Power Grid Load Forecasting using Machine Learning Approaches: Bachelor's Project

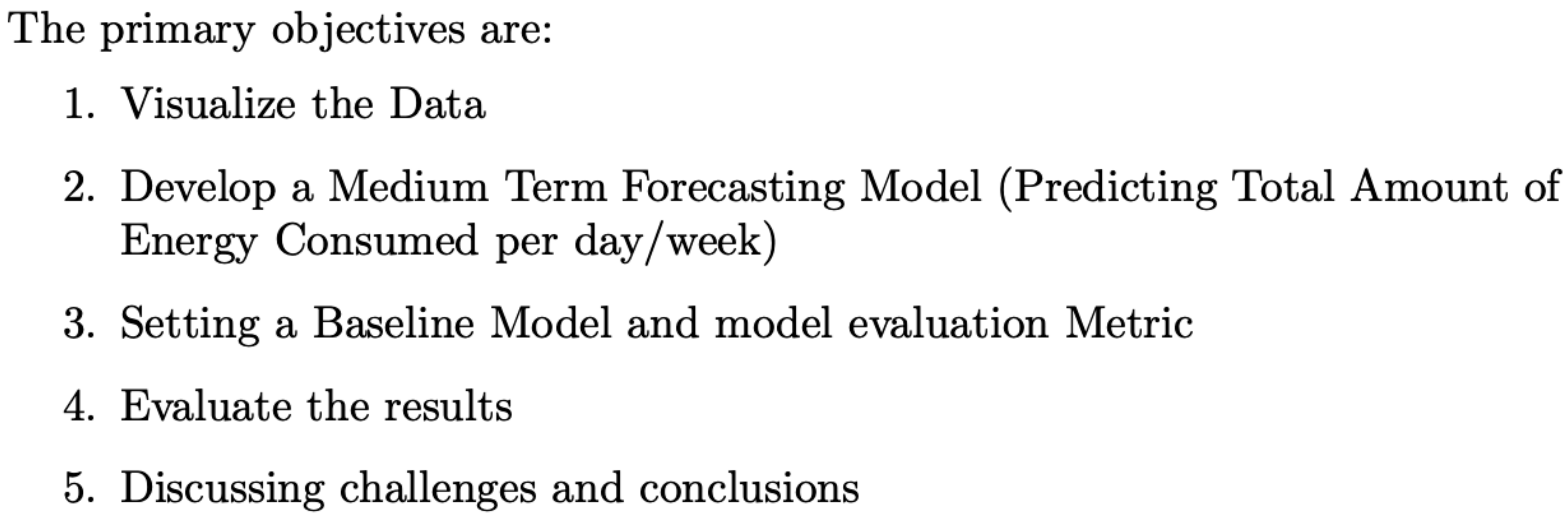
Supervised by François Fleuret, Youssef Saïed, Clément Targe

Reason

By leveraging historical energy consumption data alongside weather and seasonal variables, this project aims to create accurate medium-term energy consumption forecasts and measure their effectiveness. This project adopts time series analysis techniques, like ARIMA and SARIMA models, for medium-term forecasting.

Medium-term forecasting plays a crucial role in balancing supply and demand, scheduling energy imports or hydropower reserves, and planning grid maintenance. Time series models provide interpretable, data-driven forecasts that align well with Switzerland's structured energy market and regulatory planning needs.

Objective



The primary objectives are:

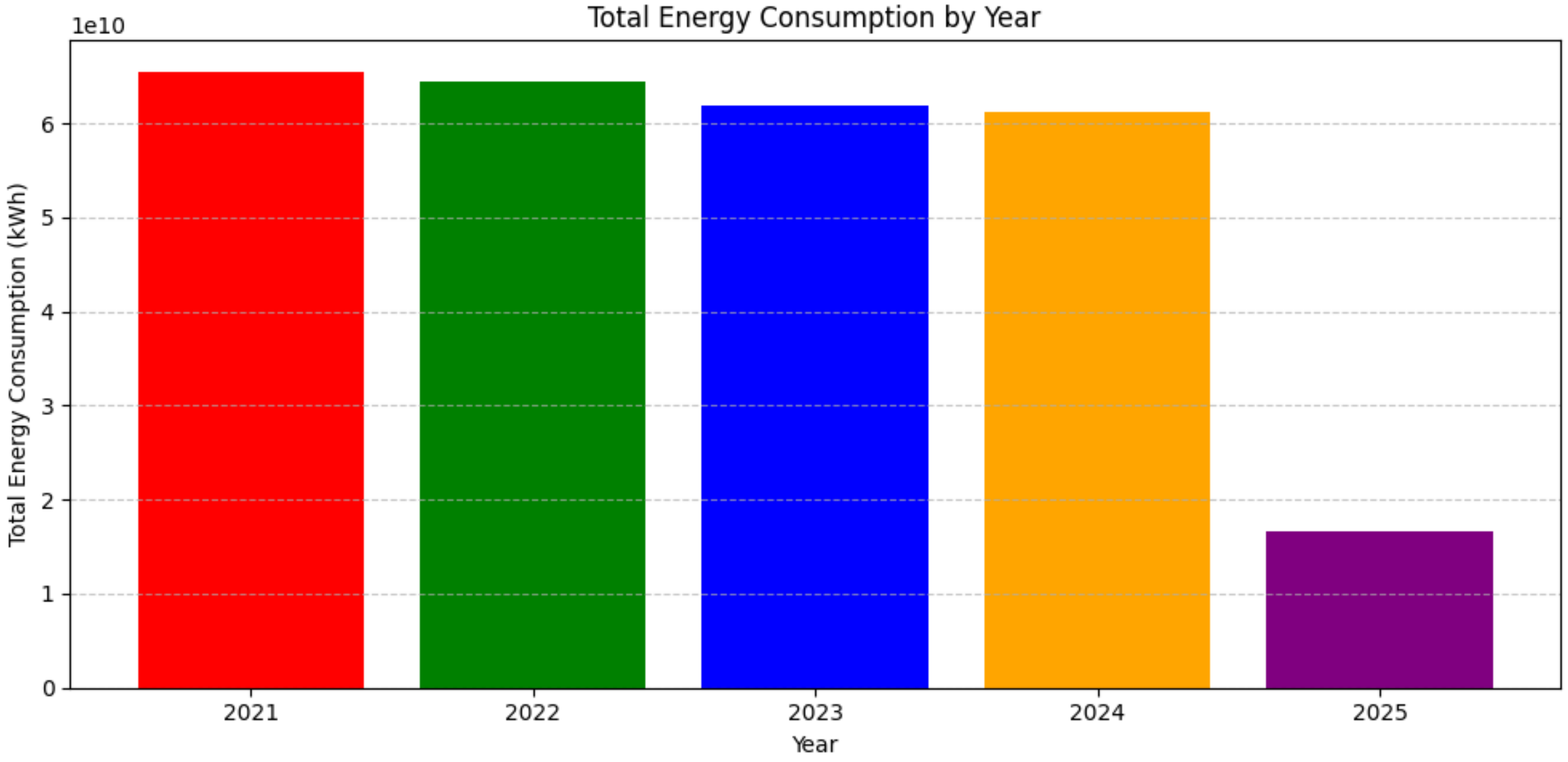
1. Visualize the Data
2. Develop a Medium Term Forecasting Model (Predicting Total Amount of Energy Consumed per day/week)
3. Setting a Baseline Model and model evaluation Metric
4. Evaluate the results
5. Discussing challenges and conclusions

Outline

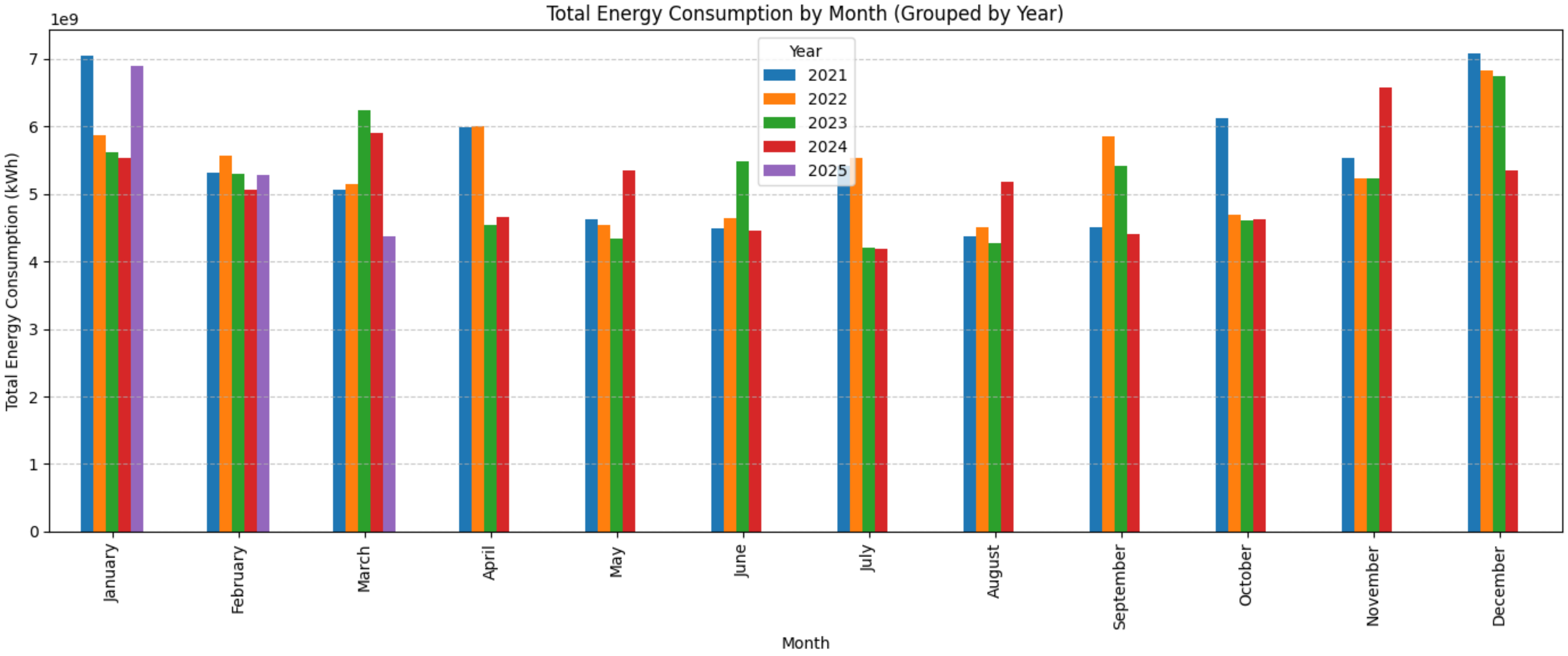
- I. Visualisation
- II. Data Cleaning
- III. Time Series Analysis
- IV. Models
- V. Evaluation
- VI. Challenges & Discussion

I. Visualisation

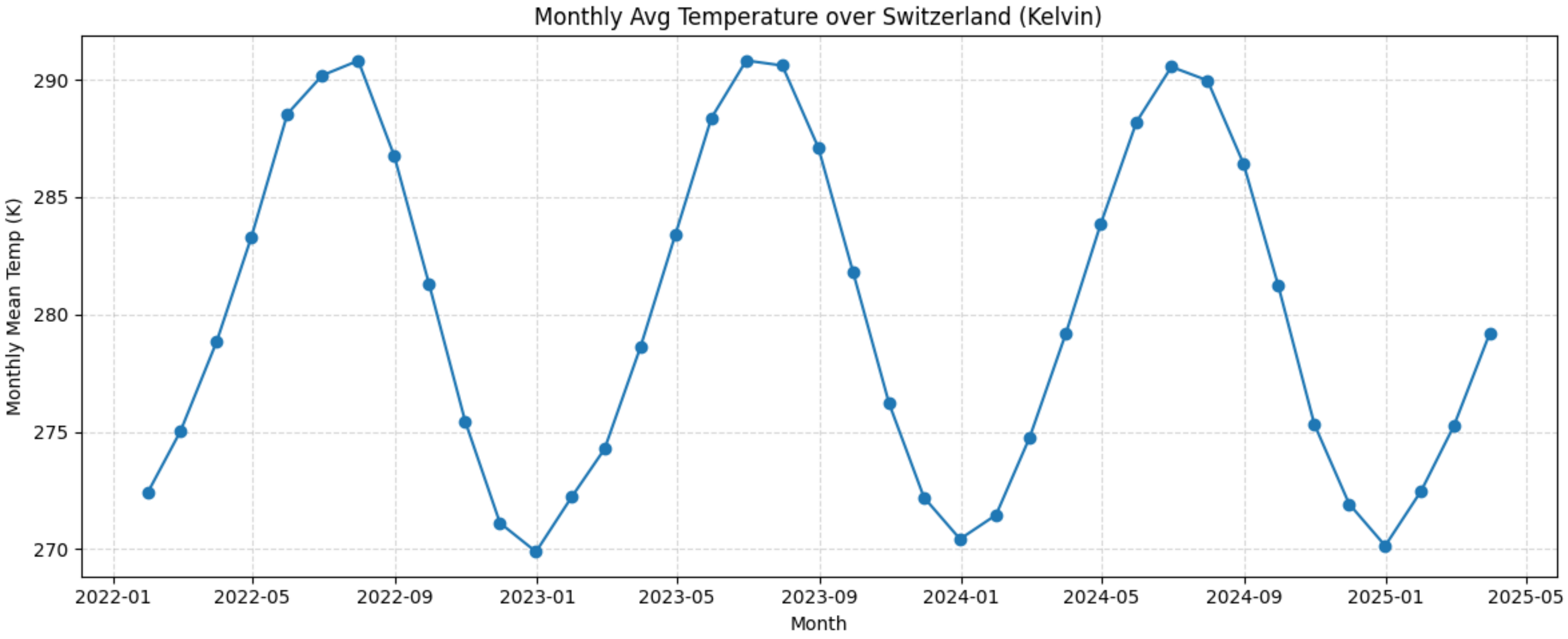
I. Visualisation



I. Visualisation



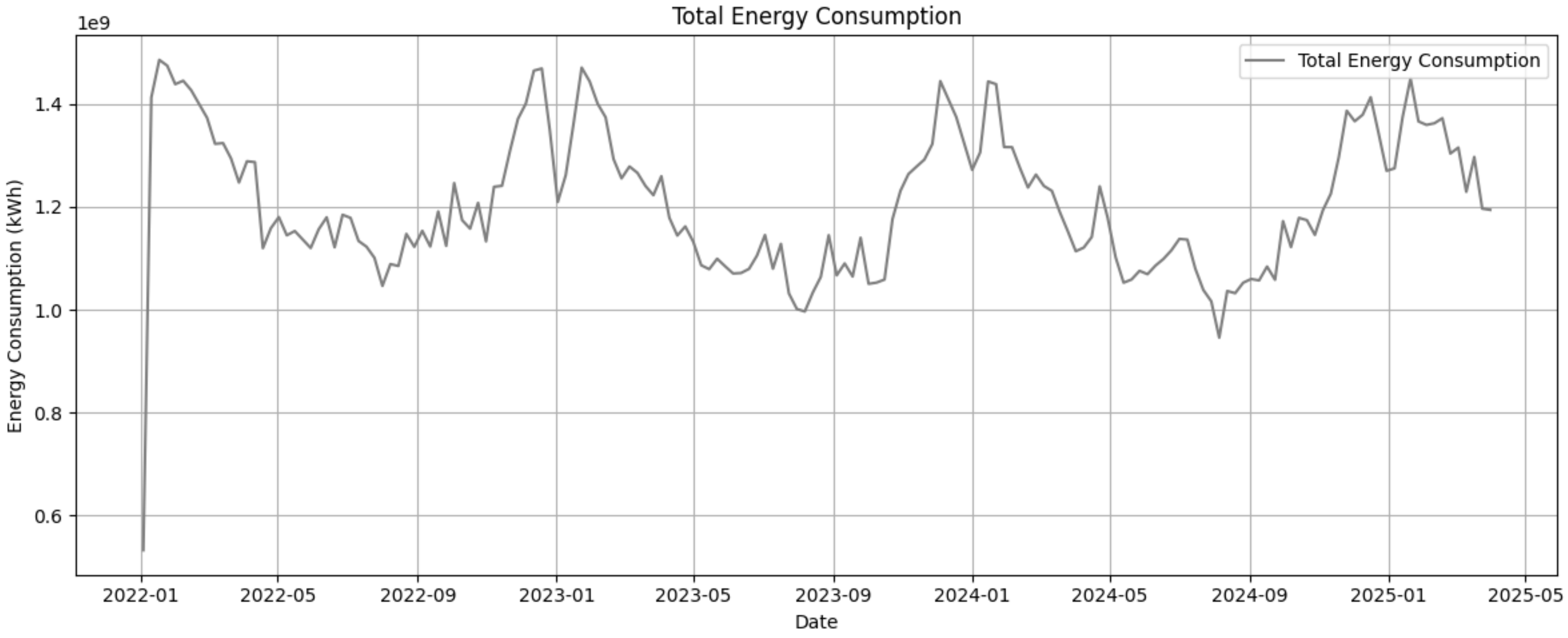
I. Visualisation of related variables, Weather Data



Variable to predict

Weekly sum

- The weekly predicted sum will be from Monday-Sunday, as is per EU energy regulations



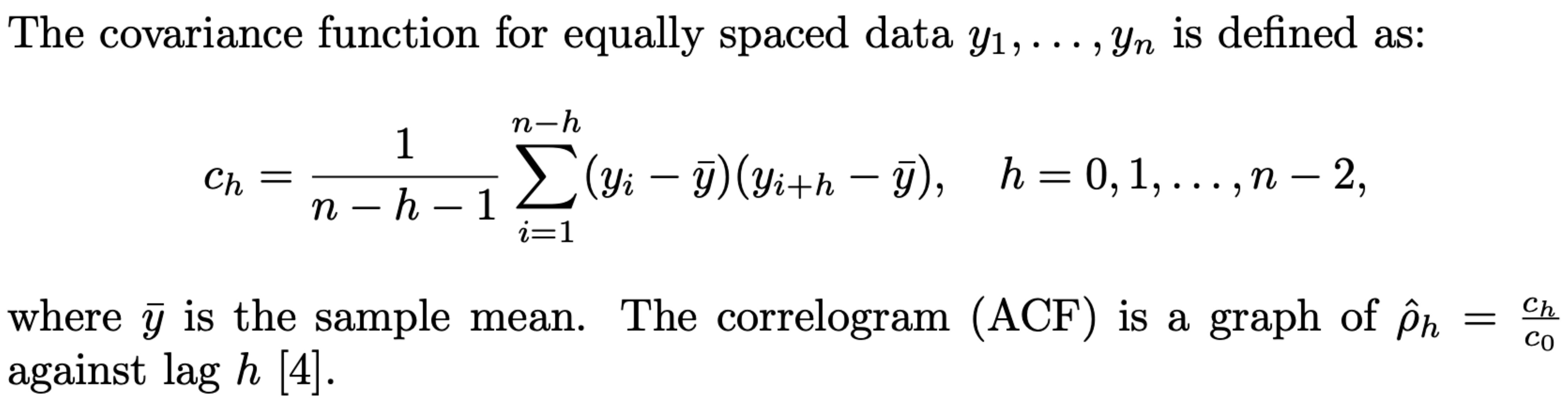
II. Time Series Analysis

• III. Time Series Analysis

- I. ACF: which lags are most important, does my data have a moving average?
- II. KPSS Test: is my data stationnary?
- III. Periodogram: Predicting seasonality, and is my data white noise?
- IV. After modeling, residual analysis, are they (white noise/normal)?

Autocovariance function

- Measures the linear dependence of a time series with a lagged version of itself
- Quantifies how much the values of a time series at different points in time are related to each other, specifically in terms of their covariance



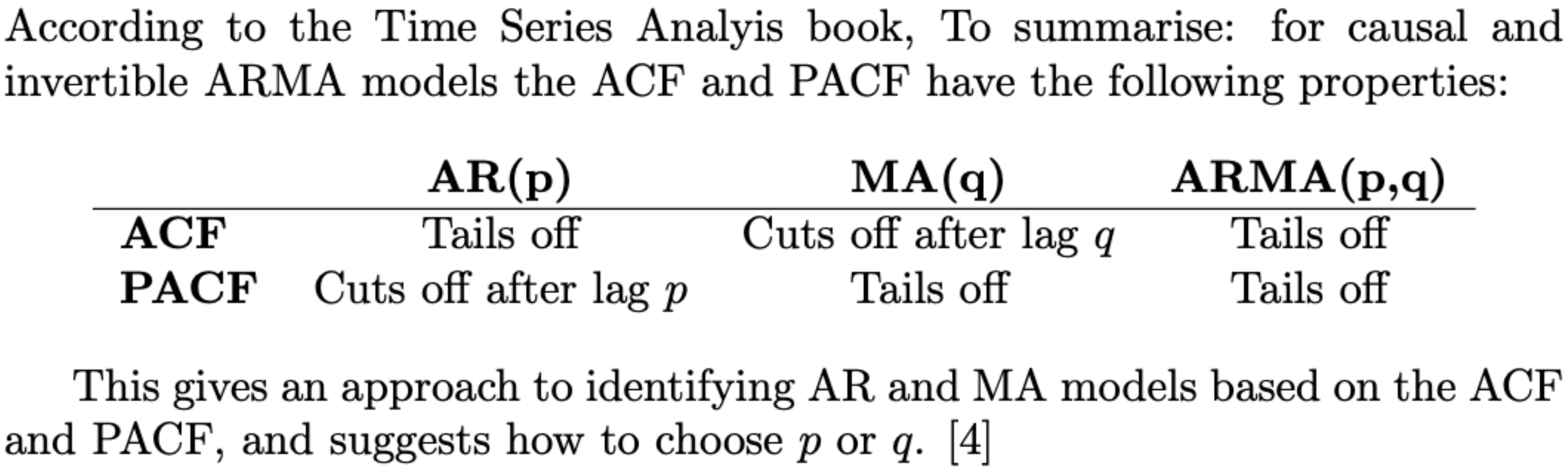
The covariance function for equally spaced data $y_1, \ldots, y_n$ is defined as:

$$c_h = \frac{1}{n-h-1} \sum_{i=1}^{n-h} (y_i - \bar{y})(y_{i+h} - \bar{y}), \quad h = 0, 1, \ldots, n-2,$$

where $\bar{y}$ is the sample mean. The correlogram (ACF) is a graph of $\hat{\rho}_h = \frac{c_h}{c_0}$ against lag h [4].

Interpretation

Source



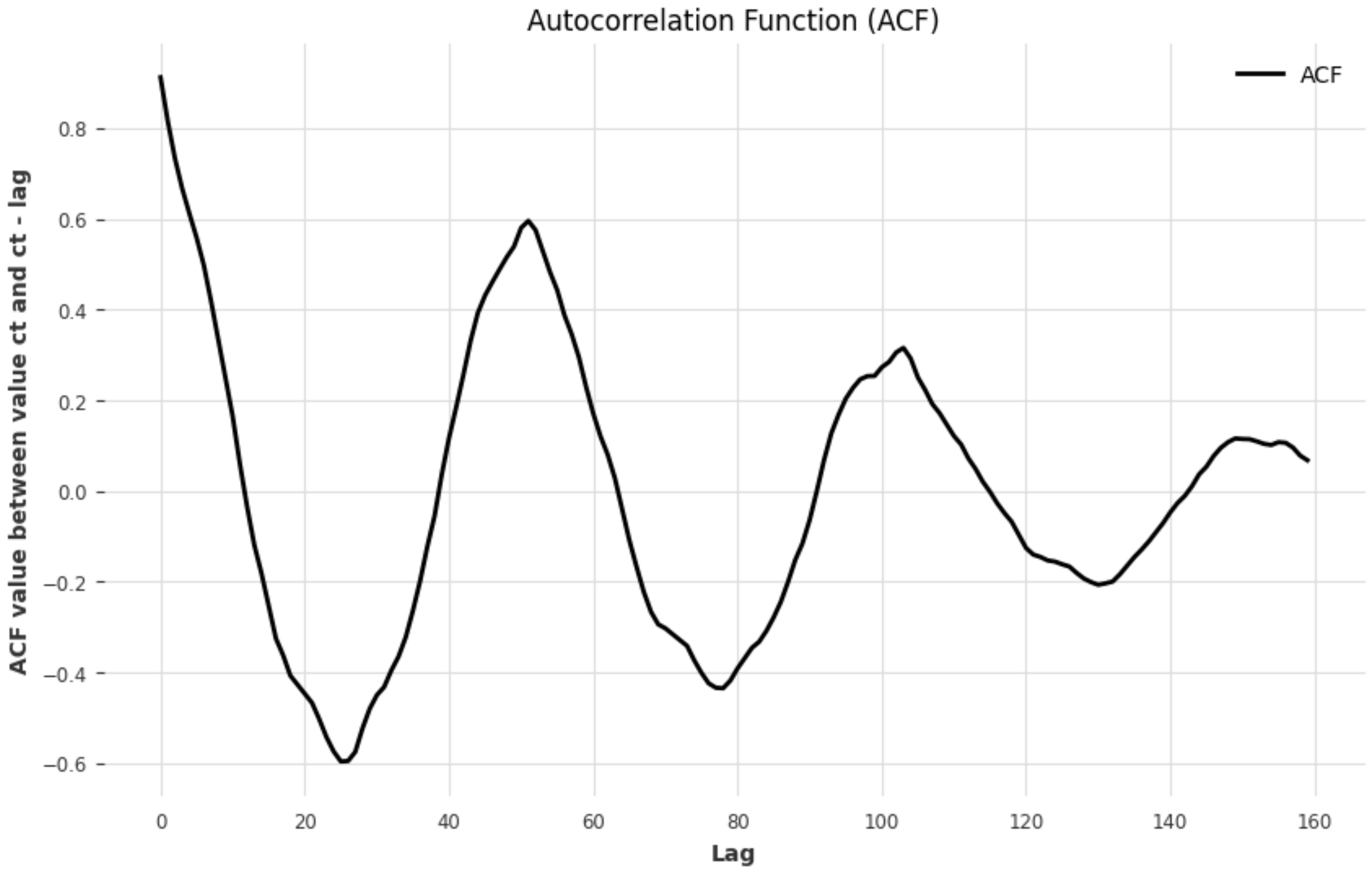
According to the Time Series Analyis book, To summarise: for causal and invertible ARMA models the ACF and PACF have the following properties:

	AR(p)	**MA(q)**	**ARMA(p,q)**
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

This gives an approach to identifying AR and MA models based on the ACF and PACF, and suggests how to choose p or q. [4]

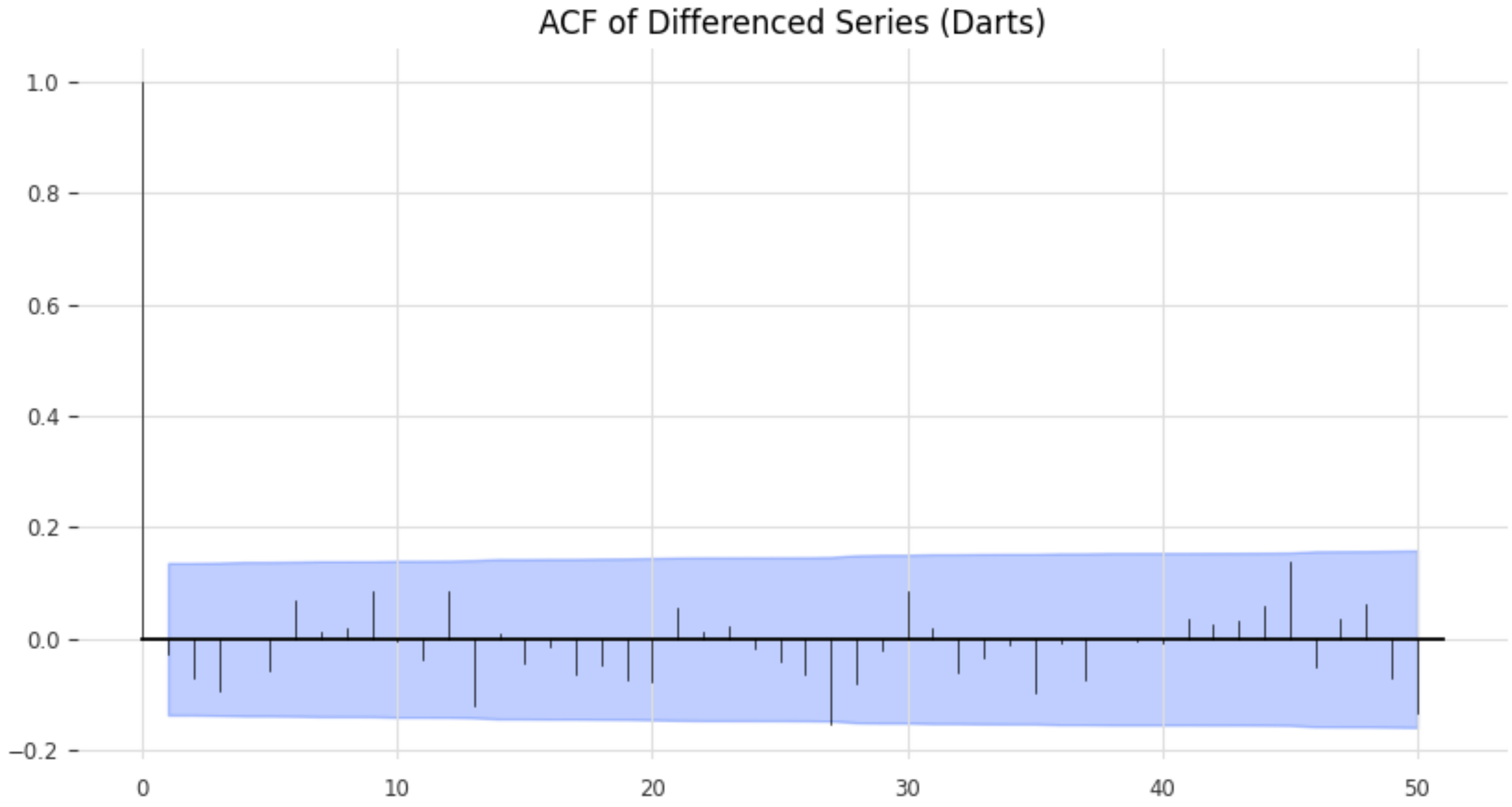
Autocorrelation function of original data

Plotted lags for 160



ACF tails off, no sharp drop, is cyclical (=> potential seasonality)
High at lag 1, then rapidly decays, likely suggesting an AR(1) model
peaking every time at a 52-week cycle

Testing for MA order
ACF of differenced data

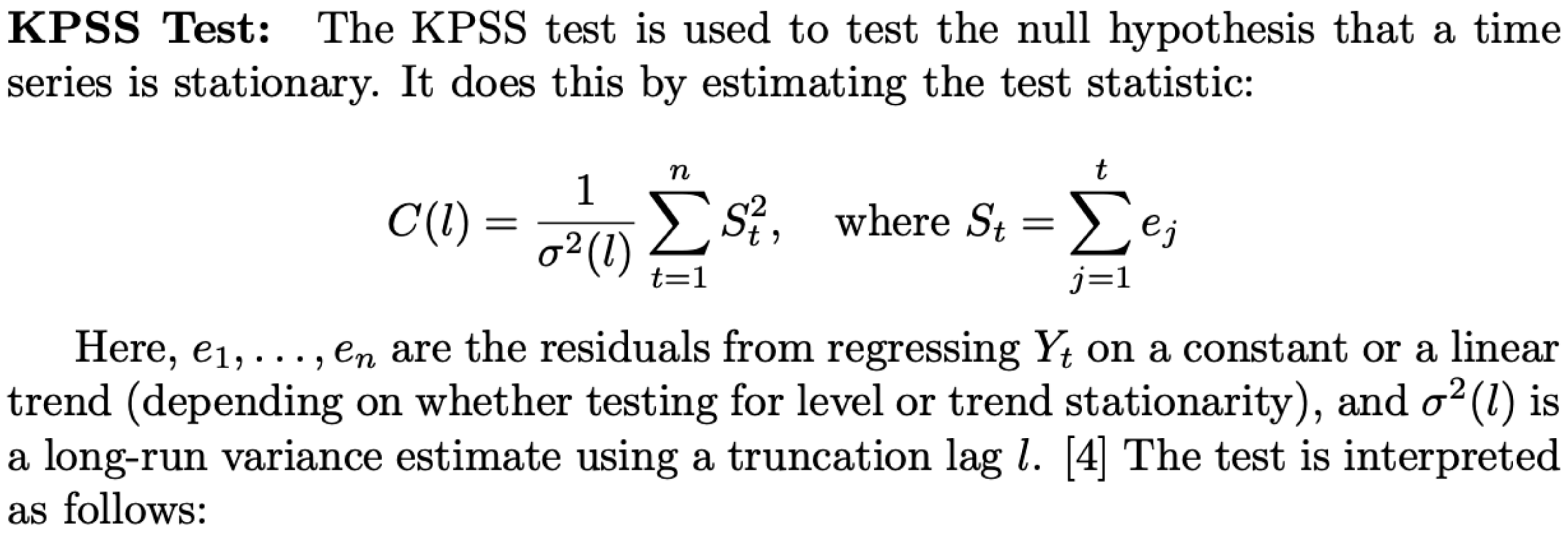


The ACF values of the deseasoned ts show no significant cutout after a certain lag, this does not suggest a ARMA model

q = 0

Testing for stationarity

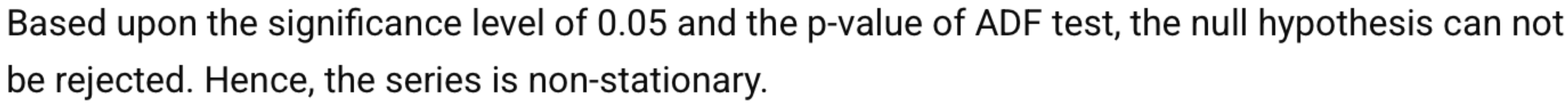
A stationary time series is a time series whose statistical properties like mean, variance, and autocorrelation are all constant over time



KPSS Test: The KPSS test is used to test the null hypothesis that a time series is stationary. It does this by estimating the test statistic:

$$C(l) = \frac{1}{\sigma^2(l)} \sum_{t=1}^{n} S_t^2, \quad \text{where } S_t = \sum_{j=1}^{t} e_j$$

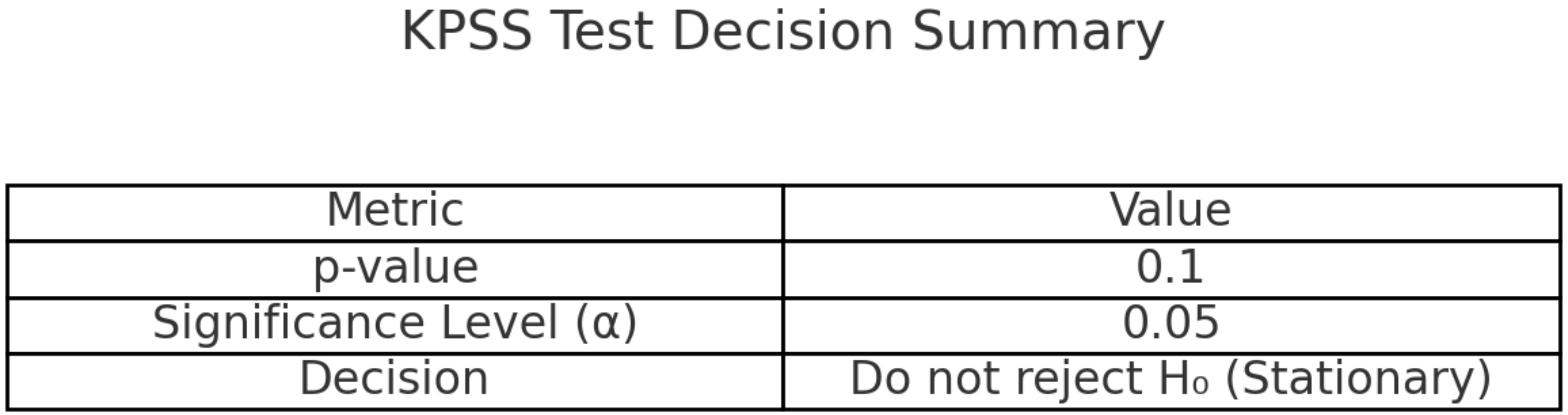
Here, $e_1, \ldots, e_n$ are the residuals from regressing Y_t on a constant or a linear trend (depending on whether testing for level or trend stationarity), and $\sigma^2(l)$ is a long-run variance estimate using a truncation lag l. [4] The test is interpreted as follows:



Based upon the significance level of 0.05 and the p-value of ADF test, the null hypothesis can not be rejected. Hence, the series is non-stationary.

If kpss result is low ==> stationary

Interpreting KPSS



KPSS Test Decision Summary

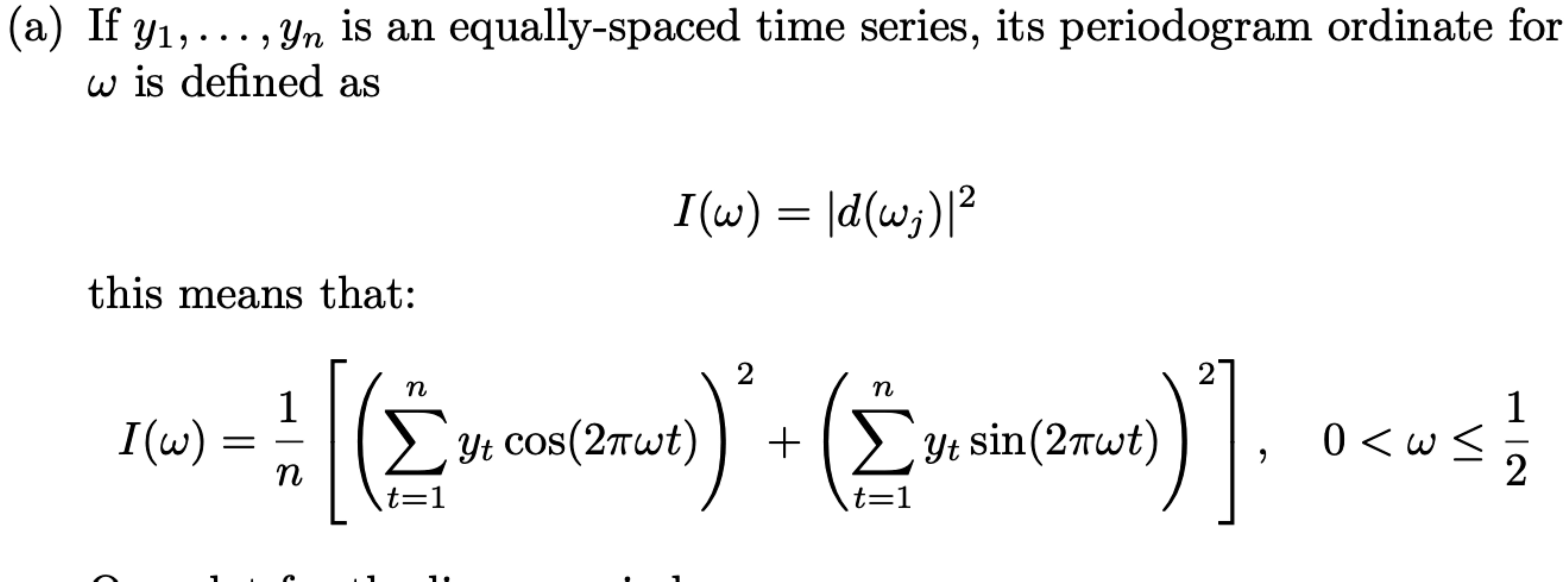
Metric	Value
p-value	0.1
Significance Level (α)	0.05
Decision	Do not reject H_0 (Stationary)

If kpss result is low ==> stationary (source: statsmodel)

No need for ARIMA then (instead of ARMA), d=0

Linear Periodogram

Definition



(a) If $y_1, \ldots, y_n$ is an equally-spaced time series, its periodogram ordinate for ω is defined as

$$I(\omega) = |d(\omega_j)|^2$$

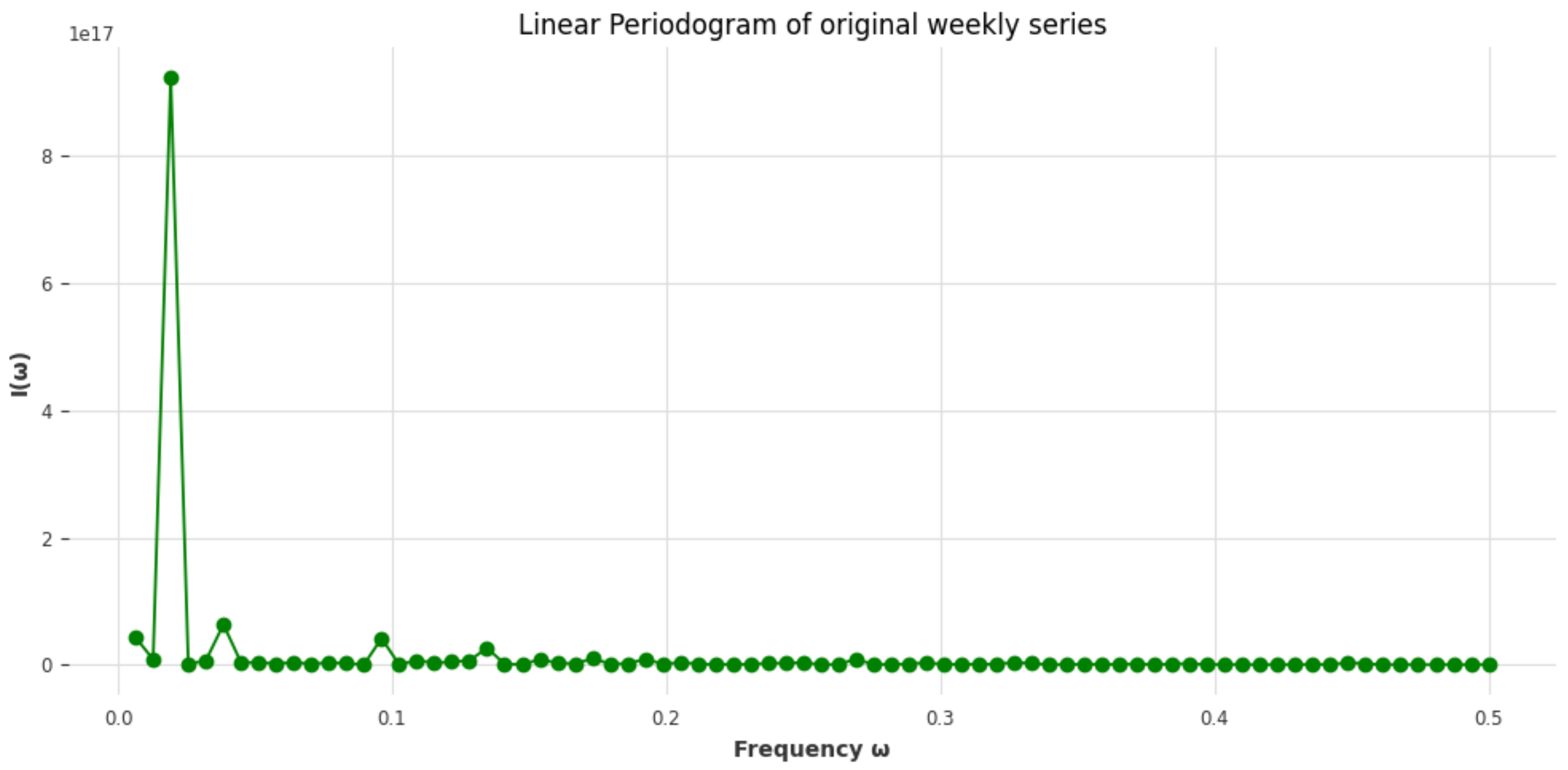
this means that:

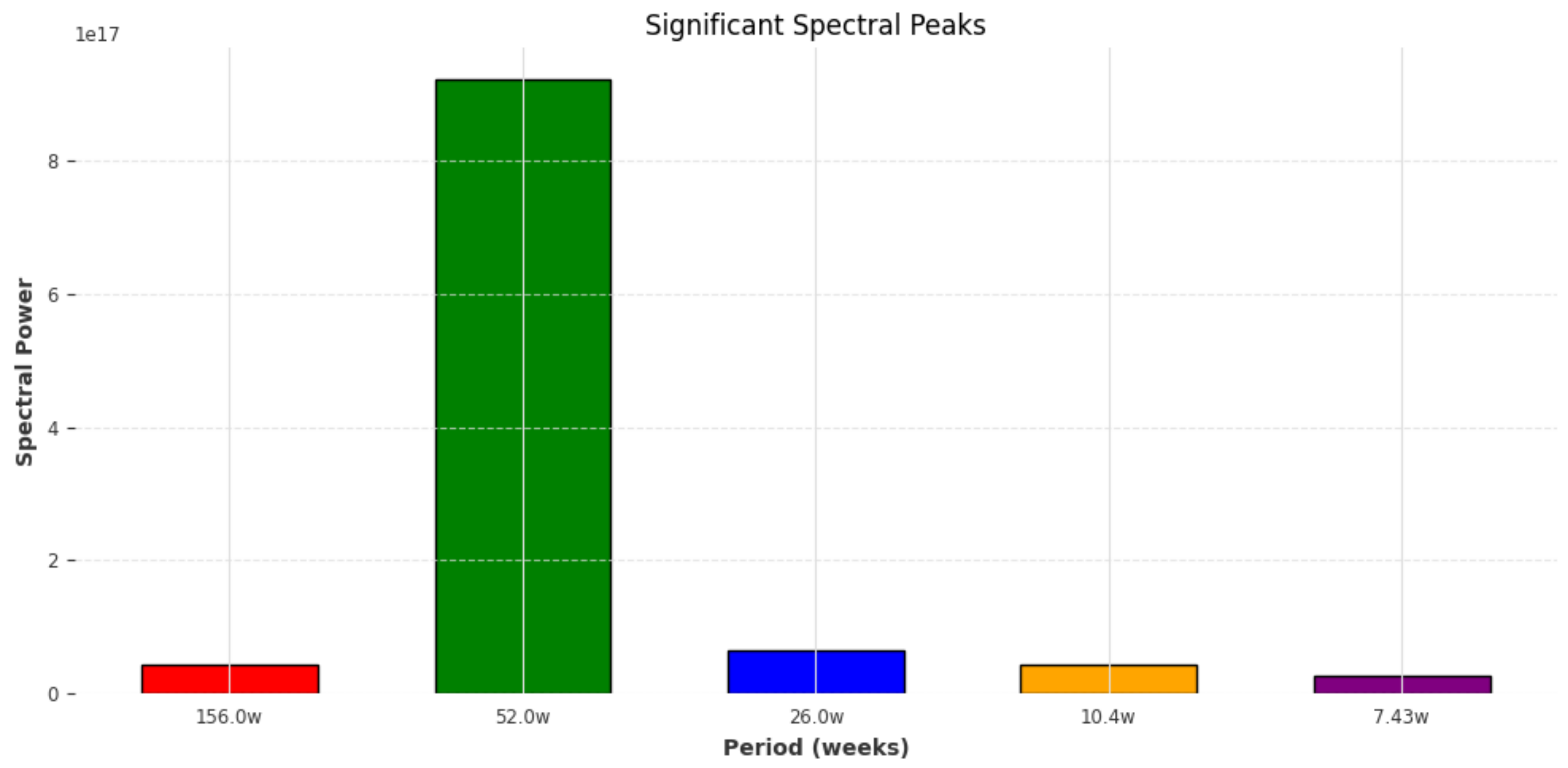
$$I(\omega) = \frac{1}{n}\left[\left(\sum_{t=1}^{n} y_t \cos(2\pi\omega t)\right)^2 + \left(\sum_{t=1}^{n} y_t \sin(2\pi\omega t)\right)^2\right], \quad 0 < \omega \leq \frac{1}{2}$$

Linear Periodogram

Results

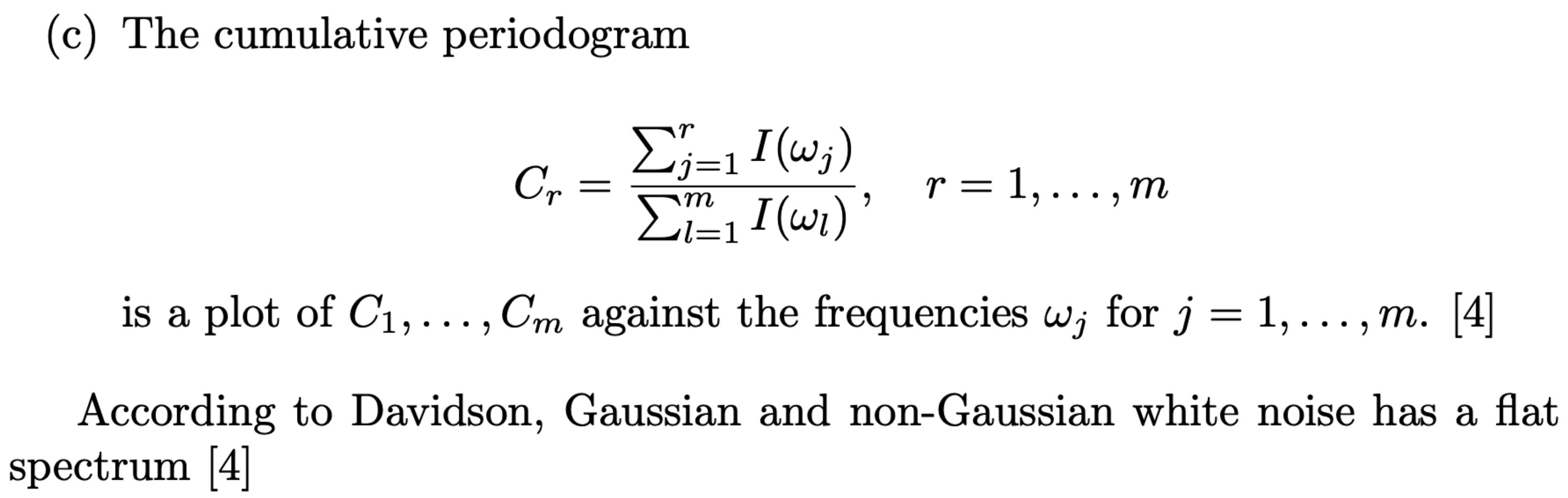
- The spectral decomposition shows that the highest frequency is of 52 weeks, this clearly demonstrates a yearly seasonality.
- The other frequencies are much weaker





Testing for white noise

Cumulative Periodogram compared to Gaussian White noise spectrum

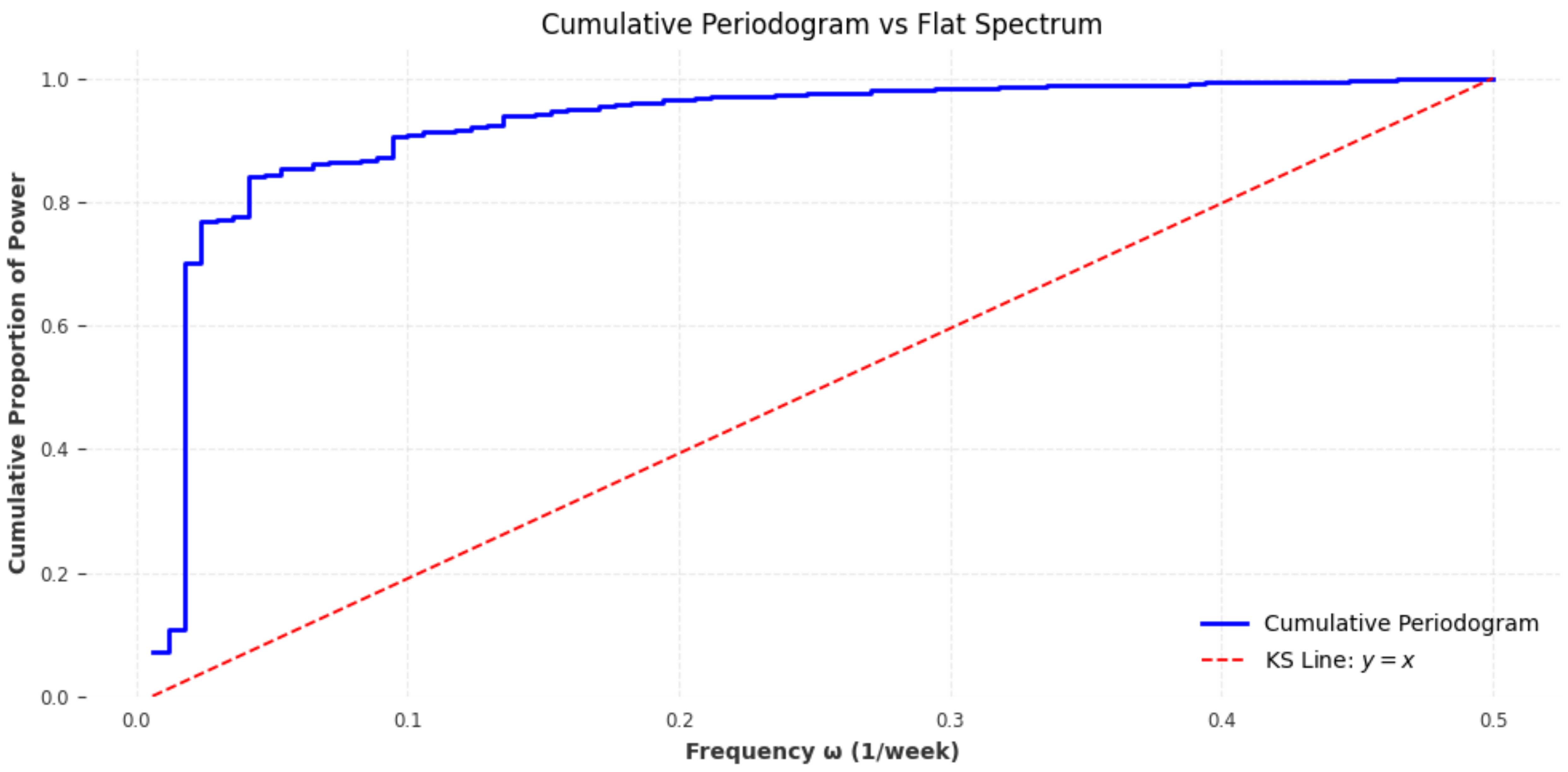


(c) The cumulative periodogram

$$C_r = \frac{\sum_{j=1}^{r} I(\omega_j)}{\sum_{l=1}^{m} I(\omega_l)}, \quad r = 1, \ldots, m$$

is a plot of $C_1, \ldots, C_m$ against the frequencies ω_j for $j = 1, \ldots, m$. [4]

According to Davidson, Gaussian and non-Gaussian white noise has a flat spectrum [4]



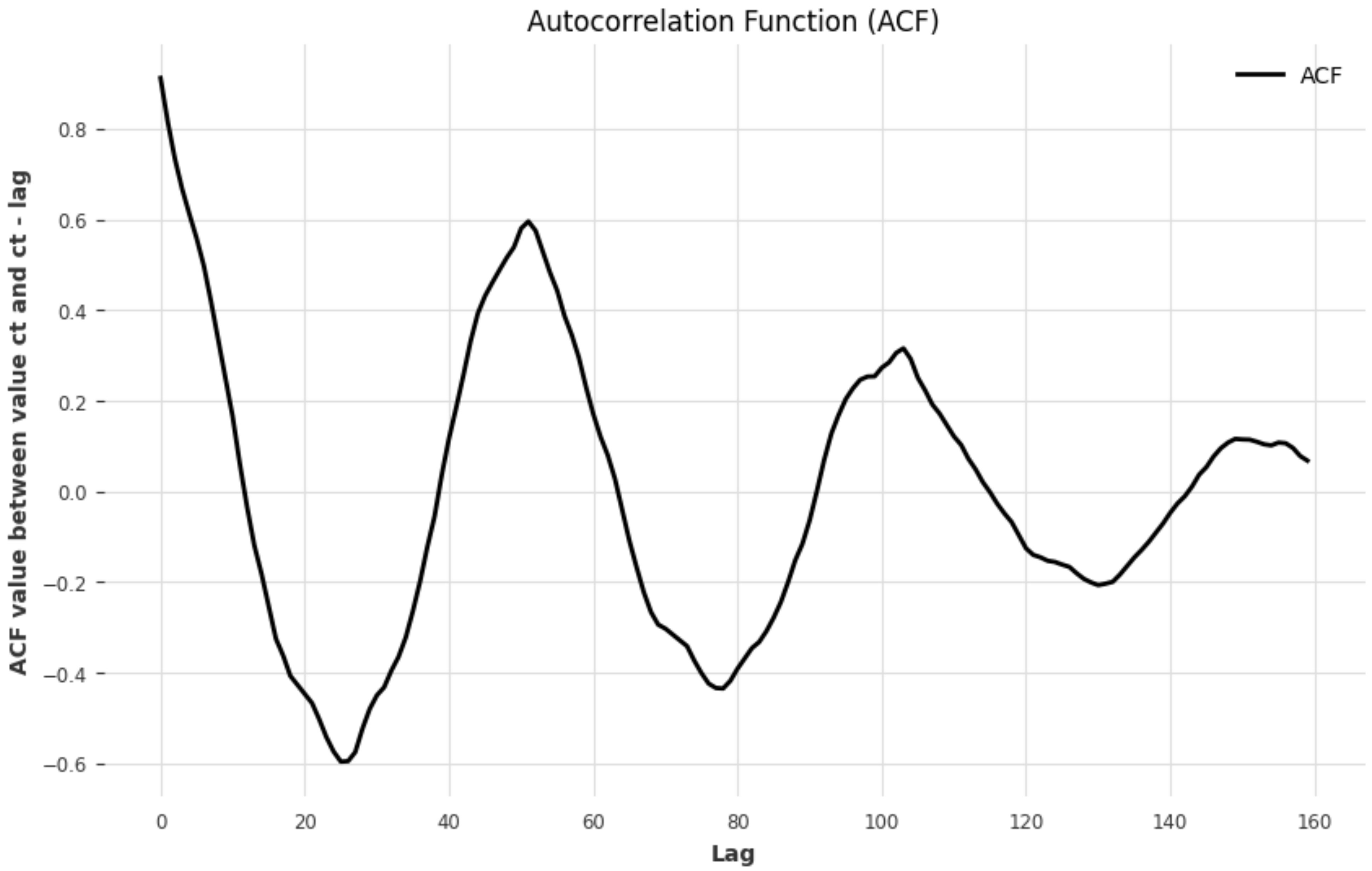
Cumulative Periodogram vs Flat Spectrum

will be redone after removing periodicity

the frequency is very high near 1/52 weeks, suggesting 52 seasonality again, and also proving that our data is again not white noise

Determining Seasonal order (P)

Original ACF

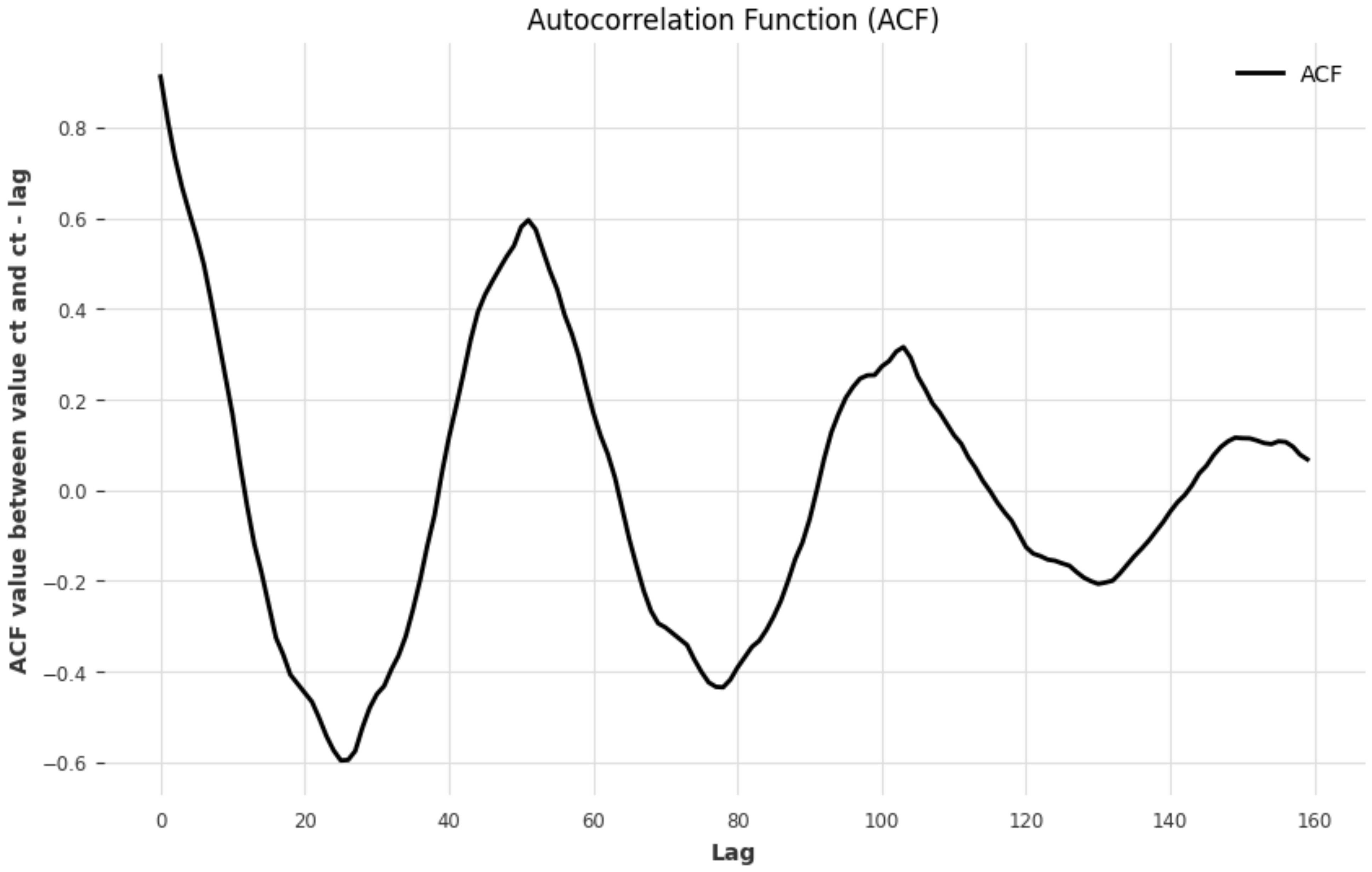


At lag 52, ACF = 0.6, the rest of the annual points are much lower and decay to 0

To determine which seasonal order to use (how many seasonal yearly lags to use I’ve set a threshhold of Autocorrelation of minimum 0.5

===> P = 1

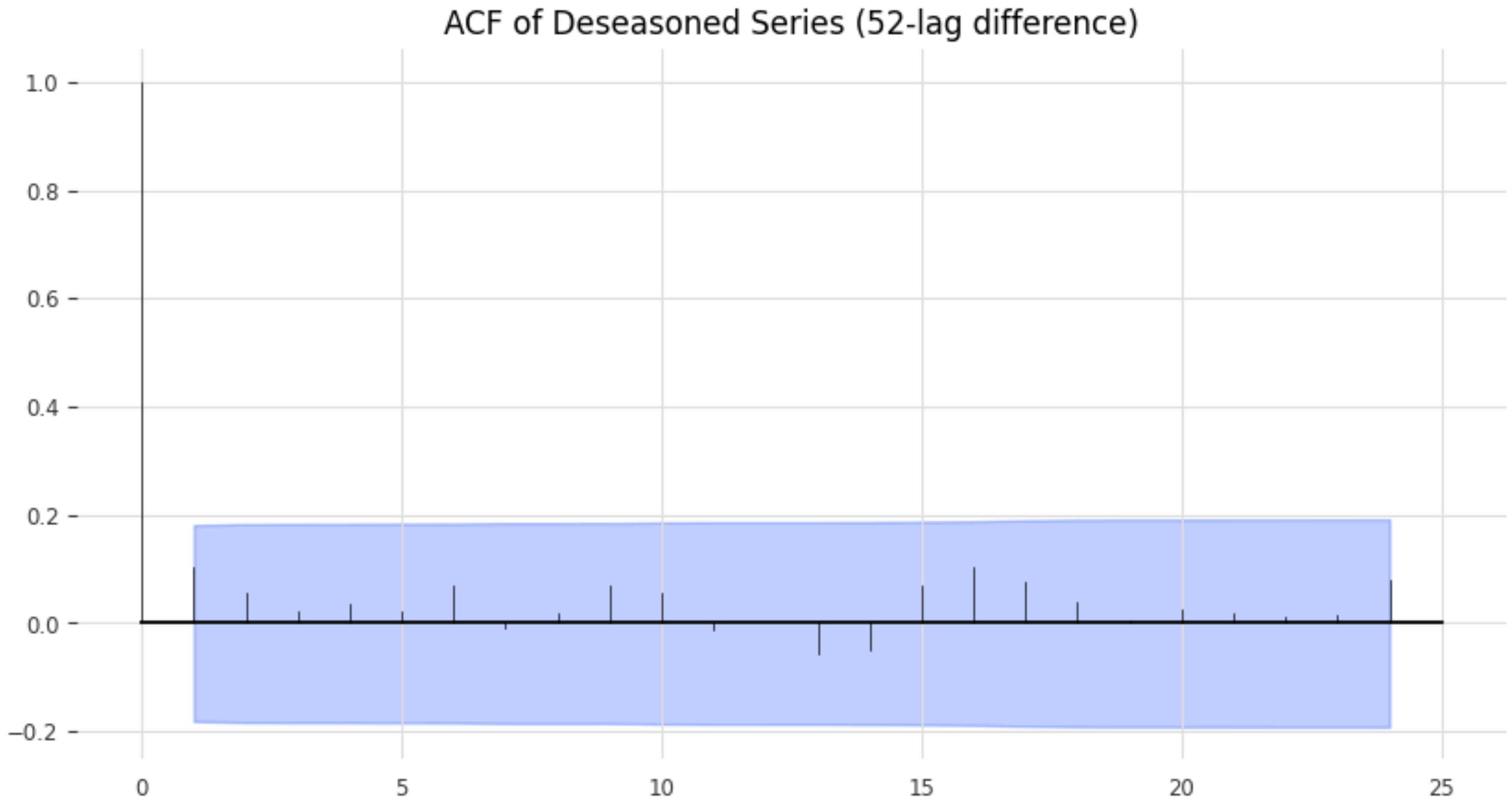
Testing for stationnarity and D order ACF of deseasoned data



The seasonal component of the original series was already stationary, as the ACF showed significant but decaying seasonal spikes at lags 52, 104, and 156. This confirmed that seasonal differencing was not necessary to achieve stationarity.

D = 0

Testing for MA order (Q)
ACF of deseasoned data



The ACF values of the deseasoned ts show no significant cutout after a certain lag, this does not suggest a ARMA model

Q = 0

Interpretation

Result

SARIMA (1,0,0 [1,0,0,52])

- Therefore, the SARIMA (1,0,0 [1,0,0,52]) model effectively captures both short-term and annual dependencies in the energy consumption data with minimal complexity.

III. Models

Model Formulas

AR(1):

$$Y_t - \mu = \alpha(Y_{t-1} - \mu) + \varepsilon_t$$

ARMA(1,1):

$$Y_t = c + \phi_1 Y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

ARIMA(1,1,0):

$$\Delta Y_t = c + \phi_1 \Delta Y_{t-1} + \varepsilon_t$$

SARIMA(1,0,0)(1,0,0,52):

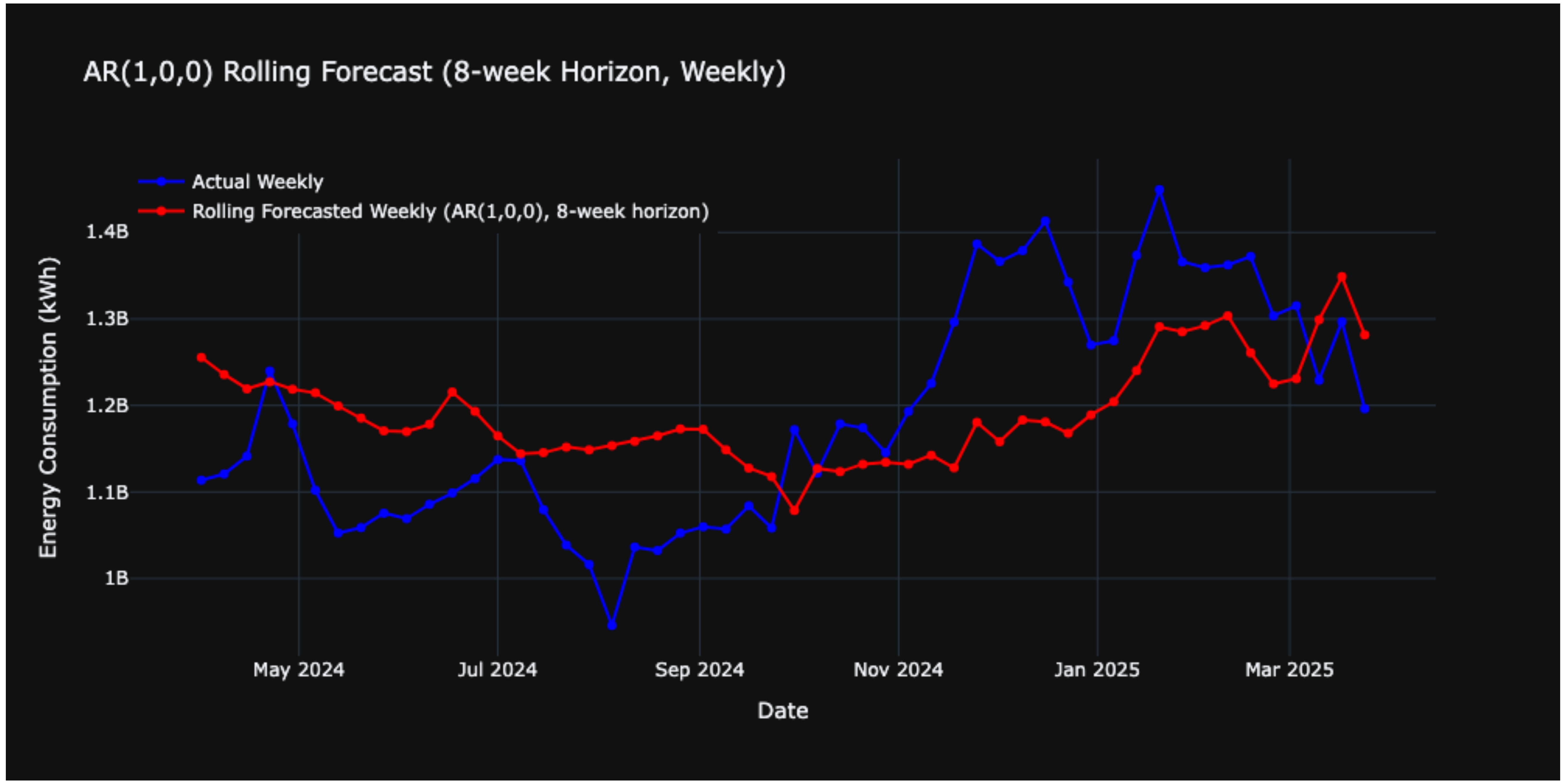
$$Y_t = c + \phi_1 Y_{t-1} + \Phi_1 Y_{t-52} + \varepsilon_t$$

SARIMAX(1,0,0)(1,0,0,52) + Temp:

$$Y_t = c + \phi_1 Y_{t-1} + \Phi_1 Y_{t-52} + \beta X_t + \varepsilon_t$$

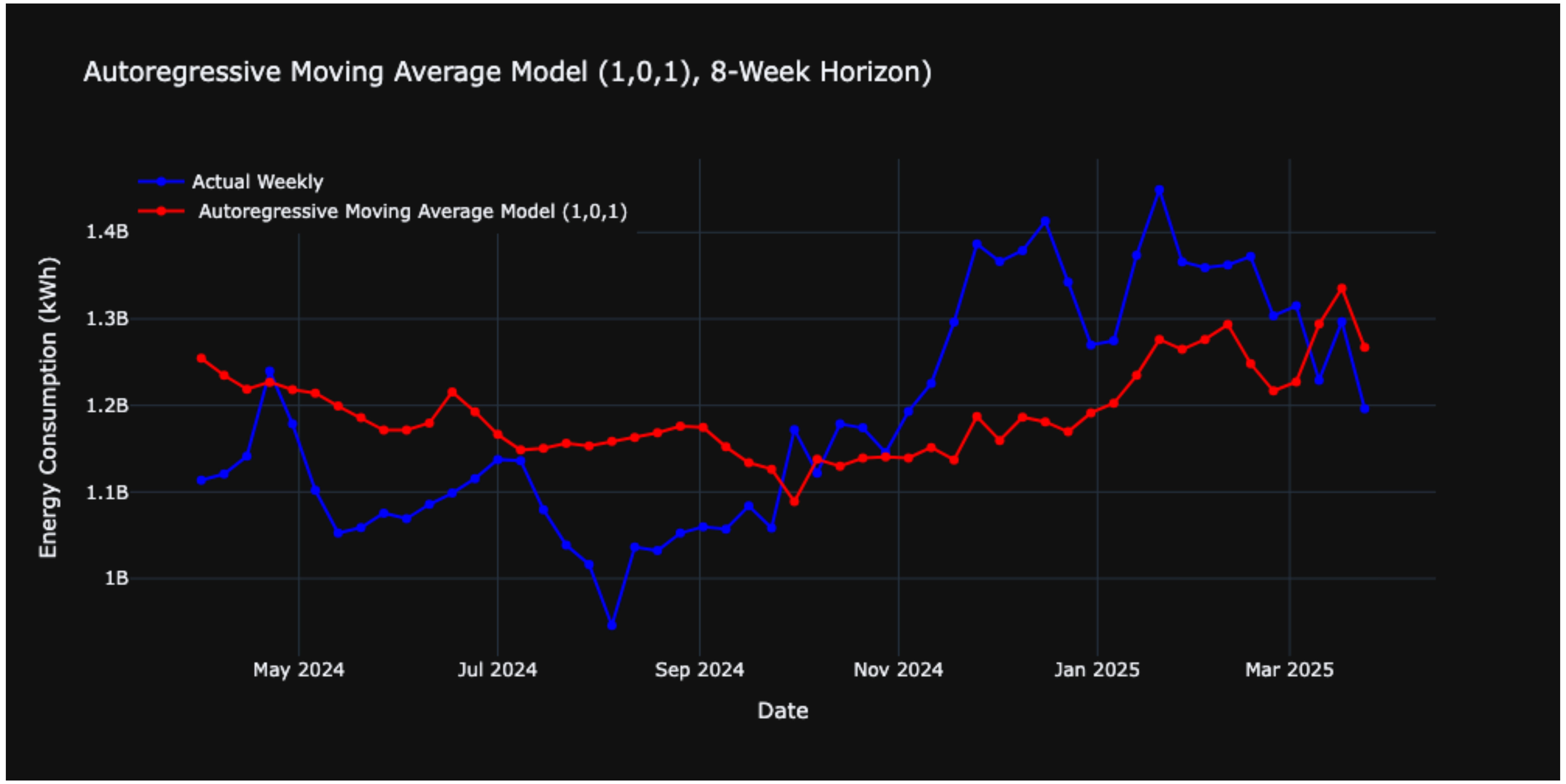
AR

Weekly sum



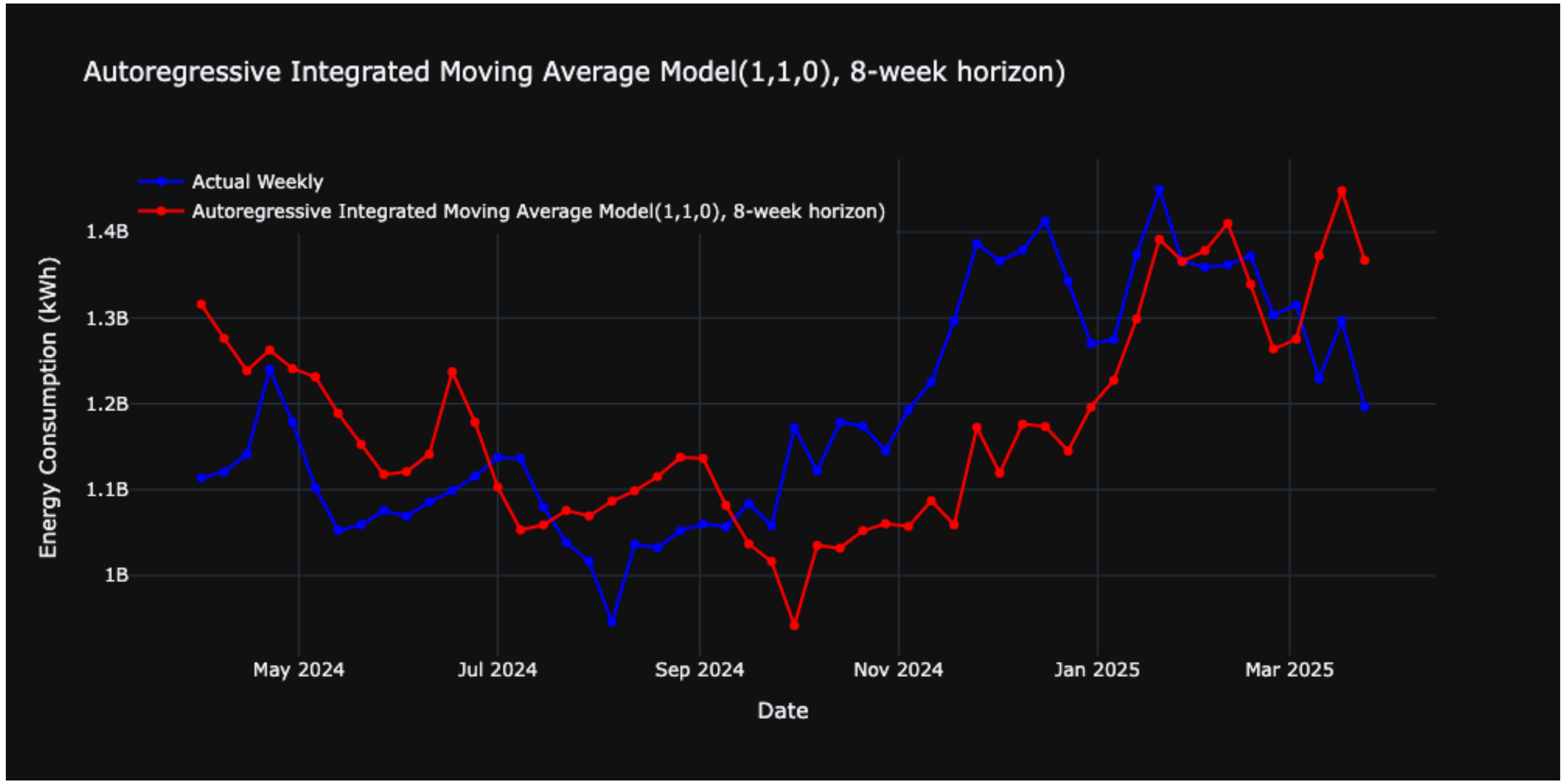
ARMA

Weekly sum



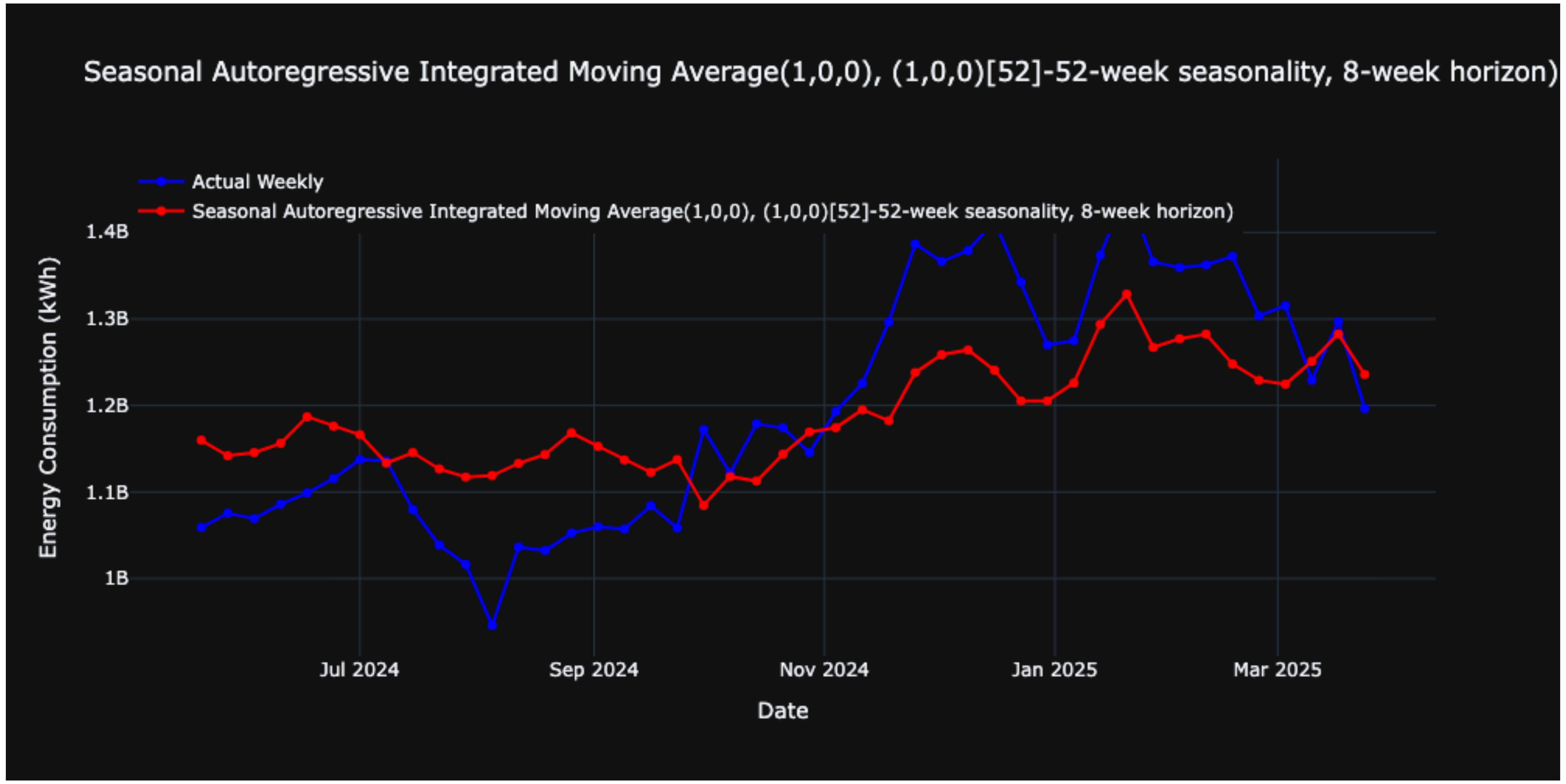
ARIMA

Weekly sum



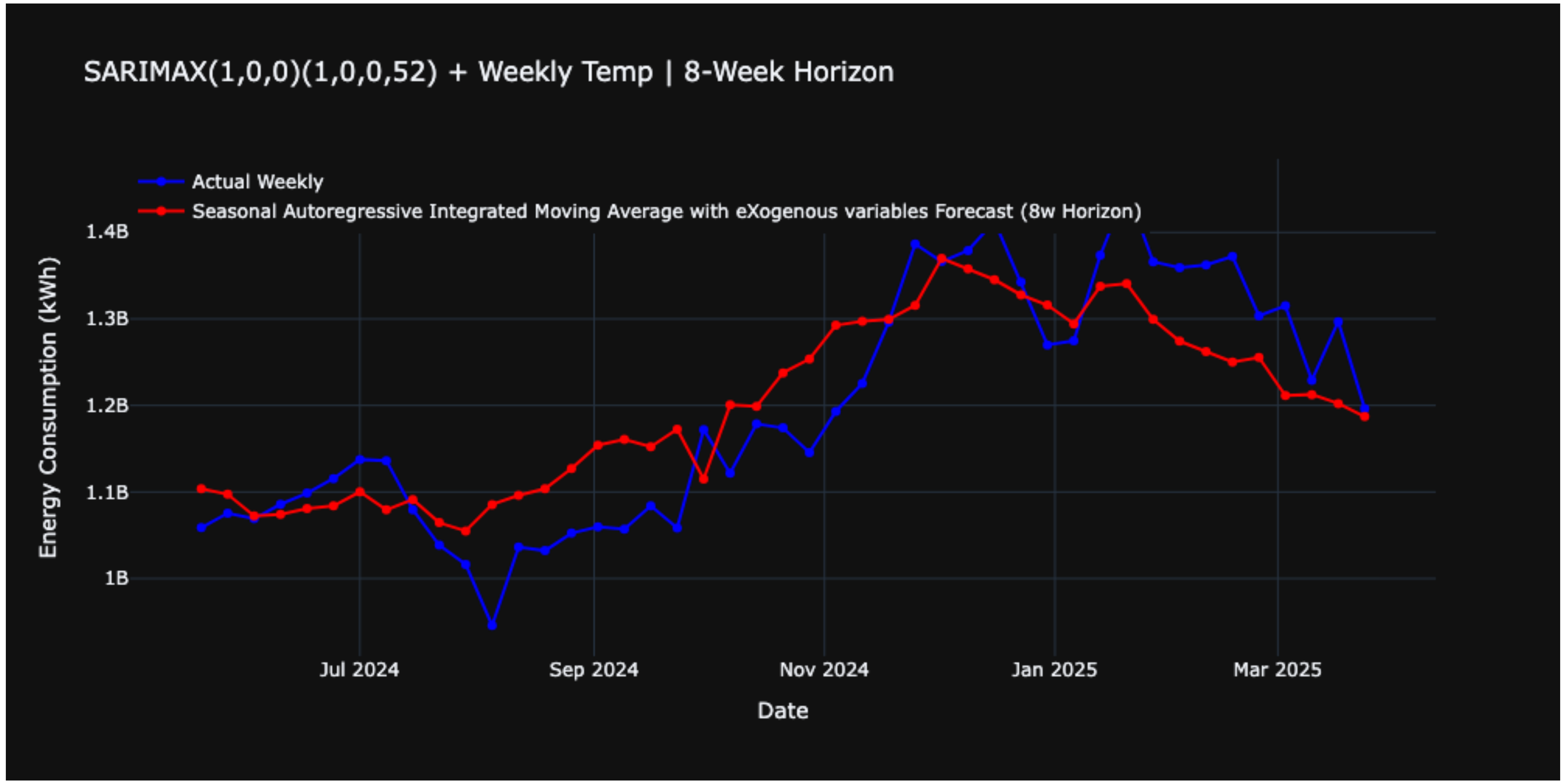
SARIMA

Weekly sum

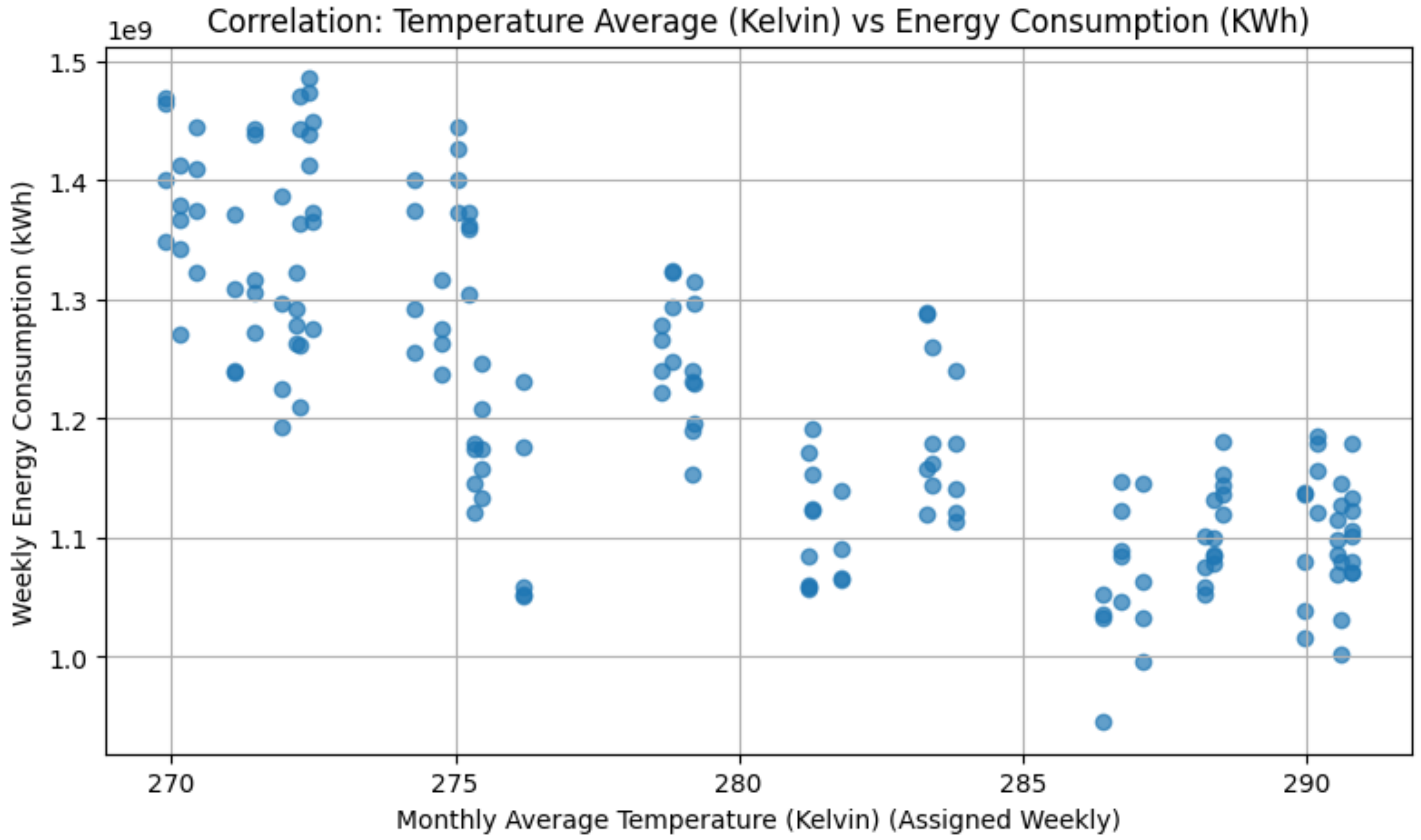


SARIMAX

Weekly sum



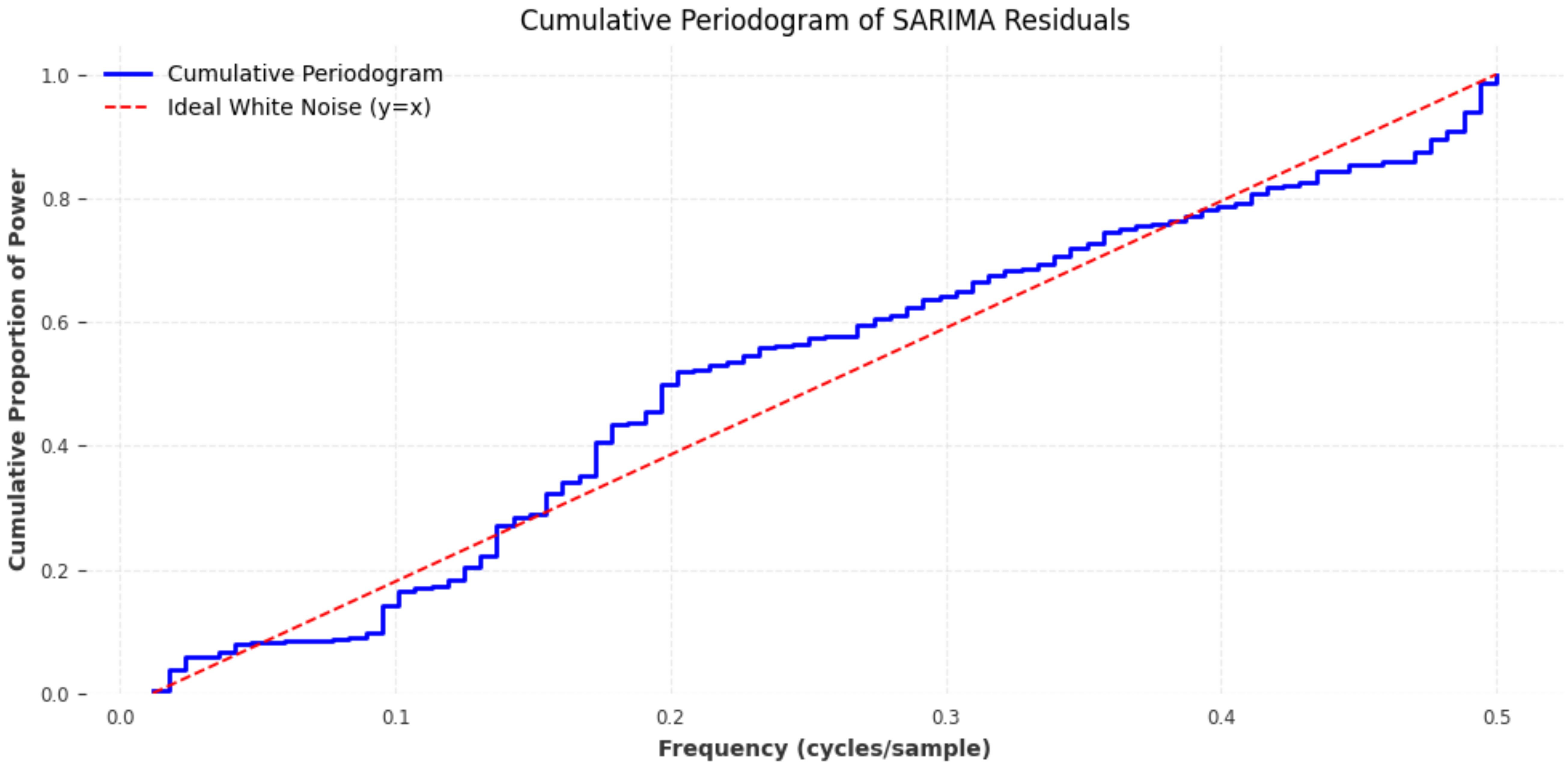
Exogenerous Variables Correlation with Weather Data



p = -0.77

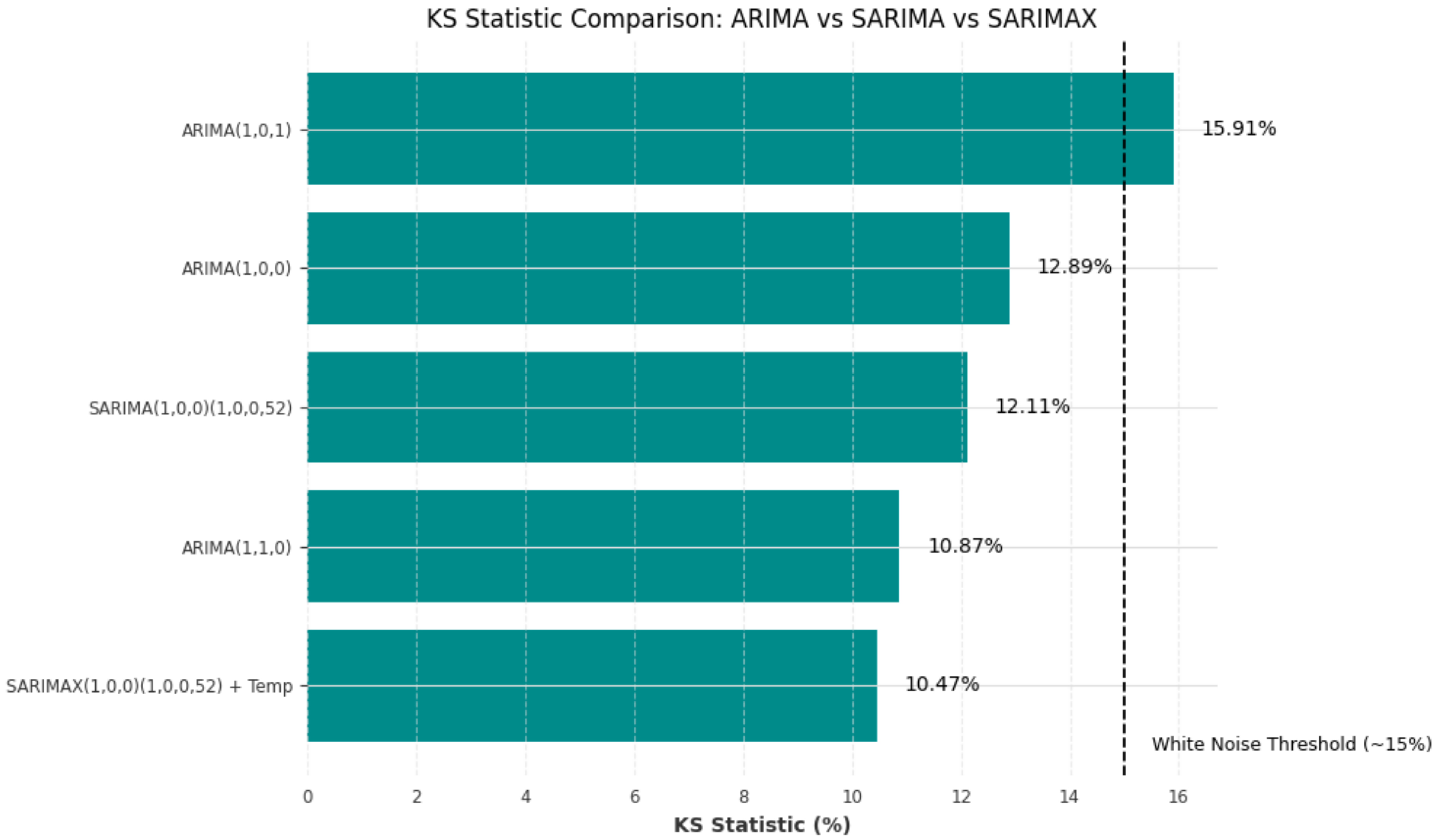
Testing model residuals for white noise

SARIMA (1,0,0,(1,0,0,[52]))



Testing model residuals for white noise

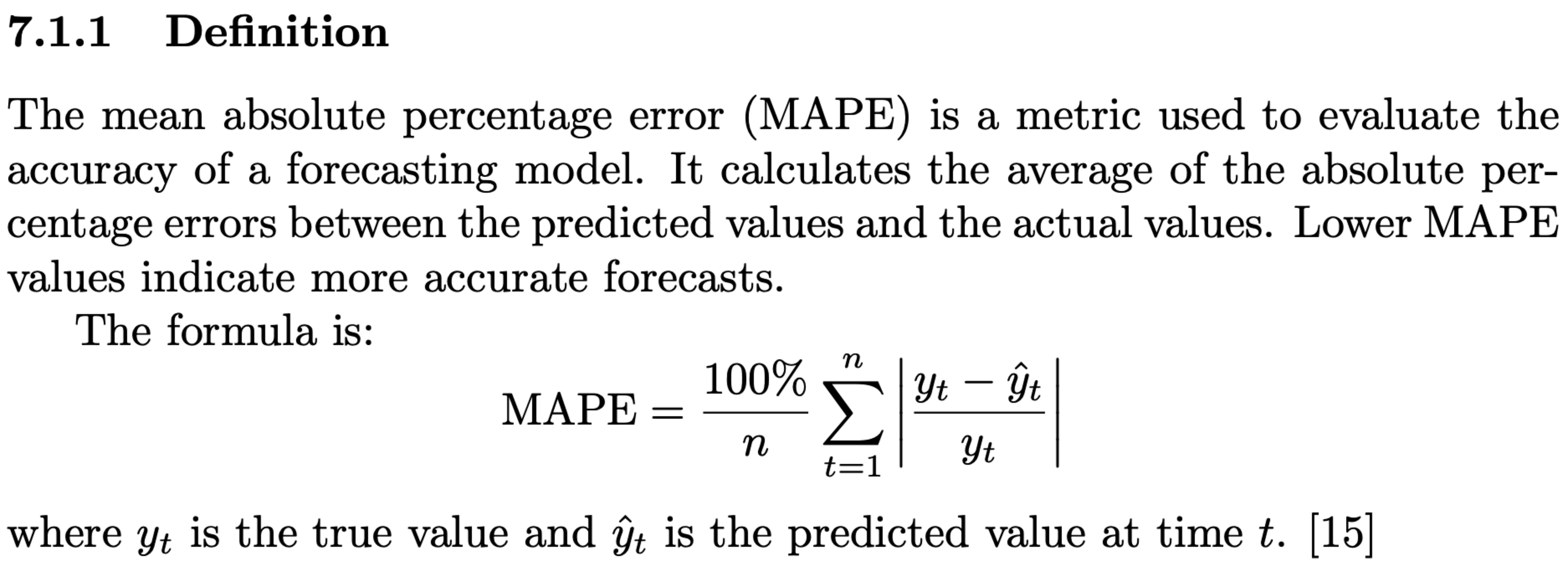
All models comparison



IV. Evaluation

MAPE

Evaluation Metric



7.1.1 Definition

The mean absolute percentage error (MAPE) is a metric used to evaluate the accuracy of a forecasting model. It calculates the average of the absolute percentage errors between the predicted values and the actual values. Lower MAPE values indicate more accurate forecasts.

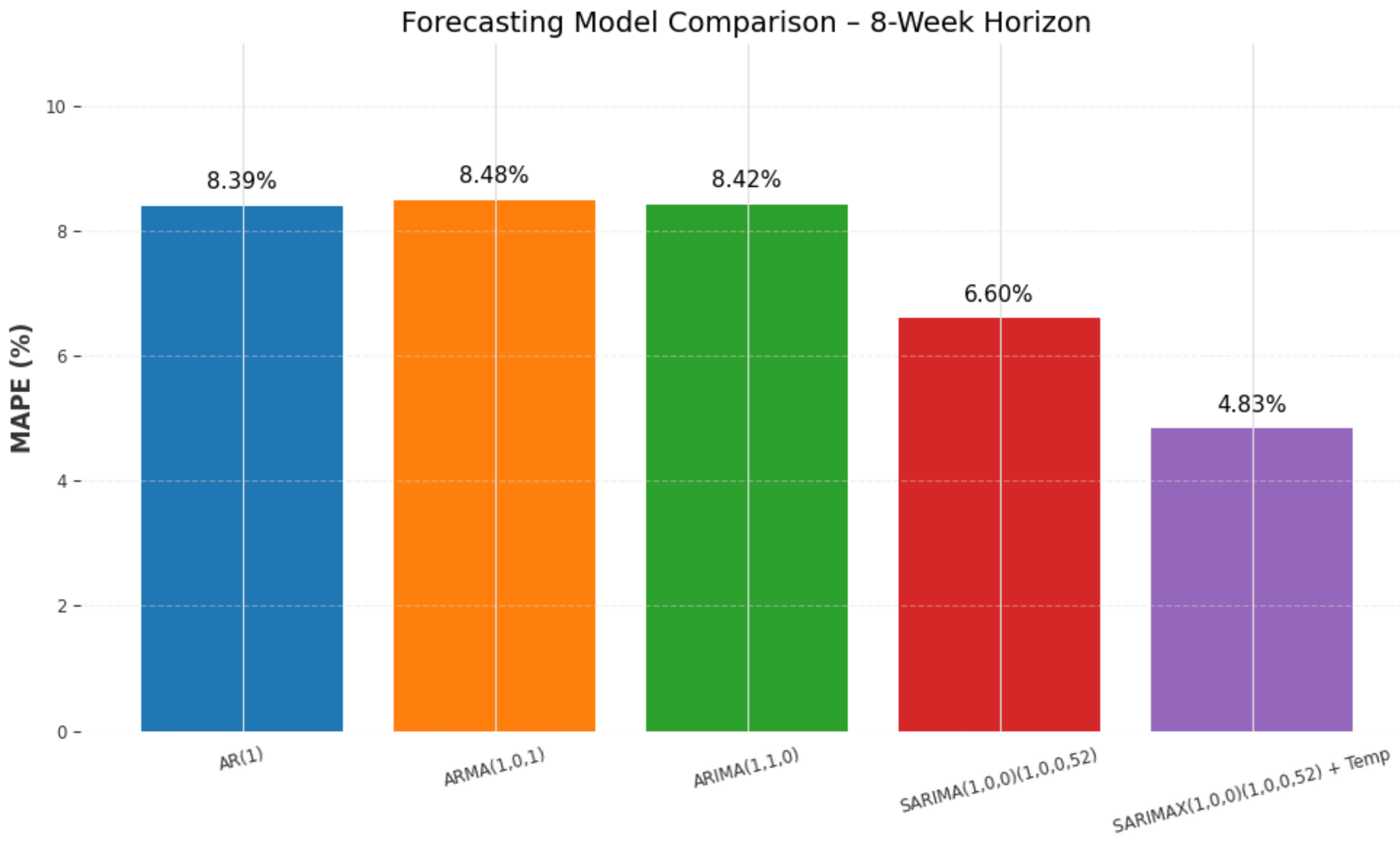
The formula is:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

where y_t is the true value and $\hat{y}_t$ is the predicted value at time t. [15]

MAPE

Evaluation Metric


Forecasting Model Comparison – 8-Week Horizon

We concluded that the SARIMAX (1,0,0,[52]) with Temp is best

V. Challenges and discussion

Challenges

- I. Understanding Time Series Analysis and what each formula represents
- II. Finding conflicting explanations online
- III. Determining AR weights using Gradient descent instead of Yule-Walker Equation
- IV. Understanding academic analysis: doing research on other papers, reports using LaTeX, etc.
- V. Backing up my approaches and results using scientific methods

Thanks for coming !