Project Report

Sasha Matveev

2025-12-02

Climate Change Analysis in the United States: The Impact of Socioeconomic and Industry on Average Temperature

Net ID: Sasha Matveev (matveev2) Vladislav Fedorov (vvf2)

Abstract

The goal of the study is to investigate the relationship between greenhouse gas emissions and the average annual temperature in the United States from 1900 to 2023 while identifying the most influential factors affecting climate change over time. We used... We found...

Introduction

Climate change is listed as one of the most critical challenges facing humanity today by the United Nations. Encompassing everything from "...shifting weather patterns that threaten food production, to rising sea levels that increase the risk of catastrophic flooding, the impacts of climate change are global in scope and unprecedented in scale. Without drastic action today, adapting to these impacts in the future will be more difficult and costly"(UN source). There are numerous social, industrial, geographic, economic, and historic factors that contribute to climate change, making it a complex issue that is difficult to model. As such, over the course of this study we will focus primarily on the industrial factors such as greenhouse gas emissions and socio-economic factors such as changes in population and gross domestic product (GDP). This study aims to analyze the social and environmental factors that contribute the most to climate change in the United States at the individual level as well as in a combined setting, withe the impact tracked by the average annual temperature in the country from 1900 to 2022. Through regression modeling and time series analysis, we look to identify the strongest contributors to climate change as well as the factors that offer the best predictive power for forecasting future temperature changes in the US so that policy makers can better understand the issue and have reliable data to base potential mitigation strategies on.

Data Overview

The original emissions data comes from Our World in Data and was compiled by Shreyansh Dangi for use on kaggle. The data includes global carbon dioxide (CO2) and other greenhouse gas emissions across countries, and sectors while also combining variables describing the countries in the dataset such as population, GDP, and energy consumption. Each row is a country and a given year so the time unit is yearly. The analysis in this study will be focusing on the United States from 1900-2022. The variables of interest will include time (year), the population, GDP (inflation adjusted and cost-of-living adjusted in dollars), CO2 emissions (million tons), methane emissions (tons), nitrous oxide emissions (tons), as well as the average annual temperature in Fahrenheit from the supplementary dataset. The supplementary dataset containing the average annual temperature in the US comes from the National Centers for Environmental Information (NCEI) and was made available on kaggle by Gia Bách Nguyen. The data is based on observations from a network of thousands of weather stations across the United States in the same exact time period (1900-2022) as the primary dataset containing emissions data and is also yearly.

Statistical Methods

Analysis A

Preliminary analysis (nature of the data, data source, time series plot, preliminary analysis (Is there a trend? Is the variance increasing? Is there any pattern? Is this data stationary?)
The temperature time series was analyzed through exploratory data analysis techniques, including both statistical and visual inspections of graphs. The time series plot in Figure 1 does not show any seasonality but does show a very weak upward trend over time, as expected based on prior background knowledge.

Figure 1: Time Series Plot of Average Annual US Temperature (1900-2022)

The variance plot does not appear to show a pattern of either increasing or decreasing over time. Furthermore, to examine the variance through a statistical manner, a Levene's test was performed by separating the data into two groups chronologically up to the halfway point. With a p-value of 0.64 there was not enough statistically significant evidence to conlcude a difference of variance between our two groups, giving more credability to the lack of change in variance over time. Based on the results of the Augmented Dickey-Fuller test (p-value = 0.0224), we have enough statistically significant evidence to reject the null hypothesis of non-stationarity and conclude that the temperature time series is actually stationary. Thus we do not need to perform any further transformations to the data to achieve stationarity prior to fitting regression models.

Regression model suggestions (at least 2) In order to examine the questions posed by the study we will fit three regression models to predict the average annual temperature in the US. The models will differ only in the variables they are fitted on, with the first model will include all potential predictors, both emissions and social including : time (year), population, GDP, CO2 emissions, methane emissions, and nitrous oxide emissions. The second model aims to test the impact of social and demographic factors on the average annual temperature in the US. Therefore the second model will include only social predictors: time (year), population, and GDP. Lastly, the third model is constructed to isolate the emissions factors and examine the relationship between time (year) and the three most common types of emissions including CO2 emissions, methane emissions, nitrous oxide emissions with the average annual temperature in the US between 1900 and 2022.A succinct summary of the variables in the three models is shown in Table 1 below.

Table 1: Summary of Variables in Regression Models | Model | Predictors | |———————|———————————————-| | Full Model | year, population, gdp, co2, methane, nitrous_oxide | | Social Factors Only | year, population, gdp | | Emissions Only | year, co2, methane, nitrous_oxide |s

The reason behind the categories of variables chosen is to test the impact of varying climate change factors both independently and together. The full model will allow us to see the combined effect of both social and emissions factors on the average annual temperature in the US. The social factors only model will help us understand how much of the change in temperature can be explained by demographic and economic factors alone, while the emissions only model will isolate the effect of greenhouse gas emissions on temperature changes, which is often attributed as the driving force behind climate change in the time period examined in this study.

Regression result analysis The full model including all predictors, both social and emissions factors, showed that only the CO2 predictor was significant at alpha =0.05 level when performing the t-test. No other predictor came even close to being significant. The adjusted R-squared value for the full model was 0.3272 , indicating that approximately 32.72 % of the variance in average annual temperature in the U.S. between 1900 and 2022 can be explained by the proposed model.

The social factors only model showed that none of the predictors were significant at alpha =0.05 level. The closest variable to being significant was gdp with a p-value of 0.0599. Like in the full model, the year and population predictors were not significant. The adjusted R-squared value for the social factors only model was relatively high at 0.3147, meaning 31.47% of the variance in the average U.S. annual temperature between 1900 and 2022 can be explained by the social factors only model, where no predictor was significant at the alpha=0.05 level.

The emissions model had two predictors that were significant at the alpha=0.05 level, including CO2 and methane emissions with p-values of 0.000531 and 0.000520, respectively. The nitrous oxide predictor was not significant with a p-value of 0.120. The adjusted R-squared value for the emissions only model was 0.3314 , indicating that 34.13% of the variance in average annual temperature in the U.S. between 1900 and 2022 can be explained by the emissions only model. The year variable was far from being significant once again with a p-value of 0.5956 but the nitrous oxide variable was much closer to being significant.

The model diagnostics for all models are relatively similar despite having different predictors. The residuals vs fitted values plot shows no signs of nonlinearity nor heteroscedasticity for all models. The Q-Q plots for all models show slight deviations from normality in the tails but given the relatively small dataset size of 123 observations this is expected. Observation 35 was noted to be a potential point to be removed due to moderate influence however based on domain knowledge we decided to keep it in the dataset as it was characteristic of the wild weather patterns of the 1930s which saw devastingly cold winters and the infamous Dust Bowl which wreaked havoc across the Midwest.

Regression with autocorrelated errors (if needed) Since the residuals of all models showed no signs of autocorrelation based on the acf plot and the Ljung-Box test p-values were all higher than alpha =0.05 , there was no need to fit regression models with autocorrelated errors.

final model selection The lowest AIC value was observed in the emissions only model with an AIC = 411.5568. The lowest BIC value belongs to the social model with a BIC of 427.6950. Lastly, the highest adjusted R^2 value belongs to the emissions only model as mentioned previously with a value of .3314. A complete breakdown of the AIC, BIC and adjusted R-squared values for all models is shown in Table 2 below.

Table 2: Model Comparison Metrics | Model | AIC | BIC | Adjusted R-squared | |———————|———|———|————————| | Full Model | 414.2226 | 436.7201 | 0.3272 | | Social Factors Only | 413.6341 | 427.6950 | 0.3147 | | Emissions Only | 411.5568 | 428.4299 | 0.3314 |

We know that lower AIC and BIC values indicate a better fitting model while higher adjusted R-squared values indicate a better fitting model. Furthermore, BIC tends to prefer smaller models. Based on the model comparison metrics, the emissions only model will be selected as the final model for forecasting purposes since it has both the lowest AIC value and the highest adjusted R-squared value while having a BIC value very close to the lowest one observed in the social model

####Forecast future 5 values

The best model found was the emissions only model which included year, co2, methane, and nitrous oxide as predictors. Using this model we forecasted the average annual temperature in the US for the next 5 years (2023-2027) based on a growth rate of 1 percent per year for each emissions predictor. The forecasted values are shown in Table 3 below.

Table 3: Forecasted Average Annual US Temperature (2023–2027) | Year | Forecasted Temperature (F) | Lower 95% CI | Upper 95% CI | |——|———————-|————|————| | 2023 | 53.48625 | 50.89185 | 56.08065 | | 2024 | 53.51607 | 50.92124 | 56.11089 | | 2025 | 53.54634 | 50.95109 | 56.14158 | | 2026 | 53.57706 | 50.98139 | 56.17273 | | 2027 | 53.60825 | 51.01214 | 56.20435 |

Since the official temperature data for 2023 and 2024 has been released by NCEI, we can compare our forecasted values to the actual observed values. In 2023 the actual average annual temperature in the US was 54.4 F while our forecasted value was 53.48625 F. In 2024 the actual average annual temperature in the US was 55.5 F while our forecasted value was 53.51607 F. Thus our model underestimated the average annual temperature in both years by approximately 0.9 F in 2023 and 1.98 F in 2024. This discrepancy is not surprising given the simplicity of our model and the fact that we assumed a constant growth rate of 1 percent per year for each emissions predictor. It should be noted however, that the point estimate is oftentimes rarely accurate in predicting future values, especially as we extrapolate further from the observations used to fit the model. In looking at the confidence intervals for our forecasted values, we can see that both actual observed values for 2023 and 2024 fall well within the 95% confidence intervals of our forecasts for the average annual temperature, indicating that while the point estimates were not accurate, the model was still able to forecast future values moderately well.

Discussion & Conclusion

Overall, the best model proposed by this study explains only 33.14% of the variance in average annual temperature in the US between 1900 and 2022. The result is indicative of the complexity of the issue of climate change as a whole, because there are numerous factors that contribute to it beyond just the three main greenhouse gas emissions explored by the study within the emissions model. Given that the year variable was not significant in the final model, the findings suggest that the increases in average annual temperature in the US over time are not driven by time components but rather by external socioeconomic factors among which are greenhouse gas emissions. This finding aligns with the stationary of the temperature time series observed in the preliminary analysis based on the adf test results. The two most significant predictors based on the lowest p-values from the t-test in the emissions model were somewhat unsurprisingly CO2 and methane emissions as they are widely regarded to be among the most common greenhouse gasses contributing to climate change.

Another potential major shortcoming of this model is its simplicity as we include a relatively small number of predictors (four) with a total of 123 observations from just a single country. The climate and socioeconomic

patterns observed within the United States are not entirely representative of global trends so the model may struggle to generalize to other countries or continents. The decision to keep the models small was mainly done for the purpose of preserving interpretability while also keeping the task manageable from a time perspective.

Future studies may explore various avenues of improving upon the results achieved by this exploration. As mentioned previously, expanding the dataset to include more countries and potentially more variables such as deforestation rates, industrial activity levels, energy production and consumption could all assist in making the model more robust and capture more of the variance in the average annual temperature. Making the model more robus may also increase the accuracy of its forecasting although more complex models are more likely to overfit so a balance should be struck. In addition, a different target variable may be used to model different aspects of climate change such as sea level rise. Furthermore, a potential area of improvement to the current modeling approach that is relatively easy to implement but was beyond the scope of the study is the use of model selection techniques such as forward, backward, or stepwise selection to remove statistically insignificant predictors from the results.

The study shows that the primary factors associated with a change in the average annual temperature in the U.S. between 1900 and 2022 were CO2 and methane emissions. Another factor that showed some association with the average annual temperature was GDP in the social factors only model which makes sense as GDP is a measure of economic activity and industry which produce greenhouse gas emissions. The emissions model fit with the CO2 and methane emissions predictors offered moderate forecasting power for the average annual temperature over the next five years with confidence intervals that contained the actual observed values for 2023 and 2024. The techniques implemented in this study highlight the complexity of climate change as an issue but highlight the potential to model and forecast its effects using statistical methods to provide lawmakers with actionable insights that could be used to alleviate its impact on humanity in the coming century.

References

Dangi, S. (2023). CO2 emissions across countries, regions, and sectors [Data set]. Kaggle. https://www.kaggle.com/datasets/shreyanshdangi/co-emissions-across-countries-regions-and-sectors

National Centers for Environmental Information. (2023). National climate report: December 2023. NOAA. https://www.ncei.noaa.gov/news/national-climate-202312

National Centers for Environmental Information. (2024). National climate report: December 2024. NOAA. https://www.ncei.noaa.gov/news/national-climate-202413

Nguyen, G. B. (2023). Average temperature from 1900 to 2023 [Data set]. Kaggle. https://www.kaggle.com/datasets/giabchnguyn/average-temperature-from-1900-to-2023

Ritchie, H., Roser, M., & Rosado, P. (2024). CO2 and greenhouse gas emissions. Our World in Data. https://ourworldindata.org/co2-emissions

United Nations. (n.d.). Climate change. United Nations. https://www.un.org/en/global-issues/climate-change

https://www.un.org/en/global-issues/climate-change

Original Emissions Data source:

https://ourworldindata.org/co2-emissions

Compiled Emissions Data:

https://www.kaggle.com/datasets/shreyanshdangi/co-emissions-across-countries-regions-and-sectors (primary data)

Compiled Temperature Data:

https://www.kaggle.com/datasets/giabchnguyn/average-temperature-from-1900-to-2023 (supplementary temperature data)

2023 us temp: https://www.ncei.noaa.gov/news/national-climate-202312

2024 us temp: https://www.ncei.noaa.gov/news/national-climate-202413

Appendix

Data dictionary

Stuff: