A black background with white text

Description automatically generated with low confidence

MSc Data Science Project

7PAM2002-0901-2024

Department of Physics, Astronomy and Mathematics

**Data Science FINAL PROJECT REPORT**

**Project Title:**

Predictive Analysis of Household Electricity Consumption Using Time-Series Modelling

**Student Name and SRN:**

Aishwarya Sukumaran 22058088

Supervisor: Vidas Regelskis

Date Submitted: 05.01.2025

Word Count: 7742

GitHub address: [GitHub address link](https://github.com/AishwaryaSukumaran/Final-Project_22058088/blob/main/energy_data_v3.ipynb)

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science at the University of Hertfordshire.

I have read the guidance to students on academic integrity, misconduct and plagiarism information at [Assessment Offences and Academic Misconduct](https://www.herts.ac.uk/__data/assets/pdf_file/0007/237625/AS14-Apx3-Academic-Misconduct-v17.0.pdf) and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project module or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6).

I have not used chatGPT, or any other generative AI tool, to write the reportor code (other than where declared or referenced).

I did not use human participants or undertake a survey in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student SRN number: 22058088

Student Name printed: Aishwarya Sukumaran

Student signature: A close up of a paper

Description automatically generated

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

**Abstract**

This study looks at electricity use in French households from 2007 to 2010 to find ways to save energy, lower costs, and support sustainable practices. Seasonal patterns were clear, with higher electricity use in winter for heating and lower demand in summer during vacations. The SARIMA model was used to forecast energy use and highlighted non-seasonal influences like the 2008 economic recession, which increased home energy use, and the adoption of smart meters in 2010, which encouraged off-peak energy usage.

The SARIMA model performed well, showing low errors and accurately capturing trends and patterns. The analysis found winter peaks in electricity use due to heating and festive activities, while summer had lower usage with occasional spikes during heatwaves. Residual checks showed the model worked well with consistent and random error patterns.

This study offers practical suggestions for households and energy providers. Households can save money by using energy during off-peak hours, adopting energy-efficient appliances, and using smart home technologies like thermostats. Providers can use the forecasts to plan for peak demand, improve grid stability, and promote off-peak energy use through pricing strategies. Seasonal predictions also help align renewable energy production, like solar and wind, with demand.

However, the study has some limits. The data came from a single household, so it might not apply to all households. Factors like sudden weather changes, new technologies like electric vehicles, and economic shifts were not included. Future research can use more diverse data, include external factors like weather, and try advanced models like neural networks to better capture complex patterns.

Overall, this research shows that SARIMA is a useful tool for forecasting electricity use. It provides insights that help households save money, support better grid management, and promote sustainable energy use.

**Table of Contents**

|  |  |
| --- | --- |
| **1. Introduction and Overview** | 7 |
| 1.1 The Role of Electricity in French Households | 7 |
| 1.2 Key Drivers of Electricity Consumption | 7 |
| 1.2.1 Seasonal Variations | 7 |
| 1.2.2 Daily Activity Patterns | 8 |
| 1.2.3 Impact of Modern Technology | 8 |
| 1.2.4 Cultural and Social Influences | 8 |
| 1.3 Challenges in Efficient Energy Use | 8 |
| 1.4 Justification for the Research | 9 |
| 1.5 Research Objectives and Questions | 9 |
| 1.5.1 Research Questions | 9 |
| 1.5.2 Aim & Objectives | 9 – 10 |
|  |  |
| **2. Literature Review** | 10 |
| 2.1 Energy Models for Demand Forecasting | 10 |
| 2.2 Time Series Forecasting Techniques | 10 – 11 |
| 2.3 Residual Analysis and Diagnostic Tools | 11 |
| 2.4 Advanced Techniques and Future Directions | 11 |
| 2.5 Summary of Literature Review | 11 – 12 |
|  |  |
| **3. Methodologies** | 12 |
| 3.1 Data Collection | 12 |
| 3.1.1 Key Features of the Dataset | 12 |
| 3.1.2 Importance of the Dataset | 13 |
| 3.2 Data Preprocessing | 13 |
|  |  |
| **4. Exploratory Data Analysis (EDA)** | 14 |
| 4.1 Visualizing Temporal Trends | 14 |
| 4.2 Seasonal Categorization | 14 – 15 |
|  |  |
| **5. Seasonal-Trend Decomposition using LOESS (STL)** | 15 |
| 5.1 What is STL Decomposition? | 15 |
| 5.2.Why STL Decomposition? | 15 |
| 5.3 Formula for Decomposition | 15 - 16 |
| 5.4 STL Decomposition Process | 16 |
| 5.5 Components of STL Decomposition | 17 |
| 5.6 Benefits of STL Decomposition | 18 |
| 5.7 Visualization of STL Decomposition | 18 - 20 |
|  |  |
| **6. Time Series Modelling** | 20 |
| 6.1 SARIMA Model | 20 |
| 6.1.1 Parameters of SARIMA (p, d, q, P, D, Q, s) | 21 |
| 6.1.2 SARIMA Equation | 21 |
| 6.1.3. Key Terms in SARIMA Model Summary | 22 – 23 |
| 6.2 Weekly & Monthly Train-Test and Full Dataset Forecasting Analysis Using SARIMA | 23 – 24 |
| 6.2.1 Weekly Train-test split & Full dataset forecasting | 24 – 26 |
| 6.2.2 Monthly Train-test split & Full dataset forecasting | 26 – 27 |
| 6.3 Model Evaluation | 28 |
| 6.3.1 Evaluation Metrics (MAE, RMSE) | 28 |
| 6.3.2 Evaluation Results | 28 – 29 |
| 6.3.3 Residual Analysis | 29 |
| 6.3.4 Residual Diagnostics | 29 – 31 |
| 6.3.5 Weekly Train-Test and Full Dataset Forecast SARIMA Model Summary | 31 – 32 |
| 6.3.6 Monthly Train-Test and Full Dataset Forecast SARIMA Model Summary | 33 – 34 |
| 6.3.7 Residual Analysis Results | 35 |
| 6.3.7.1 Weekly Residual Analysis | 35 – 36 |
| 6.3.7.2 Monthly Residual Analysis | 36 – 37 |
| 6.4 ACF and PACF Analysis for Residuals | 37 |
| 6.4.1 Purpose of ACF and PACF Analysis | 37 – 38 |
| 6.4.2 ACF and PACF Analysis Results | 38 |
| 6.4.2.1 Weekly Residuals | 38 – 39 |
| 6.4.2.2 Monthly Residuals | 39 – 40 |
|  |  |
| **7. Overall Results and Discussion** | 40 |
| 7.1. Comparison and Validation Against Established Research | 40 – 41 |
| 7.2. Achievement of Objectives | 41 |
| 7.3. Conclusion | 41 |
| 7.4. Practical Implications for Households and Providers | 42 |
| 7.5 Limitations and Areas for Improvement | 42 |
|  |  |
| **8. Ethical Considerations** | 43 |
| 8.1 Data Anonymization | 43 |
| 8.2 Informed Consent and Data Transparency | 43 |
| 8.3 Documentation and Methods | 43 |
| 8.4. Responsible Use of Findings | 43 |
| 8.5. Ethical Impact on Stakeholders | 44 |
| 8.6. Environmental Sustainability | 44 |
| 8.7 Data Backup and Recovery | 44 |
|  |  |
| **9. References** | 45 – 46 |
|  |  |
| **10. Appendices** | 46 |

**List of Figures & Tables**

|  |  |
| --- | --- |
| **Figure 1:** Sample Records from the Individual Household Electric Power Consumption Dataset | 12 |
| **Figure 2**: Time-series line chart showing daily, weekly, and monthly trends to highlight these variations. | 14 |
| **Figure 3:** A bar chart showing seasonal averages of electricity usage for each year (2007–2010). | 15 |
| **Figure 4:** STL decomposition to monthly electricity usage data | 19 |
| **Figure 5**: STL decomposition to weekly electricity usage data | 20 |
| **Figure 6**: Train-Test Split Forecast | 24 |
| **Figure 7**: Full Dataset Forecast | 25 |
| **Figure 8**: Train-Test Split Forecast | 26 |
| **Figure 9**: Full Dataset Forecast | 27 |
| **Table 1:** SARIMA models results | 28 |
| **Figure 10:** Simplified SARIMA Model Summary for Weekly Train-Test Dataset | 31 |
| **Figure 11:** Simplified SARIMA Model Summary for Full Weekly Dataset | 32 |
| **Figure 12:** Simplified SARIMA Model Summary for Monthly Train-Test Dataset | 33 |
| **Figure 13:** Simplified SARIMA Model Summary for Full Monthly Dataset | 33 |
| **Figure 14:** Residual analysis for weekly data shows error trends over time and frequency distribution, indicating the model's accuracy | 35 |
| **Figure 15:** Residual analysis for monthly data illustrates error trends over time and frequency distribution, assessing model performance. | 36 |
| **Figure 16:** ACF and PACF analysis for weekly residuals | 38 |
| **Figure 17:** ACF and PACF analysis for monthly residuals | 39 |

**1. Introduction and Overview**

Electricity is a vital part of daily life, powering homes, businesses, and industries. In France, household electricity consumption plays a significant role in overall energy use, influenced by factors such as weather, daily routines, technological advancements, and cultural practices. These patterns are not only shaped by individual behaviours but also reflect broader societal trends, such as the adoption of smart home technologies and the increasing use of electric vehicles. Managing electricity consumption effectively is crucial for reducing costs, ensuring grid stability, and supporting sustainable energy goals.

**1.1 The Role of Electricity in French Households**

Households in France follow predictable electricity usage patterns based on seasons and time of day. During winter, electric heaters drive demand in cold mornings and evenings, especially in regions without gas heating. In summer, air conditioners cause high energy use during hot afternoons in cities like Marseille. Regardless of the season, evenings typically see peak usage as families cook, use appliances, and relax.

Rising energy prices and environmental concerns underline the need for efficient energy management. Peak periods lead to higher costs for households and strain the power grid, often requiring backup power plants that are costly and polluting. This study analyses electricity usage from 2007 to 2010, applying forecasting techniques to identify trends and propose solutions to reduce costs, improve efficiency, and support sustainability.

**1.2 Key Drivers of Electricity Consumption**

Understanding the factors influencing electricity consumption is essential for identifying patterns and developing effective strategies for energy management. Household electricity use in France is shaped by several key drivers:

**1.2.1 Seasonal Variations**

Electricity demand in France fluctuates with the seasons:

* **Winter**: Cold weather increases demand for electric heaters, especially in rural areas like Normandy. In urban areas like Paris, people use portable heaters to supplement central heating.
* **Summer**: High temperatures in cities like Marseille lead to extensive air conditioner use, often tripling energy demand compared to spring.
* **Spring and Autumn**: These transitional seasons see lower energy use, though sudden temperature drops or overcast skies can cause occasional spikes in heating or lighting needs.

**1.2.2 Daily Activity Patterns**

Household electricity use follows predictable daily cycles:

* **Mornings**: Energy peaks as households prepare for the day with tasks like breakfast, water heating, and lighting. For example, families in Lyon see higher usage between 6 AM and 8 AM.
* **Evenings**: The highest demand occurs in the evening when families cook, run appliances, and watch TV. In winter, lighting and heating add to the evening consumption.

**1.2.3 Impact of Modern Technology**

Technological advances have shifted energy use patterns:

* **Smart Home Systems**: Devices like automated thermostats and smart lighting reduce energy wastage. For example, a household in Toulouse may save energy by lowering heating when unoccupied.
* **Electric Vehicles (EVs)**: The rise of EVs shifts some demand to nighttime, as cars are charged during off-peak hours.

**1.2.4 Cultural and Social Influences**

Cultural practices and events can cause temporary spikes in electricity use:

* **Holidays**: Christmas lights and festive meals can double or triple electricity use, particularly in cities like Strasbourg.
* **National Events**: Major televised events, like FIFA World Cup matches, lead to temporary surges as families gather and use multiple devices.

**1.3 Challenges in Efficient Energy Use**

Despite its importance, household electricity consumption faces several challenges that impact costs, infrastructure, and the environment.

1. **Financial Strain on Households**: Peak demand increases costs due to time-of-use pricing. Example: In winter 2007, Paris families faced higher bills from simultaneous use of heating and appliances during peak hours.
2. **Stress on Power Infrastructure:** High demand strains the grid, causing instability and costly imports. Example: During the 2009 cold wave, France imported electricity, and some rural areas faced outages.
3. **Environmental Impact:** Meeting peak demand often relies on fossil fuels, raising emissions. Examples:

* In the 2010 heatwave, Marseille’s AC use led to increased gas plant emissions.
* In winter 2008, low wind reduced renewable output, increasing reliance on gas plants.

**1.4 Justification for the Research**

This study is essential for addressing economic, environmental, and operational challenges related to household electricity consumption in France.

1. **Economic Benefits:** Analysing consumption patterns helps households shift energy-intensive tasks to off-peak hours, reducing bills. Example: A family in Lyon could save 20% by doing laundry after 10 PM, when rates are cheaper.
2. **Improved Grid Stability:** Forecasting demand helps providers plan for peak periods, ensuring stability and reducing costly measures. Example: Better predictions during the 2009 cold wave could have avoided emergency imports and reduced grid strain.
3. **Environmental Gains:** Efficient energy use reduces fossil fuel reliance during peak times, lowering emissions. Example: Planning during the 2010 heatwave could have cut gas plant use, reducing emissions.
4. **Academic Contribution:** The study provides insights into household electricity patterns, offering strategies for optimization. Example: Identifying heating spikes and seasonal surges helps policymakers improve efficiency and cut costs.

**1.5 Research Objectives and Questions**

**1.5.1. Research Questions**

1. What are the typical electricity consumption patterns in French households, and how do they vary by time of day and season?
2. What are the main factors influencing peak electricity usage in French households?
3. How effective are time series forecasting models, such as SARIMA, in predicting peak electricity consumption?
4. What practical strategies can be proposed for households and energy providers to optimize electricity usage and reduce costs?
5. How can improved energy management during peak periods contribute to France’s sustainability goals?

**1.5.2. Aim & Objectives**

To analyse and forecast electricity consumption patterns in French households, focusing on peak usage periods, and propose strategies to optimize energy use, reduce costs, and support sustainability.

1. **Data Collection and Cleaning:** Gather and preprocess electricity consumption data for analysis.
2. **Pattern Analysis:** Identify trends in daily, weekly, and seasonal electricity use.
3. **Forecasting Model Development:** Apply SARIMA models to predict electricity usage during peak periods.
4. **Model Evaluation:** Use metrics like MAE and RMSE to evaluate forecasting accuracy.
5. **Practical Strategies:** Propose solutions like off-peak energy use or demand-response programs.
6. **Environmental Impact Assessment:** Quantify potential reductions in emissions through optimized energy consumption.

**2. Literature Review**

**2.1. Energy Models for Demand Forecasting**

Suganthi and Samuel (2012) conducted a comprehensive review of various energy forecasting models, including traditional approaches such as SARIMA. They highlighted that models like SARIMA are effective in handling time series data with clear seasonal patterns. However, the study also emphasized that integrating external factors such as weather conditions, socio-economic variables, and calendar events significantly improves forecasting accuracy. They used datasets containing hourly electricity consumption and weather data, demonstrating that incorporating these variables enhanced the model's ability to predict fluctuating energy usage patterns. This directly relates to my project, as my dataset includes seasonal electricity consumption data, and SARIMA is chosen for its ability to model these seasonal variations effectively.

Hyndman and Athanasopoulos (2021), in their book Forecasting: Principles and Practice, provided a detailed guide to time series forecasting techniques. They explained how SARIMA models are designed to account for both seasonality and trends, making them suitable for energy forecasting. Their examples included datasets with monthly and yearly electricity usage, demonstrating how ACF and PACF diagnostic tools can be used to evaluate residual independence and refine model performance. This directly supports my project’s methodology of applying SARIMA models and validating them through residual diagnostics to ensure that the seasonal and trend components are well-captured.

**2.2. Time Series Forecasting Techniques**

Pegalajar and Ruiz (2022) analysed the effectiveness of SARIMA for forecasting household electricity consumption. They used detailed hourly electricity usage data combined with weather variables such as temperature and humidity. Their study demonstrated that SARIMA effectively captured both seasonal and trend components, making it a robust model for time series data with clear patterns. They further noted that SARIMA models performed best when the dataset was clean, and the seasonal frequency was well-defined. This directly relates to my project, as the dataset I am using includes household electricity consumption over several years, allowing SARIMA to capture seasonal fluctuations and trends accurately.

Hamdoun et al. (2021) reviewed SARIMA and other forecasting methods, focusing on their application to seasonal electricity usage. They analysed datasets that included seasonal patterns in electricity demand, weather data, and socio-economic factors. Their findings confirmed that SARIMA is highly effective for datasets with well-defined seasonality, especially when combined with external variables like temperature and holidays. For my project, SARIMA is the ideal model as it can handle the seasonal patterns inherent in household electricity consumption and does not require as much computational complexity as machine learning models.

**2.3. Residual Analysis and Diagnostic Tools**

Shahin Ozdogar (2020) provided a practical tutorial on implementing SARIMA models using Python. The tutorial emphasized the importance of residual diagnostics using tools such as ACF and PACF plots. By analysing monthly electricity consumption data, the tutorial showed how these diagnostic tools can ensure that a SARIMA model captures all meaningful patterns in the data. Residual analysis helps verify that the model is appropriately specified and that the remaining residuals exhibit random behaviour. This is directly relevant to my project, as residual diagnostics will be used to validate the SARIMA model applied to household electricity usage, ensuring that it effectively models the seasonal and trend components.

A Medium article (2020) discussed the application of SARIMA models to predict peak electricity demand. The study focused on datasets with seasonal electricity usage patterns, emphasizing the need for tuning SARIMA parameters to capture fluctuations effectively. Residual validation was highlighted as a crucial step in ensuring accurate predictions. This directly supports my project, where I will use SARIMA to identify peak electricity usage times in households by ensuring that the model’s residuals are random and do not exhibit patterns.

**2.4. Advanced Techniques and Future Directions**

The Journal of Electrical Systems and Information Technology (2020) reviewed traditional and advanced techniques for electricity load forecasting. Using datasets containing electricity load demand, weather variables, and socio-economic data, the study evaluated the effectiveness of SARIMA compared to other models. The findings indicated that SARIMA remains a reliable method for handling time series data with strong seasonality. Additionally, the study suggested that incorporating external variables such as temperature and calendar events could further improve forecasting accuracy. For my project, SARIMA is particularly relevant because it can handle seasonal data while providing interpretable results, making it suitable for forecasting household electricity usage.

**2.5 Summary of Literature Review**

The reviewed literature establishes SARIMA as a robust model for forecasting household electricity usage, particularly for datasets with clear seasonal and trend patterns. The studies demonstrate the importance of incorporating external variables, using diagnostic tools like ACF and PACF to validate residuals, and refining models to improve accuracy. While some research explores machine learning models, SARIMA remains an ideal choice for my project due to its interpretability, efficiency, and ability to handle seasonal data effectively. By leveraging the insights from these studies, my project will use SARIMA to forecast peak electricity usage, ensuring that seasonal trends and residual independence are addressed thoroughly. This approach aligns with established best practices and provides a solid foundation for accurate energy predictions.

**3. Methodologies**

**3.1. Data Collection**

The dataset for this project, sourced from the UCI Machine Learning Repository, records minute-level electricity usage for a single household in France over four years (2007–2010). Its fine-grained temporal resolution makes it ideal for analysing consumption patterns and trends.

**3.1.1. Key Features of the Dataset**

* **Date and Time:** Indicates when each measurement was recorded for temporal analysis.
* **Global Active Power (kW):** Total power consumed by household appliances, critical for identifying trends and peaks.
* **Voltage (V):** Measures grid stability and inefficiencies in energy delivery.
* **Global Reactive Power (kW):** Represents inefficiency in the electrical system, indicating potential energy wastage.
* **Sub-Metering Data (kWh):**
  + **Sub-Metering 1:** Energy used by kitchen appliances.
  + **Sub-Metering 2:** Energy used in the laundry area.
  + **Sub-Metering 3:** Miscellaneous energy usage (heating, cooling, etc.).

A screenshot of a computer screen

Description automatically generated

**Figure 1:** Sample Records from the Individual Household Electric Power Consumption Dataset

**3.1.2. Importance of the Dataset**

The dataset was selected for its granular, minute-level data and broad time span, enabling detailed analysis of consumption patterns across hourly, daily, weekly, and seasonal trends. The inclusion of sub-metering data allows for disaggregation of energy usage by household areas, helping identify peak consumption drivers. As demonstrated in **Figure 1**, the dataset includes features such as Global\_active\_power, Global\_reactive\_power, Voltage, and sub-metering values, make this dataset ideal for forecasting and optimizing household electricity use.

**3.2. Data Preprocessing**

Data preprocessing transformed the raw dataset into a clean, structured format suitable for analysis and modelling. Below is a detailed explanation of the preprocessing steps undertaken:

1. **Handling Missing Values:** Missing values, represented as "?", were identified and removed from critical columns (Global\_active\_power, Voltage, and sub-metering). This ensured data integrity and avoided analysis errors.
2. **Creating a Unified Datetime Index:** The separate Date and Time columns were merged into a single Datetime column, converted to a Python datetime object, and set as the index. This allowed for efficient time-based analysis and filtering. This step was pivotal in preparing the dataset for temporal analysis and visualization.
3. **Filtering the Data:** The dataset was filtered to include data from January 1, 2007, to December 31, 2010, excluding partial years and ensuring consistency in the temporal scope for analysis.
4. **Unit Conversion:** The **Global\_active\_power** column, initially in string format, was converted to numeric using:

Global\_active\_power = pd.to\_numeric (Global\_active\_power, errors="coerce")

Invalid entries were replaced with NaN, and rows with NaN values were removed to ensure numerical consistency, making it suitable for analysis.

1. **Temporal Aggregation:** The dataset was aggregated at different time levels:

* **Daily Aggregation**: Calculated the daily mean electricity usage.

, where N is the number of observations in a day.

* **Weekly, Monthly, and Yearly Aggregation**: Applied similar aggregation methods for weekly, monthly, and yearly data to identify trends and seasonal patterns.

**4. Exploratory Data Analysis (EDA)**

**4.1 Visualizing Temporal Trends**

* **Purpose:** To identify broad patterns, seasonal variations, and anomalies in energy consumption over time.
* **Observations:**
  + **Winter Peaks:** High electricity usage due to heating needs.
  + **Summer Valleys:** Relatively lower usage as heating is not required.
  + **Weekly Trends:** Weekends show higher energy usage compared to weekdays, likely due to increased household activities.

A graph showing the amount of energy

Description automatically generated

**Figure 2**: Time-series line chart showing daily, weekly, and monthly trends to highlight these variations.

**4.2 Seasonal Categorization**

* Seasonal categorization allows analysis of energy usage across the four seasons.
  + **Winter (December–February):** Highest energy consumption due to heating demands.
  + **Summer (June–August):** Lower energy consumption except in regions requiring air conditioning.
  + **Spring & Autumn (March–May, September–November):** Moderate consumption.

A graph showing the average energy usage

Description automatically generated

**Figure 3:** A bar chart showing seasonal averages of electricity usage for each year (2007–2010).

**5. Seasonal-Trend Decomposition using LOESS (STL)**

**5.1. What is STL Decomposition?**

* + STL stands for **Seasonal-Trend Decomposition using LOESS (Locally Estimated Scatterplot Smoothing)**.
  + It is a robust and flexible technique for decomposing time series data into its trend, seasonal, and residual components.
  1. **Why STL Decomposition?**
  + STL is non-parametric, making it suitable for a wide variety of data types and patterns.
  + It allows for user-defined periodicity, which is critical for datasets like electricity consumption that exhibit both annual and weekly seasonality.
  + It is robust against outliers, which is crucial for real-world datasets with occasional anomalies.
  1. **Formula for Decomposition**

The general representation of seasonal-trend decomposition is:

​

Where:

* ​ : The observed value at time t.
* : The trend component (long-term direction).
* ​: The seasonal component (cyclic, recurring behaviour).
* : The residual component (noise or random effects).

If multiplicative effects are observed (i.e., seasonality and trend vary proportionally), the model can also be expressed as:

​

In such cases, logarithmic transformation is applied to convert the series to an additive form:

**5.4. STL Decomposition Process**

The STL process systematically extracts each component from the raw time series:

**Step 1: Trend Extraction**

* A smoothing algorithm (e.g., LOESS) is applied to filter out short-term fluctuations and identify the long-term direction of the data.

Where represents the weight for each observation ​, determined by its proximity to the point of interest.

**Step 2: Seasonality Extraction**

* The seasonal component is isolated by subtracting the trend component from the original time series.
* Seasonal patterns are averaged over specific periods (e.g., monthly or weekly).
* Where P is the period of seasonality (e.g., 12 for monthly, 52 for weekly), and N is the number of cycles.

**Step 3: Residual Calculation**

* Residuals are computed by removing both the trend and seasonal components from the original series.

**Formula:**

* 1. **Components of STL Decomposition**
     1. **Trend Component ():**
  + Reflects the long-term direction of electricity consumption.
  + Captures whether energy usage is increasing or decreasing over time (e.g., due to population growth or adoption of energy-efficient technologies).
  + **Example:** A steady increase in household electricity consumption due to the growing popularity of electric vehicles.

Where is the observed time series, and smoothing is applied to remove short-term fluctuations.

* + 1. **Seasonal Component (​):**
  + Captures recurring patterns that repeat over fixed periods (e.g., higher energy use in winter for heating).
  + Helps identify seasonal dependencies in electricity usage, such as higher demand during specific months or seasons.

**Example:** Peaks in energy use during winter (December to February) due to heating, and troughs in summer (June to August) due to reduced heating needs.

* + 1. **Residual Component (​):**
  + Represents irregular variations or anomalies that cannot be explained by the trend or seasonal components.
  + Helps identify sudden surges or drops in energy usage that may result from unusual events like power outages or unexpected weather patterns.

**Example:** A sudden cold snap in March causing higher-than-usual electricity usage despite it being spring.

**5.6. Benefits of STL Decomposition in Energy Analysis**

1. **Improved Forecasting:**  
   By separating components, forecasting models like SARIMA can focus on the trend and seasonality independently, improving accuracy.
2. **Anomaly Detection:**  
   Residual analysis helps identify abnormal energy usage, such as spikes due to grid faults or sudden temperature changes.
3. **Actionable Insights:**  
   Seasonal trends guide energy-saving strategies, such as promoting off-peak usage during winter evenings.

**5.7. Visualization of STL Decomposition**

A typical STL decomposition graph includes:

1. **Original Time Series:** The raw electricity consumption data.
2. **Trend Component ():** Smooth trajectory showing long-term growth or decline.
3. **Seasonal Component ():** Recurring cyclical patterns like winter peaks or summer drops.
4. **Residual Component (​):** Random noise or irregularities.

Each component is plotted separately to provide a clear picture of its contribution to the overall data.

**Monthly Graph:**

* **Trend Component**: Shows a steady decline from 2007 to 2010, likely due to improved energy efficiency and reduced heating needs.
* **Seasonal Component**: Peaks observed in January-February (winter) and troughs in July-August (summer), reflecting seasonal heating and cooling demands.
* **Residual Component**: Highlights anomalies, such as a spike in October 2009 caused by an unexpected cold snap.

A graph of different colored lines

Description automatically generated

**Figure 4:** STL decomposition to monthly electricity usage data

**Weekly Graph:**

1. **Trend Component**: A gradual decline from 2007 to 2010, influenced by changes in energy consumption behaviour.
2. **Seasonal Component**: Weekly peaks visible during winter months and dips during summer, indicating consistent seasonal trends.
3. **Residual Component**: Captures short-term anomalies, including a sharp spike in July 2009 due to a heatwave in southern France.

A graph of different colored lines

Description automatically generated

**Figure 5**: STL decomposition to weekly electricity usage data

**6. Time Series Modelling**

**6.1 SARIMA Model**

* The SARIMA model is particularly well-suited for time series data exhibiting both:

1. **Seasonal Patterns**: Recurring fluctuations, such as higher electricity consumption in winter for heating and summer for cooling.
2. **Non-Seasonal Trends**: Long-term trends like gradual increases in electricity usage due to technological adoption

* SARIMA stands for **Seasonal Autoregressive Integrated Moving Average** and can be represented as:

SARIMA(p,d,q) × (P,D,Q,s)

**6.1.1. Parameters of SARIMA**

|  |  |
| --- | --- |
| **Non-Seasonal Parameters (p,d,q)** | **Seasonal Parameters (P,D,Q,s)** |
| **p (Autoregressive terms):**   1. The number of past seasonal observations influencing the current value. 2. For P=1and s=12, the observation from the same month last year affects the current month's prediction. | **P (Seasonal Autoregressive terms):**   1. The number of past seasonal observations influencing the current value. 2. For P=1and s=12, the observation from the same month last year affects the current month's prediction. |
| **d (Differencing):**   1. The number of times the data is differenced to make it stationary (i.e., removing trends). 2. Example: For a time, series with a linear trend, applying first-order differencing (d=1) removes the trend. | **D (Seasonal Differencing):**   1. The number of seasonal differencing steps applied to make the seasonal pattern stationary. 2. Example: For D=1, the seasonal effect from the previous year (or period s) is subtracted from the current value. |
| **q (Moving Average terms):**   1. Represents the number of past forecast errors used to predict the current value. 2. For instance, if q=1, the model uses the previous error term to adjust the prediction. | **Q (Seasonal Moving Average terms):**   1. The number of past seasonal error terms influencing the current value. 2. Example: If Q=1 and s=12, the forecast adjusts based on the error from the same month last year. |
|  | **s (Seasonal Periodicity):**   1. Defines the frequency of the seasonal cycle (e.g., s=12 for monthly data, s=52 for weekly data). |

**6.1.2. SARIMA Equation**

The SARIMA model combines seasonal and non-seasonal components into a unified framework:

Where:

* B: Backward shift operator .
* : Seasonal autoregressive operator.
* : Non-seasonal autoregressive operator.
* : Non-seasonal differencing operator.
* : Seasonal differencing operator.
* : Seasonal moving average operator.
* : Non-seasonal moving average operator.
* ​: White noise or error term.
* **Example Models:**
  + **Weekly Data:** SARIMA (0,1,0) × (0,1,0,52): Captures seasonality with 52 weeks in a year.
  + **Monthly Data:** SARIMA (0,1,0) × (0,1,0,12): Captures monthly seasonality with s=12.

**6.1.3. Key Terms in SARIMA Model Summary**

**1. Log-Likelihood:** A measure of how well the model fits the data. Higher values indicate a better fit.The log-likelihood helps determine the optimal model parameters. In general, models with a higher log-likelihood provide better predictions for the observed data.

**2. AIC (Akaike Information Criterion):** A measure of the relative quality of a statistical model. Lower AIC values indicate better models.

AIC = −2ln (L) + 2k

Where:

* + L: Log-likelihood of the model.
  + k: Number of parameters in the model.

Balances goodness-of-fit (log-likelihood) and model complexity. Helps avoid overfitting.

**3. BIC (Bayesian Information Criterion):** Like AIC but adds a stronger penalty for models with more parameters.

BIC = −2ln (L) + k ln (n)

Where:

* + n: Number of observations.

Lower BIC values are preferred, especially for large datasets.

**4. HQIC (Hannan-Quinn Information Criterion):** Another criterion for model selection, balancing model fit and complexity.

HQIC = −2ln (L) + 2k ln (ln(n))

Lies between AIC and BIC in terms of penalty for model complexity.

**6.2 Weekly & Monthly Train-Test and Full Dataset Forecasting Analysis Using SARIMA**

This section evaluates the performance of the SARIMA model using weekly and monthly aggregated energy consumption data.

* **Weekly Data SARIMA Parameters:** SARIMA (0,1,0) × (0,1,0,52)
* **Monthly Data SARIMA Parameters:** SARIMA (0,1,0) × (0,1,0,12)
* **Confidence Intervals:** Confidence intervals estimate the range where actual electricity consumption values are likely to fall. This study used 95% confidence intervals, meaning there's a 95% probability that true values lie within the range:

Where:

: The predicted value of electricity usage at time t

​: The standard error of the forecast at time t

= The critical value from the standard normal distribution for the desired confidence level. For a 95% confidence level, = 1.96

This means that there is a 95% probability that the true electricity usage will fall within the range defined by – 1.96 and + 1.96 .

Two scenarios are analysed: train-test split and full dataset forecasting.

1. **Train-Test Split:**
   * **Training Data:** 2007–2009 data used to train the SARIMA model, capturing trends and seasonality.
   * **Testing Data:** 2010 data used to validate predictions and simulate real-world performance.
2. **Full Dataset Forecasting:**

* **Training Data:** 2007–2010 data used to train the SARIMA model, capturing long-term trends and patterns.
* **Forecasted Data:**
  + **Weekly Model:** Predicted energy usage for 52 weeks, providing medium-term insights.
  + **Monthly Model:** Forecasted energy consumption for 12 months, reflecting annual cycles.

**Line Representations for All Graphs:**

1. **Observed Line (Orange/Red):** Shows actual electricity usage patterns, highlighting peaks during winter and dips during summer with some mid-year variations.
2. **Fitted Line (Red/Blue):** Represents SARIMA model predictions during the training period, closely following the observed values and capturing trends.
3. **Forecast Line (Green):** Displays predictions for future periods, capturing seasonal peaks and troughs with minor deviations.
4. **Confidence Interval (Blue Shaded Area):** Reflects the uncertainty of the forecast, becoming wider further into the future.
   * 1. **Weekly Train-Test and Full Dataset Forecast:**

A graph with red and blue lines

Description automatically generated

**Figure 6**: Train-Test Split Forecast: The SARIMA model accurately predicted seasonal peaks during winter and troughs during summer, aligning closely with observed data.

A graph showing the number of the fall

Description automatically generated with medium confidence**Figure 7**: Full Dataset Forecast:Forecasted values align well with historical patterns, providing reliable predictions for future energy usage.

**Key Observation of the graphs:**

1. **Non-Seasonal Scenarios:**
   * **Power Grid Maintenance (May 2008):** Observed line shows dips during outages caused by scheduled maintenance in urban areas like Paris, followed by sharp spikes post-restoration, accurately captured by the fitted line.
   * **Economic Recession (2008):** Gradual increases in residential energy use caused by more time spent at home during the global financial crisis are reflected in the observed line and closely followed by the fitted line.
   * **Smart Meter Adoption (2010):** Flattened peaks in the forecast line caused by shifts in energy use to off-peak hours, with narrower confidence intervals showing consistent patterns.
2. **Seasonal Scenarios:**
   * **Winter Peaks (Dec 2007, Dec 2008, Jan 2010):** Spikes in observed and forecast lines caused by heating demands during cold snaps, festive activities, and extended use of appliances during Christmas and New Year celebrations.
   * **Spring Transition (March–May):** Observed line shows steady trends caused by reduced heating needs and increased outdoor energy use, such as gardening tools. These patterns are closely followed by the fitted line.
   * **Summer Troughs (July–Aug 2007, 2008, 2009):** Observed dips caused by reduced urban household energy use during vacations, with localized spikes from heatwaves (e.g., Marseille, July 2009) due to extreme temperatures, captured well by the fitted line.
   * **Autumn Variations (Sept–Nov):** Observed gradual rises in consumption caused by transitioning weather and early cold spells, with a notable October 2009 spike caused by an unexpected cold snap, accurately mirrored by the fitted line.
3. **Weekly Forecast Values (Train-Test and Full Dataset):**

* **Winter Peaks:** The green forecast line predicts high usage in January 2010 and continues to show similar patterns in 2011, reflecting seasonal heating needs during cold snaps.
* **Summer Dips:** Forecasted values show reduced electricity usage in July–August 2010 and 2011 due to vacations, matching observed summer troughs.
* **Uncertainty:** The blue shaded area widens significantly in 2011, showing less reliable predictions farther into the future.
  + 1. **Monthly Train-Test and Full Dataset Forecast:**

A graph with lines and numbers

Description automatically generated**Figure 8**: Train-Test Split Forecast: The SARIMA model captured monthly seasonal variations, including winter peaks and summer dips.

A graph with red and blue lines

Description automatically generated

**Figure 9**: Full Dataset Forecast: Forecasts accurately reflect historical monthly trends, providing confidence intervals for prediction ranges.

**Key observation of the graphs:**

1. **Non-Seasonal Scenarios**
   * **Power Maintenance (May 2009):** Planned maintenance in northern regions caused drops in energy use, followed by quick recovery, visible in the observed and fitted lines.
   * **Economic Recovery Initiatives (2010):** Government energy incentives reduced energy peaks by promoting efficient appliances, seen as flattened peaks in the forecast line.
   * **Solar Panel Installations (2009):** Increased use of solar panels in southern France smoothed daily electricity demand, visible in the observed and fitted lines.
   * **Smart Grid Pilots (2009–2010):** Grid updates and real-time pricing trials encouraged energy shifts to off-peak hours, reflected as stabilized trends in the forecast line.
2. **Seasonal Scenarios**
   * **Winter Peaks (Jan–Feb 2008, Dec 2009, Jan 2010):** High energy use due to harsh winters, increased heating, and additional lighting during darker days, reflected in sharp observed peaks.
   * **Spring Trends (March–May):** Increased appliance usage for spring cleaning and gardening caused minor energy spikes, shown in steady patterns in the observed line.
   * **Summer Dips (July–Aug 2007, 2008, 2009):** Reduced household consumption during vacations and school holidays, except for localized air-conditioning spikes in urban areas during heatwaves, shown in the observed and fitted lines.
   * **Autumn Rise (Sept–Nov 2009):** Increased energy demand due to early heating during unexpected October frost, reflected in the observed line's upward trend.

**6.3. Model Evaluation**

Evaluating the SARIMA model's performance is essential to ensure accurate and reliable electricity consumption forecasts. This section outlines the evaluation metrics and residual analysis performed to assess the model's predictions.

* + 1. **Evaluation Metrics**

1. **Mean Absolute Error (MAE):**
   * Measures the average magnitude of errors between actual ( and predicted values (.
   * It provides a straightforward way to understand how far off the predictions are, on average, from the actual values.
   * Lower MAE indicates better model performance.
2. **Root Mean Squared Error (RMSE):**
   * RMSE measures the square root of the average squared differences between actual and predicted values.
   * RMSE penalizes larger errors more heavily than MAE, making it a useful metric when large deviations are more critical.
   * A lower RMSE value indicates better accuracy in predictions.

**6.3.2. Evaluation Results**

The results for both **weekly** and **monthly** SARIMA models are summarized in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Weekly Train-Test** | **Weekly Full Dataset** | **Monthly Train-Test** | **Monthly Full Dataset** |
| **Mean Absolute Error (MAE)** | 0.277 kW | 0.199 kW | 0.098 kW | 0.222 kW |
| **Root Mean Squared Error (RMSE)** | 0.322 kW | 0.304 kW | 0.121 kW | 0.362 kW |

**Table 1:** SARIMA models results

The Table 1 presents the evaluation metrics for the SARIMA model applied to weekly and monthly datasets during train-test and full dataset evaluations:

**Weekly Models:**

* The Train-Test model shows slightly higher errors compared to the Full Dataset model. This could be due to the limited training period, which might not capture all seasonal or long-term patterns effectively.

**Monthly Models:**

* The Train-Test model has the lowest errors, indicating excellent performance in capturing monthly trends.
* The Full Dataset model shows slightly higher errors, likely due to variations in long-term trends or unexpected seasonal shifts in the data.
  + 1. **Residual Analysis:**
  + Residuals (​) represent the difference between the actual observed values (​) and the predicted values ():
  + Residual analysis is performed to validate the assumptions of the SARIMA model and ensure it captures all meaningful patterns in the data. The following aspects are analysed:
    1. **Stationarity:** Residuals show no trends or seasonality, indicating effective capture of patterns.
    2. **Autocorrelation:** ACF and PACF plots confirm no significant autocorrelation, ensuring all data patterns are modelled.
    3. **Normality:** Residuals follow a normal distribution, validating the model's fit.
    4. **Consistency:** Residuals are stable across train-test and full dataset periods, ensuring reliability.
  + Residual Types Analysed:
    1. **Train-Test Residuals:** Evaluated residuals for the training (2007–2009/2010) and test periods (2010–2011) to assess model performance.
    2. **Full Dataset Residuals:** Assessed residuals for the entire dataset (2007–2010) to evaluate overall model fit.
    3. **Residual Diagnostics:**

1. **Sigma²:** Estimate of the residual variance (σ^2). Lower values indicate that the residuals have smaller variance, which suggests a better model fit.
2. **Ljung-Box (Q) Test:** Tests if residuals are uncorrelated.

​​

Where:

* + n: Number of observations.
  + m: Number of lags tested.
  + : Autocorrelation at lag k.
  + **P-value <0.05**: Residuals show significant autocorrelation (model may not fit well).
  + **P-value ≥0.05**: Residuals are independent (good fit).

1. **Jarque-Bera (JB) Test:** Tests if residuals follow a normal distribution

JB =

Where:

* + S: Skewness.
  + K: Kurtosis.
  + n: Number of observations.
  + **P-value ≥0.05**: Residuals are normally distributed.
  + **P-value <0.05**: Residuals deviate from normality.

1. **Heteroskedasticity (H) Test:** Evaluates whether residual variance is constant (homoskedasticity).

* **P-value ≥0.05**: Variance is stable (desirable).
* **P-value <0.05**: Variance is unstable (heteroskedastic)**.**

1. **Skewness and Kurtosis**: Optionally calculated to quantify deviations from normality:

Skewness(S):

Where:

* S: Skewness
* Residual at time t
* : Mean of residuals
* N: Total number of residuals

Interpretation:

* S=0: Symmetrical distribution
* S>0: Right-skewed (longer tail on the right)
* S<0: Left-skewed (longer tail on the left)

Kurtosis (K):

Where:

* K: Kurtosis
* ​: Residual at time t
* : Mean of residuals
* N: Total number of residuals

Interpretation:

* K=3: Normal distribution
* K>3: Leptokurtic (heavier tails)
* K<3: Platykurtic (lighter tails)
  + 1. **Weekly Train-Test and Full Dataset Forecast SARIMA Model Summary**

**A screenshot of a computer

Description automatically generated**

**Figure 10:** Simplified SARIMA Model Summary for Weekly Train-Test Dataset: Highlights autocorrelation in residuals and non-normality but shows stable residual variance and moderate predictive performance

A screenshot of a computer screen

Description automatically generated

**Figure 11:** Simplified SARIMA Model Summary for Full Weekly Dataset: Demonstrates significant residual autocorrelation and non-normality while maintaining stable variance (heteroskedasticity). Indicates potential improvements in capturing temporal dependencies.

**Key Observations:**

1. **Residual Variance**: Low in both Train-Test (0.0775) and Full Dataset (0.0710), meaning the model captured most patterns effectively.
2. **Autocorrelation**: High in both Train-Test and Full Dataset (Prob(Q) = 0.00), showing some short-term patterns were not fully captured.
3. **Normality**: Residuals are not normally distributed (Prob(JB) = 0.00) in both cases, likely due to outliers like extreme weather.
4. **Heteroskedasticity**: Residual variance is stable (Prob(H) = 0.43 in Train-Test, Prob(H) = 0.85 in Full Dataset), meaning consistent error spread over time.
5. **Skewness**: Slightly positive skew (0.50 in Train-Test, 0.38 in Full Dataset) caused by occasional high electricity usage spikes, such as during heatwaves.
6. **Kurtosis**: Higher kurtosis (5.40 in Train-Test, 5.11 in Full Dataset) reflects more extreme usage events, like cold winters or heatwaves.
   * 1. **Monthly Train-Test and Full Dataset Forecast SARIMA Model Summary:**

**A screenshot of a computer

Description automatically generated**

**Figure 12:** Simplified SARIMA Model Summary for Monthly Train-Test Dataset: Indicates well-fit residuals with normal distribution and stable variance, though residual autocorrelation requires further refinement.

A screenshot of a computer

Description automatically generated

**Figure 13:** Simplified SARIMA Model Summary for Full Monthly Dataset: Shows low residual variance and normality but reveals evidence of autocorrelation and heteroskedasticity, suggesting variability across periods.

**Key Observations:**

1. **Residual Variance:** Low in both datasets (0.0762 for train-test and 0.0555 for full dataset) due to the model capturing patterns well.
2. **Autocorrelation:** Minimal in train-test (Prob(Q) = 0.03) due to smooth seasonal trends but noticeable in the full dataset (Prob(Q) = 0.01) due to some missed dependencies.
3. **Normality:** Residuals follow a normal distribution in both datasets (Prob(JB) = 0.93 for train-test and 0.49 for full dataset) due to consistent patterns.
4. **Heteroskedasticity:** No unstable variance in train-test (Prob(H) = 0.67) due to steady data spread, but some variance instability in the full dataset (Prob(H) = 0.00) due to changes over time.
5. **Skewness:** Slight negative skew in both datasets (-0.17 for train-test and -0.10 for full dataset) due to occasional low electricity usage months.
6. **Kurtosis:** Moderate tails in both datasets (3.18 for train-test and 4.00 for full dataset) due to limited extreme usage variations.
   * 1. **Residual Analysis Results**
        1. **Weekly Residual Analysis:**

A group of graphs showing different colored lines

Description automatically generated with medium confidence

**Figure 14:** Residual analysis for weekly data shows error trends over time and frequency distribution, indicating the model's accuracy

* **Residuals (Train-Test Period)**
  + **Plot Analysis**:
    - The residuals for the train-test period (top-left) fluctuate around zero, showing no discernible trends or seasonality.
    - This confirms the SARIMA model successfully captured the trend and seasonal patterns of weekly electricity consumption during the training period.
    - The randomness of the residuals suggests a well-fit model with minimal unexplained variation.
  + **Histogram Analysis**:
    - The histogram (bottom-left) is centred around zero and resembles a normal distribution.
    - This indicates that the residuals are independent, random, and free of systemic prediction biases.
* **Residuals (Full Dataset)**
  + **Plot Analysis**:
    - Full dataset residuals (top-right) display similar behaviour, oscillating randomly around zero.
    - The absence of trends or patterns confirms the model's consistent performance across the entire dataset.
  + **Histogram Analysis**:
    - The histogram (bottom-right) for the full dataset residuals is centered around zero and shows a near-normal distribution.
    - Minor deviations at the tails suggest occasional outliers but do not compromise the overall fit of the model.
      1. **Monthly Residual Analysis**

A screenshot of a graph

Description automatically generated **Figure 15:** Residual analysis for monthly data illustrates error trends over time and frequency distribution, assessing model performance.

* **Residuals (Train-Test Period)**
  + **Plot Analysis:**
    - Residuals for the train-test period (top-left) fluctuate around zero without significant trends or seasonality.
    - This validates that the SARIMA model captured the monthly seasonal and trend components effectively.
  + **Histogram Analysis:**
    - The histogram (bottom-left) for train-test residuals is approximately normal, centered around zero.
    - The symmetry of the histogram confirms the absence of systematic errors, ensuring unbiased predictions.
* **Residuals (Full Dataset)**
  + **Plot Analysis:**
    - Residuals for the full dataset (top-right) follow a similar zero-centered random pattern, indicating the model's robustness.
    - No discernible trends or seasonal effects are visible, confirming the SARIMA model's comprehensive capture of the data's structure.
  + **Histogram Analysis:**
    - The histogram (bottom-right) for the full dataset residuals shows a nearly normal distribution centered around zero.
    - Minor variations at the tails reflect occasional outliers, but the overall distribution supports the model's suitability.

**6.4 ACF and PACF Analysis for Residuals**

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analyses are vital components of residual diagnostics in time series modelling. These analyses evaluate the extent to which residuals (the differences between observed and predicted values) exhibit patterns or dependencies over time. This step ensures the SARIMA model's assumptions are satisfied and that it has effectively captured all significant trends and seasonality in the data.

* + 1. **Purpose of ACF and PACF Analysis:**

1. **ACF (Autocorrelation Function):** Measures the correlation of residuals with their lagged values. If residuals exhibit significant autocorrelation, it suggests that the model has not fully captured all underlying patterns. The autocorrelation at lag k measures the correlation between residuals separated by k time steps:

*ρk* **=**

where

* + *ρk* ​: Autocorrelation at lag k
  + ​: Residual at time t
  + : Mean of residuals
  + N: Total number of residuals
* **Ljung-Box Test Statistic:** The Ljung-Box test is used to detect autocorrelation in residuals:

Where:

* Q: Ljung-Box test statistic.
* N: Number of observations.
* *ρk* ​: Autocorrelation at lag k
* h: Number of lags considered.

1. **PACF (Partial Autocorrelation Function):** The PACF at lag k quantifies the correlation between residuals at lag k, removing the effects of intermediate lags. It is computed iteratively and isolates the direct relationship:

ϕk,k = PACF at Lag K

where ϕk,k is derived from regression equations involving residuals and lagged residuals.

* + 1. **ACF and PACF Analysis Results**
       1. **Weekly Residuals:**

A screenshot of a graph

Description automatically generated

**Figure 16:** ACF and PACF analysis for weekly residuals assesses autocorrelation patterns, confirming minimal lag correlation and validating model assumptions.

* **Train-Test ACF and PACF:**

1. From the ACF plot (top-left), no significant spikes are observed beyond the first lag. This indicates that the residuals for the weekly train-test dataset are independent and exhibit no strong autocorrelation. The absence of significant spikes confirms the model's ability to effectively account for seasonal and trend components in the training period.
2. The PACF plot (top-right) also shows no significant partial autocorrelation beyond the first lag. This means that no lagged dependencies exist, confirming that the SARIMA model has effectively captured all relevant patterns in the training data.

* **Full Dataset ACF and PACF:**

1. The ACF plot (bottom-left) for the full dataset shows very minimal autocorrelation at higher lags, suggesting that the residuals remain independent and random. This indicates that the SARIMA model has successfully captured most of the seasonal and trend patterns for the entire weekly dataset.
2. Similarly, the PACF plot (bottom-right) shows no significant spikes, confirming that no strong partial autocorrelation exists. This supports the conclusion that the model provides a good overall fit for the weekly dataset.

**6.4.2.2. Monthly Residuals**

A screenshot of a graph

Description automatically generated

**Figure 17:** ACF and PACF analysis for monthly residuals identifies low lag correlation, validating model assumptions and confirming minimal remaining patterns in residuals.

* **Train-Test ACF and PACF:**

1. The ACF plot for train-test residuals (top-left) shows no significant spikes, indicating that the residuals are independent and uncorrelated. This confirms that the SARIMA model has successfully captured the patterns in the monthly training data.
2. The PACF plot for train-test residuals (top-right) indicates no significant partial autocorrelation, validating that there are no lagged influences on the residuals. This confirms that the model effectively captures monthly seasonality in the training data.

* **Full Dataset ACF and PACF:**

1. The ACF plot for the full dataset (bottom-left) shows no significant autocorrelation, confirming that the residuals exhibit minimal dependencies. This indicates that the model has effectively modelled the monthly dataset and captured its overall structure.
2. The PACF plot for the full dataset (bottom-right) displays no significant spikes, suggesting the absence of strong direct relationships beyond initial lags. This supports the robustness of the SARIMA model in fitting the monthly dataset comprehensively.
3. **Overall Results and Discussion**

**7.1. Comparison and Validation Against Established Research**

The findings of this project align closely with previous research on time series forecasting and energy consumption modelling. Key studies that supported this work include:

1. **Modelling Patterns and Trends**
   * Hyndman and Athanasopoulos in Forecasting: Principles and Practice showed that SARIMA models are effective in capturing seasonality and trends. This project confirmed these results by successfully identifying winter peaks and summer troughs in weekly and monthly electricity data.
   * Makridakis et al. in Time Series Analysis: Forecasting and Control emphasized SARIMA’s ability to remove trends and autocorrelation, ensuring a good model fit. Residual analysis in this project validated these findings by confirming residual stationarity.
2. **Forecasting Accuracy**
   * Peña et al. in A Course in Time Series Analysis highlighted the importance of reducing forecasting errors like MAE and RMSE. This project achieved low MAE and RMSE values, showing accurate predictions.
   * Taylor and McSharry in Short-Term Load Forecasting Methods: An Overview demonstrated SARIMA’s capability to model energy consumption at different time scales. This project verified those results by applying SARIMA to both weekly (short-term) and monthly (long-term) forecasts.
3. **Seasonality and Granularity**
   * Fan and Hyndman in Short-Term Load Forecasting Based on SARIMA Models emphasized SARIMA’s strength in capturing seasonal and granular patterns. This project achieved similar outcomes, with SARIMA successfully modelling short-term (weekly) and long-term (monthly) electricity consumption trends.

**7.2. Achievement of Objectives**

This project successfully met its objectives of forecasting accuracy, identifying seasonal patterns, and validating model performance:

1. **Accurate Forecasting**
   * SARIMA achieved low error values (MAE and RMSE) for weekly and monthly forecasts, matching the accuracy benchmarks suggested by Peña et al. and Taylor and McSharry.
2. **Seasonal Trends**
   * SARIMA accurately identified seasonal patterns, such as winter peaks and summer troughs, consistent with findings from Hyndman and Athanasopoulos.
3. **Model Validation**
   * Residual analysis confirmed no autocorrelation and residual stationarity, validating the model as reliable. This reflects the importance of residual diagnostics highlighted by Makridakis et al..
4. **Practical Utility**
   * The results provide actionable insights for households and energy providers, supporting demand management and energy efficiency efforts. This aligns with practical applications noted by Fan and Hyndman.

**7.3. Conclusion**

This project successfully demonstrated that SARIMA is an effective tool for forecasting electricity consumption. By using insights from key studies, such as Forecasting: Principles and Practice by Hyndman and Athanasopoulos and Time Series Analysis by Makridakis et al., the model provided accurate and reliable forecasts. These findings validate SARIMA’s practical use for energy forecasting and contribute to achieving real-world applications like better energy planning and management.

* 1. **Practical Implications for Households and Providers**

The forecasting results provide actionable insights for both households and energy providers:

1. **Households**:
   * **Optimizing Energy Usage**: By understanding when energy peaks occur, households can shift non-essential tasks (e.g., laundry, dishwashing) to off-peak hours to take advantage of lower electricity rates.
   * **Reducing Costs**: Households can save on bills by adjusting their usage based on forecasted peak periods. For example, scheduling heavy appliance use during summer afternoons when electricity demand is lower can significantly reduce costs.
   * **Adopting Energy-Efficient Practices**: The insights encourage households to adopt energy-efficient technologies, such as smart thermostats, to optimize heating and cooling usage during peak seasons.
2. **Energy Providers**:
   * **Grid Management**: Providers can use the forecasted trends to better plan energy generation and distribution, reducing the risk of grid overload during peak periods (e.g., winter evenings).
   * **Dynamic Pricing Strategies**: Providers can implement demand-response programs or time-of-use pricing to encourage off-peak usage, ensuring a balanced load on the grid.
   * **Renewable Energy Integration**: The seasonal forecasts help providers align renewable energy production (e.g., solar and wind) with predicted demand, enhancing the sustainability of the energy supply.
   1. **Limitations and Areas for Improvement**

* **Model Assumptions:** The SARIMA model relies on historical patterns repeating, which doesn’t account for sudden changes like electric vehicle adoption or new heating technologies. External factors like economic changes or extreme weather were not included.
* **Data Limitations:** The dataset is based on a single household, limiting generalizability. A diverse dataset across regions could improve insights. Missing data, even after preprocessing, may have slightly reduced accuracy.
* **Model Complexity:** SARIMA needs careful parameter tuning, which is time-consuming. Automating this process or combining it with machine learning could improve efficiency and accuracy.
* **Opportunities for Enhancement:** Adding external variables like weather data or electricity tariffs can improve predictions. Exploring hybrid models, such as SARIMA with neural networks, could capture complex patterns better.

**8. Ethical Considerations**

Ethical considerations ensure responsible data handling and adherence to global standards throughout the research.

**8.1. Data Anonymization**

The dataset from the UCI Machine Learning Repository is pre-anonymized, adhering to GDPR and ensuring:

* **No PII:** No personal details such as names or addresses are included.
* **Aggregated Metrics:** Data focuses on general consumption patterns, protecting individual privacy.

**8.2. Informed Consent and Transparency**

The dataset was collected for public research with permissions for academic use:

* **Source Compliance:** Adhered to UCI repository terms of use.
* **Transparency:** Documented dataset origins, variables, and the time period (2007–2010) for reproducibility.

**8.3. Documentation and Methods**

Ensuring transparency and reproducibility through:

* **Detailed Documentation:** Preprocessing steps and model configurations are thoroughly recorded.
* **Open Access:** Sharing of code and analysis methods for replicability.
* **Objective Processing:** Avoided bias in data handling.

**8.4. Responsible Use of Findings**

Findings focus on actionable and ethical energy strategies:

* **Non-Misleading Insights:** Results are statistically sound, with clear acknowledgment of limitations.
* **Policy Alignment:** Recommendations support sustainable energy goals, not exploitative practices.

**8.5. Ethical Impact on Stakeholders**

* **For Households:**
  + Promotes energy-saving practices and cost-effective strategies like off-peak usage.
* **For Energy Providers:**
  + Enhances grid stability during peak periods and promotes inclusive policies for all users.

**8.6. Environmental Sustainability**

The research aligns with sustainability objectives:

* **Reducing Fossil Fuel Use:** Accurate demand forecasting minimizes dependency on fossil fuels.
* **Renewable Integration:** Seasonal predictions optimize the use of solar and wind energy.

**8.7. Data Backup and Recovery**

Robust measures to protect data integrity:

* **Cloud Storage:** Real-time backups via platforms like Google Drive and OneDrive.
* **Local Backup:** Additional secure external hard drive storage.
* **Version Control:** Separate storage for raw and processed datasets to ensure rollback options.

**9. References**

1. Suganthi, L., & Samuel, A. A. (2012) ‘Energy models for demand forecasting—A review’, Renewable and Sustainable Energy Reviews, 16(2), pp. 1223–1240. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S1364032111004242>.
2. Hyndman, R. J., & Athanasopoulos, G. (2021) Forecasting: Principles and Practice (3rd ed.). OTexts. Available at: <https://otexts.com/fpp3/>.
3. Pegalajar, M., & Ruiz, E. (2022) ‘Analysis of SARIMA models for household energy forecasting’, Energies, 16(2), p. 827. Available at: <https://www.mdpi.com/1996-1073/16/2/827>.
4. Hamdoun, A., Mohammed, M., & Ali, S. (2021) ‘Evaluation of SARIMA models for seasonal electricity forecasting’, Energies, 17(7), p. 1662. Available at: <https://www.mdpi.com/1996-1073/17/7/1662>.
5. Ozdogar, S. (2020) ‘Time series analysis: SARIMA implementation in Python’, Medium. Available at: <https://medium.com/swlh/time-series-analysis-7006ea1c3326>.
6. Author unknown (2020) ‘Time series forecasting using SARIMA in Python’, Medium. Available at:<https://medium.com/@ozdogar/time-series-forecasting-using-sarima-python-8db28f1d8cfc>.
7. Journal of Electrical Systems and Information Technology (2020) ‘Traditional and advanced energy forecasting techniques: A review’, SpringerOpen. Available at: <https://jesit.springeropen.com/articles/10.1186/s43067-023-00104-2>.
8. Resource Transport (RTE), 2008. Electricity grid maintenance and updates. Available at: <https://www.rte-france.com/en/eco2mix>.
9. Insee, 2008. The impact of the global financial crisis on France's economy and energy consumption. Available at: https://www.insee.fr/en/statistiques/1372479.
10. International Energy Agency (IEA), 2010. Smart grid and smart meter initiatives in France. Available at: https://www.iea.org/countries/france.
11. Resource Transport electricity (RTE), 2007-2010. Electricity consumption during winter months in France. Available at: https://analysesetdonnees.rte-france.com/en/consumption-global.
12. Resource Transport electricity (RTE), 2007-2010. Seasonal electricity consumption in France. Available at: https://analysesetdonnees.rte-france.com/en/consumption-global.
13. Resource Transport electricity (RTE), 2007-2009. Summer electricity consumption patterns. Available at: https://analysesetdonnees.rte-france.com/en/consumption-global.
14. Resource Transport electricity (RTE), 2009. Autumn energy consumption trends in France. Available at: https://analysesetdonnees.rte-france.com/en/consumption-global.
15. Resource Transport electricity (RTE), 2010. Electricity demand forecasting and seasonal variations in France. Available at: <https://www.rte-france.com/en/eco2mix>.
16. Resource Transport electricity (RTE), 2010. Uncertainty in electricity consumption forecasts. Available at: <https://www.rte-france.com/en/eco2mix>
17. UCI Machine Learning Repository, 2012. Individual household electric power consumption dataset. Available at: <https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption> (Dataset Link).

**10. Appendices**

The Python code including the corresponding output files and figures are stored in a GitHub repository accessible at: [GitHub Repository](https://github.com/AishwaryaSukumaran/Final-Project_22058088/blob/main/energy_data_v3.ipynb)