MADE Report: Analysis of median household income and environmental impact metrics, deforestation and CO_2 emissions across U.S. states

Nishant Mishra, 23178030
FAU Erlangen-Nürnberg

Abstract
This research investigates how median household income relates to specific environmental measures—like deforestation and CO_2 emissions—across the states of the United States over the years from 1990 to 2021. Leveraging data at the state level, this project implements an ETL pipeline to achieve data consistency between two datasets and further analyze it. Then, it undertakes a fairly standard exploratory data analysis (EDA) to illuminate some intriguing features and disparities in the data across different regions of the United States. Correlation and regression analyses are conducted to access the temporal trends and averaged relationship between income and environmental metrics. The results establish weak and inconsistent correlations, highlighting the complexity of socioeconomic and environmental interactions and underlining the importance of integrating additional factors to inform sustainable policy development.

1 Introduction

Understanding the interplay between socioeconomic factors and environmental impacts is increasingly established in light of global efforts to achieve sustainability and development. One such avenue, examining how median household income relates to deforestation and CO_2 emissions, offers valuable insights into the dynamics of economic growth and the environment. This investigation is based on statistical theories and empirical results and tries to provide a data-based approach to analyzing correlations between these interdependencies. The study aims to examine the median income of US states and address the question: Is there a significant relationship between median household income and environmental impacts—such as deforestation and CO_2 emissions—across U.S. states?

1.1 Literature Review

The Environmental Kuznets Curve (EKC) hypothesis is a bedrock for understanding the income-environment relationship. It states that environmental degradation initially worsens with increasing income but improves once a threshold for economic development is reached[4]. Empirical studies in the literature provide mixed evidence, with some indicating that many US states have dissociated CO_2 emissions and income growth, transitioning to renewable energy sources and sustainable practices[6]. However, recent trends suggest that the EKC may have taken a U-shaped form in specific sectors, such as transportation and electricity[6].

Higher-income households significantly contribute to CO_2 emissions due to elevated consumption levels. Starr et al.[3] demonstrate that the top 10% earners in the US were responsible for approximately 40% of total emissions in 2019, primarily due to their consumption and investment patterns, highlighting the growing emissions inequality within high-income nations such as the United States[5]. Additional research suggests that alternative taxation policies, such as income-based carbon taxes, could mitigate these disparities[3].

The relationship between income and deforestation is similarly nuanced. Higher-income regions may invest more in urban tree cover and afforestation programs, thereby reducing deforestation[2]. In contrast, economic activities related to income growth, such as urbanization and agricultural expansion, often aggravate forest loss, particularly in tropical regions[1]. These opposing dynamics underscore the complexity of the significance of income, environment, and the need for targeted economic and tax policies.

This project seeks to address and find the relationship between median household income and environmental impacts, specifically focusing on deforestation and CO_2 emissions in US states. Using state-level data and employing an ETL approach, the research aims to contribute to a deeper understanding of the socioeconomic drivers of environmental change. The findings will inform policymakers about strategies to balance economic development with ecological sustainability, addressing disparities while advancing climate goals.

2 Pipeline and Exploratory Data Analysis

2.1 ETL Pipeline

The study employs a structured ETL (Extract, Transform, Load) pipeline to process and integrate datasets, ensuring clean and consistent data for analysis. Implemented in Python, the pipeline utilizes libraries like `pandas` for data manipulation, `SQLite3` for database operations, and `requests` for downloading external raw datasets into the pipeline environment.

During the **Extract** phase, median income data was sourced from the National Center for Education Statistics (NCES) website as an Excel spreadsheet containing state-level median household income data from 1990 to 2021. Additionally, deforestation data was obtained from Global Forest Watch (GFW) as an Excel file that included statistics on deforestation, tree cover loss, and CO_2 emissions due to land-use changes.

The **Transform** phase focused on data cleaning, integration, and normalization. Unnecessary rows and columns were removed, such as `summary and unnamed columns`, and column names between two datasets were standardized. State-level data was retained to ensure alignment. The datasets were then merged on the `State` and `Year` columns, with missing values dropped from the joined dataset to maintain dataset integrity.

Finally, in the **Load** phase, the processed dataset was loaded into an SQLite database named `analysis.db`, containing a single table, `US_Analysis`. This structured format facilitates seamless integration with correlation analysis tools, ensuring a reliable data processing pipeline.

2.2 EDA

The line plots in fig 1 capture regional disparities in median income trends, clearly showing differences between northern, southern, and central states. While effective in illustrating trends, the plots do not incorporate error margins or other socioeconomic indicators, which could add depth to the analysis. Adding more granular detail or supplementary metrics could enhance the objective.

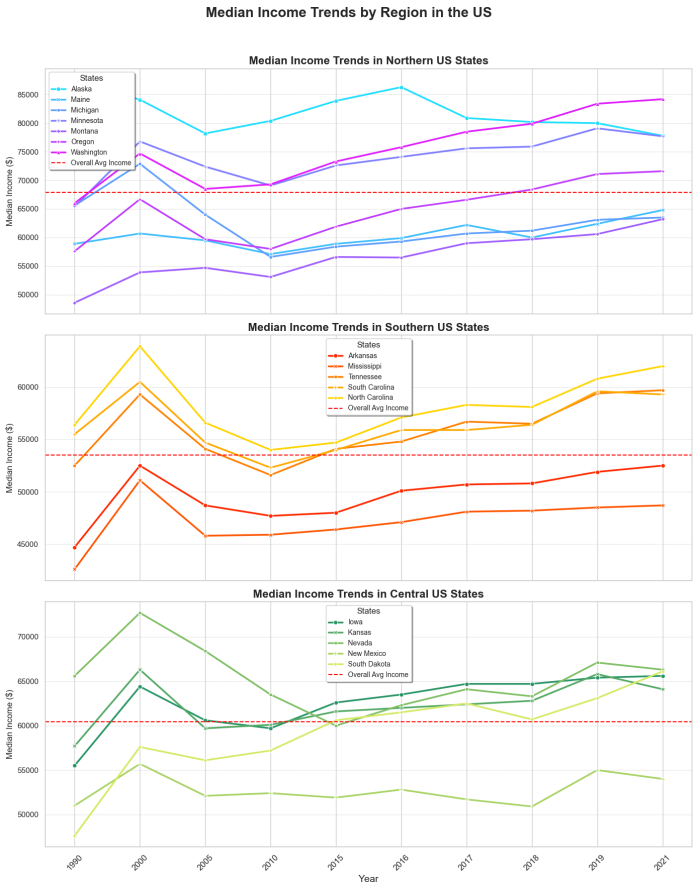

Figure 1: Regional Trends in Median Income Across Northern, Southern, and Central US States (2005–2021): The plots describe income disparities, with northern states showing higher and more stable incomes, while southern and central regions show a greater variability and generally lower income levels compared to the national average

3 Correlation Analysis

3.1 Temporal Trends in Correlation

The correlation trends reveal varying relationships between income and environmental metrics, as illustrated in fig 2. For income and deforestation, the correlation peaked at 0.246 in 2015 but declined to -0.128 in 2017, eventually rebounding to 0.180 by 2021. This inconsistency highlights the absence of a stable relationship between income and deforestation. Similarly, the income-CO_2 emissions correlations were predominantly negative, reaching their most substantial value of -0.217 in 2010 before trending slightly positive to 0.170 in 2021. These fluctuations underscore the dynamic nature of income's influence on environmental outcomes, likely driven by external factors such as policy interventions and technological advancements like the introduction of EV vehicles.

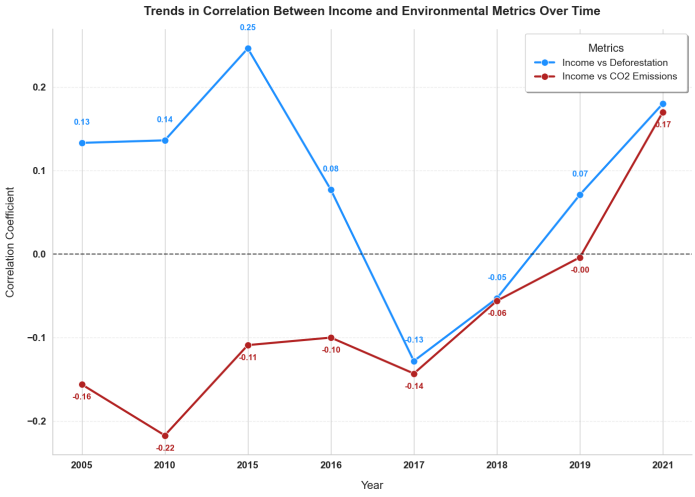

Figure 2: Temporal correlation trends between income and environmental metrics (deforestation and CO_2 emissions) from 2005 to 2021.

3.2 Regression Analysis on Averaged Data

The regression analysis on averaged income and environmental data corroborates these observations, showing weak relationships between income and environmental metrics. For deforestation, the OLS regression substantiated an R-squared value of 0.012, with a p-value of 0.609, indicating no statistically significant predictive power of average income on deforestation levels in the US. Similarly, for CO2 emissions, the regression results (R-squared = 0.006, p-value = 0.705) suggest that income explains only a negligible fraction of the variance. Both models' lack of statistical significance implies that income alone is not a strong determinant of environmental changes in the US.

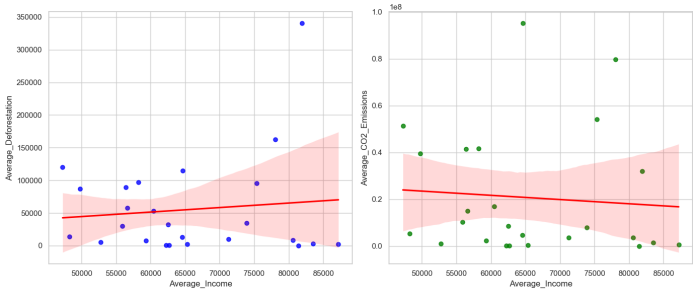

Figure 3: Scatter plots showing average income vs. deforestation (left) and CO_2 emissions (right) with regression lines and confidence intervals.

3.3 Reflections

The findings reveal two interpretative challenges. First, the low R-squared values and non-significant coefficients highlight income inadequacy as a sole predictor of deforestation or CO_2 emissions, suggesting that additional factors like industrial activity, geographic features, and/or environmental policies must be considered. Second, the large condition numbers in the regression models point to potential multicollinearity issues or numerical instability, which usually undermine the reliability of the results. Addressing these limitations through feature engineering or advanced modeling techniques like causal analysis could provide a more nuanced understanding of the socioeconomic drivers of environmental changes.

4 Conclusion

The analysis reveals a weak and inconsistent relationship between median household income and environmental metrics, with temporal trends indicating fluctuating correlations. While deforestation shows positive and negative correlations in Spearman correlation analysis with income, CO_2 emissions exhibit predominantly negative associations, which is weak and statistically insignificant. Regression analyses further affirm income's limited predictive power for these metrics, pointing to the influence of other confounding factors such as industrial activity, geographic characteristics, and policy interventions. The study's insights aim to guide policymakers and researchers in exploring these avenues and a demonstration of an ETL pipeline that can help gather multi-source data, transform it, and load it ready for further analysis.

References

[1] Marie Boltz, Philippe Delacote, and Kenneth Houngbedji. *Deforestation and Development: How Do Forests and Population Living Standards Coevolve*, page 1–22. Springer International Publishing, Cham, 2020.

[2] Robert I. McDonald, Tanushree Biswas, Cedilla Sachar, Ian Housman, Timothy M. Boucher, Deborah Balk, David Nowak, Erica Spotswood, Charlotte K. Stanley, and Stefan Leyk. The tree cover and temperature disparity in us urbanized areas: Quantifying the association with income across 5,723 communities. *PLOS ONE*, 16(4):e0249715, April 2021.

[3] Jared Starr, Craig Nicolson, Michael Ash, Ezra M. Markowitz, and Daniel Moran. Income-based u.s. household carbon footprints (1990–2019) offer new insights on emissions inequality and climate finance. *PLOS Climate*, 2(8):e0000190, August 2023.

[4] David I. Stern. *The Environmental Kuznets Curve*. Elsevier, January 2018.

[5] Drexel University. Income inequality and carbon dioxide emissions have a complex relationship.

[6] Zuyi Wang and Man-Keun Kim. Decoupling of co2 emissions and income in the u.s.: A new look from ekc. *Climatic Change*, 177(3):52, February 2024.