TIME SERIES ANALYSIS STAT 9005

VICTORY K. ODUMEH / R00277181

Time Series Analysis of Monthly Global Active Power Consumption in a Single Household

10TH DECEMBER 2025

Declaration

I, **Victory K. Odumeh**, with Student Number **R00277181**, declare that the presented work is my own and all sourced used have been acknowledged and referenced. I comply with all MTU academic honesty and integrity policies.

1 Introduction

The project aims to perform a time series analysis on the monthly averaged global active power variable from the UCI individual household electric power consumption dataset. The analysis involves cleaning the dataset, decomposing the series, selecting a classical model that describes the series, finding an ARIMA model, comparing its results with the classical models, and generating a 12-month forecast for the dataset.

1.1 Background and Dataset Description

The dataset on individual household electric power consumption contains 2,075,259 measurements taken from a single household in Sceaux, France, which is 7 kilometers from Paris. The household measurements were gathered at a one-minute sampling rate over four years, from December 2006 to November 2010. The dataset comes from the UC Irvine Machine Learning repository. The dataset consists of the following variables and their data types:

- **date (Date)**: Date in format dd/mm/yyyy
- **time (categorical)**: Time in format hh:mm:ss
- **global_active_power (continuous)**: Household global minute-averaged active power (in kilowatt).
- **global_reactive_power (continuous)**: Household global minute-averaged reactive power (in kilowatt).
- **voltage (continuous)**: Minute-averaged voltage (in volt)
- **global_intensity (continuous)**: Household global minute-averaged current intensity (in ampere).
- **sub_metering_1 (continuous)**: Energy sub-metering number 1 (in watt-hour of active energy).
- **sub_metering_2 (continuous)**: Energy sub-metering number 2 (in watt-hour of active energy).
- **sub_metering_3 (continuous)**: Energy sub-metering number 3 (in watt-hour of active energy).

The sub-metering variables refer to specific parts of the house as well as particular electrical appliances: Sub_metering_1 refers to the kitchen, which mainly consists of a dishwasher, an oven, and a microwave (the hot plates are gas-powered rather than electric); sub_metering_2 refers to the laundry room, which includes a refrigerator, a washing machine, a tumble dryer, and a light; and sub_metering_3 refers to an electric water heater and an air conditioner. The dataset has about 1.25% missing values in the measurement columns (Hebrail & Berard, 2006). The global_active_power variable was chosen as the main focus of the analysis. It has a total of 25,979 missing values, which were fixed during data cleaning using linear interpolation.

1.2 Literature Review

The UCI individual household electric power consumption dataset has been used by a variety of studies to learn about the consumption of electricity in a single household. Several methods were used to analyse the dataset and make a few-month forecast for the data. One of the methods used in studies was the ARIMA model. In a study, it was used specifically to identify patterns and trends in the household electricity dataset over real-time periods (Parate & Bhoite, 2019, 371). Another study proposes a forecasting method using Deep Neural Networks (Marino et al., n.d.). It compares two Long Short-Term Memory (LSTM) architectures: regular LSTM and LSTM-based Sequence-to-Sequence. They were evaluated on the household electricity consumption datasets with hourly and minute frequencies (Marino et al., n.d.).

2 Methodology and Results

The analysis was conducted using RStudio, an Integrated Desktop Environment (IDE) primarily used for R programming and rich with a wide variety of packages developed specifically for time series analysis and forecasting. The packages used for this analysis include dplyr, forecast, astsa, tseries, zoo, lubridate, psych, and ggfortify.

2.1 Data Cleaning and Manipulation

The household power consumption dataset was read into R using the built-in read.table function. There were missing values found in the dataset; however, they were described as the question mark symbol (?), which is why na.string="?" was used to import the dataset into R. The global_active_power feature was used in this analysis. It consists of a total of 25,979 missing values. Linear interpolation was the method used to handle the missing values in the global active power feature. The na.approx function was used for the linear interpolation process.

The aggregation process used the dmy_hms function to concatenate both the date and time features of the dataset. The global active power feature was aggregated using the new months column created using the base R format function. The mean global active power for each month was calculated for each month between December 2006 and November 2010. The aggregated data was converted into a time series using the ts function, with a frequency or periodicity of 12.

2.2 Preliminary Statistics

The mean of the monthly global active power data is 1.1, which is less than the median, 1.12. This indicates that the distribution is slightly skewed to the left. The skew value is -0.11, which also shows that the data is negatively skewed. The maximum monthly global active power is 1.9, and the minimum is 0.28, with a range of 1.63. The standard deviation is 0.3, indicating that the spread of the dataset is not too far from the mean.

```
> describe(ts_data)
   vars  n mean  sd median trimmed  mad  min max range  skew kurtosis   se
X1    1 48  1.1 0.3   1.12     1.1 0.24 0.28 1.9  1.63 -0.11     0.39 0.04
> 
```

Fig. 2.1 Summary Statistics of the monthly global active power dataset

The dataset is visualised in Fig. 2.2. The boxplot also shows that there are outliers above the upper fence and below the lower fence. The body of the boxplot also indicates that the distribution may follow a normal distribution. The histogram appears symmetric, indicating that the dataset is normally distributed.

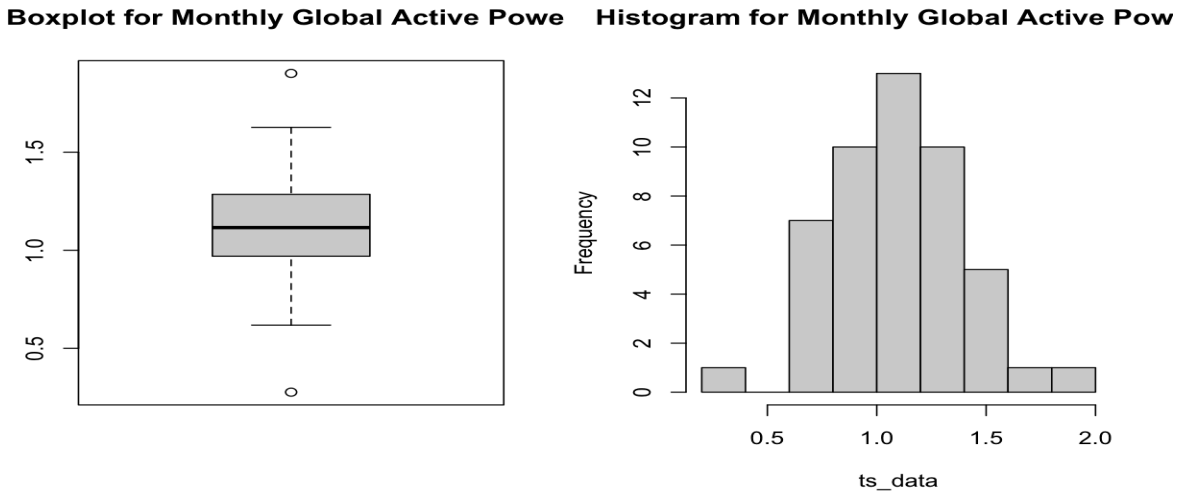

Fig. 2.2 Boxplot and Histogram for Monthly Global Active Power

Time series plot

The time series plot depicts a mild trend; however, there is some fairly constant seasonality over time, indicating a homogeneous behaviour.

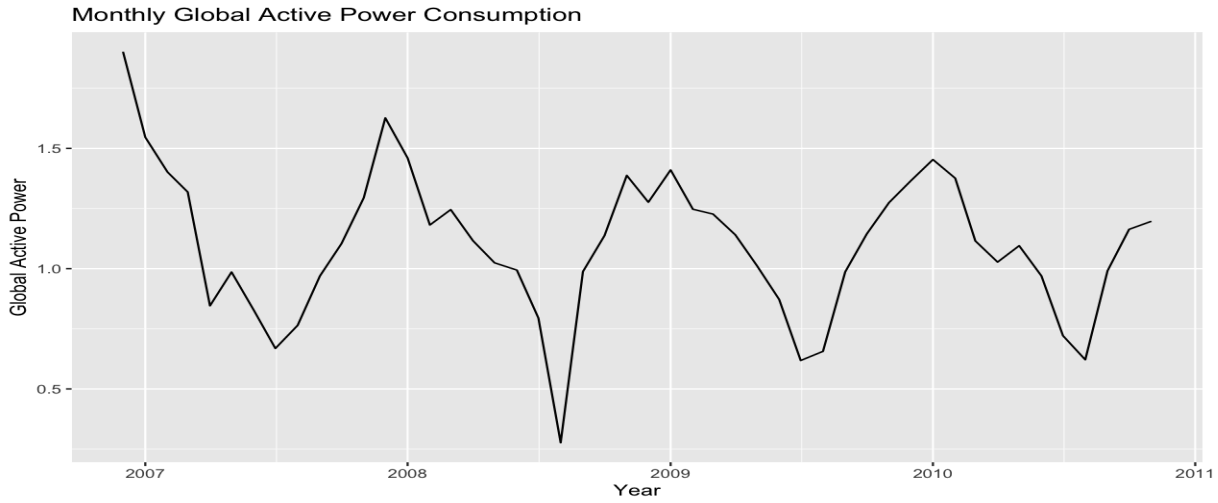

Fig. 2.3 Time series plot for Monthly Global Active Power Dataset

The correlogram shows a slow decay in the autocorrelation function (ACF) at low lags, which does indicate non-stationarity and seasonality. The partial autocorrelation function (PACF) shows a quick decay after lag 1.

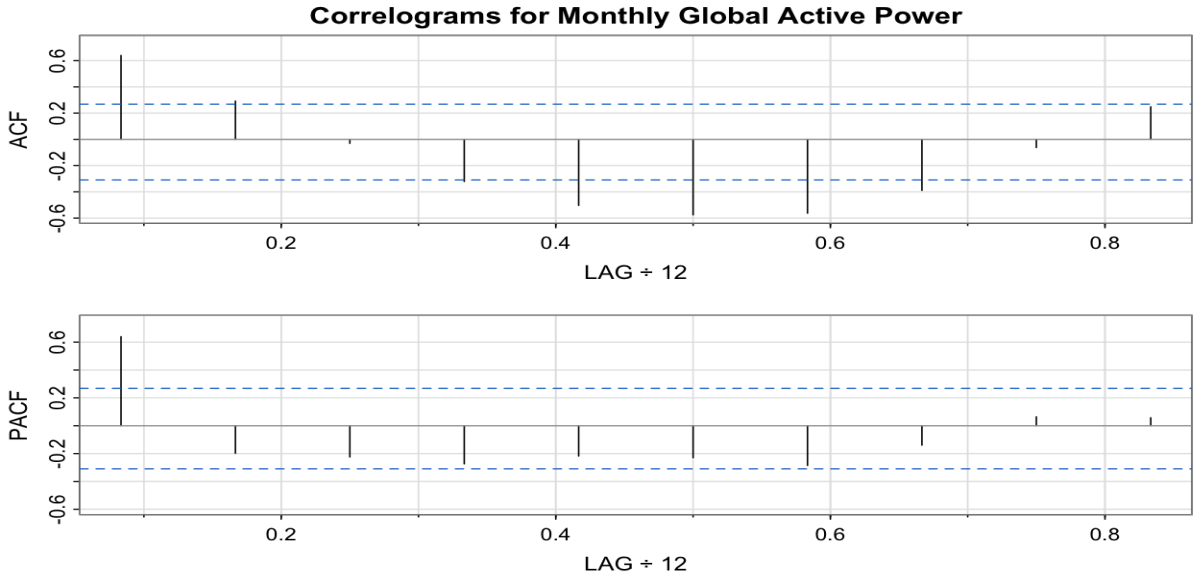

Fig. 2.4 Correlograms for Monthly Global Active Power Dataset

2.3 Decomposition

The additive and multiplicative decomposition models were compared to determine which model best described the dataset's behavior.

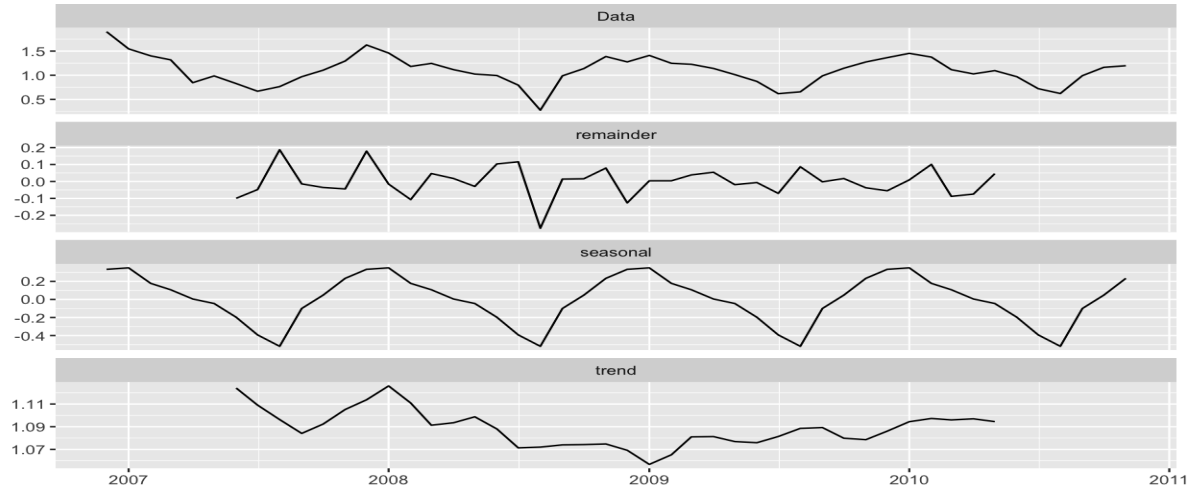

Fig. 2.5 Additive Decomposition of the dataset

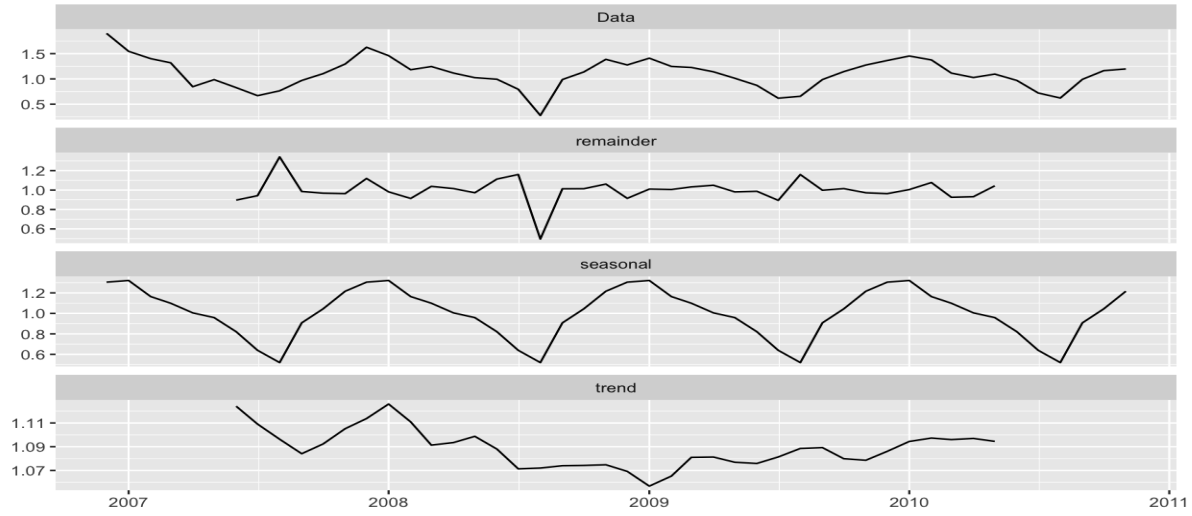

Fig. 2.6 Multiplicative Decomposition of the dataset

The additive model best describes the dataset, as the seasonal component is roughly constant over time, and the variability of the residuals for the additive model is more stable compared to the residuals of the multiplicative model. The seasonal component has more impact on the dataset than the trend, as it does not increase with the level of the series.

2.4 Classical Model Selection

The multiple seasonality trend decomposition loess function also supports the claim that the dataset consists of a mild trend and fairly consistent seasonality.

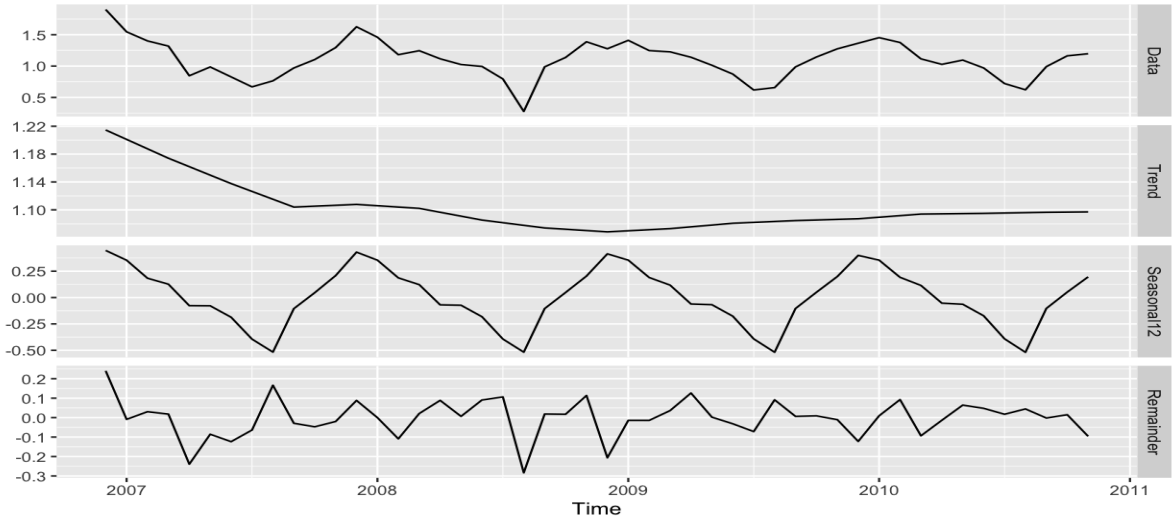

Fig. 2.7 MSTL Decomposition of Monthly Global Active Power Dataset

The exponential smoothing function, ETS, was used to confirm the claim model structure. The chosen model by the ETS function was ETS(A, N, A) (additive errors, no trend, additive seasonality), with very tiny smoothing parameters, indicating a relatively stable level and seasonal pattern. For the monthly global active power dataset, the Holt-Winters exponential smoothing technique was selected as the appropriate classical model because the dataset consists of a seasonal component, which is fairly constant over time.

For this analysis, the Holt-Winters' additive and multiplicative models were compared to decide on which model best describes the dataset.

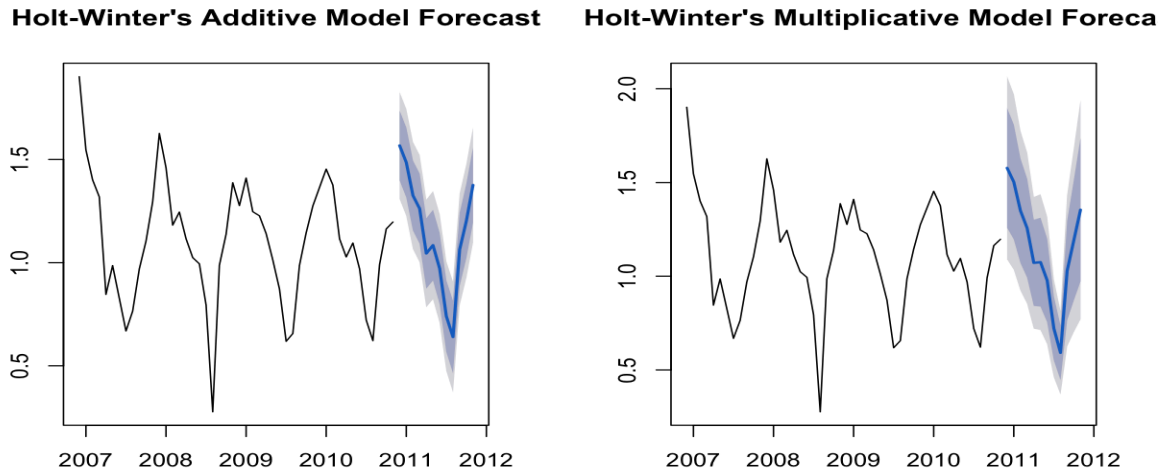

Fig. 2.8 Holt-Winters' Multiplicative Model

An accuracy comparison was also made for both the Holt-Winters Additive and Multiplicative Models.

Error Measurement	**Holt-Winters Additive**	**Holt-Winters Multiplicative**
RMSE	0.1077533	0.1085159
MAE	0.07976781	0.07821437
MPE	0.003607277	-0.8482089
MAPE	8.937339	8.960262

Table 2.1 Accuracy Measures For Holt-Winters' Models

The model selection was based on the RMSE and MAPE. The RMSE penalises larger errors more heavily in the time series dataset, while the MAPSE interprets the scales in percentages. As shown in Table 2.1, the Holt-Winters Additive Model performs best, with lower RMSE and MAPE. For these reasons, the Holt-Winters additive model was selected for forecasting.

2.5 ARIMA Models

A time series is stationary if its mean, variance, and autocorrelation are all constant. The time series plot (Fig. 2.3) suggests a small trend with fairly seasonal monthly patterns, indicating that the dataset is not stationary. The correlograms (Fig. 2.4) depict a slow decay on the autocorrelation function (ACF) and a quick decay in the partial autocorrelation function (PACF) after the first lag. The mean was assumed to be constant, given its small trend. The variance was also assumed to be constant, as the dataset portrays a homogeneous behaviour. A Dickey-Fuller test (ADF) and a level stationarity test (KPSS) were conducted to test for stationarity of the dataset. A 5% level of significance was used for the analysis. The result of the ADF test was $p = 0.058 > 0.05$, so we fail to reject H_o. The KPSS test produced a $p = 0.1 > 0.05$. So we fail to reject H_o. Since the ADF and KPSS tests show conflicting evidence on stationarity, the time series was assumed to be non-stationary.

Given that there is a fairly constant seasonal pattern in the series (Fig. 2.3 and Fig. 2.4), we differentiate with respect to seasonality.

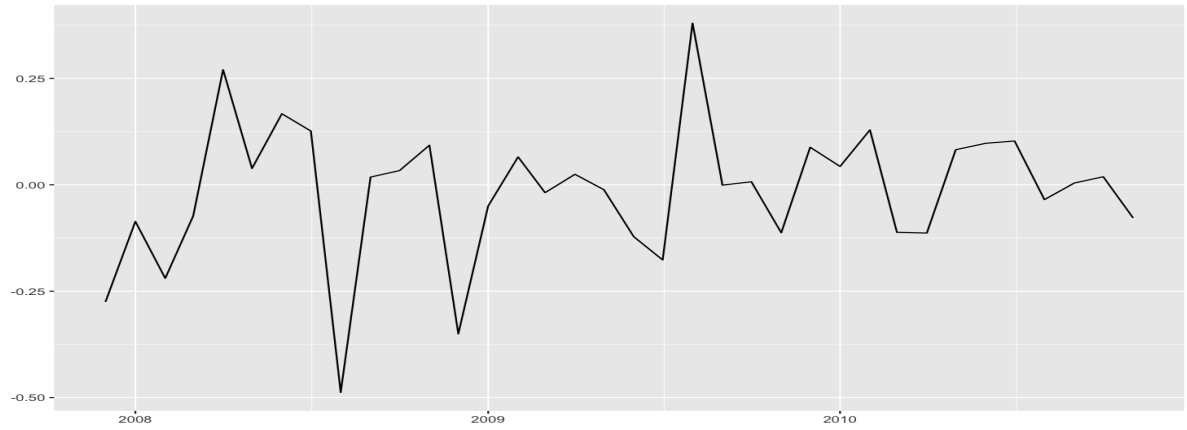

Fig. 2.8 Seasonal Differencing Plot

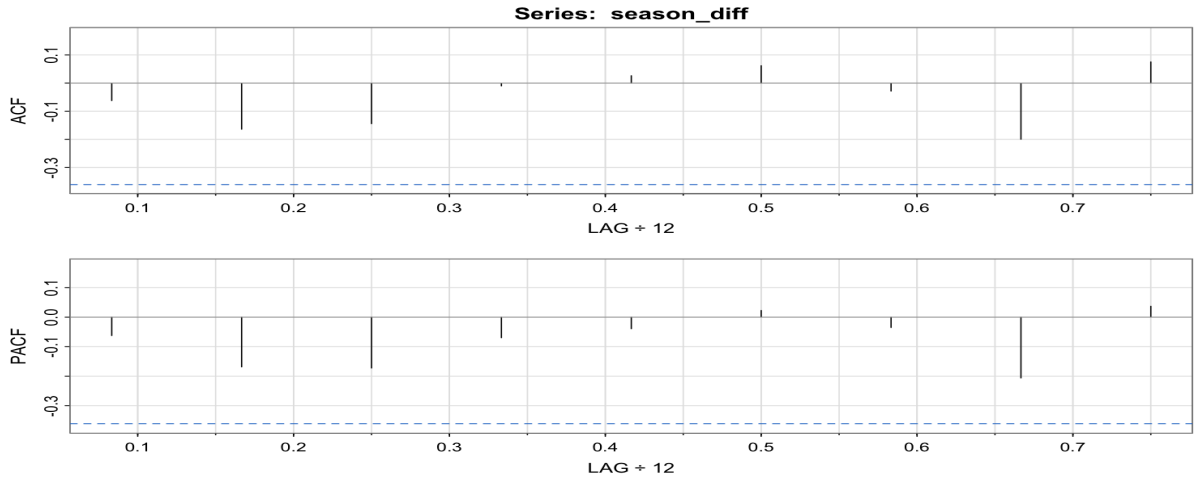

Fig. 2.9 Correlogram of Seasonal Differentiating Data

After seasonal differencing, the ACF and PACF show that all the autocorrelations are within the confidence bound, which means that there is no relationship between the current observations and the past values or residuals. The model obtained from the correlograms is **SARIMA(p=0, d=0, q=0) (P=0, D=1, Q=0) s=12**, where p is the non-seasonal order for the AR model, d is the non-seasonal differentiation order, q is the non-seasonal order for the MA model, P is the seasonal order for the AR model, D is the seasonal differentiation order, Q is the seasonal order for the MA model, s is the number of periods, and there were no more AR or MA models found in the plot**.** The auto ARIMA function also produced the same model. The seasonally differenced dataset was tested to confirm that it is stationary. The result of the ADF test was $p = 0.01 < 0.05$, so we reject H_o. The KPSS test produced a $p = 0.1 > 0.05$. So we fail to reject H_o. Since there is no conflicting evidence regarding stationarity between the ADF and KPSS tests, the time series was assumed to be stationary.

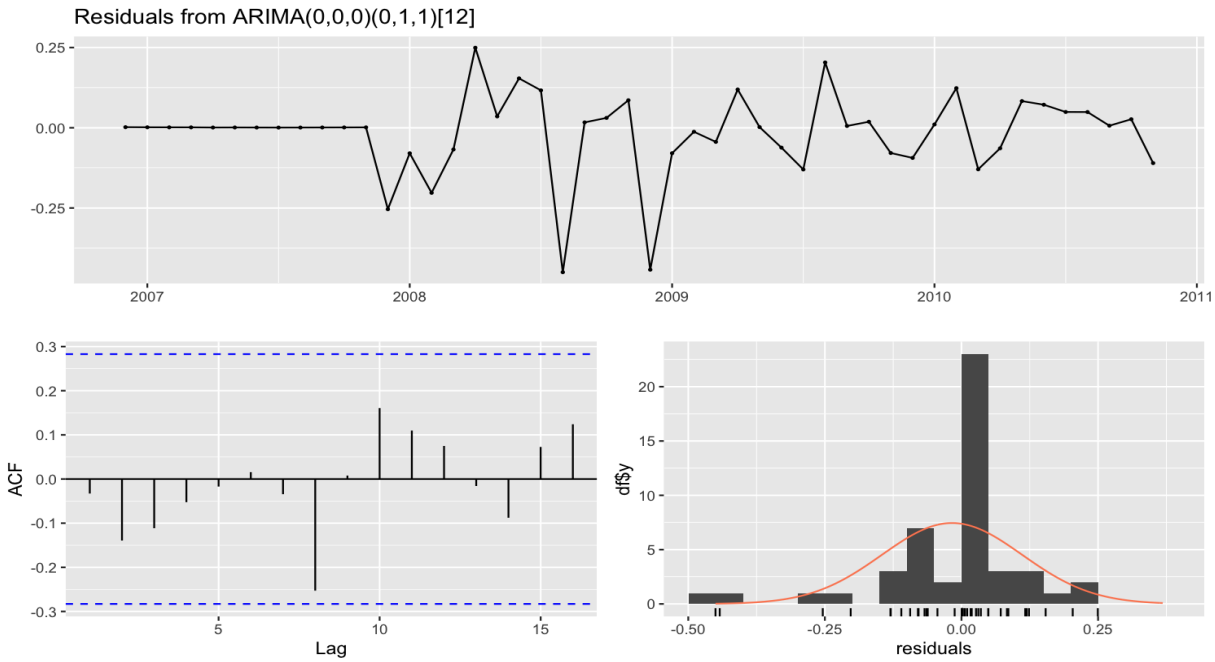

Fig. 2.10 Residual Diagnostic

A residual diagnostic was conducted to validate the derived SARIMA model, as shown in Fig. 2.13. The residuals fluctuate around a constant mean of zero, indicating that there is no trend. It also shows the behaviour of white noise. The histogram also suggests that the residuals may be normally distributed; however, there appears to be an outlier with a frequency of 20. The correlograms show that all the autocorrelations are within the confidence bounds, with no spikes at any lags. Through trial and error, other models were compared with the derived model and auto.arima. The AIC (Akaike Information Criterion) information criteria metric was used to compare the models.

Model	**AIC**
SARIMA(0,0,0)$(0,1,0)_{12}$ (model and auto arima)	-27.73
SARIMA(0,0,1)$(0,1,1)_{12}$	-27.03
SARIMA(0,0,0)$(0,1,1)_{12}$	-29.02

Table 2.2 AIC Comparison of SARIMA Models

Based on this metric, **SARIMA(0,0,0)(0,1,1)$_{12}$** is the preferred model because it has a lower AIC compared to other models. To ensure the data has constant autocorrelation, the Ljung-Box test was conducted to test whether the residuals of the time series are independently distributed. The test produced p = 0.5909 > 0.05, so we fail to reject H_o.

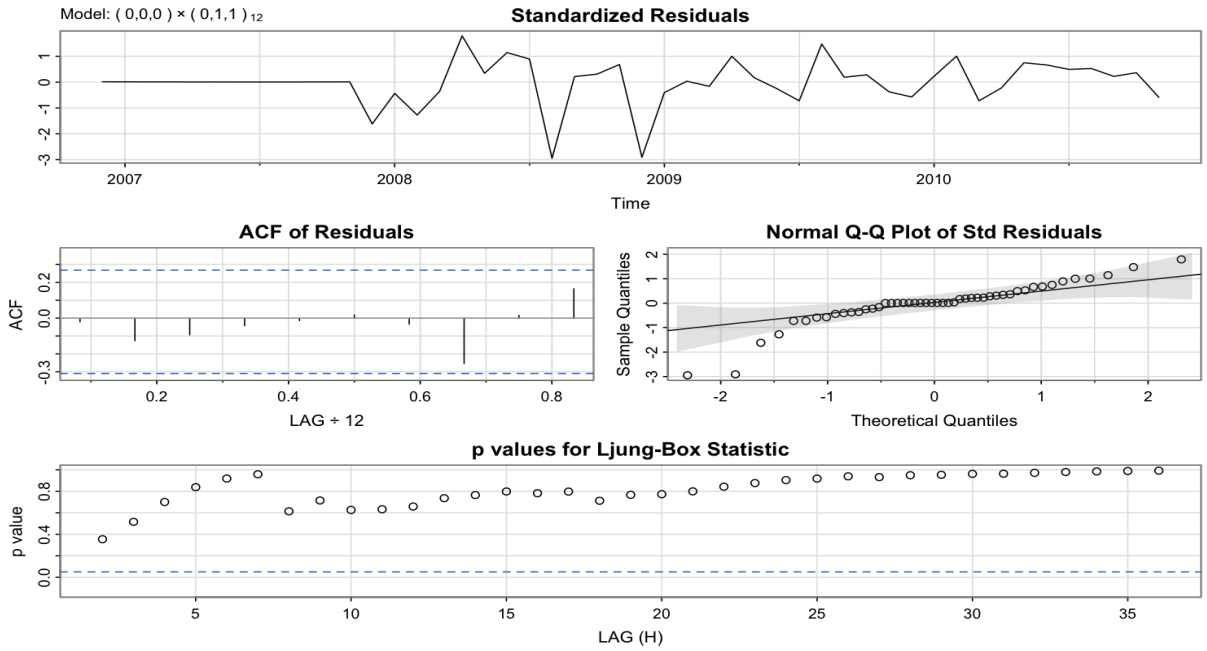

Fig. 2.11 Ljung-Box Test for independent residuals

Additionally, the Ljung-Box test p-value plot shows that all the lags are above the 5% significance level, indicating that there is significant evidence that the residuals are independent. Therefore, the dataset now follows the rule of constant autocorrelation. A 12-month forecast was produced for the derived ARIMA model, **SARIMA(0,0,0)(0,1,1)$_{12}$**. It was expected to follow a similar homogeneous behaviour, with fairly constant seasonal fluctuations.

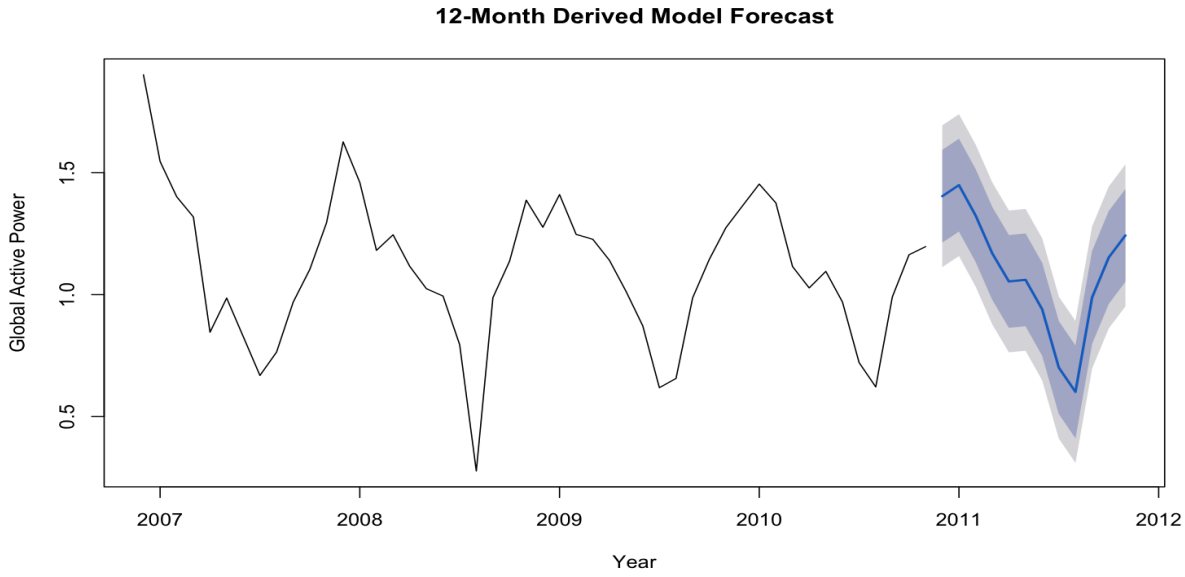

Fig. 2.12 12-month Forecast for SARIMA (0,0,0) $(0,1,1)_{12}$

2.6 Model Comparison

The error measures for the SARIMA and Holt-Winters additive models were also compared to select the best forecasting model.

Model	RMSE	MAE	MPE	MAPE
Holt-Winters Additive	0.1077533	0.07976781	0.003607277	8.937339
SARIMA	0.1283216	0.07858699	-3.446834	9.865313

Table 2.3 Accuracy comparison between Holt-Winters Additive and SARIMA model

Based on these error measures, the Holt-Winters additive model is the best model, as it has fewer errors in key error metrics (RMSE, MAE, and MAPE are low).

3 Discussion and Conclusion

The monthly global active power consumption data shows a mild trend with fairly constant seasonal patterns. The additive model was preferred in decomposition and ETS(A, N, A) analysis. The Holt-Winters additive model was chosen as the best classical model for the series since it has constant seasonal fluctuations. The ARIMA model found was **SARIMA (0,0,0) (0,1,1)12**, and the auto ARIMA function provided the same results. The error measures of both the Holt-Winters additive and the derived ARIMA model were compared, and Holt-Winters was preferred, as its error measures were lower compared to the ARIMA model.

References

Hebrail, G., & Berard, A. (2006). *Individual Household Electric Power Consumption [Dataset]*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption

Marino, D. L., Amarasinghe, K., & Manic, M. (n.d.). Building Energy Load Forecasting using Deep Neural Networks. https://www.semanticscholar.org/reader/addba7f62763606216dd1d9b34908b3c6f9158db

Parate, A., & Bhoite, S. (2019). Individual Household Electric Power Consumption Forecasting using Machine Learning Algorithms. *International Journal of Computer Applications Technology and Research*, *8*(09), 371-374. https://ijcat.com/archieve/volume8/issue9/ijcatr08091007.pdf