors in fire occurrences, climate variables (Brown, Hebda, Conder, Golinski, Hawkes, Schoups & Hebda 2017) are commonly used to predict fire occurrence, such as precipitation, temperature, wind speed, solar radiation, and continuous dry weather (Parisien, Miller, Parks, DeLancey, Robinne & Flannigan 2016). Many researchers have been interested in the analysis of fire geolocations in Brazilian biomes. Teodoro et al. have already shown that fire occurrences are associated with spatial precipitation variability in Brazilian biomes (Cerrado, Mata Alântica, and Pantanal) (Teodoro, de Oliveira-Júnior, Da Cunha, Correa, Torres, Bacani, Gois & Ribeiro 2016). Thus, this study will only focus on temporal variability to investigate the association between climate variables and fire occurrences.

2 Study Area and Material

2.1 Study Area

The study area belongs to a biome in South America that encompasses the largest tropical wetland (150,355 km^2) and flooded grassland in the world (Alho, Mamede, Benites, Andrade & SEPÚLVEDA 2019). Pantanal is mainly located in Brazil (the state of Mato Grosso do Sul), and a portion of Bolivia and Paraguay, lying between 15° 00' S and 22° 30' S of parallels, and 55° 00' W and 57° 00' W of meridians as shown in Figure 1 (Alho et al. 2019). The climate of Pantanal is tropical with two noticeable seasons: wet season (October to March) and dry season (April to September) (Da Silva & Girard 2004).

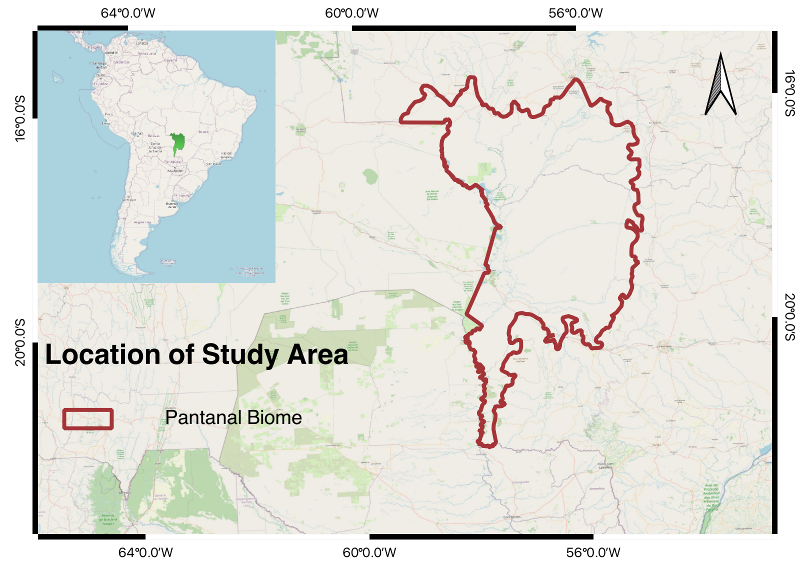


Figure 1: Location of Study Area

2.2 Fire Data Source

The first step in evaluating wildfire dynamics involves the analysis of fire information regarding geolocations and fire intensity. Moderate Resolution Imaging Spectroradiometer (MODIS) MCD14ML provides the global fire information combined with Terra and Aqua satellites on a daily basis at 1km x 1km spatial resolution (*Fire Information for Resource Management System (FIRMS)* 2021) from November 2000 to 2020 inclusively (download link `https://firms.modaps.eosdis.nasa.gov/download/`). MCD14ML data is processed by the Science Computing Facility of NASA and distributed by

FIRMS, giving a projection in the WGS84 reference coordinate system (*Fire Information for Resource Management System (FIRMS)* 2021). Fire Radiative Power (FRP) in MCD14ML represents the fire emitted energy of radiation in megawatts (Roberts & Giglio 2021), a major parameter of evaluating fire occurrences and intensity classifications. This study will focus on fire information from 2000 to 2018 inclusively in a continuous-time series with climate variables at the same time range.

2.3 Climate Data Source

WorldClim provides a variety of global climate and weather data in a full spatial resolution (Fick & Hijmans 2017). In this study, climate variable information was extracted from the monthly historical climate data (WorldClim version 2.1) at 30 seconds ($1km^2$) resolution (Fick & Hijmans 2017) for the dry season (April-September) between 1970 to 2000 observation period (download link: `https://www.worldclim.org/data/index.html`). Five climatic factors relating to the fire occurrences were evaluated that are solar radiation (kJm^{-2}/day), average temperature ($°C$), precipitation (mm), water vapour pressure (kPa), and wind speed (ms^{-1}).

3 Software and Resources

3.1 QGIS

QGIS is an Open Source software of GIS (Geographic Information System) available on most operating systems such as Linux, Unix, Windows, and Mac OS (*QGIS - Open Source Geographic Information System* 2021). QGIS supports a wide range of database formats and capable functions to edit, view, compose maps and analyse projects (*QGIS - Open Source Geographic Information System* 2021). QGIS version 3.18.3 provides the most stable functionalities and features on MAC OS for this study, and robust visualising features of climate maps for the study area (*QGIS - Open Source Geographic Information System* 2021).

3.2 R and Packages

In this study, R was used as the primary programming language for statistical tests and computation. RStduio a free software that includes more than fifteen thousand additional packages alongside an essential set of pre-installed R packages for data analysis and statistical procedures (Kurt 2020) (Muenchen 2019) (Tippmann 2015). Dplyr is an R package of a grammar set of data manipulation for users to fast and consistently solve common data operations (Wickham, François, Henry, Müller & RStudio 2018). Maptools includes a set of manipulation tools for geographic data and also allows to exchange spatial objects on interface wrappers (Roger Bivand & Pebesma 2021). Raster contains functions of basic manipulations of spatial data that analyses and models for raster and vector data (Robert J. Hijmans & Sumner 2021). The prior required packages to install are listed in Table 1.

Package	Version	Package	Version
corrplot	0.90	dismo	1.3-3
dplyr	1.0.7	ggplot2	3.3.5
Kendall	2.2	maptools	1.1-1
pROC	1.17.0.1	randomForest	4.6-14
raster	3.4-13	rattle	5.4.0
rgdal	1.5-23	rgoes	0.5-5
ROCR	1.0-11	rpart	4.1-15
rpart.plot	3.1.0	zoo	1.8-9

Table 1: R Package List

4 Methodology

4.1 Data Preprocess

In order to reduce the time to process the programme, only areas contained in the study's parameter were extracted from downloaded datasets via RStudio. This procedure applied to all fire data and climate data for the purpose of analysing dynamic fire trends over climate variables. All the climate datasets for each variable were downloaded in Tag Image File Format (TIFF), which is a standard format for storing spatial raster images (Murray & VanRyper 1996). For each climate variable, every single image represents the monthly climate data in each year from 2000 to 2018 for statistical analysis (twelve months in twelve images in nineteen years), and the average monthly climate data from 1970 to 2000 (twelve images contain the average values for 30 years corresponding to twelve months for each variable) for automatic learning. However, the fire information is presented in vector data format (Shapefile) in MODIS MCD14ML. For fire information, a single Shapefile file including all fire occurrences from 2000 to 2020 was used. The next step is to ascertain that the datasets from the individual formats are identical without losing any further information. To achieve this, the 'raster' package is able to read and edit raster data for large files without attempting to read all image information in memory on RStudio (Version 1.4.1) (Robert J. Hijmans & Sumner 2021). In addition, an extra layer of grid cells at $1km^2$ resolution was intersected with fire layers and all climatic layers so that the study area was split into 1km x 1km pixels with unique id numbers. Each grid cell has a fire parameter corresponding to climate variables in the exact location of the study area. At this point, all spatial layers of the study area were converted to the Shapefile format using a unique ID number, and ready for further analysis and modelling via RStudio.

4.2 Climate Variable Selection

Since some variables decrease prediction accuracy and cost more calculation in computation, it is always essential to select appropriate input variables for modelling prediction. Therefore, in order to carefully select variables for classification models in machine learning, the variance inflation factor (VIF) and correlation matrix were used for this study (Martínez-Álvarez, Reyes, Morales-Esteban & Rubio-Escudero 2013) (Barreto & Armenteras 2020). When multiple variables are highly correlated in a model, it might result in a less accurate prediction, which is also called multicollinearity (James, Witten, Hastie & Tibshirani 2017). In a statistical model, VIF is the quotient of the variance of a model that contains multiple input parameters at the variance of the model includes only one input parameter (Akinwande, Dikko & Samson 2015). It provides an index that evaluates how much an estimated regression coefficient variance increases with predictor correlation (Akinwande et al. 2015). Specifically, in this study, the VIF was used to evaluate the correlation between climate variables with one of the other climate variables in the R programming environment. A strong correlation between climate variables will be represented by a VIF value, which is generally greater than ten (Wu & Zhang 2013). As Table 2 shows, there is no doubt that the wind variable should be excluded from climate variables according to its VIF value in the modelling process.

Variable	VIF
Precipitation	1.9240
Solar Radiation	10.4025
Temperature	5.6874
Water Vapour Pressure	11.4655
Wind	17.0577

Table 2: Variance Inflation Factor

Additionally, the Pearson correlation analysis was used to evaluate the association of linear correlation between two sets of variables (Barreto & Armenteras 2020). The Pearson correlation coefficient is normalised to a range of between -1 and 1 (Schober, Boer & Schwarte 2018), where a positive coefficient represents a change in the same direction, and a negative coefficient represents a change in the opposite direction (*pearson's correlation - statstutor* 2008). Ideally, only a coefficient under a perfect correlation will equal to 1 (Schober et al. 2018). When a predicted value of a correlation coefficient is greater than 0.7 it will be considered as an indicator for a pair of variables being highly correlated, which will affect prediction accuracy and the modelling process (Dormann, Elith, Bacher, Buchmann, Carl, Carré, Marquéz, Gruber, Lafourcade, Leitão & et al. 2012). In Figure 2, the correlation matrix suggests a positive value of 0.88 between temperature and water vapour pressure. Since temperature is a more standard variable in fire evaluation and studies (Chowdhury & Hassan 2015), water vapour pressure was removed from the prediction variables. Additionally, a negative value of -0.95 between solar radiation data and wind data demonstrates a highly correlated relationship in Figure 2, which also proved the observed results of the VIF values in Table 2.

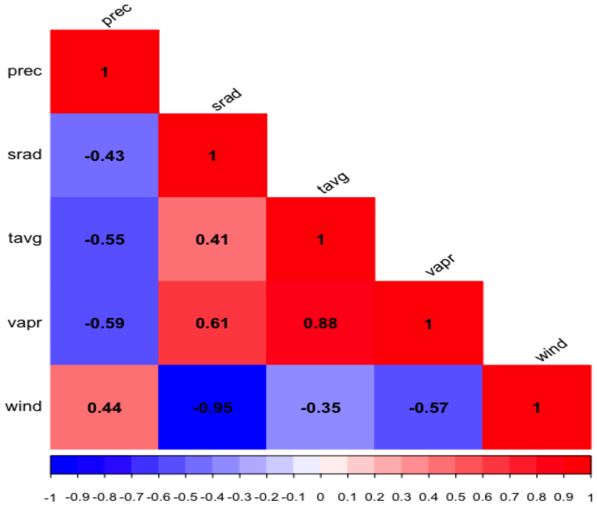


Figure 2: Correlation Matrix.

4.3 Statistical Analysis Methods

The Mann-Kendall trend test was used to investigate the trend between fire occurrences and climate variables in the time series. This is a test based on the Kendall rank correlation in a time series $z(t)$ and time t for monotonic trend. The null hypothesis is accepted at the beginning that no monotonic trend is found in the series. This non-parametric test not only considers the relative occurrence of success and the independence of data instead of the actual values, but also maintains the same distribution of probability (Pettitt 1979). Only precipitation information was considered in the Mann-Kendall trend test in this study because it is a standard and commonly used variable in research. Package 'trend' provides a collection of non-parametric tests and changing point detection in the R language environment. Subsequently, a local regression and a second regression model for each fire occurrence and precipitation dataset were created, and linear regression for fire occurrences data versus precipitation data was also established to evaluate the trends between two parameters.

4.4 Machine Learning Modelling

The data was split into 80% for training set and 20 % for test set where training set was used for training models, with the test set being used for validating accuracy and performance of models. For the climate data from 1970 to 2000 and fire information from 2000 to 2020, a layer of grid cells was intersected in the previous step containing 97,504 points, with 78,003 points for the training set and 19,501 points for the test set for each class. Ichoku et al. proposed five classes of fire based on FRP values for each point, in which category 1 (<100MW), category 2 (100-500MW), category 3 (500-1000MW), category 4 (1000-1500MW) and category 5 ($\geq$ 1500MW) identify fire intensity (Ichoku, Giglio, Wooster & Remer 2008). Five fire categories were converted into the label values from 0 to 4 during training, where each pixel represents a number corresponding to a category (Barreto & Armenteras 2020).

A Decision Tree is a tree structure classification algorithm in machine learning which is composed of root nodes, leaf nodes and branches (Jenhani, Amor & Elouedi 2008). The main feature of the decision tree consists of the building procedure and classification procedure that allows users to induce rules at each node of a tree in order to classify new branches (Jenhani et al. 2008). To build a decision tree model, recursive partitioning and regression tree (Rpart) in R were chosen to select a suitable test attribute for each node of the decision tree using index splitting function to a given factor training set (Therneau, Atkinson & Ripley 2019). In order to classify a fire category for a new instance, the classification procedure searches the decision tree from the root following the observed values in the tree corresponding to interior tree nodes until it reached a leaf (Jenhani et al. 2008). A decision tree that grows beyond a certain level could cause overfitting, therefore it is important to find the most adequate value for the complexity parameter. The best complexity parameter can be found through 'cptable', where the x-val relative error has the minimum value in the decision tree model. The last step before predicting the model is to prune the model by applying the complexity parameter found from 'cptable'.

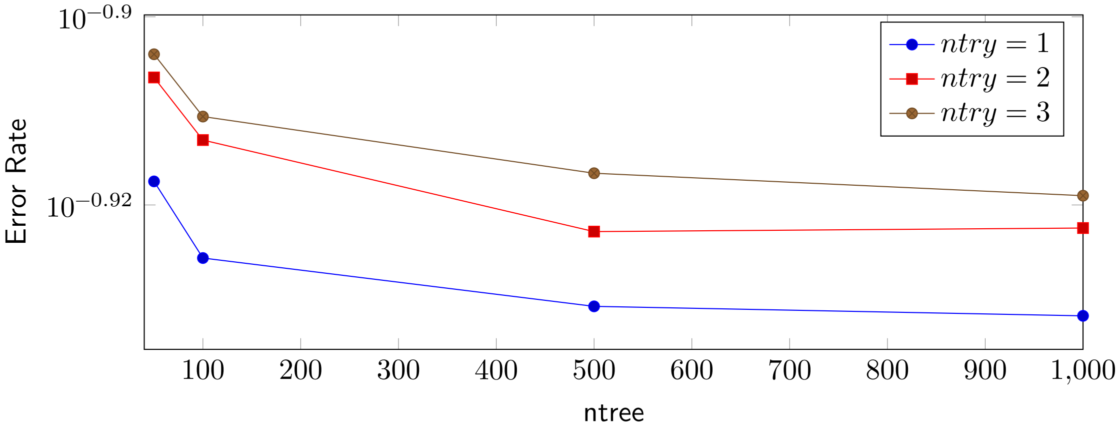


Figure 3: Results of the validation process in a group of different 'ntree' and 'ntry' values

Compared to the decision tree algorithm, the random forest algorithm reduces a large amount of branches for individual regressions by averaging the predictive values. Therefore, it is similar to the decision tree algorithm in that the same amount of elements were randomly taken for training and test sets. In a random forest algorithm model, 'ntree' and 'ntry' are two important model parameters (Tonini, D'Andrea, Biondi, Degli Esposti, Trucchia & Fiorucci 2020), which respectively represents the number of trees, and the number of variables randomly selected for splitting at each node of tree (Barreto & Armenteras 2020). The key objective of the random forest algorithm is to use a group of random subsets and choose variables in each node and combine the decision trees all together (Arpaci, Malowerschnig, Sass & Vacik 2014). A random forest model with parameters in 'ntree' of 500 and 'ntry' of 1 was initially built and simulated to investigate the features and performances of the model. Since 'ntree' and 'ntry' are two important parameters to a random forest model, and vital to identify the performance of a model avoiding overfitting, it is necessary to apply a variety of options of 'ntree' and 'ntry' values to select the best combination of 'ntree' and 'ntry' that provides a best accuracy and prediction. A group of 'ntree' values (50, 100, 500, and 1,000) and three levels of 'ntry' (1, 2, and 3) were tested with the maximum 'ntree' value being 1,000 to maintain a stable output from the research by Breiman (Breiman 2001) on "Random Forests" chapter. In Figure 3, the results of validation process using different startup values for 'ntree' and 'ntry' confirmed that the optimal combination of 'ntree' and 'ntry' was when 'ntree' = 730 and 'ntry' =1 with the lowest error rate of 0.116.

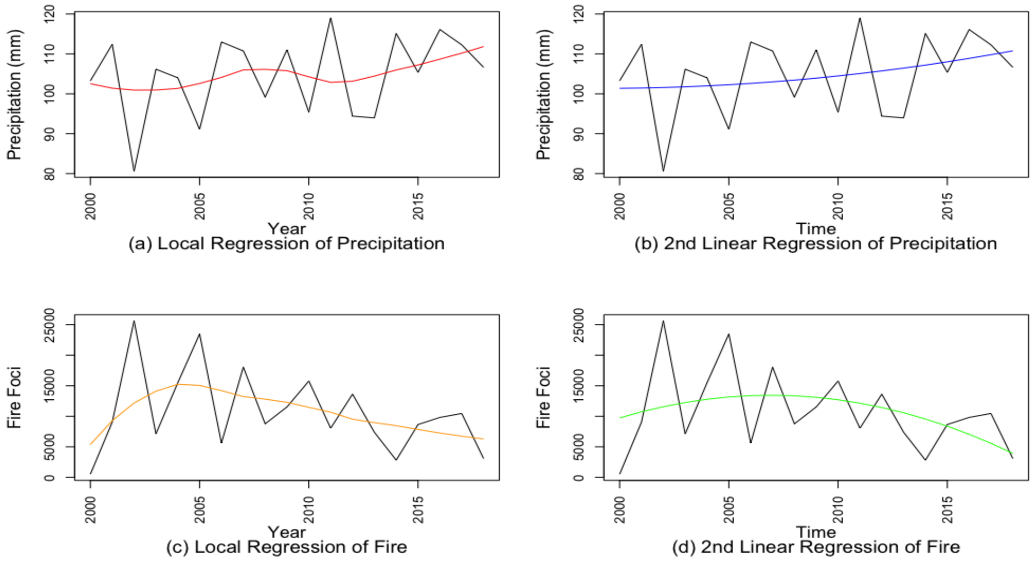
At this point, a decision tree model and a random forest model were built. Model validation is an essential process to assess the model relevance and model accuracy of predicted results (Adelabu, Mutanga & Adam 2015). Apart from the 78,003 points for the training set, 19,501 points were used for model validation. For the decision tree model, a pruned model was tested through five fire categories, and the predicted results were obtained for further accuracy and error rate calculations. In a random forest model, the Mean Decrease Accuracy was used to express the accuracy that the model losses when a particular value of a variable is randomly changed from the previous observation (Martinez-Taboada & Redondo 2020). Therefore, the Mean Decrease Accuracy defines the importance of each variable, which is very commonly used in a random forest model (Breiman 2001). The MDA plot for three climate variables is shown in Figure 5, using the importance function in the randomForest package.

As an indicator of evaluating a model, a confusion matrix associates with the predicted and observed classifications as in an N*N matrix (Visa, Ramsay, Ralescu & Van Der Knaap 2011). Each row represents the instances in an observed class and the instances in a predicted class are represented at each column, while the diagonal elements represent the number of instances when the predicted class equals the observed class (Powers 2008). The main idea of using a confusion matrix is to evaluate the model accuracy, in which only the correct predicted instances were taken into account. The accuracy in a confusion matrix with multi-class can be calculated by the correct predicted instances, divided by the total predicted instances (Piryonesi & El-Diraby 2020) (Metz 1978).

5 Results

5.1 Fire occurrences and Precipitation Through Regression Analysis

We applied local regression and second-order regression to annual precipitation data and fire occurrences during the period between 2000 and 2018 as shown in Figure 4. In Figure 4, (a) and (b) show an increasing trend of precipitation from 103 mm (2000) to 110 mm (2018). It clearly shows a negative correlation between average annual precipitation and fire occurrence between 2000 to 2018.



(a) Local Regression of Precipitation

(b) 2nd Linear Regression of Precipitation

(c) Local Regression of Fire

(d) 2nd Linear Regression of Fire

Figure 4: Regression Plot

5.2 Mann-Kendall Trend Test

The Mann-Kendall Trend test was applied to fire occurrence and precipitation data in a time series to verify the trend in the study area. The results for fire occurrences and precipitation data are shown in Table 3, where S is Kendall Score, varS is variance of S, and tau is Kendall's tau statistic.

```
Mann-Kendall Test                               Mann-Kendall Test

data: fire foci                                 data: precipitation
z = -0.83965, n = 19, p-value = 0.40            z = 1.0496, n = 19, p-value = 0.2939
alternative hypothesis:                         alternative hypothesis:
true S is not equal to 0                        true S is not equal to 0
sample estimates:                               sample estimates:
         S              varS       tau                   S            varS      tau
    -20.0000        817.0000   -0.1462             31.0000         817.000   0.1813
```

Table 3: Mann-Kendall Trend Test of Precipitation Data and Fire Occurrences

5.3 Mean Decrease Accuracy

The most important variable of the random forest model as Figure 5 shows is the average monthly temperature, which has the highest value of 57.44. Moreover, the other two variables follow very similar values, 55.73 for precipitation and 54.06 for solar radiation.

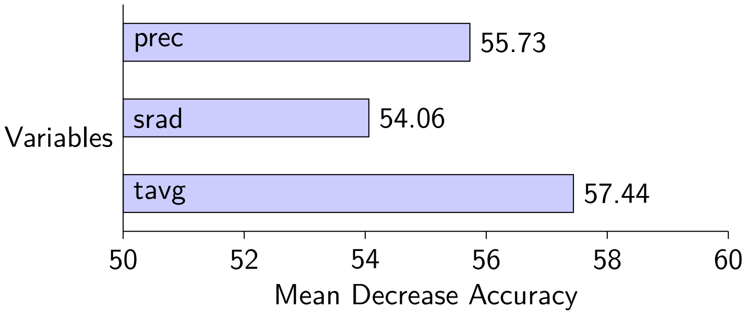


Figure 5: Mean Decrease Accuracy

5.4 Confusion Matrix and Model Accuracy

The prediction results for the decision tree and random forest models are represented in Table 4. Labels 0 to 4 on the first row and column on each confusion matrix represent the fire categories 1 to 5 respectively. The correct predicted elements are on the diagonals, with the predicted elements being identical to the observed elements from the test set. In Table 4a and 4b, the majority of the predicted elements were classified as fire category 1 (label 0) in all observed data, and a minority of the observed elements of fire category 1 (label 0) were identified as fire category 2 (label 1).

The accuracy for each model is 88.54% and 88.28%, which was calculated from the proportion of correct predicted points against the total points in the test set. We also calculated error rates for each fire category with the rate of the actual incorrect predicted points over the total predicted points in each fire category (Table 5).

	Predicted				
Observed	0	1	2	3	4
0	17235	40	0	0	0
1	2095	31	0	0	0
2	79	0	0	0	0
3	16	1	0	0	0
4	4	1	0	1	0

(a) Confusion matrix for decision tree

	Predicted				
Observed	0	1	2	3	4
0	17068	205	1	0	1
1	1986	137	1	1	1
2	67	11	0	0	1
3	13	1	0	1	1
4	4	0	0	1	0

(b) Confusion matrix for random forest

Table 4: Confusion matrix a and b.

Model	Accuracy	Error Rate	ERR_1	ERR_2	ERR_3	ERR_4	ERR_5
Decision Tree	88.54%	11.46%	0.23%	98.54%	100%	100%	100%
Random Forest	88.28%	11.72%	1.18%	93.32%	98.73%	93.75%	100%

Table 5: Model Accuracy and Error Rate

6 Discussion

In Figure 4 (a), the curve falls between 2000 and 2003, and the average annual precipitation rises and drops again from 2003 to 2011, where it reaches a local peak point in 2008. After 2011, the average annual precipitation rises again to 110 mm. From Figure 4, we can say that the local regression and second-order regression plots indicate a negative correlation between precipitation and fire occurrences from 2000 to 2018. The Mann-Kendall test has also confirmed this phenomenon in Figure 3. The statistic S suggests a decreasing in fire occurrence and an increasing in precipitation, although the magnitude of both slopes are small. The initial MK test assumption of H_0, the null hypothesis, is accepted that model has no monotonic trend (Gilbert 1987) until p-values are greater than 0.05 (Karmeshu 2012). From the results in Table 3, no monotonic trends are seen for precipitation and fire occurrences since p-values are both greater than 0.05.

It is worth noting that precipitation and temperature are associated with El Niño and La Niña phenomenons (Trenberth, Jones, Ambenje, Bojariu, Easterling, Klein Tank, Parker, Rahimzadeh, Renwick, Rusticucci et al. 2007). El Niño–Southern Oscillation (ENSO) is the periodic cycle of unusually warm and cold sea surface temperature every two to seven years (Chen, Morton, Andela, Van Der Werf, Giglio & Randerson 2017). This causes global changes in rainfall and temperature where Brazil is one of the regions most affected during El Niño and La Niña phenomenons (Grimm, Barros & Doyle 2000). Chen et al. tracked six El Niño and six La Niña events from 1997 to 2016 using satellite data concerning burned areas and responses of fire emissions (Chen et al. 2017). Compared to La Niña, decreases in precipitation and terrestrial water storage causes fire emissions to increase by 133% during and after El Niño in the pan-forest. Therefore, investigating the influence of precipitation and temperature corresponding to fire occurrences during El Niño and La Niña phenomenons is necessary for future work.

The Mean Decrease Accuracy graph shows that temperature is the most important variable to the random forest model in Pantanal. It is also important and necessary to discuss the reasons that lead temperature to the most critical climate variable. A rise in temperature would induce an increase in evapotranspiration so that it decreases soil moisture in tropical South America ((Salazar, Nobre & Oyama 2007)). The changing in temperature could cause a shift in species diversity and habitat for insects and forests (Karmeshu 2012). On the other hand, the increased temperature can lead to a heatwave change that is also another challenge to forests or vegetation that is sensitive to heat (Evans & Perschel 2009).

The result of the validation process in different values of 'ntree' and 'ntry' in random forest model

(Figure 3) suggests the best choice for the parameter of the random forest model with maximum accuracy according to the lowest error rate, and represents the performance of models with different parameter configurations to predict fire categories in Pantanal biome during the dry season. According to Figure 3, the configuration values have been chosen in which the number of trees is 730 and the number of variables is 1 for the random forest modelling process. The accuracy of the random forest model has been validated in Table 5, in which 88.28% of the predicted fire occurrences are correct against a total number of points. Meanwhile, the accuracy of the decision tree model is 88.53%. Although, the accuracy in random forest is slightly lower than decision tree by 0.25%, the error rates in Fire category 2 (100-500MW, category 3 (500-1000MW), and category 4 (1000-1500MW) decrease by around 2% - 6%. The result leans toward a better sensitivity and accuracy in category 1 but has poor sensitivity and accuracy in category 2 to 5. From the validation process, we can confirm that both models have better performance on category 1 but with poor accuracy on the other four models.

The biggest reason that low accuracy occurs to a class in machine learning is that the proportional sizes of sample classes are in substantial difference. In this study, there are 86,396 samples of category 1 (<100M) but only 399,72, and 37 samples for category 3, 4, and 5 respectively. The sample ratio of category 1 and category 3 is 217:1, with this number being much lower than category 4 and 5 because they have smaller sample numbers. Concerning this issue, the prediction performance of a classifier will potentially lean towards classes with a larger number of samples in an imbalanced dataset (Kaur, Pannu & Malhi 2019). Accuracy and Performance bias are notably large and prediction behaves differently in variables in imbalanced classes (Kaur et al. 2019). Most of the reasons behind the imbalanced classes concerned the fact that 90% of fires belong to category 1 (<100M) in the most regions globally (Ichoku et al. 2008). From category 3 to 5, only 1% of fires falls into this range (Ichoku et al. 2008). The percentage of fires allocated between category 3 to 5 is only 0.52% out of 97,505 fire occurrences in this study.

For the purpose of balancing the dataset for prediction, the total number of fires for different categories should be balanced, in which each category should have a relatively equal amount of fires. Some balancing methods in machine learning can achieve this. Batista et al. demonstrated a wide range of experimental evaluations involving ten methods to solve an imbalanced class problem where Smote + Tomek and Smote + ENN presented excellent results in order to produce a better-balanced class in machine learning (Batista, Prati & Monard 2004). These two methods used Synthetic Minority Over-sampling Technique (SMOTE) method with two data cleaning methods (Tomek links and ENN). The main objective of using over-sampling techniques and data cleaning methods is to form new minority class samples. This is completed by removing samples from both majority class and minority class, and interpolating between minority class examples where Tomek links tend to remove fewer samples than ENN does. The experiment also suggests that the proposed methods (Smote + Tomek and Smote + ENN) provide better performance on balancing data and more accurate results in machine learning.

7 Conclusion

This study has analysed the trend between climate variables and fire occurrences, and also predicted categories for fire occurrences, supported by statistical analysis methods and robust classification algorithms. A tendency of negative growth is shown in the fire occurrences in Pantanal in this study. The regression models and the Mann-Kendall test proved a noticeable negative correlation between the fire occurrences and the annual monthly precipitation. This study confirms that temperature is the most important variable out of three climate variables (precipitation and solar radiation), directly affecting fire occurrences. In addition, other relevant variables from external databases can also be used in the input of the learning models, such as soil moisture and elevation, to identify the relationship with fire occurrences.

Compared to the study model from Wei and Wang et al., two different classification algorithms were applied in this study, in which a pruned decision tree provides accurate results and better performance. Although different from the binary classification model that classifies instances into one or two classes, the models in this study can classify fire occurrences into five classes according to the fire radiation power. The decision tree and the random forest models represent excellent accuracy for the amount of correct predicted fire occurrences, but the accuracy of each category, especially categories 2 to 5, needs to be improved by a relatively balancing dataset. Also, it is worth mentioning that the location of fires in different categories has not been demonstrated and illustrated in this study, but it is vital to consider the geolocations of fire occurrences for further research. The automatic learning models in this study are demonstrated as a guide to other researchers and people interested in classification algorithms and fire dynamics analysis and prediction. With fire occurrence locating, the geolocation of future fires could be predicted so as to help optimise fire control and management.

References

Adelabu, S., Mutanga, O. & Adam, E. (2015), 'Testing the reliability and stability of the internal accuracy assessment of random forest for classifying tree defoliation levels using different validation methods', *Geocarto International* **30**(7), 810–821.

Akinwande, M. O., Dikko, H. G. & Samson, A. (2015), 'Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis', *Open Journal of Statistics* **05**(07), 754–767.

Alho, C. J., Mamede, S. B., Benites, M., Andrade, B. S. & SEPÚLVEDA, J. J. (2019), 'Threats to the biodiversity of the brazilian pantanal due to land use and occupation', *Ambiente & Sociedade* **22**.

Arpaci, A., Malowerschnig, B., Sass, O. & Vacik, H. (2014), 'Using multi variate data mining techniques for estimating fire susceptibility of tyrolean forests', *Applied Geography* **53**, 258–270.

Barreto, J. S. & Armenteras, D. (2020), 'Open data and machine learning to model the occurrence of fire in the ecoregion of "llanos colombo–venezolanos"', *Remote Sensing* **12**(23), 3921.

Batista, G. E., Prati, R. C. & Monard, M. C. (2004), 'A study of the behavior of several methods for balancing machine learning training data', *ACM SIGKDD explorations newsletter* **6**(1), 20–29.

Bond, W. J., Woodward, F. I. & Midgley, G. F. (2005), 'The global distribution of ecosystems in a world without fire', *New phytologist* **165**(2), 525–538.

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

Brown, K. J., Hebda, N. J., Conder, N., Golinski, K. G., Hawkes, B., Schoups, G. & Hebda, R. J. (2017), 'Changing climate, vegetation, and fire disturbance in a sub-boreal pine-dominated forest, british columbia, canada', *Canadian Journal of Forest Research* **47**, 615–627.

Caúla, R. H., de Oliveira-Júnior, J. F., de Gois, G., Delgado, R. C., Pimentel, L. C. G. & Teodoro, P. E. (2017), 'Nonparametric statistics applied to fire foci obtained by meteorological satellites and their relationship to the mcd12q1 product in the state of rio de janeiro, southeast brazil', *Land Degradation & Development* **28**(3), 1056–1067.

Chen, Y., Morton, D. C., Andela, N., Van Der Werf, G. R., Giglio, L. & Randerson, J. T. (2017), 'A pan-tropical cascade of fire driven by el niño/southern oscillation', *Nature Climate Change* **7**(12), 906–911.

Chowdhury, E. H. & Hassan, Q. K. (2015), 'Operational perspective of remote sensing-based forest fire danger forecasting systems', *ISPRS Journal of Photogrammetry and Remote Sensing* **104**, 224–236.

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Salas, J., Martín, M. P., Vilar, L., Martínez, J., Martín, S., Ibarra, P. et al. (2010), 'Development of a framework for fire risk assessment using remote sensing and geographic information system technologies', *Ecological Modelling* **221**(1), 46–58.

Da Silva, C. J. & Girard, P. (2004), 'New challenges in the management of the brazilian pantanal and catchment area', *Wetlands Ecology and Management* **12**(6), 553–561.

de Oliveira-Junior, J. F., Teodoro, P. E., da Silva Junior, C. A., Baio, F. H. R., Gava, R., Capristo-Silva, G. F., de Gois, G., Correia Filho, W. L. F., Lima, M., de Barros Santiago, D. et al. (2020), 'Fire foci related to rainfall and biomes of the state of mato grosso do sul, brazil', *Agricultural and Forest Meteorology* **282**, 107861.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J. & et al. (2012), 'Collinearity: a review of methods to deal with it and a simulation study evaluating their performance', *Ecography* **36**(1), 27–46.

Evans, A. M. & Perschel, R. (2009), 'A review of forestry mitigation and adaptation strategies in the northeast us', *Climatic Change* **96**(1), 167–183.

Fick, S. E. & Hijmans, R. J. (2017), 'Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas', *International journal of climatology* **37**(12), 4302–4315.

Fire Information for Resource Management System (FIRMS) (2021).
URL: *https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms*

Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L. & Justice, C. O. (2018), 'The collection 6 modis burned area mapping algorithm and product', *Remote sensing of environment* **217**, 72–85.

Gilbert, R. O. (1987), *Statistical methods for environmental pollution monitoring*, John Wiley & Sons.

Grimm, A. M., Barros, V. R. & Doyle, M. E. (2000), 'Climate variability in southern south america associated with el niño and la niña events', *Journal of climate* **13**(1), 35–58.

Ichoku, C., Giglio, L., Wooster, M. J. & Remer, L. A. (2008), 'Global characterization of biomass-burning patterns using satellite measurements of fire radiative energy', *Remote Sensing of Environment* **112**(6), 2950–2962.

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', *Environmental Reviews* **28**(4), 478–505.

James, G., Witten, D., Hastie, T. J. & Tibshirani, R. J. (2017), *An introduction to statistical learning: with applications in R*, Springer.

Jenhani, I., Amor, N. B. & Elouedi, Z. (2008), 'Decision trees as possibilistic classifiers', *International journal of approximate reasoning* **48**(3), 784–807.

Karmeshu, N. (2012), 'Trend detection in annual temperature & precipitation using the mann kendall test–a case study to assess climate change on select states in the northeastern united states'.

Kaur, H., Pannu, H. S. & Malhi, A. K. (2019), 'A systematic review on imbalanced data challenges in machine learning: Applications and solutions', *ACM Computing Surveys (CSUR)* **52**(4), 1–36.

Kurt, H. (2020), 'Frequently asked questions on r'.
URL: *https://CRAN.R-project.org/doc/FAQ/R-FAQ.html*

Martell, D. (2007), 'Forest fire management: current practices and new challenges for operational researchers. in 'handbook of operations research in natural resources'.(eds a weintraub, c romero, t bjørndal, r epstein) pp. 489–509'.

Martinez-Taboada, F. & Redondo, J. I. (2020), 'Variable importance plot (mean decrease accuracy and mean decrease gini).'.
URL: *https://plos.figshare.com/articles/figure/Variable_importance_plot_mean_decrease_accuracy_and_mean_decrease_Gini_/12060105/1*

Martínez-Álvarez, F., Reyes, J., Morales-Esteban, A. & Rubio-Escudero, C. (2013), 'Determining the best set of seismicity indicators to predict earthquakes. two case studies: Chile and the iberian peninsula', *Knowledge-Based Systems* **50**, 198–210.

Metz, C. E. (1978), Basic principles of roc analysis, *in* 'Seminars in nuclear medicine', Vol. 8, Elsevier, pp. 283–298.

Moreira de Araújo, F., Ferreira, L. G. & Arantes, A. E. (2012), 'Distribution patterns of burned areas in the brazilian biomes: An analysis based on satellite data for the 2002–2010 period', *Remote Sensing* **4**(7), 1929–1946.

Muenchen, R. A. (2019), 'The popularity of data science software', *R4statsCom* .

Murray, J. D. & VanRyper, W. (1996), *Encyclopedia of graphics file formats*, OReilly, Associates.

Parisien, M.-A., Miller, C., Parks, S. A., DeLancey, E. R., Robinne, F.-N. & Flannigan, M. D. (2016), 'The spatially varying influence of humans on fire probability in north america', *Environmental Research Letters* **11**(7), 075005.

pearson's correlation - statstutor (2008).
URL: *https://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf*

Pettitt, A. N. (1979), 'A non-parametric approach to the change-point problem', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **28**(2), 126–135.

Piryonesi, S. M. & El-Diraby, T. E. (2020), 'Data analytics in asset management: Cost-effective prediction of the pavement condition index', *Journal of Infrastructure Systems* **26**(1), 04019036.

Powers, D. M. W. (2008), 'Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation', *ResearchGate* **2**.

QGIS - Open Source Geographic Information System (2021).
URL: *https://qgis.org/*

Robert J. Hijmans, J. v. E. & Sumner, M. (2021), 'Raster: Geographic data analysis and modeling'.
URL: *https://cran.r-project.org/web/packages/raster/index.html*

Roberts, G. & Giglio, L. (2021), 'Ceos land product validation subgroup'.
URL: *https://lpvs.gsfc.nasa.gov/Fire/Fire_home.html*

Roger Bivand, N. L.-K. & Pebesma, E. (2021), 'Maptools: Tools for handling spatial objects [r package maptools version 1.1-1]'.
URL: *https://cran.r-project.org/web/packages/maptools/index.html*

Salazar, L. F., Nobre, C. A. & Oyama, M. D. (2007), 'Climate change consequences on the biome distribution in tropical south america', *Geophysical Research Letters* **34**(9).

Schober, P., Boer, C. & Schwarte, L. A. (2018), 'Correlation coefficients', *Anesthesia, Analgesia* **126**(5), 1763–1768.

Teodoro, P. E., de Oliveira-Júnior, J. F., Da Cunha, E. R., Correa, C. C. G., Torres, F. E., Bacani, V. M., Gois, G. & Ribeiro, L. P. (2016), 'Cluster analysis applied to the spatial and temporal variability of monthly rainfall in mato grosso do sul state, brazil', *Meteorology and Atmospheric Physics* **128**(2), 197–209.

Therneau, T., Atkinson, B. & Ripley, B. (2019), 'Recursive partitioning and regression trees'.
URL: *https://cran.r-project.org/web/packages/rpart/rpart.pdf*

Tippmann, S. (2015), 'Programming tools: Adventures with r', *Nature News* **517**(7532), 109.

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).
URL: *https://www.mdpi.com/2076-3263/10/3/105*

Trenberth, K. E., Jones, P. D., Ambenje, P., Bojariu, R., Easterling, D., Klein Tank, A., Parker, D., Rahimzadeh, F., Renwick, J. A., Rusticucci, M. et al. (2007), 'Observations. surface and atmospheric climate change. chapter 3'.

Visa, S., Ramsay, B., Ralescu, A. L. & Van Der Knaap, E. (2011), 'Confusion matrix-based feature selection', *MAICS* **710**, 120–127.

Wickham, H., François, R., Henry, L., Müller, K. & RStudio (2018), 'dplyr'.
URL: *https://www.rdocumentation.org/packages/dplyr/versions/0.7.8*

Wu, W. & Zhang, L. (2013), 'Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern puerto rico', *Applied Geography* **37**, 52–62.