**Evaluating Electricity Generation from Coal and Natural Gas: A Predictive Analysis of Future Consumption**

Yewon Lee, George Jiang, and Christine Sangphet

**1. Introduction**

**1.1 Problem Definition**

Nonrenewable energy sources, such as coal and natural gas, remain central to global electricity generation but come at a significant environmental cost. The combustion of these fuels releases pollutants like sulfur dioxide (SO₂), nitrogen oxides (NOₓ), and fine particulate matter (PM2.5), which degrade air quality and pose severe risks to public health. These emissions not only contribute to respiratory illnesses but also accelerate climate change, creating a pressing need for cleaner energy solutions. Despite strides toward renewable energy, the reliance on nonrenewables continues to challenge regions striving for sustainability. In California, natural gas plants often operate to stabilize the grid during periods of high energy demand, such as heat waves, leading to localized spikes in air pollution. Meanwhile, New England faces its own hurdles, with seasonal dependence on natural gas for heating and electricity generation putting pressure on air quality during winter months. These challenges highlight the intricate relationship between energy consumption, environmental impact, and public health, underscoring the urgency of analyzing nonrenewable energy trends to inform cleaner, more sustainable policies.

**1.2 Objectives**

The primary objectives of this project are to:

1. **Analyze electricity generation by coal and natural gas**: Evaluate the trends and contributions of coal and natural gas in electricity generation across New England and Los Angeles (LA) from 2020 to 2023, focusing on the impact of these energy sources on air quality.
2. **Examine historical AQI data**: Investigate historical Air Quality Index (AQI) data for New England and LA during the same period (2020-2023) to understand the relationship between energy production and air pollution levels.
3. **Predict AQI for 2024**: Develop a model to predict AQI values for 2024 based on historical data from 2020 to 2023, taking into account fluctuations in electricity generation and its environmental consequences.
4. **Compare predicted and actual AQI for 2024**: Assess the accuracy of the 2024 AQI predictions by comparing them to the actual AQI values for the year, identifying trends and discrepancies to better understand the efficacy of prediction models.
5. **Inform sustainability policies**: Provide insights on the potential for cleaner energy solutions in New England and LA, helping policymakers make informed decisions to reduce air pollution and mitigate the environmental and public health impacts of nonrenewable energy sources.

**1.3 Scope of Project**

1. Azure Resource Setup

* Create the necessary Azure resources for the project

1. Data Architecture Design

* Develop a data architecture diagram to outline the data flow and integration points

1. Data Ingestion Pipelines

* Build pipelines for ingesting historical data
* Develop pipelines for real-time data ingestion

1. Stream Processing

* Implement stream processing and store results in target data storage location

1. Data Preprocessing

* Implement workflows to clean, transform, and prepare data for analysis and modeling

1. Exploratory Data Analysis (EDA)

* Perform EDA to uncover patterns, trends, and insights from the data

1. Batch Data Processing

* Execute batch processing to handle large datasets efficiently

1. Data Loading

* Load processed data into the target storage location

1. Predictive Modeling

* Develop and implement a predictive model to generate actionable insights

1. Data Visualization

* Create interactive visualizations using Power BI to effectively communicate insights and findings

**2. Methodology**

**2.1 Data Collection**

Energy generation data was sourced from the [U.S. Energy Information Administration](https://www.eia.gov/opendata/browser/electricity/rto/fuel-type-data) (EIA) via their official platform, eia.gov. Specifically, “hourly generation by energy source” filtered by coal and natural gas was chosen to best match our daily air quality data integration from [OpenWeather](https://openweathermap.org/history). Boston and Los Angeles was our focus areas for analysis, as these cities provided contrasting environmental and energy usage patterns, with Boston experiencing seasonal dependence on natural gas for heating and Los Angeles relying heavily on coal and natural gas for electricity generation during peak demand periods.

For real-time streaming data, [ISO New England](https://webservices.iso-ne.com/docs/v1.1/) provided real-time data on natural gas and coal energy generation, while [Ambee](https://www.getambee.com/api/air-quality) supplied real-time air quality information. These pipelines ensured the continuous integration of static and dynamic datasets, providing a robust foundation for deeper analysis.

**2.2 Data Processing**

Workflows in Azure Data Factory were designed to process the datasets collected from the data pipelines. This process included defining data types for each column and filtering out unnecessary or null values to ensure data quality in Azure Data Flow. After cleaning, the data was ingested into Databricks, where we performed aggregation operations to prepare it for further analysis and prediction. To accommodate the varying requirements of different predictive models, additional adjustments were made to the data structure locally.

To integrate and process real-time data, we implemented Azure Function Apps and Event Hubs within our architecture. These components allowed us to efficiently capture streaming data from sources such as ISO New England and Ambee in near real-time. The streaming data was funneled into Event Hubs, which served as a centralized ingestion layer, and was subsequently processed and stored for downstream analysis. This hybrid approach enabled seamless handling of both static and real-time data, ensuring that our models operated on the most current and relevant information.

**2.3 Exploratory Data Analysis**

For our initial exploratory data analysis (EDA) and predictive modeling, we employed both traditional statistical methods and modern machine learning techniques. To analyze patterns and relationships within the time series data, we plotted monthly energy production from coal and natural gas for New England and Los Angeles. To determine stationarity, we conducted the Augmented Dickey-Fuller (ADF) test, which checks if a time series’ mean, variance, or autocorrelation remains constant over time. A series was considered stationary if the ADF test statistic was lower than the critical value (α = 0.05). All our time series data—AQIs and energy generation from coal and natural gas for both regions—were confirmed to be stationary, ensuring compatibility with forecasting methods that require stability over time.

We also performed seasonal decomposition to analyze annual and quarterly trends. The energy data (January 2019 to November 2024) provided sufficient length to detect annual seasonality and trends. For AQI data (January 2021 to November 2024), we analyzed quarterly trends due to the shorter time span. Limited access to recent AQI data motivated us to predict the 2024 AQI for later validation.

Due to the shorter length of the AQI dataset, we merged the energy generation dataset into the AQI dataset based on the AQI’s time period. We splitted our dataset into training and test dataset with a ratio of 8:2. So, models predicted from January 2024 to November 2024. We evaluated 10 predictive models: VectorAutoRegressor, RandomForestRegressor, AdaBoostRegressor, LightGBMRegressor, CatBoostRegressor, XGBoostRegressor, MultiLayerPerceptron (MLP) Regressor, ElasticNet, Support Vector Regressor (SVR), and Long Short-Term Memory (LSTM). By comparing mean absolute percentage error (MAPE), we identified the most accurate model. Additional metrics, such as mean error(ME) and mean percentage error(MPE), provided insights into whether the models tended to overestimate or underestimate the actual values.

**3. Data Analysis**

**3.1 Descriptive Analytics:** analyze historical data to identify patterns and relationships

When compared to LA's AQI trend, there does not appear to be a strong correlation between coal usage and AQI, as AQI levels remain consistently high throughout the period despite a sharp reduction in coal reliance. This suggests that the reduction in coal production alone has not been sufficient to significantly improve air quality in LA. On the other hand, natural gas usage may exhibit a positive correlation with AQI, as periods of higher natural gas electricity production align with elevated AQI levels. However, given the complexity of LA's pollution sources, such as transportation, industrial activities, and wildfires, natural gas electricity production alone cannot be definitively identified as the primary contributor to poor air quality.

The New England’s Natural Gas Trend exhibits a relatively stable trajectory with a gradual upward trend in electricity production, especially toward the later part of the timeline. In contrast, the Coal Trend shows a dramatic and consistent decline, stabilizing at a very low level for most of the observed period. When considering AQI trends alongside these energy trends, there appears to be a potential inverse correlation between AQI and coal usage: as coal-based electricity production significantly drops, AQI remains relatively stable with fewer fluctuations. This suggests that reduced reliance on coal in New England may contribute to improved or stable air quality. On the other hand, increasing natural gas usage, which is considered cleaner than coal but still emits pollutants, does not appear to have caused significant spikes in AQI.

**3.2 Diagnostic Analytics: ​​**a data analysis technique that helps organizations understand the root causes of events, behaviors, and outcomes

The decomposition of LA's AQI data into observed, trend, seasonal, and residual components provides valuable insights into its behavior over time. The observed data reveals significant short-term fluctuations, with AQI values frequently peaking above 150, indicating intermittent episodes of poor air quality. The trend component highlights a clear long-term pattern, showing a decline in AQI values from early 2022 to mid-2022, followed by a steady upward trend throughout 2023 and into 2024. This suggests that air quality in LA has been deteriorating over the past year, which may be linked to increased emissions, weather conditions, or seasonal influences. The seasonal component displays a recurring pattern with AQI peaking cyclically, likely due to seasonal factors such as higher summer temperatures, reduced air circulation, or wildfire activity, which are common contributors to worsening air quality in LA. Lastly, the residual component shows considerable variability, indicating the presence of irregular events or anomalies that are not captured by the trend or seasonal patterns. These residual spikes could correspond to isolated pollution events, such as industrial activity, traffic surges, or localized wildfires. Overall, the analysis suggests a combination of long-term worsening trends, cyclical seasonal effects, and random external factors driving AQI variations in LA.

The diagnostic analysis of New England's AQI based on the time series decomposition reveals several key insights into its behavior over time. The observed data exhibits substantial short-term fluctuations, indicating frequent variations in air quality, which could be from seasonal and environmental factors. The trend component shows a gradual rise in AQI values from early 2021 to mid-2022, peaking around 79, followed by a moderate decline toward the end of 2023. This trend suggests a temporary deterioration of air quality during specific periods, likely caused by external factors such as increased emissions or weather patterns. The seasonal component reveals clear periodic patterns, with AQI peaking during the warmer months (e.g., mid-year), potentially due to increased pollutant concentration from industrial activity, vehicle emissions, or atmospheric conditions like temperature inversions. Finally, the residual component shows considerable random noise, indicating the presence of irregular events or anomalies not captured by the trend or seasonality. Overall, this decomposition highlights the combined impact of cyclical, long-term, and random factors contributing to AQI variations in New England, with notable seasonal peaks and a slightly declining trend after mid-2022.

**3.3 Predictive Analytics:** a data analysis process that uses machine learning, statistical models, and artificial intelligence to forecast future outcomes

For each model, we calculated MAPE, ME, mean absolute error(MAE), MPE, root mean square error(RMSE), and correlation coefficient(Corr). Overall, CatBoostRegressor and ElasticNet tend to perform well across both NE and LA with relatively low MAPE, MAE, and RMSE values. They also exhibit decent correlation scores. In contrast, MLPRegressor and LSTM stand out with high errors (MAPE, RMSE) and weak correlations, indicating poor performance. LSTM in LA is particularly problematic with an extremely high MAPE (0.8069) and RMSE (105.2394).

Models like ElasticNet for New England(NE) and SVR show smaller ME and MPE values, indicating minimal systematic bias. Models with higher correlation scores include SVR (NE) (0.4025), ElasticNet (NE) (0.3366), and CatBoostRegressor (NE) (0.2693), indicating a strong linear relationship with actual values. Due to less bias and more linear relationship with actual values, ElasticNet fails to follow the trends or seasonality. MLPRegressor and LSTM exhibit large negative ME and MPE values, highlighting severe underestimation of the targets, especially in LA.

Models perform slightly better in NE compared to LA, as reflected in lower errors and stronger correlations for NE models. This suggests AQI trends in NE may be easier to predict due to more stable patterns compared to the higher volatility in LA AQI.

**3.4 Prescriptive Analytics:** seeks to answer the question, “What should we do?”

To improve air quality in Los Angeles, efforts should be made to reduce emissions from July to October and November when ozone levels and wildfire activity exacerbate pollution. This can be achieved by promoting the adoption of renewable energy sources like solar and wind to reduce reliance on natural gas and coal during high-demand seasons. Also, implementing stricter wildfire management and emission control policies can mitigate wildfire season spikes in AQI. In addition, implementing stricter regulations on industrial and vehicle emissions during high-risk periods is recommended.

In New England, the diagnostic analysis shows that a sharp decline in coal electricity production coincides with stable AQI. Policies should ensure that coal plants remain phased out, while maintaining reliability of cleaner alternatives like natural gas or renewable. The upward trend in natural gas usage indicates increasing dependency, which, while cleaner than coal, still contributes to pollution. Gradually transitioning to renewable energy sources such as offshore wind and hydropower can reduce emissions.

By targeting the root causes with season-specific interventions or random factors through real-time monitoring and transitioning to renewable energy, both regions can achieve better air quality.

Every visualizations or evidence related to above analytics can be also found in [our github](https://github.com/george962/EnergyAQAnalysis).

**4. Conclusion**

**4.1 Summary of Findings**

The analysis highlights regional differences in air quality management and the factors influencing air quality index (AQI) trends in Los Angeles and New England. In Los Angeles, the significant reduction in coal usage has not resulted in noticeable improvements in air quality, suggesting that coal reliance is not the primary factor influencing AQI levels. Instead, natural gas usage shows a possible relationship with AQI, particularly during seasons with higher emissions. However, natural gas is only one contributor among several significant factors, including transportation emissions, industrial activities, and wildfires. Seasonal variations, especially in the summer, and irregular events such as wildfires further add complexity to air quality challenges in Los Angeles.

In contrast, New England demonstrates a clearer connection between reduced coal-based electricity generation and stable or improving AQI levels, indicating that the decline in coal usage has had a positive impact on air quality. Although natural gas usage has increased over time, its impact on AQI appears minimal due to its relatively cleaner combustion properties compared to coal. The patterns in New England's AQI are more consistent, making them easier to predict and manage compared to the more erratic trends observed in Los Angeles.

**4.2 Key Insights**

* **Model Performance**: Predictive models perform more effectively in New England due to its consistent AQI patterns, while the more unpredictable trends in Los Angeles present greater challenges for accurate forecasting. Among the evaluated models, CatBoostRegressor and ElasticNet demonstrated the best overall performance for predicting air quality. In contrast, models such as MLPRegressor and LSTM struggled, particularly in Los Angeles, where they exhibited high errors and weak correlations with actual AQI data.
* **Season-Specific Strategies:** Addressing air quality issues requires targeted, season-specific interventions. For Los Angeles, efforts should focus on reducing emissions during high-risk summer and wildfire seasons by adopting renewable energy, implementing stricter emission regulations, and improving wildfire management. In New England, continuing the transition away from coal while enhancing reliance on renewable energy sources such as offshore wind and hydropower will further improve air quality and reduce dependency on natural gas.

**4.3 Recommendations**

* **For Policymakers:**
  + **Grid Modernization and Energy Storage:** Upgrading the energy grid to accommodate a larger share of renewable energy and enhancing energy storage infrastructure are essential for maintaining a stable and clean energy supply during peak demand seasons. Investments in advanced battery technologies and smart grid systems will improve both efficiency and reliability.
  + **Mitigating Urban Heat Islands with Smart City Design:** Investing in smart city technologies can help reduce the urban heat island effect, which worsens air quality, particularly in cities like LA. Solutions such as reflective pavements, green spaces, and rooftop gardens can be implemented, which not only lower temperatures but also absorb CO2. By mitigating the heat island effect, temperatures will decrease, and air quality will improve, as lower temperatures reduce the formation of ground-level ozone, a major air pollutant. A creative application of this approach would be launching a "smart cooling" initiative, where sensors throughout the city monitor temperature and air quality, dynamically adjusting cooling systems and maintaining green infrastructure to further reduce pollution.
* **For the Public:**
  + **Localized Air Quality Sensors and Personal Tracking:** Low-cost air quality sensors can be deployed in both urban and rural areas of Los Angeles and the Northeast to collect hyper-local pollution data. These sensors can also be integrated into wearable devices, enabling individuals to monitor their personal exposure. Real-time data on pollutant exposure helps people take preventive measures, such as wearing masks or avoiding certain areas during high-pollution periods. This information can also guide policy decisions. A useful application would be pairing sensors with a mobile app that alerts users when air quality is poor and suggests actions like staying indoors or using an air purifier. Public health officials could leverage this data to target pollution-reduction efforts more effectively.

**4.4 Project Evaluation**

The real-time insights gained from this project are incredibly valuable for grid operators, enabling them to make immediate, informed decisions about energy usage and air quality management. The ability to act in real time is critical, as it allows operators to respond proactively to dynamic conditions such as sudden changes in energy demand, unexpected renewable energy generation fluctuations, or deteriorating air quality levels. In the future, our model can integrate additional real-time inputs, such as weather forecasts or live traffic data for a better understanding on Air Quality impact. Not only would operators be able to take immediate action, but they are also better equipped to address challenges.

A key limitation of the project was the availability of historical data. The OpenWeatherApp, which served as the primary data source for air quality information, only provided data starting from 2019. This restricted the analysis to a relatively short period, preventing the exploration of longer-term trends or more recent shifts in air quality that may have resulted from changing environmental policies, energy generation patterns, or other factors. The lack of a more extensive historical dataset limited the ability to identify and evaluate long-term impacts and trends, which could have provided more comprehensive insights into air quality dynamics over time.

One of the primary challenges faced during this project was the difficulty in finding a reliable and efficient stream API that could consistently provide the required real-time data. Accessing accurate, continuous data for air quality and energy generation is essential, but many available APIs either offered limited data, were challenging to integrate, or suffered from data quality issues. These constraints hindered the ability to perform more comprehensive real-time analysis and reduced the overall effectiveness of the project in providing timely insights into air quality and energy trends.

To address the limitations and challenges, exploring alternative API options could be an effective strategy. There are several other data providers that might offer broader or more recent datasets, such as government or academic databases that could provide comprehensive historical air quality and energy generation data. Open-source APIs or partnerships with environmental organizations might also offer better access to data. However, when we looked into some of these alternatives, we found that they often required paid subscriptions or had high usage fees, which is why we chose not to use them for this project.