The impact of Amazon deforestation on Brazil's carbon footprint

Ronja Michel - 23478672

November 27, 2024

Abstract

Forests play a key role in climate change mitigation by absorbing CO_2 from the atmosphere and storing it as carbon. Tropical rainforests, such as the Amazon, are particularly effective in this regard, absorbing far more CO_2 than native forests. The Amazon is the largest tropical rainforest in the world and plays a key role in the global carbon cycle. Around 60 per cent of the Amazon is located in Brazil, making the country a key player in the fight against climate change.

This project investigates how deforestation in the Amazon affects Brazilian CO_2 emissions by analysing deforestation and emissions data over time. The aim is to show the links between rainforest deforestation and rising CO_2 emissions in Brazil. This analysis helps to clarify the importance of the Amazon for the global climate and to understand how important protecting this ecosystem is for reducing CO_2 emissions.

Question

How does the deforestation of the Amazon rainforest affect CO_2 emissions in Brazil?

Data Sources

Description of the Data

Two data sets were used to answer the key question. Firstly, the data set 'Brazilian Amazon Rainforest Degradation' [1], which comes from Kaggle, a platform that contains a large number of data sets for machine learning. The dataset includes fires, deforestation area and climatic phenomena of the Amazon rainforest in Brazil by year and state for the period from 1999 to 2019. The dataset was selected due to the relevance of the included region for the project.

The second dataset is 'CO_2 and Greenhouse Gas Emissions' [2]. The dataset includes CO_2 emissions data globally and by country, including emissions from various sectors such as land use, from 1751 to 2022. Due to its comprehensive database and the possibility of detailed country breakdown, especially with regard to the CO_2 emissions sector, it represents a valuable resource for answering the key question. The dataset was provided by Our World in Data, a platform that provides scientifically sound datasets on the historical development of human living conditions.

Structure of the Data

The data set 'Brazilian Amazon Rainforest Degradation' comprises three files. The key question is answered using the file 'def_area_2004_2019'. This documents the deforestation areas in the Amazon region between 2004 and 2019. The data is structured in tabular form in CSV format. The dataset comprises a total of 17 rows, including a header row and 16 rows for the years 2004 to 2019. The data set consists of 11 columns. The first column contains the respective year, while the last column, which is labelled 'AMZ LEGAL', shows the sum of the deforestation areas of all federal states. Columns 2 to 10 show the area of deforestation in the individual states. These are Acre (AC), Amazonas (AM), Amapá (AP), Maranhao (MA), Mato Grosso (MT), Para (PA), Rondonia (RO), Roraima (RR) and Tocantins (TO). The values in the fields are all positive and whole numbers. With the exception of years, the unit is square kilometres (km^2).

The 'CO_2 and Greenhouse Gas Emissions' dataset is available as structured data in CSV, XLSX and JSON formats. In order to create a standardised basis for further analysis, the CSV data type is also used. The data set comprises a total of 47,416 lines. The first row represents the header, while the remaining rows are listed per year and location. The dataset also includes 79 columns containing data on CO_2 emissions (annual, per capita, cumulative and consumption-based), other greenhouse gases, the energy mix and other relevant key figures. Only the columns relevant to the project are presented below.

- **"Country"**: Geographic location.
- **"Year"**: Year of observation.
- **"co2"**: Annual total emissions of CO_2, excluding land-use change.
- **"co2_including_luc"**: Annual total emissions of CO_2, including land-use change.
- **"cumulative_co2"**: Total cumulative emissions of CO_2, excluding land-use change, since the first

year of available data
- **"cumulative_co2_including_luc"**: Total cumulative emissions of CO_2, including land-use change, since the first year of available data.
- **"cumulative_luc_co2"**: Cumulative emissions of CO_2 from land-use change since the first year of available data
- **"land_use_change_co2"**: Annual emissions of CO_2 from land-use change

Data Quality

The data of the 'Brazilian Amazon Rainforest Degradation' data set was collected by the 'National Institute for Research in Brazil' and processed using the 'PRODES' programme. The high accuracy of the data set is due to the use of high-resolution satellite images and a transparent methodology. The data set is complete and has no missing values. The number formats are consistent. The timeliness of the data is guaranteed for the period from 2004 to 2019, although more recent values are missing, which limits the informative value of the data. Nevertheless, the data is relevant for answering the question.

The data from the 'CO_2 and Greenhouse Gas Emissions' dataset also comes from a credible source and is based on scientifically sound methods. Although the dataset has a certain number of missing values, the columns that will be extracted at a later stage as part of the data pipeline are complete. The data is organised in a structured and consistent way in terms of columns and number formats. The data set covers the period from 1751 to 2022 and therefore represents a comprehensive database. The relevance of the data for answering the research question results from the existence of a column containing CO_2 emissions from land use. A separate column that exclusively records CO_2 emissions from deforestation would be desirable, but no corresponding data could be found. It should be noted that the term 'land use' also includes other activities such as agriculture or livestock farming in addition to deforestation.

Licensing Information

The dataset 'Brazilian Amazon Rainforest Degradation' is licensed under the Creative Commons licence CC0: Public Domain [1]. This implies that the author completely relinquishes all rights to a work and leaves it to the public for free use. Consequently, the data may be used for this project.

The second dataset 'CO_2 and Greenhouse Gas Emissions' is licensed under the Creative Commons BY licence [2]. Use of the data is permitted provided that the name of the author and the licence information are correctly stated. I will comply with the licence conditions of the Creative Commons BY licence by correctly indicating the name of the author and the licence information when using the data from the 'CO_2 and Greenhouse Gas Emissions' dataset.

Data Pipeline

Overview of the Data Pipeline

Python 3.12.6 is used for the data pipeline. The process follows the so-called ETL model, whereby the individual steps are as follows: Extract, Transform and Load. The first step is to authenticate the Kaggle API in order to download the data set on deforestation in the Brazilian Amazon region. The dataset on CO_2 and greenhouse gas emissions is downloaded from an external URL via the requests library. Both data sets are then loaded, cleaned and transformed using the Pandas library. The cleansed data is stored in two separate SQLite databases using the SQLite3 library. The os library is used to manage the file paths and execute operations. The individual steps of the pipeline are explained in more detail below.

Data Extraction The data pipeline begins with the authentication of the Kaggle API. This takes place via the KaggleAPI client, whereby the user's API access data is used via the `authenticate()` method. This ensures access to the data records hosted by Kaggle. The data set on deforestation in the Amazon region can therefore be downloaded from Kaggle. Furthermore, the dataset on CO_2 and greenhouse gas emissions is downloaded via a URL from GitHub. Both data sets are available in CSV format and are saved in the local folder '../data'. The data is then processed further in Panda's dataframes.

Data Cleaning As part of the further analysis, only a selection of columns from the 'CO_2 and Greenhouse Gas Emissions' dataset is taken into account, so that the relevant data is focussed on. As part of the data reduction, the columns of the data set mentioned in the 'Structure of the data' section are retained. Furthermore, the 'country' column is filtered for Brazil and the 'year' column for the years 2004 to 2019, as the data set on the degradation of the Brazilian Amazon rainforest only covers these years. It is not necessary to adjust the Brazilian Amazon Rainforest Degradation dataset, as it does not contain any columns or missing values that are irrelevant for the project.

Data Loading The adjusted data set on the degradation of the Brazilian Amazon rainforest is stored in an SQLite database (deforestation_data.sqlite). This is done using the `to_sql()` method provided by Pandas, which saves the data as a table in an SQLite

country	year	co2	co2_including_luc	cumulative_co2_including_luc	cumulative_luc_co2	land_use_change_co2
Brazil	2004	361.434	2560.971	105928.805	97239.344	2199.537
Brazil	2005	364.371	2279.558	108208.359	99154.523	1915.187
Brazil	2006	368.871	2091.557	110299.922	100877.211	1722.686
Brazil	2007	390.573	1906.509	112206.43	102393.148	1515.936

Figure 1: *first five columns of the file 'co2_data.sqlite'.*

database. If an existing table already exists, the new data is integrated into this table so that it replaces the current data. The corrected data on CO_2 and greenhouse gas emissions are stored in a separate SQLite database (co2_data.sqlite) in the same way.

Meta-Quality Measures

The pipeline uses data validation to ensure that the required columns are complete and that there are no empty data frames. In the event of errors, such as missing data or network problems, error handling is initiated, which is supported by try-except blocks and meaningful log messages. Robustness in the event of missing data is ensured by preventive validation so that only complete and correct data is saved. Comprehensive logging ensures transparency by documenting all steps and errors in the pipeline.

Results and Limitations

Output Data

The data pipeline is output in the form of two SQLite databases, which contain the cleansed and filtered data from the original CSV tables. The database is complete and has no missing values. The SQLite format was chosen for the output of the data pipeline as it is characterised by low complexity and can be easily integrated into various programming languages, including Python and R. Another advantage of SQLite databases is that they are portable and can be easily shared with others. For better comprehensibility, the first five lines of the data sets are shown in the figures 1 and 2.

Ano/Estados	AC	AM	AP	MA	MT	PA	RO	RR	TO	AMZ LEGAL
2004	728	1232	46	755	11814	8870	3858	311	158	27772
2005	592	775	33	922	7145	5899	3244	133	271	19014
2006	398	788	30	674	4333	5659	2049	231	124	14286
2007	184	610	39	631	2678	5526	1611	309	63	11651

Figure 2: *first five columns of the file 'deforestation_data.sqlite'.*

Critical Reflection

A critical reflection on the data reveals several potential challenges. One key aspect is the limited up-to-dateness of the data set on deforestation. The data set ends in 2019, meaning that more recent developments cannot be taken into account in this context. This can affect the validity of the results, especially if significant political or environmental changes have taken place since 2019. Another problem is the diversity of the time periods to which the data relates. For example, one data set covers the period from 2004 to 2019, while the other data set covers the period from 1751 to 2022. This means that a considerable part of the data set on CO_2 data, which covers the longer period, cannot be used. Another problem is the lack of specialization of the CO_2 dataset. The dataset only includes CO_2 emissions from land use, but no specific values for CO_2 emissions from deforestation. This makes it difficult to derive a direct correlation between deforestation and emissions.

References

[1] Mariana Boger Netto. *Brazilian Amazon Rainforest Degradation*. https://www.kaggle.com/datasets/mbogernetto/brazilian-amazon-rainforest-degradation. Accessed: 2024-11-27. 2019.

[2] Pablo Rosado Hannah Ritchie and Max Roser. *CO2 and Greenhouse Gas Emissions*. https://ourworldindata.org/co2-and-greenhouse-gas-emissions. Accessed: 2024-11-27. 2024.