Deforestation Correlation with Air Pollution in Amazon

Introduction

Human activity has led to environmental challenges like deforestation and air pollution. The Amazon plays a crucial role in climate regulation, but deforestation threatens its functions. This led to the following question:

How does deforestation in the Amazon correlate with pollution outcomes in Brazil, specifically in São Paulo?

Data Sources

Global Forest Watch - Amazon Deforestation (SAD Alerts)

• **Source**: Global Forest Watch

• **Data Description**: The dataset contains geo-referenced deforestation and degradation alerts for the Brazilian Amazon from 2008 to 2018, with attributes like date, area affected, and spatial data.

• **Reason for Choosing:** relevance to understanding deforestation's impact on global environmental changes, including pollution in São Paulo.

• **Data Structure and Quality**: It includes time-stamped records with geographic coordinates and may have some inaccuracies, addressed during preprocessing. The dataset is updated hourly and available as a CSV.

• **License**: Open Data Commons Attribution License (ODC-By), which requires proper attribution in any published work.

Kaggle- Air Pollution at São Paulo, Brazil, Since 2013

• **Source**: Kaggle

• **Data Description**: This dataset provides hourly air pollution data for São Paulo, Brazil, covering pollutants like MP10, O3, NO2, CO, MP2.5, SO2, benzene, and toluene, spanning from 2013 onward.

• **Reason for Choosing**: It's relevant for understanding the impact of Amazon deforestation to pollution levels.

• **Data Structure and Quality**: The data includes time-stamped readings with some gaps, stored in a CSV format for easy integration with the deforestation data.

• **License**: Open Data License (Kaggle dataset terms apply) which requires proper attribution in any published material.

Data Pipeline Overview

The pipeline automates the download, cleaning, transformation, and storage of two datasets: deforestation from Global Forest Watch and air pollution from Kaggle. It is implemented in Python, using libraries such as pandas, requests, matplotlib, seaborn, and sqlite3.

Pipeline Steps:

1. **Data Acquisition:**

   Deforestation data is fetched via the Global Forest Watch API, and air pollution data is retrieved using the Kaggle API.

2. **Data Cleaning & Transformation**:
   - **Deforestation Data:** Filtered for major deforestation events, columns renamed, time format standardized, and data aggregated by month. Missing values were interpolated.
   - **Pollution Data:** Columns renamed, missing rows dropped, and pollutant levels averaged monthly. Time format standardized, and missing pollutant values filled with zero.

3. **Data Storage**:

   The cleaned data is saved into SQLite databases: deforestation.db for aggregated deforestation data and air_pollution.db for averaged pollutant levels.

4. **Visualizations**:

Several plots were created to analyze trends and correlations:

- **Deforestation Trend Plot:** A line plot of the total affected area over time.
- **Pollutant Trend Plots:** Line plots for each pollutant (e.g., PM10, CO, NO2).
- **Correlation Heatmap:** The correlation between deforestation and pollutants.
- **Scatter Plots:** Deforestation vs. pollutants (e.g., PM10, PM2.5) to explore correlations.

5. **Meta-Quality Measures & Error Handling**:

- **Error Handling:** The pipeline retries failed downloads to ensure reliability.
- **Changing Input Data:** Pipeline fetches the latest data, avoiding issues.
- **Quality Assurance:** Data is validated after transformations to handle missing values and anomalies**.**

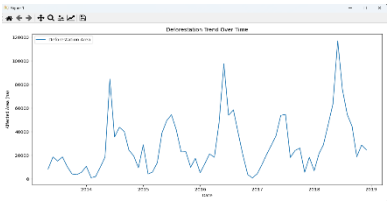


Fig. 1. Deforestation trend (monthly)

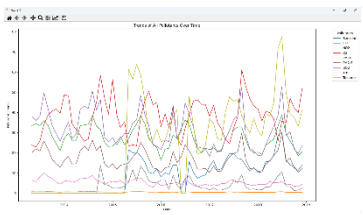


Fig. 2. Trend of all pollutants

Fig. 3. Correlation heatmap

Results, Limitations, and future prospect

Output Data Structure and Quality:

- **Structure**: The final output consists of two SQLite databases: deforestation.db and air_pollution.db. Each stores cleaned and aggregated data in structured tables.
    - **Deforestation Data**: Includes columns for date ("Date") and total affected area ("AffectedArea") from 2013 to 2018.
    - **Pollution Data**: Contains pollutants (e.g., PM10, PM2.5, CO, NO2) with their monthly averages from 2013 to 2018.
- **Quality**: The data is clean, with missing values interpolated and aggregated monthly. Data quality depends on the integrity of the source, especially with missing values.

Output Data Format and Reasoning:

- **Format**: The output is stored in SQLite format, chosen for its simplicity and efficiency in handling structured data. It enables easy querying and integration with other tools, making it ideal for local data storage.
- **Reasoning**: SQLite is used because it doesn't require a server, is easily queried, and ensures data consistency and reliability.

Critical Reflection on Data and Limitations:

- **Data Limitations**:
    - Deforestation data has gaps or errors due to missing alerts or delayed reporting, affecting trend accuracy.
    - The overlapping period between deforestation and pollution data may not fully capture the long-term effects of deforestation on pollution.
    - Pollution data may be incomplete or inconsistent, addressed by filling gaps with zeros or interpolation.
- **Anticipated Issues**:
    - **Correlation Validity**: External variables like transportation or seasonal weather effects may distort the correlation between deforestation and pollution. Possible solutions include studying longer periods, filtering industry-related pollutants, and using better statistical methods or Machine Learning approaches.
    - **Data Resolution**: Monthly aggregation may obscure short-term variations. Solutions include studying different temporal resolutions, using advanced interpolation methods, and investigating specific events like forest fires.

Output Visualization:

The final output is stored in deforestation.db and air_pollution.db, and a sample visualization will be provided.

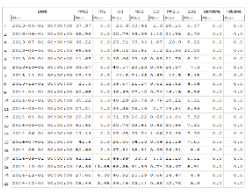
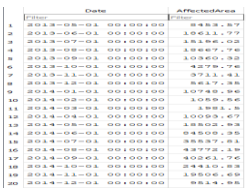


Fig. 5. Pollution.db



	Date	AffectedArea
1	2013-05-01	9483.57
2	2013-06-01	10611.77
3	2013-07-01	15366.03
4	2013-08-01	13667.76
5	2013-09-01	10360.32
6	2013-10-01	4279.76
7	2013-11-01	3711.41
8	2013-12-01	5547.36
9	2014-01-01	10749.96
10	2014-02-01	1059.56
11	2014-03-01	1993.5
12	2014-04-01	10093.67
13	2014-05-01	15502.93
14	2014-06-01	14500.35
15	2014-07-01	9587.81
16	2014-08-01	13772.19
17	2014-09-01	10261.76
18	2014-10-01	2481.83
19	2014-11-01	1956.65
20	2014-12-01	9514.96

Fig. 4. Deforestation.db