*AN EXPLORATORY DATA ANALYSIS OF FIRES IN DEMOCRATIC REPUBLIC OF CONGO*

*ARE FIRES IN DEMOCRATIC REPUBLIC OF CONGO SOLELY BASED ON BAD FARMING METHODS AND HARSH WEATHER CONDITIONS?*

Authors: Esther Mensah, Gideon Ahiadzi, & James Owusu-Appiah

Contents

[INTRODUCTION 3](#_Toc50542262)

[HYPOTHESIS AND RESEARCH OBJECTIVES 3](#_Toc50542263)

[RESEARCH QUESTIONS 3](#_Toc50542264)

[BRIEF DESCRIPTION OF DATASET 3](#_Toc50542265)

[METHODOLOGY 4](#_Toc50542266)

[EXPLORATORY DATA ANALYSIS 4](#_Toc50542267)

[I. UNDERSTANDING THE DATA 4](#_Toc50542268)

[II. CLEANING THE DATA 5](#_Toc50542269)

[III. FEATURIZATION 5](#_Toc50542270)

[IV. ANALYSIS OF RELATIONSHIP BETWEEN VARIABLES 6](#_Toc50542271)

[USE OF MICROSOFT AZURE TO CREATE PREDICTIVE MODEL 6](#_Toc50542272)

[DEPLOYMENT OF PREDICTIVE MODEL AS A SERVICE 6](#_Toc50542273)

[DISCUSSION AND INSIGHTS 6](#_Toc50542274)

[RECOMMENDATIONS AND CONCLUSIONS 6](#_Toc50542275)

[REFERENCES 6](#_Toc50542276)

# INTRODUCTION

The Congo Basin has been described the earth’s “second green lung” after the Amazon region. Each year, Central Africa record thousands of Agricultural burns. According to several international reports, Angola and DR Congo is ranked first in terms of the extent of the current wildfires. The country records on average 8,000 bushfires every year since 2000 to 2016.

On 24th August,2019, a NASA satellite was said to have detected 3395 cases in Congo. The Congo Basin’s forests absorb massive amounts of carbon dioxide making them a key player in the fight against climate change. According to Global Forest Watch fires, since 1 July, there have been nearly one million sources of fires in DR Congo. The practice of “slash and burn” agriculture has been hazardous and one of the major causes of Agricultural burn in DR Congo.

Slash-and-burn is also a leading cause of deforestation in many parts of the world. There is a time every year, traditional farmers in Central Africa set fire to the remains of old crop fields to clear them and prepare for new planting, to renew pasture, or to clear forest for new agricultural use.

While this practice is over 12,000 years old and still practiced by small farmers, in more recent time larger operations also have adopted fire to clear fields, especially in forested land. This “slash-and-burn” agriculture is useful for the farmer, but the high levels of smoke can be hazardous to human and animal health. Slash-and-burn is also a leading cause of deforestation in many parts of the world.

# HYPOTHESIS AND RESEARCH OBJECTIVES

In this report, an exploratory data analysis and predictive model is used to estimate the various probabilities and relationships between the percentage of an area that will be affected based on some particular environmental and atmospheric indicators. As the data becomes more complete and further analysis is done, this analysis can be extended to estimate the percentage of an area that will get burnt in case of a fire outbreak in not only the Democratic Republic of Congo but other countries across the globe. The main objective of this EDA is to determine the correlation between the percentage of an area that will get burnt in case of a fire outbreak based on the vapor pressure, soil moisture, and other indicators both environmental and atmospheric.

The following are the hypothesis for our analysis:

1. Does areas with less soil water experience high area burnt when there is an agricultural burn?
2. Are the burns caused by farmers?
3. Do atmospheric conditions of an area affect the percentage of the area that will get burnt in case of a fire outbreak?
4. Is there a significant difference in the area that will get burnt and area that will not when there is an agricultural burn?

# RESEARCH QUESTIONS

More specifically, the following research questions will be investigated under the various hypothesis given:

1. What percentage of an area will get burnt in case of a fire outbreak?
2. What is the probability of an area getting burnt based on particular atmospheric indicators?
3. What is the probability of an area with small amount of water and hot and dry atmospheric conditions getting burnt?
4. What relationship exists between the atmospheric vapor pressure of an area and the percentage of area burnt?
5. What is the probability of an urban area having a higher percentage of area burnt?

# BRIEF DESCRIPTION OF DATASET

The dataset used in this analysis is an aggregated data on burned areas across the whole of Democratic Republic of Congo since 1st April, 2020. The dataset consists of 764,200 rows and 32 columns which represents the attributes of the data. The features of the dataset include:

1. **ID:** The IDs take the form of [area ID]\_yyyy-mm-dd. There are 3821 area squares each with a unique ID ranging from 0 to 3820.
2. **area:** Area ID
3. **date:** The date that the data is aggregated over
4. **lat:** Latitude of the center of the area
5. **lon:** Longitude of the center of the area
6. **burn\_area:** Percentage of the area burnt
7. **climate\_aet:** Actual evapotranspiration, derived using a one-dimensional soil water balance model
8. **climate\_def:** Climate water deficit, derived using a one-dimensional soil water balance model
9. **climate\_pdsi:** Palmer Drought Severity Index
10. **climate\_pet:** Reference evapotranspiration (ASCE Penman-Montieth)
11. **climate\_pr:** Precipitation accumulation
12. **climate\_ro:** Runoff, derived using a one-dimensional soil water balance model
13. **climate\_soil:** Soil moisture, derived using a one-dimensional soil water balance model
14. **climate\_srad:** Downward surface shortwave radiation
15. **climate\_swe:** Snow water equivalent, derived using a one-dimensional soil water balance model
16. **climate\_tmmn:** Minimum temperature
17. **climate\_tmmx:** Maximum temperature
18. **climate\_vap:** Vapor pressure
19. **climate\_vpd:** Vapor pressure deficit
20. **climate\_vs:** Wind-speed at 10m
21. **elevation:** Land elevation
22. **landcover\_0:** Water Bodies: at least 60% of area is covered by permanent water bodies.
23. **landcover\_1**
24. l**andcover\_2:** Evergreen Broadleaf Vegetation: dominated by evergreen broadleaf and palmate trees and shrubs (>1m). Woody vegetation cover >10%.
25. l**andcover\_3:** Deciduous Needleleaf Vegetation: dominated by deciduous needleleaf (larch) trees and shrubs (>1m). Woody vegetation cover >10%.
26. **landcover\_4:** Deciduous Broadleaf Vegetation: dominated by deciduous broadleaf trees and shrubs (>1m). Woody vegetation cover >10%.
27. **landcover\_5**
28. **landcover\_6:** Annual Grass Vegetation: dominated by herbaceous annuals (<2m) including cereal croplands.
29. **landcover\_7:** Non-Vegetated Lands: at least 60% of area is non-vegetated barren (sand, rock, soil) or permanent snow/ice with less than 10% vegetation.
30. **landcover\_8:** Urban and Built-up Lands: at least 30% impervious surface area including building materials, asphalt, and vehicles.
31. **population\_density:** The estimated number of persons per square kilometer.
32. **precipitation:** Merged microwave/IR precipitation estimate

# METHODOLOGY

The tools used in the exploratory analysis is the Python Jupyter notebook together with some functionalities of the Microsoft Azure cloud. Major python libraries that were used include NumPy, Pandas, MatplotLib, Seaborn, Missingno.

The main processes used in the exploration of the data, analysis, and predictive model creation includes:

1. Understanding the Data
2. Cleaning the Data
3. Featurization
4. Analysis of relationship between variables

# EXPLORATORY DATA ANALYSIS

## UNDERSTANDING THE DATA

Understanding the data is a crucial part of exploratory data analysis.

1. The necessary libraries and modules were imported. These included NumPy, Pandas, Collections, Matplotlib, Seaborn, Missingno and SciPy.
2. The train and test datasets were read into a Pandas data frame.
3. The columns and rows of the dataset was explored, and the total count of rows determined.
4. The various data types of each column were checked, and they were all floats except the date which was a date type.

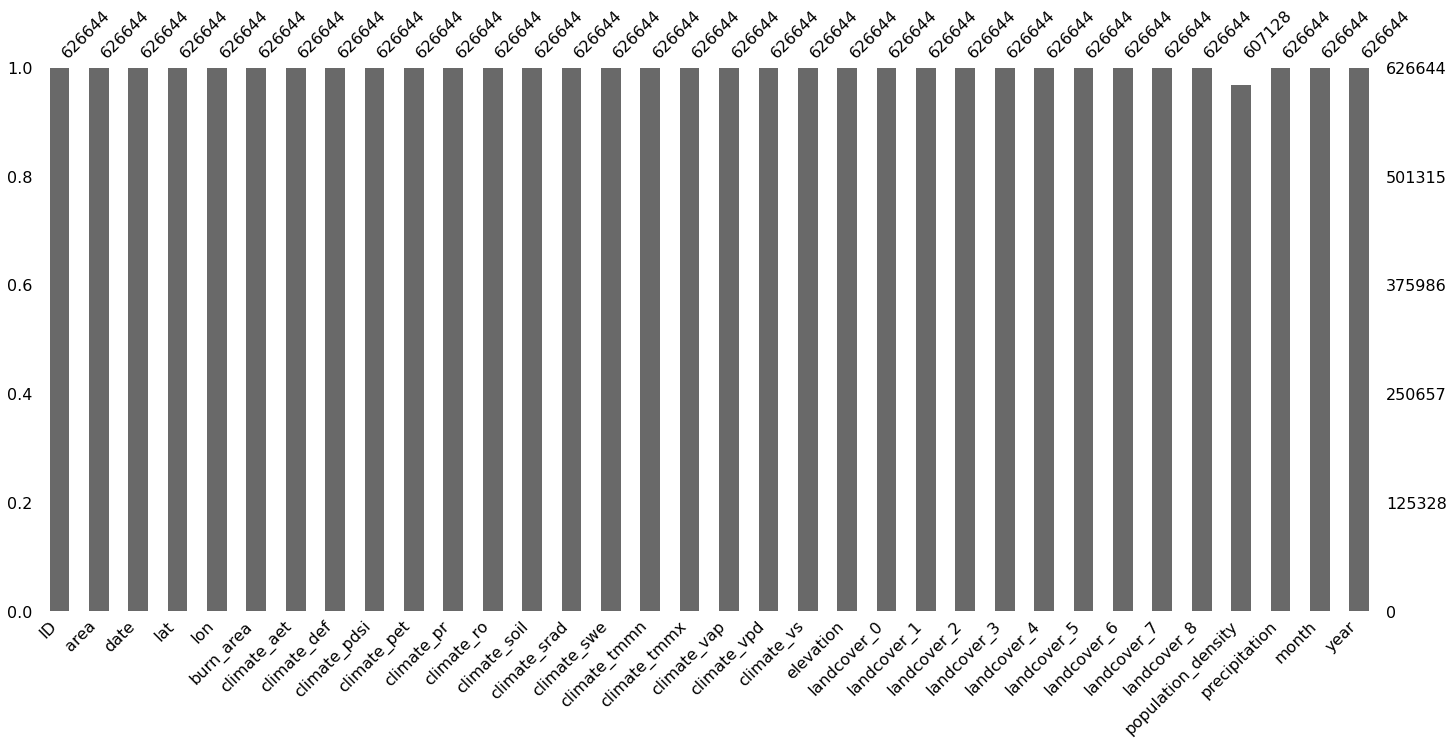
Insights drawn include:

1. The dataset is made up of 626644 rows and 32 columns which means it has 626644 records with 32 features.
2. The dataset has 19516 rows with null values in the population\_density column which is 3% of the total dataset.
3. All the columns in the dataset are of float data type except ID and date which are object and date respectively.

## CLEANING THE DATA

Cleaning the data is important for making a predictive model that is highly accurate.

1. The final dataset from the combination of the train and test datasets was checked for missing values, duplicates and outliers.
2. The dataset was checked for duplicates and there was none.
3. The dataset was checked for inconsistent and irrelevant data and there was none.
4. The dataset was wrangled to transform it into a form that can be good for visualizations
5. All columns of the dataset had no null values except the population\_density column which had 19516 null values.



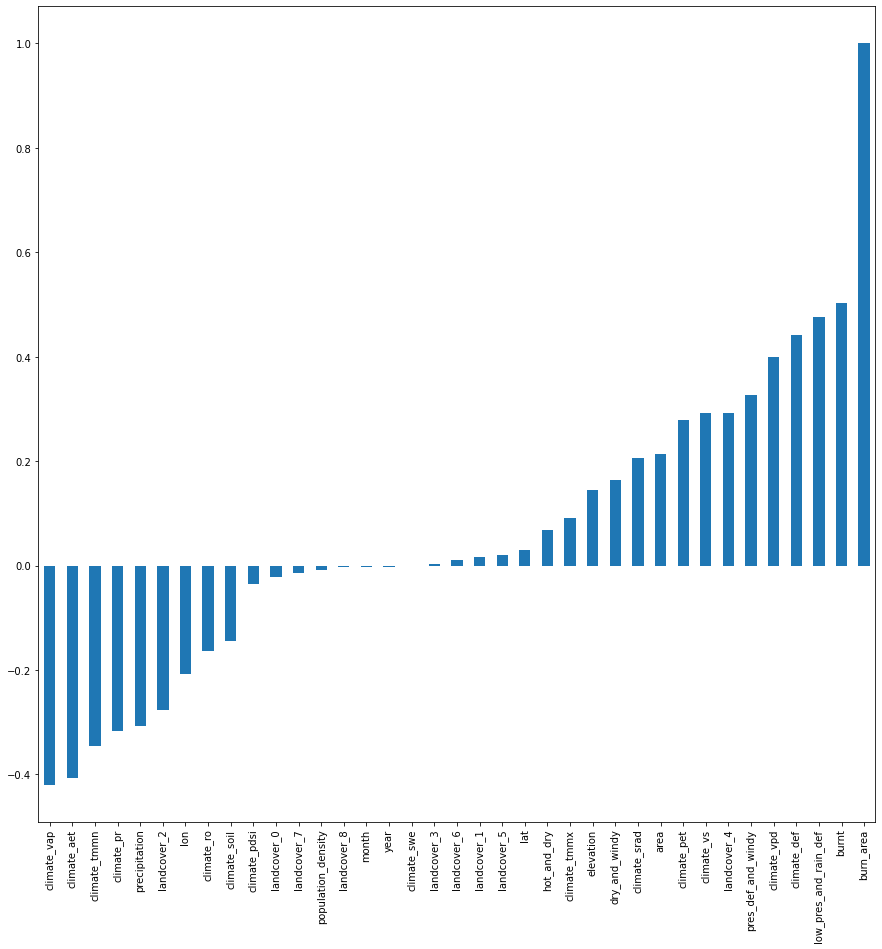
**Figure 1: This visualization shows that only the population\_density column has null values**

1. The rows with null values were removed which formed a total of only 3 % of the total data to help in the analysis of the data for effective building of a predictive model.

## FEATURIZATION

Featurization is using domain knowledge of the data to create features that help machine learning algorithms to learn better.

1. Month and year were extracted from the date column to create month and year columns in addition to the original columns given.
2. Addition of other columns such as dry\_and\_windy, hot\_and\_dry, low\_pres\_and\_rain\_def, pres\_def\_and\_windy and burnt.
3. The dry\_and\_windy column was created by calculating the inverse of the climate\_soil column values and multiplying it by the climate\_vs column values. The reason was because dryness and wetness are inversely proportional and since the climate\_soil represents the amount of water in the soil then an inverse of that will give the dryness of the soil.
4. The hot\_and\_dry was created by multiplying the climate\_srad with the inverse of the climate\_soil column values. The reason was the climate\_srad represents the downward surface shortwave radiation which is synonymous to how hot the place is and the inverse of the climate\_soil column values gives how dry the place is.
5. The low\_pres\_and\_rain\_def was created by multiplying the climate\_def column values with the inverse of climate\_vap column values. The reason was because the inverse of the climate\_vap column values gives low pressure values and that multiplied by the water deficit which is represented by the climate\_def column values gives a new column that shows values of low pressure and rain.
6. The pres\_def\_and\_windy was created by multiplying the climate\_vs column values with the inverse of the climate\_aet column values. The reason was because the climate\_vs column values represent wind-speed at 10m and the since the climate\_aet column values represent actual evapotranspiration which takes places under low atmospheric pressure. An inverse of that gives the actual pressure in the atmosphere since both are inversely proportional.
7. The level of correlation between each column and the target variable(burn\_area) was checked using visualization.
8. The burnt column has values of 0 and 1, which shows whether the area was burnt or not based on the value of the burn\_area column.



**Figure 2: The visualization showing the positive and negative correlation between the target variable (burn\_area) and the other columns of the dataset.**

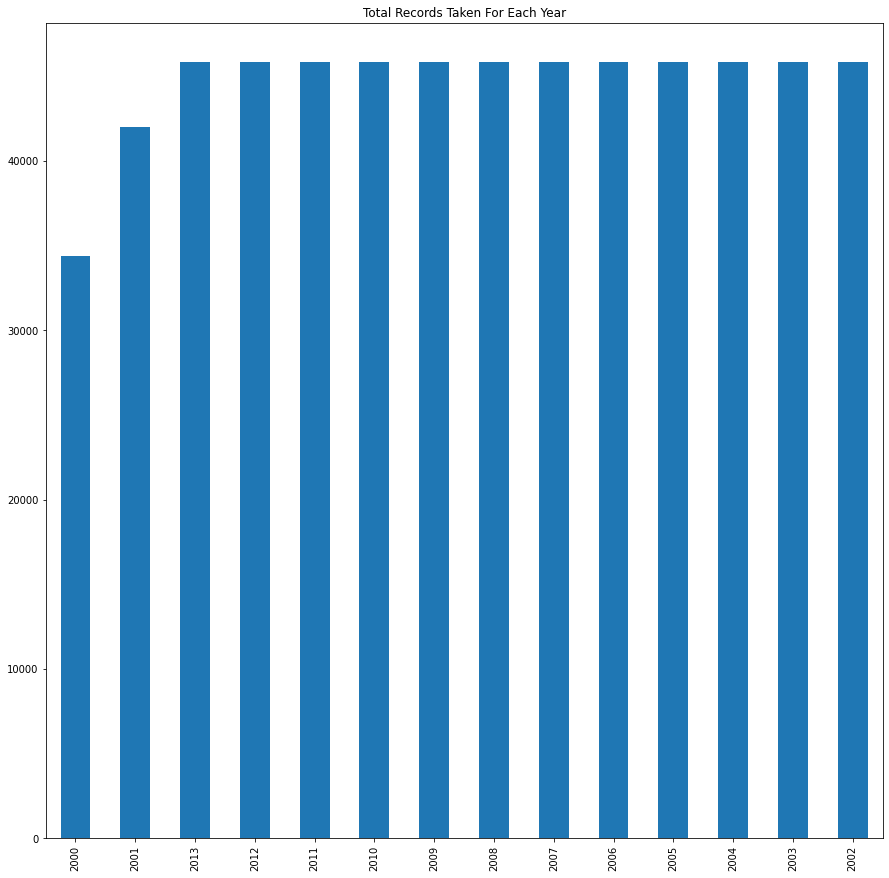
Insights drawn:

1. The visualization shows that the low\_pres\_and\_rain\_def shows a high positive correlation with the burn\_area. This means that when there is low pressure and low rainfall then the area density in case of a fire will be high.
2. The visualization also shows that the climate\_vap has a high negative correlation with the burn\_area. This means that when the vapor pressure is high then there the area density affected will be less.

## ANALYSIS OF RELATIONSHIP BETWEEN VARIABLES BY VISUALIZATIONS

The various columns of the dataset have a particular relationship with other columns of the dataset and this can be well noted through visualizations.

**Figure 3: The visualization showing the total records taken for each year in ascending order.**



**Figure 4: The visualization showing the total count of actual fires that happened each year in descending order**

# 



# USE OF MICROSOFT AZURE TO CREATE PREDICTIVE MODEL

Microsoft Azure cloud services was effectively utilized in the creation of an efficient predictive model. The main purpose of the predictive model is to predict the percentage of an area that will get burnt in if there were to be a fire outbreak based on indicators such as the wind speed, soil moisture, evapotranspiration rate and many others. The various processed used to get the predictive model up and running was mainly with the help of Microsoft Azure Designer. The main reason was to have a visual representation of how the flow will be like for our predictive model.

1. Created an Azure Machine Learning workspace to get access to the needed resources for the creation of a good predictive model. Used the enterprise edition to make the use of Designer possible.
2. Created a compute instance to serve as the major workstation for creation of the predictive model and give the needed support for the model.
3. Created a compute cluster with a maximum of 2 nodes to allow parallel run and faster processing of data.
4. Uploaded the dataset that we have already worked on with Python Jupyter Notebook from local files to help in easy analysis. Placed it into the After the analysis and featurization, a refined dataset was uploaded to Azure. This dataset had new features added to enhance in the training of our model.
5. Creation of a new pipeline from the Design tab of the Azure Machine Learning studio GUI by doing the following:

* Dragged the Fires dataset from the Dataset tab onto the navigator.
* Dragged the Select Column in Dataset module from the Modules tab and connected it to the Fires dataset. Selected all columns except ID, area and date.
* Dragged the Clean Missing Data module from the Modules tab and connected it to the Select Column in Dataset module.
* Selected Clean Missing Data module and in the edit column added population density since it has null values and set the cleaning mode to remove entire row.
* Dragged the Normalize Data module from the Modules tab and connected it to the Cleaning Missing Data module. Set the transformation method to MinMax and all columns selected to be transformed except the burn\_area, ID and date.
* Dragged the Split Data module and split the data in the ratio 0.7 for train and 0.3 for test. Connected it to the Normalize Data module.
* Dragged the Train module and connected one end to the Split Data module and the other end to the Linear Regression module.
* Dragged the Score module and connected one end to the Train module and the other to Split Data module.
* Dragged the Evaluate module and connected it to the Score module. Used it to check the effectiveness of all the algorithms used.
* Continued to change the algorithms to see the best fit.

1. Evaluation results of Linear Regression algorithm when used:

* Mean\_Absolute\_Error = 0.023599
* Root\_Mean\_Squared\_Error = 0.0514
* Relative\_Squared\_Error = 0.686903
* Relative\_Absolute\_Error = 0.894758
* Coefficient\_of\_Determination = 0.313097

1. Evaluation results of Boosted Decision Tree Regression when used:

* Mean\_Absolute\_Error = 0.013055
* Root\_Mean\_Squared\_Error = 0.039449
* Relative\_Squared\_Error = 0.404602
* Relative\_Absolute\_Error = 0.494994
* Coefficient\_of\_Determination = 0.595398

1. Evaluation results of Decision Forest Regression when used:

* Mean\_Absolute\_Error = 0.007284
* Root\_Mean\_Squared\_Error = 0.026836
* Relative\_Squared\_Error = 0.187245
* Relative\_Absolute\_Error = 0.276179
* Coefficient\_of\_Determination = 0.812755

1. Evaluation results of Neural Network Regression when used:

* Mean\_Absolute\_Error = 0.018405
* Root\_Mean\_Squared\_Error = 0.045379
* Relative\_Squared\_Error = 0.535397
* Relative\_Absolute\_Error = 0.697828
* Coefficient\_of\_Determination = 0.464603

1. Afterwards, changed the design of the pipeline by bringing in Tune Model Hyperparameter module and Filter Based Feature Selection to substitute the Train module. The Decision Forest Regression was used as the Machine learning algorithm.
2. Evaluation results of new pipeline with Decision Forest Regression when used:

* Mean\_Absolute\_Error = 0.007008
* Root\_Mean\_Squared\_Error = 0.025717
* Relative\_Squared\_Error = 0.171951
* Relative\_Absolute\_Error = 0.265696
* Coefficient\_of\_Determination = 0.828049

# DEPLOYMENT OF PREDICTIVE MODEL AS A SERVICE

The Microsoft Azure Machine Learning studio has the needed functionalities to help deploy a predictive model as a service for others to access it and get the needed results. The final predictive model will be used by a lot of countries to help predict how susceptible they are to fires based on the conditions available in their jurisdiction.

1. An inference cluster was created to help with the deployment of the predictive model as a service. Created in a region different from the compute instance.
2. Created a real-time pipeline to be able to take inputs from users and predict the percentage of an area that will get burnt in case of a fire outbreak.
3. Modified the Select Columns module to exclude the burn\_area column since it is the target variable.
4. Dragged the Execute Python Script and added some codes for it to only produce the required output.
5. Added some data to Enter Data Manually module for prediction with our model.

# DISCUSSION AND INSIGHTS

From the Exploratory Data Analysis, in the year 2010, the Democratic Republic of Congo recorded the highest fires since 2000. It recorded the highest area density of 0.95 burnt. DR Congo recorded the highest climate water deficit of 2039.10 during the year 2010 and also recorded the highest population density of 9514. These indicators contributed to the numerous burns in the year 2010. Generally, in 2010, the atmosphere was hot and dry throughout the year and the less humid, this contributed to the major increase in the fires that occurred throughout the year.

There is an inconsistent trend in the total number of fires that occur annually from 2000 to 2013. From 7292 fires in 2000 and decreased to 6645 fires in 2001. The number spiked up to 8361 in 2002 and increased slightly to 8794 fires in 2003, and 8811 in 2004. In 2005, the numbers increased to 9037 throughout the year and dropped to 8812 in 2006. The fires decreased slightly in 2007 by over 300 giving 8436 total fires and 8779 in 2008. In 2009, it dropped again to 8325 but again reached its peak in 2010 when it recorded 9278. In 2011, there were 8493 fires throughout the year and dropped in the next year to 8264. 2013 brought about 8584 fires. This trend clearly shows that the depletion of the ozone layer is not a major reason why there are so many fires in Democratic Republic of Congo. This is due to the fact that, as the years went by the fires if were to be influenced by the ozone layer depletion should have increased since the ozone layer is getting worse as the year increases.

# RECOMMENDATIONS AND CONCLUSIONS

1. Check with local authorities for any local regulations concerning agricultural burns.
2. Human activities and the operation of certain machinery can increase the potential for agricultural fire ignition. Policies and procedures should consider the potential for agricultural fire ignition, particularly during the days where weather conditions are conducive to Wildfire spread.
3. Farmers should put off fire after the “slush-and-burn" practice.
4. Farmers should store flammable or combustible materials such as woodpiles and rubbish away from their farmland.
5. Putting up warning signals at areas that are prone to burns.
6. Areas in DR Congo with high temperatures should practice afforestation since plants help the area to be cool.

# REFERENCES

<https://www.independent.co.uk/environment/climate-crisis-amazon-fires-brazil-congo-rainforest-africa-bolsonaro-nasa-macron-a9080741.html>

<https://www.bbc.com/news/world-africa-49471644>