© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

U.S. State Energy Transition Dynamics: Renewable Adoption, Fossil Displacement, And Forecasting with Ensemble Machine Learning

CHRISTOPHER ODEDINA[1], TOCHUCKWU AKAEGBUSI[2]

Abstract- ***The study investigates the dynamics of the United States' energy transition in all 50 states (including the District of Columbia), over the period 1990 to 2024, using the United States Energy Information Administration's State Energy Datasets. Using a combination of panel econometrics, clustering, and machine learning, this study aims to identify major patterns in the uptake of renewable energy sources in the United States, along with their economic effects. The preliminary results show significant cross-state heterogeneity and structural change in 2008 as the major turning point in the transition path. In addition, four major archetypes of states are identified, namely: early movers, progressive large-scale adopters, gradually transitioning states, and fossil-locked states, which are characterized by their high dependence on conventional sources of energy. Furthermore, a higher share of renewable energy reduces total energy expenditure, taking into account factors such as energy demand and prices. Based on the machine learning models, this study demonstrates a high degree of predictive accuracy ($R^2 \approx 0.77$). Also, the interpretability analysis of the results demonstrates that lagged renewable share is the dominant predictor of future adoption. This implies a high degree of path dependence in which historical energy structures strongly constrain current transition dynamics. Thus, the U.S. energy transition is heterogeneous, cost-reducing, and structurally persistent, with inertia in renewable adoption representing a key barrier to accelerated decarbonization.***

Index Terms- ***Energy transition; Renewable energy; Fossil fuels; U.S. state panel data; XGBoost; LightGBM; SHAP; Structural break; Panel regression; Decarbonisation***

I. INTRODUCTION

The global energy system is in the midst of a structural transformation from fossil fuels to renewable energy sources. This is driven by the confluence of the objectives of climate stabilization, security of energy supply, and economic efficiency. However, the pace of this transformation is uneven across the states in the United States (US) owing to the county diversity in its resource base, regulatory environment, industrial composition, and political economy. The US is one of the largest consumers of global energy, which accounts for around 15% of the total global primary energy demand [1]. Although the renewable energy capacity in the country has grown significantly in recent decades with the support of policy instruments such as the Production Tax Credit, Renewable Portfolio Standards, the American Recovery and Reinvestment Act of 2009, and the Inflation Reduction Act of 2022, the pace of the transformation is significantly heterogeneous in the country [2, 3].

The economic consequences of this process are also noteworthy. For instance, energy costs are a significant portion of household budgets, business expenses, and public service bills. Renewable technology supporters argue that a decrease in costs and reduced exposure to fuel price volatilities will ultimately lead to a decrease in long-run energy costs [4], while opponents point to the high costs of integrating renewable technology as a factor for increasing costs [5]. The empirical verification of these arguments at the state level for a period of several decades is still an open issue. For instance, while some researchers have focused on aggregate analysis at the national level, failing to capture regional differences, others have conducted regional analysis but have failed to provide a comprehensive view. From a methodological perspective, researchers have primarily relied on either econometric approaches, which are easy to interpret but less flexible, or machine learning approaches, which are flexible but less interpretable. Moreover, the role of structural breaks, especially policy-induced breaks as seen in the ARRA, and the use of interpretable

machine learning methods, including SHAP [6], is still unexplored.

Against the backdrop aforementioned, this study aims to bridge this gap by offering a comprehensive analysis of the state-level US energy transition dynamics from 1990 to 2024. The study seeks to provide empirical evidence of the renewable energy revolution and fossil fuel consumption, detect structural breaks and transition dynamics, and classify US states according to their specific archetypes of transition. It estimates the relationship between the composition of energy consumption and expenditures using panel econometric methods, and supplements this by using machine learning techniques to predict renewable adoption. Most importantly, this analysis seeks to uncover the structural and dynamic determinants of transition by using SHAP and linking this to economic theory on path dependence and resource scarcity.

Thus, the study develops a broad and exhaustive panel dataset of states and years, covering 34 years and 52 jurisdictions, using a scalable data pipeline. From a methodological perspective, the study combines rigorous diagnostics with econometric and machine learning techniques, ensuring the robustness of inference and prediction. Also, the study offers the panel-based results demonstrating the statistical significance of 2008 as a year of structural change in the US energy transition, corresponding to the ARRA stimulus. It establishes that path dependence, defined by lagged renewable adoption, is the dominant driver of transition dynamics, with significant implications for the rate and possibility of decarbonization. The rest of the paper follows this structure: literature review, data and methods, result presentation, conclusion, recommendations, and future work.

II. LITERATURE REVIEW

2.1 Theories of Energy Transition

The study is grounded in three major paradigms according to the Multi-Level Perspective (MLP) theory explains an energy transition as an interaction between innovations, an existing sociotechnical regime, and exogenous pressures on the subject [7, 8]. In this perspective, the transition from fossil-based to renewable-based energy is an example of a regime transition driven by technological, political, and social drivers. Although widely used in qualitative research, there is limited empirical application of MLP theory in subnational panel analysis. Contrary to the MLP theory, York and Bell [5] argued an alternative theory of an *"addition model,"* where renewable energy is added to the total pool of energy without replacing fossil-based energy. This is supported by recent trends showing an increase in renewable-based energy without a corresponding decrease in fossil-based energy. Further, the theoretical implications of the study are extended by Arman et al. [9], which views the energy transition as a physical and digital revolution, focusing on the importance of data-driven systems in facilitating renewable energy integration. Also, the Technological Innovation Systems (TIS) framework provides the third perspective, which focuses on learning by doing and policy-driven cost reduction. Kavlak et al. [4] noted that the significant reduction in the costs of solar photovoltaics is driven by policy-induced market expansion rather than external technological progress, which may be a basis for the expectation of policy-driven structural breaks in the energy transition process.

2.2 Empirical Studies on U.S. State-Level Energy Dynamics

Empirical research on the dynamics of energy in the US states highlights the significance of heterogeneity: in their study, Delmas and Montes-Sancho [2] reveal the systematic relationship between the early adoption of Renewable Portfolio Standards and resource endowments, suggesting the existence of selection effects in the evaluation of the effectiveness of the RPS, and hence the use of the fixed effects model to address the heterogeneity issue. Similarly, Wiser et al. [3] reveal the existence of benefits from the adoption of renewables, which are not necessarily reflected in the price savings but in the improvements in public health and environmental externalities. This can explain the low estimates of the effects of the adoption of renewables based on the savings in energy expenditure. Herrera et al. [10] noted the heterogeneity of the states in the US by revealing the significant differences in the dynamics of energy prices in different regulatory and market regimes, where linear models are ineffective in capturing the complexity of the relationships. This

limitation has prompted the application ensemble models, which have shown promise in enhancing the accuracy of predictions [9].

Recent developments support the application of ML and interpretation in energy research. For instance, Guo et al. [11] showed the efficacy of the application of gradient-boosting models, along with SHAP, in emissions prediction, while Hao and He [12] stated that non-parametric neural network models perform better than traditional linear models in forecasting fossil energy consumption. The significance of interpretation in ML models in energy research is also supported by the need to move from prediction to policy, which is achieved through SHAP. Moreover, recent evidence from cointegration models, as shown in Nica et al. [13], serves as robust evidence that there are equilibrium relationships between energy consumption and macroeconomic variables in the long run, which is supported by the unit root behavior observed in energy consumption.

2.3 Panel Econometric Approaches in Energy Research

The classical linear specification with fixed effects remains a popular choice, although it may not be able to account for nonlinearities, regime shifts, and irregularities in the underlying energy distributions, which are inherent in multi-decade datasets. Herrera et al. [10] noted that, although linear baseline specifications are useful for establishing a reference point, they consistently perform poorly in the presence of structural shocks and price regime shifts, thereby highlighting the importance of a diagnostic-driven modelling approach. The treatment of non-stationarity is critical in the context of the energy sector, as most variables are characterized by persistence and stochastic trends in the data. The results of stationarity tests on the relevant datasets determine the modelling frameworks and choice of incorporating differencing and lag structures. Nica et al. [13] show that autoregressive distributed lag frameworks may be able to recover long-run relationships in the presence of non-stationarity, and MacKinnon [14] provides critical value tables for the Augmented Dickey–Fuller tests in such contexts. Moreover, the limitations of purely parametric models are also underlined in the recent work by Guo et al. [11] who points out that linear regressions are not sufficient to identify threshold effects and nonlinear interactions in emission prediction. Thus, the need to combine econometric models and machine learning in a unifying approach [15].

2.4 Machine Learning in Energy Forecasting

This rapid development of machine learning in the field of energy forecasting is a result of the necessity of better handling nonlinear dynamics and high-dimensional interactions compared to traditional time-series-based approaches. Benti et al. [15] demonstrated hybrid/ensemble machine learning models as the current frontier in renewable energy forecasting, as these models outperform other machine learning models in terms of accuracy and robustness. Within this framework, gradient-boosted models have shown promising results. Guo et al. [11] demonstrated the applicability of gradient-boosted trees with SHAP-based interpretability in attributing policy-relevant outcomes of the model predictions. Hao and He [12] showed the superiority of the performance of non-parametric neural networks in fossil energy consumption forecasting compared to linear models. Herrera et al. [10] have shown the advantages of hybrid econometric/machine learning models in the context of energy forecasting. The benefits of ensemble modeling are also supported by Arman et al. [9], who demonstrate the effectiveness of reducing the error in the forecasting of renewable energy generation, and Benti et al. [15], who highlight the benefits of better uncertainty quantification. The challenge of interpretability, which has been identified as a significant limitation of machine learning in energy analytics [26], is addressed through the use of SHAP-based decomposition, which allows for the economically relevant contributions of complex outputs to be determined [11, 15].

Despite these advances, the majority of the extant literature has failed to combine rigorous diagnostics, fixed effects panel econometric techniques, unsupervised machine learning, and ensemble machine learning in a unified framework. Furthermore, while the impacts of policy-induced structural changes on the energy system are a common theme in the literature, the effects of the American Recovery and Reinvestment Act on the adoption of renewables in the states of the USA have

yet to be formally investigated using a comprehensive multi-state Chow structural break analysis. While SHAP has been used in the modelling of greenhouse gas emissions and decarbonization pathways [11], its application in the decomposition of renewable share forecasts in the states of the USA and the investigation of path dependence as a primary mechanism of the transition has yet to be explored. This paper seeks to fill these gaps in the literature by combining econometric and machine learning techniques in a unified framework spanning multiple decades and jurisdictions.

III. METHODOLOGY

3.1 Theoretical Framework

This study is informed by the *energy transition and structural change* approach, which views the transition from fossil fuel-based energy systems to renewable-based ones as a path-dependent, policy-driven, and heterogeneous process across different jurisdictions [7, 8]. From an economic point of view, total energy expenditures are a function of both energy costs and volumes, whereas the composition of the energy mix affects long-run cost efficiency and sustainability. Following dynamic transition theory and empirical evidence on learning and induced technological changes in response to policies [4], renewable energy adoption is found to display strong temporal dependence driven by inertia in physical infrastructures and policies. Hence, lag structures and nonlinear approaches are used to capture strong transition effects.

3.2 Data and Study Area

3.2.1 Data Source and Coverage

For the empirical analysis, the dataset was obtained from the United States Energy Information Administration State Energy Data System (SEDS), which contains comprehensive state-level information on energy production, consumption, prices, and expenditures [1]. The dataset was accessed via a memory-efficient streaming pipeline, which allows for the direct handling of NDJSON data and facilitates the inclusion of the entire sample period from 1990-2024 without memory constraints. The sample comprises all 50 U.S. states and the District of Columbia, resulting in a dataset of 1,820 state-years, which is a largely balanced panel, although some missing values are present in early solar series before 2005, and U.S. territories are excluded due to incomplete data availability.

3.2.2 Variable Description and Construction

The study develops 34 variables in five categories. The variables in the first category, *Energy Consumption*, include total energy consumption, renewable energy consumption, fossil fuel consumption, and fuel-specific variables (coal, natural gas, petroleum, hydropower, wind, solar, geothermal, and biomass) expressed in British thermal units. The second category, *Price and Expenditure*, comprises variables including total energy expenditure, fuel-specific prices, and retail electricity prices. The third category, *Ratio Variables*, is included in the study as a way of mitigating multicollinearity in volumetric variables, including renewable share and fossil share variables. The variables in the fourth category, *Engineered Variables*, include three-period lag variables, growth rates, temporal indices, decade variables, a structural break variable for post-2008, and a state cluster membership indicator variable.

3.2.3 Training and Validation Split

Model evaluation is done using a temporal validation strategy: the training set is composed of the observations from 1990 - 2017, which equals 1,456 state-years, and the validation set is composed of the observations from 2018 - 2020, which equals 364 state-years.

3.3 Analytical Framework

The study uses a five-stage empirical approach: first, preliminary diagnostics are conducted to identify the statistical properties and the corresponding inference procedures. Second, the analysis of the regime shifts caused by the government's interventions, or the 'structural breaks,' is carried out. Third, the unsupervised learning method of clustering the states based on their transition archetypes is conducted. Fourth, the panel econometric regression method is used to identify the determinants of the energy expenditure. Finally, the ensemble machine learning method with the 'SHAP' interpretability technique is used to identify the determinants.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

3.3.1 Preliminary Statistical Diagnostics

For normality, the Shapiro-Wilk test [16] and the D'Agostino K-squared test [17] are applied. For stationarity, the Augmented Dickey-Fuller test [18] and the KPSS test [19], along with critical values from MacKinnon [14], are applied. Cross-state heterogeneity is investigated by means of the Kruskal-Wallis H test [20], along with eta-squared effect size measures. Multicollinearity is investigated by means of variance inflation factors, whereas heteroscedasticity is investigated by means of the Breusch-Pagan test [21], and autocorrelation is investigated by means of the Durbin-Watson test.

3.3.2 Structural Break Analysis

For identifying the structural breaks in the renewable share, the Chow F-test is employed, as suggested by Chow [22], and iterated for different structural breaks between 2003 and 2018. The test for the presence of a structural break at a given time point t^* is conducted by comparing the results of the pooled regression and the regressions for the subsamples before and after the hypothesized break point. The test statistic is defined as:

$$F = \frac{(RSS_0 - (RSS_1 + RSS_2))/k}{(RSS_1 + RSS_2)/(n-2k)} \quad \text{---------- (1)}$$

where RSS_0 is the residual sum of squares from the pooled regression, $RSS_1$and $RSS_2$are the residual sums of squares from the pre- and post-break regressions, $k$is the number of estimated parameters, and n is the total number of observations. The break year that maximises the Chow F-statistic is selected as the dominant structural shift. A binary indicator variable, $Post2008_t$, equal to one for years after the identified break and zero otherwise, is subsequently incorporated into the econometric and machine learning models to capture regime change effects.

3.3.3 State Transition Archetype Clustering

Unsupervised clustering is accomplished through a K-means clustering algorithm with a principal component reduced feature space. Seven state-level mean values for the 2010-2024 period are standardised before dimension reduction. The number of clusters is determined based on silhouette score calculations over a range of k from two to seven. The clusters define unique state-level energy transition archetypes.

3.3.4 Panel Econometric Model

Economic determinants of expenditure on energy are estimated using a two-way fixed effects ordinary least squares model, which is robust to heteroscedasticity of unknown form, specifically HC3, as proposed by White [23]. The equation is represented as follows:

$$\ln(Expenditure_{it}) = \alpha + \beta_1 RenewShare_{it} + \beta_2 FossilShare_{it} + \beta_3 \ln(TotalEnergy_{it}) + \beta_4 CoalPrice_{it} + \beta_5 PetrolPrice_{it} + \beta_6 ClusterID_i + \beta_7 Post2008_t + \gamma_i + \delta_t + \varepsilon_{it}$$

----------(2)

Where i indexes states and $t\ denotes$ time. $RenewShare_{it}$ and $FossilShare_{it}$ measure energy mix composition, $\ln(TotalEnergy_{it})$ captures scale effects, $CoalPrice_{it}$ and $PetrolPrice_{it}$ control for fuel price variation, $ClusterID_i$ represents transition archetypes, and $Post2008_t$ captures structural regime change. The terms γ_i and δ_t denote state and year fixed effects, respectively, while ε_{it} is the idiosyncratic error term. This specification isolates the effect of energy transition dynamics on expenditure while controlling for persistent cross-state heterogeneity and common macroeconomic shocks.

3.3.5 Machine Learning Forecasting Framework

Two algorithms for implementing a gradient-boosting model are XGBoost [24] and LightGBM [25]. XGBoost uses a depth-wise tree growth method along with L1 and L2 regularisation, whereas LightGBM uses a leaf-wise method along with a histogram-based splitting technique, and both techniques incorporate an internal early stopping method for a holdout sample. A technique of equal-weight ensemble is used for averaging the results of

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

both algorithms. The performance of the model is checked by using the RMSE, MAE, and R^2.

3.3.6 SHAP Interpretability

Model interpretability is achieved through SHAP values using the Tree Explainer algorithm [6]. Global feature importance is achieved through the mean absolute SHAP values, while individual contributions result in variation. The analysis tests for path dependence by comparing the importance of renewable share variables and structural and policy features.

IV. RESULTS AND DISCUSSION

4.1 Preliminary Statistical Diagnostics

4.1.1 Descriptive Statistics

Descriptive statistical results reveal a large cross-state variation in renewable energy adoption, as presented in Table 1 (Fig. 1). The mean renewable energy share is 7.6%, with a standard deviation of 7.1% and a coefficient of variation of 0.94%. The range of renewable energy share varies from 1.2 to 25.2%, indicating a large geographical inequality (Fig. 2). The strong right skewness of the energy variables (all with a skewness of more than 3) and heavy tails (all with a kurtosis of more than 12) indicate the influence of large-state outliers, Texas, and California. The observed distributions of the energy variables are consistent with the reported energy heterogeneity among U.S. states [2, 10]. The degree of cross-state variation in renewable energy adoption is similar to the heterogeneity issues reported in national-level analytics, where aggregated modeling conceals unit-level structural differences.

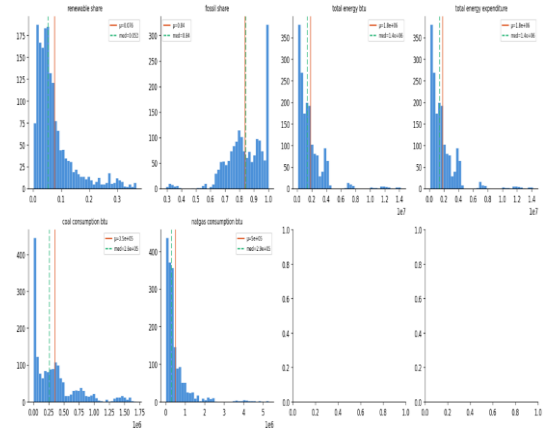

Fig. 1: Distribution of key study variables (all states, 1990–2024)

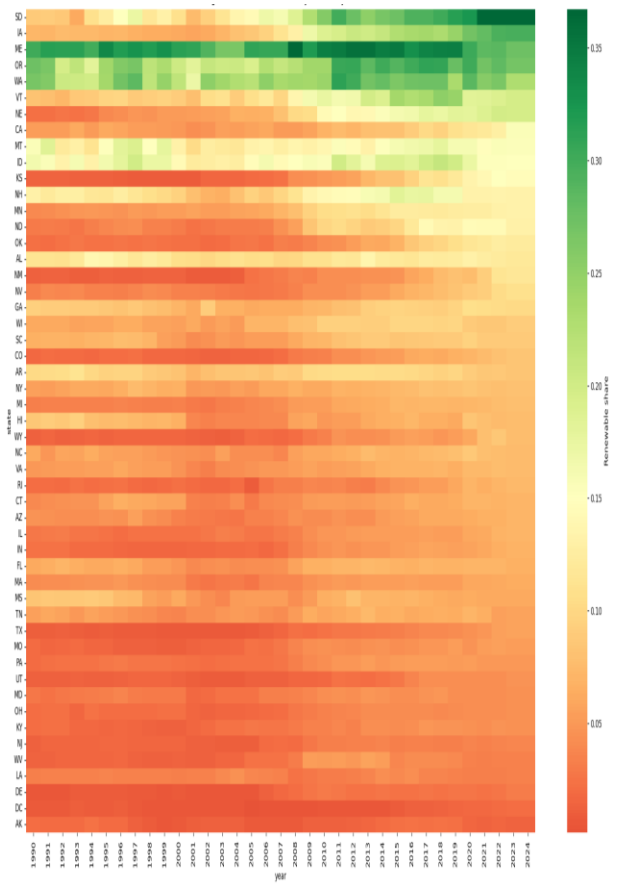

Fig. 2: Renewable share by state and year (1990–2024)

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

Table 1. Descriptive Statistics of Key Study Variables

Variable	Mean	SD	CV	Skewness	Kurtosis	5th pct.	95th pct.
Renewable Share	0.076	0.071	0.939	1.920	3.525	0.012	0.252
Fossil Share	0.838	0.134	0.160	−1.055	2.002	0.642	1.000
Total Energy (Btu)	1,822,445	2,002,264	1.099	3.071	12.428	200,299	4,257,256
Total Energy Expenditure	1,822,286	2,001,828	1.099	3.072	12.437	200,299	4,257,261
Coal Consumption (Btu)	353,918	377,477	1.067	1.453	1.742	99	1,286,332
Natural Gas Consumption (Btu)	499,574	692,694	1.387	3.568	15.634	21,868	1,673,836

Note. n = 1,820 state-year observations (1990–2024; 52 states/territories). Btu values in thousands. Expenditure in thousands of current USD. CV = coefficient of variation (SD/|mean|)

4.1.2 Normality Tests

Table 2 shows the results of the Shapiro-Wilk and D'Agostino K^2 tests, which strongly reject the null hypothesis of normality for each variable, confirming a p-value lower than 0.001. In addition, the Shapiro-Wilk statistic ranges between 0.607 (natural gas) and 0.909 (fossil share). These results are in line with previous studies on energy forecasting, which reported that the distribution of energy types is not normally distributed [10, 12]. In terms of methodology, this study supports the use of Kruskal-Wallis tests for between-state comparisons and HC3 heteroskedasticity-robust inference in panel regressions

Table 2. Normality Tests via Shapiro-Wilk and D'Agostino

Variable	n	SW Stat.	SW p-value	K^2 Stat.	K^2 p-value	Normal
Renewable Share	1,785	0.780	0.000	659.52	0.000	No
Fossil Share	1,785	0.909	0.000	316.54	0.000	No
Total Energy (Btu)	1,785	0.691	0.000	1195.03	0.000	No
Energy Expenditure	1,785	0.691	0.000	1195.32	0.000	No
Coal Consumption (Btu)	1,785	0.833	0.000	433.51	0.000	No
Nat. Gas Consumption (Btu)	1,785	0.607	0.000	1361.76	0.000	No

Note. SW = Shapiro-Wilk statistic (Shapiro & Wilk, 1965). K^2 = D'Agostino-Pearson omnibus test (D'Agostino et al., 1990). $\alpha = 0.05$.

Fig. 3: U.S.-aggregate trend of key variables

4.1.3 Panel Stationarity Tests

Table 3 (Fig. 3) shows the results of the ADF-KPSS tests, which classify the renewable share as I(1) since the ADF test yields a p-value of 0.968 and the KPSS test yields a p-value of 0.016. Similarly, the fossil share is found to be I(1) since the ADF test yields a p-value of 0.809 and the KPSS test yields a p-value of 0.024. Finally, the total energy consumption is found to have a near unit root since the ADF test yields a p-value of 0.067 and the KPSS test yields a p-value of 0.060. These findings align with the results in the literature on the energy transition, which use the ARDL methodology to account for the presence of unit roots and long-run relationships in the data [13].

Table 3. Panel Stationarity Tests - Augmented Dickey-Fuller and KPSS

Variable	ADF Stat.	ADF p-value	KPSS Stat.	KPSS p-value	Conclusion
Renewable Share	0.132	0.968	0.674**	0.016	I(1)
Fossil Share	−0.833	0.809	0.587**	0.024	I(1)
Total Energy (Btu)	−2.741*	0.067	0.439	0.060*	I(1)

Note. ADF null: unit root (Dickey & Fuller, [18]). KPSS null: stationary (Kwiatkowski et al. [19]). Critical values from MacKinnon [14].

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

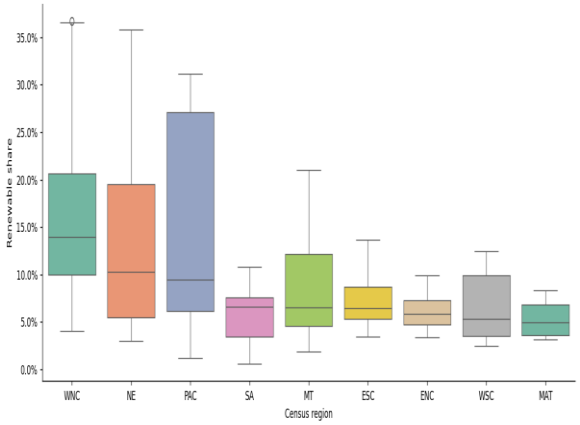

Fig. 4: Renewable share by Census region (2010–present)

4.1.4 Cross-State Heterogeneity

The Kruskal-Wallis results in Table 4 confirm extreme cross-state heterogeneity (Fig. 4) in renewable share ($H = 1{,}328.50$, $\eta^2 = 0.737$) and energy expenditure ($H = 1{,}763.44$, $\eta^2 = 0.988$). State identity explains 73.7% of renewable share variance and 98.8% of expenditure variance, indicating that structural characteristics dominate outcomes. These effect sizes exceed those reported by Delmas and Montes-Sancho [2], who attribute 40–60% of variation to resource endowment. The dominance of structural context parallels evidence that system-level characteristics determine efficiency outcomes in complex analytics environments [27].

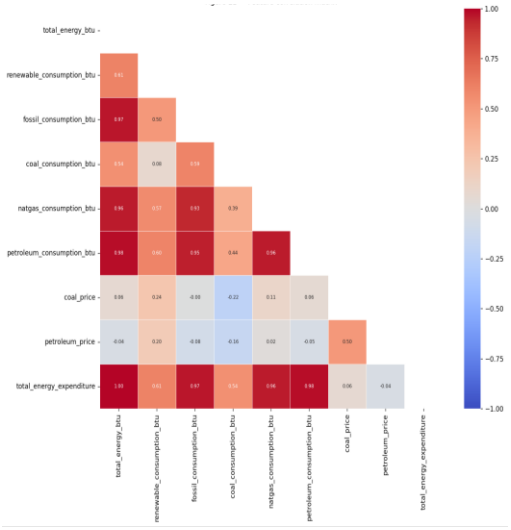

Fig. 5: Feature correlation matrix

Table 4. Cross-State Heterogeneity (Kruskal-Wallis Test)

Variable	H Statistic	p-value	η^2	Significant
Renewable share	1,328.50	< 0.001	0.737	Yes ***
Total energy expenditure	1,763.44	< 0.001	0.988	Yes ***

Note. $\eta^2 = (H - k + 1) / (N - k)$. *** $p < 0.001$

4.1.5 Multicollinearity Diagnostics

Variance inflation factors point to the high multicollinearity between volumetric energy variables, as indicated in Table 5 (Fig. 5). Total energy consumption and expenditure have VIFs above 400,000, implying perfect linear association, while petroleum, fossil, and natural gas consumption have VIFs ranging from 20 to 56. This issue is addressed by substituting raw volumes with log and share-based transformations, which solves the specification bias and stability, as suggested for the panel modeling of emissions in Guo et al. [11].

Table 5. Multicollinearity Diagnostics - Variance Inflation Factors (VIF)

Feature	VIF	Status
Total Energy (Btu)	439,242	High (>10)
Total Energy Expenditure	438,684	High (>10)
Petroleum Consumption (Btu)	55.82	High (>10)
Fossil Consumption (Btu)	22.92	High (>10)
Natural Gas Consumption (Btu)	20.33	High (>10)
Coal Consumption (Btu)	5.02	Moderate (5–10)
Renewable Consumption (Btu)	2.54	Acceptable (<5)
Coal Price	1.52	Acceptable (<5)
Petroleum Price	1.47	Acceptable (<5)

Note. VIF > 10: exclude from OLS. 5-10: moderate. <5: acceptable.

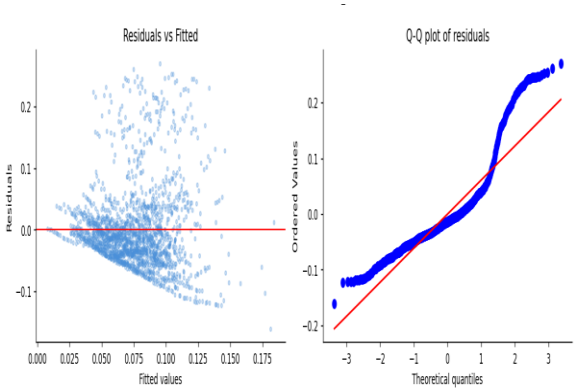

Fig. 6: OLS diagnostic plots — residuals vs. fitted and Q-Q plot (renewable share baseline)

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

4.1.6 Heteroskedasticity and Autocorrelation

As shown in Table 6 (Fig. 6), the results of the diagnostic tests indicate a high level of heteroskedasticity (Breusch-Pagan LM = 144.37, $p < 0.001$) and positive autocorrelation (Durbin-Watson = 0.085). The extremely low DW statistic points to a high level of serial dependence, which is characteristic of long-horizon state panels. HC3 standard errors for heteroskedasticity are employed throughout, which account for both forms of violation [21, 23].

Table 6. Heteroskedasticity and Autocorrelation Diagnostics

Test	Statistic	p-value	Conclusion
Breusch-Pagan LM	144.37	< 0.001	Heteroskedasticity present
Breusch-Pagan F	39.16	< 0.001	Confirms LM result
Durbin-Watson	0.085	N/A	Strong autocorrelation

Note. Breusch-Pagan test [21]: null = homoskedastic errors. DW statistic: values near 2 indicate no autocorrelation.

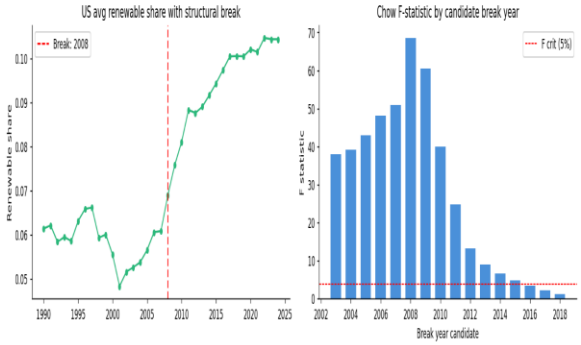

Fig. 7: Structural break analysis (Chow F-statistic by candidate break year)

4.1.7 Structural Break Analysis

In Table 7 (Fig. 7), Chow tests identify 2008 as the primary point of structural change in national renewable share ($F = 68.40$), with surrounding years 2007-2009 forming a policy-driven transition window. The new regime following 2008 displays a permanent shift in renewable adoption, which is consistent with the idea of investment shocks driving permanent economies of scale and learning effects [4]. The persistence of this shift suggests a form of hysteresis in energy transition dynamics, where policy interventions permanently shift adoption trajectories.

Table 7. Structural Break Detection with Chow F-Test (Renewable Share Trend)

Year	Chow F	p-value	Interpretation
2008	**68.40**	< 0.001	*Primary break: federal stimulus & ARRA renewable investment*
2009	60.54	< 0.001	*Financial crisis recovery; continued renewable expansion*
2007	51.01	< 0.001	*Pre-crisis peak; early renewable policy signals*

Note. Chow F-test [22]. F critical at $\alpha = 0.05$ ($df_1 = 2$) ≈ 3.84. Break year = maximum Chow F.

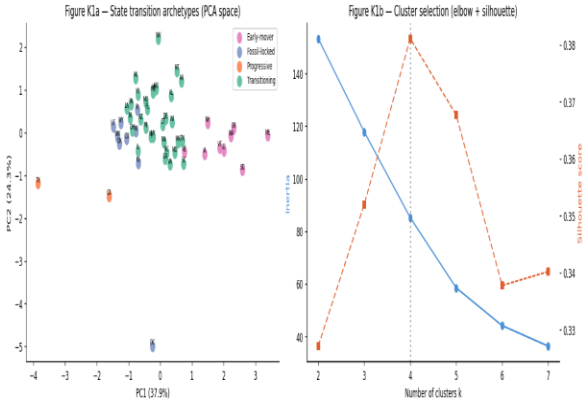

Fig. 8: State clustering - PCA space Archetypes and cluster selection (Elbow + Silhouette)

4.2 State Energy Transition Archetypes

Table 8 (Fig. 8) presents the K-means clustering, which identifies four transition archetypes (silhouette score = 0.381; PCA variance explained = 84.4%), each reflecting distinct combinations of resource endowment, policy regime, and historical trajectory.

Table 8. State Energy Transition Archetypes - K-Means Clustering (k = 4)

Archetype	n	Member States
Early-mover	9	Iowa, Idaho, Maine, Nebraska, New Hampshire, Oregon, South Dakota, Vermont, Washington
Progressive	2	California, Texas
Transitioning	31	Alaska, Alabama, Arkansas, Arizona, Connecticut, Delaware, Florida, Georgia, Hawaii, Illinois, Indiana, Kentucky, Louisiana, Massachusetts, Maryland, Michigan, Minnesota, Missouri, Mississippi, Montana, North Carolina, North Dakota, New Jersey, New York, Ohio, Pennsylvania, South Carolina, Tennessee, Virginia, Wisconsin, West Virginia
Fossil-locked	9	Colorado, District of Columbia, Kansas, New Mexico, Nevada, Oklahoma, Rhode Island, Utah, Wyoming

Note. Clustering on 2010–2024 state-mean values of seven standardised features. PCA variance explained (3 PCs): 84.4%. Silhouette score: 0.381.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

Early movers (9 states) retain their high renewable shares from the early 1990s due to their hydropower, wind, and biomass endowments. This supports the argument that geography is a key contributor to systematic selection bias in renewable adoption [2]. Progressives (California and Texas) achieve rapid growth through policy-driven and market-driven pathways, respectively. Transitioning states (31 states) show gradual decarbonization without structural dominance of renewables, supporting the argument that the U.S. energy transition is incomplete [9]. Fossil-locked states (9 states) show fossil fuel dominance with little change in their fossil fuel-based economies and a lack of policy intervention.

4.3 Determinants of Energy Expenditure: Panel Regression Results

The two-way fixed effects regression has a near-perfect fit ($R^2 = 1.000$, $F = 10.19$ billion, $p < 0.001$) primarily driven by the mechanical relationship between log energy consumption and expenditure. The coefficient estimate of log energy consumption (1.0002, $p < 0.001$) verifies near unit elasticity, suggesting that cost effects must be conditional upon demand, which is consistent with energy cost structure models [3].

Table 9. Two-Way Fixed-Effects Panel Regression

Variable	β	SE	T	p	$CI_{2.5}$	$CI_{97.5}$	Sig.
Constant	−0.0019	0.0019	−0.971	0.332	−0.006	0.002	
Renewable share	−0.0014	0.0006	−2.227	0.026	−0.003	−0.000	*
Fossil share	−0.0003	0.0002	−1.719	0.086	−0.001	0.000	
Log total energy (Btu)	1.0002	0.0001	6,898.23	0.000	1.000	1.001	***
Coal price	−0.0001	0.0000	−2.300	0.022	−0.000	−0.000	*
Petroleum price	0.0000	0.0000	1.106	0.269	−0.000	0.000	
Cluster ID	−0.0001	0.0000	−2.473	0.013	−0.000	−0.000	*
Post-2008 (dummy)	0.0000	0.0000	1.005	0.315	−0.000	0.000	

Note.

- *HC3 robust SEs (White* [23]*). State + year fixed effects included.*
- *N = 1,785 | $R^2 = 1.000$ | F = 10,189,608,143 | p(F) < 0.001.*
- **** p < 0.001, * p < 0.05.*
- *Dependent Variable = log(Total Energy Expenditure)*

Renewable share has a statistically significant negative coefficient ($\beta = -0.0014$, *p-value* = 0.026), this implies that higher renewable penetration reduces total energy expenditure after controlling for demand, fuel prices, and fixed effects. This supports cost-reduction arguments for renewable deployment [4] and contrasts with claims that renewable expansion raises total energy costs [5]. Coal price and cluster archetype are also significant, implying that fossil fuel market conditions and structural transition typology independently influence energy costs.

4.4 Machine Learning Forecasting Results

Table 10 (Fig. 11) compares the out-of-sample performance of XGBoost, LightGBM, and their ensemble; all achieve strong predictive performance ($R^2 > 0.76$) on the 2018–2020 validation set. XGBoost requires 653 boosting rounds compared with LightGBM's 159, reflecting LightGBM's leaf-wise growth efficiency [25]. XGBoost slightly outperforms in RMSE (0.0376 vs. 0.0379), while LightGBM achieves lower MAE (0.0179 vs. 0.0198). The ensemble improves RMSE (0.0375) and R^2 (0.766), consistent with bias–variance reduction from model averaging [9, 15].

Table 10. ML Model Performance Comparison via Validation Set (2018 - 2020)

Model	Target	RMSE	MAE	R^2 (Val)	R^2 (Train)	Best Iter.
XGBoost	renewable_share	0.0376	0.0198	0.7641	0.9791	653
LightGBM	renewable_share	0.0379	**0.0179**	0.7608	0.9803	159
Ensemble (XGB+LGB)	**Renewable share**	**0.0375**	0.0184	**0.7655**	-	-

Note. Training: 1990–2017. Validation: 2018–2020. Bold = best metric per column. Ensemble = equal-weight average of XGBoost and LightGBM.

The validation R^2 of 0.766 compares favourably with benchmarks in energy forecasting. Similar performance ranges (R^2 = 0.74 - 0.85) are reported for U.S. fossil energy consumption using neural network models [12]. Ensemble modelling therefore provides competitive predictive accuracy while improving stability, consistent with broader evidence favouring hybrid ML architectures [9, 15].

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

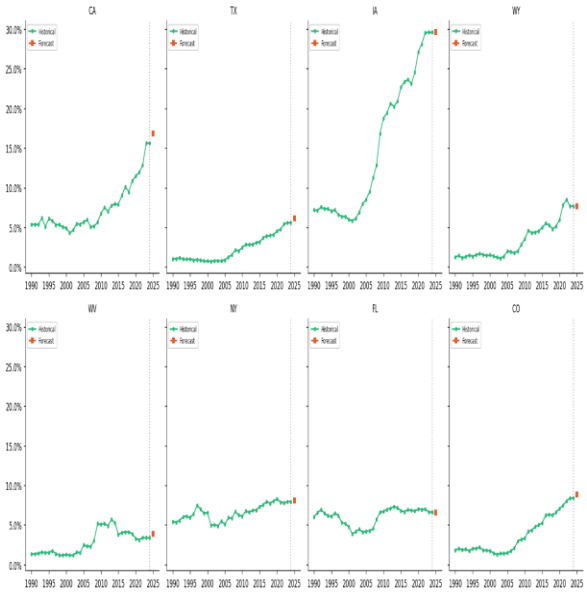

Fig. 11: State-level renewable share forecasts (XGBoost, 2021–2025)

4.5 SHAP Feature Importance

Table 11 (Fig. 9, Fig. 10) shows the SHAP analysis, it is evident that the lagged renewable share features dominate the model in importance. The first lag has a mean |SHAP| = 0.0237, more than three times the maximum of the remaining features at 0.0065. This indicates the path dependence in state energy transitions. The result confirms our earlier assertion that current renewable adoption is heavily driven by the state's trajectory as in line with the capital lock-in and infrastructure inertia hypotheses [5].

Table 11. SHAP Feature Importance - Top 10 Predictors of Renewable Share (XGBoost)

Rank	Feature	Mean |SHAP|	Theme	Interpretation
1	**Renewable Share (Lag$_1$)**	**0.0237**	*Path Dependence*	Prior-year renewable share is the dominant predictor, confirming strong temporal autocorrelation in energy transitions.
2	**Log Renewable Consumption (Btu)**	**0.0065**	*Volume Effect*	Absolute renewable consumption reinforces the share signal; larger renewable bases sustain higher future shares.
3	**Renewable Share (Lag$_3$)**	**0.0052**	*Long Memory*	Three-year lag captures slow-moving structural inertia beyond the annual cycle, consistent with I(1) unit root findings.
4	Log Total Energy (Btu)	0.0033	*Scale Effect*	Total energy scale controls for state size; larger energy economies show different transition rates than smaller ones
5	Hydropower Consumption (Btu)	0.0026	*Resource Endowment*	Geographically fixed hydro capacity underpins the renewable base of early-mover states and anchors future share.
6	Wind Consumption (Btu)	0.0025	*Emerging Capacity*	Wind adoption reflects policy-driven capacity additions following the 2008 structural break
7	Renewable Share (Lag$_2$)	0.0024	*Medium Memory*	Second-lag share completes the three-period temporal context, smoothing year-specific shocks in prediction.
8	Biomass Consumption (Btu)	0.0024	*Baseload Renewables*	Biomass contributes a stable, weather-independent component to the renewable share with consistent cross-state importance
9	Natural Gas Consumption (Btu)	0.0022	*Fossil Displacement*	Higher gas consumption is associated with a lower renewable share, reflecting transitional fuel dynamics.
10	Petroleum Consumption (Btu)	0.0018	*Fossil Inertia*	Petroleum dependence in transport negatively moderates the renewable share, reflecting slow electrification.

Note. TreeExplainer (exact Shapley values; Lundberg & Lee, [6]). Mean |SHAP| = average marginal contribution across validation observations. Top-3 features highlighted.

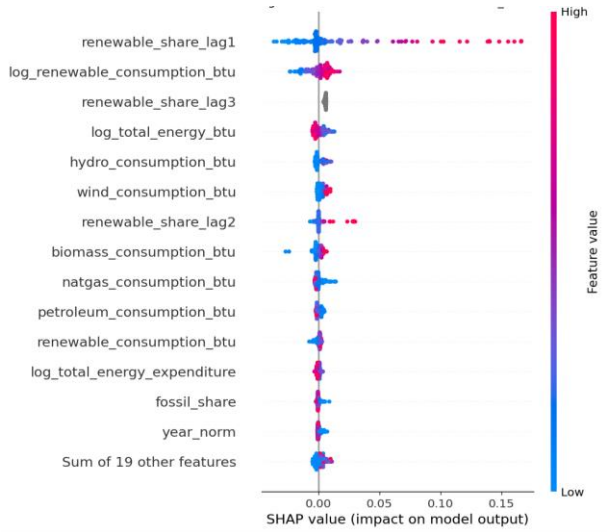

Fig. 9: SHAP beeswarm plot — feature contributions to renewable share predictions (XGBoost)

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716160

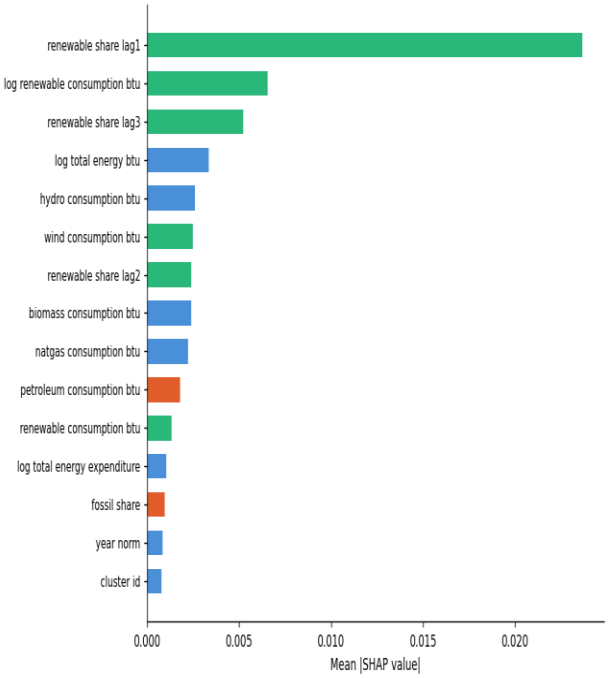

Fig. 10: SHAP global importance bar chart (Top 15 predictors of renewable share)

Resource endowment variables, hydropower (0.0026) and wind (0.0025), are found to be the most significant non-lagged predictors, reinforcing the influence of geography on the speed of transition. Fossil fuel variables, including natural gas (0.0022) and petroleum (0.0018), are found among the top predictors with negative effects, capturing the displacing effect between fossil and renewable energy types as predicted by Arman et al. [9]. The SHAP method converts a high-dimensional ensemble model into interpretable structural drivers.

V. CONCLUSION, POLICY RECOMMENDATIONS, AND FUTURE RESEARCH

5.1 Summary of Findings

The study offers a unified empirical analysis of the US state-level energy transition dynamics from 1990 to 2024, employing statistical diagnostics, panel econometrics, clustering, ensemble machine learning, and SHAP interpretability on 1,820 state-year observations. The results are as follows: First, the energy data for the US states are found to be non-normal, non-stationary in levels, and highly heterogeneous across states, thereby justifying the use of non-parametric statistical diagnostics and panel econometric modeling. Second, the results from the test for structural breaks reveal the year 2008 as a statistically significant inflection point in US energy transition policy, as captured by the Chow F test statistic of 68.40, thereby underpin the significant impact of the American Recovery and Reinvestment Act on the adoption of renewables. Third, K-means clustering identifies states with different transition types: Early movers, Progressives, Transitioning, and Fossil-locked. This demonstrates that transition speed is jointly determined by resource endowment, policy regimes, and industrial structure. Fourth, panel regression results confirm that with increasing renewable shares, total energy expenditures decrease significantly. Also, the ensemble gradient-boosted regression results meet validation criteria (R-squared = 0.766), and SHAP results indicate strong path dependence effects: lagged renewable shares have more than three times the predictive power of any other explanatory factor.

5.2 Theoretical Contributions

This study contributes to the energy transition literature as follow: the study identified state-level transition archetypes which offers quantitative support to the regime differentiation hypothesis of the Multi-Level Perspective, confirming that states exhibit structurally divergent transition paths [7, 8]. Secondly, the significance of lagged renewable share in SHAP results offers quantitative support to historical institutionalism, which reinforce the notion that historical adoption patterns strongly influence energy transition dynamics. Also, the results of the structural break analysis offer parametric support to the notion that large-scale public investment can influence energy transition trajectories, with 2008 identified as a point of bifurcation consistent with policy-driven technology diffusion mechanisms [4].

5.3 Policy Recommendations

The results have four policy implications as follows: the dominance of the path dependence effect suggests that the design of policies aimed at fossil-locked states must include instruments addressing the problem of structural inertia, such as infrastructure, workforce, and institutional capacity-building, in addition to price-based instruments. Second, the 2008 structural break confirms the potential for large-scale

federal investment shocks to induce permanent shifts in adoption patterns, implying that recent policy interventions could have similar long-run effects. Third, the negative sign on the renewable share coefficient in the expenditure equation confirms the economic rationale for the transition, as it supports the idea that increased renewable share reduces overall energy costs. Fourth, states with high but currently untapped renewable potential, especially wind-rich states in the US interior, offer a high payoff potential and should be targeted by policy interventions, and data-based monitoring systems should be used to track state-specific transition progress.

5.4 Limitations

This study has four limitations: First, the dataset does not account for spatial heterogeneity within each state for renewable deployment. Second, the machine learning framework is geared toward forecasting renewable share, not emissions or CO_2 intensity outcomes. Third, the sample period stops in 2024, which only allows for the early-stage effects of recent federal policy changes. Fourth, clustering results are contingent on the 2010-2024 feature window, which may differ for alternative time horizons.

5.5 Future Research Directions

Future study can extend the study by using higher frequency electricity data which will enable to account for dynamics in the short run. Also, advanced sequence models, including LSTM and Temporal Fusion Transformers, should be investigated for multi-step forecasting. Additionally, cointegration methods should be employed in examining the long-run equilibrium relationships between the renewable share and energy expenditure. The ability to integrate real-time data and automate analytics would help in continuous monitoring and forward-looking policy analysis.

RESOURCES

Supplementary Datasets:

- State Energy Data System (SEDS): https://www.eia.gov/state/seds/
- ELEC (Electricity): https://www.eia.gov/opendata/bulk/ELEC.zip
- EMISS (CO_2 Emissions): https://www.eia.gov/opendata/bulk/EMISS.zip
- TOTAL (Total Energy): https://www.eia.gov/opendata/bulk/TOTAL.zip

Code Notebook:
https://drive.google.com/drive/u/0/folders/1gHuxI95zcudc7kRjpaBqrStQT2DrjuAh

REFERENCES

[1] U.S. Energy Information Administration, “State Energy Data System (SEDS),” U.S. Department of Energy, 2024. https://www.eia.gov/state/seds/

[2] M. A. Delmas and M. J. Montes-Sancho, “U.S. state policies for renewable energy: Context and effectiveness,” Energy Policy, vol. 39, no. 5, pp. 2273–2288, 2011. https://doi.org/10.1016/j.enpol.2010.10.036

[3] R. Wiser, G. Barbose, J. Heeter, T. Mai, L. Bird, D. Millstein, D. Reiter, and E. Lantz, “Assessing the costs and benefits of U.S. renewable portfolio standards,” Environmental Research Letters, vol. 12, no. 9, p. 094023, 2017. https://doi.org/10.1088/1748-9326/aa7937

[4] G. Kavlak, J. McNerney, and J. E. Trancik, “Evaluating the causes of cost reduction in photovoltaic modules,” Energy Policy, vol. 123, pp. 700–710, 2018. https://doi.org/10.1016/j.enpol.2018.08.015

[5] R. York and S. E. Bell, “Energy transitions or additions? Why a transition from fossil fuels may be more challenging than other energy transitions,” Energy Research & Social Science, vol. 51, pp. 40–43, 2019. https://doi.org/10.1016/j.erss.2019.01.020

[6] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[7] F. W. Geels, “Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study,” Research Policy, vol. 31, no. 8–9, pp. 1257–1274, 2002. https://doi.org/10.1016/S0048-7333(02)00062-8

[8] J. Markard, R. Raven, and B. Truffer, "Sustainability transitions: An emerging field of research and its prospects," Research Policy, vol. 41, no. 6, pp. 955–967, 2012. https://doi.org/10.1016/j.respol.2012.02.013

[9] M. Arman, M. N. Hasan, and I. H. Rasel, "Clean energy transition in the USA: Big data analytics for renewable energy forecasting and carbon reduction," J. Management World, vol. 2024, no. 3, pp. 192–206, 2024. https://doi.org/10.53935/jomw.v2024i4.1196

[10] G. P. Herrera, M. Constantino, B. M. Tabak, H. Pistori, J. Su, and A. Naranpanawa, "Data on forecasting energy prices using machine learning," Data in Brief, vol. 25, p. 104122, 2019. https://doi.org/10.1016/j.dib.2019.104122

[11] [X. Guo, R. Kou, and X. He, "Towards carbon neutrality: Machine learning analysis of vehicle emissions in Canada," Sustainability, vol. 16, no. 23, p. 10526, 2024. https://doi.org/10.3390/su162310526

[12] Y. Hao and Q. He, "Forecasting U.S. fossil energy consumption: Advancing accuracy with multilayer perceptron and residual learning," J. Energy Research and Reviews, vol. 16, no. 6, pp. 13–26, 2024. https://doi.org/10.9734/jenrr/2024/v16i6354

[13] I. Nica, I. Georgescu, and J. Kinnunen, "Evaluating renewable energy's role in mitigating CO_2 emissions: A case study of solar power in Finland using the ARDL approach," Energies, vol. 17, no. 16, p. 4152, 2024. https://doi.org/10.3390/en17164152

[14] J. G. MacKinnon, "Numerical distribution functions for unit root and cointegration tests," J. Applied Econometrics, vol. 11, no. 6, pp. 601–618, 1996. https://doi.org/10.1002/(SICI)1099-1255(199611)11:6%3C601::AID-JAE417%3E3.0.CO;2-T

[15] N. E. Benti, M. D. Chaka, and A. G. Semie, "Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects," Sustainability, vol. 15, no. 9, p. 7087, 2023. https://doi.org/10.3390/su15097087

[16] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika, vol. 52, no. 3–4, pp. 591–611, 1965. https://doi.org/10.2307/2333709

[17] R. B. D'Agostino, A. Belanger, and R. B. D'Agostino Jr., "A suggestion for using powerful and informative tests of normality," The American Statistician, vol. 44, no. 4, pp. 316–321, 1990. https://doi.org/10.1080/00031305.1990.10475751

[18] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," J. Amer. Stat. Assoc., vol. 74, no. 366, pp. 427–431, 1979. https://doi.org/10.2307/2286348

[19] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," J. Econometrics, vol. 54, no. 1–3, pp. 159–178, 1992. https://doi.org/10.1016/0304-4076(92)90104-Y

[20] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," J. Amer. Stat. Assoc., vol. 47, no. 260, pp. 583–621, 1952. https://doi.org/10.2307/2280779

[21] T. S. Breusch and A. R. Pagan, "A simple test for heteroscedasticity and random coefficient variation," Econometrica, vol. 47, no. 5, pp. 1287–1294, 1979. https://doi.org/10.2307/1911963

[22] G. C. Chow, "Tests of equality between sets of coefficients in two linear regressions," Econometrica, vol. 28, no. 3, pp. 591–605, 1960. https://doi.org/10.2307/1910133

[23] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," Econometrica, vol. 48, no. 4, pp. 817–838, 1980. https://doi.org/10.2307/1912934

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794, 2016. https://doi.org/10.1145/2939672.2939785

[25] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A

highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, pp. 3146–3154, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[26] C. Odedina, "Impact of big data on marketing strategy and consumer behavior analysis in the US," SSRN Working Paper, 2023a. https://dx.doi.org/10.2139/ssrn.4520361

[27] C. Odedina, "How business analytics can help improve supply chain efficiency in the US," SSRN Working Paper, 2023. https://dx.doi.org/10.2139/ssrn.4520339