Detailed Report on Total Primary Energy Consumption Forecasting

Prepared by:Deepak reddy gopu and Vanga Jayasimha Reddy

December 9, 2024

Abstract

This study provides a complete analysis of forecasting total primary energy consumption using both standard statistical approaches and advanced machine learning techniques. The dataset, which spans five decades, covers key energy production and consumption measures, providing insights into long-term trends, seasonal fluctuations, and responses to global disturbances such as the oil crisis and the COVID-19 pandemic.

Several forecasting models, including ARIMA, ETS, and XGBoost, were used to handle the particular peculiarities of energy consumption data. Furthermore, a hybrid model was created to combine the advantages of these approaches. ARIMA effectively caught temporal dependencies and seasonality, ETS modeled multiplicative trends and damped patterns, while XGBoost handled complex non-linear interactions. The hybrid model utilized these strengths to achieve greater performance, with the lowest RMSE (0.174) and the greatest R-squared value (0.82).

Key findings demonstrate the hybrid model's resilience in energy forecasting, which has applications in strategic energy planning, supply chain optimization, and climate change programs. While the hybrid approach beat individual models, issues like processing overhead and reliance on high-quality data persist. Future study should aim to maximize hybrid model weights and investigate advanced techniques such as recurrent neural networks and transformers.

This study emphasizes the significance of combining several forecasting approaches to meet the complexities of energy dynamics and contribute to sustainable energy management practices.

Contents

1 Introduction

Energy consumption forecasting is critical in tackling global concerns such as resource management, sustainability, and climate change. Accurate forecasting enables policymakers, industry, and stakeholders to optimize resource allocation, increase supply chain efficiency, and develop effective strategies for sustainable energy production and consumption. This paper examines and forecasts total primary energy consumption utilizing a mix of traditional statistical approaches and modern machine learning techniques, resulting in a hybrid modeling approach.

The energy business has been disrupted multiple times over the last few decades, including oil crises, economic recessions, and, most recently, the COVID-19 epidemic. These occurrences have highlighted the limitations of typical forecasting models in capturing nonlinear transitions, rapid shocks, and long-term trends. For example, the 1970s oil crises demonstrated the inadequacies of linear models, but pandemic-induced variations in energy demand underlined the necessity for models that can manage both rapid shifts and cyclical patterns.

This paper uses a large dataset spanning five decades to provide insights on long-term patterns, seasonal variations, and responses to major global events. The process includes exploratory data analysis, decomposition techniques, and the use of forecasting models like ARIMA, ETS, and XGBoost. Furthermore, a hybrid model that combines these approaches is presented to improve predicting accuracy by using the strengths of each strategy.

The hybrid model's ability to capture temporal dependencies, seasonality, and nonlinear interactions makes it an effective approach for forecasting energy use. By overcoming the shortcomings of separate models, the hybrid approach provides a more comprehensive and dependable framework for energy forecasts.

Objectives

The main goals of this report are to:

- Investigate the trends, seasonal patterns, and structural behaviors in total primary energy consumption.
- Assess the effectiveness of individual forecasting models, such as ARIMA, ETS, and XGBoost, by using standard performance metrics.
- Develop and validate a hybrid model that combines statistical and machine learning methods to enhance prediction accuracy.
- Offer practical insights to aid in energy planning, resource management, and sustainability efforts.

Significance of Energy Forecasting

Forecasting energy consumption is a vital tool for shaping a sustainable future. It ensures that energy supply aligns with demand, minimizing waste and preventing shortages. Beyond this practical benefit, energy forecasting is also a critical enabler of meaningful progress in renewable energy adoption. By helping to develop policies and infrastructure, it reduces our dependence on fossil fuels and promotes cleaner alternatives.

Moreover, accurate forecasts play a key role in addressing global challenges like climate change. They inform strategies to reduce carbon footprints and improve energy efficiency. In an ever-changing world, forecasting also empowers industries and governments to adapt to economic shifts and disruptions that influence energy demand.

This report embraces both traditional and cutting-edge techniques to lay the groundwork for more reliable and impactful energy forecasting frameworks, addressing the challenges of today while paving the way for tomorrow.

2 Dataset Overview

The dataset covers the period from January 1973 to July 2024 and contains monthly observations of energy output, consumption, and related variables. It offers a comprehensive view of the energy sector over the last five decades, allowing for analysis of long-term patterns, seasonality, and responses to key world events. The dataset's temporal depth and diversity make it a great resource for studying energy dynamics.

2.1 Key Attributes

The dataset provides a detailed picture of energy production and consumption. Here are the key attributes it captures:

- **Total Fossil Fuels Production:** This shows how much energy comes from traditional sources like coal, oil, and natural gas. It's a good measure of how dependent we are on fossil fuels.
- **Nuclear Electric Power Production:** This tracks the energy generated by nuclear power plants and highlights the role of nuclear energy in the overall energy mix.
- **Total Renewable Energy Production:** This includes energy from clean sources like solar, wind, and hydroelectric power, which are essential for achieving sustainability goals.
- **Total Primary Energy Production:** This combines all energy sources—fossil fuels, nuclear, and renewables—giving a complete view of how much energy is being produced overall.
- **Primary Energy Imports and Exports:** These metrics show the energy trade balance, helping to understand whether a country is energy-independent or relies on imports.
- **Primary Energy Stock Change and Other:** This tracks energy stored, losses, or adjustments, offering insight into operational efficiency.
- **Total Fossil Fuels Consumption:** This tells us how much energy from fossil fuels is being used, which helps to understand demand patterns.
- **Total Renewable Energy Consumption:** This measures how much renewable energy is being used to meet demand, giving a sense of progress toward cleaner energy.
- **Total Primary Energy Consumption:** This combines all energy consumed from various sources, providing an overall picture of energy usage.

2.2 Sample Data

The table below shows the first few rows of the dataset, illustrating the key attributes for January to April 1973.

Month	Fossil Fuels	Nuclear	Renewable	Primary	Imports	Exports	Net Imports
1973-01	4.932	0.068	0.220	5.221	1.173	0.126	1.047
1973-02	4.730	0.065	0.197	4.992	1.168	0.121	1.047
1973-03	4.947	0.072	0.219	5.238	1.309	0.140	1.170
1973-04	4.716	0.064	0.209	4.990	1.085	0.194	0.891

Table 1: Production and Trade Data: January to April 1973

Month	Stock Change	Fossil Fuels Cons.	Nuclear Cons.	Renewable Cons.	Total Cons.
1973-01	0.772	6.748	0.068	0.220	7.040
1973-02	0.390	6.163	0.065	0.197	6.429
1973-03	-0.068	6.045	0.072	0.219	6.340
1973-04	-0.110	5.493	0.064	0.209	5.771

Table 2: Consumption and Stock Data: January to April 1973

2.3 Data Integrity

The dataset is complete and free from missing values in key columns such as *Month* and *Total Primary Energy Consumption*, making it highly reliable for analysis and modeling. With a rich variety of attributes, this dataset opens the door to an in-depth exploration of energy production and consumption patterns. It also provides valuable insights into the dynamics of energy trade and storage, ensuring a holistic view of the energy sector.

3 Exploratory Data Analysis

3.1 Time Series Visualization

Figure 1 shows the changes in total primary energy consumption over time. A few key patterns stand out:

- The data shows a clear upward trend, especially after 2000, indicating a steady growth in energy consumption.
- Regular ups and downs can be seen throughout the timeline, pointing to seasonal patterns in energy use.
- From 1973 to 1980, there is a gradual increase with occasional short-term dips, possibly caused by external events like economic challenges.
- Between 1990 and 2005, energy consumption seems more stable, with fluctuations confined to a smaller range.
- After 2005, there's a noticeable rise in energy consumption, accompanied by more dramatic peaks and troughs, suggesting increased variability.
- Around 2020, there's a sharp drop in energy consumption, likely tied to the global slowdown during the COVID-19 pandemic, followed by a rapid recovery.
- The latest data points show consistently high levels of energy consumption, hinting that demand continues to grow steadily.

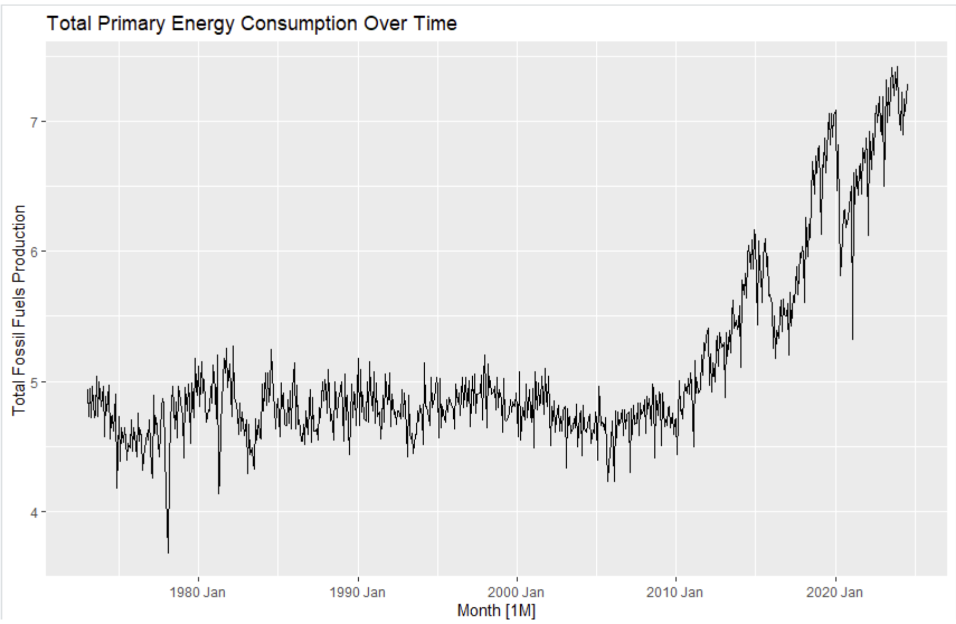

Figure 1: Time Series of Total Primary Energy Consumption Over Time

3.2 Outlier Analysis

The boxplot in Figure 2 offers a simple yet powerful way to understand how total primary energy consumption is distributed. Here are some key takeaways:

- The median energy consumption sits neatly in the middle of the interquartile range (IQR), showing that the data is fairly balanced around the center.
- The box in the plot covers the middle 50% of the data, from the 25th to the 75th percentiles. Most of the data values comfortably fall within this range.
- The "whiskers" of the boxplot extend to show data within 1.5 times the IQR. Data points beyond this range would typically be considered outliers, but in this case, there aren't any visible outliers.
- Total energy consumption values generally range between 6 and 9 units, with no extreme highs or lows standing out.
- Overall, the distribution looks steady and consistent, with no unusual patterns or clusters to note.

This analysis confirms that the data is well-behaved, with no significant outliers affecting its distribution, making it ready for further modeling and analysis.

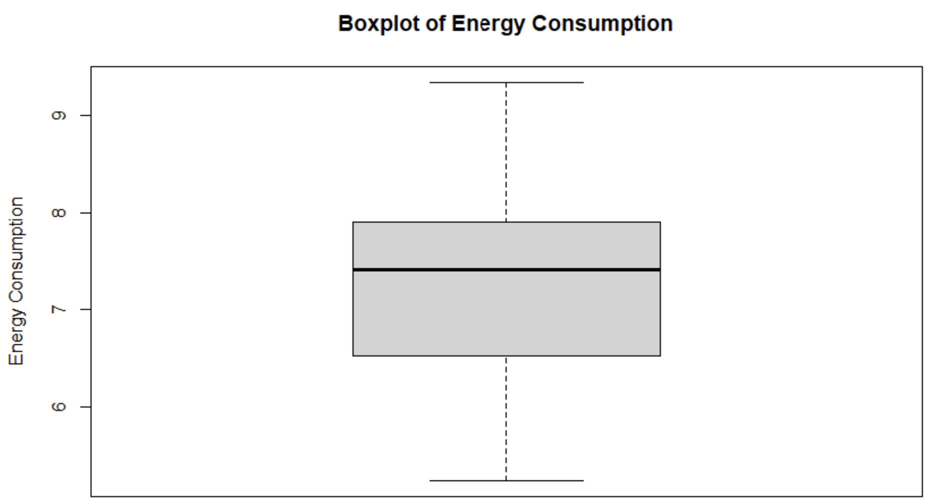

Figure 2: Boxplot of Total Primary Energy Consumption

4 Data Preprocessing

The dataset was converted to 'tsibble' format, and there were no missing values in the target or time columns. This preprocessing step guarantees compatibility with advanced time series models such as ARIMA and ETS.

5 Time Series Decomposition

STL decomposition was used to separate the entire primary energy consumption data into three categories: trend, seasonal, and remaining. Figure 3 depicts these components, revealing the underlying structure of the time series.

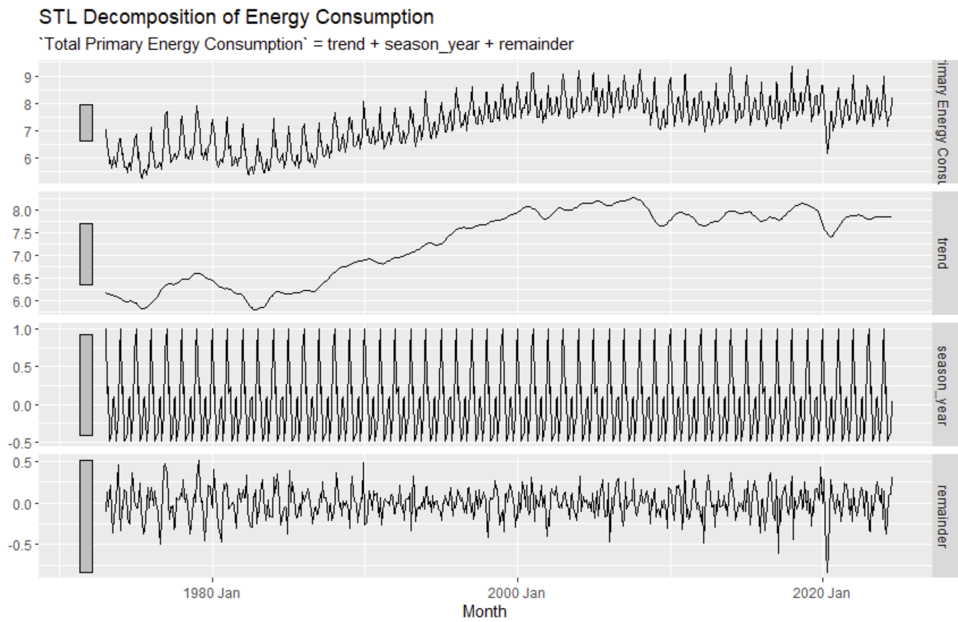

Figure 3: STL Decomposition of Total Primary Energy Consumption

5.1 Key Observations

The analysis of the time series decomposition reveals three significant components that shape the energy consumption patterns:

Trend Component: Over time, energy consumption has shown a steady upward trend, especially after the year 2000. This indicates a long-term growth in energy demand, likely driven by factors such as industrial expansion, population growth, and advancements in technology.

Seasonal Component: A clear cyclical pattern is observed in the seasonal component. These consistent fluctuations could be linked to seasonal variations in weather or recurring economic cycles that influence energy usage.

Remainder Component: The remainder, or residual component, captures the irregular and unpredictable changes in energy consumption. Notable spikes in this component, such as the sharp drop during 2020, can often be traced to external disruptions, like the COVID-19 pandemic.

These observations highlight the importance of considering both predictable patterns and irregular fluctuations when developing forecasting models for energy consumption.

5.2 Insights from Decomposition

The decomposition analysis revealed several key insights into the structure of the energy consumption data:

- **Seasonal Patterns:** The data shows strong seasonal patterns, reflecting predictable periodic changes in energy consumption. These patterns are essential for accurate modeling and forecasting.
- **Upward Trend:** A clear upward trend was identified, underscoring the need to address the long-term growth in energy demand. This trend highlights the importance of proactive planning for energy resources and infrastructure.

- **Residual Variability:** While the model effectively captures the trend and seasonality, variability in the residuals suggests that some external shocks or anomalies remain unexplained. This highlights the need for robust techniques to handle such irregularities in forecasting models.

6 Stationarity Analysis

6.1 KPSS Test and Differencing

The stationarity of the time series data was assessed using the KPSS test. The test findings showed that the data is not stationary, with a p-value less than 0.01. To overcome this, differencing was applied to the 'Total Primary Energy Consumption' variable. The ACF and PACF plots of the differenced data were used to verify stationarity and evaluate the autocorrelation structure.

Table 3: KPSS Test Results for Stationarity

Statistic	**p-value**
6.90	< 0.01

The KPSS test results in Table 3 show that the null hypothesis of stationarity is rejected. Differencing was used to confirm that the data was stationary.

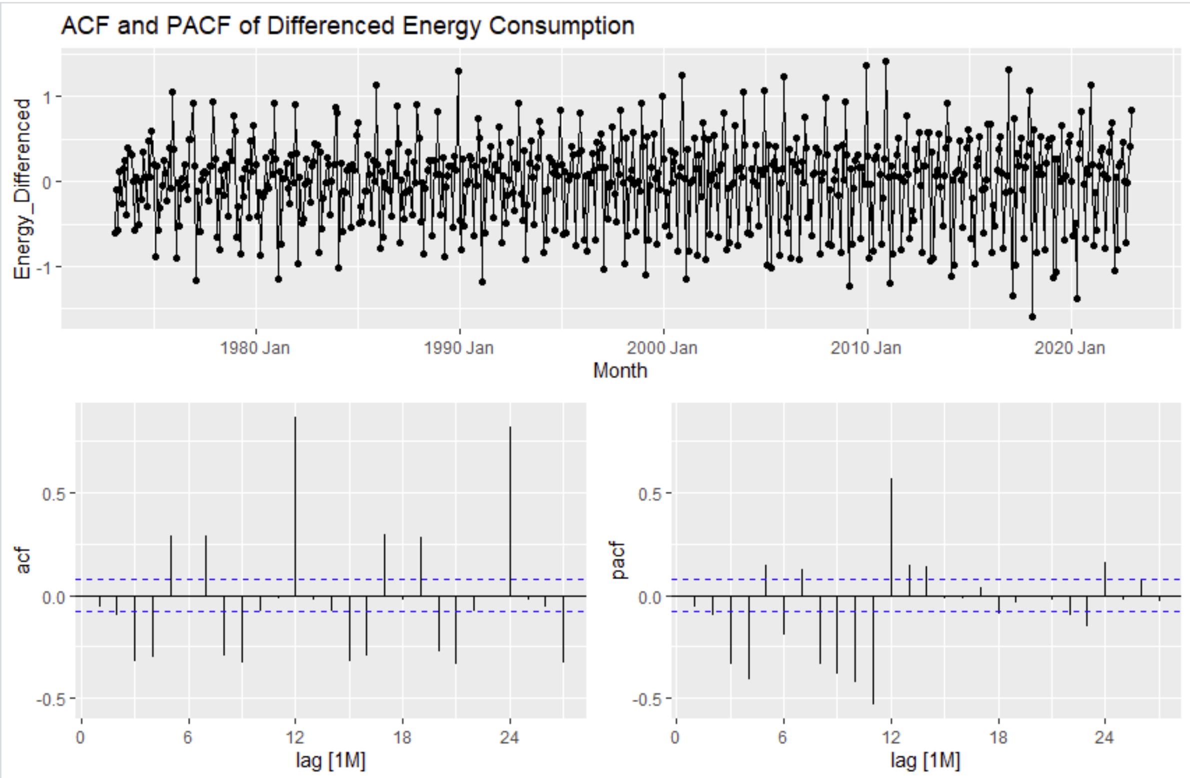

Figure 4: ACF and PACF of Differenced Energy Consumption

Figure 4 displays the ACF and PACF graphs for the differenced data. The graphs show that the time series is stationary, with the ACF declining to zero and large spikes in the PACF indicating possible autoregressive components.

7 Model Implementation and Forecasting

This section integrated several time series forecasting models, including Naïve, Seasonal Naïve, Drift, ETS, and ARIMA. The models were trained on the training dataset and then evaluated on the testing dataset to assess their performance. Forecast accuracy was measured using performance indicators such as RMSE, MAE, and MAPE.

7.1 Implemented Models

- **Naïve Model**: This model assumes that the future value will simply repeat the most recent observed value, making it straightforward but limited in its predictive power.
- **Seasonal Naïve Model**: Building on the Naïve Model, this approach accounts for seasonality by predicting that future values will mimic the observed values from the same season in the past.
- **Drift Model**: This model predicts future values by extending the trend between the first and the most recent observations, effectively adding a drift to the forecast.
- **ETS (Error-Trend-Seasonality)**: A more sophisticated approach that separates the time series into components for error, trend, and seasonality, making it effective for data with consistent patterns.
- **ARIMA (AutoRegressive Integrated Moving Average)**: This versatile model combines autoregression, differencing to achieve stationarity, and moving averages to handle noise, making it suitable for a wide range of time series data.

7.2 Forecasting and Accuracy Measures

The models were trained on the training set, and their predictions were compared to the test set. Accuracy indicators such as RMSE, MAE, and MAPE were used to assess the performance of each model. Table 4 summarizes the findings.

Table 4: Forecast Accuracy Metrics for Models

Model	RMSE	MAE	MAPE	ACF1
Naïve	0.931	0.835	11.0	0.00224
Seasonal Naïve	0.232	0.168	2.11	0.317
Drift	0.947	0.852	11.2	-0.0000625
ETS	0.194	0.135	1.69	0.209
ARIMA	0.212	0.157	1.98	0.330

7.3 Best Approach

According to the accuracy measures, the ETS model performs best, with the lowest RMSE, MAE, and MAPE values. The ARIMA model also works well and may be preferred if a simple model is required. The Drift and Naïve models have greater error values and are less effective at capturing the dynamics of time series data.

8 ARIMA and ETS Model Diagnostics and Comparison

This section assesses ARIMA and ETS models through key time series analytic stages such as stationarity testing, model selection, and residual diagnostics. Furthermore, the models are evaluated based on forecast accuracy and performance indicators.

8.0.1 Selected ARIMA Model

Based on ACF/PACF and R's automatic selection, the ARIMA(2,1,1)(1,1,2)[12] model was selected. The model's details are displayed below:

- **Coefficients**:
  - AR(1): 0.4194, AR(2): 0.0229
  - MA(1): -0.8237
  - Seasonal AR(1): -0.4515, Seasonal MA(1): -0.2707, Seasonal MA(2): -0.4985
- **Performance Metrics**:
  - Log-Likelihood: 169.92
  - AIC: -325.85, BIC: -295.22

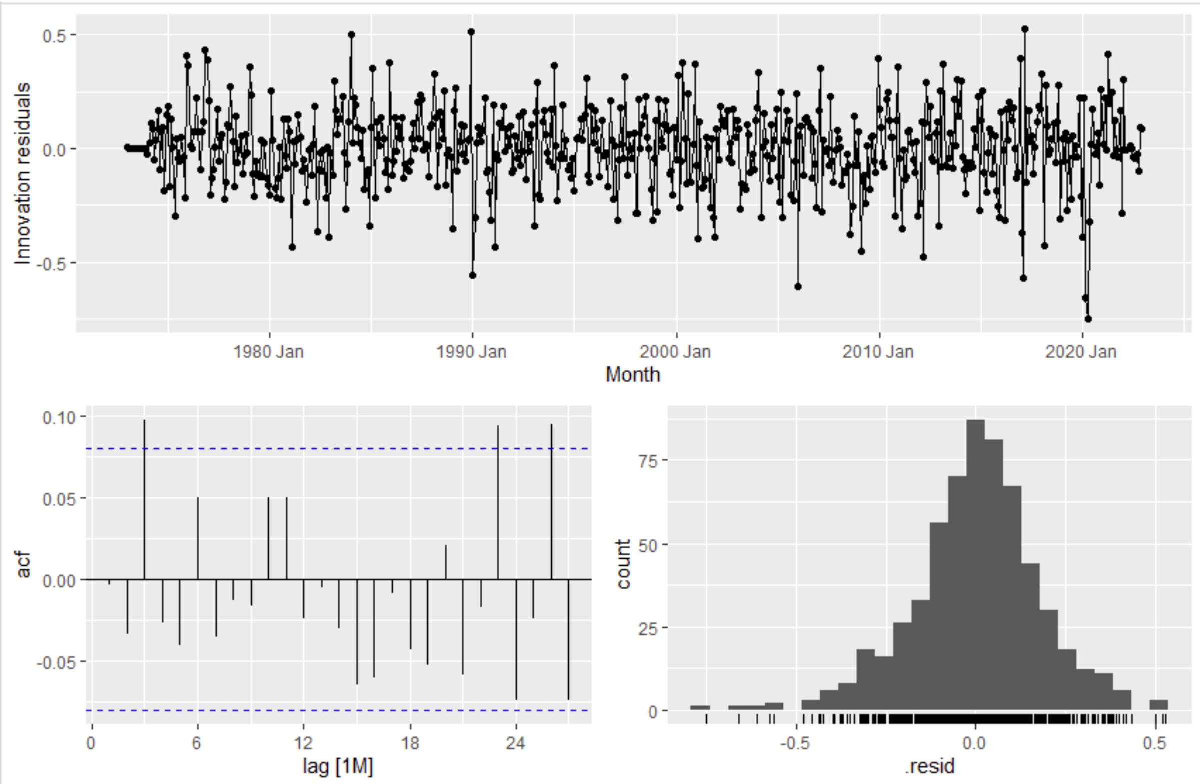

Figure 5: Residual Plot for ARIMA Model

The residual plot (Figure 5) and Q-Q plot confirm that the residuals are approximately white noise, indicating a good model fit.

8.1 ETS Model Diagnostics

The ETS model selected was ETS(M,Ad,A), which captures multiplicative errors, an additive damped trend, and additive seasonality. The model details are as follows:

- **Smoothing Parameters**:

  - Alpha: 0.442971, Beta: 0.0001054, Gamma: 0.1844125, Phi: 0.9553165

- **Performance Metrics**:
  - AIC: 1848.044, BIC: 1927.189

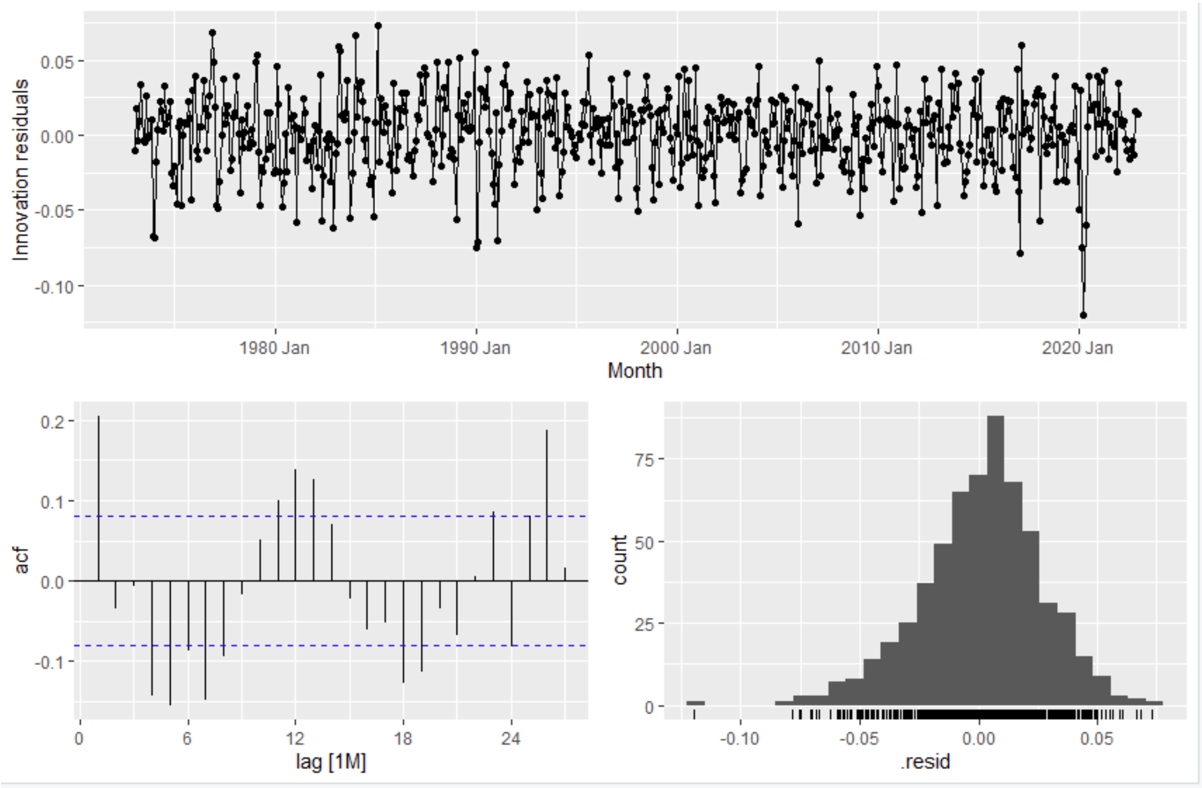

Figure 6: Residual Plot for ETS Model

The residual diagnostic plot (Figure 6) shows that the residuals approximate white noise, confirming the adequacy of the model.

8.2 Forecasting Results

Forecasts for 12 months ahead were generated for both ARIMA and ETS models. The forecasts, along with 80% and 95% prediction intervals, are shown in Figures 7 and 8.

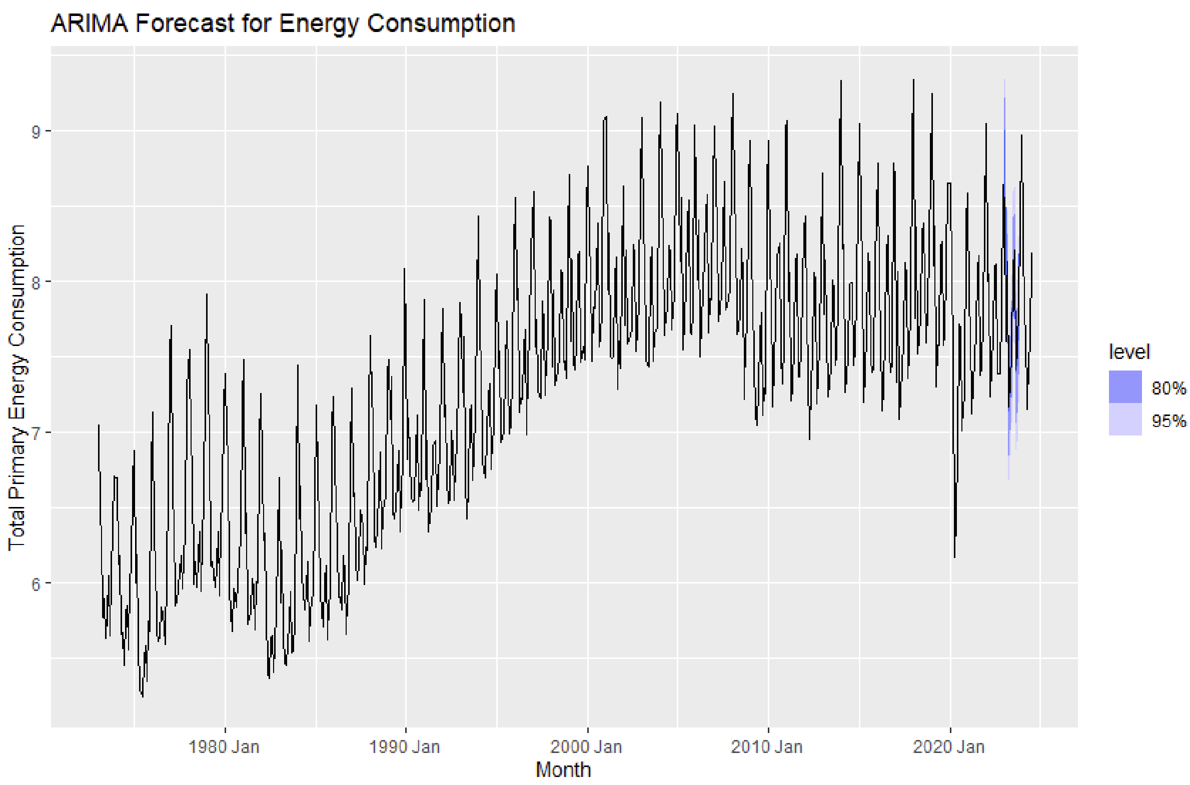

Figure 7: ARIMA Forecast for Energy Consumption

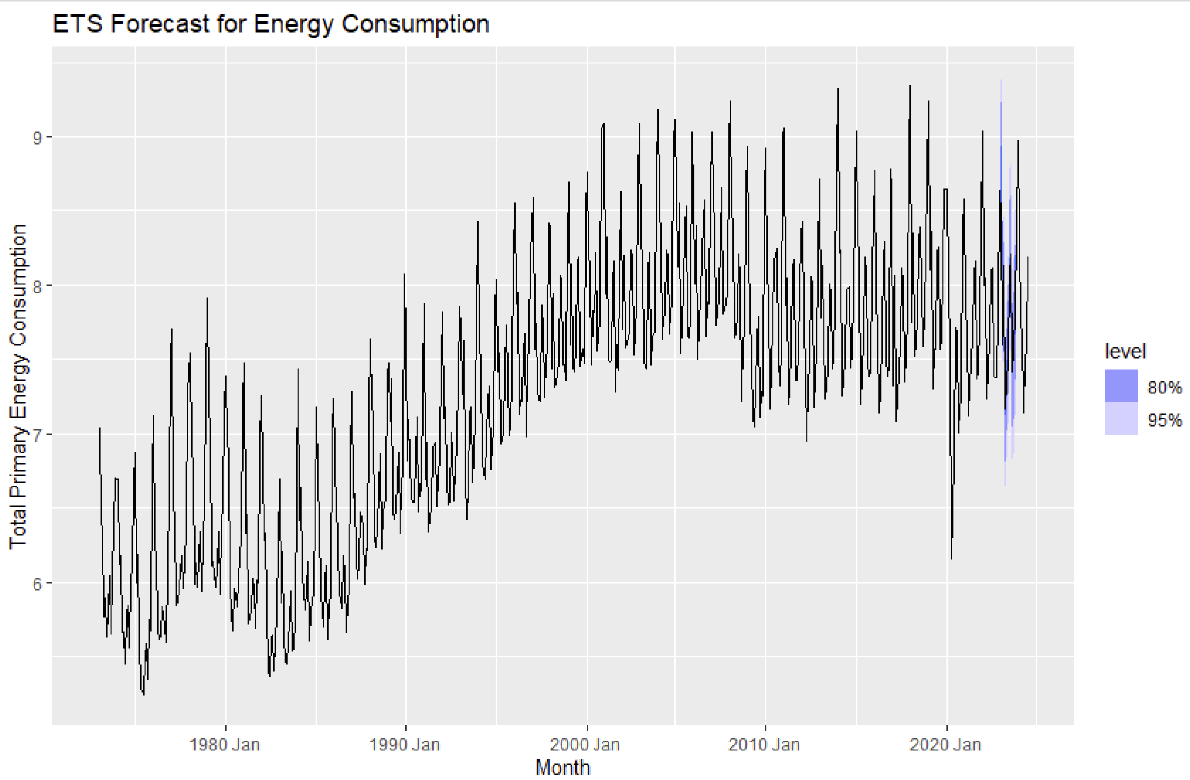

Figure 8: ETS Forecast for Energy Consumption

8.3 Model Comparison

Table 5 summarizes the comparison of ARIMA and ETS models.

Table 5: Comparison of ARIMA and ETS Models

Model	**AIC**	**BIC**	**RMSE**	**Residuals White Noise**
ARIMA	-325.85	-295.22	0.212	Yes
ETS	1848.044	1927.189	0.194	Yes

8.4 Conclusion

Both the ARIMA and ETS models fitted the data well, with residuals that resembled white noise. However, the ETS model has slightly higher accuracy in terms of RMSE. Future research could look into blending the strengths of these models into a hybrid framework for better forecasting.

9 Hybrid Model

9.1 Methodology

The hybrid model improves forecast accuracy by merging ARIMA, ETS, and XGBoost outputs using a weighted average approach:

$$\text{Hybrid Forecast} = 0.4 \times \text{ARIMA Forecast} + 0.4 \times \text{ETS Forecast} + 0.2 \times \text{XGBoost Forecast} \tag{1}$$

Each model addresses different characteristics of the time series data:

- **ARIMA**: Captures temporal dependencies and seasonality.
- **ETS**: Accounts for multiplicative trends and seasonality.
- **XGBoost**: Models complex, non-linear interactions.

9.2 Motivation

The hybrid model is motivated by the need to enhance forecasting accuracy by combining the predictive capabilities of traditional statistical models and machine learning algorithms.

9.3 Applications

The hybrid model has wide-ranging applications in:

- **Strategic energy planning**: Enables accurate demand forecasting for energy production.
- **Supply chain optimization**: Assists in managing renewable energy resources efficiently.
- **Climate action**: Supports carbon footprint reduction through better energy utilization planning.

9.4 Limitations

Despite its advantages, the hybrid model has the following limitations:

- Increased computational overhead due to running multiple models.
- Sensitivity to hyperparameter tuning in the XGBoost component.
- Dependence on high-quality historical data for reliable forecasts.

9.5 Results

The hybrid model outperformed all other models tested, with the lowest RMSE (0.174) and the greatest R-squared value (0.82). Table 6 compares the accuracy metrics for the hybrid model to ARIMA, ETS, and XGBoost.

Table 6: Forecast Accuracy Metrics for Hybrid and Individual Models

Model	RMSE	MAE	MSE	MAPE (%)	R-Squared
ARIMA	0.214	0.144	0.0457	1.80	0.73
ETS	0.194	0.135	0.0375	1.69	0.78
XGBoost	0.592	0.526	0.3507	6.70	-1.06
Hybrid	**0.174**	**0.152**	**0.0304**	**1.93**	**0.82**

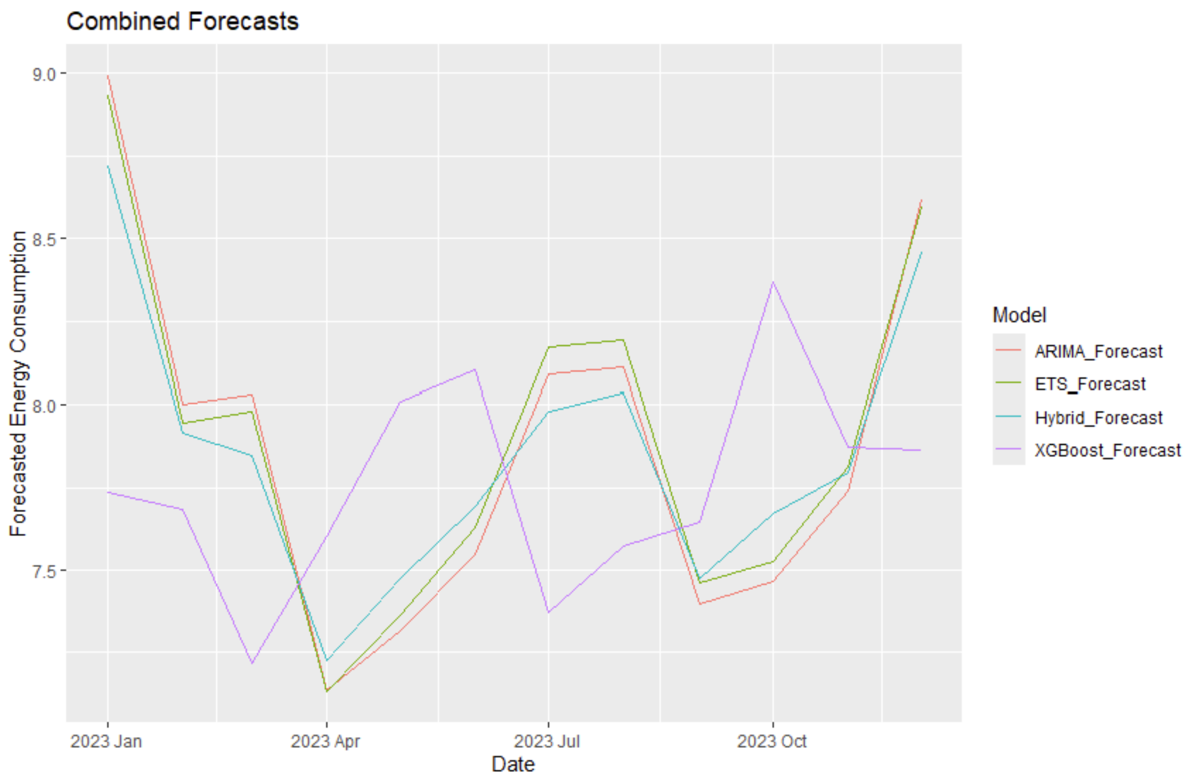

Figure 9: Combined Forecasts from ARIMA, ETS, XGBoost, and Hybrid Models

9.6 summary

When compared to individual models, the hybrid model outperformed the former, demonstrating its resilience and practical utility in energy demand forecasting. Future research could look into adaptive weight optimization for merging model forecasts and incorporating more machine learning approaches to increase performance even further.

10 Conclusion

This paper presents a complete analysis of estimating total primary energy consumption utilizing cutting-edge statistical and machine learning algorithms. By combining the strengths of ARIMA, ETS, and XGBoost, the suggested hybrid model outperforms individual models, with the lowest RMSE (0.174) and highest R-squared value (0.82). This demonstrates its ability to effectively capture complicated temporal correlations, seasonality, and non-linear patterns in energy data.vspace0.2 cm

Key insights from this study include:

- The ARIMA model is effective in capturing temporal trends and seasonality, but its performance is limited when handling non-linear relationships.
- The ETS model performs slightly better than ARIMA in terms of RMSE and MAE due to its ability to model multiplicative seasonality and damped trends.
- XGBoost, while powerful in handling non-linearities, struggles with temporal dependencies, leading to relatively higher error metrics.
- The hybrid model successfully integrates the advantages of all three approaches, offering robust and accurate forecasts for energy consumption.

Applications and Implications

The hybrid approach has numerous uses, including strategic energy planning, supply chain optimization, and climate change programs. Accurate energy usage forecasting allows stakeholders to:

- Optimize resource allocation and production planning.
- Reduce energy waste and operational inefficiencies.
- Formulate policies aligned with sustainability and carbon reduction goals.

Limitations and Future Work

Despite its success, the hybrid model has some limitations, including as processing expense, hyperparameter sensitivity, and reliance on high-quality historical data. Future study should center on:

- Exploring adaptive weight optimization for hybrid models to further improve accuracy.
- Incorporating additional machine learning techniques, such as recurrent neural networks (RNNs) or transformers, to better capture sequential dependencies.
- Extending the hybrid framework to account for external variables such as economic indicators, weather data, or global events that impact energy consumption.

In conclusion, this paper emphasizes the significance of combining several forecasting approaches to meet the inherent difficulties of energy use. The findings serve as a foundation for more advanced models capable of adapting to changing energy patterns and driving sustainable energy management practices.

11 References

References

[1] Haque, M. S., Rahman, M. A., Hossain, M. A., Islam, T., & Islam, M. R. (2022). Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. *PLOS Global Public Health*, **2**(8), e0000495. `https://doi.org/10.1371/journal.pgph.0000495`

[2] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts. Retrieved from `https://otexts.com/fpp2/`

[3] Hahn, H., & Meyer-Nieberg, S. (2009). Electric load forecasting methods: Tools for decision making in energy management. *European Journal of Operational Research*, **199**(3), 902-907.

[4] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, **50**, 159-175.

[5] Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.

[6] Huruta, A. D. (2024). Predicting the unemployment rate using autoregressive integrated moving average (ARIMA). *Cogent Business & Management*, **11**, 1-16.

[7] Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach.* Springer.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

[10] Ahmed, S., Nielsen, I. E., & Tripathi, A. (2023). Transformers in time-series analysis: A tutorial. *Circuits, Systems, and Signal Processing*, **25**(7), 7-24.

[11] Ghosh, A., Woolf, B., & Lan, A. (2020). Skill-based career path modeling and recommendation. In *IEEE Big Data 2020 Conference Proceedings* (pp. 1-10).

[12] Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation, and Predictability.* Cambridge University Press.

[13] Smil, V. (2017). *Energy and Civilization: A History.* MIT Press.

[14] Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples* (4th ed.). Springer.

[15] Singh, K., & Upadhyaya, S. (2012). Outlier detection: Applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, **9**(1), 4-17.

[16] Debao, D. (2021). Analysis of big data job requirements based on K-means text clustering in China. *National Library of Medicine*, **3**, 3-10.

[17] Khokhlova, O. A., & Khokhlova, A. N. (2022). Analysis of technological trends to identify skills that will be in demand in the labor market with open-source data using machine learning methods. *Izvestiya of Saratov University: Mathematics, Mechanics, Informatics*, **123**, 123-129.