The Effect of Deforestation on CO_2 Emissions

Julian Agolini and Rashed Rifat

December 11, 2022

Abstract

Prior studies have shown the effect of deforestation on carbon dioxide emissions to be significant. However, such studies have not attempted to predict the degree to which deforestation contributes to an increase in carbon dioxide emissions. This study reviews global data and formally investigates the relationship between deforestation and carbon dioxide emissions in a given region. In this study, big data tools were used to analyze global sets and produce insights. The results suggest that deforestation is a key component of climate change and should be targeted to preserve the future health of our planet.

Introduction

Tree coverage is an extremely important metric to monitor when studying climate change and more specifically carbon emissions within a specific area. Trees store carbon dioxide, regulate climate, and provide vital resources to an ecosystem. In this study, we investigated the relationship between tree coverage by hectares of loss to carbon dioxide emissions by tons to better understand this relationship. To advance our analysis, we utilized two large data sets. When completed, we gained insights into this relationship and auxiliary ones related to eliminating climate change. Ultimately, climate science is incredibly important yet equally complex. Continual investigation into proven methods of alleviating the adverse effects of Climate change is paramount for the future of our planet. As part of this crucial intervention, the prevalence of large data sets and the utilization of big data tools are monumental in providing valuable insights in an expedited fashion. In our study, we used several of these big-data tools under the Apache Hadoop ecosystem. Below, a flow of our datapath can be seen which depicts the journey our data took on its way to producing our analytics.

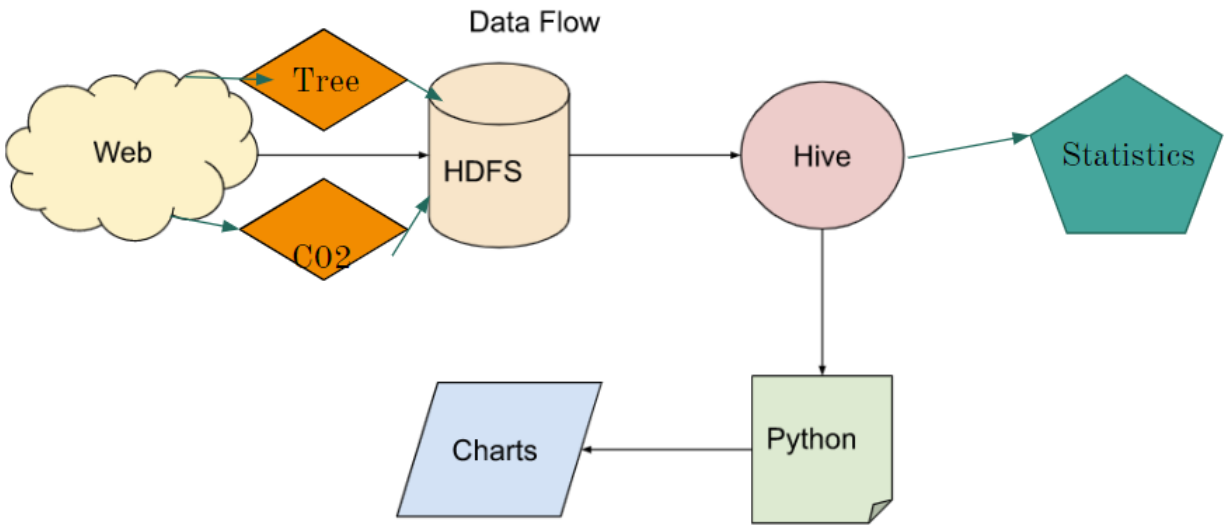

Figure 1

In this figure, the processing of our data can be observed. Stemming from the web, we first downloaded our data sets and after some preliminary cleaning and profiling which will be detailed later in the paper, we incorporated our data into Hadoops HDFS (Highly Distributed File System). Next, we used another Apache service, Hive, to perform queries to further our analysis. In using Hive, we generated some statistics that can be easily interpreted and provide insight. Finally, we used python libraries to generate some visualizations of our findings.

Motivation

While Climate change is as pressing an issue as ever with increased temperatures and rising sea levels reported worldwide, the field is unfortunatly still riddled with controversy. Stances are often firm, with little room for debate. Given this state of contention, we chose to study a sector of the field that is as uncontroversial as possible and attempted to build off of prior knowledge to produce convincing analytics. In utilizing this method, we feel this is the best way to convince those with dissenting opinions. Starting with proven facts allows the case for prescriptive strategies be possible. Two uncontested facts of climate science are that carbon dioxide warms the atmosphere, and trees take in carbon dioxide. By studying the amount of deforestation in a given area, we can account for a certain amount of carbon dioxide emissions. Although other reports on deforestation have been made, they often are not accessible to the layman. We believe that education through, simple, uncontroversial, and convincing analytics is the most optimal way to advance the field.

Related Work

In a related report focused on deforestation efforts entitled “High-Resolution Global Maps of 21st-Centurary Forest Cover Change”, researchers M.C Hansen and P.V Potapov, found that global forestry levels have been plummeting for many years. In this report, the researchers used a similar approach to that conducted in this study but its primary focus was centered on the sources of deforestation rather than its effects. This report was tremendously valuable in providing insights into the sources of deforestation. In order to make prescriptive conclusions surrounding paths forward in eliminating the negative externalities of deforestation, it is vital to first understand what deforestation can be eliminated. Additionally, some sources of deforestation can be recursive in the cycle of climate change and are especially important to understand and rationalize. For example, Hansen and Potapov found that one of the leading drivers of deforestation globally is forest fires. These fires even if occasionally started directly by humans, are always in part fueled by dry conditions as a result of climate change.

In a second report focused on the Effects of Forestry on Carbon Emissions in China, researchers Zaijun Li and Zouheir Mighri showed that managing and controlling forests in China is an effective means of regulating carbon dioxide emissions. This result is predicted but relieving, without it hopes to curtail the emissions due to carbon dioxide quickly dwindling. Importantly, the methods that Li and Mighri suggest to control forestry in an optimal way do not negatively impact China’s economic outcomes. Without a reduction in GDP, there become decreasingly few reasons to not implement such solutions. As Li and Mighri note, carbon dioxide emissions and economic activity are inseparable in the current economic climate. China, in particular, has seen monstrous increases in carbon dioxide emissions coupled with its rapid economic development. In our analysis, we found that net carbon dioxide emissions were incredibly related to economic activity. As more developed nations, those with high levels of economic activity also have more resources to implement some of the suggested strategies.

In both cases, the related research reinforces our results with two key findings: carbon emissions can be controlled and managed by preserving forestry and these managements are feasible at a country-wide scale. While the science behind forestry and carbon emissions may seem trivial, it is of utmost importance to have adamantine grounds to build the case for a greener future for our planet.

Description of Datasets

For the purposes of this project, we use two datasets, which we will introduce shortly. The first dataset we gathered was deforestation data, from Global Forest Watch, an organization that "offers the latest data, technology, and tools that empower people everywhere to better protect forests" (Global Forest Watch). This dataset recorded data about a country (denoted in ISO 3166 Country Codes) and its loss of tree cover (in hectares) and CO2 emissions within the lost area (measured in milligrams) from 2001 to 2021. The dataset was structured in such a way that each record contained a unique combination of country and year, which we used to join our datasets further down the line. In terms of statistics, this dataset contained over 4,0000 rows of data and measured 136 KB in size.

The second dataset that we gathered for this project was data relating to the total greenhouse gas emissions (measured in equivalent metric tons of CO2) within each country, differentiated by year, source, sector, and gas. This data was gathered from Climate Watch, an organization dedicated to offering "open data, visualizations, and analysis to help policymakers, researchers and other stakeholders gather insights on countries' climate progress" (Climate Watch). As such, the source column within this dataset all referred to CAIT, an abbreviation for Climate Watch. The sector column denoted the sector that this record referred to - the two sectors that we focused on were related to the land-use change and forestry sector (LUCF) and the total greenhouse gas emissions sector. Gas denoted the gas that we were measuring for this record. Years, in this dataset, referred to a set of columns labeled from 1990 to 2019. These columns denoted the number of emissions for that particular country and sector.

As you might imagine, this meant that the initial greenhouse gas emissions data we gathered was wide (as opposed to the long format described in the deforestation dataset). We will address this later on within our data profiling and cleaning stages. The greenhouse gas emissions dataset, when cleaned, measured 22,000 rows and occurred 2.54 MB in size.

Analytic Stages and Process

We began the initial data ingestion process by first loading the datasets into HDFS, where we ran some simple profiling and cleaning programs. In terms of profiling, for both datasets, we simply counted the number of rows that each dataset had before and after running our initial cleaning steps. We were only interested in the number of records that we had as we wanted to make sure that we were not dropping an inappropriate number of records. Additionally, the profiling that we were interested in doing could not be performed easily with MapReduce and is shown later on.

Cleaning our datasets required creating separate procedures for both datasets. For our deforestation data, our data was well structured and in a long format - given this, we simply chose to drop some records for which there was no information as to the CO2 emissions column. We chose to do this as a value of 0 within this column usually denoted a data quality control issue, where there was no relevant information for the record.

Cleaning our CO2 emissions dataset took greater effort. This data was structured in a wide format, with over 20 columns named as years from 1990 to 2019 (each of the year columns had a value denoting the tons of CO2 emissions for that record). Our first challenge was to filter this dataset so that the header column was dropped and that only records related to total and LUCF emissions were kept. We did this by simply skipping rows that did not contain these values. Our second challenge was to then convert our wide-format data into long-format data. This was done by writing multiple values within our mapper for each initial record within the dataset, where the values contained only a single year of data. While this did make our dataset much larger, it was crucial to do so to be able to join our disparate datasets together. Finally, we formatted some of the string values within our dataset by converting them to lowercase letters and replacing spaces with underscores. This would help in easily applying certain filters later on.

After cleaning our data and shaping them into the right format, we proceed to rename our out files (for convenience) and upload them to appropriate directories within HDFS once more. We then loaded these datasets as tables within Hive, where we joined our datasets on iso and year. This resulted in a combined dataset from which we created multiple other tables and statistics for our eventual visualization. We note that we used Hive for the bulk of the analysis within our project. This was accomplished by writing complex Hive queries applied to user-generated tables. For more information regarding these specific queries, refer to the hive_queries.sh file included within our project GitHub.

We began our analysis by running some simple queries to check that our data was in the format and shape that we expected. After confirming that the ingestion process did complete successfully, we began by first taking a look a the top ten countries by deforestation per year. We

performed this query by grouping by year and sector and the ranking countries by their amount of tree loss cover (in descending order). We then took the results of this subquery and filtered out for the first three ranks. The results of this query were then written in comma-separated format within HDFS, which we moved into our local environments (after renaming the files in an appropriate manner).

We then repeated this same process to rank the top countries by total greenhouse gas emissions per year, with the only key difference being that we ranked over CO2 emissions and applied an additional filter to select records only from the total greenhouse gas emissions sector. A similar query was applied to retrieve the top countries by CO2 emissions within the LUCF sector only. We pulled these results down to our local environments for later before moving onward to create our "rates" table.

Within our project, we wanted to examine the relationship between tree cover loss and the increased rate of CO2 emissions. To do this, we created a metric that roughly measures the amount by which CO2 emissions increased per hectare acre of deforestation. To do this, we simply divided our CO2 emissions by our deforestation value for records that from the LUCF sector - since our CO2 measurements denoted the amount of greenhouse gas emissions could be attributed to deforestation, dividing by the amount of tree cover loss gave us a good estimate of this rate. We saved this as a table so that we could glean further insights from it, starting with gathering some rankings based on rate.

One of the first analyses that we ran on our rates table was to gather the average global rate of CO2 emissions per hectare of deforestation per year. To do so, we grouped our dataset by year and then ran a simple average on those groups. For full disclosure, we only looked at records where the tree cover loss was greater than or equal to zero, as otherwise, our metrics would not have made much sense.

We then went on to rank the countries with the highest rates per year. To do so, we first grouped our data by year and country, where we then applied an average on the rates. Note that this average is mainly superfluous and was written mostly to catch any possible multiple records per country, although there should not have been any within our dataset. Regardless, we then ranked the average rate (in descending order) and gathered the top three countries per year. This data was then saved for further visualization.

After examining the rates of CO2 emissions per hectare of deforestation, we moved on to the final part of our analysis which was to look at the year-over-year change in deforestation (tree cover loss) and greenhouse gas emissions within LUCF and all sectors. We did this by first generating three different tables within our Hive environment, each of which measured the year-over-year change in the factors mentioned above. We generated these tables by applying a

lag function to the factor we were examining at the moment over a partition by country and grouped by year.

To then look at the YOY change for all years within our dataset, we generated a percentage change (as compared to the year prior) for all records within a specific factor table. We then grouped this data by year and took the average of the percent change, which would give the YOY change in deforestation, total greenhouse gas emissions, and LUCF greenhouse gas emissions globally. These results were then saved to our local environment for later visualization.

One final query that we looked at was to get the top three countries by YOY growth in deforestation per year- this would tell us the countries that were the most destructive towards their greenery. To do so, we repeated the steps noted above to generate our YOY percentages. We then grouped these percentages by year and then ranked them. Finally, we selected the top three countries per year within our ranking.

After we had completed our analysis, we took the data that we had collected from our Hive queries and created some visualization within Python. For the sake of brevity, we will not mention the code used to create these graphics, although you may view the code within our project GitHub. We examine these graphs and the insights that we can draw from them below.

Graphs and Insights

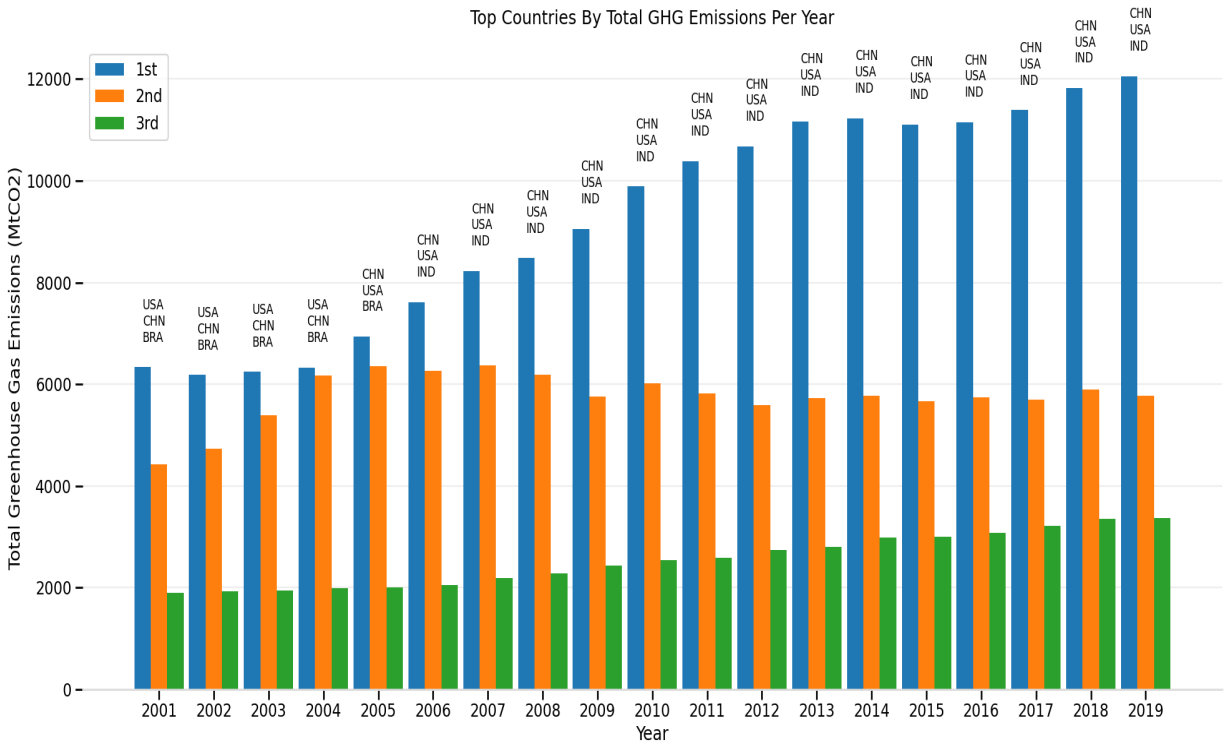

Figure 2

In this graphic, we examine the top three countries (as ranked by total greenhouse gas emissions) per year. The small legend denotes the colors representing what ranking and the text above each group of bars represent, in order, the ISO codes of the countries featured within the rankings. By examining the top three countries over a period of years, we note several interesting characteristics. The first of which is the fact that China, the USA, and India all dominate the rankings from 2006 onwards. This is somewhat surprising given that we may expect a little more variability from other developed nations. This suggests that there are a select group of countries that are responsible for the majority of the emissions overall. Exploring this further, we note that China has consistently grown its emissions year to year and has far outpaced the rate of emissions from the USA or India. India has also increased it rate of emissions consistently, however, the USA remains quite flat in terms of its suggestions. This could signal that there are certain developed countries that have taken efforts to curb their emissions in the pursuit of a greener world while other nations are still ramping up their industry. Looking into the causes for such a reason could warrant another follow-up study.

We would also like to note that the findings reported here map well to economic activity, which serves as a measure of validity for our project. We expect to see that the more developed nations would emit more CO2 and be able to view this trend lends further credibility to the results.

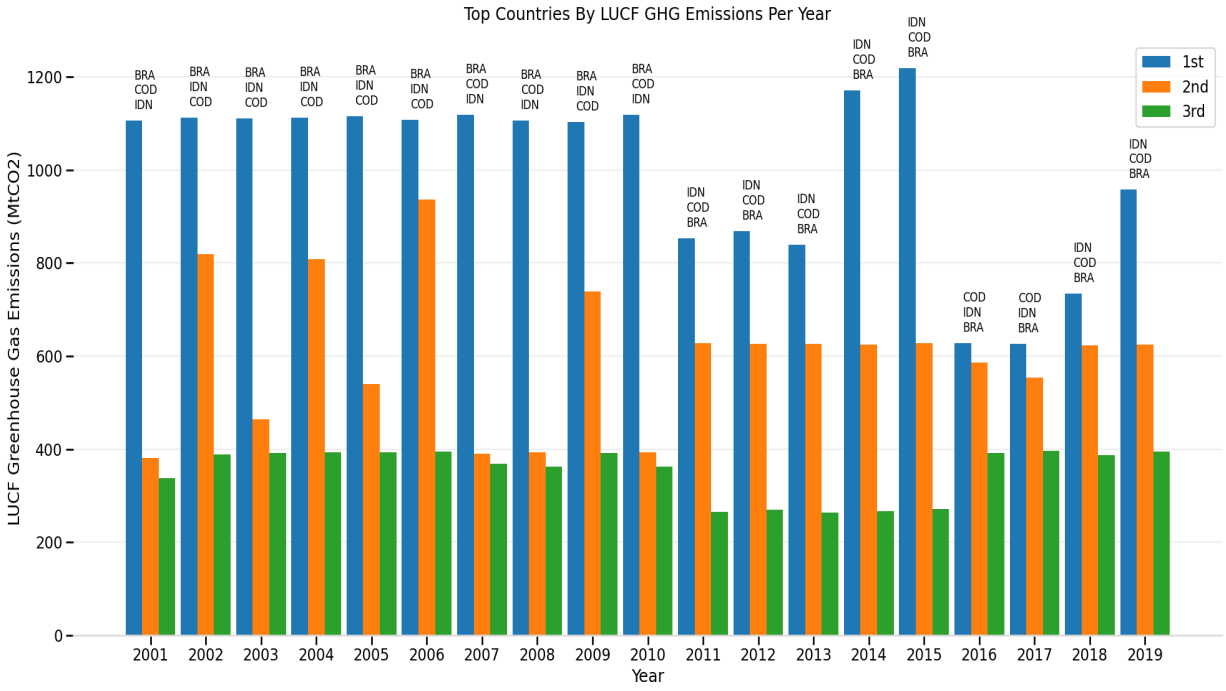

Figure 3

The visualization here mirrors much of the same format as the graph presented beforehand - in this case, however, we examine the top countries by greenhouse gas emissions related only to the LUCF sector specifically. The trends that we have seen previously still play a large part here as well; we note that a few countries dominate the rankings as well. The top countries, in descending order, are Russia, Brazil, and the Democratic Republic of the Congo. It is interesting to note that Brazil has appeared both times in the analytics we have seen so far - it not only was a world leader in terms of total GHG emissions, but it is also prominent in terms of LUCF emissions as well. This repeated pattern merits a closer look at why Brazil is featured so prominently within the climate discussion, perhaps in a follow-up study.

We do note that this graph features much more variability in terms of year-over-year change in LUCF emissions - there is no clear pattern in this case. We expect to see this level of variability given that we are looking at one sector of each country's emissions; while in aggregate the level of emissions tends to increase, the individual sectors may fluctuate. This is particularly prevalent given that we are looking at emissions related to land use, where countries may frequently adjust their level of activity depending on their needs.

Still, the pattern of a few countries outputting such a large proportion of emissions is something that should be researched further - why is it that these countries are relevant to the discussion? What can we learn from their policies and execution that we could use to better guide ourselves to a cleaner world?

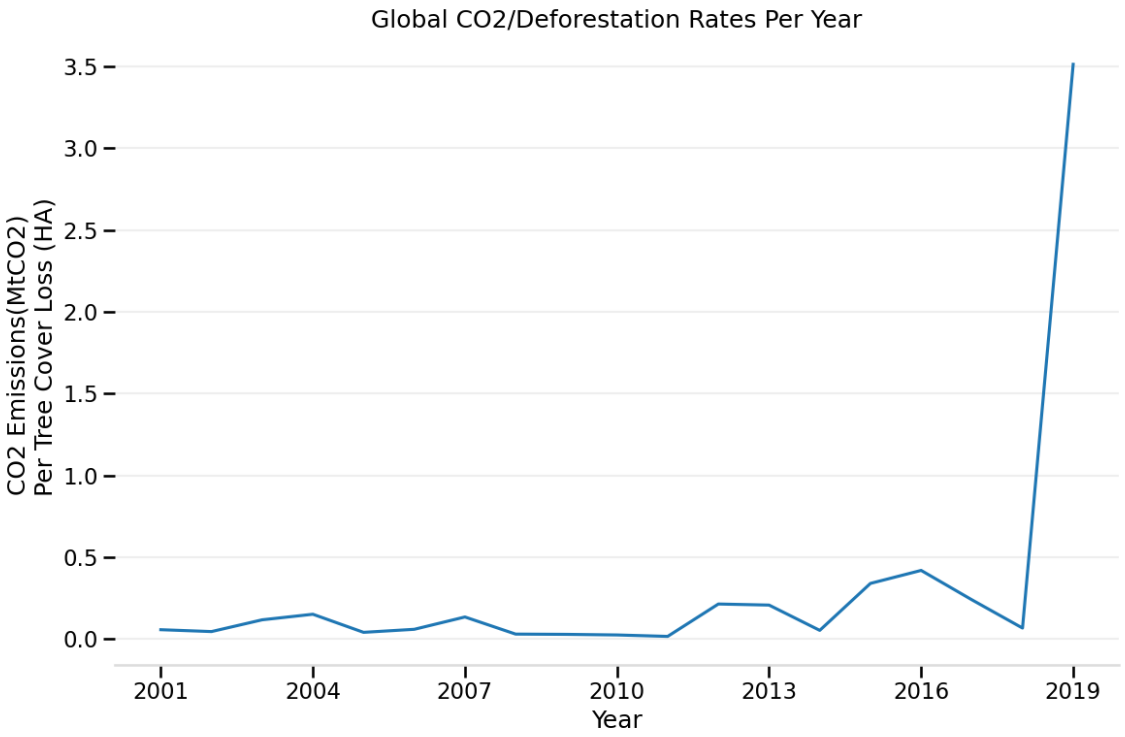

Figure 4

We also took a look at the rate of CO2 emissions per hectare of tree cover loss. Examining this rate could tell us the importance of deforestation in terms of greenhouse gas emissions and this graph certainly speaks to the need for moderation of deforestation. We have seen a consistent increase in the amount of CO2 gas missions relating to the LUCF sector, suggesting that as we continue to increase our use of wooded areas, the amount of CO2 that the process generates is exponentially increasing. This would suggest that deforestation is becoming a larger and larger issue and merits taking action by policymakers and the general public.

We take a moment to note here the drastic increase in the rate of emissions per deforestation in the year 2019 is certainly quite irregular. Within our analysis, we were unable to find any errors that could have led to such a drastic increase (if there were any errors, to begin with). We suspect that this may have had something to do with the quality of the data itself, as opposed to any analysis. Regardless, we note that this is something to follow up on, in either data quality or examining the reason for why such a drastic shift occurred.

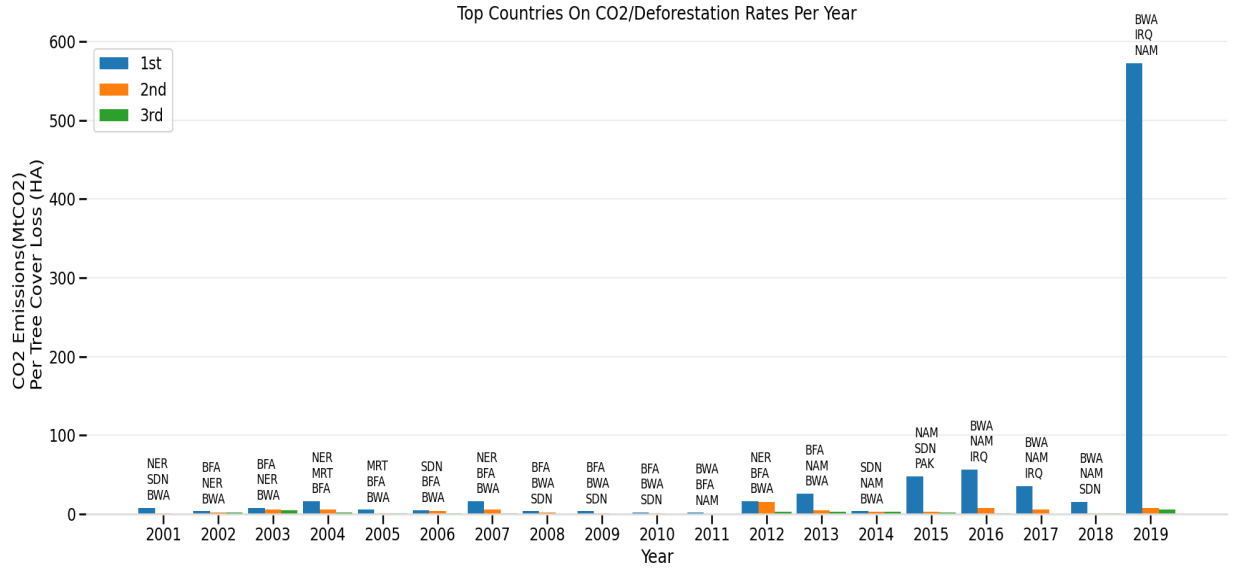

Figure 5

Finally, within this paper, we take a look at the countries which had the highest rates within our study. We note that while these rankings are not as consolidated, Botswana, Iraq, and Namibia all dominate the rankings. More specifically, Botswana is responsible for a large increase in the amount of CO2 emissions that each hectare of tree cover loss that it incurs. It is worthwhile to examine why exactly this is the case.

This graph also sheds light on those countries that are in dire need of reexamining their land use policies. As we have illustrated beforehand, deforestation, as it relates to greenhouse gas emissions, is a serious concern. Addressing this issue directly in a speedy manner is required to tackle this issue and prevent any further damage on a global scale.

Conclusion

As we progress into the oncoming years, climate change continues to become a more pressing issue as we face the consequences of years of ignored policy changes. Deforestation, in particular, is a bipartisan issue that contributes to the worsening state of the world. As we have shown in our analytics, there are a select group of countries that lead the world in terms of greenhouse gas emissions and deforestation. However, some of these nations have been able to flatline their year-over-year increase in CO2 produced while others rapidly increased instead. This, coupled with the fact that deforestation continues to become more and more relevant to the climate change discussion, suggests that we should focus on learning from the countries identified within this study, both in terms of procedures to follow and policies to review. We have shown that while tackling this issue may be difficult, it is possible and that there is a sense of urgency within these matters. We urge both the public and these nations, to consider the steps that they may be taking to preserve a planet that has existed far longer than the lifespan of humanity and will continue to persist long after our lifetimes.

Most of the literature within climate change discussions often offers a bleak outlook of the future - we would like to supply an opposite viewpoint. Greenhouse gas emissions and deforestation are certainly important but they should be met with a calm vigor and clear mind. The full force of humanity and all of its greatest inventions all have a vested interest in seeing a solution come to light. We hope that this study, and this paper, have provided you with the countries and sectors that we should focus on. Let us tackle perhaps one the gravest obstacles that humanity has had to face as we should: together.

Bibliography

1. Hansen, M. C., et al. “High-Resolution Global Maps of 21st-Century Forest Cover Change.” *Science*, vol. 342, no. 6160, 2013, pp. 850–853., https://doi.org/10.1126/science.1244693
2. Li, Zaijun, et al. “Effects of Forestry on Carbon Emissions in China: Evidence from a Dynamic Spatial Durbin Model.” *Frontiers in Environmental Science*, vol. 9, 2021, https://doi.org/10.3389/fenvs.2021.760675
3. Climatewatchdata.org. (n.d.). Retrieved December 11, 2022, from https://www.climatewatchdata.org/
4. Vizzuality. (n.d.). *Forest Monitoring, land use & deforestation trends*. Global Forest Watch. Retrieved December 11, 2022, from https://www.globalforestwatch.org/