Patterns and Projections: Exploring Predictors of Greenhouse Gas Emissions

Group 6:
Luke Hazelton, Trang Pham, Sierra Iverson, Davis Salvador, Riley Spas, Matthew Ardon
CSCI 385
Dec 13, 2023

Abstract

Over the last few decades, environmental scientists and climate experts have concluded that the planet's continual emission of greenhouse gasses (GHG) would lead to irreversible climate change if not addressed properly. This study focuses on examining recognized drivers of GHG emissions and predicting GHG emissions using a moving average for the time-series data we collected. Correlation of individual variables to GHG has proven to be significantly difficult, but predicting GHG emissions based on the collection of a subset of drivers worked with relative accuracy.

Introduction

The pressing issue of Greenhouse Gas (GHG) emissions, encompassing CO2, methane, nitrous oxide, and others, stands as a critical catalyst in climate change [1]. Global governments increasingly prioritize emissions reduction and accurate tracking as a cornerstone of climate mitigation efforts. Numerous industries and processes contribute to these emissions, necessitating an understanding of the most significant drivers. To address this, our focus revolves around meticulously tracking and analyzing a specific subset of GHG emission drivers in the US. This endeavor aims not only to comprehend their correlation with emissions but also to predict future GHG emissions accurately. Through meticulously exploring datasets encompassing the US Cattle Inventory, Population Data, Median Household Income, GDP, and Gross GHG Emissions, we undertake a comprehensive time series Exploratory Data Analysis (EDA), followed by developing and evaluating predictive models. Beyond data analysis, this endeavor offers insights crucial for informed policy-making and sustainable practices, particularly within sectors like agriculture, population dynamics, economic indicators, and emissions. Ultimately, this pursuit strives to pave the way for proactive measures towards a more sustainable and environmentally conscious future. Our goal is to track this specific subset of drivers of GHG

emissions, analyze their relationship to GHG emissions, and accurately predict GHG emissions in the US.

Related Work

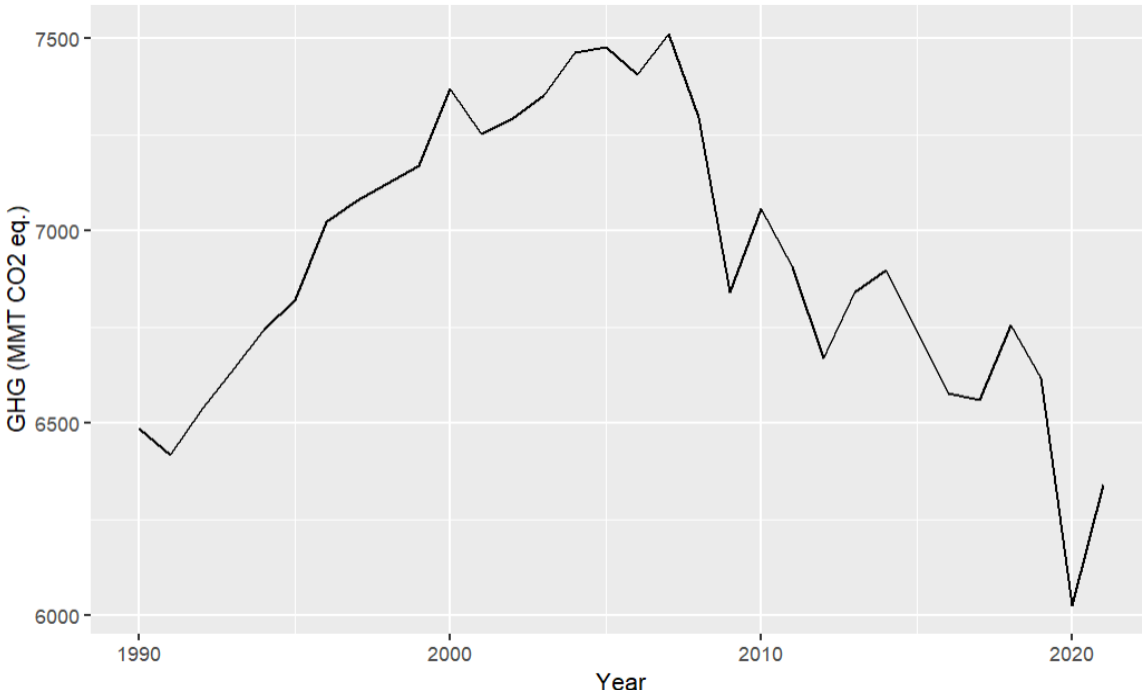
Identifying the drivers our team wanted to focus on required a review of the literature of various studies related to environmental science. After examining the impact of the agricultural sector on GHG emissions and recognizing that livestock alone accounts for four-fifths of GHG emissions associated with agriculture [2], our team decided to include that as one of our drivers. Other studies our team examined highlighted the significance of the affluence or wealth of a population (as well as its size) on the impact that population has on the environment [3], which is why we included data about population, US median household income, and US GDP to track our target population.

Methods

We couldn't find any data sets available that already included all of these variables together. So, next we had to find data on these variables somewhat individually and make sure that they cohesively fit together in a data set. After finding the required data, we then pivoted and trimmed each of the data sets until we were left with what variables we needed and joined them all together. After this, we did more data cleaning on our new tibble. We got rid of N/As and formatted data types to be the correct types, as these are important when making models and figures later on. From here we examined the data and came up with some simple figures to get a better grip on the data. From here we followed a straightforward path of building models using specific/a combination of variables to predict GHG emissions. This followed a cycle of testing, revision, and building. Eventually, we ran into a wall where we couldn't improve our model anymore, from our limited amount of data. Then with some help from Professor Tillquist, we got some good direction on a better more "data efficient" model to use.

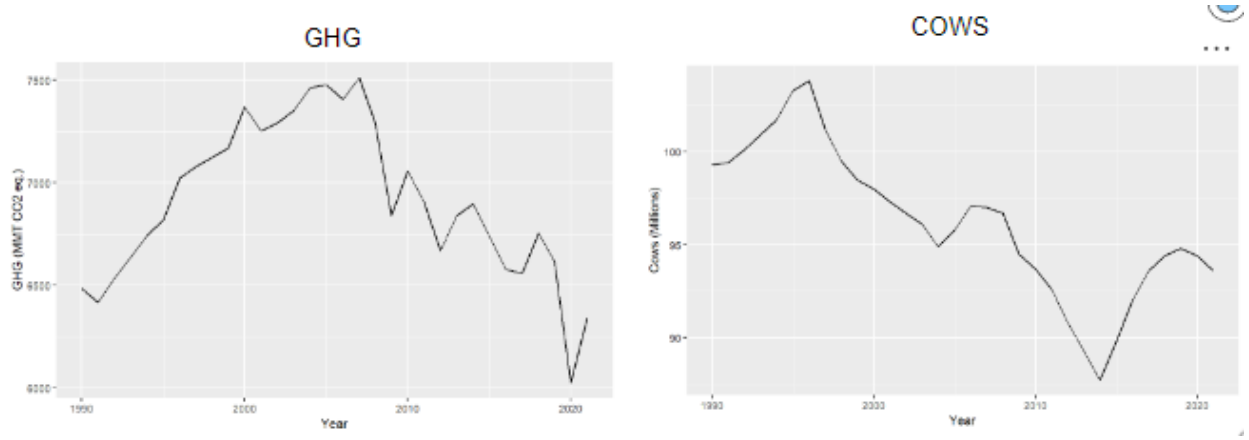
Results/Discussion

As we began our EDA process we looked at time series graphs for all our variables against the year. To our surprise, total GHG has been on the decline since around 2008.

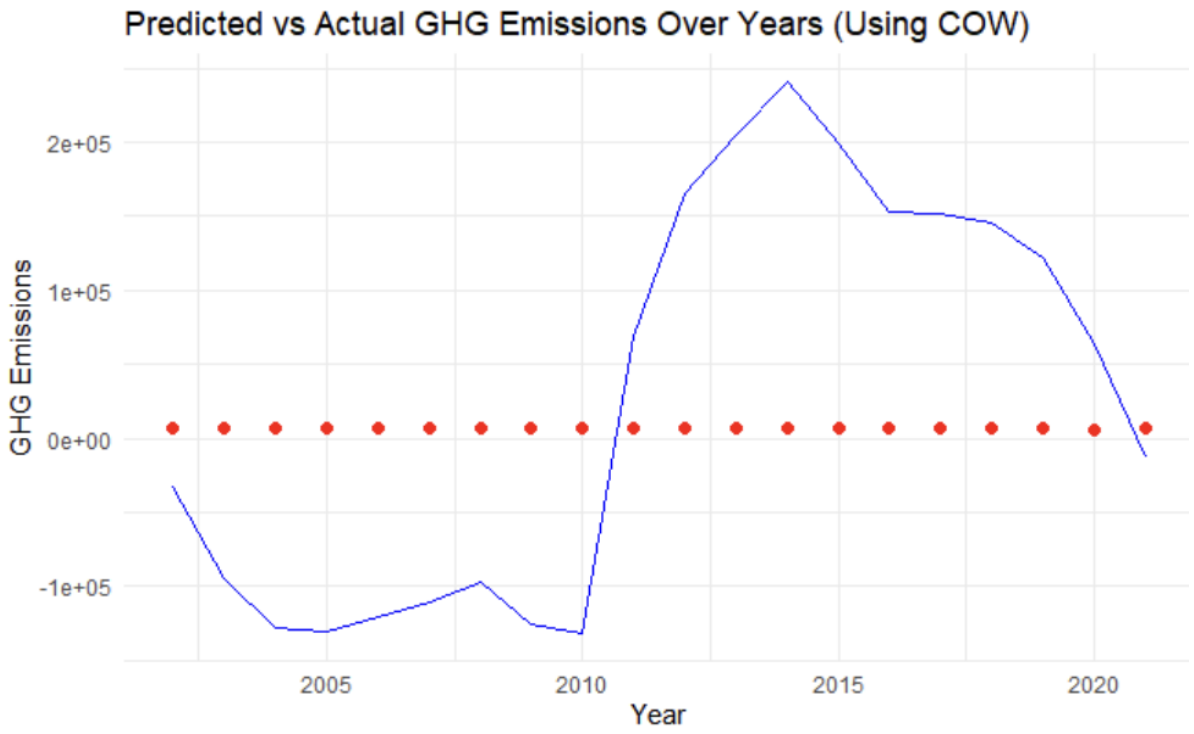


Originally we thought GHGs would have more of a linear increase throughout the years, so seeing that it's been at its lowest point since 1990 was a happy surprise.

Of note were the curves for GHG and cow inventory because they had rises and falls that seemed to coincide with each other.

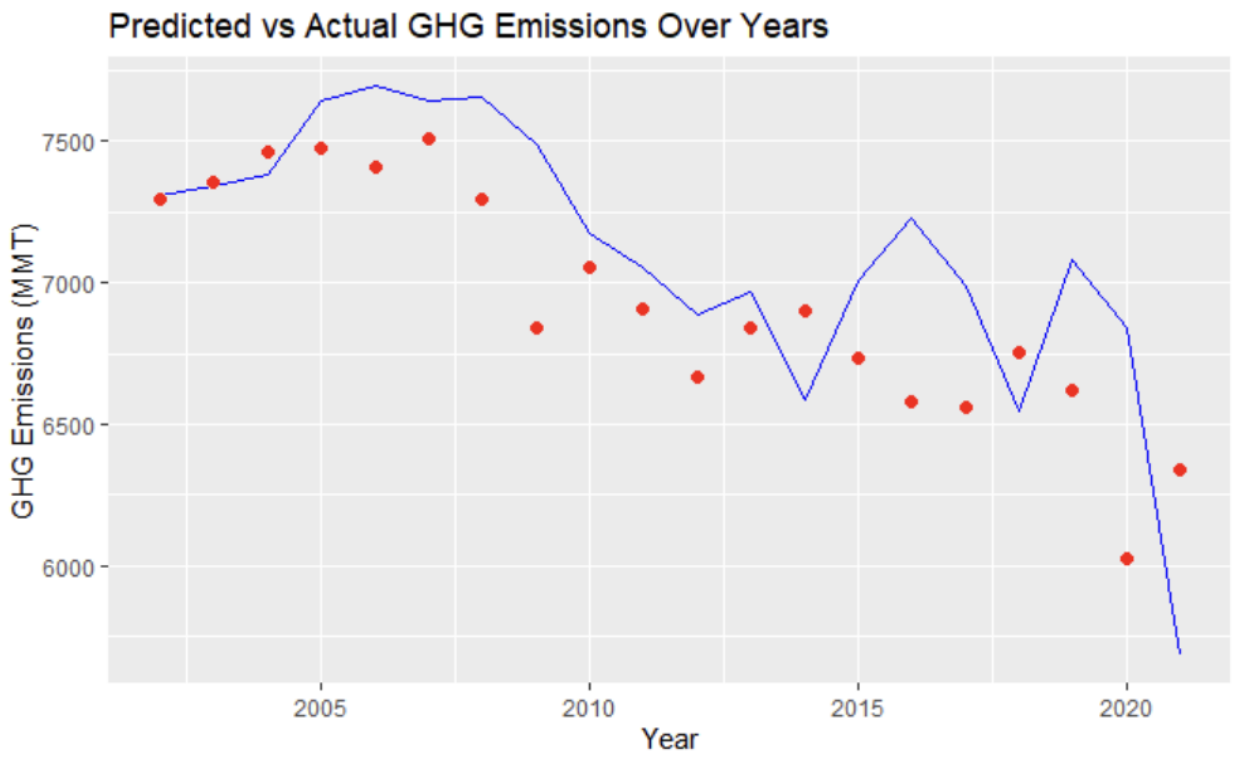


This led us to believe that perhaps they were correlated since cows are known to produce most GHGs associated with the agricultural industry [2]. However, when we created a moving linear model predicting GHG by cow inventory the R-squared associated with it was -111,461. This was due to an error when training the model which resulted in the true values of GHG for the specified years being equal to 0.



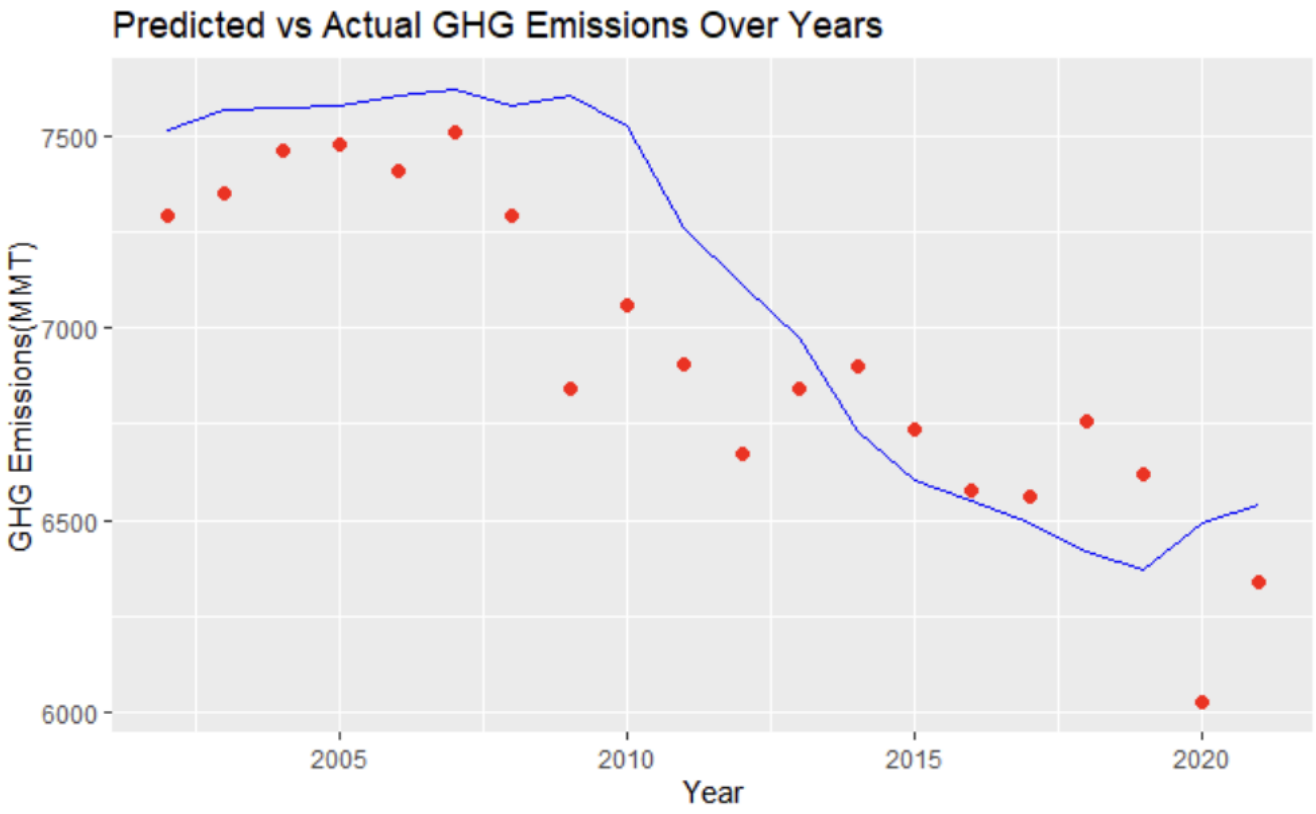
No doubt, this was why the R-squared returned by the model was so bad. Unfortunately, due to time constraints, we were not able to review the model and adjust it further. Our failure here does not, however, indicate that cow inventory is a poor predictor of GHG so it would be worth exploring the relationship between the two again in the future.

While the aforementioned model yielded poor results, we began to find success with our next model. We decided to throw in all the variables we had–year, median household income, population, cow inventory, and GDP–into a linear model and try to predict GHGs.



We can see that the model never predicts the actual GHG emissions except for the first two years, but the predicted line seems to follow the actual values to a moderate degree. However, the R-squared for this model was approximately 0.1198. Given this low R-squared, we should be weary about relying on this model to predict future GHG emissions, though we can visually see that our predictions weren't ever too far off from what was actually.

To round out our analysis, we decided to use GHG emissions to predict itself. The resulting model yielded us an R-squared of 0.4292, the highest out of all our other models.



Visually, it does appear that our predictions more closely coincide with the actual values of GHG emissions, with the model having more of a balance between under and over-predicting when compared to the previous model. However, with such a low R-squared value it is difficult to recommend this model as an accurate predictor of GHG emissions. Furthermore, this model has a lag effect, where once the previous few years start heading in a particular direction we can see our prediction line head in that direction. This leads us to believe the model has a similar outcome as a rolling average forecast.

Limitations

It should be noted that while our project scope remained consistent throughout the entire lifecycle of this project, we had to pivot towards the end by working on a completely new data set. Previously, we were analyzing data on GHG emissions for specific facilities located all over the U.S., but we realized that the predictions generated using said data would only be at the facility level. Even if we were to aggregate the results, the emissions only accounted for around half of all GHG emissions in the U.S., so we would be missing the full picture. This, paired with some of the feedback that we received during our in-class EDA presentation about how some of the visualizations we were making were hard to read and understand, influenced our decision to change directions, even if it was during the final days we had to work on this assignment.

Ultimately, we had far less time to analyze the data set we ended up using for our project, though the interpretations of our analysis were much clearer compared to the analysis of the old data set.

Future Work

Given what we found during this project, there are a few things that would be interesting to take a look at or include in future work. One thing is looking into more factors that impact GHG emissions such as technology - especially those that aid in reducing emissions [4], natural factors/disasters, fossil fuels, and transportation. It might be difficult to quantify some of these factors but if possible, it would make the project more insightful and comprehensive. The more emission sources examined, the more accurate the results as to what factors have the greatest impact on GHG emissions. For instance, adding data for transportation into the mix could produce some interesting results considering how much emphasis is typically placed on the ramifications of car pollution. Factors such as technology would make for some interesting and possibly newly discovered results. It would also be beneficial to do this project with proprietary data to possibly lead to more accurate and clean results because the quality of data does matter. This was not done in this project just due to availability, cost restraints, and the scope of the project. In general, it would also be beneficial to focus more on producing accurate models to get the best results. Part of getting more accurate models could involve creating a baseline model by creating a moving average forecast of GHG emissions.

Conclusion

We found that a pure linear regression model will not work, but a moving average is somewhat good at predicting GHG emissions. Also, correlating individual variables to GHG emissions is near impossible, due to the scope of the topic. However, predicting GHG emissions based on the collection of variables ultimately worked with relative accuracy.

Contributions

Luke: coded a lot, and helped create presentations. Davis: found articles to use for citations, formatted references in GSA format, and focused on providing context in intro sections of the report and presentation. Riley: assisted with data cleaning/transformation, and helped create presentation material. Sierra: helped with finding data and research, helped with presentations, and formulated data science questions. Trang: help with collecting data and preprocessing.

Matthew: Found the data sets we used in our final analysis, created time series visualizations and assisted with formatting the presentation.

References

[1]: Cassia, R. et al., 2018, Climate Change and the Impact of Greenhouse Gasses: CO2 and NO, Friends and Foes of Plant Oxidative Stress: Frontiers In Plant Science, v. 9, doi: 10.3389/fpls.2018.00273

[2]: Friel, S. et al., 2009, Public health benefits of strategies to reduce greenhouse-gas emissions: food and agriculture: The Lancet, v. 374, p. 1953 - 1955.

[3]: Bruckner, B. et al., 2022, Impacts of poverty alleviation on national and global carbon emissions: Nature Sustainability, v. 5, p. 311 - 320.

[4]: Hunter, P., 2016, The potential of molecular biology and biotechnology for dealing with global warming: EMBO Reports, v. 7, p. 946 - 948