Graduate Project Report

Topic 2: Climate and the Environment

Greenhouse Gas Emissions Prediction of Stationary Emitters in the US

Jiayin Guo, Jazmyn Huang

Dec 2021

Abstract

Climate and environmental data provide a pool of valuable perspectives for Environmental Policymaking. In this project, we reflected on historical Greenhouse Gas (GHGs) emissions data from direct emitters with respect to three features: company, sector, and state. We applied Feature Engineering and used three Linear Regression models to predict Greenhouse Gas (GHGs) emissions in a given facility in a given year, and add them to calculate the total CO2e emission nationwide. As a result, model3 that properly OHE on industrial identity performs the best among the three models. Additional societal data relating to the fast-pacing GHGs policies are needed to improve total CO2e emissions prediction to combat climate change.

1. Introduction

1.1 Background and Motivation

The 2021 United Nations Climate Change Conference, more commonly referred to as the COP26, was held in Glasgow, Scotland from October 31 to November 13 this year. The Glasgow Climate Pact negotiated through consensus of the representatives of the 197 attending parties was the first climate deal to explicitly commit to reducing the use of coal and included wording that encouraged more urgent Greenhouse Gas (GHG) emissions cuts. As one of the world's largest carbon emitters and the world's largest carbon emissions per capita, the United States has been committed to further emission reductions and transformation to the use of clean energy at various climate conferences in recent years. But have these promises been fulfilled?

In this report, we are interested in the total carbon dioxide equivalent (CO2e) emissions in the U.S. in recent 10 years. The data we used is from the Greenhouse Gas Reporting Program (GHGRP) held by EPA, which tracks facility-level emissions across industries in the United States. In the United States, though mobile emitters such as automobiles are the largest source of carbon dioxide emissions, the emissions of stationary emitters like industrial facilities can better reflect the performance of each state's environmental policy, and it also allows us to explore the environmental friendliness of different industries.

1.2 Research Objectives

Our research problem can be stated as follows:

How did the total CO_2 emissions of the recorded facilities in the U.S. change in recent years and how will its trend be in the near future? To be more specific, what can CO2e emissions profiles with respect to **Company**, **Sub-sector**, and **State** can tell us about the past and the future?

To address this problem, we first explored the relationships between a facility's total carbon dioxide emissions and its parent company, sub-sector info, and the state it locates. Next, we use a linear regression model to predict the CO_2 emissions of a given facility in a given year, and the sum of the emissions from each facility would be the total emissions across the states.

1.3 Innovations and Distinctions from Other Research

A large number of findings and research can be found on the EPA's official website (https://www.epa.gov/ghgreporting). Most of them focus on the facilities' performance every single year, especially the most recent year. In our report, we concentrate on the fluctuations in carbon dioxide emissions over time (years) and predict the total emission using the time, location, and industry information by machine learning tools, through which we try to reach the conclusion whether or not the U.S. has fulfilled its promises to reduce carbon emissions.

2. Data Description

2.1 The gas type data

The "gas type" dataset is provided on the EPA website. It is a summary table of gas information of emitters, which contains over 200,000 observations for a total of 16 features, with each row representing a record of one emitter. Among its columns, there is vast information about the type of emission gases, location of emitters, year, etc. The most important column is called V_GHG_EMITTER_GAS.CO2E_EMISSIONS, which shows the emissions of greenhouse gases in carbon dioxide equivalent (CO2e) value in unit of ton. **For future modeling purposes, we are most interested in total CO2e emission, emitter's state, emissions year, and gas type.**

Table 1: Key Columns Profile of "Gas Type" Dataset

V_GHG_EMITTER_GAS.GAS_NAME	Biogenic CO2, Methane, Nitrous Oxide, Carbon Dioxide, Sulfur Hexafluoride, Other Fully Fluorinated GHGs, PFCs, HFCs, Very Short-lived Compounds, HFEs, Nitrogen Triflouride, Other
V_GHG_EMITTER_GAS.STATE	WI, NY, LA, MS, KY, GA, UT, AL, IL, HI, AZ, PA, TX, MI, CO,

	CA, FL, ID, MA, AR, OH, NC, WY, CT, OK, DE, SC, IN, MT, VA, WA, IA, NV, MO, MD, MN, OR, NJ, KS, NE, AK, NM, WV, TN, VT, ND, SD, ME, PR, RI, DC, NH, VI, GU
V_GHG_EMITTER_GAS.YEAR	2010 - 2019
V_GHG_EMITTER_GAS.CO2E_EMISSION	Measured in unit of ton

2.2 The gas facility data

Similarly, the "gas facility" dataset contains over 77,106 observations for 21 features. Each row represents a facility record. Compared to the gas type dataset, this table includes more details on the facility itself. For example, it indicates whether the facility is EPA verified, its parent company, and it is noticeable that the NAICS code column contains useful information about the industrial property of the facility. Intuitively, GHGs emissions are often associated with the industry type. Besides, the facility id serves as a useful joining key between these first two tables. **For modeling purposes, we are mainly interested in the NAICS code and numerical facility features.**

Table 2: Key Columns Profile of "Gas Facility" Dataset

V_GHG_EMITTER_FACILITIES.FACILITY_ID	Unique identifiers for emitter facilities
V_GHG_EMITTER_FACILITIES.PRIMARY_NAICS_CODE	The primary NAICS code reported by the facility to represent the activity they are engaged in that is the principal source of revenue
V_GHG_EMITTER_FACILITIES.PARENT_COMPANY	Parent company names

2.3 The NAICS data

Following the NAICS code column in the "gas facility dataset", we downloaded the additional NAICS code table (updated 2017) from the United States Census Bureau. NAICS Code is a classification within the North American Industry Classification System, which is widely used for the collection, analysis, and publication of statistical data related to the US Economy. NAICS uses a two-through-six-digit hierarchical structure, with the first 2 digits indicating the economic sector, followed by the third digit designating the subsector, the fourth for the industry group, the fifth and sixth digits for NAICS industry and national industry respectively. Our NAICS table contains 2,196 rows with each row representing a distinct NAICS code and its corresponding industry. This code bridges the emission to the national economy, which may provide insightful results for policymakers. **For modeling purposes, we are most interested in the facility's sub-sector level information, which is not too general or too detailed for classification.**

3. Data Cleaning

At this stage, the imported tables remain redundant and untidy. We hope to remove unnecessary features, format the content, handle outliers (sensitive for future modeling), and merge tables. Therefore, the main steps in our data cleaning are as follows:

(1) Imputation

Records with NaNs are invalid for modeling. Our measure for NaNs handling is to drop these records.

(2) Letters uppercasing

Letter cases vary due to human input habits. For uniformity, we uppercase strings.

(3) Columns remaming

This step is conducted for clarity and tidiness.

(4) Datasets merging

The final goal for data cleaning is to include all useful features in one table.

(5) Log transformation

This step is applied to handle the long-tailed property of features without dropping useful information in outliers.

4. Exploratory Data Analysis (EDA)

4.1 Company-wise CO2e emissions

We first start exploring the data with respect to the column of the parent company. Considering the structure of the national energy market, it is a common phenomenon that some conglomerate companies make the main contributors to CO2e emissions. In other words, the energy industry always tends to be an oligopoly due to its nature. As a result, we can see from Figure 4.1(a) that the top 20 companies having the most CO2e emissions bar plot that Vistra Energy Corp is the biggest emitter company nationwide, followed by American Electric Power Co Inc. It is noticeable that energy companies lead the GHG emissions.

4.2 Sector-wise CO2e emissions

Another key column feature is the NAICS code which, as a thread, gives us an understanding of how each sector produces CO2e emissions. Similarly, we grouped the table by its sub-sector. From the result in Figure 4.1(b), it is not surprising to see that the top 5 industries turn out to be Utilities, Oil and Gas Extraction, Petroleum and Coal Production Manufacturing, Chemical Manufacturing, and Paper Manufacturing.

4.3 State-wise CO2e emissions

The geographical distribution of CO2e emissions may present useful information as well. The bar plot shows a one-dimensional comparison of state-level total CO2e emissions. It can be noticed that Texas is the most polluted state and its emission doubled that of Pennsylvania which ranks at second place. We utilized *Census Map* as well as *GeoPandas* to visualize CO2e emissions in Figure 4.2, from which we noticed that the western and southern US states generate more CO2e emissions compared to other regions. Specifically, these states are Texas, Pennsylvania, and Ohio.

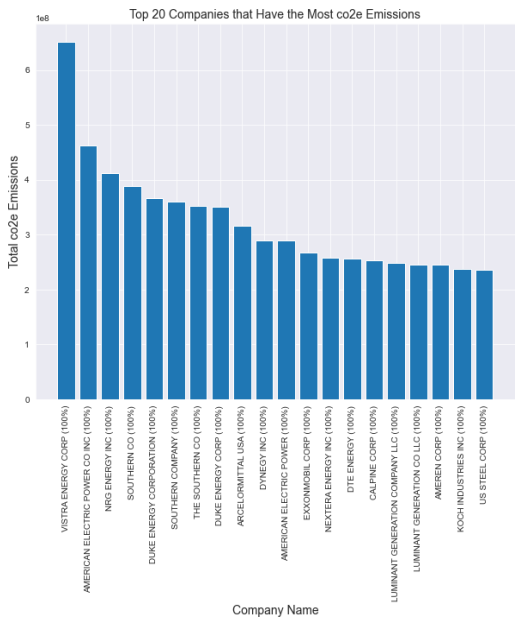

Figure 4.1 (a) Top 20 Companies of CO2e Emissions

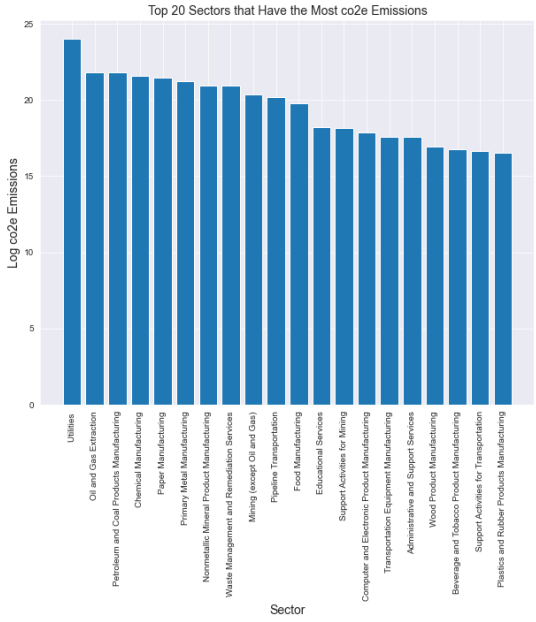

Figure 4.1 (b) Top 20 Sectors of CO2e Emissions

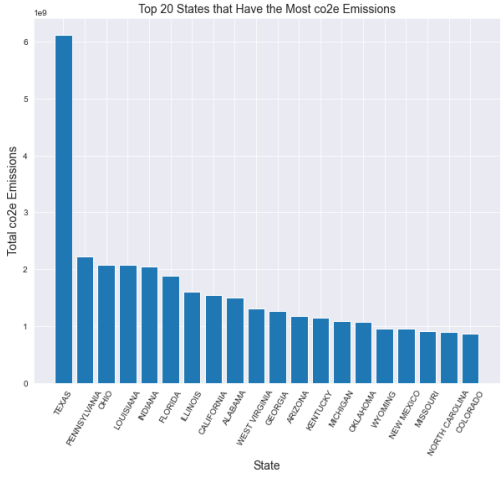

Figure 4.1 (c) Top 20 States of CO2e Emissions

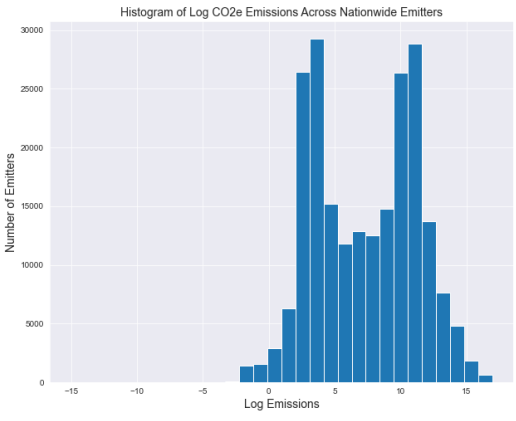

Figure 4.1 (d) Histogram of Log CO2e Emissions for Emitters

4.4 The Bimodal Property

We also applied log transformation on the CO2e column in Figure 4.1(d). It can be seen that the CO2e distribution across emitters is bimodal, which indicates that emitters can be roughly clustered into two major groups. Some specific companies lead the emissions while the other emitters remain at a relatively low emission level.

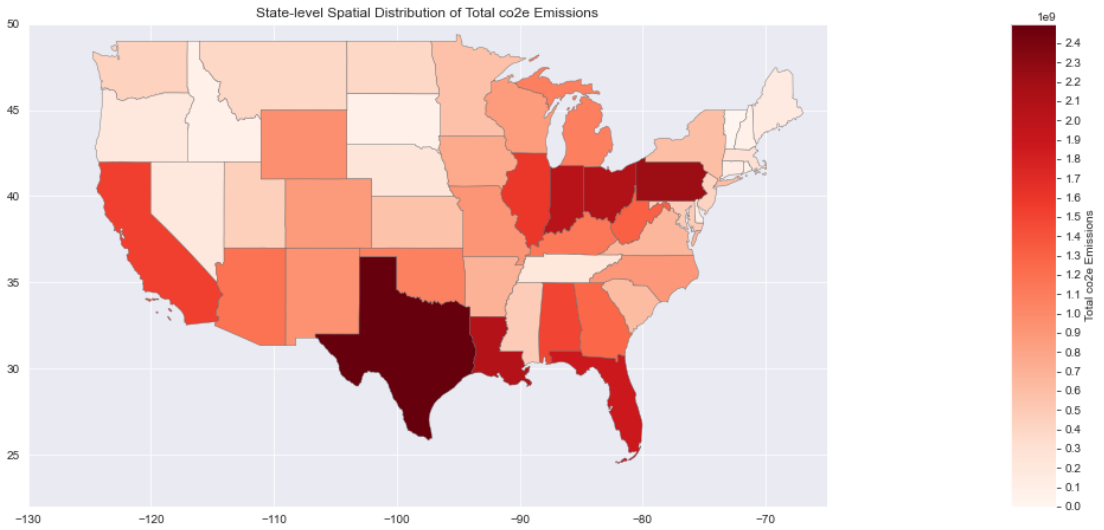

Figure 4.2 Geographical Profile of State-wise CO2e Emissions (Continental US)

5. Modeling and Inference Techniques

From the EDA section, it's clear that the CO2e is largely affected by the home state of the facility and the industry (sub-sector) of the facility. The gas type the facility emits may also exert some influence. In the modeling part, we use the Linear Regression model to:

(1) predict the CO2e per emitter given in the historical table.

(2) sum CO2e across emitters to get the total emission for each year.

5.1 Training-testing Split

Split our data into training and testing sets, with the testing set having a 20% size.

5.2 Three Linear Regression Models

We use the linear regression model to approach this prediction. Since all the features other than *gas-year* are in text type, the most important job in the feature engineering part is to transform the text features into the numeric type. We considered two possible approaches:

(1) Directly One-Hot-Encoding the Feature.

OHE can be applied to gas_type, state, and industry. An example is shown in Figure 5.1.

```
# feature engineering 1: OHE gas type
def OHE_gas_type(df):
    from sklearn.preprocessing import OneHotEncoder
    oh_enc = OneHotEncoder()
    oh_enc.fit(df[['gas_code']])
    dummies = pd.DataFrame(oh_enc.transform(df[['gas_code']]).todense(),
                           columns=oh_enc.get_feature_names(),
                           index = df.index)
    return df.join(dummies)
```

Figure 5.1 OHE code

(2) Introducing a New Metric: Contribution Score

We know from EDA that GHGs emissions vary by state and sub-sector. To measure the impact of the location and industry of the facility, other than OHE, we introduce a new metric for each emitter: the Contribution Score. The score represents the percentage of the emissions of the facility's state (sub-sector) to the total emissions of all the states (sub-sectors). An example of state contribution score calculation is shown in Figure 5.2.

```
# feature engineering 2: add contribution score
def add_contribution_score(df):
    # state contribution
    state_contribution = df.groupby("gas_state").sum()[["co2e_emission"]]
    state_contribution["total_emission"] = sum(state_contribution["co2e_emission"])
    state_contribution["state_contribution_score"] = ((state_contribution["co2e_emission"]/state_contribution["total_emission"]*100)).round(2)
    state_contribution = state_contribution.reset_index()

    with_state = df.merge(state_contribution[["gas_state", "state_contribution_score"]], on="gas_state", how="outer")
```

Figure 5.2 Adding state contribution score to data

By applying these two techniques we can convert the three text type features into numeric ones, then we build our models choosing different features.

Although state and sub-sector information will affect the emission, without doubt, it's not clear whether or not gas-type really has an impact, so we build models with and without the *gas_code* to compare the performance. We also want to know which of the OHE and contribution scores is the more efficient method to transform text values, so we also built models to compare.

A total of three models with different features were created:

(1) **Model1**: use four features including *gas_year*, *gas_code* (gas type), *gas_state,* and *naics_title* (sub-sector), with the *gas_code* being One-Hot-Encoding and the state and sub-sector information recorded as "contribution score".

(2) **Model2**: use three features including *gas_year*, *gas_code*, *gas_state,* and *naics_title* (sub-sector) with the state and sub-sector information recorded as "contribution score".

(3) **Model3**: use three features including *gas_year*, *gas_code*, *gas_state,* and *naics_title* (sub-sector) with the state and sub-sector information being One-Hot-Encoding.

6. Analysis of Results

To evaluate the performance of the models, we calculated the MSE and RMSE of both the training set and testing set of each model, and drew the true CO2e emission (Y) v.s predicted CO2e emission (Y_hat) plot. To see how well the model could capture the total emission every year and thus predict the future emission, the plot of the total CO_2 emissions as a function of the year was also shown.

6.1 Model Evaluation

6.1.1 Y v.s. Y_hat plot

First, we looked at the Y v.s Y_hat plot of the training set of each model. The size of the data is too large, to avoid overplotting, we sampled 1% points out of all the data points in the training set. The three plots are shown in Figures 6.1 (a) (b) (c).

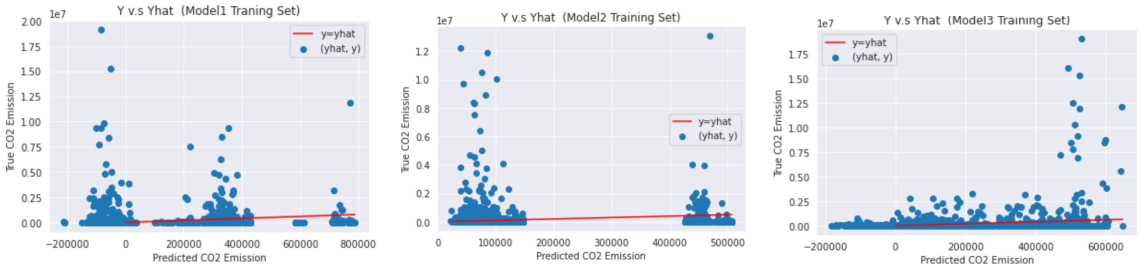

Figure 6.1 (a) (b) (c) Y v.s. Y_hat plot of 1% the sampled training sets, model1, model2, and model3

From the plots, it's clear that **the third model did the best capturing the Y ground truth value**. While the first and the second model did predictions more discrete, the predictions are more smoothly distributed in the third model.

6.1.2 RMSE of the training and testing sets

The RMSE values of the three pairs of training and testing sets are shown in Figure 6.2.

		RMSE	RMSE in le7	Difference Rate
Model1	Training Set	880098	0.088	
	Testing Set	**903774**	**0.090**	**2.7%**
Model2	Training Set	856929	0.086	
	Testing Set	**882722**	**0.088**	**3.0%**
Model3	Training Set	822155	0.082	
	Testing Set	**846932**	**0.085**	**4.0%**

Figure 6.2 RMSE of training and testing sets of three models

From model 1 to model 3, the RMSE is decreasing, indicating that the **performance of the model is getting better**. Also notice that there's not a big difference between the RMSE of the training set and that of the testing set, meaning that we are **not overfitting**. In fact with the limited features we have, there's a big chance that we are still at the very left side of the bias-variance trade-off, thus no cross-validation or regularization is needed.

6.2 Model Performance on Prediction of the Total CO2e

The actual/predicted total CO2e as a function of time of the three models are shown in Figure 5.7 to Figure 6.3 (a) (b) (c).

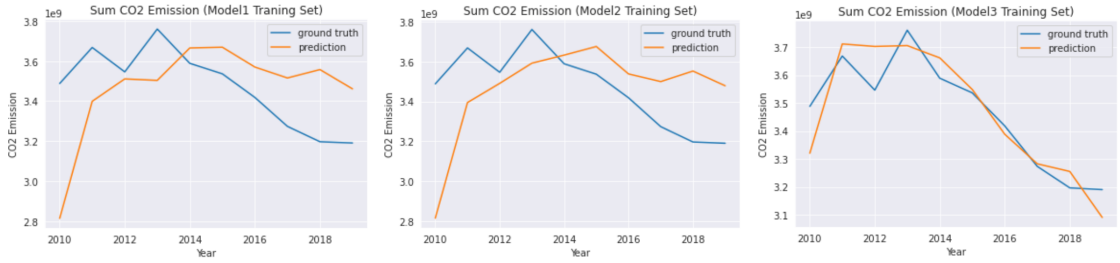

Figure 6.3 (a) (b) (c) Total CO2e as a function of the time, model1, model2, and model3

All three models succeed in showing a peak at around 2014, but clearly, only model 3 has captured the significant drop from 2013. Although none of the models demonstrated a perfect fit with the actual value, model 3 could be most trusted with the task of predicting the total CO2e of the future.

To sum up section 6.1 and section 6.2, the fact model 2 performed better than model 1 indicating that the *gas_code* (gas type) feature is not strongly correlated with the CO2e; while model 3 performed better than model 2 means that when dealing with state and industry information in this project, it's better to use OHE than to add a new, self-defined numeric feature based on the EDA to represent the contribution. The mechanism behind is in need of further investigation.

7. Discussion

7.1 Limitation of The Method

There are two major limitations of the presented model:

(1) The important extreme values cannot be well captured.

When looking close to the Y v.s. Yhat plots, we can see that in all three models, the presence of outsiders is very significant, a result of which is that we did not remove the extreme values during data cleaning. This is because the "outliers" of carbon dioxide emissions are the most important contributors to the greenhouse gas emissions—some energy giants, such as utility facilities and fuel companies, tend to emit hundreds of times more carbon dioxide than other facilities in the light-polluting industry. Since we are using the Linear Regression model, our models tend to be located in the area where most of the values are located, thus cannot predict the extremely high emissions of these polluting monsters well, but removing them will lead to serious bias.

(2) Limited available features.

Another thing to notice is that even our best-performing model (model 3) is not a perfect predictor of the fluctuation of carbon dioxide emissions over time. This is because the emission of all pollutants including carbon dioxide is greatly affected by national policies and national economic levels. For example, the leap in emissions from 2010 to 2011 might be due to the fact that it took the United States almost three years to recover from the 2008 economic crisis; and the dramatic and long-lasting drop started from 2015 is almost certainly a result of the Paris Climate Accords.

While it's possible to import policy and economic related features in our data, it's not easy to include them into our model, especially the policies—digitizing policies has always been a hard task in the data science field. In this respect, our model will be much more effective when being used along with other non-numeric features.

7.2 Potential Societal Impacts and/or Ethical Concerns

Summarizing our model, it predicts the carbon dioxide emissions of a facility based on its state, industry, and emission year. If used properly, it can help the (state and federal) government to **predict the future greenhouse gas emissions and take related corresponding procedures in advance**.

But we should also pay close attention to the **timeliness of this model**—and potentially, all models in the environmental field—as the policy changes quickly when referring to the sensitive topic relating to the environment. For example, when we design this model, the oil and gas industry is one of the second-largest contributors to CO2e. But if the United States introduced policies in 2030 that require 80% of American cars to become electric vehicles, then this model's evaluation of the oil and gas industry will no longer be valid.

7.3 Surprising Discoveries and Future Work

Reflecting on the project, it is surprising that the log CO2e emissions histogram displays a bimodal property. That means the emitters can be roughly clustered into two groups which show significant differences in emissions level. But what feature can best categorize these two groups? Should we associate with dimensions such as state, sector, parent company directly, and to what extent are they correlated? In the meantime, how can we take into consideration of big emitters' economic contributions while downplaying their pollution to the environment? Will it be a divide for the government to implement different emission standards for these two groups that might function differently in society? Can we incorporate additional economic data such as revenue and GDP to support the enactment of a more reasonable policy, e.g. environmental tax?

Therefore, in a wish to curve down the GHGs emissions, it is necessary for policymakers to identify and properly handle the following two relationships:

(1) **Equity**: Emitters who are more responsible for emissions and those who are less.
(2) **Balance**: Tradeoffs between environmental externalities and economic development.

For future work, it is plausible to answer the question: **will it be possible to formalize a carbon trading market between these two major groups of emitters?** It will be an insightful attempt to mobilize available economic data from the Bureau of Economic Analysis and keep pace with international carbon centrality policies, on top of climate and environmental data exploration. Beyond stationary emitters, we should also have an eye on mobile emitters such as automobiles to gain a bigger picture of GHGs emissions cutoff.

References

[1] Garg, A., Bhattacharya, S., Shukla, P. R. et Dadhwal, V. K.. (2001). Regional and sectoral assessment of greenhouse gas emissions in India. *Atmospheric environment*, *35*(15), 2679-2695.

[2] Ghazouani, A., Jebli, M. B. et Shahzad, U.. (2021). Impacts of environmental taxes and technologies on greenhouse gas emissions: contextual evidence from leading emitter European countries. *Environmental science and pollution research*, *28*(18), 22758-22767.

[3] Liang, S., Wang, H., Qu, S., Feng, T., Guan, D., Fang, H. et Xu, M.. (2016). Socioeconomic Drivers of Greenhouse Gas Emissions in the United States. *Environmental science & technology*, *50*(14), 7535-7545.

[4] Heidari, N. et Pearce, J. M.. (2016). A review of greenhouse gas emission liabilities as the value of renewable energy for mitigating lawsuits for climate change-related damages. *Renewable and sustainable energy reviews*, *55*, 899-908.

[5] Fischer, C., Kerr, S. et Toman, M.. (1998). USING EMISSIONS TRADING TO REGULATE U.S. GREENHOUSE GAS EMISSIONS: AN OVERVIEW OF POLICY DESIGN AND IMPLEMENTATION ISSUES. *National tax journal*, *51*(3), 453-464.