Analyzing Energy Consumption Trends by State and Sector in the U.S.

Konstantinos Sarinopoulos
Data Science
ksarinop@umich.edu

Neil Anand Mankodi
Data Science
nmankodi@umich.edu

Abstract*—Understanding energy consumption trends is vital for fostering sustainable development and efficient resource management. This project introduces a robust suite of time-series analysis tools designed to unravel complex energy usage patterns across U.S. states and sectors. With over 200 time series representing 50 states and four sectors, we demonstrate the flexibility and power of our framework by applying it to Michigan as a case study. However, the methodologies we propose are adaptable and scalable, capable of handling diverse scenarios across any state or sector. Our forecasting models, driven by hyperparameter-optimized ARIMA and SARIMA techniques, excel in capturing intricate seasonal and trend dynamics, delivering precise predictions with exceptional accuracy. For Michigan, one of our models achieved an excellent R^2 of 0.958, showcasing their effectiveness in predicting energy demands. These forecasts empower stakeholders to make informed decisions on energy production, storage, and policy planning. Complementing this, our time-series clustering analysis harnesses Euclidean distance and Dynamic Time Warping (DTW) to uncover state-level consumption similarities. By identifying optimal clusters through silhouette score maximization, we reveal regional trends that encourage collaboration and shared energy strategies. Visualized through state-level maps, these clusters provide intuitive and actionable insights for policymakers and utility companies. Our project highlights the immense potential of advanced data mining techniques in tackling energy challenges, offering an innovative and adaptable approach to drive sustainable energy practices and informed decision-making.

I. Introduction

Energy consumption in the United States has grown exponentially over the past several decades, reflecting the nation's economic expansion, technological advancements [1], and population growth. From powering homes and industries to fueling transportation, commerce, and, more recently, the growing demands of AI, energy remains central to modern life. Yet, this increasing demand poses significant challenges, especially as traditional methods of energy production—predominantly reliant on fossil fuels—strain finite natural resources and contribute to environmental degradation. The urgency of transitioning to sustainable energy production, efficient consumption practices, and equitable distribution systems has never been more apparent. Without such measures, the consequences of unsustainable energy use—ranging from severe climate change impacts to economic instability—could irreversibly affect both the environment and society.

To address these challenges, advanced analytics and data-driven insights have become indispensable. By understanding energy consumption trends and identifying actionable patterns, stakeholders such as policymakers, utility companies, and governments can make informed decisions to optimize energy use, enhance efficiency, and plan for future demand. Our project aims to contribute to this critical effort by developing a suite of time-series analysis tools designed to uncover intricate energy consumption trends across the United States, providing both high-level and granular insights.

Our project focuses on two core methodologies: time-series forecasting and time-series clustering. The forecasting models employ hyperparameter-optimized ARIMA and SARIMA techniques to analyze energy consumption patterns in Michigan across four key sectors: residential, commercial, industrial, and transportation. These models accurately capture complex seasonal trends and long-term variations, achieving remarkable predictive accuracy. For instance, our industrial sector model achieved an R^2 of 0.958, with an RMSE of 42,152 and an MAE of 15,379, underscoring its reliability and precision. The forecasts empower stakeholders to make informed decisions regarding energy production, storage, and distribution to meet future demand effectively.

In parallel, our time-series clustering analysis extends the scope beyond any one chosen state, analyzing energy consumption across all 50 states. Using similarity measures such as Euclidean distance and Dynamic Time Warping (DTW), we clustered states based on their energy usage trends. By maximizing silhouette scores to determine optimal clusters, we identified meaningful groupings that reveal shared energy behaviors and consumption patterns. These insights facilitate regional collaboration and unified energy policies, ensuring more efficient resource allocation and sustainable practices.

The tools developed in this project are versatile and scalable, capable of analyzing energy trends for any state or sector. They provide robust support for addressing critical questions about energy sustainability, efficiency, and policy development.

The rest of this report is structured as follows:

1) Data - This section describes the energy consumption dataset, its source, structure, and the preprocessing steps taken to prepare it for analysis.
2) Methodology - Here, we outline the techniques used for time-series forecasting (ARIMA and SARIMA) and clustering (Euclidean distance, DTW). We also explain model validation, evaluation metrics, and the computational framework to ensure reliable results.
3) Results - The results section presents the forecasting outcomes for Michigan's energy consumption and the clustering analysis across states. We highlight key performance metrics and provide visualizations to show trends and similarities in energy usage.
4) Conclusions - This section summarizes the insights from the analysis, emphasizing the practical implications for sustainable energy practices, policy making, and the effectiveness of our models in uncovering key trends.
5) Future work - The final section explores potential directions for extending this project.

II. Data

The dataset originates from the U.S. Energy Information Administration (EIA) and is derived from the EIA-861 report [2]. It provides detailed insights into energy sales, revenue, and prices across all U.S. states, segmented by sector. The four sectors covered are:

- Residential: Energy used by private households for personal activities such as heating, cooling, and powering appliances.
- Commercial: Energy consumed by buildings and facilities providing services, including offices, retail stores, etc.
- Industrial: Energy used in manufacturing, mining, agriculture, and other large-scale industrial operations.
- Transportation: Energy used for transportation purposes, including electric trains, subways, and electric vehicles.

Key Metrics

- Revenue: Measured in thousands of U.S. dollars, representing the total earnings from energy sales.
- Sales: Recorded in megawatt-hours (MWh), indicating the total volume of energy sold.
- Customer Count: The total number of customers receiving energy supply within each sector.
- Price: Expressed in cents per kilowatt-hour (cents/kWh), representing the average cost of energy for end-users.

The dataset spans a time series from 2010 to 2024, with data reported on a monthly basis. This temporal granularity allows for both long-term trend analysis and seasonal pattern identification. This dataset provides a robust foundation for analyzing state-level energy dynamics, sector-specific trends, and broader patterns in energy consumption, pricing, and revenue generation.

III. Methodology

A. *Time Series Forecasting*

This section outlines the approach used to develop and evaluate the time series forecasting for predicting energy consumption. The steps are presented in the order of their execution to maintain consistency and clarity.

Models:
ARIMA (AutoRegressive Integrated Moving Average): ARIMA is a widely used model for time series forecasting. It is designed for data that exhibit trends but no strong seasonality. The model is characterized by three parameters:

- p: The order of the autoregressive (AR) part, representing the number of lag observations.
- d: The degree of differencing, indicating how many times the data need to be differenced to make it stationary.
- q: The order of the moving average (MA) part, representing the number of lagged forecast errors in the prediction.

SARIMA (Seasonal ARIMA): SARIMA extends ARIMA to handle seasonal patterns in time series data. It incorporates the same parameters as ARIMA but adds seasonal components:

- P, D, Q, S: These parameters represent the seasonal autoregressive (SAR), seasonal differencing (SD), seasonal moving average (SMA), and the period of seasonality (S), respectively.

Data Split:
For training and testing the models, the dataset was split chronologically. The first 80% of the data was used for training the models, and the remaining 20% was used for testing. This chronological split is critical for time series analysis to ensure that the testing data remains unseen and accurately reflects real-world forecasting conditions.

Stationarity Check:
A key assumption for ARIMA and SARIMA models is that the time series data must be stationary. Stationarity implies that the statistical properties of the data, such as mean and variance, are constant over time. The following steps were used to check for stationarity:

- Rolling Mean and Rolling Standard Deviation: Plots were generated to visualize whether the mean and standard deviation remained constant over time.
- ADF Test (Augmented Dickey-Fuller Test): This test was conducted to statistically check for stationarity. A p-value less than 0.05 indicates that the series is stationary.
- KPSS Test (Kwiatkowski-Phillips-Schmidt-Shin Test): This test was used to further validate the stationarity of the series. A p-value greater than 0.05 indicates stationarity.

Stationarity Transformation (if necessary):
If the time series data was found to be non-stationary, several techniques were employed to achieve stationarity:

- First-order Differencing: This technique involves subtracting the previous observation from the current observation to remove trends.
- Seasonal Differencing: For data with seasonal patterns, we differenced the series at the seasonal period.

ACF and PACF Plots:
Once stationarity was achieved, we analyzed the AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) plots. These plots are crucial for identifying the characteristics of the time series data and help inform the model-building process.

- ACF Plot: Helps identify the appropriate q (MA) parameter by showing the correlation between a time series and its lagged values.
- PACF Plot: Helps determine the p (AR) parameter by showing the partial correlation between the series and its lags.

Hyperparameter Optimization:
To optimize the ARIMA and SARIMA models, we used the auto arima library, which automates the process of selecting the best model parameters. This method performs a smart grid search over possible values of the parameters (p, d, q, s for ARIMA, and p, d, q, s, P, D, Q, S for SARIMA). The optimization is based on minimizing the Akaike Information Criterion (AIC), which balances model fit and complexity. A lower AIC indicates a better-fitting model. We also implemented a manual grid search as a sanity check and found that the automated method was more efficient and yielded similar results.

Model Evaluation:
The best-performing ARIMA and SARIMA models were evaluated using the following performance metrics:

- RMSE (Root Mean Squared Error): Measures the average magnitude of errors between predicted and actual values, with lower values indicating better model performance.
- MAE (Mean Absolute Error): Measures the average of the absolute differences between predicted and actual values, indicating how far off are the predictions.
- R^2 (Coefficient of Determination): Represents the proportion of variance in the dependent variable that is explained by the model. A higher R^2 indicates a better fit.

These metrics were computed on the test set to assess the predictive power of the models, ensuring that the selected model accurately represents the energy consumption patterns and can reliably forecast future values.

Final Model Selection & Forecasting:
After evaluating the models, the best-performing model was selected based on the lowest RMSE, MAE, and highest R^2 scores. This model was then used to generate forecasts on the test set, as well as for a one-year forecast into the future.

Visualization of Results:
The final step involved plotting the true values, predicted values, and the one-year ahead forecasts. The plots visually demonstrate how well the selected model captured historical trends and seasonal patterns in the data, as well as its ability to forecast future energy consumption.

B. Time Series Clustering

Time series clustering was used to segment U.S. states based on their energy consumption, revenue, and pricing trends. The methodology consists of five major steps: calculating similarity measures, determining the optimal number of clusters, performing sector-specific clustering, conducting post-clustering analysis, and exploring alternative cluster configurations [3].

Similarity Measures:
To cluster the time series data, similarity measures were calculated to quantify the relationship between states' energy profiles. Two measures were employed:

- Euclidean Distance: A simple measure that calculates the direct, point-to-point distance between time series. It serves as a baseline for identifying simple patterns.
- Dynamic Time Warping (DTW): A more flexible measure that aligns time series with temporal shifts, capturing complex and misaligned patterns.

Rationale: Using both Euclidean Distance and DTW ensures that the clustering captures both straightforward and nuanced relationships, providing a robust basis for subsequent analysis.

K-Means Clustering and Optimal Group Determination:
The K-Means clustering algorithm was applied to group states based on their similarity measures. This algorithm partitions data into clusters by minimizing within-cluster variance. To determine the optimal number of clusters, Silhouette Scores were calculated for a range of cluster counts. This metric evaluates how well each point fits within its cluster versus others. A score analysis determined that two clusters were optimal for all sectors.

Rationale: Identifying the optimal number of clusters ensures that the data is grouped meaningfully, balancing interpretability and detail.

Sector-Specific Clustering:
Once the optimal number of clusters was determined, K-Means clustering was performed for each sector:

- Residential
- Commercial
- Industrial
- Transportation

Rationale: Clustering each sector separately allows for a detailed exploration of sector-specific trends and patterns in energy consumption, revenue, and pricing.

Post-Clustering Analysis:
Post-clustering analysis was conducted to validate and interpret the clusters:

- Statistical Summaries: Key metrics (e.g., means, medians, variances) were calculated for each cluster to understand their characteristics.
- Geographic Visualization: Clusters were plotted on a U.S. map to explore their spatial distribution and trends.

Rationale: Statistical and geographic analysis enhances understanding of the clustering results, providing both numerical and visual validation of the groupings.

Alternative Clustering with Non-Optimal Clusters:
To assess the robustness of the clustering approach, the K-Means algorithm was repeated using non-optimal numbers of clusters for each sector. Statistical summaries and geographic distributions of these clusters were compared with the optimal clustering results.

Rationale: This step tests the sensitivity of the clustering process, ensuring that the chosen number of clusters is well-grounded and reliable.

Visualization of Results:
The final clustering results were visualized through:

- Cluster Summary Plots: Highlighting statistical differences between clusters.
- Geographic Maps: Displaying the spatial distribution of clusters across U.S. states.

Rationale: Visualization provides a clear, intuitive understanding of the clustering results, emphasizing their practical implications and regional trends.

IV. Results

A. *Time Series Forecasting*

Overview of Model Performance Metrics:
The table below presents the key performance metrics — RMSE, MAE, and R^2—alongside the target column's standard deviation, offering a concise comparison of the models' accuracy and fit relative to the variability in the target values.

Sector	R2	RMSE	MAE	SD Target
Residential	0.75	209,445	151,761	483,162
Commercial	0.83	98,414	80,545	276,385
Industrial	0.96	42,152	15,379	208,203
Transport	0.29	74	55	137

TABLE I
Forecasting Results

Michigan's Residential Sector:
The ARIMA model successfully captured the seasonal patterns in residential energy consumption. Future forecasts indicate a predictable cyclical trend, with seasonal dips expected at the end of each year followed by a rise at the beginning of the next. Notably, peak consumption appears to be gradually declining, suggesting a potential shift in energy usage behavior. Recommendations for this sector include optimizing energy production to align with the predictable cyclical patterns and exploring energy efficiency initiatives to address the declining peaks.

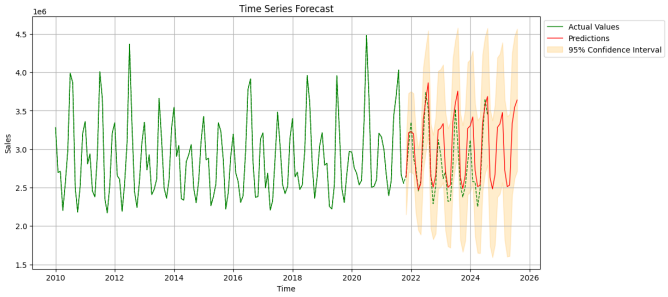

Fig. 1. Forecasting - Residential Sector

Michigan's Commercial Sector:
The SARIMA model effectively replicated the unique "triple dip" pattern observed in the commercial sector's energy consumption. Future forecasts project a continuation of this trend into early 2025, with a sharp rise towards the latter half of the year. These insights are critical for planning energy production and storage, ensuring sufficient resources during peak periods while preparing for significant declines. This sector benefits from precise, seasonally aware models to accommodate its unique fluctuations.

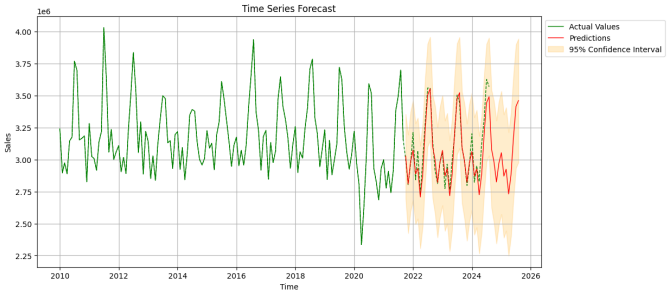

Fig. 2. Forecasting - Commercial Sector

Michigan's Industrial Sector:
The SARIMA model successfully addressed the faint but complex seasonal trends in industrial energy consumption. Future forecasts suggest a continuation of the "double drop" pattern followed by increased usage through 2025. Adjusting for model bias significantly improved the predictive accuracy, underscoring the importance of validation during the modeling process. These forecasts can guide resource allocation during periods of high demand and steep declines, optimizing industrial sector energy management.

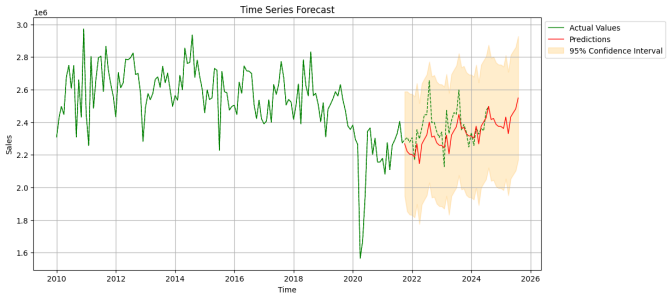

Fig. 3. Forecasting - Industrial Sector

Michigan's Transport Sector:
The transport sector's time series presented challenges due to its smaller magnitude and lack of clear seasonal trends. The ARIMA model predicted steady energy consumption, reflecting limited growth or stagnation in this sector. Recommendations focus on identifying external factors driving this stability and exploring opportunities to enhance energy efficiency tailored to transport systems.

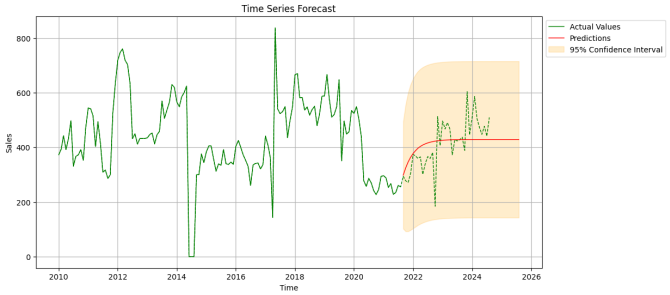

Fig. 4. Forecasting - Transport Sector

B. Time Series Clustering

Overview of Cluster Analysis Metrics:
The table below summarizes the clustering performance metrics for each sector, including the optimal number of clusters and silhouette scores. These metrics highlight the separation quality and robustness of the clustering models.

Residential Sector Clustering:
The analysis revealed two primary clusters for residential energy demand, with a silhouette score of around 0.8. High-demand states—California, Texas, New York, and Florida—are distinctly separated from lower-demand states, driven by population size and economic activity. Non-optimal clustering revealed geographical patterns, where energy demand varies between eastern and western states. Further segmentation into three or four clusters slightly reduced separation quality but offered deeper regional insights. Below is a 2-means post-cluster analysis, we see one cluster has higher sales, revenue, and prices.

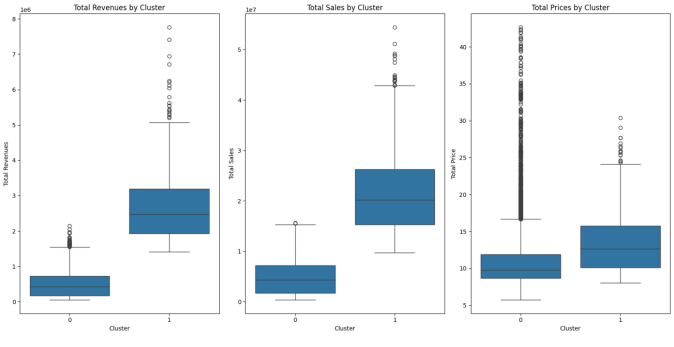

Fig. 5. Post Cluster Analysis - Residential Sector

Commercial Sector Clustering:
Similar to the residential sector, the commercial sector exhibited two well-separated clusters with a silhouette score of around 0.8. The same four high-demand states consistently formed the high-energy cluster. Alternative configurations with three or four clusters provided additional granularity but did not significantly alter key findings. Below is the cluster map of Commercial DTW with 2 groups.

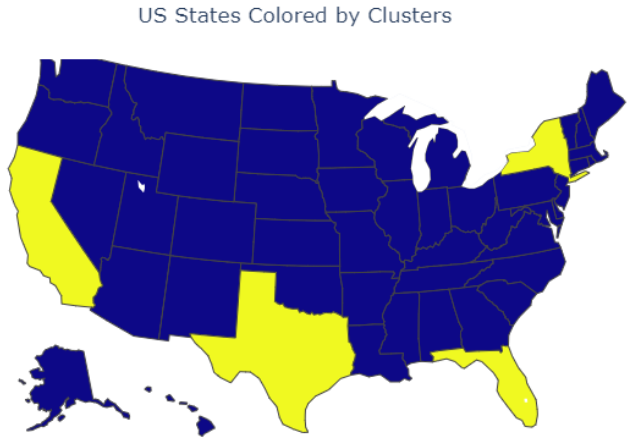

Fig. 6. Cluster Map - Commercial Sector

Industrial Sector Clustering:
Industrial energy demand followed the same clustering patterns as residential and commercial sectors. The results demonstrated strong inter-sectoral correlations, emphasizing the influence of economic activity and population. A silhouette score of around 0.8 confirmed the robustness of the clustering. Below is the cluster map of Industrial DTW with 3 groups.

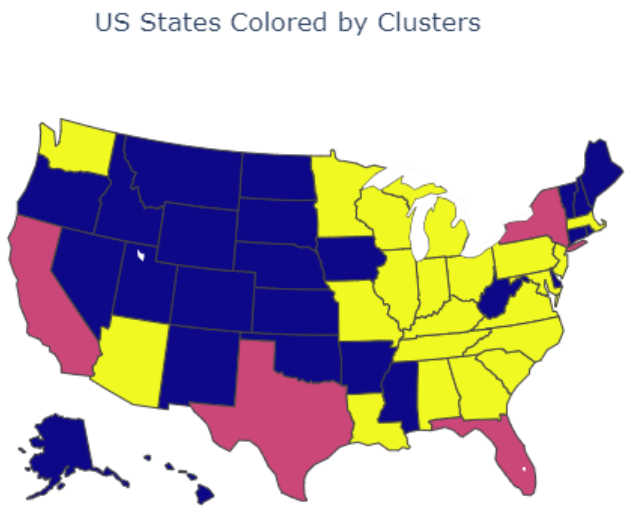

Fig. 7. Cluster Map - Industrial Sector

Transportation Sector Clustering: Transportation energy demand clustering reflected similar segmentation, with two clusters and a silhouette score of around 0.8. High-demand states once again emerged as a distinct group, underscoring the consistent role of population and economic factors across sectors. Below is the cluster map of Transportation DTW with 4 groups

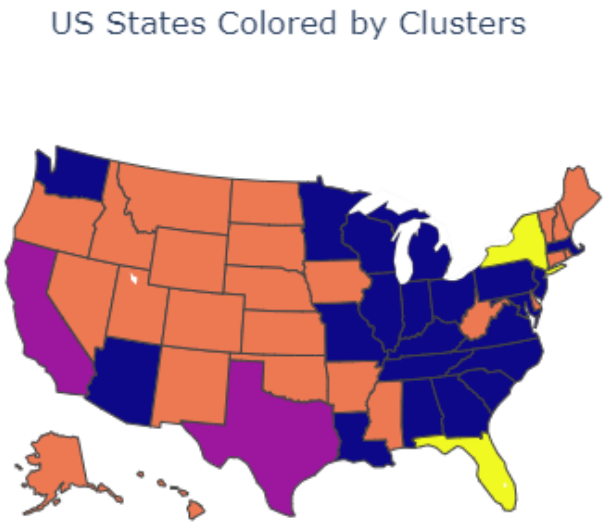

Fig. 8. Cluster Map - Transport Sector

Cross-Sectoral Insights and Robustness: The consistent clustering patterns across sectors underscore the significant influence of population and economic activity on energy demand. The use of different similarity measures (DTW and Euclidean distance) yielded identical results, affirming the robustness of the clustering methodology. While energy demand correlates strongly across sectors, energy pricing appears to be influenced more by economic factors than direct demand.

Overall Recommendations:

1) Leverage clustering insights to design state-specific energy policies that address demand across all sectors.
2) Explore opportunities for cross-sectoral energy optimization in high-demand states.
3) Investigate the economic factors influencing energy pricing and explore innovative pricing models to support sustainable energy consumption.

V. Conclusion

The forecasting analysis demonstrated the effectiveness of tailored time series models in capturing sector-specific energy trends in Michigan and any other state. Residential, commercial, and industrial sectors exhibited clear seasonal patterns, enabling accurate predictions and actionable insights for optimizing energy production and storage. In contrast, the transport sector presented steady consumption with limited variability, highlighting the need for alternative approaches to uncover underlying drivers. These findings emphasize the importance of leveraging appropriate models to identify and address unique sectoral characteristics, ensuring efficient energy planning and resource allocation while identifying opportunities for long-term improvements in energy management and sustainability.

The clustering analysis effectively segmented U.S. states based on sector-specific energy demand patterns, revealing clear distinctions driven by population size and economic activity. Residential, commercial, industrial, and transportation sectors consistently identified high-demand states—California, Texas, New York, and Florida—highlighting their outsized influence on national energy consumption. Robust clustering metrics, including strong silhouette scores, validated the separation quality and offered actionable insights into energy management across regions. While energy demand correlated strongly across sectors, emphasizing the role of economic factors. These findings underscore the value of clustering techniques in understanding energy usage dynamics, enabling targeted strategies for efficient resource allocation, sustainable policy development, and long-term energy optimization.

VI. Future Work

Expanding the analysis to differentiate between renewable and non-renewable energy sources provides a deeper understanding of state-level energy dynamics and consumption trends. This approach enables an evaluation of how the energy mix evolves over time and its implications for sustainability, energy policy, and infrastructure planning. A focus on production and consumption trends can include metrics for renewable energy sources such as solar, wind, hydroelectric, geothermal, and biomass, alongside non-renewable sources like coal, natural gas, oil, and nuclear energy. Analyzing energy transition patterns, particularly shifts from non-renewable to renewable consumption, highlights progress toward sustainability goals. Additionally, examining the sector-specific energy mix offers insights into the dependence of residential, commercial, industrial, and transportation sectors on different energy types, revealing opportunities for renewable integration and reductions in non-renewable reliance. Geographic variations in energy mix, such as hydroelectric dependency in the Pacific Northwest or coal reliance in the Midwest, can uncover regional trends and inform strategies for shared renewable infrastructure investments among states with similar energy profiles.

Incorporating a broader range of external variables enhances clustering and forecasting models by capturing the nuanced influences on energy consumption. Socioeconomic indicators, including population growth and density, income levels, employment rates, and urbanization trends, provide context for understanding regional and sectoral energy demand. Climate data, such as temperature patterns, heating and cooling degree days, and extreme weather events, allow for the assessment of seasonal energy consumption fluctuations. Accounting for geographic differences through climate zones further contextualizes variations in energy use for heating, cooling, and transportation needs. This comprehensive inclusion of external variables improves the models' ability to reflect the diverse drivers of energy usage, supporting more targeted energy management and policy-making tailored to state-specific circumstances.

References

[1] United States: What sources does the country get its energy from? https://ourworldindata.org/energy/country/united-stateswhat-sources-does-the-country-get-its-energy-from.

[2] Annual Electric Power Industry Report, Form EIA-861 detailed data files. https://www.eia.gov/electricity/data/eia861/.

[3] Liao, T. Warren. "Clustering of time series data—a survey." *Pattern Recognition*, vol. 38, no. 11, 2005, pp. 1857–1874. Elsevier.

Team Contribution

Konstantinos Sarinopoulos:

- Data Sourcing
- Data Cleaning & Processing
- Time Series Clustering
- Sector Based Clustering Analysis
- Final Report Preparation
- Final Video Preparation

Neil Anand Mankodi:

- Data Sourcing
- Data Cleaning & Processing
- Time Series Forecasting
- Sector Based Forecast Analysis
- Final Report Preparation
- Final Video Preparation