Machine Learning Approaches to Modeling Carbon Emissions Based on Political Affiliation in the U.S.

Ayal M. Yakobe Thomas Y. Chen

May 11, 2025

1 Introduction

Climate change has long been a topic of debate in the United States, where policy decisions are suspected to correlate with—among other environmental factors—carbon emissions. [1] Over the past several decades, the political landscape has evolved, with Democrats championing climate conscious policies while Republicans remain focused on economic growth, often with drastically less emphasis on environmental concerns. [2]

This study examines the potential relationship between political party affiliation and carbon emissions in the United States. Utilizing a combination of election results, state-level emissions data, political affiliation metrics, and classical machine learning techniques, our analysis does not reveal a clear correlation between political affiliation and carbon emissions across states.

2 Methodology

2.1 Data Collection & Preparation

The data for this study was sourced from several reputable databases. Federal election data (1976 - 2020) was obtained from the MIT Election Lab, providing insight into political associations across states. State-level carbon dioxide emissions data (measured in million metric tons of carbon) from the U.S. Energy Information Administration (1970-2022) was used to analyze carbon footprints. [3] Both population data (1970-2020) and geographic size data for each state were sourced from the U.S. Census Bureau to support a per capita and per square mile emissions analysis. [4, 5]

Originally, our analysis included legislative data from LegiScan, covering the years 2004 to 2020, and we were attempting to scan state-level bills using sentiment analysis techniques. The goal was to assign a score to each bill based on how environmentally conscious its language and intent appeared to be, and

then aggregate these scores annually for each state. This would have enabled an assessment of how legislative trends reflected environmental priorities over time. [6] However, due to the extent of missing data - which would have required excessive imputation or the removal of a substantial number of rows - we ultimately abandoned this effort.

All datasets underwent mean imputation to handle missing values. To account for gaps arising from four-year election cycles, Python's *pandas* library was used to apply linear interpolation for our feature engineered political columns, estimating missing values separately for each state over time.

Finally, data was Z-score normalized across columns excluding year and state in linear / logistic regression and clustering algorithms, where scaling is important to ensure comparability across features. Normalization was not applied for random forest as tree-based models are insensitive to feature scaling.

2.2 Variables

Several variables were feature-engineered and used in our analysis.

The first is our *political affiliation score*, which synthesizes information from multiple elections—presidential, House, and Senate races—to represent each state's political landscape over time. This score incorporates factors such as the longest winning streak for a party, average vote percentages, and the most recent election outcomes, and is assigned to a specific year. The results are then normalized to a scale from -1 to 1, with Democrats arbitrarily assigned positive values and Republicans negative, reflecting relative partisan leaning.

To normalize for differences in state size and population, we engineered two more variables: *emissions per capita*, defined as $\frac{\text{emissions}}{\text{population}}$, and *population density*, defined as $\frac{\text{population}}{\text{size}}$. These transformations help control for scale effects, making emissions and demographic comparisons more meaningful across states and improving model interpretability.

Lastly, we engineered the feature *emissions change* to capture the annual rate of carbon emissions change per state. It was computed by grouping data by state and applying a first-order difference to the *emissions* variable, highlighting year-over-year shifts and capturing dynamic trends.

2.3 Machine Learning Methods

2.3.1 Linear Regression

Two linear forecasting models were implemented.

A linear regression model predicts state-level carbon emissions based on lagged emissions values, *population*, *state*, *size*, and *party affiliation score*s. It uses a time series cross-validation framework with $k = 5$ splits to respect temporal ordering and is evaluated using *Mean Squared Error* (MSE), R^2, and *explained variance*. Feature importance is determined based on *sklearn*'s built-in tools and is calculated by averaging the absolute value of coefficients across folds. We also test forecasting performance using sliding windows of 5 and

10 years to capture potential changes in temporal dynamics. While additional window configurations could offer further insight, exploring them fell outside the scope of this paper. We also implemented a ridge regression model with a standard regularization strength ($\alpha = 0.05$) to mitigate potential overfitting.

The logistic regression model, by contrast, predicts binary political party affiliation using lagged emissions data and demographic features such as population and geographic size. To avoid circularity, we excluded the *party affiliation score* variable, instead labeling states as Republican or Democrat based on whether their affiliation score was positive or negative in a given year. Its performance is assessed via *accuracy score*, and variable importance is similarly derived from average coefficient magnitudes across folds and states. We classify each year as Republican if the *party affiliation score* is negative, and Democrat if the score is positive. This classification approach resulted in 89% of state-year observations being labeled as Republican, and only 11% as Democrat. To address this imbalance, we apply random downsampling of the majority class within each training fold, ensuring a more balanced learning signal. To test robustness, we shuffled the labels as a baseline to ensure the model was capturing genuine signal rather than noise.

2.3.2 Random Forest

A random forest regressor was used as a non-linear alternative to test whether capturing complex interactions improved predictive performance. The model followed the same structure and configurations as the linear regression, using lagged emissions and demographic features, and was implemented with *sklearn*'s default hyperparameters for the random forest algorithm.

2.3.3 Clustering

To identify groups of states with similar political, demographic, and environmental characteristics, we implemented a clustering pipeline using the *KMeans* algorithm. The function first scales relevant state-level features—including R^2 scores from our predictive model, party affiliation, emissions change, population, geographic size, and total emissions—using *StandardScaler* to ensure comparability. To determine the optimal number of clusters, we apply the *Elbow Method* by computing the *sum of squared errors* (SSE) across a range of k values, and identifying the "elbow point" using Python's *KneeLocator* package. While useful as a guideline, the Elbow Method is a heuristic approach that can be subjective and sometimes inconclusive; it is best interpreted as part of broader exploratory data analysis rather than a definitive clustering criterion.

Once the "optimal" k is selected, *KMeans* is applied to assign states to clusters. The function returns average feature values per cluster and highlights discriminative variables based on the range of feature means across clusters.

2.4 Weighted Graph Inclusion

2.4.1 Weighted Graph-Based Spatial Modeling

To investigate whether geographic proximity and demographic similarity lead to spatial spillovers in carbon emissions, we augment the feature matrix with an explicit graph structure $G = (V, E, \mathbf{W})$. Each vertex $v_i \in V$ represents a U.S.state (N= *51* including the District of Columbia), and the edge–weight matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ encodes how strongly a pair of states i and j are expected to affect one another.

2.4.2 Graph Construction

Contiguity backbone. We first build a binary contiguity matrix $\mathbf{C}$ such that $c_{ij} = 1$ if states i and j share a land border (rook contiguity) and $c_{ij} = 0$ otherwise. Alaska and Hawaii have no contiguous neighbors and are therefore connected only through a small constant ε discussed below.

Feature-aware weights. Purely topological contiguity ignores heterogeneity across states. To model varying influence we define

$$w_{ij} = c_{ij}\,\exp\!\Big\{-\gamma_1\Big(\tfrac{|\mathrm{pop}_i-\mathrm{pop}_j|}{\sigma_{\mathrm{pop}}}\Big)^2 - \gamma_2\Big(\tfrac{|\mathrm{area}_i-\mathrm{area}_j|}{\sigma_{\mathrm{area}}}\Big)^2\Big\}, \qquad i\neq j,\; w_{ii}=0. \tag{1}$$

Here pop$i$ and area$i$ denote (log-scaled) population and land area of state i, σpop and σarea are the corresponding sample standard deviations, and $(\gamma_1, \gamma_2) \in \mathbb{R}> 0^2$ are bandwidth hyper-parameters tuned on a validation grid. The exponential kernel enforces that demographically similar neighbours exert stronger influence. To guarantee the graph is connected we add a small fully-connected component $w_{ij} \leftarrow w_{ij} + \varepsilon$ with $\varepsilon = 10^{-4}$.

Row–stochastic normalisation. Let $\mathbf{D} = \mathrm{diag}(\mathbf{W1})$ be the weighted degree matrix. We use the random-walk normalisation

$$\widetilde{\mathbf{W}};=;\mathbf{D}^{-1}\mathbf{W}, \tag{2}$$

so that rows of $\widetilde{\mathbf{W}}$ sum to one and can be interpreted as transition probabilities in a Markov chain.

2.4.3 Spatially-Regularised Regression

Given a vector of annual emissions $\mathbf{y} \in \mathbb{R}^N$ and covariates $\mathbf{X} \in \mathbb{R}^{N\times d}$ for a fixed year, we learn regression coefficients $\boldsymbol{\beta}$ via

$$\min_{\boldsymbol{\beta}};\underbrace{|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|2^2}\,\text{data-fit} + \underbrace{\lambda\sum_{i,j} w_{ij}\big(\mathbf{x}_i^\top\boldsymbol{\beta}-\mathbf{x}j^\top\boldsymbol{\beta}\big)^2}\,\text{graph Laplacian penalty}, \tag{3}$$

where $\lambda > 0$ controls smoothness across adjacent states. Equation (3) can be solved in closed form as $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{X}^\top\mathbf{L}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ with $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

2.4.4 Graph Neural Network Variant

For a fully non-linear alternative we treat emissions forecasting as a node-level regression problem on a spatio-temporal graph. Let $\mathbf{H}^{(0)}! =!\mathbf{X}$ be the input feature matrix and apply K Graph Convolutional Network (GCN) layers:

$$\mathbf{H}^{(k+1)} = \sigma\big(\widehat{\mathbf{D}}^{-1/2}\widehat{\mathbf{W}}, \widehat{\mathbf{D}}^{-1/2}\mathbf{H}^{(k)}\mathbf{\Theta}^{(k)}\big), \quad k = 0, \ldots, K-1, \tag{4}$$

where $\widehat{\mathbf{W}} = \mathbf{W} + \mathbf{I}$ adds self-loops, $\widehat{\mathbf{D}}$ is its degree matrix, $\mathbf{\Theta}^{(k)}$ are learnable weight matrices, and σ is a non-linearity (ReLU). The final prediction is $\hat{\mathbf{y}} = \mathbf{H}^{(K)}\boldsymbol{\theta}_{\text{out}}$.

2.4.5 Temporal Splitting and Training

We roll out the graph model for each year t, sharing weights across years. Training uses data from 1970–2015, validation on 2016–2018 to tune $(\gamma_1, \gamma_2, \lambda, K)$, and testing on 2019–2022. Loss is Mean Squared Error with ℓ_2 regularisation on parameters.

2.4.6 Diagnostic Checks

To verify that spatial correlation has been captured we compute Moran's I statistic on model residuals and compare it to a null permutation distribution. A value of I not significantly different from zero indicates that the graph-based model has removed spatial autocorrelation that remained in baseline approaches.

2.4.7 Hyper-parameter Summary

Symbol	Description
γ_1	population kernel bandwidth (grid $\{0.5, 1, 2, 4\}$)
γ_2	area kernel bandwidth (same grid as γ_1)
λ	Laplacian regularisation weight ($10^{\{-3,-2,-1,0\}}$)
K	number of GCN layers ($\{1, 2, 3\}$)

Table 1: Hyper-parameters explored during cross-validation.

3 Results

3.1 Linear Forecasting

3.1.1 Temporal Linear Regression Forecasting

Using a 10-year sliding window. The linear regression model performed well, averaging an 0.9947 R^2. The most influential predictor in 60% of folds was the *party affiliation score*, with the highest average coefficient (1.71) and

greater variability across folds (std = 0.92), suggesting a context-dependent influence. This was followed by *population*, which also contributed significantly to the model's predictions.

Performance varied across states, with an average stratified R^2 of 0.6070. Additionally, *party affiliation score* was the top predictor in only 6.82% of state-level models, underscoring its selective but notable impact.

Using a 5-year sliding window: Results remained consistent overall, though the average stratified R^2 improved to 0.7902, with *party affiliation score* emerging as the top predictor in 17.78% of state models.

Note: The overall change in emissions was relatively modest - only 2.52% (102.54 million metric tons) annually on average across states - which likely explains the high accuracy of the model. With limited variation in the predictor, the model learns stable trends that may not reflect meaningful changes in political affiliation.

3.1.2 Regularizing Using Ridge Regression

To fix the suspiciously high R^2 score in our plain linear regression model, we used a ridge regression model and achieved a strong but lower global fit.

Using a 10-year sliding window: The ridge regression model scored an average R^2 of 0.9727 across all folds, demonstrating stronger generalization and the ability to capture overall emission trends. This time *population* had the highest average coefficient (4.32), followed by *party affiliation score* (1.91) which was never the most influential feature in any of the folds. Stratified R^2 performed poorly, with an average of -5.2432, indicating the model failed to generalize across states. Notably, *party affiliation score* emerged as the top predictor in 18.18% of state models—an increase from the standard linear model.

Using a 5-year sliding window: Model performance remained stable across metrics, with *party affiliation score* the top predictor in 22.22% of state models. However, its average coefficient (3.49) was notably lower than that of *population* (7.59), indicating lesser influence. Notably, *size* showed a negative coefficient for the first time, aligning with the tendency of larger states to lean Republican and thus exhibit lower *party affiliation scores.* [7]

3.1.3 Feature Engineering to Correct for *size* and *population*

To control for the outsized influence of *population* and *size*, we introduced feature-engineered variables—*emissions per capita* and *population density.* However, this adjustment significantly reduced model performance. The average R^2 dropped to -0.0478, with a stratified R^2 of -0.0742, indicating the model performed worse than a naive baseline. These results show that removing absolute demographic scale limits predictive power in this context.

3.1.4 Logistic Regression Analysis

Given the high accuracy of our continuous emissions forecasting model, we further explored the relationship between emissions patterns and political alignment by applying a logistic regression model to classify party affiliation.

The model achieved approximately 0.5050 accuracy and 0.4434 stratified accuracy across both 5- and 10-year sliding windows, indicating that it is essentially performing at the level of random guessing and that there is no clear relationship between carbon emissions and our feature set with this model configuration.

3.2 Clustering

To explore whether states with high predictive performance shared common characteristics, we clustered states based on their R^2 scores and our previously used features. The results revealed distinct groupings: The cluster with the highest average R^2 (0.926), was marked by high emissions, large geographic size, and strong positive emissions change, while the cluster with the lowest R^2 (0.409), was associated with smaller populations and moderate emissions.

An analysis of feature ranges across clusters identified *emissions*, *emissions change*, and *population* as the most discriminative variables, whereas *party affiliation score* contributed minimally to cluster separation.

3.3 Random Forest Model

3.3.1 Attempting Regression Using a Non-Linear Model

To test whether emissions forecasting could benefit from a non-linear modeling approach, we implemented a random forest regression model using the same feature set as our linear models. The model achieved strong overall performance, with an average R^2 of approximately 0.94 for both 10- and 5-year sliding windows. The most important feature by a large margin was the previous year's emissions, though *population* and *size* also contributed modestly. However, stratified R^2 scores revealed substantial underperformance at the state level, averaging -0.5874 and -0.2407 across states for the 10- and 5-year windows, respectively. These results suggest that while random forests are effective at capturing national-level trends, they struggle to generalize across individual states—likely due to data sparsity or state-specific variation not well captured by the available features. Damningly, *party affiliation score* ranked last in average feature importance across all folds, indicating minimal influence in the random forest model's decision-making process.

3.4 Adding Spatial Structure

Incorporating spatial structure through adjacency graph features as described in section 2.4 generally improved model performance across both linear and nonlinear approaches. The enhancement was most notable for ridge regression

and random forest models under the 5-year window, suggesting that spatial dependencies are more influential when temporal data is limited. However, not all configurations benefited; for instance, the 10-year OLS model showed a marginal decrease in accuracy with the graph features, indicating potential redundancy or overfitting when temporal coverage is sufficient.

Model	**Window (yrs)**	**Spatial graph**	$\mathbf{R^2}$	**RMSE**
OLS	10	–	0.9947	5.93
OLS	10	✓	0.9952	5.66
OLS	5	–	0.9945	6.09
OLS	5	✓	0.9943	6.19
Ridge	10	–	0.9727	15.82
Ridge	10	✓	0.9765	15.17
Ridge	5	–	0.9700	17.32
Ridge	5	✓	0.9712	17.18
RF	10	–	0.9339	41.14
RF	10	✓	0.9290	41.23
RF	5	–	0.9383	36.79
RF	5	✓	0.9440	35.90

Table 2: Cross–validated performance on state-level CO_2 emissions forecasting. A "✓" indicates that the weighted adjacency matrix from Eq. (1) was incorporated either as a Laplacian penalty (OLS/Ridge) or as an additional input feature set (RF). Introducing spatial structure helps in most—but not all—settings, suggesting that the benefit depends on both model capacity and the temporal context.

4 Discussion

Our logistic regression and feature engineering efforts to control for state size and population were largely uninformative and served primarily exploratory purposes. The more compelling insights emerged from the shifting role of *party affiliation score* across modeling approaches. In plain linear regression, its overall importance increased across folds as the sliding window lengthened; however, at the state level, its influence diminished with longer windows. In contrast, ridge regression revealed that while *party affiliation score* retained some prominence with shorter windows, its overall importance declined across folds and it was no longer the top predictor. Most notably, in the random forest model—which achieved R^2 scores comparable to the linear models—*party affiliation score* exhibited little to no importance. This finding was echoed in our clustering analysis, where it played only a minor explanatory role in identifying high R^2 clusters.

The inclusion of spatial structure appears to enhance predictive accuracy in

most cases, highlighting the value of accounting for geographic dependencies in housing market modeling. This benefit was especially pronounced when using shorter time windows or more complex models. Nonetheless, the occasional decline in performance suggests that spatial features should be applied judiciously, possibly requiring regularization or selection to avoid introducing noise in already well-specified models.

Taken together, these results suggest that while political affiliation may correlate with emissions patterns under certain linear assumptions, its explanatory power weakens under regularized and non-linear modeling, indicating that it may act more as a proxy variable than a direct driver of emissions trends.

5 Conclusion

While political affiliation demonstrated some predictive power in our models, its influence was inconsistent and overall weak. In the absence of longer-term data and additional contextual variables—such as economic indicators or policy sentiment—our analysis cannot provide definitive evidence of a meaningful relationship. For now, the possibility of a correlation between political party affiliation and carbon emissions remains open and promising as a result of our linear models, but our results do not offer clear or conclusive support.

References

[1] J. Basseches, R. Bromley-Trujillo, M. Boykoff, T. Culhane, G. Hall, N. Healy, D. Hess, D. Hsu, R. Krause, H. Prechel, J. Roberts, and J. Stephens, "Climate policy conflict in the u.s. states: a critical review and way forward," *Climatic Change*, vol. 170, no. 3-4, p. 32, 2022. Epub 2022 Feb 16.

[2] I. Gallup, "Record party gap on environment-economic growth tradeoff," 2024. Accessed: 2025-04-02.

[3] U.S. Energy Information Administration, "State carbon dioxide emissions data," 2023.

[4] U.S. Census Bureau, "State population totals and components of change: 2020-2023," 2024. Accessed May 6, 2025.

[5] U.S. Census Bureau, "State area measurements and internal point coordinates," 2010.

[6] LegiScan LLC, "Legiscan," 2025.

[7] E. Data, "County electoral map: Land vs. population," 2020.