Received: 3 October 2024 | Revised: 31 December 2024 | Accepted: 3 April 2025 | Published online: 23 April 2025

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2025, Vol. 00(00) 1–13
DOI: 10.47852/bonviewJDSIS52024482

BON VIEW PUBLISHING

Forecasting CO_2 Emission in the US Using Regression Models

Kamil Samara[1,*], Yunhwan Jeong[1] and Thomas H. Beaupre[1]

[1]*Department of Computer Science, University of Wisconsin-Parkside, USA*

Abstract: Our paper examines CO_2 emissions resulting from energy consumption across key sectors: commercial, industrial, transportation, residential, and electrical. We emphasize predicting CO_2 output associated with diverse fuel types used within these sectors. Leveraging extensive datasets from the Energy Information Administration and the Environmental Protection Agency, we utilize energy consumption data, measured in Trillion Btu units, to build predictive models that forecast future CO_2 emissions. By analyzing the correlation between energy use and CO_2 output, our study provides critical insights into the environmental impact of different fuel sources. We incorporate the polluting factors of each energy type to estimate their individual contributions to overall emissions. These models empower stakeholders to make informed decisions regarding energy use, fostering proactive environmental control. Our work advocates for sustainable energy practices by identifying opportunities for reducing CO_2 emissions while emphasizing the importance of mindful consumption. The findings encourage transitioning to environmentally friendly energy alternatives and promote collective action toward mitigating climate change. Ultimately, this research underscores the need for balancing energy demands with environmental stewardship, aiming to inspire practices that contribute to a greener, more sustainable future for current and future generations.

Keywords: carbon dioxide emission, greenhouse gas emission, energy consumption, environmental prediction, supervised learning, regression models, XGBoost model

1. Introduction

1.1. Background

The motivation for conducting this research comes from several reasons. First, carbon dioxide is a major greenhouse gas that traps heat in the Earth's atmosphere, contributing to global warming and climate change. By understanding and monitoring CO_2 emissions, we can better predict and mitigate their impact on the environment. Second, reducing CO_2 emissions is essential for achieving climate goals [1]. The U.N. climate science panel has emphasized the need to cut CO_2 emissions significantly by 2030 to limit global warming to 1.5°C and avoid the worst impacts of climate change [2]. Finally, studying CO_2 emissions helps inform policy decisions and technological advancements. By identifying the major sources of CO_2 emissions, we can develop targeted strategies to reduce them, such as transitioning to renewable energy sources and improving energy efficiency in industries [3, 4].

Addressing climate change has taken center stage in recent years, and research activities are pointing towards understanding and reducing CO_2 emissions. Energy consumption is considered one of the serious facilitators of economic activity. The high level of energy consumption contributes a great deal to the greenhouse effect, mainly through CO_2 emissions. Commercial, industrial, transportation, residential, and electrical sectors have different energy consumption patterns depending on the prevailing fuel types. The comprehensive crux, in understanding the trend of energy consumption and its highly significant relation to CO_2 emissions, is crucial at a time when most societies are struggling hard to achieve the goal of sustainable development. Our effort of forecasting the rate of CO_2 emission, based on this perspective, is grounded on extensive datasets from the Energy Information Administration (EIA) and the United States Environmental Protection Agency (EPA). We will be able to give insight into the future dynamics of CO_2 emissions by analyzing historical data emanating from a wide array of sectors, thereby advocating environmentally friendly energy consumptions across all domains [5–7].

This study aims to explore the accuracy and effectiveness of machine learning regression models, such as spline regression and XGBoost, in forecasting CO_2 emissions across various US sectors, including commercial, industrial, transportation, residential, and electrical. It investigates the primary factors driving CO_2 emissions, particularly the correlation between energy consumption patterns and fuel types like coal, petroleum, natural gas, and biomass. By examining sector-specific contributions to emissions and the impact of transitioning to cleaner energy sources, the research highlights trends and their implications for achieving climate goals, such as net-zero emissions by 2050. Additionally, the study assesses the role of advanced predictive models in improving policy decisions and promoting sustainable energy practices through targeted emission reduction strategies.

*Corresponding author: Kamil Samara, Department of Computer Science, University of Wisconsin-Parkside, USA. Email: samara@uwp.edu

1.2. Plan

We scraped monthly data from the Energy Information Agency and EPA. The data are collected by sectors—transportation, residential, industrial, and commercial. For each sector, energy consumption and CO_2 emission data that are sequential from January 1973 to August 2023. Also, the energy consumption and CO_2 emission data are categorized by fuel type such as coal, natural gas, petroleum, and biomass fuel. We will focus on finding the coefficient of CO_2 emission based on energy consumption. The coefficient shows which sector contaminates the environment mostly and how each fuel impacts on the environment. To interpret the data more conveniently, we are going to visualize the collected data using trendline and scatter plot.

2. Literature Review

The domain of CO_2 emission prediction has been extensively studied. In this literature review, we will examine several significant contributions in the field, highlighting their relevance to our research.

The entitled "A Novel Grey Verhulst Model and its Application in Forecasting CO_2 Emissions," by Mingyu Tong, Huiming Duan, and Leiyuhang, proposes an improved version of the Grey Verhulst model applicable in the forecast of carbon dioxide emissions. According to the authors, the traditional Grey Verhulst model has the weakness of being primarily applicable to saturated S-type data, while suggesting a new model that lessens such dependence. Then, they applied the optimized Verhulst model to forecast the CO_2 emission in China and Russia and compared its result with that from the traditional Verhulst and ARIMA models. The result indicated that the new model improved the prediction accuracy by more than 10% compared to the traditional model, proving to be better than the ARIMA model, which tends to underestimate such emissions. The conclusions from this study are that the new model has better capabilities of simulation and prediction, thus might be useful to inform government policy on energy conservation and emission reduction [8].

In the article "A Study on the CO_2 Emissions of Mainland China Using Deep Learning Models," Lexing Huang examines carbon emissions in China from 1997 to 2017 and will make projections into what they could look like from 2030 to 2060, using various deep learning and statistical models. In the paper, RNNs, CNNs, LSTM networks, and ARIMA models have been combined to predict the trend of emission. The present study demonstrated that economic and demographic variables were strongly associated with emission variables; their highest shares originated from provinces like Shandong, Jiangsu, and Hebei. These results demonstrate the efficiency of RNNs with a mean average percentage error of 2.56% against 11.23% obtained from the ARIMA model. The study concludes that in order for China to meet its carbon peak in 2030 and carbon neutrality in 2060, the country is in dire need of technological innovation and policy measures [9].

The paper "Evolving Dynamic Bayesian Networks for CO_2 Emissions Forecasting in Multi-Source Power Generation Systems," by Talysson M. O. Santos, Michel Bessani, and Ivan N. Da Silva, is aimed at proposing a methodology for the exact forecast of CO_2 emissions with evolving discrete DBNs. The obtained model adapts continuously to new data, adjusting the network structure based on the frequency that the data occurs, hence increasing the robustness and accuracy of making such predictions. The concept was applied to real data for power generation systems of Belgium, Germany, Portugal, and Spain, outperforming conventional DBN, ANN, and XGBoost. The paper highlights the strengths of EDBN regarding handling variability and uncertainty typical of multisource energetic systems, relevant for computational efficiency in real-time CO_2 emissions forecasting at the heart of energy management for sustainability [10].

The paper "Forecasting of CO_2 Emission in Iran Based on Time Series and Regression Analysis" by Seyed Mohsen Hosseini et al. explores the emission of the gas in Iran and, employing a time series and regression analysis, forecasts future trends according to two conditions: business as usual and the Sixth Development Plan. The research forecasts the emissions until 2030 using the MLR and MPR models. The results show that Iran, under a BAU scenario, may not achieve its pledge under the Paris Agreement due to persistent high energy intensity and insufficient policy measures in place. However, under the full implementation of SDP, with its ambitious targets on renewable energy growth, GDP increase, and energy efficiency, the emissions could go down drastically. The authors have brought to the fore the need for pragmatic, attainable policies towards these environmental goals [11].

The article "Forecasting the Path of U.S. CO_2 Emissions Using State-Level Information" by Maximilian Auffhammer and Ralf Steinhauser explores the accuracy of various reduced-form models for predicting CO_2 emissions in the U.S. Using a panel data set of state-level emissions from 1960 to 2001, they compare the performance of existing models against a large universe of potential models. The study finds that many models commonly used in literature perform poorly out-of-sample compared to the best model identified through their comprehensive search. The authors emphasize the importance of selecting models based on out-of-sample loss measures over aggregate emissions rather than in-sample fit or per capita emissions. Their best model forecasts U.S. emissions for 2011 to be significantly lower than some standard predictions, highlighting potential biases in existing emissions forecasts. The paper underscores the value of using disaggregated state-level data to improve forecasting accuracy and suggests that better forecasts can lead to more informed policy decisions regarding climate change mitigation [12].

The paper "Forecasting CO_2 Emissions in Saudi Arabia Using Artificial Neural Network, Holt-Winters Exponential Smoothing, and Autoregressive Integrated Moving Average Models" reviewed the performance of various models on the forecast of CO_2 emission in Saudi Arabia. Based on data from 1960 to 2014, this study has integrated artificial neural network ANN, Holt-Winters exponential smoothing H-W, and autoregressive integrated moving average ARIMA in modeling for the forecast of the emission. The ARIMA model of order (2,1,2) was found to be most appropriate for the forecasting of CO_2 emission, showing much better accuracy compared to the ANN and H-W models. These results reflect that CO_2 emission would vary in Saudi Arabia by 2025. The findings of this study will help both the researchers and government agencies during strategic planning and perceive the tendency of the emission, laying great emphasis on how model performance could vary with characteristics of data [13].

The "Forecasting of CO_2 Emissions, Renewable Energy Consumption and Economic Growth in Vietnam using Grey Models" article applies gray prediction models, GM(1,1), and DGM(1,1), by Hong-Xuyen Thi Ho in forecasting the CO_2 emissions, renewable energy use, and GDP of the Vietnamese economy in the period 2010–2019. It has emerged from the study

that these variables play a central role in shaping energy efficiency, economic growth, and climate change in Vietnam. The results forecast that in 2019, CO_2 emissions will see an increase of 3% while renewable energy is consumed minimally, and GDP will increase by 5% over 2010. From this perspective, the DGM(1,1) model performed better than GM(1,1); hence, it is more suitable for policymakers to reach a three-in-one objective in terms of energy efficiency, economic development, and environmental protection [14].

In "Smart Forecasting: Harnessing Machine Learning for Accurate CO_2 Emission Predictions," A. Hency Juliet, P. Malathi, and N. Legapriyadharshini discuss how to use the ML models in order to correctly estimate the CO_2 emission using real data. In this work, the performances of four different ML classifiers, namely LR, GPR, MLP, and SMOreg, will be evaluated based on a dataset available in Kaggle. The dataset consists of 935 instances with 12 attributes that entail the specifications of vehicles and their emissions. The best performance, in terms of the accuracy of prediction, can be provided by the SMOreg classifier as supported by metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), correlation coefficient, and root relative squared error (RRSE). The study underlined the capability of ML algorithms in order to improve the CO_2 emission forecast that, on its turn, may be contributory in mitigating climate change. It is observed that SMOreg outperforms other models; hence, the tool at this early stage of detection and strategic decision-making in carbon mitigation is very valuable [15].

The "Thailand Carbon Dioxide Emission Forecasting using Stacked LSTM-Based Prediction Model" conducted by Yamin Thwe, Dechrit Maneetham, Padma Nyoman Crisnapati, and Myo Min Aung expounds on the application of a stacked long short-term memory model in forecasting Thailand's CO_2 emissions from 1987 to 2021. The study also emphasizes the high CO_2 emission by Thailand and the need for accurate forecasting relevant to assessing and improving environmental policy. These experiments also present the performance of the LSTM model in comparison to more traditional models, namely, VAR and ARIMA, outperforming them with a lower RMSE, MAE, and MAPE result. Hence, the result shows that the LSTM model provides the more accurate forecast, and therefore it is represented further using TensorFlow.js in a web application. Advanced machine learning techniques predominantly stand out in this study as the future of improving CO_2 emission predictions and providing substantial information on climate change mitigation [16].

The paper "A Comparative Study of Statistical and Machine Learning Models on Carbon Dioxide Emissions Prediction of China" reviewed the performance of different models capable of the forecast of daily CO_2 emissions in China. These have considered three statistical models—GM, ARIMA, and SARIMAX—and three machine learning models—ANN, Random Forest, and LSTM—using univariate time series data from January 2020 to September 2022. The performances of the selected models have been compared based on MSE, RMSE, MAE, MAPE, and R^2. This study finds that the machine learning models outperform the statistical models, and the best performance for all metrics is LSTM; hence, it is the most suitable for near real-time daily CO_2 emission prediction. The paper then concludes with some policy recommendations toward cutting emissions in China, underlining how the accurate short-term prediction of emissions will be necessary for timely policy adjustments [17].

The paper, "Carbon Footprint Monitoring System Using Machine Learning and Deep Learning Techniques" reviewed the use of IoT devices and multiple machine learning models that might support a GHG monitoring and prediction system in real time. Data collected from sensors across various locations that report measured gases such as CO_2, CH_4, and N_2O are relayed to the system. The collected data was then preprocessed and analyzed to run the Birch clustering algorithm and the LSTM model for time series prediction. It shall help the system bring insights to organizations in tracking and reducing their carbon footprint. The Birch did improve in accuracy to about 72%. The LSTM model scored an accuracy of 60%. The authors have also pointed out the possible contribution of such a system to businesses in their effort to cut down emissions through real-time actionable data [18].

The paper "The United States Energy Consumption and Carbon Dioxide Emissions: A Comprehensive Forecast Using a Regression Model" explores in detail the relationship between energy consumption and CO_2 emissions in the U.S., derived mainly from fossil fuels, and also provides a forecast of future trends up until the year 2050. Using regression models, the study examines CO_2 emissions for the main sectors—transportation, electricity generation, and industry—and forecasts emission savings for different states. These results are indicative that, though the majority of states and especially the energy sector will see a decline in emissions, there should be continued effort if the U.S. is to meet its ambition of net-zero emissions by 2050. The study highlights the role that can be played by renewable energy, electric vehicles, and carbon capture and storage topical technologies in mitigating future emissions [19].

This study builds on existing research by providing a comprehensive evaluation of CO_2 emissions forecasting using regression models, with a focus on sector-specific contributions and fuel type impacts. Unlike earlier works, such as Tong et al.'s Grey Verhulst model for forecasting emissions or Auffhammer and Steinhauser's use of state-level data, this research integrates detailed datasets from the EIA and EPA to analyze sectoral and temporal energy consumption patterns in the United States. By employing advanced machine learning models like XGBoost and spline regression, it captures complex, non-linear trends more effectively than traditional models, addressing gaps highlighted in studies such as those by Hosseini et al. and Thwe et al. Furthermore, this research offers actionable insights into the shift towards cleaner energy sources, echoing trends noted in the transition studies by Li and Zhang, while advancing the policy relevance of emission forecasting. Ultimately, this study not only reaffirms findings from the literature on the importance of data-driven decision-making in climate policy but also contributes novel methodologies and sectoral analyses that enhance the predictive accuracy and interpretability of CO_2 emission forecasts.

3. Proposed Work

The proposed work will develop and evaluate predictive models to forecast multi-sector energy consumption-based CO_2 emissions. We further intend to integrate data of different fuel types to analyze their respective impacts on CO_2 emission variations. The final result would be to suggest some practical implications for individuals in making better decisions towards sustainable energy utilization.

3.1. Models

We employed several machine learning models to predict CO_2 emissions. Linear regression was used as a baseline model for capturing linear relationships. Polynomial regression was

employed to capture higher-order relationships, while spline transformation regression was used for modeling intricate non-linear patterns. XGBoost regression, known for its robustness in handling complex interactions, was also utilized. Additionally, Lasso regression was implemented to promote sparsity and interpretability in the model coefficients.

3.2. Data

The primary data sources for our research are extensive datasets provided by the EIA and the United States EPA. These datasets encompass detailed information on energy consumption and CO_2 emissions across various sectors, including commercial, industrial, transportation, residential, and electrical domains. The data covers multiple fuel types, such as coal, natural gas, petroleum, geothermal, solar, and biomass, measured in Trillion Btu units. This comprehensive dataset provides a robust foundation for our analysis and modeling efforts.

3.3. Data preprocessing

The preprocessing of the datasets in this study focuses on the cleanliness, consistency, and analysis readiness of the data. Datasets have been loaded into pandas DataFrames, where missing or non-numeric values, such as "Not Available" or "No Data Reported," are standardized as pd.NA. Missing data were contextually imputed; for instance, zeros represent absent emissions or energy consumption, while means are applied for balance in features such as fuel consumption. For example, coal consumption in recent years is missing, this is due to the consumption being near zero for some sectors. Key transformations included extracting the Year from the Month column and removing features that were irrelevant. Datasets were combined where needed, and non-numeric entries were coerced to NaN for consistency. This was to provide a robust base for the analysis providing easy access for multiple analysis while preserving data integrity.

3.4. What we did

In this project, we have been working on the implementation of various machine learning models with the purpose of analyzing and predicting energy consumption along with carbon dioxide emissions among different sectors. Our work started with a baseline model of linear regression, which was used to discern the linear relationships between the year and other features. To capture less smooth, non-linear trends, we started by applying polynomial regression that let us model more complex relationships by adding polynomial features. We also used the spline transformation in an effort to obtain a more flexible fit by estimating piecewise polynomials to capture non-linear patterns smoothly. Another algorithm employed was the robust gradient boosting XGBoost regressor, which would make accurate predictions by leveraging the collective strength of the weak learners. Lasso regression further allowed feature selection and dimensionality reduction, hence improving model interpretability and generalization. The performance metrics to be evaluated included MAE, MSE, RMSE, and R-squared score, all backed by cross-validation in ensuring model stability and reliability. Visualizations such as scatter plots, line plots, regression plots, and residual plots further elucidated model insights and trends. We thus sought, with this wide approach, to make an important contribution: first, to the understanding of energy consumption dynamics and, secondly, to environmental impacts.

4. Simulation and Results

4.1. Tools

In this project, a variety of tools were employed to ensure comprehensive data analysis and model development. Python was the primary programming language due to its versatility and extensive library support for data analysis, machine learning, and visualization. Pandas and NumPy were fundamental for data manipulation, cleaning, and numerical computations. For machine learning tasks, Scikit-Learn provided essential algorithms and model evaluation techniques, while XGBoost was used to build an optimized gradient boosting model capable of capturing complex patterns in the data. Visualization was crucial for analyzing and presenting results; hence, Matplotlib and Seaborn were utilized to create detailed and informative plots. Statsmodels played a key role in implementing spline transformations and conducting in-depth statistical analyses. The entire development process was carried out using a mix of Google Colab and Visual Studio Code. Git ensured robust version control and collaboration, while Microsoft Excel was used for initial data exploration and quick manual inspections of data trends and anomalies.

XGBoost, short for Extreme Gradient Boosting, is a powerful and popular machine learning algorithm used for supervised learning tasks such as classification and regression. It builds a predictive model by combining the predictions of multiple weak models, typically decision trees, in an iterative manner. This ensemble learning technique improves model accuracy and performance by focusing on correcting the errors of previous models in the sequence [20].

4.2. Results

4.2.1. Commercial sector

The commercial sector includes office buildings, retail spaces, warehouses, and other business-related properties. This sector is crucial for understanding economic activity and energy usage trends, as it reflects the dynamics of business operations and their impact on energy consumption.

As seen in Figure 1, the spline transformation model provided a more flexible fit by using piecewise polynomials. This method effectively captured non-linear trends and abrupt changes in the data, as seen in the smoother curves that follow the data points closely. This model offered a balance between flexibility and overfitting, making it suitable for capturing complex energy consumption patterns. As seen with a variety of the energy types there are extreme cases of variance.

The line plot in Figure 2 shows the trends in energy consumption over time for different fuel types. Natural gas and electricity sales to customers exhibit a steady increase, while coal and other consumption sources show a declining trend. This visualization highlights the shift towards cleaner energy sources over the years, while still demonstrating our heavy reliance on petroleum.

The predictions shown in Figure 3 showcases the predictive capabilities of the spline transformation model. The curves in the plots follow the data points closely, indicating the model's ability to capture non-linear patterns and provide accurate forecasts for future data points without overfitting.

The analysis of the commercial sector reveals several important trends and insights. The spline transformation model offered superior predictive performance by capturing complex patterns. The shift towards natural gas and electricity, as shown in the line plots,

Figure 1
Spline transformation using commercial sector energy consumption data

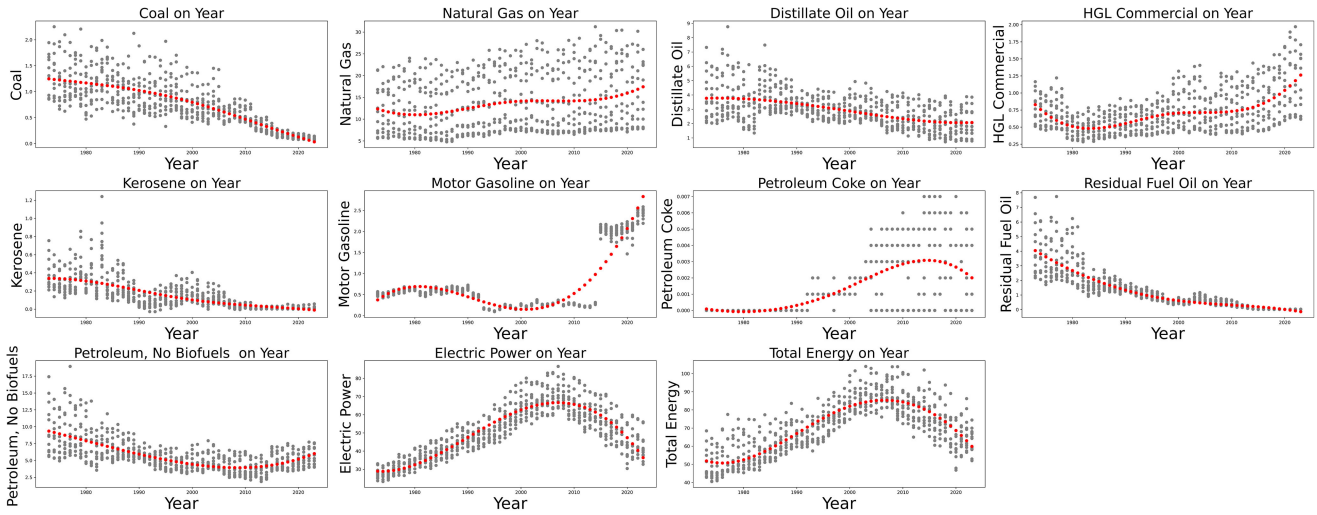
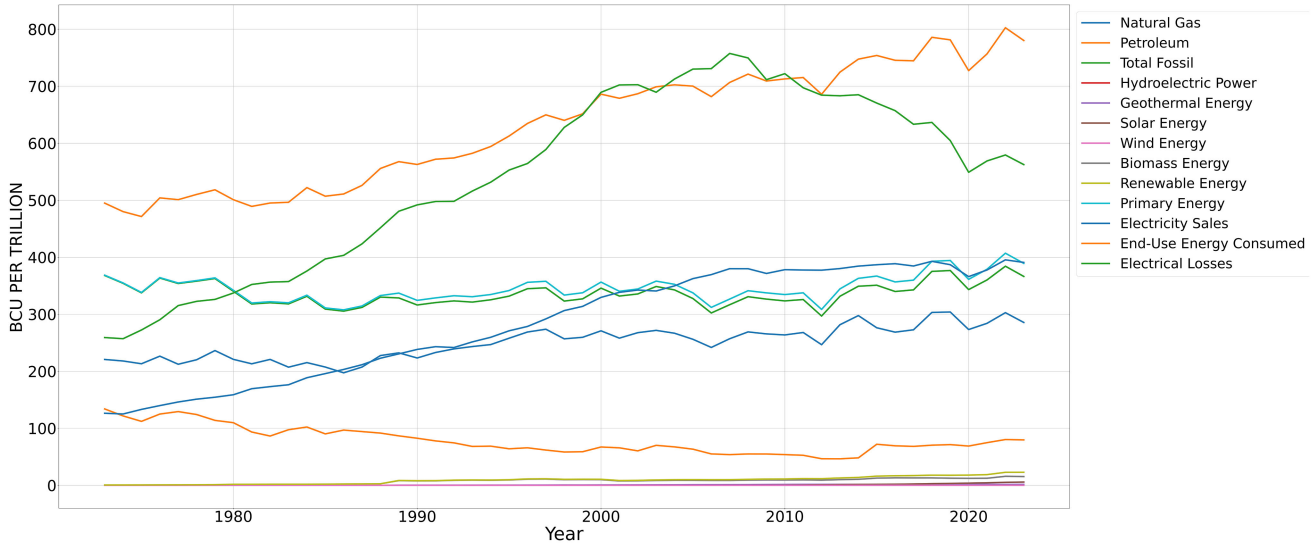


Figure 2
Energy consumption trends in the commercial sector



indicates a move towards cleaner energy sources. The residual and RMSE plots highlight areas where the models perform well and where further refinement is needed. Overall, the commercial sector's energy consumption patterns reflect broader economic and environmental trends, emphasizing the importance of continuous monitoring and model improvement to enhance predictive accuracy.

4.2.2. Industrial sector

The industrial sector includes manufacturing plants, factories, and other facilities where raw materials are processed into finished goods. This sector is a significant energy consumer due to its energy-intensive processes, making it essential to understand its energy consumption patterns and identify opportunities for efficiency improvements.

As shown in Figure 4, the spline transformation model effectively handled the non-linear trends in the industrial sector data. By using piecewise polynomials, this model captured abrupt changes and complex patterns in energy consumption, resulting in smoother curves that align closely with actual data points. This method's flexibility, while avoiding overfitting, made it ideal for modeling industrial energy consumption patterns.

The line plot in Figure 5 illustrates trends in energy consumption for various fuel types over time. Notably, petroleum products and natural gas have shown significant usage, with petroleum demonstrating higher variability and a general decline in recent years. This visualization underscores the sector's reliance on fossil fuels, while the consistent use of natural gas suggests a shift towards cleaner energy sources. The minimal contribution from renewables indicates an area for potential growth in sustainability efforts.

Figure 3
Commercial sector energy consumption predictions

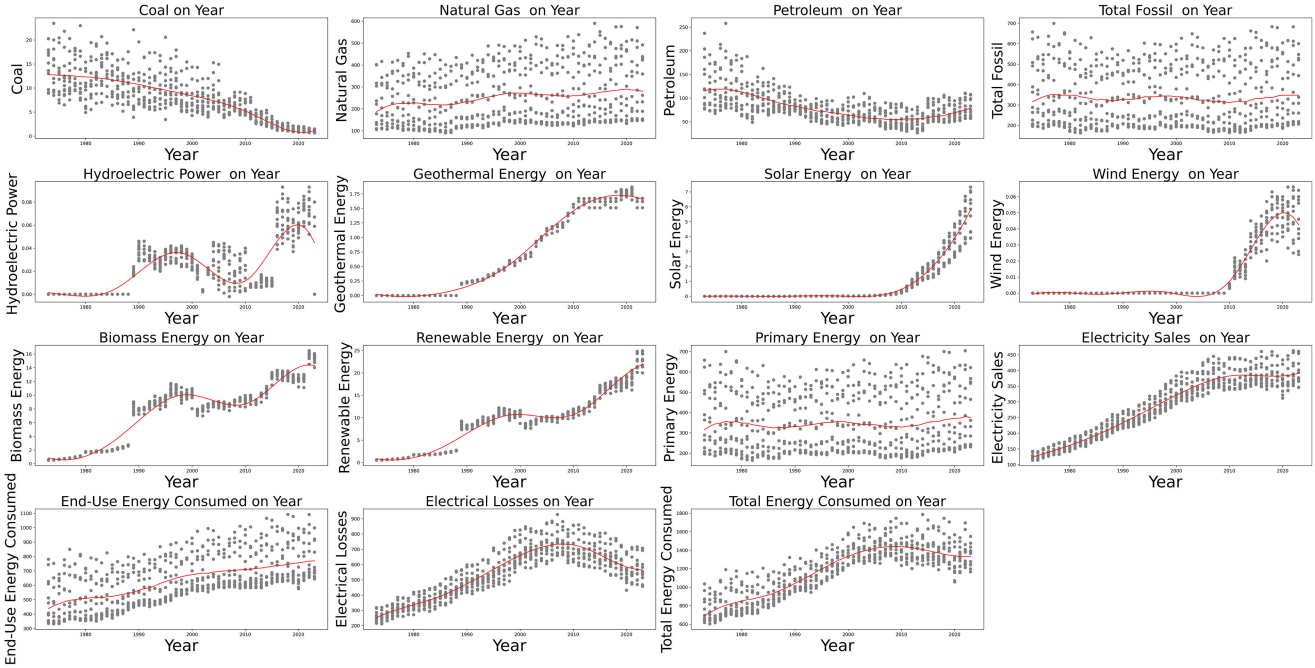


Figure 4
Spline transformation using industrial sector energy consumption data

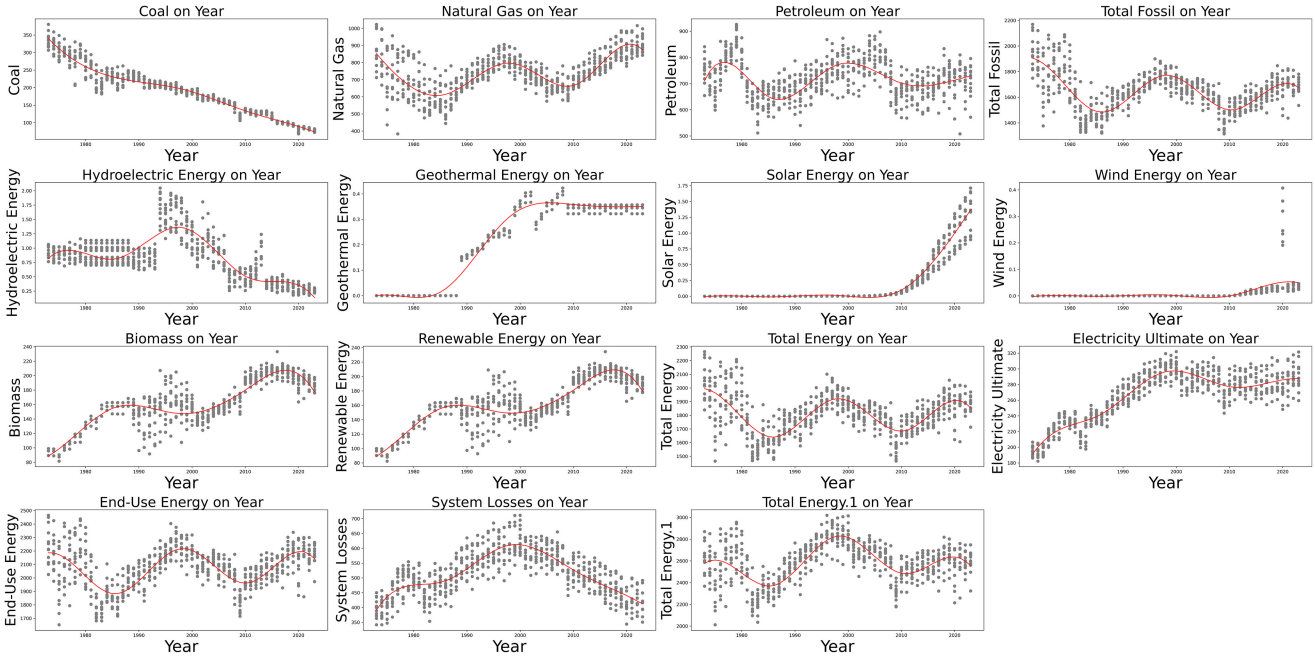


Figure 6 shows the XGBoost model's predictive capabilities with the actual versus predicted plot and residual plot. The actual versus predicted plot shows a strong correlation, with the predictions closely matching the actual values, highlighting the model's effectiveness. The residual plot further illustrates the model's performance, indicating areas where the predictions deviate from the actual values. These insights are crucial for refining the model and improving its accuracy in predicting industrial energy consumption.

The prediction shown in Figure 7 demonstrates the predictive power of the model. The curves closely follow the actual data points, indicating the model's ability to capture complex,

Figure 5
Energy consumption trends in the industrial sector

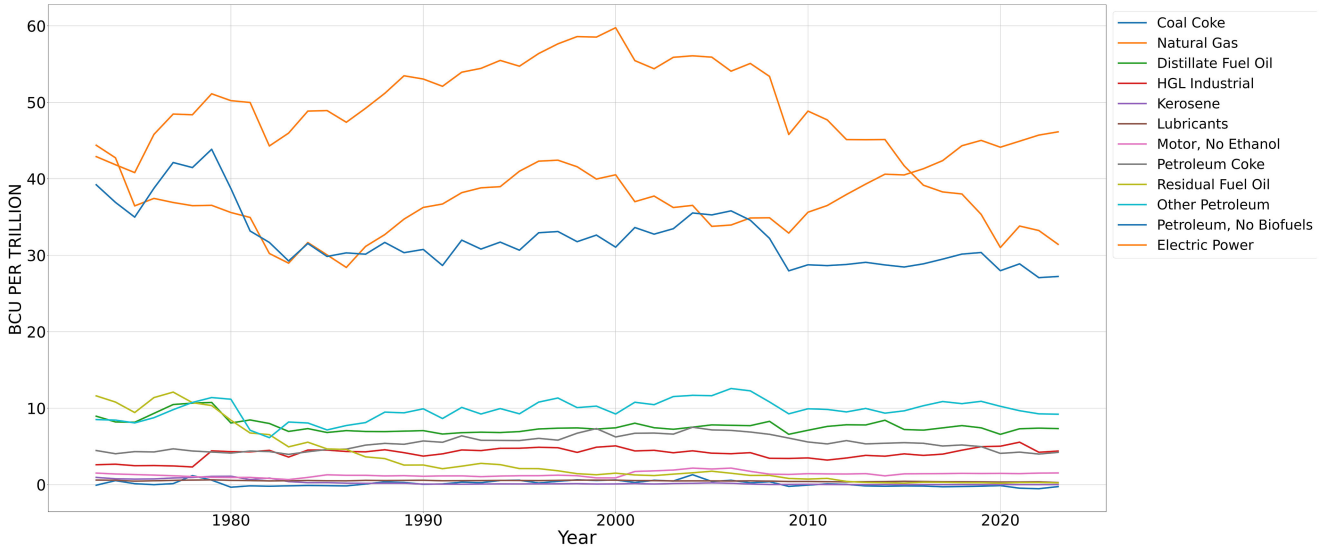
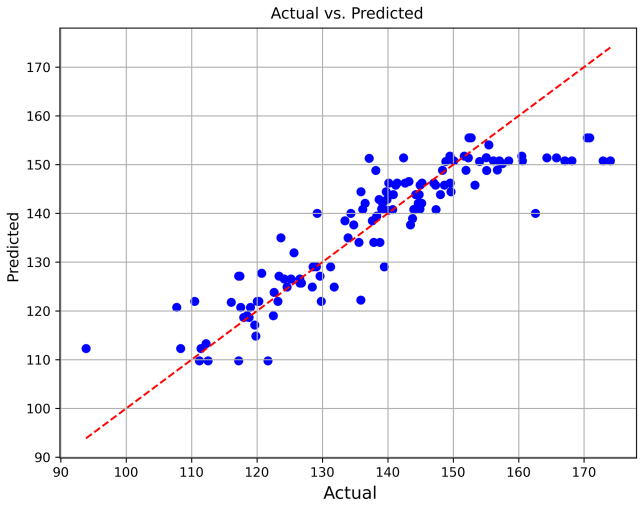


Figure 6
Industrial sector energy consumption prediction using XGBoost



non-linear consumption patterns and provide accurate forecasts. This capability is crucial for making reliable predictions in the industrial sector, where energy usage can fluctuate significantly due to varying production levels and the differing energy sources.

The industrial sector analysis gives a number of important messages. The spline transformation model did an excellent job in capturing the peculiar consumption pattern typical in this industry. The main dependence in accordance with the line plots consists of petroleum and natural gas, with a noticeable decline in petroleum usage over time, hence showing the shift towards cleaner fuels. The XGBoost plots provide the opportunity to identify points where the performance of the models should be further improved. Therefore, in the industrial sector, patterns of energy use embody broader economic and environmental trends that call for sustained monitoring and model refinement to improve predictive accuracy and inform energy efficiency initiatives.

4.2.3. Transportation sector

The transportation sector includes various modes of transport such as cars, trucks, airplanes, and ships. Energy consumption in this sector is driven by fuel usage and is a critical component in understanding overall energy demand and CO_2 emissions.

Figure 8 shows how the spline transformation model fits the data using piecewise polynomials. In this respect, there is a capturing of non-linear trends and sudden changes; thus, yielding smoother curves that trace the actual data points. The flexibility of the model along with the capability to avoid overfitting even makes it appropriate for modeling elaborately the energy consumption patterns in the transport sector. Some key observations that can be clearly drawn from Figure 8 include high dominance by motor gasoline and gradual rise in natural gas usage within the time dimension.

The line plot in Figure 9 reveals trends in energy consumption across different fuel types over time. Motor gasoline remains the dominant fuel, showing a steady increase, while aviation gasoline and distillate fuel oil exhibit fluctuations. The usage of natural gas has significantly increased, reflecting a shift towards cleaner energy sources. This visualization highlights the heavy reliance on petroleum products and the gradual adoption of alternative fuels in the transportation sector. Attention should also be drawn to covid-19 years as a substantial dip in consumption is observable.

Figure 10 shows the XGBoost model's strong predictive capabilities. The actual vs. predicted plot shows that the model closely follows the actual data points, indicating its ability to capture non-linear patterns and provide accurate forecasts.

The transportation sector analysis highlights several significant trends and insights. The spline transformation model excelled in capturing complex consumption patterns, with notable increases in natural gas and fluctuations in motor gasoline usage. Line plots indicate a heavy reliance on petroleum products, with a notable increase in natural gas consumption over time.

The XGBoost model provided strong predictive performance, although some refinement may be needed for higher values. Overall, the energy consumption patterns in the transportation sector reflect economic and global trends, continuous monitoring will continue to enhance the model's predictive capabilities.

Figure 7
Industrial sector energy consumption predictions

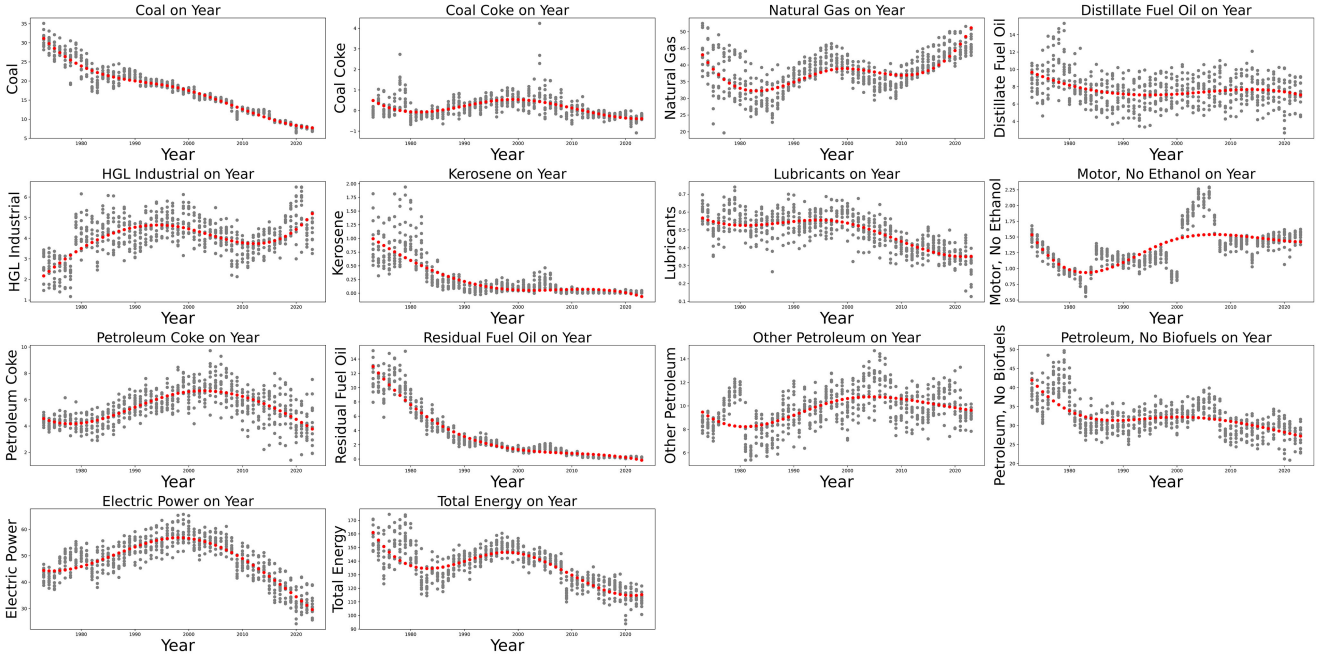
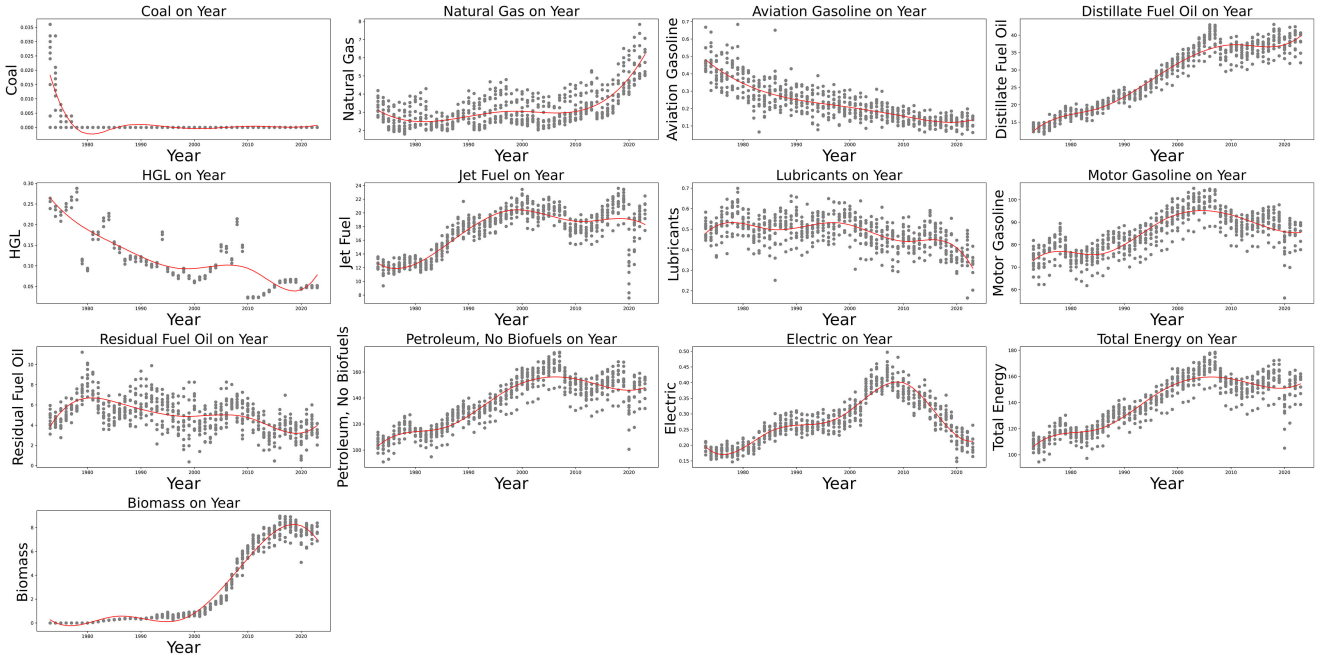


Figure 8
Spline transformation using transportation sector energy consumption data



4.2.4. Residential sector

The residential sector includes private households and dwellings, where energy consumption is driven by various activities such as heating, cooling, lighting, and appliance usage. Understanding energy patterns in this sector is essential for developing effective policies to promote energy efficiency and reduce emissions.

As shown in Figure 11, the spline transformation model proved effective by using piecewise polynomials to fit the data, similar to the commercial sector. This approach captured non-linear trends and sudden changes, resulting in smoother curves that align closely with the actual data points. This balance between flexibility and preventing overfitting makes spline transformation an ideal method for modeling complex energy consumption patterns in the

Figure 9
Energy consumption trends in the transportation sector

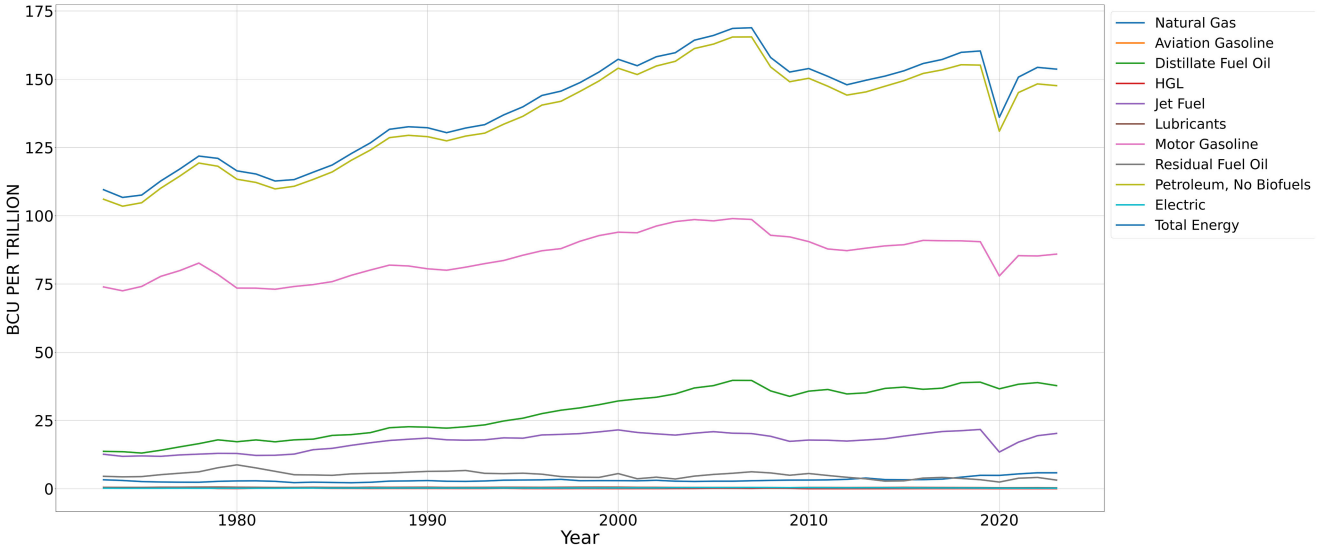
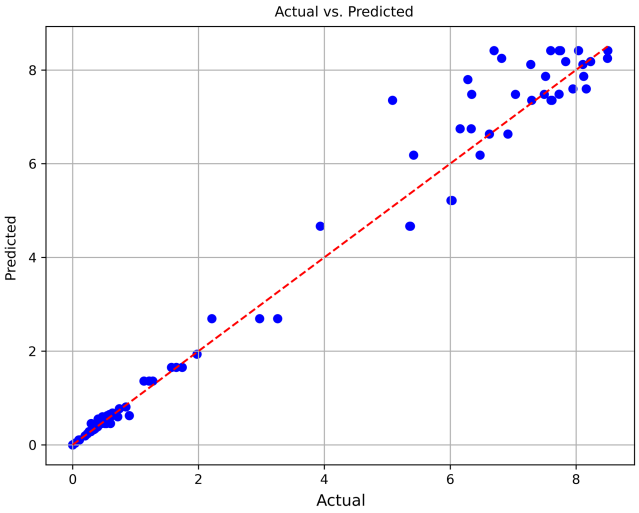


Figure 10
Transportation sector energy consumption prediction using XGBoost



residential sector. The graphs reveal key points of information, such as the transition to cleaner energy sources and the halt of coal usage for residential homes.

The line plot in Figure 12 reveals trends in energy consumption across different fuel types over time. Natural gas shows steady trends and makes up the majority of consumption, while petroleum usage has significantly declined over the decades. Biomass energy, although fluctuating, remains relatively stable. This visualization underscores our heavy reliance on fossil fuels over the years, while also showing a slight decline in their usage.

The prediction shown in Figure 13 highlights the predictive accuracy of the spline transformation model. The model's curves closely match the actual data points, demonstrating its capability to capture non-linear patterns and provide reliable forecasts without overfitting. This accuracy is crucial given the variability in the data, which can pose challenges for prediction.

The residential sector analysis highlights several significant trends and insights. The spline transformation model excelled in capturing complex consumption patterns. The line plots indicate a shift towards cleaner energy sources such as natural gas and the abrupt end of coal use in favor of cleaner alternatives. Overall, the

Figure 11
Spline transformation using residential sector energy consumption data

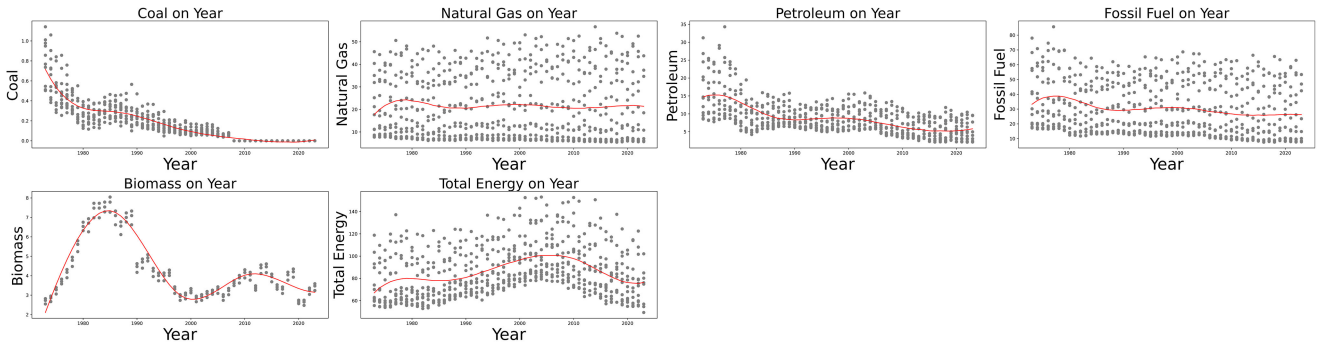


Figure 12
Energy consumption trends in the residential sector

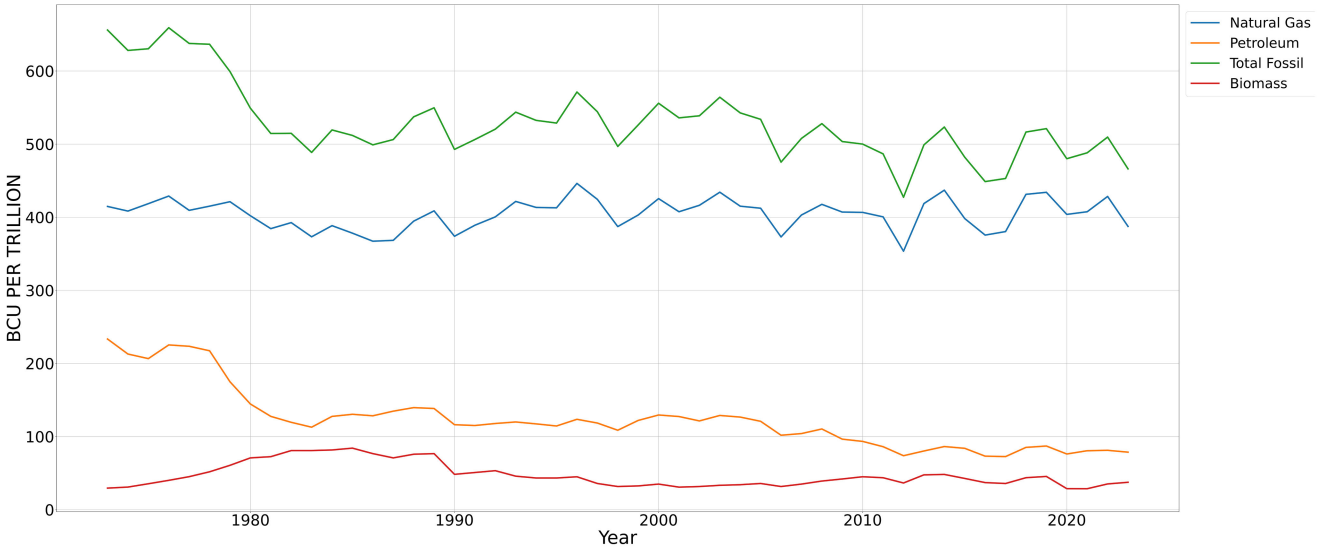


Figure 13
Residential sector energy consumption predictions

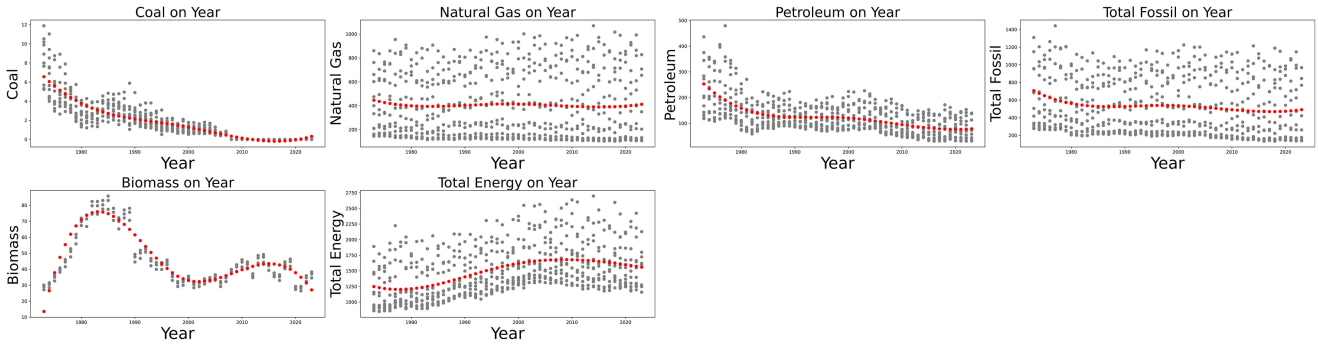


Figure 14
Spline transformation using electrical sector energy consumption data

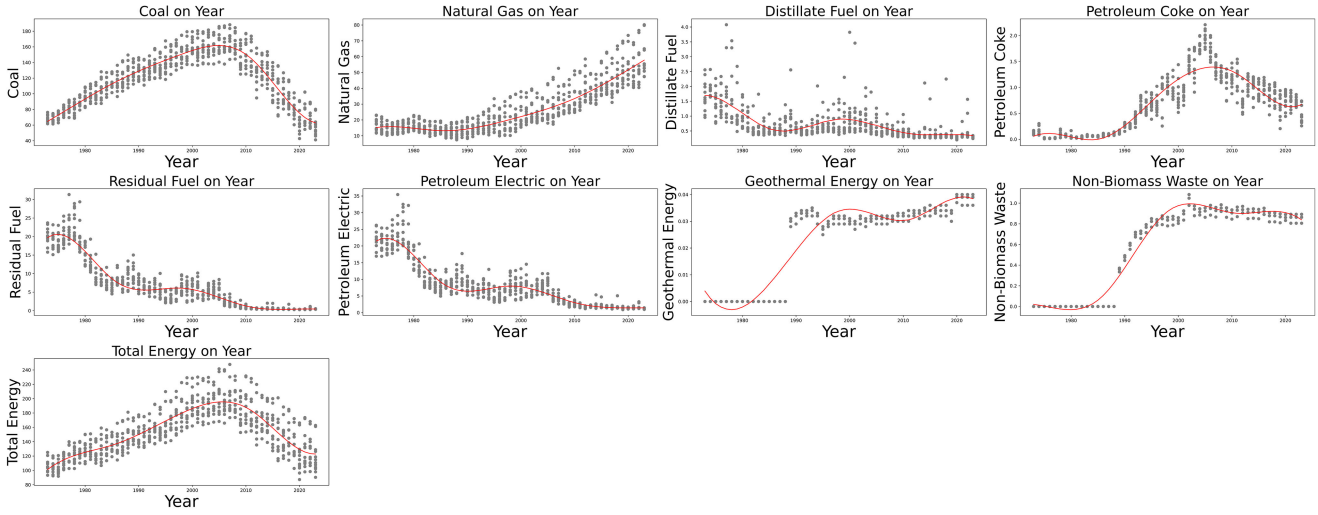


Figure 15
Energy consumption trends in the electrical sector

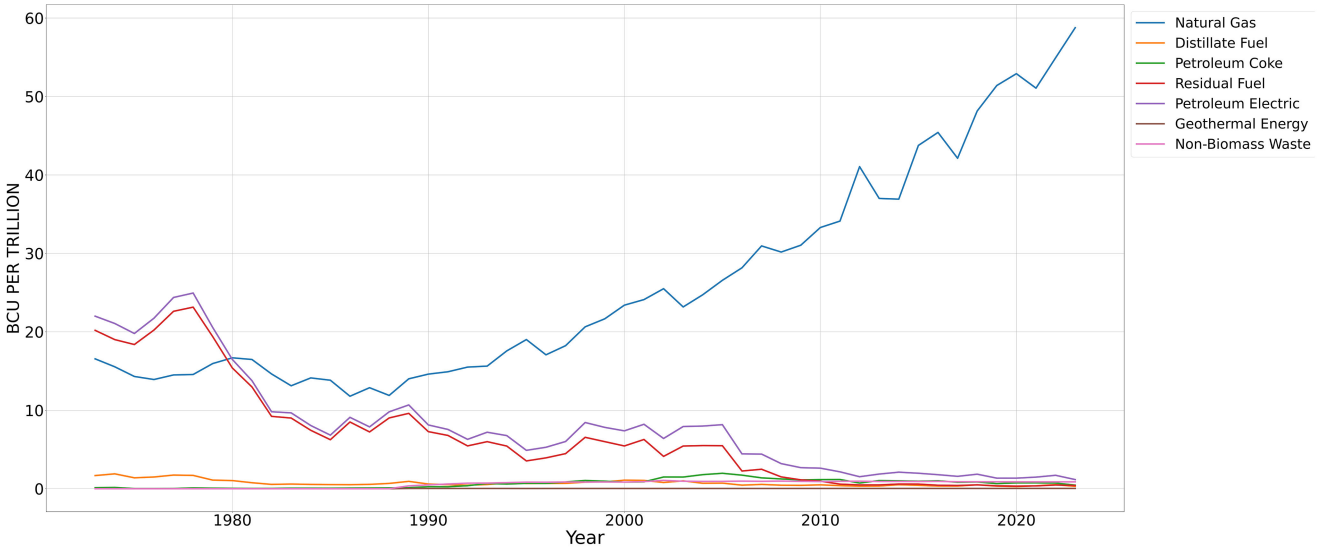
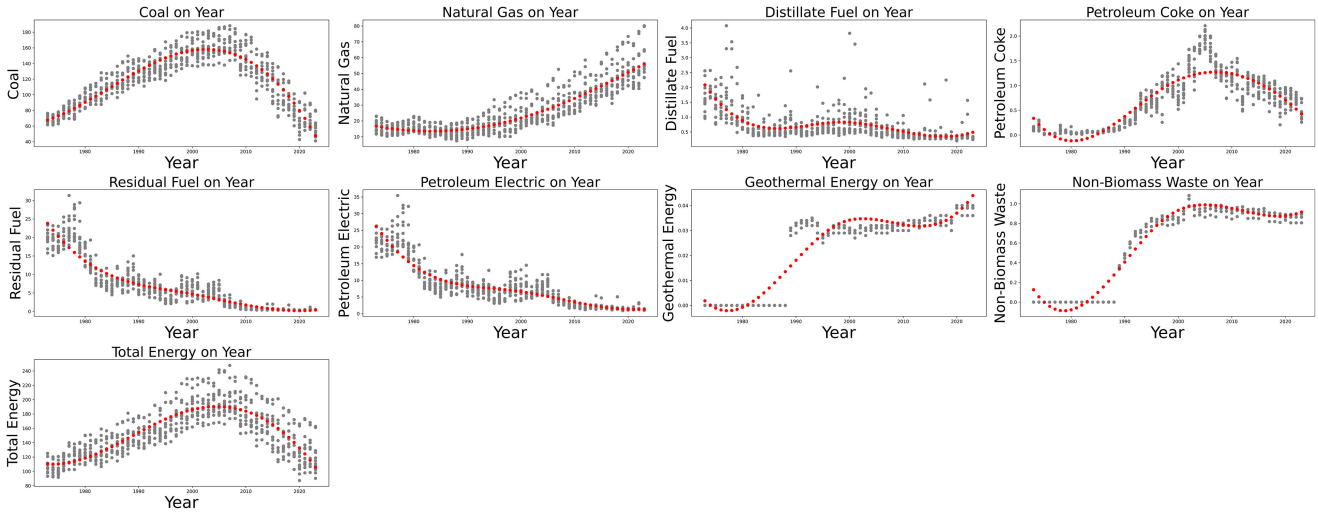


Figure 16
Electrical sector energy consumption predictions



energy consumption patterns in the residential sector reflect broader economic, seasonal, and environmental trends, underscoring the need for continuous monitoring and model enhancement to improve predictive accuracy.

4.2.5. Electrical sector

The electrical sector is responsible for generating and distributing electricity. This sector includes power plants, substations, and the electrical grid. Understanding the trends in energy consumption and emissions in this sector is crucial for developing sustainable energy policies and reducing the environmental impact of electricity generation.

As shown in Figure 14, the spline transformation model uses piecewise polynomials to fit the data, effectively capturing non-linear trends and sudden changes. This results in smoother curves that closely follow the actual data points. This balance between flexibility and avoiding overfitting makes spline transformation a valuable method for modeling complex energy consumption patterns in the electrical sector. Key points include the substantial rise in natural gas usage over recent decades, while other energy sources such as coal and petroleum exhibit a declining trend.

The line plot in Figure 15 demonstrates trends in energy consumption across various fuel types over time. Natural gas usage has significantly increased, becoming the predominant fuel for electricity generation, while coal and petroleum usage have declined. This shift highlights the move towards cleaner energy sources, although the continued reliance on fossil fuels remains evident. The steady or declining trends of other fuel sources, such

as Distillant and non-biomass waste, further emphasize the sector's transition.

The scatter plots depicted in Figure 16 offer a detailed view of the relationship between actual and predicted values for different fuel types. These plots highlight areas where the models perform well and identify potential areas for refinement. The consistency in prediction accuracy across various fuel types underscores the robustness of the models used.

Key trends and insights highlighted in the analysis of the electrical sector include some critical ones: The most complicated pattern in consumption was captured by the model spline transformation, which worked best in terms of predictability. Line plots show that there is a surge towards natural gas in generating electricity, while coal and petroleum have fallen drastically. In the scatter plots, areas were identified that need further refinement in order to maintain accuracy in the predictions. Overall, electrical energy consumption indicates economic and environmental trends. Continuous data provision needs to be enhanced in order to arrive at more accurate predictions and add to the development of sound sustainable energy policy.

5. Conclusion

We conclude that inclusive study of CO_2 emission forecasting through regression models contributes much from various fuel types and sectors to the overall emissions. Using large data sets from the EIA, we have brought forward predictive models that provide an in-depth understanding of the future output of CO_2. Our work puts forth an underlying understanding of energy consumption patterns, which will lead down the road of relevant decision-making and foster sustainable behavior.

Our findings indicate that the commercial and industrial sectors are moving towards cleaner sources of energy, such as natural gas. In contrast, the transportation sector remains very dependent on petroleum products, though it gradually increased in its consumption of natural gas. In addition, the residential sector has greatly reduced its use of coal and is replacing it with cleaner alternatives. Regarding the electrical sector, there is a huge increase in the consumption of natural gas; this shows movement into cleaner energy sources to generate energy.

It has been shown that spline transformation and XGBoost models are efficient in the capturing of complex consumption patterns, hence rendering quite accurate forecasts. These models are thus of utter importance for continuous monitoring and refinement to enhance predictive accuracy and inform energy efficiency initiatives.

For policymakers, the results offer actionable intelligence to prioritize sector-specific strategies, such as incentivizing the adoption of renewable energy in industrial operations or accelerating the deployment of electric vehicles to reduce petroleum dependency in transportation. Industry professionals can utilize these predictive models to assess the environmental impact of their energy use and adopt technologies that align with sustainability goals.

These insights align with global CO_2 reduction targets, including the Paris Agreement, by emphasizing scalable approaches to mitigate emissions. By fostering data-driven policies and encouraging investments in renewable energy and energy efficiency, this research contributes to the broader objective of limiting global temperature rise to below 2°C, as outlined in international climate initiatives. The study advocates for collaborative action among stakeholders to accelerate progress toward a sustainable, low-carbon future.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the U.S. Energy Information Administration at www.eia.gov. The data that support the findings of this study are openly available in the United States Environmental Protection Agency at www.epa.gov.

Author Contribution Statement

Kamil Samara: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Yunhwan Jeong and Thomas H. Beaupre:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization.

References

[1] Ren, Q., Wei, S., & Du, J. (2023). Research progress and perspectives on carbon capture, utilization, and storage (CCUS) technologies in China and the USA: A bibliometric analysis. *Environmental Science and Pollution Research*, *30*, 76437–76454. https://doi.org/10.1007/s11356-023-27749-w

[2] de Gouveia, M., & Inglesi-Lotz, R. (2021). Examining the relationship between climate change-related research output and CO_2 emissions. *Scientometrics*, *126*, 9069–9111. https://doi.org/10.1007/s11192-021-04148-x

[3] Zuhal, M., & Göcen, S. (2024). The relationship between CO_2 emissions, renewable energy, and economic growth in the US: Evidence from symmetric and asymmetric spectral Granger causality analysis. *Environment, Development and Sustainability*, *26*, 1–22. https://doi.org/10.1007/s10668-024-05002-9

[4] Gyimah, J., Hayford, I. S., Nwigwe, U. A., & Opoku, E. O. (2023). The role of energy and economic growth towards a sustainable environment through carbon emissions mitigation. *PLOS Climate*, *2*(3), e0000068. https://doi.org/10.1371/journal.pclm.0000116

[5] U.S. Energy Information Administration. (n.d.). *Commercial sector energy consumption*. Retrieved from: https://www.eia.gov/totalenergy/data/browser/index.php?tbl=T02.03

[6] U.S. Energy Information Administration. (n.d.). *Industrial sector energy consumption*. Retrieved from: https://www.eia.gov/totalenergy/data/browser/index.php?tbl=T02.04

[7] U.S. Energy Information Administration. (n.d.). *Transportation sector energy consumption*. Retrieved from: https://www.eia.gov/totalenergy/data/browser/index.php?tbl=T02.05

[8] Tong, M., Duan, H., & He, L. (2021). A novel Grey Verhulst model and its application in forecasting CO_2 emissions. *Environmental Science and Pollution Research*, *28*, 31370–31379. https://doi.org/10.1007/s11356-020-12137-5

[9] Huang, L. (2023). A study on the CO_2 emissions of mainland China using deep learning models. In *Proceedings of the 4th*

International Conference on Big Data & Artificial Intelligence & Software Engineering, 256–265.

[10] Santos, T. M. O., Bessani, M., & Da Silva, I. N. (2023). Evolving dynamic Bayesian networks for CO_2 emissions forecasting in multi-source power generation systems. *IEEE Latin America Transactions*, *21*(9), 1022–1031. https://doi.org/10.1109/TLA.2023.10251809

[11] Hosseini, S. M., Saifoddin, A., Shirmohammadi, R., & Aslani, A. (2019). Forecasting of CO_2 emissions in Iran based on time series and regression analysis. *Energy Reports*, *5*, 619–631. https://doi.org/10.1016/j.egyr.2019.05.004

[12] Auffhammer, M., & Steinhauser, R. (2012). Forecasting the path of U.S. CO_2 emissions using state-level information. *The Review of Economics and Statistics*, *94*(1), 172–185. http://www.jstor.org/stable/41349167

[13] Alam, T., & AlArjani, A. (2021). Forecasting CO_2 emissions in Saudi Arabia using artificial neural networks, Holt-Winters exponential smoothing, and ARIMA models. In *Proceedings of the International Conference on Technology and Policy in Energy and Electric Power*, 125–129.

[14] Ho, H. X. T. (2018). Forecasting CO_2 emissions, renewable energy consumption, and economic growth in Vietnam using Grey models. In *Proceedings of the 4th International Conference on Green Technology and Sustainable Development*, 452–455.

[15] Juliet, A. H., Malathi, P., & Legapriyadharshini, N. (2024). Smart forecasting: Harnessing machine learning for accurate CO_2 emission predictions. In *Proceedings of the 11th International Conference on Reliability, Infocom Technologies and Optimization*, 1–7.

[16] Thwe, Y., Maneetham, D., Crisnapati, P. N., & Aung, M. M. (2023). Thailand carbon dioxide emissions forecasting using stacked LSTM-based prediction model. In *Proceedings of the 11th International Conference on Cyber and IT Service Management*, 1–6.

[17] Li, X., & Zhang, X. (2023). A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China. *Environmental Science and Pollution Research*, *30*, 117485–117502. https://doi.org/10.1007/s11356-023-30428-5

[18] Priya, N., Srinidhi, K., & Kousalya, T. (2023). Carbon footprint monitoring system using machine learning and deep learning techniques. In *Proceedings of the 12th International Conference on Advanced Computing*, 1–8.

[19] Krishnamurthy, B. K., Shi-Wei, W., Mu-En, W., & Thangavelu, K. (2023). The United States energy consumption and carbon dioxide emissions: A comprehensive forecast using a regression model. *Sustainability*, *15*, 7932. https://doi.org/10.3390/su15107932

[20] Harrison, M. (2023). Effective XGBoost. *MetaSnake*.

How to Cite: Samara, K., Jeong, Y., & Beaupre, T. H. (2025). Forecasting CO_2 Emission in the US Using Regression Models. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS52024482