University of California, Berkeley

Data 102: Data, Inference, and Decisions

May 10, 2025

Statistical Models for a Cleaner Grid: Parametric, Non-Parametric, and Causal Insights into Carbon Reduction

Group: 64

Contributors: Bryan Pac, Ralph Castaneda

Table of Contents

1 Data Overview

This project uses data from the U.S. Environmental Protection Agency's **eGRID2023** (Emissions & Generation Resource Integrated Database). The dataset provides environmental characteristics of electric power generation in the U.S., including *emission rates, resource mix, and grid loss values.*

The original dataset was quite large and available only in `.xlsx` format, containing multiple sheets with distinct information. Downloading this dataset into a Jupyter Notebook posed technical challenges. Each relevant sheet had to be manually extracted, saved as its own workbook, and then exported as a `.csv`. Although tedious, this process had advantages: specifically, it allowed for clean formatting of numeric values directly in Excel, which simplified the preprocessing of dozens of columns and avoided the need for custom Python parsing functions.

One important decision was our choice of granularity. The EPA dataset contains both state-level data and **eGRID subregion**-level data. While there are over 50 states, there are only 26 eGRID subregions across the country. Modeling at the eGRID level was a deliberate choice to reduce noise: state-level data introduces variability due to differences in geography, population, and policy enforcement. In contrast, eGRID subregions span multiple states and reflect grid-level organization, offering more generalizable patterns and reducing the impact of outliers.

We did not drop any features, even those with many zeros, because the models we developed, particularly non-parametric ones—are inherently designed to determine which features are important. Features with little or no signal are assigned low or zero importance, allowing the model to decide what matters. Furthermore, one of our relations contained 169 features, and we chose to retain them to create the highest-dimensional model possible, incorporating as many energy and emission parameters as available.

However, during the cleaning phase, we also noticed the dataset recorded emissions for multiple greenhouse gases (e.g., SO_2, NO_x, CH_4, Hg). To streamline our model and focus on policy-relevant outcomes, we isolated $\mathbf{CO_2}$, which accounts for approximately 80% of all greenhouse gas emissions. The rationale was simple: if policy is to be effective, it must target the most harmful and abundant pollutant. After filtering, we retained 70 features, stilll a high-dimensional set that reflects the full complexity of regional energy production.

There were no missing values in our working dataset, which allowed for seamless integration into modeling workflows. No imputation or deletion was necessary during preprocessing.

Limitations. The data are cross-sectional for 2023 and do not include longitudinal trends. Additionally, some contextual variables that might influence emissions (e.g., policy stringency, weather, or regional economic indicators) are not included.

To conduct the causal portion of our analysis, we used annual state-level emissions data

from the U.S. Energy Information Administration's (EIA) Form 923 dataset. The raw Excel dataset contained over 49,000 rows and included the following key variables: `Year`, `State`, `Producer_Type`, `Energy_Source`, and emissions for several pollutants, including our variable of interest, `CO2_Metric_Tons`.

We grouped the data by year and state and aggregated total CO_2 emissions using a pivot table to construct a clean panel dataset spanning 50 U.S. states and Washington, D.C. over time. Puerto Rico was excluded due to persistent missing values. The EIA's *Massachusetts Electricity Profile 2022* reported 9.098 million metric tons of CO_2 emissions from the electric power sector—a value that exactly matches our result, confirming that our generation-based data and preprocessing pipeline produced accurate and credible estimates[1].

Limitations. The dataset originates from the EIA-923 emissions survey, which reflects plant-level self-reported data collected by the U.S. Energy Information Administration (eia.gov/electricity/data.php). While more detailed than administrative aggregates, survey data may still be subject to reporting inconsistencies across states and years. Our analysis does not control for variation in state-level reporting practices, nor does it capture sub-annual dynamics.

.....

[1] `https://www.eia.gov/electricity/state/archive/2022/massachusetts/`

2 Research Questions

Research Question 1: What features of electricity production and resource usage are the most predictive of CO_2 emissions at the eGRID subregion level?

This question is motivated by the need to prioritize which aspects of energy infrastructure should be targeted by climate policy. By identifying which features of electricity generation (e.g., coal use, total heat input, nuclear generation) most strongly predict CO_2 emissions, decision-makers can craft more effective regulations aimed at curbing emissions in the most impactful ways.

To answer this question, we apply both parametric and non-parametric modeling approaches, including a Generalized Linear Model (GLM) and Random Forest Regression. These models are well-suited because they not only produce accurate predictions, but also provide estimates of feature importance, directly addressing the research goal.

Limitations. GLMs assume a linear relationship between predictors and the outcome, which may not hold for complex energy systems. Non-parametric models, while flexible, can be prone to overfitting and may obscure interpretability if too many features are irrelevant or collinear. Additionally, the models are limited by the absence of time-series data and contextual covariates (e.g., policy enforcement intensity, weather, or socioeconomic indicators), which may confound the true drivers of emissions.

Research Question 2: What was the causal impact of Massachusetts' 2003 Renewable Portfolio Standard (RPS) climate policy on electricity-sector CO_2 emissions?

This question is motivated by the broader goal of evaluating whether climate legislation leads to measurable emission reductions. To address this question, we applied the synthetic control method, comparing Massachusetts' post-treatment emissions to a weighted combination of similar states that did not implement comparable policies in 2003. We assess model fit using pre-treatment RMSE, evaluate robustness via placebo tests and RMSPE-ratio distributions, and estimate the magnitude of the treatment effect by comparing observed and synthetic emissions in 2022.

Limitations. Due to the small donor pool, the statistical power of the placebo tests is limited. Additionally, our analysis relies on strong assumptions—including no unobserved time-varying confounders and accurate matching of pre-treatment trends—which may not always hold in practice.

.....

3 Prior Work

A study titled "Regression analysis and driving force model building of CO_2 emissions in China" by Zhou et al. (2021) provides a comprehensive look at the factors influencing carbon emissions in China using a multivariate linear regression approach. This study identifies three major sources of emissions—energy production, fuel combustion in other industries, and industrial processes—and constructs predictive models for each using panel data spanning from 1990 to 2017. The authors forecast CO_2 emission trends and evaluate the impact of various policy and economic indicators over time.

This article is particularly relevant to our project because it also applies parametric modeling—specifically, multiple linear regression—to understand how specific features drive carbon emissions. Like the authors, we used multivariate linear models to identify key drivers of CO_2 emissions. However, our study differs in scope and structure. While Zhou et al. leveraged panel data and forecasted emissions over time, our dataset is cross-sectional, capturing emissions data for the year 2023 only. This limits our ability to make temporal forecasts but allows for higher granularity in spatial comparisons within the U.S.

Another point of departure lies in the dataset dimensionality. Our EPA eGRID dataset included over 70 features, enabling the construction of a high-dimensional model that emphasized feature selection and interpretability. Zhou et al., in contrast, focused on selecting specific economic and policy indicators with theoretical significance and applied normalization to compare relative impacts.

Finally, although our dataset does not include policy enforcement metrics or temporal variance, the two projects share a common motivation: to uncover the structural factors most responsible for CO_2 emissions and to inform data-driven policy decisions. Our work complements Zhou et al.'s study by offering a U.S.-focused, high-dimensional, and cross-sectional counterpart to their longitudinal, China-based model.

Another key influence on our work is the study by Abadie and Gardeazabal (2003), which introduced the synthetic control method to estimate the economic costs of terrorism in the Basque Country. Their approach constructs a weighted average of untreated units to serve as a counterfactual, minimizing the distance between the treated unit and its synthetic counterpart in the pre-intervention period. We adopted this framework to estimate the impact of Massachusetts' 2003 electricity policy on CO_2 emissions, using the simplified version of the method that relies solely on outcome variables and omits covariates. This aligns with the original formulation in their paper, where weights $\mathbf{W}^*$ are chosen to minimize the squared distance $\|Z_1 - Z_0\mathbf{W}\|^2$ between treated and control units prior to the intervention.

We also implement placebo tests by iteratively reassigning the treatment to control states and recalculating synthetic gaps. This allows us to benchmark Massachusetts' post-

treatment deviation against those of unaffected units and assess whether the observed effect is unusually large. While their study focused on GDP trends, we apply the same logic to CO_2 emissions, using post/pre-treatment RMSPE ratios to quantify effect significance and compute a test statistic.

.....

4 Exploratory Data Analysis

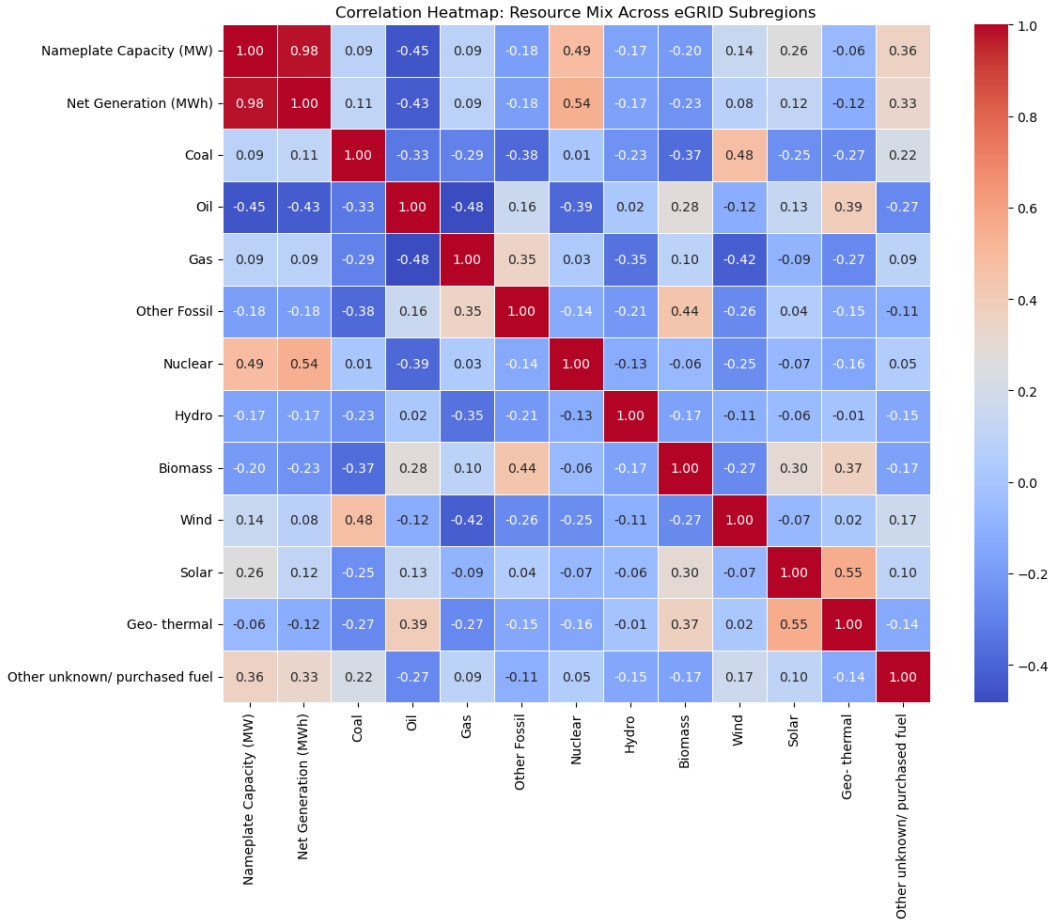


Figure 1: Correlation matrix showing pairwise relationships between variables.

This correlation matrix highlights relationships between fuel types and generation metrics across eGRID subregions. As expected, nameplate capacity and net generation are highly correlated, since capacity limits potential output. Natural gas shows a moderate positive correlation with net generation, suggesting it may be a key contributor to total output. In contrast, coal and hydro show weaker or negative correlations, possibly due to more localized use.

These insights motivate our model by pointing to gas as a potentially high-impact driver of CO_2 emissions. Although the plot is not causal, it helps surface which variables may be most relevant to predictive modeling.

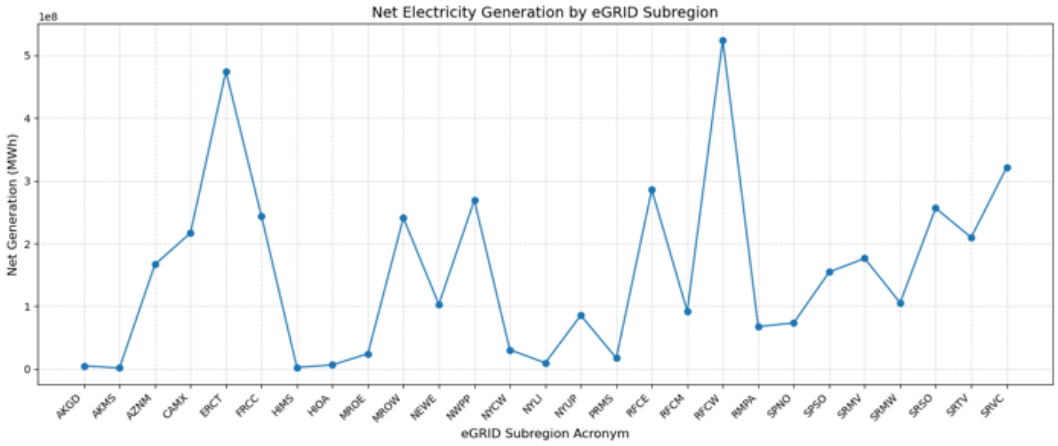


Figure 2: Net electricity generation (in MWh) across eGRID subregions

This plot visualizes how total electricity output varies by eGRID subregion, highlighting regional disparities in generation capacity. The plot shows large spikes in areas such as ERCOT (Texas) and RFC West, suggesting those areas contribute disproportionately to national totals. These differences are important because they may influence CO_2 emission levels and model outcomes, especially when fitting regression models to predict emissions or energy mix.

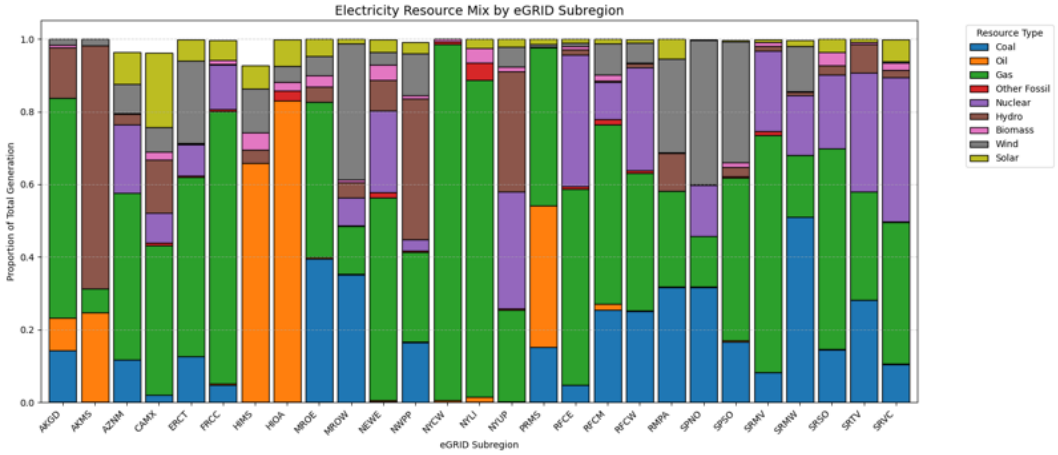


Figure 3: Stacked bar chart showing the electricity resource mix across eGRID subregions.

This stacked bar chart shows the electricity generation mix across eGRID subregions, with each bar segmented by fuel type (coal, gas, hydro, wind, etc.). Regions like ERCT, FRCC, and SRVC rely heavily on natural gas, while others like NWPP and CAMX have a more balanced mix with greater renewable use.

These patterns help motivate our modeling choices. Subregions dominated by gas may be linked to higher CO_2 emissions, while those with more renewables may emit less. While this

visualization is not causal, it highlights categorical variables (region, fuel type) and quantitative relationships (proportion of fuel mix) relevant to our research on emission drivers.

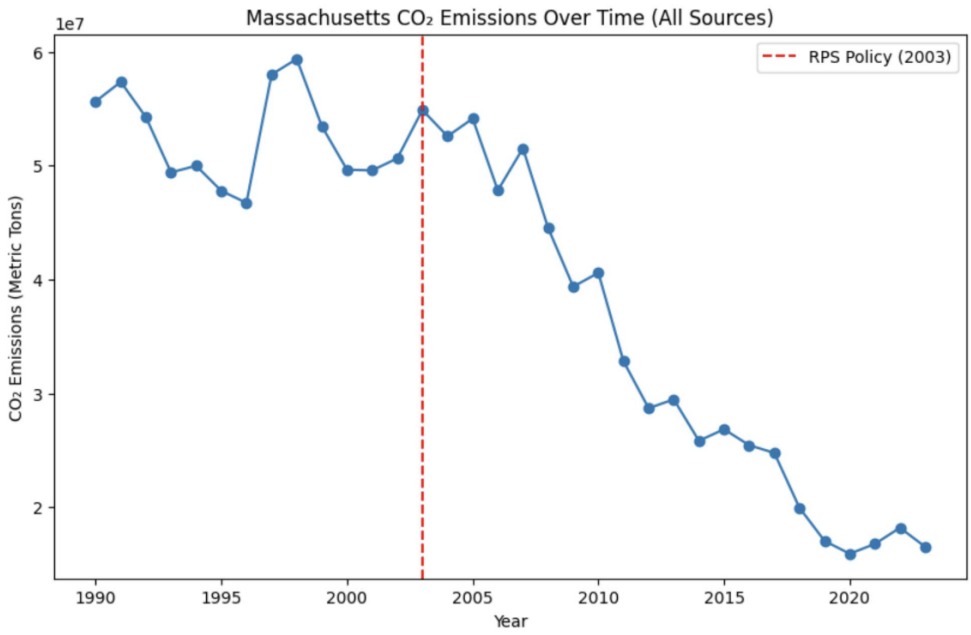


Figure 4: CO_2 emissions from electricity generation in Massachusetts (1990–2022)

Before 2003, CO_2 emissions in Massachusetts were relatively stable, fluctuating between 40 to 60 million metric tons annually. Following the 2003 introduction of the Renewable Portfolio Standard (RPS), there is a clear and sustained downward trend, with emissions falling below 20 million metric tons by 2020.

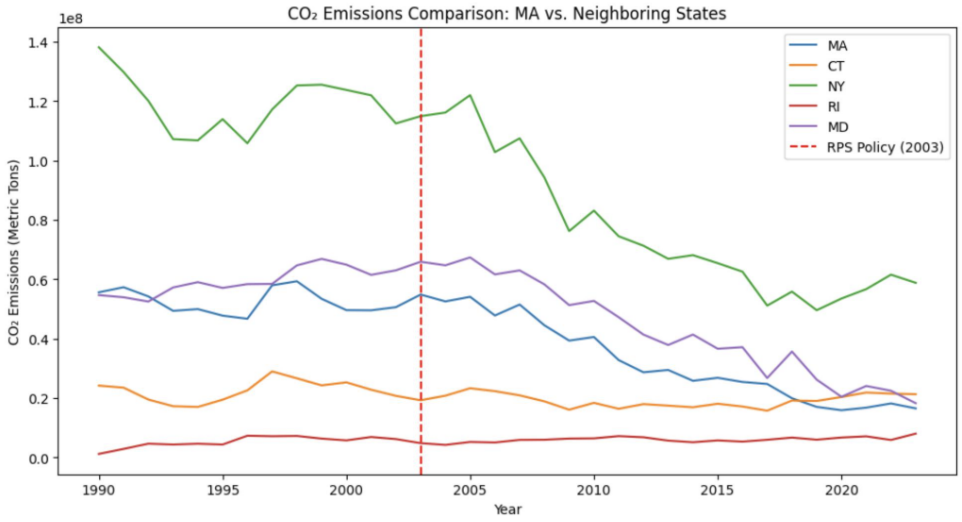


Figure 5: CO_2 emissions comparison between Massachusetts and neighboring states

This plot shows that while most states experienced gradual declines in emissions, Massachusetts's drop after 2003 appears sharper and earlier than some neighboring states.

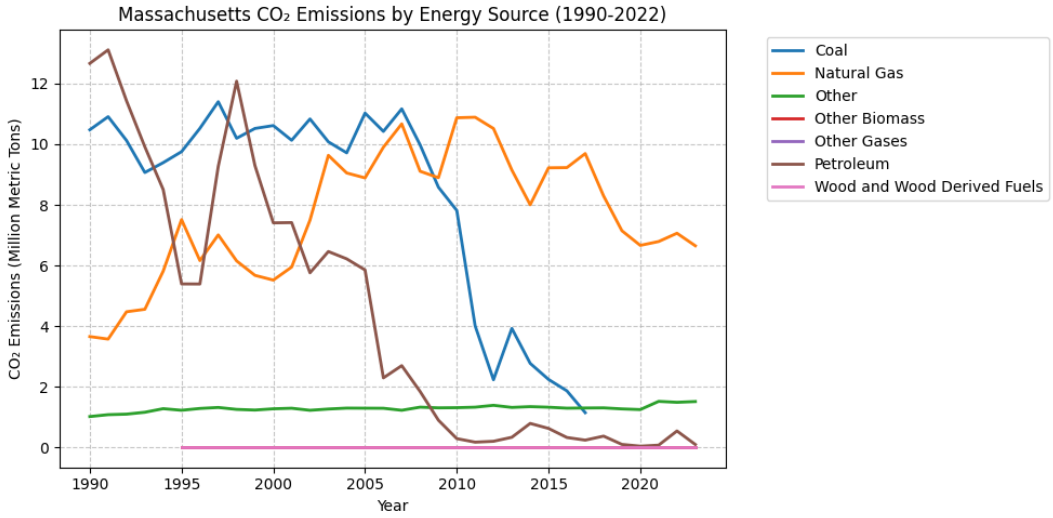


Figure 6: Massachusetts CO_2 emissions by primary energy source from 1990 to 2022.

This plot reveals that Massachusetts's emissions decline after 2003 was driven primarily by sharp drops in coal and petroleum use—two of the most carbon-intensive fuels. Their replacement by cleaner-burning natural gas aligns with the timing of the 2003 RPS policy.

.....

5 Research Question 1: GLM & Nonparametric Models

5.1 Methods

For our nonparametric model, we implemented a random forest regressor to predict eGRID subregion annual CO_2 emissions (tons). We intentionally included all 70 available numeric features—spanning electricity generation metrics, fuel types, and heat input—rather than preselecting a limited subset. This choice was driven by the concern that restrictive feature selection might exclude important sources of variation. By letting the model determine which variables mattered most, we avoided imposing assumptions that could hinder generalizability or obscure emission drivers.

We performed additional preprocessing to convert all features to numeric types and dropped non-informative identifiers like subregion names. This enabled the random forest to work effectively on the full dataset, taking advantage of its ability to handle high dimensionality, nonlinear relationships, feature interactions, and implicit variable selection, without requiring standardization or assumptions about distributions.

Model performance was evaluated using an 80/20 train-test split, with Mean Squared Error (MSE) and Mean Absolute Error (MAE) computed for each partition. To further assess generalization, we used the Out-of-Bag (OOB) R^2 score as an internal validation metric. Residual and actual vs. predicted plots were examined to check for calibration and structural issues. Finally, we leveraged the random forest's impurity-based feature importance rankings to identify variables that were the most significant contributors to CO_2 emissions, serving as a comparative benchmark for our parametric analysis.

For our parametric model, we implemented Principal Component Regression (PCR). After standardizing all features, we applied PCA and selected the top components that together explained approximately 92% of the total variance. These principal components were used as inputs to a multivariate Ordinary Least Squares (OLS) regression model.

To assess the relevance of each component, we evaluated the statistical significance of the OLS coefficients. Only PC1, PC3, and PC5 were statistically significant and positively associated with the target. Based on this result, we re-fit the OLS model using only these three PCs and evaluated its performance. Residual and actual vs. predicted plots showed that this simplified model performed comparably to the full model while yielding a more stable and interpretable structure. This re-fitting step provided an additional validation of model assumptions and helped focus our interpretation on the most reliable dimensions of variation.

Finally, to interpret the PCR model in terms of original features, we back-projected the

coefficients from PC space to the original feature space. This revealed which variables most strongly contributed to CO_2 emissions, enabling a clearer, more grounded understanding of the underlying relationships.

6 Results

6.1 Non-Parametric Model: Random Forest

```
OOB Score (R² estimate): 0.7103
Training MSE: 143,856,593,445,521
Test MSE: 79,971,402,901,299
Training MAE: 6,420,664
Test MAE: 6,312,150
R² score: 0.9547
```

Figure 7: Regression validation.

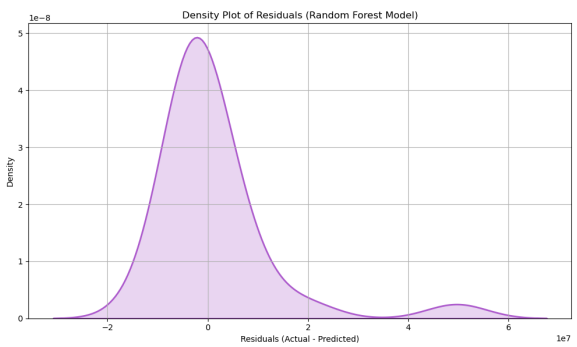
Despite the Random Forest model's tendency to overfit training data, our implementation appears to generalize well. We used an 80/20 train-test split to assess performance on unseen data. The Mean Squared Error (MSE) on the test set (≈79.97 billion) is actually lower than on the training set (≈143.86 billion), which is somewhat unexpected and suggests the model may have benefitted from noise reduction or regularization effects due to the ensemble structure. Additionally, the Mean Absolute Error (MAE) is slightly lower in the test set (6.31 million) compared to the training set (6.42 million), further supporting strong generalization.

To validate that these results are not an artifact of a favorable split, we also examined the Out-of-Bag (OOB) R^2 score, which was 0.7103, compared to a full test R^2 of 0.9547. While the OOB estimate shows a drop in performance on unseen data, the relative decline in explained variance is approximately 25.6%, which is not unexpected for Random Forests. This gap quantifies the model's loss in predictive power when moving from known to new data, but the OOB score still supports that the model retains substantial explanatory strength.

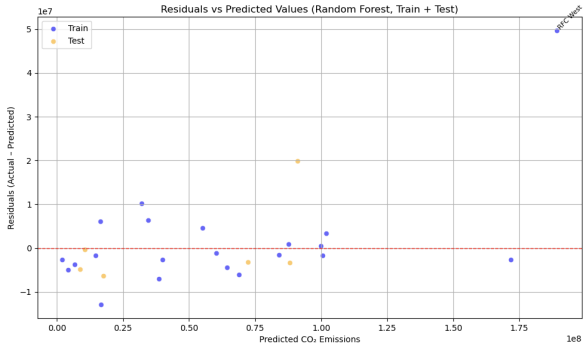
Overall, this behavior—lower test errors, modest OOB drop, and strong R^2 is typical of a well-tuned Random Forest and reflects the model's ability to capture complex patterns while maintaining generalizability.

The residual analysis provides insight into the quality and limitations of the Random Forest model's predictions. A well-behaved residual plot should show points randomly dispersed around a horizontal line at zero, with no visible structure or trend. This pattern suggests that the model is not systematically over or under-predicting and that its assumptions hold across the range of predicted values. In this case, the scatter plot does exhibit this ideal behavior, indicating that the model generally fits the data well.

However, the residual density plot reveals a slight right skew, caused by a single outlier in the predictions. This suggests that while the model performs well overall, it has limited capacity to handle extreme cases, subregions where CO_2 emissions deviate significantly from the general trend. The presence of this outlier highlights a common limitation of ensemble



(a) Density plot of residuals (Random Forest).



(b) Scatter Plot of residuals

Figure 8: Comparison of residual distribution and model fit for the Random Forest model.

methods like Random Forests: their performance can degrade when exposed to inputs that fall outside the dominant patterns learned during training.

In sum, the residual patterns affirm the model's reliability on most observations but also point to a need for caution when interpreting results for extreme or atypical subregions.

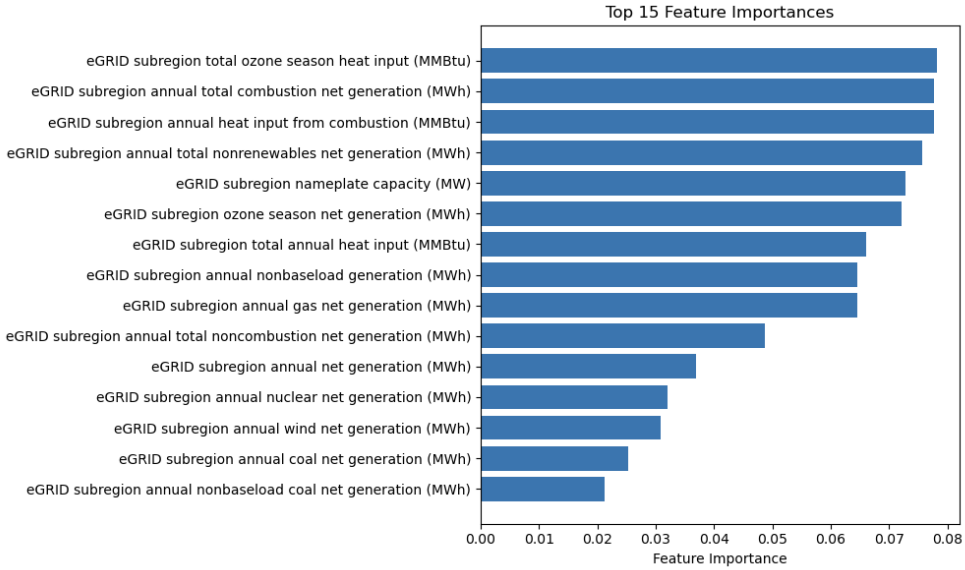


Figure 9: Top 15 features contributing to CO_2 emissions, ranked by importance in the Random Forest model.

The feature importance plot reveals which variables contributed most to the Random Forest model's ability to predict CO_2 emissions across eGRID subregions. Notably, the most influential features are related to heat input (e.g., total ozone season heat input, total combustion heat input, and heat input from combustion), along with nonrenewable generation sources, such as coal and gas generation in megawatt-hours (MWh). These features directly reflect the intensity and type of energy used, aligning with expectations that CO_2 emissions

are driven by both the quantity of electricity produced and the carbon content of the fuel sources.

Also prominent are variables like total nonrenewables net generation and nameplate capacity, which further reinforce that emission levels are strongly tied to overall generation scale and infrastructure size. Interestingly, renewable sources such as wind and solar appear much lower in importance, which is consistent with their minimal or zero direct emissions.

This distribution of feature importance confirms that the model is learning a pattern consistent with domain knowledge: emissions are largely driven by the volume of fossil fuel-based generation and associated heat input. It also suggests that limiting these variables or replacing them with low-carbon alternatives would have the most significant impact on reducing emissions.

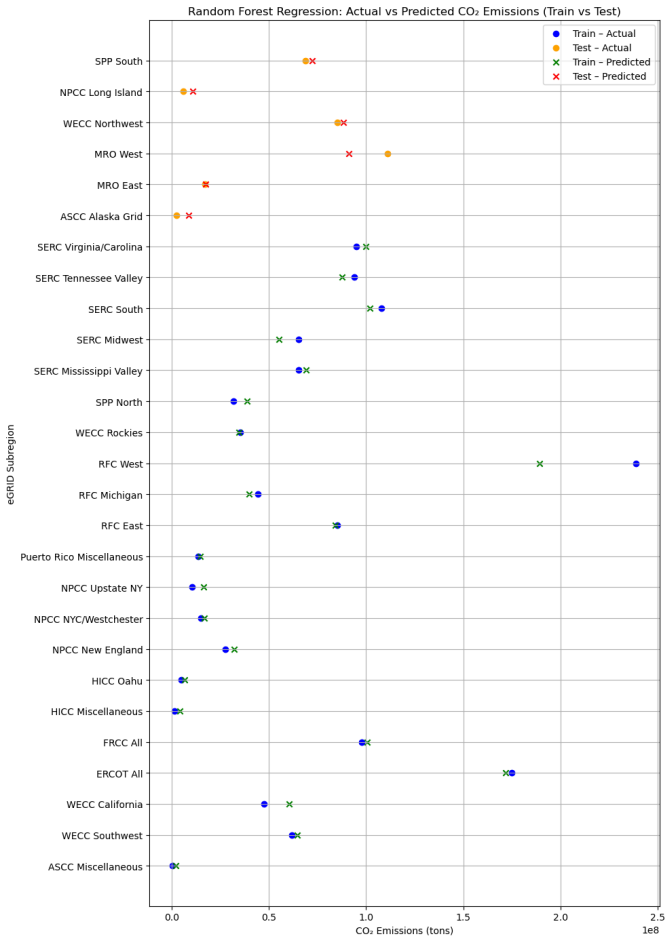


Figure 10: Actual vs. Predicted CO_2 emissions across eGRID subregions for both training and test sets. The close alignment with the diagonal line indicates strong predictive performance.

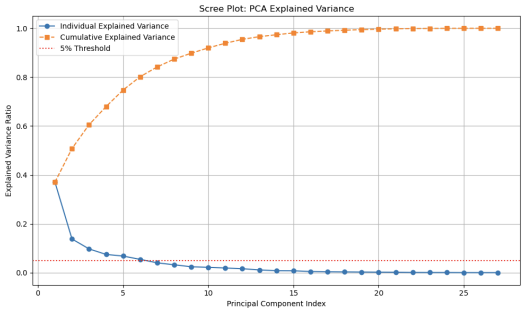
This plot compares the actual and predicted CO_2 emissions for each eGRID subregion, separated by training and test sets. Each subregion is represented by a pair of markers: one for the true CO_2 emissions and one for the model's prediction. Ideally, a good model will produce predicted values that align closely with the actual values, forming near-vertical groupings and clustering around a consistent diagonal alignment across subregions.

In this case, the predictions show strong alignment with the actual values, indicating the model captures the underlying variation in emissions well. For most subregions, especially those with moderate emission levels, the predicted values closely match the true observations, demonstrating that the model generalizes well across the range of inputs.

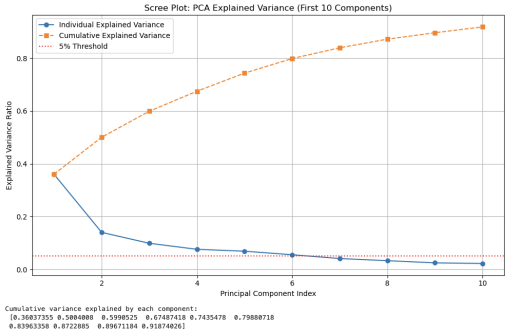
However, in a few cases—particularly subregions with extremely high or low emissions,

the predictions slightly diverge from the actual values. These deviations highlight regions where the model may be more sensitive to noise or where subregion-specific factors (not captured in the data) may contribute to CO_2 output.

6.2 Parametric Model (GLM): PCR



(a) Scree plot showing explained variance across all PCs.



(b) Scree plot showing Around 92% of total variance is explained within top 10 PCs.

Figure 11: These plots were used to determine how many components to retain for the regression.

OLS Regression Results

Dep. Variable:	eGRID subregion annual CO2 emissions (tons)	R-squared:	0.973
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	35.53
Date:	Sun, 11 May 2025	Prob (F-statistic):	1.77e-06
Time:	16:55:33	Log-Likelihood:	-367.53
No. Observations:	21	AIC:	757.1
Df Residuals:	10	BIC:	768.5
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.774e+07	3.46e+06	16.672	0.000	5e+07	6.55e+07
x1	1.163e+07	7.07e+05	16.446	0.000	1.01e+07	1.32e+07
x2	1.65e+06	1.39e+06	1.187	0.263	-1.45e+06	4.75e+06
x3	6.378e+06	1.27e+06	5.024	0.001	3.55e+06	9.21e+06
x4	-6.903e+04	1.44e+06	-0.048	0.963	-3.29e+06	3.15e+06
x5	5.969e+06	1.54e+06	3.880	0.003	2.54e+06	9.4e+06
x6	-1.665e+06	1.9e+06	-0.878	0.401	-5.89e+06	2.56e+06
x7	2.492e+05	2.01e+06	0.124	0.904	-4.23e+06	4.72e+06
x8	-1.542e+04	2.72e+06	-0.006	0.996	-6.07e+06	6.04e+06
x9	-1.417e+06	3.3e+06	-0.429	0.677	-8.78e+06	5.94e+06
x10	-3.193e+06	3.25e+06	-0.983	0.349	-1.04e+07	4.04e+06

Omnibus:	3.359	Durbin-Watson:	1.872
Prob(Omnibus):	0.187	Jarque-Bera (JB):	1.714
Skew:	-0.648	Prob(JB):	0.424
Kurtosis:	3.528	Cond. No.	7.01

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

(a) OLS Results Using Top 10 PCs

OLS Regression Results

Dep. Variable:	eGRID subregion annual CO2 emissions (tons)	R-squared:	0.965
Model:	OLS	Adj. R-squared:	0.959
Method:	Least Squares	F-statistic:	155.3
Date:	Sun, 11 May 2025	Prob (F-statistic):	1.50e-12
Time:	17:21:44	Log-Likelihood:	-370.17
No. Observations:	21	AIC:	748.3
Df Residuals:	17	BIC:	752.5
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.863e+07	2.7e+06	21.716	0.000	5.29e+07	6.43e+07
x1	1.193e+07	5.77e+05	20.681	0.000	1.07e+07	1.31e+07
x2	6.58e+06	1.08e+06	6.083	0.000	4.3e+06	8.86e+06
x3	5.988e+06	1.3e+06	4.601	0.000	3.24e+06	8.73e+06

Omnibus:	1.957	Durbin-Watson:	1.523
Prob(Omnibus):	0.376	Jarque-Bera (JB):	0.723
Skew:	0.395	Prob(JB):	0.697
Kurtosis:	3.450	Cond. No.	4.73

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

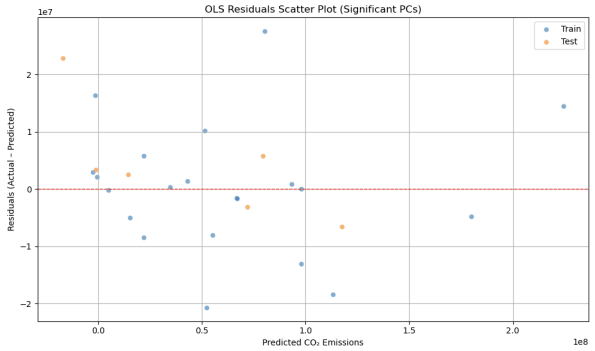
(b) OLS Results Using Significant PCs

Figure 12: Comparison of full vs. significant-only principal component models in the PCR estimator.

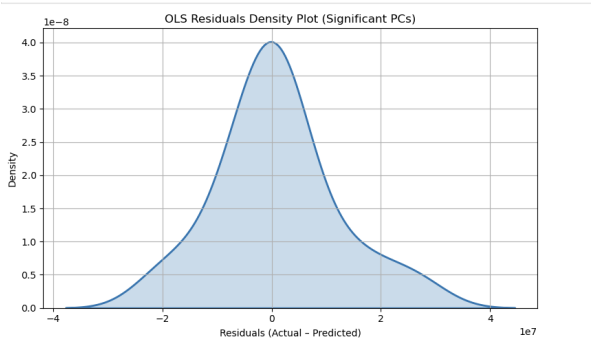
This table summarizes the Ordinary Least Squares (OLS) regression results from our Principal Component Regression (PCR) model. Initially, we selected the top 10 principal components to explain approximately 92% of the total variance—slightly below our original 95% target—to avoid overfitting. The resulting model achieved a high R^2 of 0.973, closely matching the performance of the Random Forest model. However, upon inspecting the coefficient estimates, we observed that several components had 95% confidence intervals that included zero and exhibited negative signs. This indicates that those components may contribute to reductions in predicted CO_2 emissions and are statistically insignificant.

To address this, we refit the OLS model using only the three principal components—PC1, PC3, and PC5—that were both statistically significant and positively associated with the target. This simplified model retained strong predictive performance with an R^2 of 0.965, while improving model parsimony and interpretability. The minimal drop in R^2 confirms that most of the predictive power is concentrated in these components, validating their use

for subsequent feature interpretation and dimensionality reduction.



(a) Scatter plot of residuals using significant PCs.



(b) Density plot of residuals using significant PCs.

Figure 13: Residual analysis of PCR model using only statistically significant PCs. The residuals remain randomly scattered and approximately symmetric, indicating that the simplified model still satisfies the assumptions of linear regression.

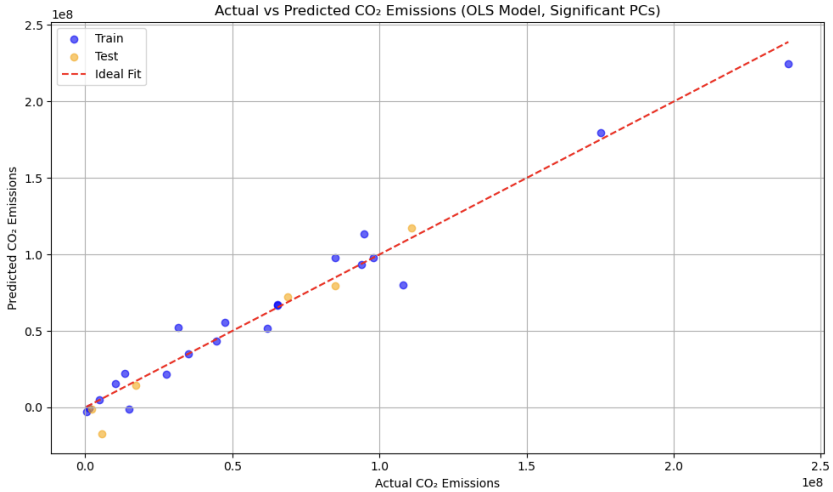


Figure 14: Actual vs. Predicted CO_2 emissions using the OLS model

The close alignment with the ideal fit line suggests that the reduced model retains strong predictive capability while improving interpretability and avoiding overfitting.

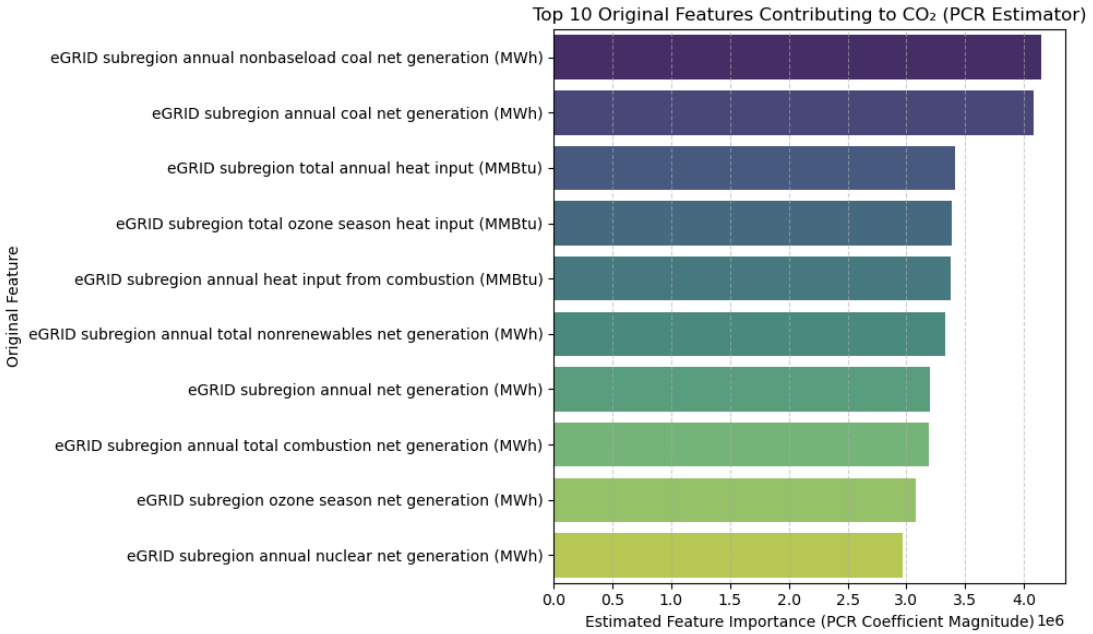


Figure 15: Top 10 original features contributing to CO_2 emissions

These estimates are generated from the back-projected coefficients of the statistically significant principal components (PC1, PC3, PC5). The most influential features are primarily tied to coal-based electricity generation and heat input, indicating that fossil fuel-related variables are dominant contributors to regional CO_2 output. These estimates are derived from a Principal Component Regression (PCR) model restricted to significant components, enhancing interpretability while preserving model performance.

7 Discussion

This project compared two modeling approaches for predicting regional CO_2 emissions: a non-parametric Random Forest model and a parametric Principal Component Regression (PCR) model using a reduced set of significant principal components (PC1, PC3, PC5). Each model brought distinct advantages and limitations in terms of fit, interpretability, and generalization.

The Random Forest model achieved an R^2 of 0.9547 on the test set, reflecting high predictive performance. The PCR model using only significant PCs achieved a comparable R^2 of 0.965, with both models producing visually similar actual vs. predicted plots and residual patterns. Importantly, the Random Forest model slightly outperformed in mean squared error (MSE), but the PCR model offered better transparency in terms of feature contribution. Overall, both models generalize well on unseen data, but Random Forest showed stronger robustness due to its ability to capture non-linear interactions.

The PCR model was implemented as a linear regression on selected principal components to avoid multicollinearity and reduce dimensionality. Originally, 10 PCs were retained to capture $\approx$92% variance, but this introduced overfitting. Reducing the number to 3 statistically significant PCs (PC1, PC3, PC5) produced a better tradeoff between interpretability and predictive performance. The AIC and BIC scores for this reduced model (AIC: 748.3, BIC: 752.5) were slightly lower than the full-PC version, indicating a more parsimonious model.

Both models fit the data well. The Random Forest model is non-parametric and robust to overfitting due to internal ensembling (e.g., bootstrap aggregation). In contrast, the initial PCR model showed signs of overfitting when using too many PCs. By restricting the model to statistically significant components, we maintained a high R^2 while controlling complexity. Residual plots from both models exhibited random scatter around zero with no clear structure, supporting the assumption of valid fits.

The PCR model enabled explicit interpretation of feature importance through the back-projection of statistically significant principal components (PC1, PC3, and PC5). The most influential features were consistently tied to coal-based electricity generation and thermal energy input, such as nonbaseload coal net generation, total annual heat input, and coal net generation. These variables are not only highly interpretable in the context of emissions but also reflect known mechanisms of CO_2 production, strengthening the validity of our PCR-derived insights.

In contrast, the Random Forest model computed feature importance through impurity-based measures, which, while effective for predictive performance, are heuristic and lack a direct statistical interpretation. Nevertheless, its top-ranked features also included heat input, combustion net generation, and nonrenewable generation, aligning closely with those

surfaced by the PCR model.

Interestingly, the PCR model emphasized more specific and narrowly defined features (e.g., coal-specific generation types), whereas Random Forest distributed importance across broader energy categories (e.g., total combustion or nonrenewables). This suggests that while both models identify the same general emission drivers, PCR provides more granularity and precision in determining which aspects of the energy system contribute most to emissions.

The statistical grounding of the PCR model—especially the ability to tie feature importance to significant PCs—enables clearer and more defensible interpretations of model behavior. This is particularly useful in policy-facing applications, where explainability is crucial for targeting emissions reduction strategies.

The Random Forest model is a black box, while performant, it lacks interpretability and can be sensitive to noise. The PCR model is interpretable but assumes linear relationships and may underperform if the true data-generating process is non-linear. Both models depend on the quality and representativeness of the training data.

More granular and updated data on regional energy production (e.g., monthly observations, fuel-type breakdowns, emissions policies) would likely enhance predictive power. Including weather data or socioeconomic factors could provide further insight into regional variation in emissions.

Both models produced residuals that were roughly symmetrically distributed and centered near zero, indicating overall low bias. However, the PCR model using significant PCs demonstrated a subtle but important improvement in handling outliers. While the Random Forest model slightly underpredicted the highest CO_2 emission region, the PCR model came closer to the actual value, suggesting it generalized better to this extreme case. This supports the robustness of the PCR model despite its simpler structure and linear assumptions. The moderate uncertainty in both models can be attributed to the small sample size and the potential presence of latent factors not captured in the current dataset.

The results from our Principal Component Regression (PCR) model show partial alignment with prior work examining CO_2 emissions in China. The previous study emphasized macroeconomic and policy-oriented driving forces such as energy intensity, industrial structure reforms, nonmetallic mineral production, and industrial technology innovation. While our model did not directly include such high-level variables, many of our most important features, such as total annual heat input, nonbaseload coal net generation, and total nonrenewables net generation, reflect the operational outcomes of those broader forces.

For example, the PCR model identified coal-based generation and nonrenewable energy sources as top contributors to CO_2 emissions, echoing the prior study's emphasis on reducing energy intensity and reforming industrial energy structures. Similarly, our finding that nuclear and renewable generation played a minor or negative role in emissions supports the

earlier work's call for an energy transition away from fossil fuels.

However, our model departs from the prior study in its data granularity and framing. Rather than focusing on national-level policy indicators or predictions to 2030, our analysis concentrated on subregional electricity generation characteristics and their direct quantitative relationship to emissions. This makes our model more tailored for policy translation within the U.S. electricity sector, but also less generalizable to broader economic sectors or international contexts.

While both models performed well, we favor the PCR model using statistically significant PCs. It offered strong predictive accuracy ($R^2 = 0.965$), competitive with the Random Forest model, but with significantly greater interpretability and parsimony. Most importantly, it generalized slightly better to an extreme outlier in the dataset, a strong indicator of robustness in a small-sample context. These strengths, along with its ability to isolate specific feature contributions via back-projection, make it a more reliable and informative model for guiding future emissions policy and research.

8 Causal Inference

8.1 Methods

8.1.1 Data and Variables

- **Treatment**: Binary indicator for Massachusetts post-2003

$$D_t = \begin{cases} 1 & \text{if } t \geq 2003 \text{ and state} = \text{MA} \\ 0 & \text{otherwise} \end{cases}$$

- **Outcome:** Y_{jt} = total mass of CO_2 emitted by electricity generators *within* state j during year t (million metric tons, MMT).

$$Y_{jt} \;=\; \sum_{m} \text{CO}_{2,\ \text{Metric Tons}_{j,m,t}}$$

where the sum is taken over all primary fuel categories m *after* we restrict the raw EIA table to rows satisfying

`Producer_Type` = "Total Electric Power Industry" and `Energy_Source` = "All Sources".

We drop rows that are not "All Sources" to avoid double-counting and divide by 10^6 to convert tons to MMT. Because the EIA dataset does not record interstate power transactions, Y_{jt} is a *generation-based*—not consumption-based—emissions measure.

8.1.2 Validation Against EIA Benchmark (2022)

To assess the accuracy of our data aggregation, we benchmarked our computed value for Massachusetts' electric sector CO_2 emissions in 2022 against the value officially reported by the U.S. Energy Information Administration (EIA).

eia Menu

Massachusetts Electricity Profile 2022

Table 1. 2022 Summary statistics (Massachusetts)

Item	Value	Rank
Primary energy source		**Natural gas**
Net summer capacity (megawatts)	**12,767**	**33**
Electric utilities	1,045	42
IPP & CHP	11,722	9
Net generation (megawatthours)	**21,026,161**	**41**
Electric utilities	529,283	43
IPP & CHP	20,496,878	19
Emissions		
Sulfur dioxide (short tons)	2,449	41
Nitrogen oxide (short tons)	8,129	41
Carbon dioxide (thousand metric tons)	9,098	40

Figure 16: Official EIA-reported electricity sector CO_2 emissions for Massachusetts in 2022

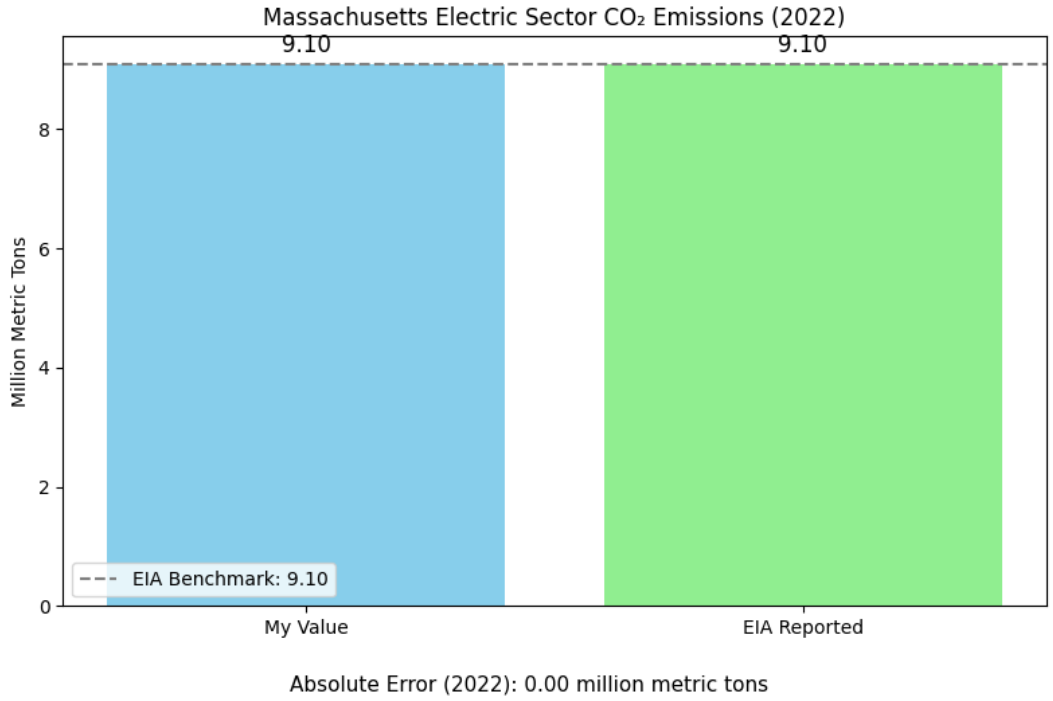


Figure 17: Comparison of Massachusetts Electric Sector CO_2 Emissions in 2022. Our computed value is shown against the official EIA benchmark (dashed line).

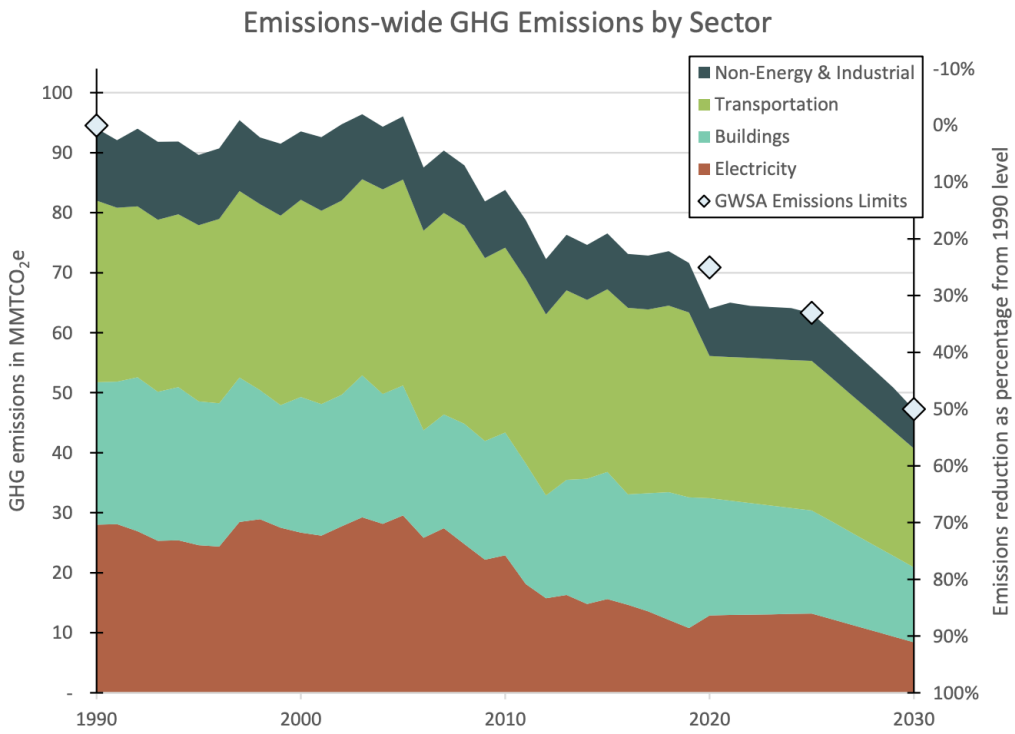


Figure 18: Massachusetts Greenhouse Gas Emissions by Sector (1990–2030)

Figure 18 illustrates the broader policy context for our analysis. The electricity sector (shown in red) experienced reductions in emissions beginning in the early 2000s—consistent with the timing of Massachusetts' 2003 Renewable Portfolio Standard (RPS).

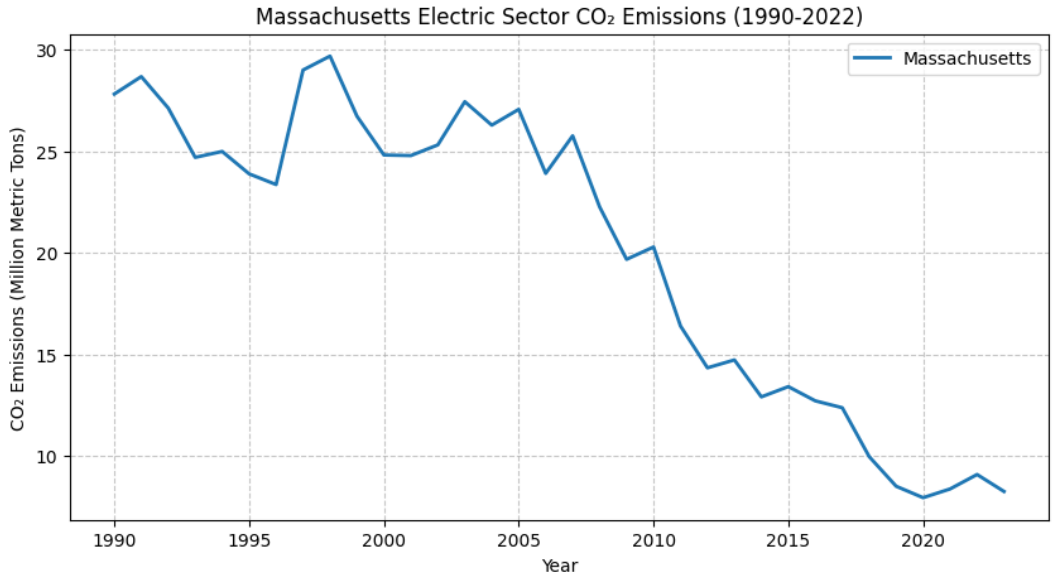


Figure 19: Massachusetts Electric Sector CO_2 Emissions from 1990 to 2022.

Our plot closely mirrors the official EIA trend, confirming a substantial decline in emissions following 2003. This alignment validates our processed dataset against real-world reported values. While our study isolates the electricity sector only, the state's overall GHG trajectory—including buildings, transportation, and non-energy sources— helps validate that observed changes are not isolated anomalies but part of a broader decarbonization effort.

8.1.3 Causal Assumptions

To interpret the estimated gap between real and synthetic Massachusetts emissions, we rely on the following assumptions:

1. **Stable Unit Treatment Value Assumption (SUTVA):**

   There is no interference between units, and each unit has only one version of the treatment. For our setting:

   $$Y_{jt}(D_j) \quad \text{depends only on} \quad D_j, \quad \text{not on } D_k \text{ for any } k \neq j,$$

   where $Y_{jt}(D_j)$ is the potential outcome for state j in year t under treatment status $D_j \in \{0, 1\}$.

2. **No Anticipation:**

   Units do not alter their behavior in anticipation of treatment:

   $$Y_{jt}(0) = Y_{jt}(1) \quad \text{for all } t < 2003$$

That is, Massachusetts emissions before 2003 follow the untreated potential outcome trajectory.

3. **Synthetic Control Provides an Unbiased Counterfactual:**

   The synthetic control unit constructed from donor states closely approximates Massachusetts' untreated potential outcomes:

$$Y_{1t}(0) \approx \sum_{j=2}^{J+1} w_j^* Y_{jt}(0), \quad \text{for all } t < 2003$$

   Thus, for $t \geq 2003$, the post-treatment gap:

$$\text{Effect}_t = Y_{1t}(1) - \sum_{j=2}^{J+1} w_j^* Y_{jt}(0)$$

   can be interpreted as the causal effect of the policy, assuming the pre-treatment fit is strong.

4. **No Time-Varying Unobserved Confounders:**

   All relevant confounding factors that influence CO_2 emissions evolve similarly across treated and donor units, and are captured during the pre-treatment period. That is, any unobserved variable U_t satisfies:

$$\mathbb{E}[U_t^{\text{MA}} - \sum_j w_j^* U_t^{(j)}] \approx 0 \quad \text{for all } t$$

5. **Linearity and Convexity of Weights:**

   The counterfactual outcome is assumed to lie within the convex hull of donor outcomes:

$$\sum_j w_j = 1, \quad w_j \geq 0$$

   This ensures interpretability and stability of the synthetic control estimator.

8.1.4 Data Smoothing

We applied 3-year centered rolling average to address volatility:

$$\tilde{Y}_{jt} = \frac{1}{3}\left(Y_{j,t-1} + Y_{jt} + Y_{j,t+1}\right)$$

8.1.5 Synthetic Control Construction

To estimate the counterfactual trajectory of Massachusetts' CO_2 emissions in the absence of its 2003 climate policy, we apply the synthetic control method (Abadie et al., 2010). The core idea is to construct a weighted combination of donor states whose pre-treatment emissions closely track those of Massachusetts. This synthetic unit is then used to predict post-treatment outcomes in the absence of the intervention.

8.1.6 Mathematical Formulation

Let:

- Y_{1t} be Massachusetts' CO_2 emissions in year t
- Y_{jt} for $j = 2, \ldots, J+1$ be emissions from donor states
- $\mathbf{Z}_1 \in \mathbb{R}^{T_0 \times 1}$ be the pre-treatment emissions vector for MA
- $\mathbf{Z}_0 \in \mathbb{R}^{T_0 \times J}$ be the matrix of donor emissions
- $\mathbf{w} \in \mathbb{R}^{J \times 1}$ be the vector of weights

The optimal weights solve:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} \left(\mathbf{Z}_1 - \mathbf{Z}_0\mathbf{w}\right)' \left(\mathbf{Z}_1 - \mathbf{Z}_0\mathbf{w}\right), \quad \text{where} \quad \mathcal{W} = \left\{ \mathbf{w} \in \mathbb{R}^J : w_j \geq 0,\ \sum_j w_j = 1 \right\}$$

To prevent scale differences from distorting the solution, we normalize all columns:

$$Z_{jt}^{\text{norm}} = \frac{Y_{jt} - \bar{Y}_j}{\sigma_j}$$

8.1.7 Donor Selection Procedure

We began with a donor pool of 50 states and D.C., then filtered the top five based on Pearson correlation with Massachusetts' pre-treatment trajectory (1990–2002). These were:

```
NY, CT, NH, MT, VT
```

We then applied a greedy forward selection algorithm:

1. Start with the top 2 states by correlation.

2. Iteratively add one new donor at a time and solve the synthetic control optimization problem.

3. Retain the configuration yielding the lowest pre-treatment RMSE.

4. Stop when RMSE does not improve.

The final optimal donor set was:

$$\text{Donors} = \{\text{NY, CT, NH, MT}\} \quad \text{with pre-treatment RMSE} = 0.6499$$

The optimized weights were:

$$\text{NY} : 0.546, \quad \text{CT} : 0.268, \quad \text{NH} : 0.186, \quad \text{MT} :\approx 0$$

8.1.8 Visual Assessment of Fit

Figure 20 shows the alignment between the real and synthetic Massachusetts emissions trajectories prior to 2003. This supports the assumption that the synthetic control provides a valid counterfactual in the post-policy period.

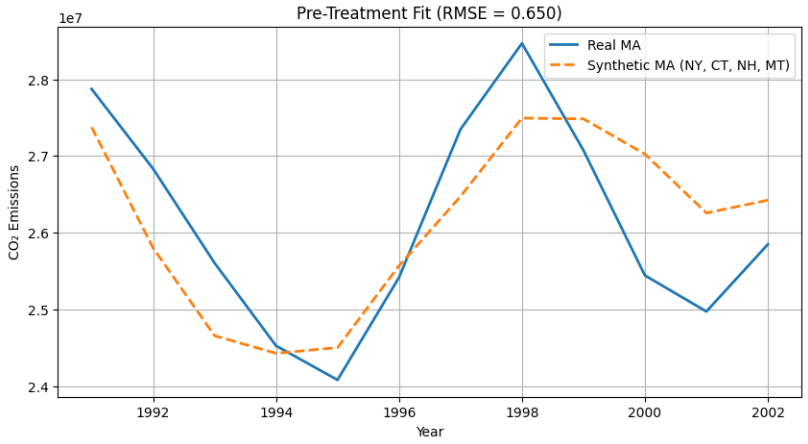


Figure 20: Pre-Treatment Fit: Real vs Synthetic Massachusetts CO_2 Emissions

Figure 21 compares the real and synthetic emissions trajectories from 1990 through 2022.

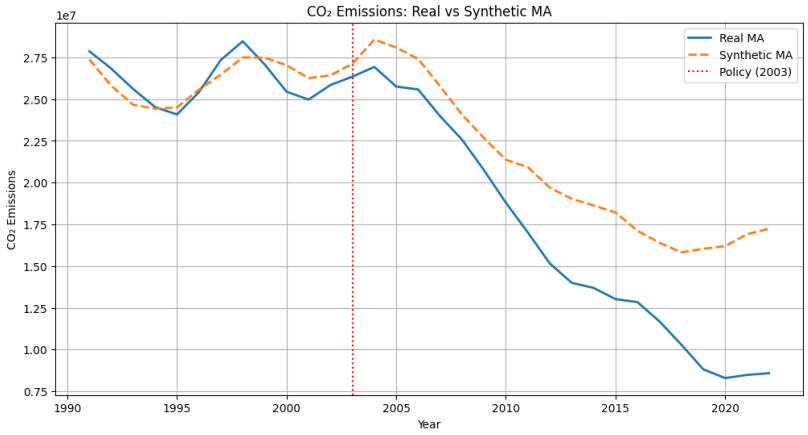


Figure 21: Full Timeline: Real vs Synthetic Massachusetts CO_2 Emissions (1990–2022)

To visually assess the quality of our pre-treatment match between Massachusetts and its synthetic counterpart, we also normalize both time series to a 0–1 scale and plot them over the pre-policy period.

The normalization transformation is given by:

$$\text{normalize}(x) = \frac{x - \min(x)}{\max(x) - \min(x)},$$

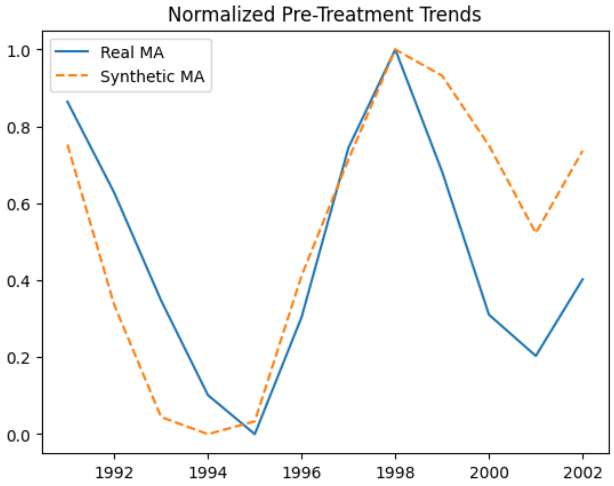


Figure 22: Normalized Pre-Treatment Trends: Real vs Synthetic Massachusetts CO_2 Emissions.

As shown in Figure 22, the synthetic trajectory replicates the turning points and volatility of the actual Massachusetts emissions trend across the entire pre-treatment window.

8.1.9 Placebo Tests

To assess the statistical significance and robustness of our estimated treatment effect for Massachusetts, we conduct *placebo tests* (Abadie et al., 2010). These tests check whether similarly large post-treatment gaps would appear for other states *not* subject to the 2003 policy intervention.

For each state $s \in \{\text{NY}, \text{CT}, \text{NH}, \text{MT}\}$, we re-run the synthetic control procedure treating s as if it were the treated unit and using the other donor states to construct its synthetic version. We compare actual emissions to this synthetic counterfactual in the post-treatment period to compute the placebo gap:

$$\text{Gap}_{s,t} = \hat{Y}_{s,t}^{\text{synthetic}} - Y_{s,t}^{\text{real}}, \quad \text{for } t \geq 2003$$

The synthetic control is built using the same optimization problem:

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \geq 0,\, \sum w_j = 1} \left\| \frac{Y_s^{\text{pre}} - \bar{Y}_s}{\sigma_s} - \sum_j w_j \cdot \frac{Y_j^{\text{pre}} - \bar{Y}_j}{\sigma_j} \right\|^2$$

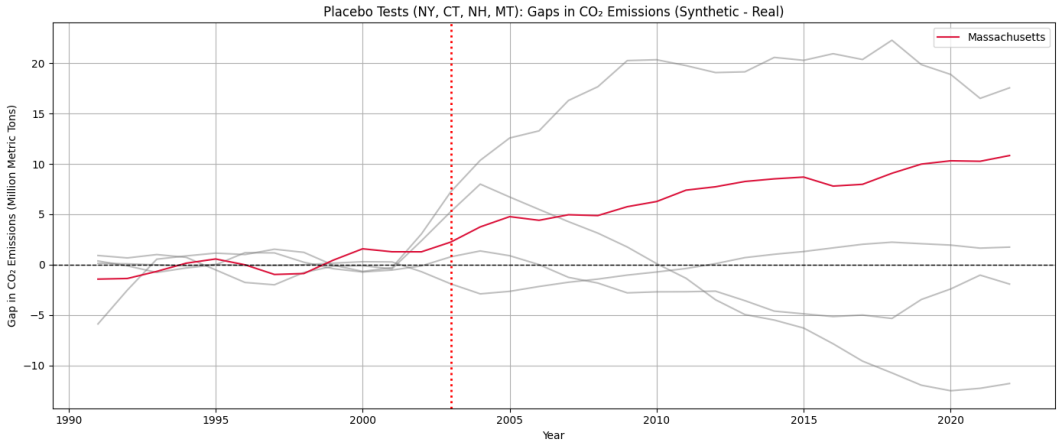


Figure 23: Placebo Tests: Post-treatment gaps in real vs. synthetic CO_2 emissions for Massachusetts and control states.

As shown in Figure 23, the gap between real and synthetic CO_2 emissions for Massachusetts increases notably after the policy implementation in 2003, suggesting a substantial treatment effect due to lower real emissions. In contrast, the placebo gaps for the donor states (gray lines) remain relatively flat or display idiosyncratic movements. Notably, New York exhibits an upward trend post-2003, while New Hampshire shows a downward trajectory around the same period. These divergent placebo paths highlight that the magnitude and direction of Massachusetts' treatment effect is distinct and not mirrored in the control units other than New York.

8.2 Results

8.2.1 Post-Treatment Effects and Statistical Inference

To quantify the magnitude and statistical significance of the observed treatment effect, we compute the ratio of post-treatment to pre-treatment Root Mean Squared Prediction Error (RMSPE) for each unit:

$$T_j = \frac{\text{RMSPE}_j^{\text{post}}}{\text{RMSPE}_j^{\text{pre}}} = \frac{\sqrt{\frac{1}{T_{\text{post}}} \sum_{t \geq 2003} \left(\hat{Y}_{jt} - Y_{jt}\right)^2}}{\sqrt{\frac{1}{T_{\text{pre}}} \sum_{t < 2003} \left(\hat{Y}_{jt} - Y_{jt}\right)^2}}$$

Here, $\hat{Y}_{jt}$ denotes the synthetic control prediction for unit j, and Y_{jt} is the observed value. For Massachusetts, we found a test statistic of:

$$T_{\text{MA}} = 7.46$$

We conduct a permutation-style inference by comparing this value to placebo states (NY, CT, NH, MT), each of which was treated as if it were the treated unit.

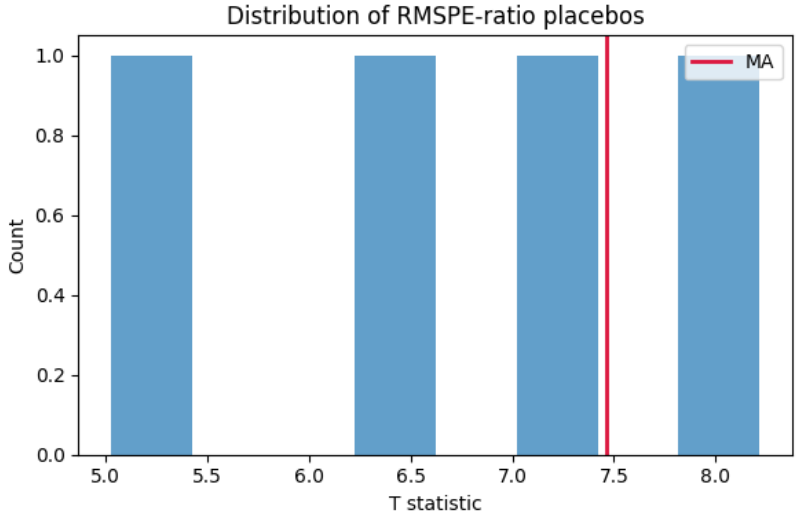


Figure 24: Distribution of RMSPE-ratio for placebo states.

We compute the pseudo-p-value using the formula:

$$\hat{p} = \frac{1 + \sum_{j \neq \text{MA}} \mathbb{1}\left[T_j \geq T_{\text{MA}}\right]}{1 + |\text{Donor Pool}|}$$

Given that the donor pool consists of 4 states (NY, CT, NH, MT) and that only one of

them produced a test statistic at least as large as Massachusetts ($T_{\text{MA}} = 7.46$), we have:

$$\hat{p} = \frac{1+1}{1+4} = \frac{2}{5} = 0.40$$

This implies that 25% of the donor states had a test statistic as extreme or more extreme than Massachusetts. The addition of 1 in the numerator and denominator ensures a conservative estimate, especially with a small number of donor units.

8.2.2 Magnitude of the Estimated Effect

By 2022, real emissions in Massachusetts were roughly 9.1 million metric tons, while the synthetic counterfactual predicts emissions near 19.41 million metric tons, suggesting a reduction of approximately:

$$\text{Effect}_{2022} \approx 10.31 \text{ million metric tons of } CO_2$$

This represents a policy-relevant drop in generation-based emissions attributable to the 2003 intervention. Our estimated 10.31 MMT reduction in 2022 emissions comes with a high p-value of 0.40, reflecting low statistical significance likely due to a small donor pool and limited placebo comparisons. Violations of key assumptions (e.g., pre-treatment fit, unobserved confounders) introduce uncertainty. According to the official EIA data, Massachusetts reduced its electricity sector CO_2 emissions by an estimated 18.7 million metric tons between 1990 and 2022. Our synthetic control estimate for the 2022 treatment effect was approximately 10.31 MMT. While our estimate is smaller, it is of the same order of magnitude and captures over half of the officially reported decline. This suggests our causal estimate is directionally consistent with the broader decarbonization trend, though it may be attenuated by methodological limitations, unobserved policy effects, or post-treatment dynamics not fully captured by our model.

8.3 Discussion

One key limitation of our synthetic control approach lies in the small size of our donor pool—only four states (NY, CT, NH, MT) were selected. We chose this subset by minimizing the pre-treatment Root Mean Squared Error (RMSE) between Massachusetts and potential donor units. While this data-driven method improved fit, it risks overfitting to noise (even after smoothing) and excludes states that might better serve as controls when policy context or energy profiles are considered. Additionally, we used a simplified form of the synthetic control algorithm that omitted time-varying covariates, which could have better accounted for unobserved confounders.

To better estimate the policy's effect, we would benefit from incorporating richer time-series covariates such as fuel prices, electricity demand, GDP per capita, and local weather conditions. These could improve the validity of the parallel trends assumption and allow for a more accurate model of untreated potential outcomes. Access to plant-level emissions or data on utility-level generation portfolios would also allow for more granular causal identification.

While we observe an 10.31 million metric ton reduction in CO_2 emissions by 2022 in Massachusetts relative to its synthetic control, the pseudo-p-value of 0.40 suggests this result is not statistically significant. The high p-value, combined with a marginal pre-treatment fit, indicates that the result could arise from chance, and we therefore cannot confidently attribute the reduction solely to the 2003 RPS policy. Other plausible explanations include market-driven shifts from coal to natural gas, technological improvements, or parallel federal policies. In particular, the Electric Industry Restructuring Act of 1997 introduced competitive retail electricity markets, utility divestiture, and municipal aggregation programs that may have influenced the state's emissions trajectory. These changes, along with other contemporaneous policies such as mandatory energy efficiency programs, complicate attribution to the RPS alone. Furthermore, policy effects often take time to manifest, making it difficult to precisely isolate the impact of a single intervention from overlapping long-term reforms.

Compared to Abadie and Gardeazabal (2003), our analysis uses a simpler model with fewer controls and a smaller donor pool. Their study incorporated covariates and achieved a strong pre-treatment fit, leading to more credible causal claims. In contrast, our choice to use outcome-only matching and to optimize donor states solely via RMSE minimization introduces more uncertainty. This difference shows how algorithmic decisions and covariate inclusion can substantially affect inference quality and the interpretability of causal estimates.

9 Conclusions

Our analysis investigated the causal effect of Massachusetts' 2003 Renewable Portfolio Standard (RPS) policy on electric sector CO_2 emissions. Using a synthetic control framework, we estimated that the policy was associated with a reduction of approximately 10.31 million metric tons of CO_2 by 2022. However, this effect was not statistically significant (pseudo-$p = 0.40$), and the confidence in a causal interpretation is limited. Our dataset excluded contextual covariates, all of which could influence emissions. A critical gap in our analysis was the lack of granular insight into how overlapping policies such as the *Electric Industry Restructuring Act of 1997* and subsequent market reforms may have confounded the observed CO_2 reductions. To untangle these effects, we would ask a domain expert: **"Which regulatory or structural shifts around 2003 (e.g., wholesale market liberalization,**

natural gas pipeline expansions, or utility incentive reforms) could have driven emissions changes *independently* of the Renewable Portfolio Standard (RPS)?" Their response would shape our approach by (1) distinguishing the RPS's marginal impact from concurrent reforms (strengthening causal inference via synthetic controls), (2) refining the donor pool (e.g., excluding states without analogous market changes to avoid bias), and (3) contextualizing the RPS's role if exogenous factors like gas prices dominated early reductions.

Our conclusions were highly sensitive to the choice of donor states and the exclusion of covariates in the synthetic control algorithm. For instance, changing the donor pool—such as removing a state with heavy optimal weights like New York—shifted the estimated treatment effect from 19.41 to 21.15 MMT, and increased the RMSPE-ratio test statistic for Massachusetts to $T = 8.27$, yielding a permutation p-value of approximately 0.25. This highlights that minor donor pool changes can affect both effect magnitude and statistical inference, with treatment effect shifts on the order of ± 2.5 MMT depending on the donor pool. We also used a simplified synthetic control that matched on pre-treatment outcomes alone. Incorporating covariates could have improved pre-treatment fit and causal validity. A user interpreting our result should recognize that our method prioritizes transparency and interpretability at the cost of reduced statistical power.

Our findings are specific to Massachusetts and may not generalize to other states with different energy mixes, regulatory histories, or policy contexts. Nonetheless, our methodological approach—using synthetic control methods and benchmarking against official emissions data—can be applied to other case studies assessing energy policy impacts.

Now, regarding the prediction research question, the EPA dataset lacked several potentially important drivers of CO_2 emissions, including policy implementation, industrial activity, and time-varying trends. These limitations may have constrained the models' ability to fully capture causal mechanisms. Additionally, without domain-specific insight into the regulatory environment, such as how state-level mandates or energy regulations affect emissions, our interpretations remained purely data-driven. Consulting an energy policy expert could have clarified whether observed patterns reflected intentional interventions or coincidental market behavior.

The conclusions of our analysis appear reasonably robust to modeling choices. In the Random Forest, controlling the number of features per tree reduced the risk of dominant variables skewing predictions. For PCR, we validated our component selection by evaluating statistical significance and re-fitting the model using only PCs that were both significant and positively correlated with CO_2 output. This step confirmed that the final model preserved predictive accuracy while improving interpretability. That said, a more complex modeling framework or different selection thresholds could lead to slightly different conclusions.

Our results are most generalizable to U.S. electricity subregions with similar data availability and generation structures. The features identified—primarily related to coal usage and heat input—likely reflect underlying physical drivers common to centralized, fossil-based grids. However, regions with different energy mixes (e.g., hydro-dominant) or in countries with less granular emissions tracking may require different approaches.

Future studies could enhance generalizability and causal insight by integrating longitudinal data, emissions policy indicators, and industrial production metrics. A natural extension would be to analyze how emissions respond to regulatory changes over time or to model emissions at a finer sectoral level.

Given the strong influence of coal and inefficient heat sources identified in both models, we recommend targeted interventions such as coal taxes, heat input efficiency standards, or renewable energy subsidies. These could reduce CO_2 emissions without requiring a wholesale shift in generation infrastructure. However, such policies may disproportionately affect coal-dependent regions and workforces. Ethical implementation should include support mechanisms to ensure a just transition, such as retraining programs and regional reinvestment. While our findings support decisive emissions reduction, they also underscore the importance of aligning environmental goals with social equity.

10 Citations

Bair, Eric, Trevor Hastie, Debashis Paul, and Robert Tibshirani. *Prediction by Supervised Principal Components.* Journal of the American Statistical Association, vol. 101, no. 473, 2006, pp. 119–137. https://doi.org/10.1198/016214505000000628.

Song, F., Guo, Z., and Mei, D. *Feature Selection Using Principal Component Analysis.* 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization, Yichang, China, 2010, pp. 27–30. https://doi.org/10.1109/ICSEM.2010.14.

Zhou, Jian, Luo, Wenbo, Liu, Haoran, Song, Xiaoming, and Jin, Shiyun. *Regression analysis and driving force model building of CO_2 emissions in China.* Scientific Reports, vol. 11, no. 1, 2021, pp. 6727. Nature Publishing Group. https://doi.org/10.1038/s41598-021-86183-5

Abadie, Alberto and Gardeazabal, Javier. *The Economic Costs of Conflict: A Case Study of the Basque Country.* American Economic Review, vol. 93, no. 1, 2003, pp. 113–132. https://doi.org/10.1257/000282803321455188

U.S. Energy Information Administration. *Detailed State-Level Emissions Data (EIA-*

923), available at https://www.eia.gov/electricity/data.php.

U.S. Energy Information Administration. *Massachusetts Electricity Profile 2022.* Available at: https://www.eia.gov/electricity/state/archive/2022/massachusetts/.

Massachusetts Executive Office of Energy and Environmental Affairs. *Clean Energy and Climate Plan for 2025 and 2030.* Available at: https://www.mass.gov/doc/clean-energy-and-climate-plan-for-2025-and-2030/download.

A Appendix