Exploratory Data Analysis of Climate and Land-Use Data

Andressa Silva de Oliveira
Insper Institute of Education and Research
São Paulo, Brazil

Fabrício Neri Lima
Insper Institute of Education and Research
São Paulo, Brazil

Isabelle Moschini Murollo
Insper Institute of Education and Research
São Paulo, Brazil

Vinicius Grando Eller
Insper Institute of Education and Research
São Paulo, Brazil

***Abstract*—This study investigates the growing frequency and intensity of extreme weather events, analyzing the influence of environmental variables such as deforestation, urbanization, and land-use changes. By utilizing the FAO's Land Cover and Forest Area dataset and the EM-DAT disaster database, the research examines global disaster patterns, with a focus on the relationship between environmental factors and disaster occurrences. A comprehensive analysis was conducted, including data cleaning, correlation analysis, and regression modeling, particularly targeting Brazil's climate behavior. The results underscore the critical role of land-use changes in shaping disaster vulnerability and highlight the potential of machine learning models to predict temperature trends. The findings emphasize the need for region-specific mitigation strategies and demonstrate how data science can inform climate resilience efforts and sustainable policies. This work provides meaningful contributions to the ongoing dialogue on climate change, sustainable environmental management, and disaster risk mitigation.**

***Index Terms*—extreme weather events, climate change, deforestation, land-use, disaster risk, machine learning, regression analysis, Brazil, environmental policy**

I. Introduction

In the current context, extreme weather events have become increasingly frequent and intense, leading to devastating consequences for populations, biodiversity, and global economies. This trend is linked to global climate change, influenced by factors such as deforestation, rapid urbanization, and greenhouse gas emissions. Investigating the relationship between these events and the regions in which they occur is therefore crucial to understanding the associated environmental and socioeconomic impacts, enabling the development of more effective mitigation strategies.

This project explores the connection between extreme weather events and various environmental variables, such as the expansion of agricultural areas and forest degradation. To achieve this, the following datasets were utilized: FAO (Land Cover and Forest Area), which provides detailed information on land cover and forest areas, and EM-DAT, which documents natural disasters on a global scale. These robust data sources facilitate the assessment of global patterns, regional comparisons, and historical analyses, contributing to the identification of climate trends and hotspots that demand urgent attention.

The aim of this study is to understand the causes and consequences of interactions between natural and human factors and to provide insights that can support public policies and sustainable initiatives. Moreover, analyzing these relationships is particularly relevant in the present circumstances, where humanity faces challenges such as increasing climate inequalities and the urgent need to adopt more responsible environmental practices. This project seeks not only to generate insights of relevance to this context but also to highlight the importance of data science in creating solutions for global climate challenges.

II. Methodology

A. Data Sources

To conduct the analyses in this project, two primary datasets were utilized: FAO (Land Cover and Forest Area) and EM-DAT (Emergency Events Database).

- FAO (Land Cover and Forest Area): This dataset contains annual data from 1992 to 2020 on land use, including types of land cover and climate-related indices. It was published in 2022 by FAOSTAT (Food and Agriculture Organization Statistics). FAOSTAT is a global database maintained by the United Nations' Food and Agriculture Organization (FAO) that has been providing agricultural, food, and environmental statistics from countries since 1961.
- EM-DAT (Emergency Events Database): The EM-DAT database presents data on the number and types of natural disasters, grouped annually, enabling analyses of the frequency and intensity of weather events over time. It was published in 2023 by the Emergency Events Database (EM-DAT), developed by the Centre for Research on the Epidemiology of Disasters (CRED). EM-DAT is a global database that records and analyzes data on natural and technological disasters.

These datasets provide a robust foundation for examining long-term trends and performing comparative regional analyses.

B. Data Processing and Analysis

The main steps regarding the processing and analysis of the data acquired are listed below.

- Data Cleaning: Addressed missing values and outliers.
- Exploratory Data Analysis: Utilized summary statistics, correlation analysis, and visualizations (e.g., histograms and scatter plots).
- Temporal Analysis: Trends in deforestation and agricultural activity were examined using time-series models.
- Correlation and Regression: Statistical tests and regression models quantified relationships between land-use factors and disaster frequencies.
- Analysis Focused on Brazil: Prediction of climate and environmental behaviors in Brazil.

III. Results

The following section outlines the key findings of the analysis, emphasizing significant trends and patterns in natural disaster occurrences globally.

A. Most Affected Countries in Recent Years

Over the past five years, the United States recorded the highest number of natural disasters, totaling 154 events. This finding emphasizes the vulnerability of highly developed and urbanized nations to extreme weather events, likely influenced by their large geographic size and diverse climates.

B. Major Types of Natural Disasters

The most frequent types of natural disasters globally include:

- Floods: 4,548 events.
- Storms: 3,047 events.
- Landslides: 558 events.
- Extreme Temperatures: 520 events.
- Droughts: 489 events.
- Wildfires: 369 events.

Floods and storms alone account for the majority of disasters, underscoring the role of weather-related phenomena as key drivers of natural disasters.

C. Year with the Most Natural Disasters

The year 2005 reported the highest number of disasters, reaching 406 events. This aligns with a particularly active hurricane season in the Atlantic that year, as well as other global extreme weather patterns.

D. Increasing Trend in Disaster Frequency

A clear increasing trend in the number of natural disasters over time was observed, supported by statistical analysis:

- Slope (Rate of Increase): 4.09 disasters per year.
- R^2 (Goodness of Fit): 0.37.
- p-value: 0.0003 (statistically significant).

The graph in 'Fig. 1" highlights a steady rise in disasters, indicating worsening global conditions influenced by factors such as climate change and deforestation.

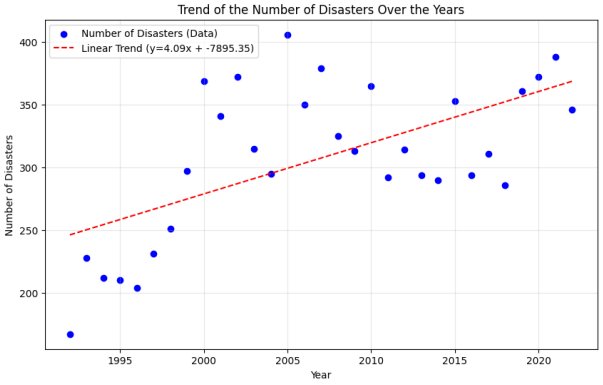

Fig. 1. Trend graph of disaster frequency.

E. Analysis and Predictions Focused on Brazil

In order to better understand the correlation between different environmental aspects and natural disasters, it was necessary to choose a country to be carefully analyzed. Thus, the country chosen was Brazil due to its continental size and various ecosystems and biomes.

Based on the heatmap matrix from 'Fig. 2" and the utility of the variables analyzed, certain variables are more meaningful for prediction. Specifically, the cumulative temperature change stands out as an important target variable, as it captures how temperatures have been increasing or decreasing over the years.

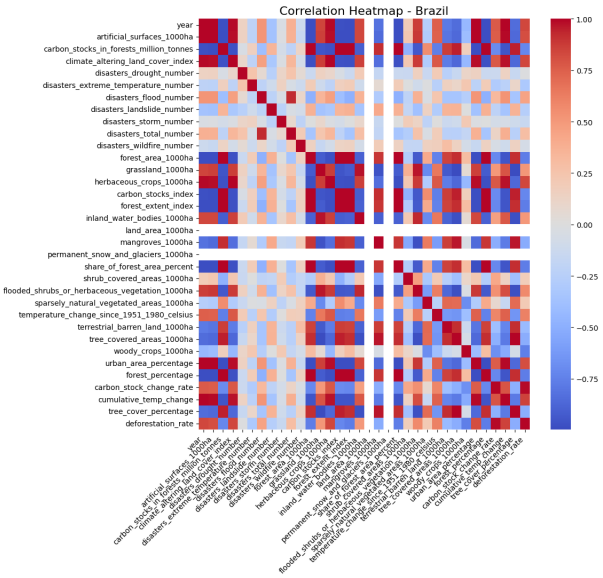

Fig. 2. Correlation Heatmap for Brazil.

The analysis identifies all columns with a relationship to cumulative temperature change as relevant features. These include variables directly or indirectly associated with environmental changes, such as artificial surfaces, forest areas, carbon stocks in forests, deforestation rates, and tree-covered areas. These variables are particularly relevant as they provide

insights into both anthropogenic and natural factors affecting temperature variations.

For this analysis, the cumulative temperature change was defined as the target variable. The selected features from the dataset that are closely linked to this target include:

- Artificial surfaces (in 1,000 hectares);
- Forest area (in 1,000 hectares);
- Carbon stocks in forests (in million tonnes);
- Deforestation rate;
- Tree-covered areas (in 1,000 hectares).

These features align with the overarching goal of the study, which is to model and predict the cumulative effects of climate-related factors on temperature dynamics. The inclusion of these variables ensures a robust approach to understanding the complex interplay between land use, forest management, and climate change.

Afterwards, several prediction models were applied and evaluated by its performance. The dataset was divided into independent variables (features) and the target variable. Subsequently, the data was split into training and testing sets to ensure unbiased evaluation. Given that the target variable is numerical, a selection of regression models was tested to identify the most suitable predictor.

A grid search methodology was employed to optimize model hyperparameters systematically. This process aimed to maximize performance by evaluating a set of candidate regressors, including Linear Regression, Ridge, Lasso, Random Forest, and Gradient Boosting. The models were assessed based on their Mean Squared Error (MSE) and R^2 score, ensuring a robust comparison, as shown in the bar chart 'Fig. 3", providing a clear visualization of their relative performances.

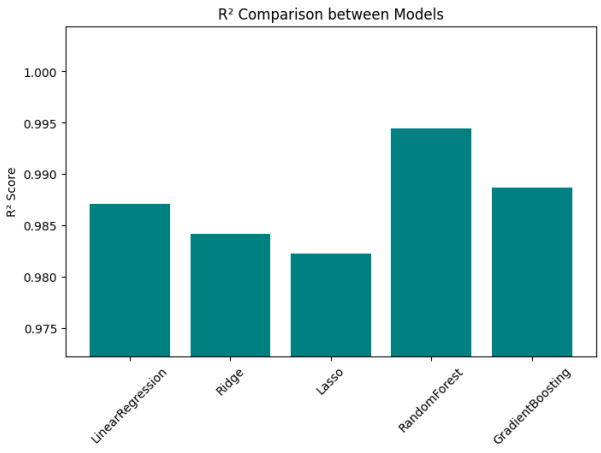

Fig. 3. R^2 score comparison across prediction models.

The performance comparison revealed that the Random Forest Regressor achieved the best results, with the highest R^2 score, albeit with only a marginal improvement over other models. This indicates that Random Forest effectively captures the complex relationships within the dataset while maintaining low prediction error.

Additionally, a scatter plot of actual versus predicted values for the best-performing model (Random Forest) was generated ('Fig. 4"). This plot demonstrated a strong agreement between the real and predicted values, showcasing the model's ability to describe the expected behavior of the data.

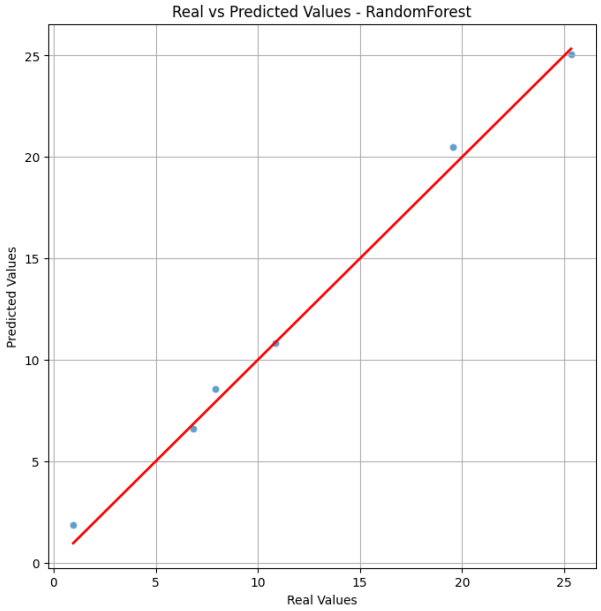

Fig. 4. RandomForest model: comparison between predicted and real values.

Despite the limited dataset size, the Random Forest model showed considerable potential in predicting temperature changes in Brazil. The selected features—spanning land use, forest coverage, and deforestation rates—proved effective in capturing relevant trends. With this model, it is feasible to forecast cumulative temperature changes, aiding in future climate assessments and decision-making processes for environmental management.

IV. Discussion

The results provide critical insights into global natural disaster patterns, their drivers, and trends over recent decades. This analysis highlights the urgent need for targeted mitigation and adaptation strategies, especially in high-risk regions such as the United States, which is notably vulnerable to extreme weather events due to its geographic size and climatic diversity.

Floods and storms dominate as the most frequent disaster types globally, underscoring the increasing impact of weather-related phenomena. These findings emphasize the intricate relationship between human-induced land-use changes and the vulnerability to extreme events. For example, deforestation and urbanization may exacerbate flooding and temperature anomalies, highlighting the importance of integrating sustainable land management policies with climate resilience efforts.

The study of Brazil, focusing on cumulative temperature change as the target variable, demonstrated the utility of

data-driven approaches to understanding and predicting environmental behavior. By isolating relevant features—such as deforestation rates and forest area—the analysis provided a localized context for predicting temperature changes. This reinforces the notion that global datasets must often be adapted for regional analyses, as environmental behaviors and their consequences are highly localized.

Moreover, the predictive modeling revealed the potential of machine learning techniques like Random Forest to effectively describe and anticipate temperature trends. These models are valuable tools for exploring complex environmental interactions and forecasting future scenarios, even when dataset sizes are limited.

Overall, this research underscores the importance of leveraging regional data for addressing specific environmental challenges. By combining land-use policies with advanced predictive models, countries can better prepare for and mitigate the impacts of climate change, ensuring more resilient ecosystems and communities.

V. Conclusion

This study illustrates the importance of leveraging global datasets to understand the dynamics of extreme weather events and land-use changes. The results underscore the necessity for coordinated policy interventions, emphasizing sustainable land-use practices and climate resilience. Furthermore, the analysis highlights the critical role of data science in tackling global climate challenges.

This study highlights the power of data-driven approaches in understanding and addressing the increasing frequency and severity of natural disasters. By leveraging comprehensive datasets and advanced analytical techniques, key insights were derived into the trends, drivers, and localized impacts of environmental changes. Globally, weather-related disasters such as floods and storms emerged as the most prevalent, underscoring the growing influence of climate change and land-use practices.

The focused analysis on Brazil demonstrated the critical importance of regional perspectives, where cumulative temperature change was effectively modeled using relevant features like deforestation rates, forest area, and land-use metrics. The Random Forest model stood out in predicting temperature trends, illustrating the potential of machine learning in forecasting environmental phenomena despite data limitations.

The study's comprehensive methodology—ranging from data cleaning and exploratory analysis to regression modeling—provides a replicable framework for analyzing environmental and climate data. By integrating these insights with targeted policy interventions, nations can enhance resilience, prioritize resource allocation, and mitigate the adverse effects of climate change. This work emphasizes the need for localized analyses within global contexts to better understand and address the multifaceted challenges posed by natural disasters.

References

Citation of Data Sources

[1] FAO, 2022. *FAOSTAT Land, Inputs and Sustainability, Land Use.* Available at: https://www.fao.org/faostat/en/#data/RL.

[2] The Emergency Events Database (EM-DAT), Centre for Research on the Epidemiology of Disasters (CRED), UCLouvain.

Articles Using the Land Cover and Forest Area Database (FAO)

[3] Rosan, T. M. et al. *A multi-data assessment of land use and land cover emissions from Brazil during 2000–2019.* Environmental Research Letters, 16(074004), 2021. Available at: https://iopscience.iop.org/article/10.1088/1748-9326/ac08c3/meta.

[4] Dooley, K. et al. *Over-reliance on land for carbon dioxide removal in net-zero climate pledges.* Nature Communications, 15:9118, 2024. Published on October 23, 2024. Available at: https://www.nature.com/articles/s41467-024-53466-0.

[5] Reiner, F. et al. *More than one quarter of Africa's tree cover is found outside areas previously classified as forest.* Nature Communications, 14:2258, 2023. Published on March 29, 2023. Available at: https://www.nature.com/articles/s41467-023-37880-4.

Articles Using the Extreme Climate Disasters Database (EM-DAT)

[6] Newman, R.; Noy, I. *The global costs of extreme weather that are attributable to climate change.* Nature Communications, 14:6103, 2023. Published on September 29, 2023. Available at: https://www.nature.com/articles/s41467-023-41888-1.

[7] Dong, C. *Indo-Pacific regional extremes aggravated by changes in tropical weather patterns.* Nature Geoscience, 17:979–986, 2024. Published on October 4, 2024. Available at: https://www.nature.com/articles/s41561-024-01537-8.

[8] Balaian, S. K.; Sanders, B. F.; Qomi, M. J. A. *How urban form impacts flooding.* Nature Communications, 15:6911, 2024. Published on August 19, 2024. Available at: https://www.nature.com/articles/s41467-024-50347-4.