CITADEL/CORRELATION ONE

BEYOND OIL: AN ANALYSIS OF CLIMATE IMPACT, OIL SENTIMENT, AND PUBLIC POLICIES

AUTHORS

ISHITA JAIN [1]

CHENGYI TANG [1]

ZARA COOK [2]

ANJIE LIU [3]

Team 5

Spring Invitational Datathon

APRIL 2024

[1] *Harvey Mudd College*

[2] *University of Waterloo*

[3] *University of Texas at Austin*

Contents

1

Executive Summary

It has been known that fossil fuels are detrimental to the environment. Many solutions have been proposed, that target the reduction of the consumption of these fossil fuels. For example, a common idea is to transition to cleaner energy for our needs such as solar energy. However, these solutions haven't had the impact they hoped to because it is a difficult problem to solve. Many industries and economies rely directly on the use of these fossil fuels and it is not as simple as simply reducing consumption. Alternative energies are more expensive, not feasible, unreliable, and often end up using fossil fuels in their generation anyway. Additionally, there simply isn't enough of alternative energy to supply the world's energy needs.

Our team was motivated by alternative solutions that can be found to address fossil-fuel-caused climate change. We hope to propose specific ideas that can help solve this issue by focusing on what has helped climate change in the past.

So, we posed a multi-step question: 1) What causes fuel to be detrimental for the environment and which ones are detrimental? 2) Does Public Opinion affect the consumption or price of that fuel? 3) How does policy in the US impact carbon emissions, and how effective is policy at reducing emissions?

We summarize our findings below:

What causes fuel to be detrimental for the environment and which fuels are detrimental?

1. Oil is the worst fuel for the environment because it has a high impact on Emissions in the US and because consuming it causes anomalies in temperature to rise. Additionally, it takes 28 months for the consumption of oil to affect anomalies in temperature.
2. We also found that emission per consumption also impacts anomalies in temperature. Additionally, it takes 59 months for the decrease of emission per consumption to decrease anomalies in temperature. We hypothesize this change could occur due to innovations in technologies, perhaps spurred by regulations.

Does public opinion affect the consumption or price of oil?

Previous literature has shown that public and corporate sentiments affect oil prices. Here we find that sentiments expressed by U.S. presidents in their public speeches can be used to model and predict oil prices. Specifically:

1. There is a strong positive correlation between the maximum (most extreme) oil-related sentiments in presidential speeches and oil prices.
2. In contrast, climate-related sentiments and oil consumption/demand had weaker and more variable relationships with oil price.
3. We were able to simulate oil price trajectories that captured the major price spikes and crashes observed historically, better than a purely deterministic model.

This suggests that the strongly expressed opinions and priorities of policy makers, as reflected in their public speeches, can significantly impact oil market dynamics and pricing. The sentiments of high-level policy makers appear to be an important factor driving oil price movements.

How effective is policy in the US in impacting carbon emissions?

1. State-level environmental policies cause a reduction in emissions in the US.
2. Aggressive policies like the ones in New York and California resulted in a significant difference between pre-policy and post-policy emissions.
3. However, the amount of change the policies have still depend on the state's economic circumstances.

Taking these insights, we propose the best solution to climate change would be to have more educational about its dangers because that would help change public sentiment on these issues. These educated individuals must then lobby their local state government (because it is easier to enact changes on a smaller scale and it has impact) to tighten restrictions and reduce both emissions per consumption in the long term but also to encourage industries to transition away from the consumption of Oil to cleaner fossil fuels such as Natural Gas. Governments should also encourage growing businesses in this area that build solutions to reduce consumption or decrease the emission-per-consumption ratio of Oil. Finally, they should create policies that are targeting their specific economic demands and dependencies rather than following other states.

2

Technical Exposition

We first discuss our data derivation and cleaning procedure and communicate our decision-making process.

2.1 Data Gathering

We used the following provided datasets:

- The **Emissions** dataset from Energy Data: We used CO_2 emission data to explore what energy sources are correlated with global warming. CO_2 causes global warming due to the greenhouse effect (United Nations, n.d.).
- The **Consumption** dataset from Energy Data: We used energy consumption data to explore both how policymaker sentiment influences consumption of different energy sources and how consumption of oil as well as how other fossil fuels affect temperature and emissions.

We merged two datasets on their date and energy type to create a single data frame with emission and consumption data for each energy source. We decided to drop the missing values since we had a large dataset of 53,000 rows. Interpolating the missing values would undermine the insights, as climate change can lead to data volatility. However, we found that the nulls in the sector column were aggregate data over a particular group. We then realized the fuel types had different energy units. Using a source from the U.S. Energy Information Administration, we converted the units to a common standard of British Thermal Units (BTU) to measure the heat content of the fuels. Through dimension analysis, we identified the conversion constants to transform the original units of Thousand Barrels per Day, Billion Cubic Feet, Thousand Short Tons, and Trillion Btu into the standard Quadrillion BTU. We also multiplied the oil production constant by 30 to get the total barrels per month. To get the correct total fuel consumption and emissions across industries, we had to aggregate the data. We then created a new feature called“Emission per Consumption" to explore the efficiency of fuel use and its relationship with other variables.

To build on the insights we found in Exploratory Data Analysis, we supplemented our paper with datasets explored in other sections.

2.2 Exploratory Data Analysis

We first wanted to understand the data and explore potential questions that could lead to insights.

Consumption and Emission Over Time

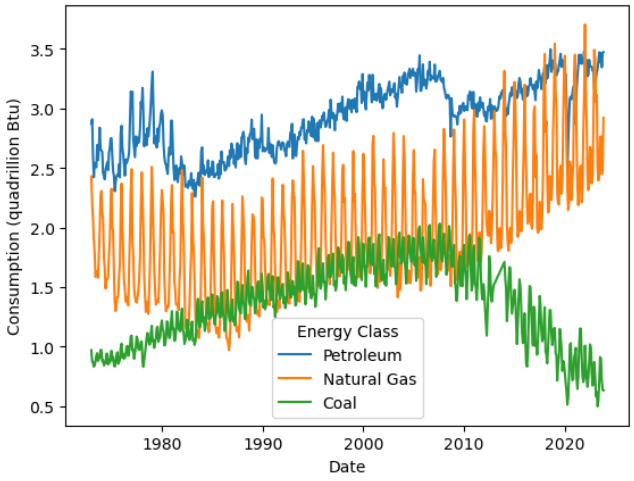


Figure 2.1: *Consumption over time of different energy classes*

Consumption Trends: As seen in Figure 2.1, energy consumption has an overall upward trend, which agrees with intuition, because of growing energy needs over the years. There is high seasonality in the different energy classes, due to the seasonal nature of energy needs. Natural Gas especially is highly seasonal because it is primarily used for generating electricity, used to heat homes. The other energy classes are possibly seasonal due to industrial needs. There are certain peaks in the oil data that can be explained by major oil events. In the late 1970s, the Iran Oil Crisis resulted in low energy consumption. There was also an oil glut in the 1980s that could correlate to the decrease in consumption. The dip in 2008 corresponds to the 2008 financial crisis. Lastly, there was an oil glut in the 2010s that could correspond to the dip in consumption as well. The increase in Natural Gas usage in 2010 may stem from the exploitation of new natural gas reserves, such as shale gas, coupled with the gas's burgeoning role in replacing coal for electricity generation and its heightened utilization in residential and commercial heating systems.

Emission Trends: Figure 2.2 shows the emissions over time of the different energy classes. Similar to consumption, emissions also fluctuate over time but with more volatility. There are also clear visual relationships between emissions and consumption over time. The peaks in consumption correspond to peaks in emissions. This is a potential relationship to explore. In petroleum, the emissions trajectory indicates a slight upward trend until the late 2000s, which then stabilizes and possibly begins to decline. This pattern might reflect improvements in petroleum utilization efficiency or a strategic pivot towards alternative energy sources. The stabilization and potential reduction in petroleum emissions highlight the impact of renewable energy advancements, energy conservation initiatives, and enhanced emissions control technologies.

Correlation and Causality: While the graphs can indicate the relationship between consumption and emissions, it doesn't provide direct evidence of causality. To establish a more concrete link, a more

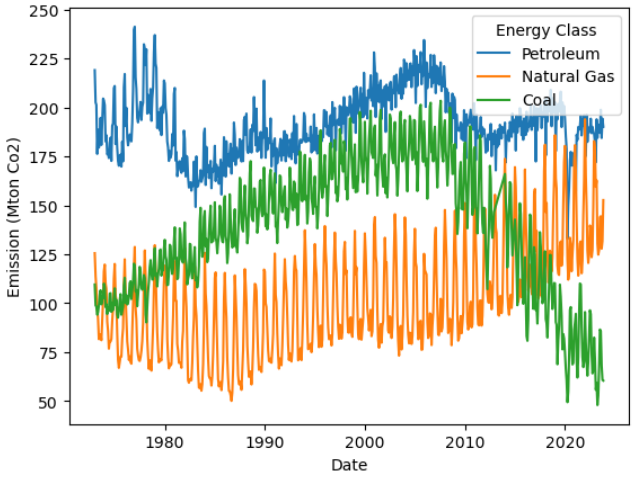


Figure 2.2: *Emission over time of different energy classes*

detailed analysis considering the types of energy consumed and the emissions factors for each type would be needed. One start would be to explore each fuel's emission over consumption ratio.

Emission per Consumption for Different Energy Classes

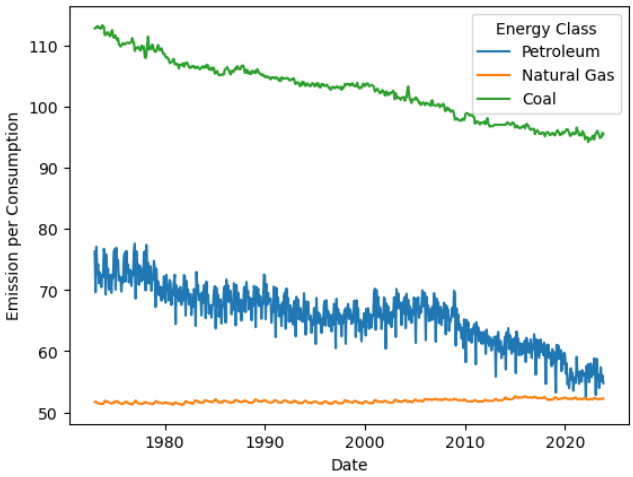


Figure 2.3: *Emission per consumption over time*

Emission per Consumption, as shown in Figure 2.3, in general, shows the volume of emissions per unit of energy consumed. It can be used to understand how the emission per fuel type affects the other factors we will study without considering how much they are consumed.

Trend Analysis: A downward trajectory in the emissions per unit of consumption is evident from the plot, signaling an enhancement in the efficiency of energy utilization across Petroleum and Coal. This trend may also reflect a strategic pivot towards energy sources with lower carbon footprints, thereby contributing to a reduction in the overall emissions intensity of consumed energy. This could be because of the implementation of emissions reduction strategies or the adoption of more stable and efficient energy consumption practices. Such a shift is indicative of progress towards a more sustainable and environmentally friendly energy landscape.

A crucial aspect of our analysis is the comparison of emissions per unit of consumption across energy classes: Petroleum, Natural Gas, and Coal. Petroleum has undergone the most volatility and the most change in recent times since the 1970s, suggesting that there have been many advances in combustion technology, purification of crude oil, and other technologies.

Natural Gas exhibits the lowest emission per consumption values across the three energy classes. This aligns with the growing body of research suggesting natural gas as a 'cleaner' fuel alternative. The work by Burnham et al. (2012) supports this, stating that natural gas combustion generally results in lower emissions of CO2, NOx, and SO2 compared to coal and oil for the same energy output (Burnham et al., 2012).

Coal's emission per consumption is significantly higher than the other energy classes. This is supported by current literature, which verifies that we cleaned the data and interpreted its results correctly (Crawford et al., 2017). It is widely recognized that coal combustion is more carbon-intensive, corroborated by the declining use of coal in favor of cleaner energy sources as seen in recent trends (International Energy Agency (IEA), 2021).

2.3 Which Fuels Matter for Global Warming?

First, we wanted to conduct sanity checks on our data by verifying already supported work in the literature. We also wanted to explore what exactly makes certain fossil fuels bad for the environment and which fossil fuels are bad for the environment.

We used the Temperature from external sources to supplement the provided data. The dataset used in this study was obtained from Berkeley Earth, an independent non-profit organization focused on climate data science (Berkeley Earth, 2016). Berkeley Earth was one of the few sources that published monthly temperature anomaly data from 1960 to 2016. The temperature anomalies, reported in Celsius with a 95% confidence interval, represent the deviation of temperature from a long-term average. While the use of anomalies helps mitigate the effects of geolocation and contextual factors on temperature, it introduces potential biases due to the unknown calculation of the long-term average, which may not reflect the actual expected temperature in a given year.

The temperature data were derived by aggregating measurements from weather monitoring stations, calculating the average temperature, and then integrating over the surrounding area to obtain a regional average. To assist in easier merging between this and the provided dataset, the date was standardized to be the first date of each month.

Our conclusions related to temperature in this paper are based on pre-Covid data, as it is far more beneficial to have long-term insights about oil. As Kumar et al. demonstrate, Covid did not cause a significant long-term impact on energy consumption (Kumar et al., 2022).

After this analysis, the next step was to identify the most detrimental fossil fuel to the Earth and understand the underlying reasons. By integrating temperature anomalies with emissions and consumption data based on their dates, we utilize emissions and temperature deviations as key indicators of global warming.

We plotted the temperature anomaly over time and saw a clear trend upward with volatility. The volatility can be attributed to the fact that global warming affects temperature in the extremities.

Then, we wanted to see how different energy types play a role in emissions and temperature

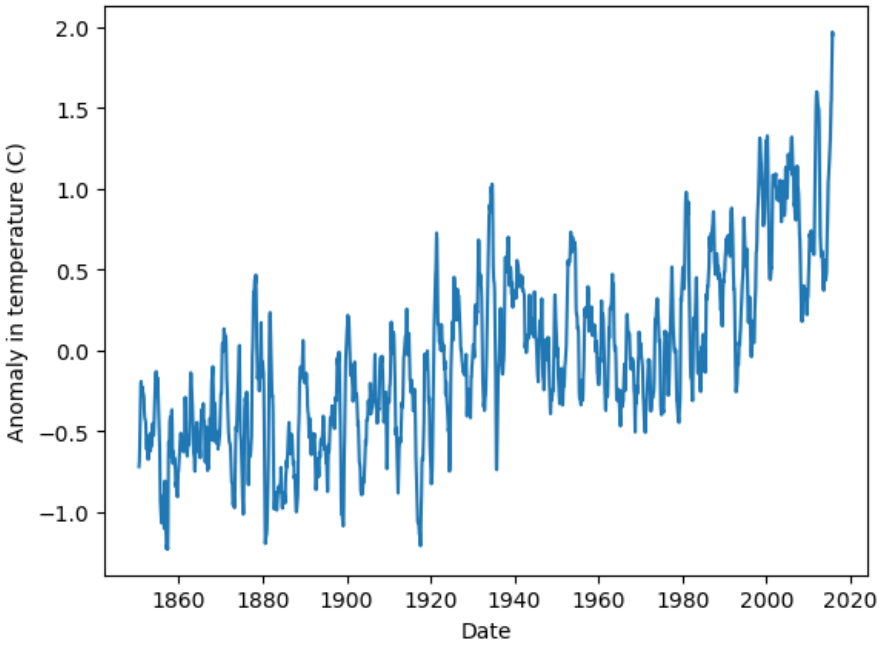
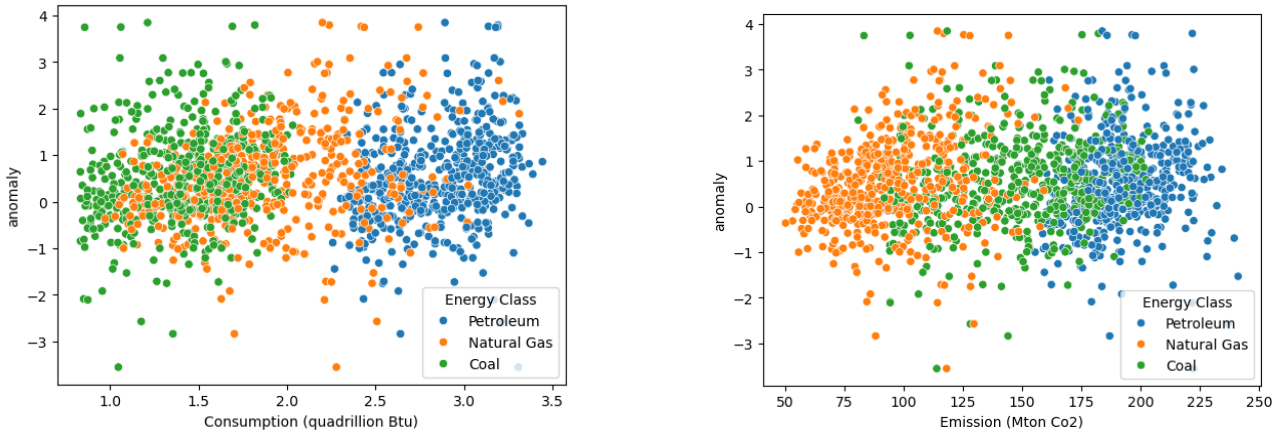


Figure 2.4: *Anomaly temperature over time*



(a) *Consumption versus anomaly* **(b)** *Emission versus anomaly*

Figure 2.5: *Consumption and Emission versus anomaly*

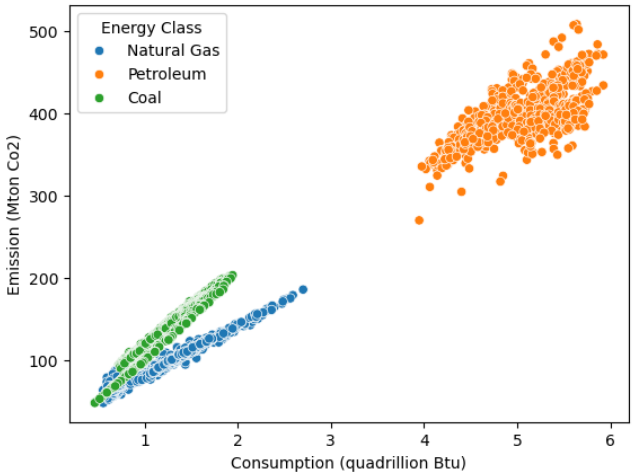


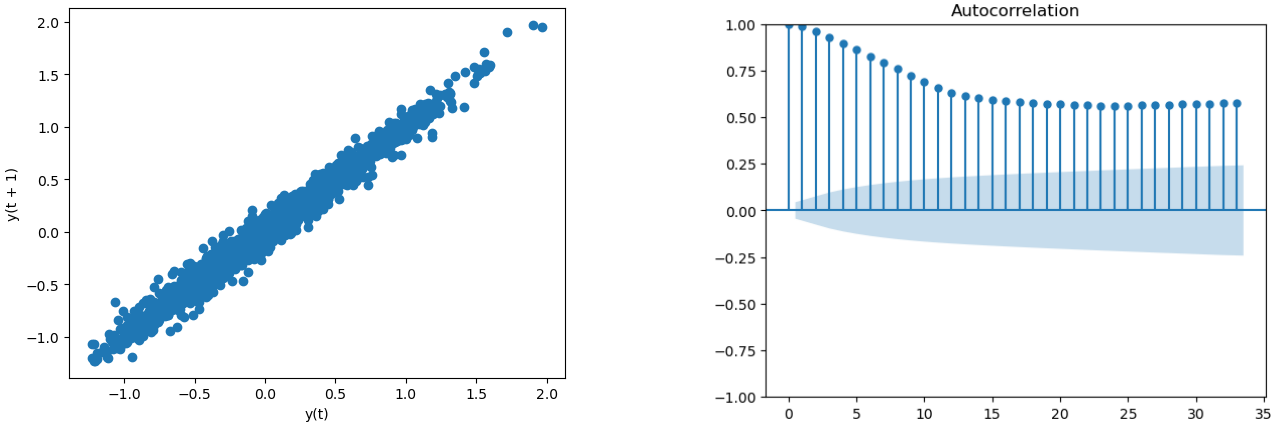
Figure 2.6: *Consumption versus emission for energy classes*

anomalies. Many clusters of petroleum and coal consumption do not affect anomaly. There seems to be a slight upward correlation. However, to truly explore this more, we wanted to examine whether there is perhaps a lag between consumption and anomaly.

A similar pattern is detected in the emission of energy classes over time. There seems to be a slight upward correlation but we intuitively hypothesized that it would make more sense to have lags in this data and see whether there is more correlation between lagged emission and anomaly data.

Finally, there is a very strong positive linear correlation between consumption and emission. As we are using emission as a signal to indicate climate change, we have established that there is a strong positive correlation between the consumption of these fossil fuels and climate change. This is supported by the Pearson correlation coefficient, a measure of linear correlation, being 0.75.

There was no linear correlation between temperature and other variables. Thus, it is worth exploring whether the data is linear or non-linear. To do that, because this is a time series graph, we graphed the data on a lag plot.



(a) *Lag plot of temp*

(b) *Anomaly autocorrelation plots*

Figure 2.7: *Anomaly lag and autocorrelation plots*

The lag plot is a useful information technique for time series data because it shows the randomness

of the data, and patterns that demonstrate how much of a variable depends on the previous days' values.

The temperature anomaly lag plot has a very clear discernible pattern which means it depends highly on previous days' anomaly. This clear pattern was observable even when we changed the lag time. This suggests that the data is linear. Additionally, the results from the autocorrelation plot also show that the temperature anomaly is highly dependent on the lags. Hence, when the anomaly temperature falls, it is likely to keep falling or when it rises, it is likely to keep rising.

After taking the log of each of the variables to make the data stationary, the time lag plot shows that emission per consumption, emission, and consumption were not dependent on their previous day's values.

To prove causation, we first started by proving correlation between:

1. Consumption and Emission
2. Consumption and Anomaly Temperature

for each of the three main fossil fuels.

To prove a high correlation between consumption and emission, we used the Pearson correlation. This is because the relationship between consumption and emission was linear, and the data was homoscedastic. Their Pearson coefficient was 0.75 and thus there is a strong correlation between them.

To prove the relationship between consumption and temperature, we used Mutual Information. Mutual Information is a way to assess the nonlinear relationship between two variables. Since during EDA, there was no linear relationship between any of the variables and anomalies in temperature, we knew that we needed a metric that would capture non-linear relationships. Mutual Information measures how much one random variable tells us about another, by specifically measuring the decrease in uncertainty of predicting a variable given another. Intuitively, we also hypothesized that there might be a lag in the effects of these variables. Thus, we tested the ideal lag time and the mutual information at the same time. We tested the lag from 0 to 60 months for each of the variables and assessed its accuracy with the mutual information. Here is a table that summarizes the results:

X Variable	Y Variable	Correlation	Time Lag (months)	Mutual Information
Emission	Anomaly Temp	0.10	32	0.08
Consumption	Anomaly Temp	0.12	28	0.13
Emission/consumption	Anomaly Temp	-0.03	59	0.16

Table 2.1: *Relationship Information for Anomalies in Temperature*

This shows that consumption and anomaly temperature have mutual information overall. Additionally, the effect of consuming fossil fuels on temperature anomalies is the most 2 years after.

Interestingly, if emission per consumption reduces, perhaps due to laws, it takes 5 years for its effects to be felt.

Additionally, emission and consumption have a 0.75 Pearson correlation and consumption has a -0.52 Pearson correlation with emission per consumption. The first can be explained by the fact that consuming fossil fuels produces a lot of emissions and the latter can be explained by lawmakers targeting emission per consumption rather than consumption because it is easier to make fuel efficient

than to reduce consumption. By making fuels more efficient, consumption is likely to decrease. Additionally, these laws might affect public sentiment and thus that might change consumption. This will be explored in later sections. Decreasing emission per consumption affects anomaly temperatures 5 years after the effect.

We wanted to find this ideal lag as a control for comparing the other fuels, which is why we first found this table for the aggregated datasets.

To prove weak causation with both of these factors for energy classes, we used Granger Causality. Granger Causality tells us how well a time series can predict another time series. It requires that the data is stationary and that there is no inter-dependency between its variables. Our data meets both of its requirements, that it is stationary and only has one-way causality. A caveat also is that Granger Causality doesn't prove total causality, just gives a stronger relationship between two variables. We let our threshold be 0.05.

2.3.1 Petroleum

X Variable	Y Variable	Granger Causality p-value
Emission	Anomaly Temp	0.24
Consumption	Anomaly Temp	0.02
Emission per consumption	Anomaly Temp	0.00

Table 2.2: *Petroleum Granger Causality*

We let our null hypothesis be that the X variable does not cause the Y variable. Because the p-value of consumption and anomaly temperature and emission per consumption and anomaly temp is less than 0.05, we can reject the null hypothesis. Thus, consumption has reason to cause anomaly temperature to rise. Additionally, emission per consumption has reason to cause anomaly temperature to decay. For each of the Granger tests, we utilized the results of the model previously to find the ideal lag time and evaluated the ssr-based chi2 test because it can be used to prove weak causality and doesn't assume their variances are equal like the F-test. We satisfy the requirements of the Chi-squared test that the data has a large sample size, and that the measurements are independent.

2.3.2 Natural Gas

X Variable	Y Variable	Granger Causality p-value
Emission	Anomaly Temp	0.00
Consumption	Anomaly Temp	0.00
Emission per consumption	Anomaly Temp	0.07

Table 2.3: *Natural Gas Granger Causality*

We let our null hypothesis be that the X variable does not cause the Y variable. We can reject the null hypothesis for both emission and anomaly temperature and consumption and anomaly

temperature. Thus, consumption has reason to cause anomaly temperature and emission has reason to cause anomaly temperature.

2.3.3 Coal

X Variable	Y Variable	Granger Causality p-value
Emission	Anomaly Temp	0.00
Consumption	Anomaly Temp	0.02
Emission per consumption	Anomaly Temp	0.01

Table 2.4: *Coal Granger Causality*

We let our null hypothesis be that the X variable does not cause the Y variable. We can reject the null hypothesis for all these variables. Thus, the emission has reason to cause anomalies in temperature, consumption has reason to cause anomalies in temperature, and emission per consumption has reason to cause anomaly temperature.

We see that consumption has reason to cause anomalies in temperature for all 3 fuel types. It is also correlated to emissions, and current literature supports causation between consumption and emissions (n.d.). Interestingly, coal's emission and consumption variables were good predictors for anomaly temperature. However, because coal is not as widely used as Petroleum and doesn't cause as much net emission and consumption, we focus on oil. We do this simply because oil is consumed the most and produces the most emissions. Additionally, it is easier to develop technologies to reduce its emission per consumption ratio. Therefore, we find that oil and coal are bad for the environment fossil fuels because their consumption causes an increase in temperature anomaly and their consumption causes a rise in emissions.

We also found that the emissions per consumption ratio of fuel matters when for Petroleum and Coal. Thus, it is important to continue to develop solutions that expand this technology.

Thus, through this exploration, we performed a sanity check first, ensuring that our unit conversions matched what exists in the literature. For example, coal has the highest emission per consumption but oil is consumed the most and produces the highest emissions (Lindsey, n.d.). We provided a reason for why we focus on primarily oil throughout this report: because it has the biggest environmental impact. Lastly, we establish that oil consumption can cause temperature anomalies, or more generally global warming, and that consumption causes emissions.

2.4 Sentiment Analysis: Whose Opinions Matter?

In tackling the problems brought forth by oil, we need to look at how people think about oil and oil-related policies to see if it affects oil consumption and price. We can only propose viable solutions to global warming problems when we know whose opinions matter the most. The three most important groups to consider are 1. policymakers, 2. the business community (corporations), and 3. ordinary citizens and investors. We believe that a measure of importance can be seen from how much their attitude/speech/actions affect the cost of buying oil and consumption. Assuming market

equilibrium, the consumption could also represent the demand for oil in the United States. We will perform regression tasks on different groups' opinions (we will define how we quantify this) and compare the standard regression coefficients. The group that has a high correlation with oil prices/consumption/demand is likely important in the energy policy-making and market-making process, either because their action directly caused the changes or because they are reacting strongly to the changes. To analyze which is the case, we will perform causality tests on the respective time series plots to see if they lead or lag the changes in oil prices/consumption/demand.

2.4.1 Literature Review: Public and Corporate Sentiments

Before we move on to analyzing the implications of our sentiment analysis results, we will talk about public and corporate sentiments about oil. We were unable to find publicly available datasets about public and corporate sentiments, but there is a plethora of literature that explores how public and corporate sentiments affect the oil market. Here we shall provide a summary of the major results. There was, however, no analysis of how the sentiments of policymakers influence the market. Thus we will explore this in the next section to fill this gap in literature.

In classical energy economic theory, investor sentiment does not play any role in oil prices or their volatility (Qadan et al., 2018). However, multiple researchers have found that there is indeed a correlation between investor sentiments and oil prices. Deeney et al. suggested creating sentiment indices from market activities such as asset volatility, put call ratio for equity options (market fear), IPO volume and returns (speculation), and market volatility (Deeney et al., 2015). Through constructing time series models, they concluded that sentiment is an important consideration when explaining crude oil prices. Similarly, Qadan et al. found that investor sentiment, captured by nine different proxies, has a significant effect on oil prices. They expanded upon the solely financial approach proposed in Deeney et al. and included sentiment data about consumer confidence, sentiment, and uncertainty. Through their analysis, they were able to show that investor sentiment has significant effects on oil prices and their volatility. Contu et al. have also shown that sentiments about climate change impact the consumption of oil-related products (Contu et al., 2021).

In an International Monetary Fund (IMF) report in 2023, Bogmans et al. used text-based firm-level measure of climate policy exposure to find that climate policies have led to a global decline of 6.5 percent in oil and gas investment between 2015 and 2019 (Bogmans et al., 2023). This suggests that firms may have pre-emptively cut investment in response to downward shifts in expected future demand for fossil fuels. This finding is supported by economists Areszki et al. (Arezki et al., 2022).

What the above literature informs us is that energy/climate policies are important in influencing the oil market. This gives an increased incentive to explore how policies and policymaker sentiments relate to the oil market.

2.4.2 Data Collection

Now we present how we collect and quantify the abstract concept of "opinions" for policymakers.

We believe that the U.S. President, the highest-ranked policy-maker, is a good representation of the class of policymakers. This is because

1. every law passed in Congress requires the signature of the president, and the president has the power to veto any law they dislike;
2. the president can appoint people to key executive positions such as the Departments of Commerce, Energy, Treasury, and Transportation, which all have key links to oil;
3. the President has ample media coverage such that their public opinions could affect market trends (U.S. Bank, 2024).

As such, we believe that public speeches from presidents are adequate representations of the beliefs of the majority of policymakers.

The Miller Center of Presidential Scholarship at the University of Virginia (UVA) archives a comprehensive list of public speeches by U.S. presidents (UVA Miller Center, 2024). From the website, we scraped and cleaned every significant public presidential speech available since the year 1973 (the year in which our oil emission data started). Other than the text, we also recorded the date of the speech. We know the speaker from the date since we know exactly the tenure of each president.

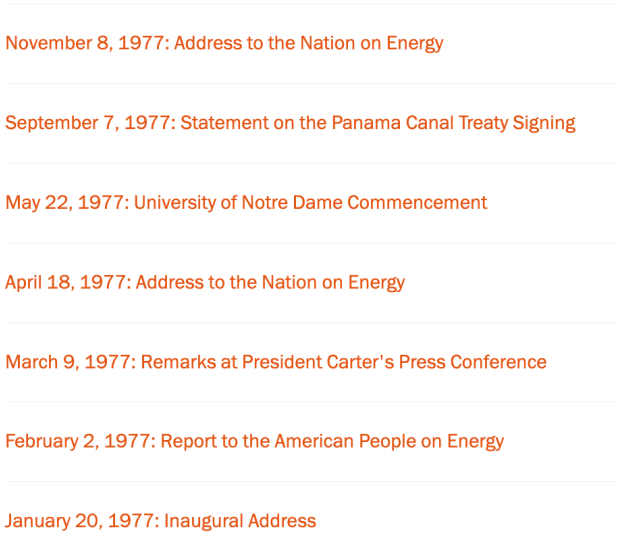


Figure 2.8: *A Sample List of Speeches by President Jimmy Carter (D). The Iran Oil Crisis happened during Carter's term; indeed Carter made a lot of public speeches about energy.*

To obtain their opinions on energy-related topics, we curated a list of words that closely relate to oil/oil-related products, as well as another set of words that closely relate to climate change topics. Some sample words relating to oil are *oil, petroleum, OPEC, gasoline, crude, refinery*, etc. Some sample words relating to climate are *climate, green, sustainable, pollution, renewable, geothermal*, etc. [1] For each of the groups of words, we found each occurrence of the word in each of the speeches and recorded the sentence the word is in. If multiple oil/climate-related words appear in the sentence, we record it multiple times. This is because multiple appearances suggest that this sentence is highly related to oil/climate topics and thus merits extra attention. We were able to obtain a total of 1700 sentences each for oil and climate. For each piece of speech, we also record how many oil/climate-related sentences there are in the whole speech as an additional metric for how relevant the speech is about

[1] For a complete list of words used, reference appendix A.2.

energy/climate concepts. The more sentences there are, the more relevant. In future analysis, we will use this as a measure of relevance.

A good sentence example would be the sentence *"We can continue using scarce oil and natural gas to generate electricity and continue wasting two-thirds of their fuel value in the process"* from President Carter's 1977 Address to the Nation on Energy. An example of a bad sentence would be *"Listen to Master Sergeant J.P. Kendall of the 82d Airborne: 'We're here for more than just the price of a gallon of gas.'"* which comes from President Bush Sr.'s 1991 address to the Nation on the Invasion of Iraq. The main subject of this sentence is not gasoline, but it is still misclassified as oil-related.

To improve upon our model, we will need to train context-aware language models that can recognize if a sentence is truly about gas. One may first think of natural language processing techniques such as topic modeling, but topic models are not as good at generating themes as checking themes of different sentences. Here, we describe another scheme in which we can achieve this: we can encode each of the sentences as an embedding of n-dimensional vector that describes the context of the sentence (existing tools such as the OpenAI or Vertex AI APIs can achieve this). Then, we can perform unsupervised clustering on the embeddings of different sentences. By the nature of embeddings, sentences that are similar in meaning have similar embeddings, which means that we should be able to see clusters of energy/climate-related sentences. This way we can identify our target sentences with high accuracy. However, training such a model on hundreds of thousands of sentences would require more than ten hours of embedding time and hours of fine-tuning (without high-performance computing tools), as opposed to the keyword identification method we used which takes less than 10 seconds. Upon manually inspecting around 100 randomly sorted sentences, we saw that we were able to identify actual oil/climate-related sentences with an accuracy of 95%. Thus we believe that the data we obtained is a good measure of presidential speech about energy/climate topics.

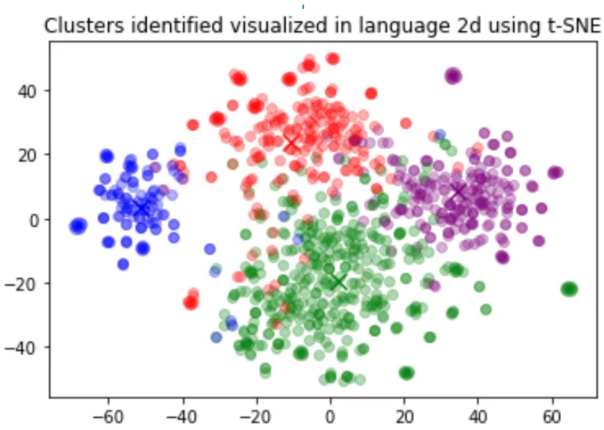


Figure 2.9: *An example of clustering sentences based on embeddings, reduced to 2D by t-SNE. Different clusters may relate to climate/energy/foreign policy/national defence, etc. Image courtesy of OpenAI.*

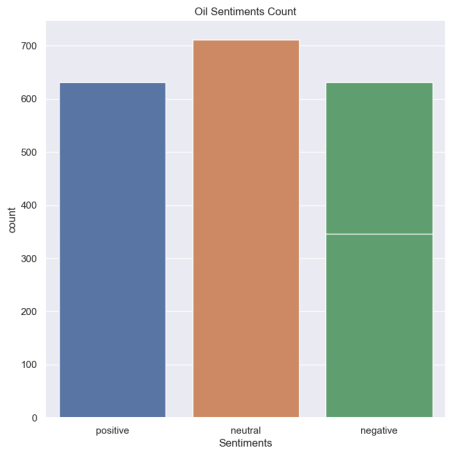
2.4.3 Sentiment Analysis

Now that we have sentences related to oil/climate, how can we extract useful information from these sentences? A simple but insightful model we can perform is sentiment analysis. Sentiment analysis involves analyzing text to determine the sentiment expressed, whether it's positive, negative,

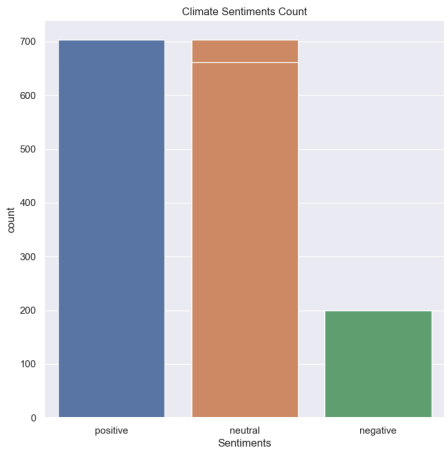
or neutral. In our context, sentiment analysis can help us identify the opinions a president has about energy/climate-related topics. The Natural Language ToolKit (nltk) package has a dictionary of pre-classified sentences that we could train on, which greatly simplifies the embedding and unsupervised clustering scheme proposed earlier. We were able to train the model within 10 minutes.

Here we describe in detail the process of sentiment analysis. To preprocess our sentences, we convert all of them to lower spaces, strip any extra spaces, remove any HTML tags potentially left in the scraping process, remove punctuations and special characters, and expand contractions. We also remove stopwords (words without actual meanings) such as 'a', 'the', 'and', etc. Finally, we take the stem of all the words we have (e.g. precipitating turns into precipitate), which correspond to existing tokens[2] in the nltk package. Upon preprocessing, we feed all our transformed sentences into the pre-trained sentiment analysis model.

In the model, we use a pre-existing sentiment lexicon, which is a collection of words or phrases associated with positive, negative, or neutral sentiment. According to the lexicon, we analyze the sentences and assign a sentiment score to each sentence, paragraph, or document based on the sentiment lexicon. The model returns a polarity measure of how positive or negative a sentence is. If the polarity score is positive, we categorize the text as positive; if the polarity score is negative, we categorize it as negative. Else we categorize it as neutral. For our analysis, however, we keep both the polarity measure and the sentiment classification. This way, we can analyze both general sentiments and the degree of the sentiments. Below we present the cumulative sentiment analysis results for oil and climate-related speech:



(a) *Cumulative Sentiment of Oil*



(b) *Cumulative Sentiment of Climate*

Figure 2.10: *Sentiment Analysis for Oil and Climate-related Speech*

Upon inspection, we can immediately see that **most presidential speeches are positive/neutral about the environment while for oil-related concepts it is more of an even split** (figure 2.10). For comparison, when we performed the same analysis on 1000 recent tweets sampled over X (previously known as Twitter) which we obtained from the Twitter API, we found a much clearer negative sentiment (65% negative) towards oil. However, before we perform further analysis, we need to make

[2] A token refers to a single, atomic unit of text. It could be a word, a part of a word (subword), or even a punctuation mark. Tokenization is the process of breaking down a text into individual tokens.

sure that our sentiment analysis does produce valid sentiments. To do this, we shall perform a few anomaly-detection tasks to show that our sentiment records the data we need.

We can use our data to see if we can identify key oil events from sentiments. To get the sentiment of a speech, we take the average sentiment of the sentiments of each energy/climate-related sentence. Ranking our results by the degree of negative oil sentiment and then relevance to energy/climate, we see that the top 3 speeches relate to (a) the Iran Oil Crisis, (b) the Iraq war, (c) the 1980s oil glut, all of which are major oil events that negatively affect the United States. To illustrate that we can identify key events, we plot the oil sentiments over time (figure 2.11). Indeed, they correspond to periods of turmoil in the energy market (Iran Oil crisis, the 1980s oil boom and glut, the Gulf War, the Iraq War, the 2008 Financial Crisis, and COVID). Another interesting thing to notice is that **major events are associated with greater polarities instead of a specific sentiment**. We will return to this insight with an analysis of how it relates to oil prices and consumption.

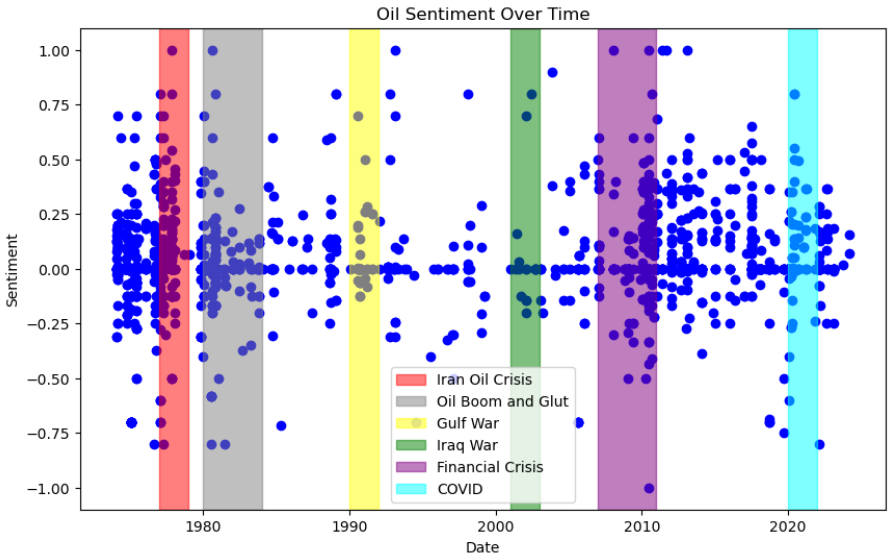


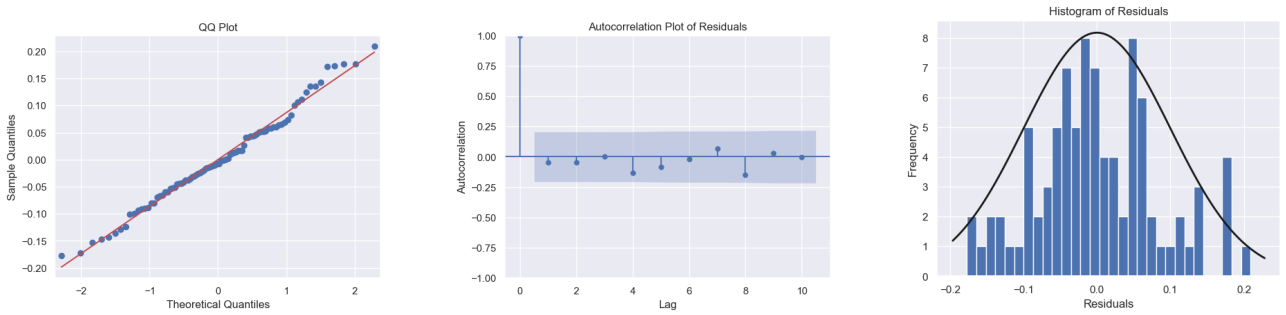
Figure 2.11: *Sentiment about oil over time plotted with major events. Periods of turmoil in the U.S. energy market usually correspond to high polarity in sentiments.*

With the sentiments verified, we can start analyzing it in detail. The first natural thing to do would be to compare how presidential speeches lean towards energy and climate. We initially hypothesized that there should be a negative correlation between oil and climate sentiments. That is, presidents who have high affinities for oil would have low or little sentiments for climate. This is because the development of climate-friendly policies often necessitates cutting carbon-based fuels such as petroleum, and vice versa.

The Jarque-Bera test for normality of residuals returns 0.90 and low skew and kurtosis. The Breusch-Pagan test for homoscedasticity returns a p-value of 0.98, which means that we reject that our data is heteroscedastic. Finally, the Durbin-Watson test returns a statistic of 2.07, indicating no autocorrelation. Looking at our three plots below, we can verify that is true.

Thus, we can use linear regression on our data.

In figure 2.13 we plot sentiments for oil vs. climate for all speeches. The hue of the plot indicates the year the speech was in while the size of the dots indicates the sum of numbers of oil/climate-related



(a) *QQ Plot* **(b)** *Autocorrelation Plot* **(c)** *Residual Plot*

Figure 2.12: *Linear Model Gauss-Markov Assumptions Test.*

sentences. To obtain a single sentiment for each speech, we again take the average sentiment of the sentiments of each energy/climate-related sentence. Looking at the plot, however, we realize that oil sentiments and climate sentiments are not correlated in general. We performed the regular F-test and obtained a p-value of 0.20. At a significance level of $\alpha = 0.05$, we reject that there is a correlation between oil sentiments and climate sentiments in general.

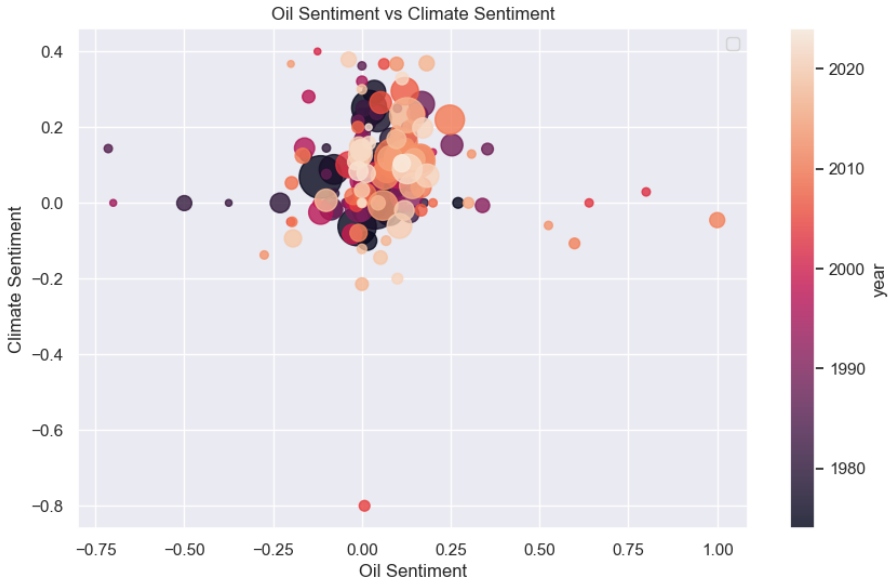


Figure 2.13: *Climate Sentiments vs. Oil Sentiments. The hue indicates the date of the speech; the size of the points indicates how relevant they are to energy and climate. There is no correlation at all.*

But notice that there is a significant number of outliers with small sentence counts, which are weighted *equally* as the speeches with large sentence counts. Speeches with less relevant sentences are more likely to have been misclassified as energy/climate-related. To avoid this bias, we decided to look at speeches only with a significant amount of oil/climate-related sentences (a total of more than 15 energy/climate-related sentences). Now, **there appears to be a strong correlation between oil sentiments and climate sentiments** (figure 2.14). We chose 15 to be the threshold since the 25th percentile for the number of oil/climate-related sentences for both is around 7. We checked that there was at least one highly energy/climate-related speech per year, totaling 90 speeches over 50 years. These speeches also tend to lean significantly and positively toward oil, meaning that relevant

American presidential speeches often express a leaning toward oil. The climate sentiments, however, remained normally distributed.

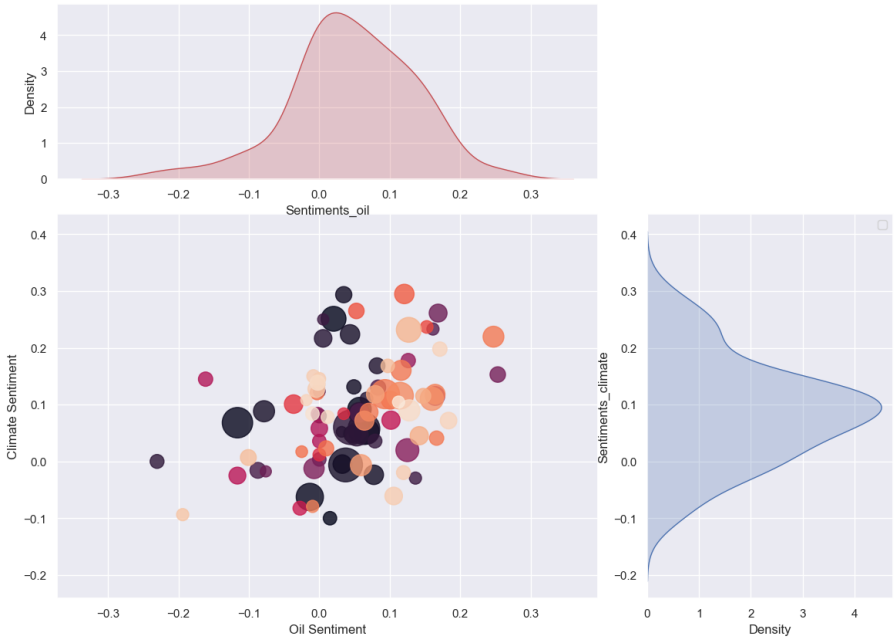


Figure 2.14: *Filtered Graph for Oil Sentiments vs Climate Sentiments. The years are similarly color-coded and the sizes still represent how relevant they are to energy and climate.*

The new linear regression model provides a p-value of 0.00295 for the F-test along with a linear Pearson's r correlation of 0.373. The standardized coefficient is 0.38, which means that climate sentiments lean to change less than oil sentiments.

The result was initially shocking. How can presidents who support oil also support preserving the climate at the same time, especially with a clear linear correlation? However, after investigating individual sentences that were classified as positive/negative, we reached a new hypothesis. Instead of measuring the degree of positive/negative sentiments, **the magnitude of the sentiment polarity score measures how strongly an opinion was stated**. Conceptually this makes sense - a speech about oil could either be praising the U.S. for stabilizing the oil market and bringing significant amounts of oil-related jobs back [3], or dreaming how the U.S. can reach energy independence through the development of green energy [4]. Both sentences would have a positive sentiment score (0.05 and 0.03), but their connotations are exactly opposite. However, both sentences do express strong opinions about energy. Similar situations might happen for climate-related sentences. Thus, instead of thinking of the linear relationship above as a statement of "presidents who support oil love support climate," we should see it as "*speeches often contain a similar level of intensity in sentiment regarding both climate and energy-related issues*".

[3] Example: "This weekend, the United States also helped facilitate an unprecedented agreement among the 23 nations of OPEC Plus—that's OPEC plus additional energy-producing nations—representing many of the world's largest oil-producing countries to stabilize oil markets." (Trump, "Coronavirus Task Force Briefing")

[4] Example: "Because we know we can't power America's future on energy that's controlled by foreign dictators, we are taking big steps down the road to energy independence, laying the groundwork for new green energy economies that can create countless well-paying jobs." (Obama, "Remarks on the American Recovery and Reinvestment Act")

A way to show that this claim is true is to perform paired t-tests on aggregate positive and negative sentiments of each speech to show that positive and negative sentiments have similar polarity magnitude distributions. The assumptions for the paired t-test are satisfied as we have sentences that come from the same set of speeches. We have also previously demonstrated homoscedasticity and normality. The independence assumption is not satisfied for specific pairs, but we could aggregate negative and positive sentiments in a speech to reduce dependence. The t-test returned a p-value of 0.058, which is greater than our significance value of $\alpha = 0.05$. Thus we reject that there is a statistically significant difference between the magnitudes of positive and negative sentiments in general. However, it is worth noticing that the p-value is barely over the significance value, meaning that we should proceed with this claim with caution.

2.4.4 Oil Price can be Modeled with Sentiment

Now we will relate our sentiment analysis results back to oil consumption and prices. Remember that we were trying to see what impact, if any, do opinions of policymakers have on oil prices and consumption. We were given oil price data from the U.S. Energy Information Administration. The prices of three commodities were provided: Natural Gas, West Texas Intermediary Crude Oil (WTI), and Brent Crude Oil prices. WTI represents intermediary crude oil prices in the U.S. while Brent Crude Oil prices measure oil prices across the world. Since we are only analyzing data in the United States, we used the WTI as a measure of oil prices in the U.S. From now on, “oil prices" would refer to WTI in units of dollars per barrel. The consumption data we use is the same that has been explored earlier in this report. Since oil prices are only recorded after 1986 and till 2024 we cut our sentiment and consumption data to after 1986 and before 2024.

The oil prices are recorded daily, but consumption data is recorded monthly and sentiment data is recorded nonuniformly. To account for missing values, we used linear interpolation to fill in the data for consumption data. For consumption data, it makes sense to use linear interpolation because consumption varies (in most cases) continuously. However, using linear interpolation for sentiment may not be as good of a choice. Thus, we decided to use spline interpolation, specifically utilizing a 2nd order spline, to more accurately estimate the missing sentiment values. This spline-based approach fits a piecewise polynomial curve to the available sentiment data points, producing a more nuanced estimate of sentiment trends. While the interpolation may still introduce some bias, the spline method is a far more suitable choice than the linear alternative when dealing with the erratic nature of the sentiment time series.

To see oil price trends, we plotted oil sentiments with major events again, but this time we juxtaposed it with oil prices as well (figure 2.15). We can see that drastic changes in oil prices correspond to the major events outlined earlier, with the addition of the 2010s Oil Glut. Thus we hypothesize that extreme sentiments would be able to help us model oil prices and consumption. To further differentiate between extreme and moderate sentiments, we exponentiated max sentiments. A preliminary analysis of the correlation (Pearsons’s r) between (exponentiated) max oil/climate sentiment and oil price gave coefficients of correlations of 0.24 and 0.15. The correlations are not high. If we use the statistical interpretation of Pearson’s r, we would get that the square of the correlations, or around 6% of variations in oil price can be explained by oil sentiments. Similarly, only 3% of oil

variations can be explained by climate sentiments. We performed similar tests on consumption and found that the correlations are 0.25 and 0.24 - higher than the correlations for the price, but still low overall.

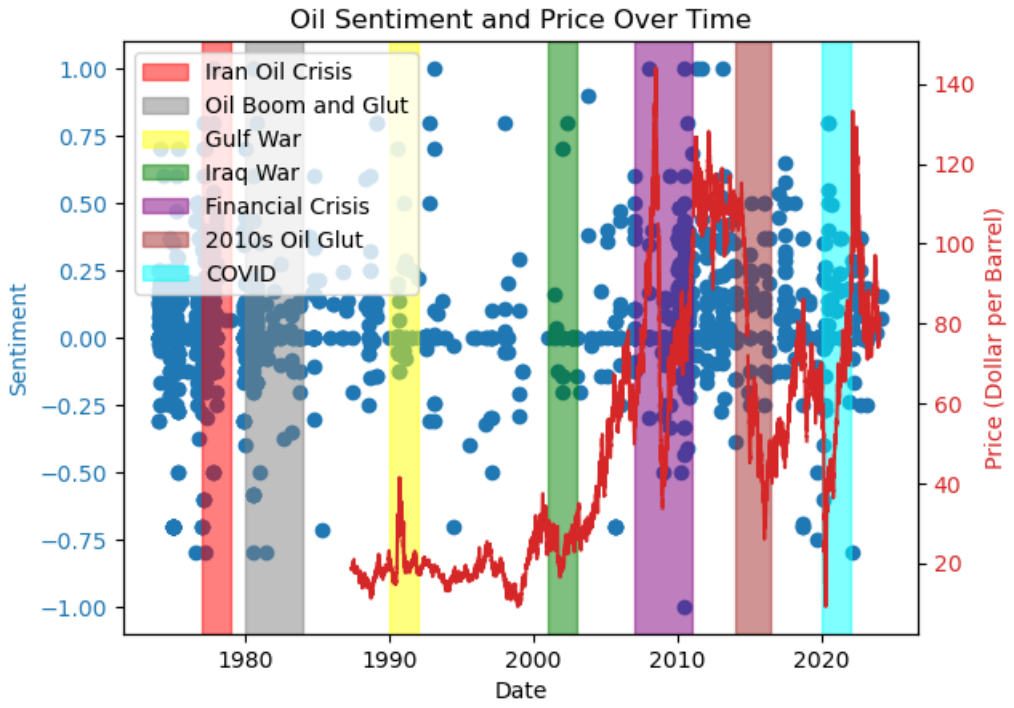


Figure 2.15: *Oil price plotted with oil sentiment and major oil events.*

Does it mean, then, that sentiments are just unable to model oil prices? We believe that the answer is no. Correlations assume independent observations; but for time series data such as oil prices and consumption, every piece of data necessarily depends on the data from the last observation. Thus, we need to employ time series models for our time-dynamic data. Specifically, we employ a dynamic linear model (DLM) on oil prices (Laine, 2019). A dynamic linear model can be seen as a general regression model where the coefficients can vary in time [5]. Most time series models assume stationarity of the underlying process; however, dynamic linear models are adept at handling non-stationary processes, missing values, and non-uniform sampling as well as observations with varying accuracies. However, this also means that DLMs would smooth data to extract its general trends. To account for the non-trend parts, we employ a stochastic part to account for extreme changes. Financial data often exist in fat-tailed distributions, i.e. extreme events such as financial crises happen more often than regular models such as normal distributions predict (Taleb, 2008). A stochastic part would deal with the fat-tailed parts.

The stochastic model can be understood with the following formula:

$$\text{Price}(t + \mathrm{d}t) = \text{Price}(t) + (\alpha \text{OS}(t) + \beta \text{CS}(t) + \gamma \text{D(t)})\,\mathrm{d}t + v\,\mathrm{d}W_t\,,$$

where OS is the most polar oil sentiment, CS is the most polar climate sentiment, D is the demand/consumption, W_t is a Wienner process or Brownian motion and v is the variability of this

[5] In this sense, a DLM is not a linear model since the the way that the coefficients vary may not be linear

Brownian motion. The equation above is exactly an arithmetic Brownian motion on the price of oil (Bachelier, 1900). Now we show that in fact **the U.S. oil market follows an arithmetic Brownian motion**. This implies that the price of oil can be described as a linear stochastic process, where the changes in oil price are driven by a Wiener process (W_t) and the variability of this process (v).

The key implications of this model are:

1. The oil price is subject to continuous, random fluctuations, with the magnitude of these fluctuations proportional to the variability parameter (v). This captures the inherent uncertainty and volatility in the oil market.
2. The relationship between the oil price and the underlying factors (OS, CS, D) is linear, rather than exponential. This suggests that changes in these factors have a direct, proportional impact on the oil price.
3. On short timescales, it may be difficult to distinguish between linear and exponential models, as the high variance of the stochastic process can obscure the underlying functional form.

This analytic result can be obtained through the logarithm of a General Brownian Motion via Itô's formula; the derivation of this model will be presented in Appendix A.1.

Now we describe how we perform a DLM on oil prices. Notice from above that other than max oil/climate sentiments, we also included demand as a factor in the regression. As a basic principle of economics, the price of a good at equilibrium naturally depends on the demand for this good. The process of a DLM is similar to that for a regular linear regression, but instead of training the model on the whole dataset, we trained the model on a sliding window of 30 days. This way we can make sure that our model does not overfit to oil prices in the past 40 years or train on data that is no longer relevant. Our regression model did not include stochasticity because it is hard to estimate a fixed volatility constant for a non-stationary time series with dynamic mean and variance. Instead, we will add stochastic features after we finish training the DLM and test different volatilities with simulations of oil prices and consumption.

The assumptions of DLM are the same as that of general linear models. All of the Gauss-Markov assumptions were passed except for the Jarque-Bera test for normality. This means that we need to introduce stochasticity into our model.

Now we talk about the results of our regression. Since coefficients are time-dependent, we provide the 95% confidence interval for the coefficients. The confidence intervals for the values of α, β, and γ with their standard coefficients are:

Coefficient	Variable	2.5%	50%	97.5%	Standard Coefficient
α	Max Oil Sentiment	0.026	0.0694	0.112	around 1
β	Max Climate Sentiment	-0.070	-0.0124	0.045	around 0
γ	Consumption (quadrillion BTU)	-0.017	0.0650	0.147	around 0

Table 2.5: *Confidence Intervals for DLM Coefficients with Standard Coefficients*

Notice that max oil sentiment is the only dynamic variable that has constantly positive coefficients (and thus must have an impact). Max climate sentiment and consumption coefficients span positive and negative numbers, suggesting that they have variable impacts on oil prices. Additionally, since

both contain 0 in their 95% confidence interval and the coefficients are relatively small, their impact on oil prices may be mild. The difference in standard coefficients supports this point. The standard coefficients β^* measure how much oil prices (in standard deviations) would change if we change a regression variable by unit variance. Thus our standard coefficients indicate that oil price would change by 1 unit standard deviation when max oil sentiment changes by 1 unit standard deviation. This suggests that **max oil sentiment is indeed a good predictor**, contrary to what the initial correlation test indicated.

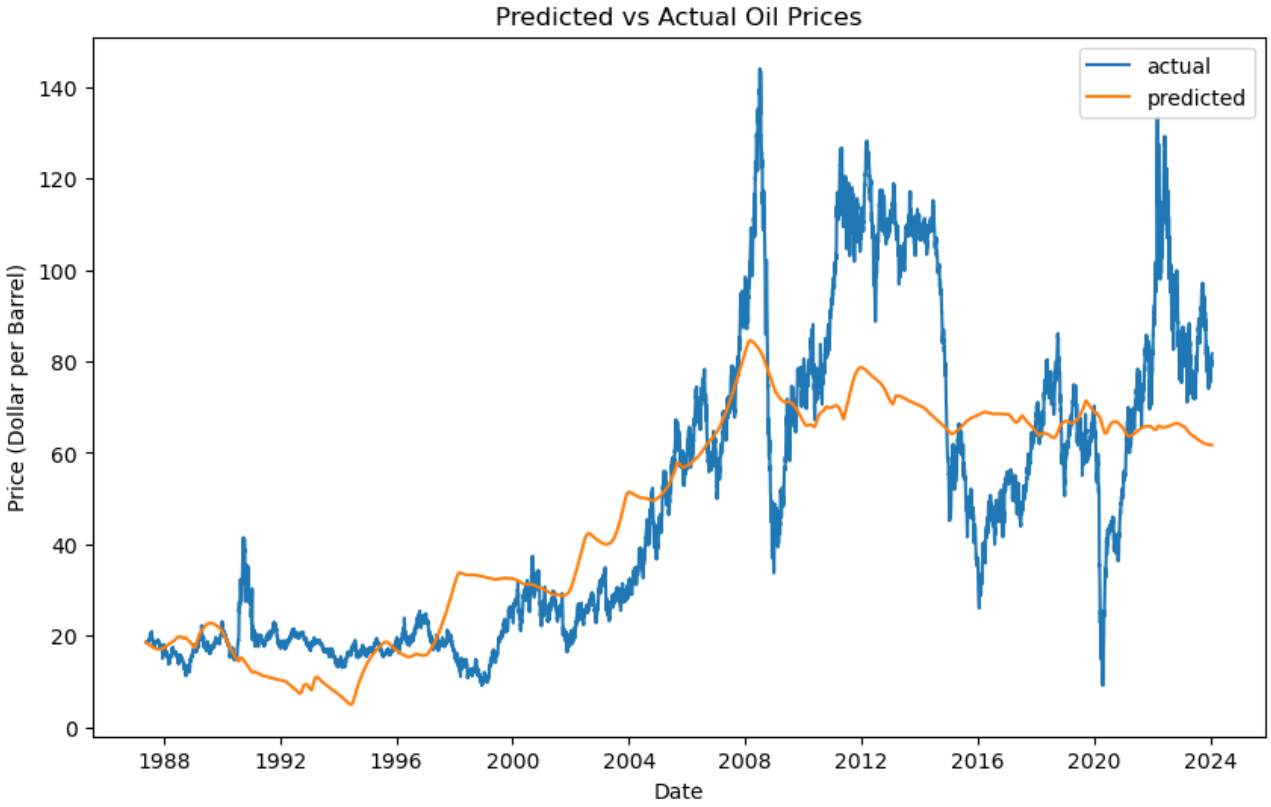


Figure 2.16: *Oil prices predicted by our DML model vs. actual oil prices.*

In Figure 2.16, we present a visualization of the oil prices predicted by our model against the actual observed prices. As expected, our model was unable to precisely identify the sharp peaks in oil prices over the time period analyzed. However, the model demonstrates a commendable ability to capture the general trends in oil price movements. The strengths of our Dynamic Linear Model (DLM) approach are particularly evident in the model's performance from 1986 to around 2008, where it successfully tracks the underlying exponential trend in oil prices. This is a direct consequence of employing the DLM framework, which is well-suited for modeling and forecasting such gradually evolving, non-stationary time series data. Nonetheless, the model's shortcomings become apparent in the post-2008 period, where it fails to keep pace with the more drastic, volatile changes in oil prices. This might be attributed to the train-test-validation data-splitting approach used during model development. Since extreme price spikes tend to be short-lived phenomena, it is possible that these critical events were not adequately represented in the training data, preventing the model from learning to properly handle such abrupt price movements. However, we will be able to add these

features with our stochastic model.

2.4.5 Stochastic Simulations Reproduce Oil Market Trends

Our full model is of the form $\text{Price}(t + \mathrm{d}t) = \text{Price}(t) + (\alpha \text{OS}(t) + \beta \text{CS}(t) + \gamma \text{D(t)})\,\mathrm{d}t + v\,\mathrm{d}W_t$. In the previous part, we tackled the DLM regression part and retrieved α, β, and γ. Since our price data is recorded in days, it is a natural length scale to define the infinitesimal length scale as a day. We initialize the prize to the price of oil on January 4, 1986 (the first day of our dataset) and pass it into our deterministic DLM along with our interpolated values for max oil/climate sentiment and consumption. However, now, instead of setting the oil price on January 5, 1986, as the value predicted by our model, we add a normal random variable with mean 0 and standard deviation v and set that to be the predicted value to add stochasticity. For the next day, we use the new value with stochastic noise as input to the model and add another random normal value. Thus, the deviation from the initial deterministic model would be amplified through time. The bigger v is, the more our prediction is going to deviate from the expected path.

Another feature we implemented for our simulations was a price control system. We implemented this because of two reasons: (1) the random noise we added might force oil prices very low by chance; (2) in reality, the price of oil does not fall below a certain value, or oil organizations such as OPEC would start measures to control production to raise oil prices (Bromberg, Michael, 2023). This lower price also grows with time. In our model, we set the initial lowest price to be 15 dollars per barrel and the growth rate to be around 0.1 dollars per year. This way we can strike a balance between the accuracy and reality of the model.

In figure 2.17, we graph 4 simulations of oil prices with our stochastic model with variability 0, 0.1, 0.5, and 1. With increasing variability, the stochastic predictions grow further apart from the deterministic prediction. Notice that our price control system was triggered multiple times between 1986 and 1996, indicating that it successfully curbed oil prices from dropping to unexpectedly low values.

Among all our predictions, the variability-1 model was best able to capture extreme events such as the Iraq War, the Gulf War, the financial crisis, and COVID. As such, we can infer that **the variability of the U.S. oil market is around 1 dollar per barrel per day** – which is around 1% to 5% of the actual price. Also, notice that the oil prices predicted from our model often peaks/troughs a few months before actual prices. This might be a result of the model being acute to sudden changes; the stochasticity it has only contributed to amplifying the change. This could also hint at a causal relationship between sentiment and oil prices, knowing that sentiment results from a few months ago could reproduce results now.

In figure 2.18, we train and simulate the same model but for consumption. Notice that there is much less variability in consumption data just by the nature of consumption being rigid. The variabilities we chose are 1/100 for that of prices although the scale is only 1/10 of prices. Due to the rigidness of prices, the deterministic prediction here was the best at capturing the trend in consumption. The high variability model was successful at predicting consumption until 2008 (just like before), but continued the downward trend until 2012. However, it was able to catch the COVID dip in 2020 although no other models were able to predict it.

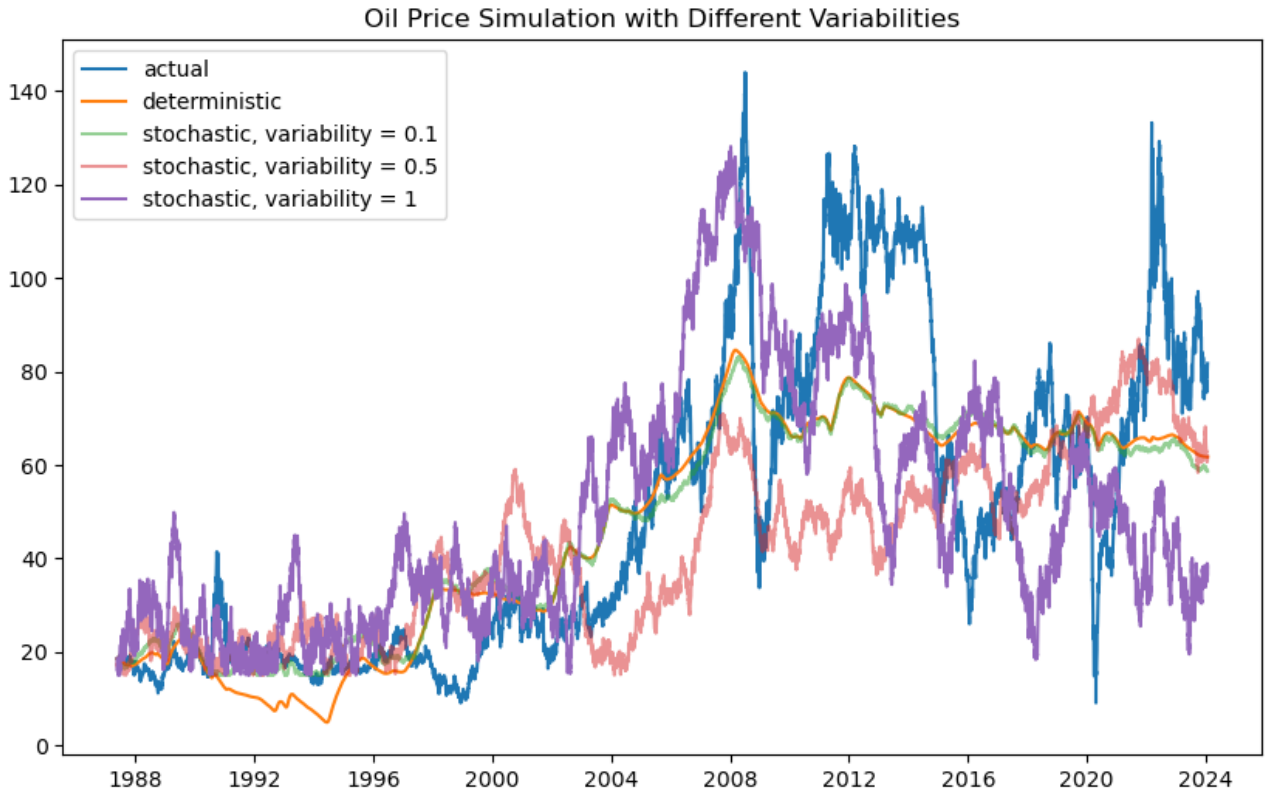


Figure 2.17: *Simulations of oil prices with different variabilities.*

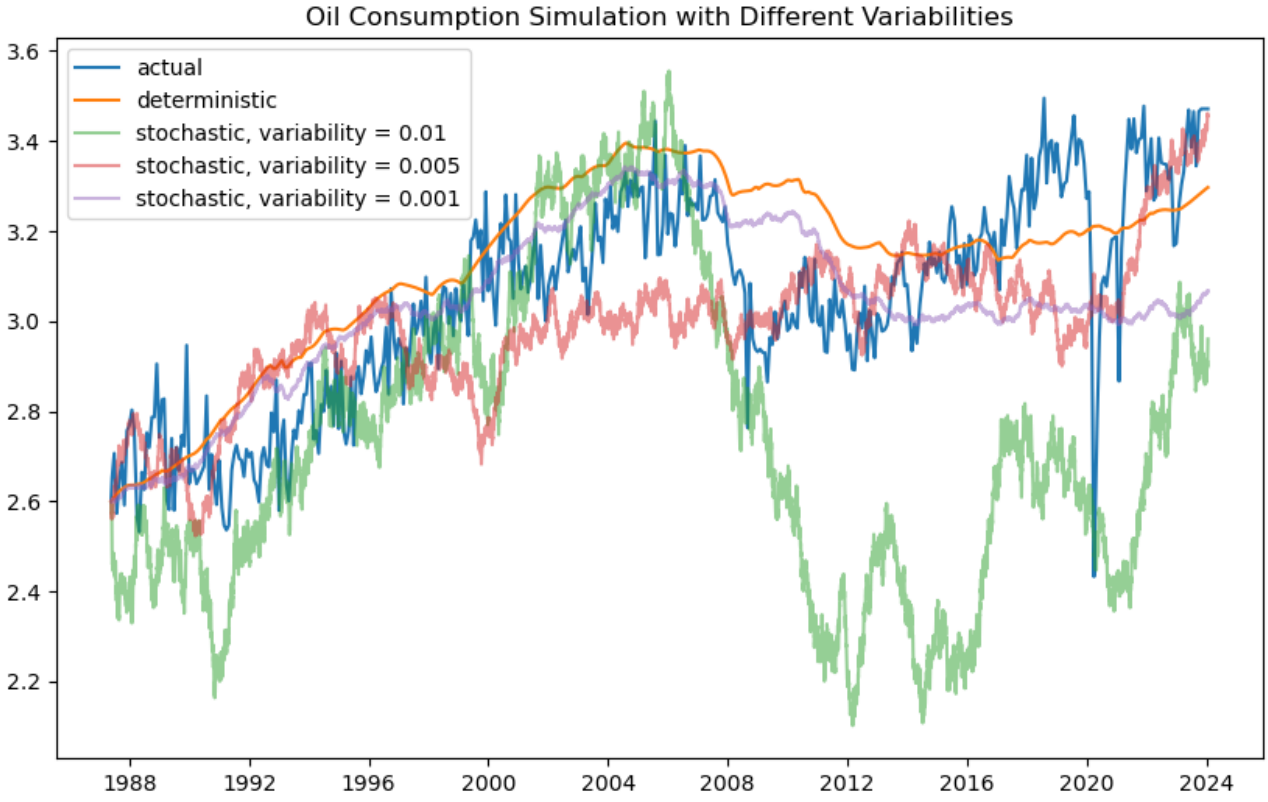


Figure 2.18: *Simulations of oil consumption with different variabilities.*

We found that policy maker sentiments, as reflected in presidential speeches, have a significant impact on oil prices and consumption because our model was highly accurate with 14.38% RMSE accuracy. This suggests that the priorities and perspectives of lawmakers play a crucial role in shaping energy market dynamics and, by extension, the trajectory of greenhouse gas emissions and climate change.

With this in mind, we will analyze how effective certain policies are at addressing climate issues, especially the emission of carbon dioxide.

2.5 How Does Policy Affect Emissions?

This section investigates the impact of environmental policies on state-level CO2 emissions and tries to determine whether or not policies are effective in reducing emissions. The best way to do this was to look at states in the US because they are more of a controlled environment to see the effect of policies in the same geopolitical region, with the same legislative frameworks. We also used emissions to assess the effectiveness of policies to target global warming because emissions don't depend on geolocation factors. We selected a diverse set of states, each with its unique environmental challenges and legislative landscapes, in order to get a general understanding of how policy impacts emissions at the state level. We utilized a Synthetic Control Model (SCM) as an analytical tool in order to determine what emissions would look like for the selected states in the absence of policy intervention and to establish potential causation between policy and emission.

2.5.1 Data Collection

Our primary focus was on CO2 emissions data, sourced from the "Methodology Report: Inventory of U.S. Greenhouse Gas Emissions and Sinks by State: 1990-2021" by the U.S. Environmental Protection Agency (United States Environmental Protection Agency, 2022). The dataset initially contained various greenhouse gases, but we filtered the dataset to exclusively include CO2 emissions. This was necessary in order to align the data with the rest of our analysis focused on the impact of policies on CO2 emissions.

In the initial dataset, there were some instances of missing values. To deal with this, the rows with missing CO2 emission values were removed. We did not want to fill in values as that would bias the emission values to the median. Additionally, the dataset included multiple entries for some states within the same year. To address this, we aggregated these entries by summing up the CO2 emissions, thereby representing the total CO2 emissions for each state per year. Figure 2.19 shows the emissions for all states.

An integral part of transforming the raw emissions data into a more analytically useful form was the conversion of total emissions into emissions per capita. We wanted to do this to account for different energy needs and thus emissions of different states. This involved obtaining historical state population data from the U.S. Census Bureau's Annual Estimates of the Population for the U.S. and States (U.S. Census Bureau, n.d.). The population data spanning from 1990 to 2019 was used to provide a comprehensive overview of the population for each state throughout our study period. This step enabled us to standardize emissions data on a per capita basis, facilitating a more equitable

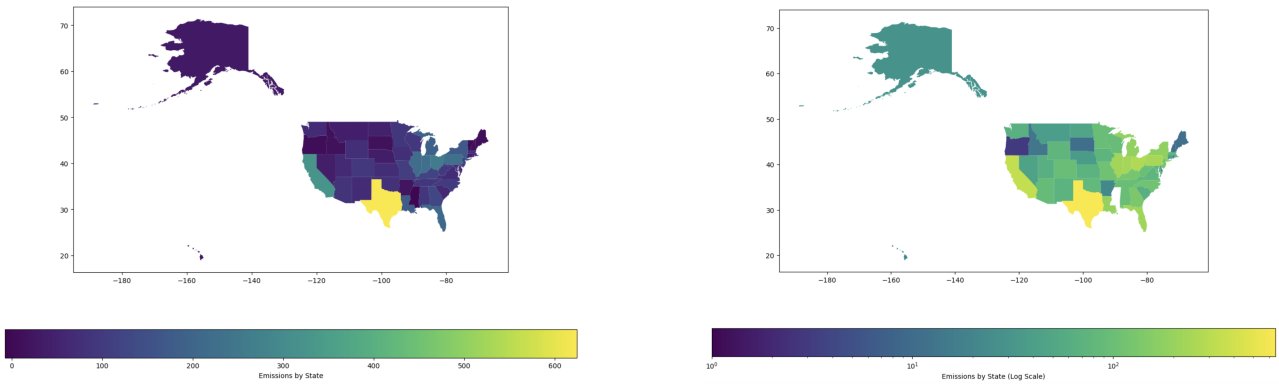
comparison across states with varying population sizes.

We found our data on state laws and incentives from the Alternative Fuels Data Center (Alternative Fuels Data Center, 2022) and employed a state-specific filter to extract a list of relevant laws and incentives enacted in each state. Figure 2.20 shows the amount of policies enacted by state from 1990-2019. This granular approach allowed us to closely examine the legislative landscape influencing vehicle emissions and alternative fuel adoption across different states and aggregate our findings to the US overall.

2.5.2 Feature Selection

Our analysis encompasses a diverse set of states—New York, California, Texas, Ohio, Washington, and Florida—each chosen for their unique environmental, economic, and political landscapes. These states provide a comprehensive representation of nationwide policy impacts.

New York and **California** are states with aggressive climate policies and high urban density, offering insights into the effect of environmental legislation. **Texas**, with its economic foundation in the oil industry, juxtaposes the previous two by reflecting the interaction between economic priorities and environmental initiatives. As an industrial heartland, **Ohio** shows the challenges and opportunities in transitioning manufacturing sectors to greener practices. **Washington**, with its significant investment in hydroelectric and renewable energy, showcases the potential for state-led clean energy transitions. Lastly, **Florida**'s vulnerability to climate change impacts such as rising sea levels provides a context for evaluating policy urgency and effectiveness.

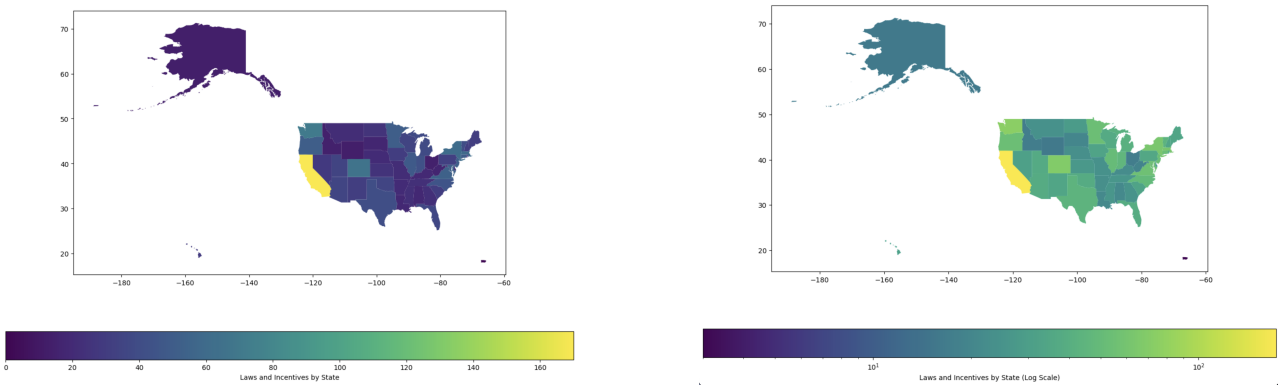


(a) *Total State Emissions* **(b)** *Total State Emissions (Log Scale)*

Figure 2.19: *Total State Emissions 2019 in million metric tons*

2.5.3 Synthetic Control Model

The Synthetic Control Model (SCM) is a statistical approach used for causal inference, particularly useful in evaluating one X variable as compared to different Y variables over time. It has been frequently used in evaluating the effectiveness of policies. It constructs a synthetic control by selecting a combination of control states that closely resemble the treated state's pre-intervention characteristics, without the implementation of the specific policy under study. SCMs are able to create a realistic counterfactual scenario, which illustrates what the emissions trajectory of the treated state would have



(a) *Laws and Incentives Count* **(b)** *Laws and Incentives Count (Log Scale)*

Figure 2.20: *Laws and Incentives Count by State*

been had the policy not been enacted. This counterfactual is constructed by assigning optimal weights to control states, ensuring the synthetic control mimics the pre-policy emissions trend of the treated state as accurately as possible.

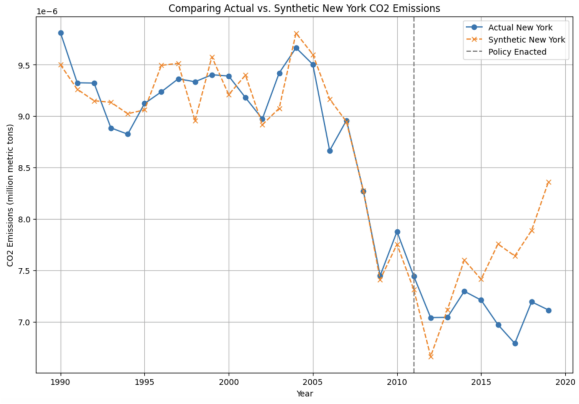
Control State Selection: Control states are chosen based on whether or not they have a similar policy enacted as the the treated state. This identification process involves analyzing historical policy data to pinpoint states that do not have similar policies implemented during the same time frame as the analysis.

Linear Regression to Approximate Weights: Initial weights for constructing the synthetic control are derived from fitting a linear regression model. We can use a linear regression model because this data satisfies the main assumptions. This model uses emissions data from the control states as predictors and the treated state's emissions as the dependent variable, focusing solely on the pre-intervention period. The regression coefficients obtained from this model provide an initial approximation of the weights, indicating the relative contribution of each control state to the synthetic emissions trajectory. This step is crucial in creating a synthetic control that mimics the treated state's pre-policy emissions.

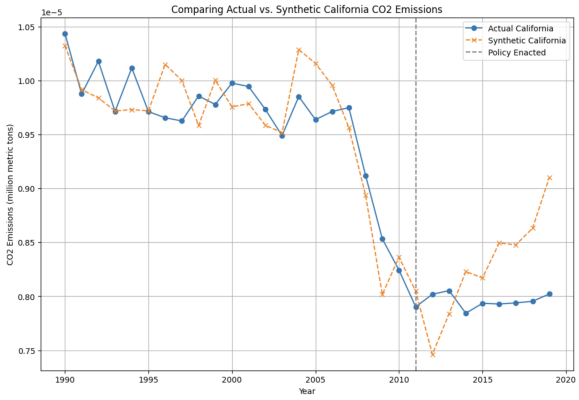
Optimization for Weights Calculation: After obtaining initial weight approximations, we make use of convex optimization techniques to refine these weights. The optimization is subject to two constraints: the weights must sum up to one, ensuring that the synthetic control is a weighted average of control states, and they must be non-negative, preventing any state from inversely affecting the synthetic control. The objective is to minimize the discrepancy between the treated state's actual pre-intervention emissions and the synthetic control's estimated emissions. This refinement process enhances the accuracy of the synthetic control, ensuring that it represents a credible counterfactual scenario. The resulting optimized weights dictate the composition of the synthetic control.

Construction of the Synthetic Control: With optimized weights, we calculate the synthetic emissions trajectory for the treated state, reflecting what emissions would have looked like in the absence of the policy intervention. This synthetic trajectory is then compared to the actual emissions observed after the policy's implementation.

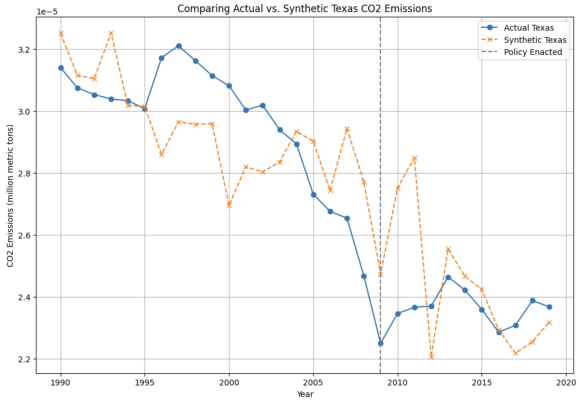
For **New York** and **California**, the analysis reveals a clear divergence between actual and synthetic emissions after the implementation of vehicle emissions policies as shown in Figure 2.21a and Figure 2.21b. This widening gap provides strong evidence that the policies have been effective in reducing



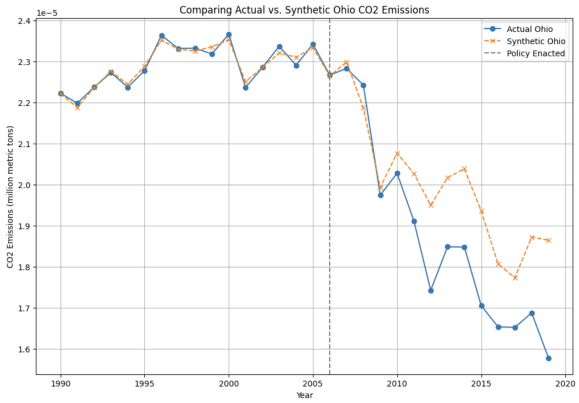
(a) *New York*



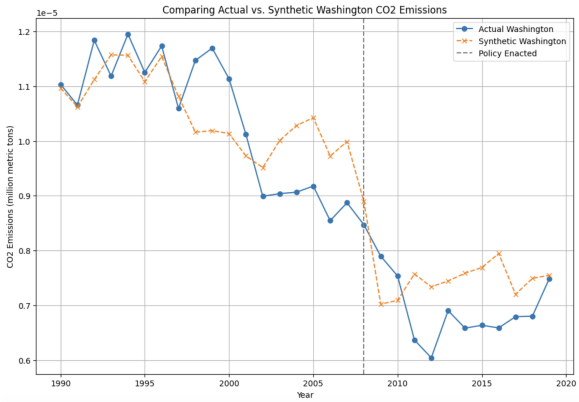
(b) *California*



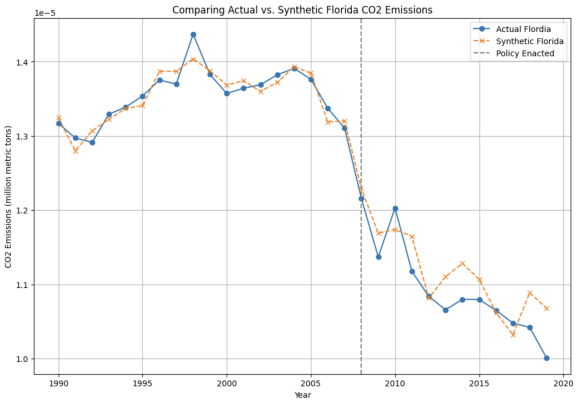
(c) *Texas*



(d) *Ohio*



(e) *Washington*



(f) *Florida*

Figure 2.21: *Synthetic Control Models for Six States*

per capita emissions. It illustrates a direct link between policy action and environmental outcomes, suggesting that these states' specific regulations have led to significant improvements in emissions standards. Conversely, for **Ohio** and **Texas**, shown in Figure 2.21c and Figure 2.21d, the differences between actual and synthetic emissions are more nuanced, hinting at either a gradual effect of the policies or the existence of competing factors that influence emissions independently of policy measures. **Washington** and **Florida**, shown in Figure 2.21e and Figure 2.21f, demonstrate that policies can reduce emissions even when state-specific circumstances may lead to an overall rise in emissions. This reinforces the idea that policy impacts must be contextualized within the broader state-specific economic and demographic landscape.

These findings illustrate the potential of environmental policies to shape emissions trajectories positively. They provide compelling evidence for the impact of policy in reducing emissions and emphasize the necessity of strategic policy design to effectively combat climate change.

2.5.4 Analysis of Results and Causal Inference

The observed trends allow us to infer causality between policy implementation and emissions reduction in the US. Our analysis underscores the importance of a tailored approach to policy-making that considers unique state attributes such as its economic dependencies.

The relationship between policy implementation and emissions reduction is highlighted through the analysis of state-level interventions. States like California and New York have set the standard with their aggressive environmental policies, such as California's Zero Emission Vehicle (ZEV) mandates and New York's diesel emissions regulations. The California Air Resources Board reports a noticeable impact of the ZEV program on reducing greenhouse gas emissions, evidencing the policy's direct contribution to the state's environmental achievements (Board, 2021). Similarly, New York's Diesel Emissions Reduction Act (DERA) of 2006 has effectively reduced emissions from diesel engines (Environmental Conservation, 2021).Thus, these states have clear reduction of emissions after their policies have been implemented.

In Texas and Ohio, however, there is a more complex relationship between policy and emissions outcomes. Texas, with its significant industrial base and status as a leading energy producer, encounters unique challenges in reconciling environmental policies with its economic interests. The state's industrial activities, heavily inclined towards oil and gas extraction, refining, and chemical manufacturing, exert a substantial influence on its emissions profile. This industrial composition necessitates a balanced approach to policy-making that accounts for both environmental sustainability and economic vitality (Environmental Quality, 2020). Thus, its policy didn't have a big impact on its emissions. Ohio's situation, characterized by its manufacturing-centric economy and reliance on coal for electricity generation, further illustrates the complexity of policy impact. The state's emissions trends are influenced by a combination of regulatory measures, economic cycles affecting industrial production, and transitions in energy sources (Ohio Environmental Protection Agency, 2021). Thus, it also didn't have a significant gap in the synthetic model and the reality after the policy was implemented.

The results from using Washington and Florida as the treated states show that the results from SCM closely align with the actual emissions or those two states. To investigate why this is the case, we

looked deeper into the vehicle emissions and energy policies in Florida and Washington. We found that Florida's growing population and high dependency on road transport present ongoing challenges to emissions reduction efforts. Despite aggressive policies aimed at reducing vehicle emissions, these challenges remain (Environmental Protection, 2020). Washington's Clean Air Rule and incentives for electric vehicles have made strides in emissions reduction, yet the state's industrial growth such as in Big Tech and consumer behaviors continue to exert pressure on emissions levels (Ecology, 2021).

Another point to consider when interpreting the results shown in the figures is the potential lag in policy impact. Studies have indicated that the effects of environmental policies on emissions may not be immediate but rather manifest over time as industries adapt and technologies evolve (Environmental Policy et al., 2019). This is also supported from our analysis in Section 2.3.

Our findings indicate a clear relationship between state-level environmental policies and a reduction in carbon emissions. However, in implementing new policies, states must consider their own economic makeup and consumer behavior about what policies are economically feasible for them and that target their specific consumption of fossil fuels. The efficacy of policies in inducing such changes underscores the importance of sustained and contextually designed legislative efforts to mitigate the adverse effects of climate change.

3

Solutions and Further Research

3.1 Recommended Solutions

Addressing climate change will require a multifaceted approach combining public education, advocacy, policy changes, business innovation, and technological advancement. Educational workshops will play a vital role in raising awareness about the climate crisis and changing public opinion, and also encourage people to talk to their local lawmakers (Blomgren, 2018). These educational efforts can empower citizens to lobby state and local governments to enact meaningful emissions reduction policies in areas like electricity generation, transportation, and building efficiency (Lutsey et al., 2008).

While tightening restrictions is important, policies should also include incentives that make clean energy and sustainability economically viable (Edenhofer et al., 2015). A transition away from oil toward cleaner fossil fuels like natural gas can potentially reduce emissions in the nearer term as a "bridge" fuel (Alvarez et al., 2012).

Governments encouraging entrepreneurship and investment in clean technology businesses can drive innovation to reduce emissions while making low-carbon solutions more affordable and accessible (Creutzig et al., 2014). Policy mechanisms like carbon pricing and public-private partnerships can accelerate growth in this sector (Carbon Pricing, 2017). Ultimately, adequately mitigating climate change will likely require coordination across multiple levels - individual, local, state, national, and global (IPCC, 2022).

3.2 Further Research

Our research mainly focused on the United States, due to available data and the fact that the US is one of the largest Greenhouse Gases producers in the world. Further research should focus on this analysis on a global scale, emissions from other sources such as Carbon Monoxide, and what types of policies are effective in reducing the effects of climate change. It is also important to research the cleaner fossil fuels and find which ones are more feasible in the short term use.

References

(N.d.). In: *Where greenhouse gases come from - U.S. Energy Information Administration (EIA)* (). URL: https://www.eia.gov/energyexplained/energy-and-the-environment/where-greenhouse-gases-come-from.php.

Alternative Fuels Data Center (2022). *State Laws and Incentives*. https://afdc.energy.gov/laws/state.

Alvarez, Ramón A et al. (2012). "Greater focus needed on methane leakage from natural gas infrastructure". In: *Proceedings of the National Academy of Sciences* 109.17, pp. 6435–6440.

Arezki, Rabah and Adnan Mazarei (Aug. 2022). *Climate Policies Are Becoming a Casualty of High Oil Prices*. English. URL: https://www.barrons.com/articles/climate-policies-high-oil-prices-energy-transition-94ce47da.

Bachelier, Louis (1900). "Théorie de la Spéculation". In: *Annales Scientifiques de l'Ecole Normale Supérieure 17*, pp. 21–88.

Berkeley Earth (June 2016). *United States Climate Data*. URL: https://berkeleyearth.org/temperature-region/united-states.

Blomgren, Anne-Marie (2018). *Climate change education*. Stockholm University.

Board, California Air Resources (2021). "California's Zero Emission Vehicle Program: Progress, Challenges, and Future Directions". In: *Environmental and Energy Study Institute*. https://www.arb.ca.gov/zev-program.

Bogmans, Christian, Andrea Pescatori, and Ervin Prifti (2023). *The Impact of Climate Policy on Oil and Gas Investment: Evidence from Firm-Level Data*. Working Paper. FEEM Working Paper No. 92.2007.

Bromberg, Michael (Apr. 2023). *OPEC's Influence on Global Oil Prices*. https://www.investopedia.com/ask/answers/060415/how-much-influence-does-opec-have-global-price-oil.asp.

Burnham, Andrew et al. (2012). "Life-cycle greenhouse gas emissions of shale gas, natural gas, coal, and petroleum". In: *Environmental science & technology* 46.2, pp. 619–627.

Carbon Pricing, High-Level Commission on (2017). *Report of the High-Level Commission on Carbon Prices*. Tech. rep. World Bank.

Contu, Davide, Ozgur Kaya, and Ilker Kaya (2021). "Attitudes towards climate change and energy sources in oil exporters". In: *Energy Strategy Reviews* 38, p. 100732. ISSN: 2211-467X. DOI: https://doi.org/10.1016/j.esr.2021.100732. URL: https://www.sciencedirect.com/science/article/pii/S2211467X21001188.

Crawford, Maxx, Holly Gray, and Abby Coppinger (July 2017). "Oil booms and busts: What causes them?" In: *EnergyHQ*. URL: https://energyhq.com/2017/07/oil-booms-and-busts-what-causes-them/.

Creutzig, Felix et al. (2014). "Catching two European birds on renewable energy: mitigating climate change and Eurozone crisis by an energy transition". In: *Renewable and Sustainable Energy Reviews* 38, pp. 1015–1028.

Deeney, Peter et al. (Jan. 2015). "Sentiment in Oil Markets". In: *International Review of Financial Analysis* 39. DOI: 10.1016/j.irfa.2015.01.005.

Ecology, Washington State Department of (2021). "Washington's Clean Air Rule and Electric Vehicle Policies: An Analysis". In: https://ecology.wa.gov/Air-Climate/Air-quality/Climate-change/Clean-Air-Rule.

Edenhofer, Ottmar et al. (2015). *Climate change 2014: mitigation of climate change: Working Group III contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Vol. 3. Cambridge University Press.

Environmental Conservation, New York State Department of (2021). "Diesel Emissions Reduction Act (DERA) of 2006: Implementation Plan and Progress". In: https://www.dec.ny.gov/dera.

Environmental Policy, Journal of and Planning (2019). "Assessing the Time Lag Effect of Environmental Policies on Carbon Emissions". In: 21.5, pp. 539–553.

Environmental Protection, Florida Department of (2020). "Challenges to Reducing Transportation-Related Emissions in Florida". In: https://floridadep.gov/air.

Environmental Quality, Texas Commission on (2020). "The Impact of Industrial Emissions on the Environment: A Case Study of Texas". In: https://www.tceq.texas.gov/airquality.

International Energy Agency (IEA) (2021). *Coal Information: Overview*. https://www.iea.org/reports/coal-information-overview.

IPCC (2022). *Climate Change 2022: Mitigation of Climate Change*. Intergovernmental Panel on Climate Change.

Kumar, Abhinandan et al. (Feb. 2022). "Impact of covid-19 on Greenhouse Gases Emissions: A critical review". In: *The Science of the total environment*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8445775/.

Laine, Marko (Aug. 2019). "Introduction to Dynamic Linear Models for Time Series Analysis". In: *Springer Geophysics*. Springer International Publishing, pp. 139–156. ISBN: 9783030217181. DOI: 10.1007/978-3-030-21718-1_4. URL: http://dx.doi.org/10.1007/978-3-030-21718-1_4.

Lindsey, Rebecca (n.d.). "Climate change: Atmospheric carbon dioxide". In: *NOAA Climate.gov* (). URL: https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide.

Lutsey, Nicholas and Daniel Sperling (2008). "America's bottom-up climate change mitigation policy". In: *Energy Policy* 36.2, pp. 673–685.

Ohio Environmental Protection Agency (2021). "Understanding Emissions Trends in Ohio: Policy and Industrial Perspectives". In: *Ohio EPA Division of Air Pollution Control*. https://epa.ohio.gov/dapc.

Qadan, Mahmoud and Hazar Nama (2018). "Investor sentiment and the price of oil". In: *Energy Economics* 69, pp. 42–58. ISSN: 0140-9883. DOI: https://doi.org/10.1016/j.eneco.2017.10.035. URL: https://www.sciencedirect.com/science/article/pii/S0140988317303766.

Taleb, Nassim Nicholas (2008). *The black swan: The impact of the highly improbable*. Harlow, England: Penguin Books. ISBN: 9780141034591.

U.S. Bank (2024). *How presidential elections affect the stock market*. https://www.usbank.com/investing/financial-perspectives/market-news/how-presidential-elections-affect-the-stock-market.html.

U.S. Census Bureau (n.d.). *Annual Estimates of the Population for the U.S. and States*. https://www.census.gov/programs-surveys/popest.html. Accessed: insert date of access here.

United Nations (n.d.). "Causes and Effects of Climate Change". In: *united nations* (). URL: https://www.un.org/en/climatechange/science/causes-effects-climate-change.

United States Environmental Protection Agency (2022). *Methodology Report: Inventory of U.S. Greenhouse Gas Emissions and Sinks by State: 1990-2021*. https://www.epa.gov/ghgemissions/methodology-report-inventory-us-greenhouse-gas-emissions-and-sinks-state-1990-2021.

UVA Miller Center (2024). *Famous Presidential Speeches*.

A

Appendix

A.1 Arithmetic Brownian Motion

A stochastic process S_t is said to follow a geometric Brownian motion (GBM) if it satisfies the following stochastic differential equation (SDE):

$$\mathrm{d}S_t = \mu S_t \,\mathrm{d}t + \sigma S_t \,\mathrm{d}W_t \,,$$

where W_t is a Wiener process/Brownian motion, μ is the percentage drift and σ is the percentage volatility. In this SDE, μ represents the overall trends of the differential equation while σ represents the degree of randomness. Note that without the stochastic part W_t, the differential equation just represents an exponential. This models the stock market well due to the compounding effect.

One of the greatest results of stochastic calculus was Itô's formula, which describes the behavior of the functional of a stochastic process. Itô's formula provides a way to calculate the differential of a function of a stochastic process, taking into account the randomness and volatility of the underlying process.

Itô's formula states that if a stochastic process X_t follows the SDE:

$$\mathrm{d}X_t = \mu(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}W_t \,,$$

and $f(t, X_t)$ is a twice differentiable function, then $f(t, X_t)$ also follows an SDE given by:

$$\mathrm{d}f(t, X_t) = \left(\frac{\partial f}{\partial t} + \mu\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 f}{\partial x^2}\right)\mathrm{d}t + \sigma\frac{\partial f}{\partial x}\,\mathrm{d}W_t \,,$$

where $\mathrm{d}x = \mathrm{d}X_t$. Applying this formula, we get

$$\mathrm{d}(\ln S_t) = (\ln S_t)'\,\mathrm{d}S_t + \frac{1}{2}(\ln S_t)''\,\mathrm{d}S_t\,\mathrm{d}S_t = \frac{\mathrm{d}S_t}{S_t} - \frac{1}{2}\frac{1}{S_t^2}\,\mathrm{d}S_t\,\mathrm{d}S_t \,.$$

Substituting in the form of S_t for geometric Brownian motion above, we get that the quadratic variation of the SDE is

$$\mathrm{d}S_t\,\mathrm{d}S_t = \sigma^2 S_t^2\,\mathrm{d}{W_t}^2 + 2\sigma S_t^2\mu\,\mathrm{d}W_t\,\mathrm{d}t + \mu^2 S_t^2\,\mathrm{d}t^2 \,.$$

The quadratic variation measures the amount of deviation a Brownian motion process has from a linear function over a given time interval. Since we are in the differential limit as $\mathrm{d}t \to 0$ and the

Brownian motion converges slower to 0 (specifically $\mathrm{d}{W_t}^2 \sim \mathrm{d}t$), we can ignore the deterministic parts and write

$$\mathrm{d}S_t \, \mathrm{d}S_t \approx \sigma^2 S_t^2 \, \mathrm{d}t \, .$$

Plugging this back into the result from Itô's formula with the differential form of GBM, we have

$$\mathrm{d}(\ln S_t) = \frac{\mu S_t \, \mathrm{d}t + \sigma S_t \, \mathrm{d}W_t}{S_t} - \frac{\sigma^2 S_t^2}{2S_t^2} \, \mathrm{d}t = \left(\mu - \frac{\sigma^2}{2}\right) \mathrm{d}t + \sigma \, \mathrm{d}W_t \, .$$

Notice that applying Itô's formula has allowed us to succesfully remove S_t dependence from the SDE. We can easily solve this SDE by integrating over both sides:

$$\ln S_t - \ln S_0 = \ln \frac{S_t}{S_0} = \left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t$$

The greatest intuition stemming from this formula is that the log of the ratio $\ln \frac{S_t}{S_0}$ is normally distributed with mean $\left(\mu - \frac{\sigma^2}{2}\right) t$ (take the expectation value of both sides and noticing that Brownian motions W_t have 0 mean) and variance σ^2 (Brownian motions have unit variance 1). Notice that the mean here is smaller than the expectation value of μ without the stochastic part, meaning that stochasticity curbs average growth.

To take advantage of this normal distribution, we can define a new variable $X_t = \ln \frac{S_t}{S_0}$ which follows the SDE

$$\mathrm{d}X_t = \left(\mu - \frac{\sigma^2}{2}\right) \mathrm{d}t + \sigma \, \mathrm{d}W_t \, ,$$

or more genrally

$$\mathrm{d}X_t = m \, \mathrm{d}t + v \, \mathrm{d}W_t \, .$$

This is in accordance with the linear assumptions we made for regressions earlier.

A.2 List of All Keywords Used for Filtering Speeches

For oil, we used 'oil', 'petroleum', 'OPEC', 'crude', 'gasoline', 'gas', 'fuel', 'energy', 'drilling', 'pipeline', 'refinery', 'barrel', 'barrels', 'platform', 'offshore', and 'onshore'.

For climate, we used 'climate', 'Paris', 'sustainable', 'green', 'carbon', 'emission', 'renewable', 'solar', 'wind', 'hydro', 'geothermal', 'nuclear', 'clean', 'pollution', 'sustainability', 'environment', 'ecology', 'conservation', 'recycle', 'recycling', and 'renewal'.

Bias would necessarily be introduced through multiple connotations of the same word.