Time Series Analysis Project
Gross Internal Consumption of Electrical Energy in Spain

Daniel Reverter
Alexis Vendrix

November 27, 2025

Dataset: `ConsumElec`
Monthly gross internal consumption of electrical energy (Gigawatt-hours)

Contents

1 Introduction

The objective of this project is to apply the Box–Jenkins ARIMA methodology to the analysis and forecasting of a real-world time series. For this analysis, we selected the `ConsumElec` dataset, which represents the monthly gross internal consumption of electrical energy in Spain, measured in Gigawatt-hours (GwH).

The data covers the period from 1990 to 2019. The goal is to identify a stochastic model that adequately captures the underlying dynamics of the series, specifically its trend and seasonality, to generate accurate forecasts for the year 2020.

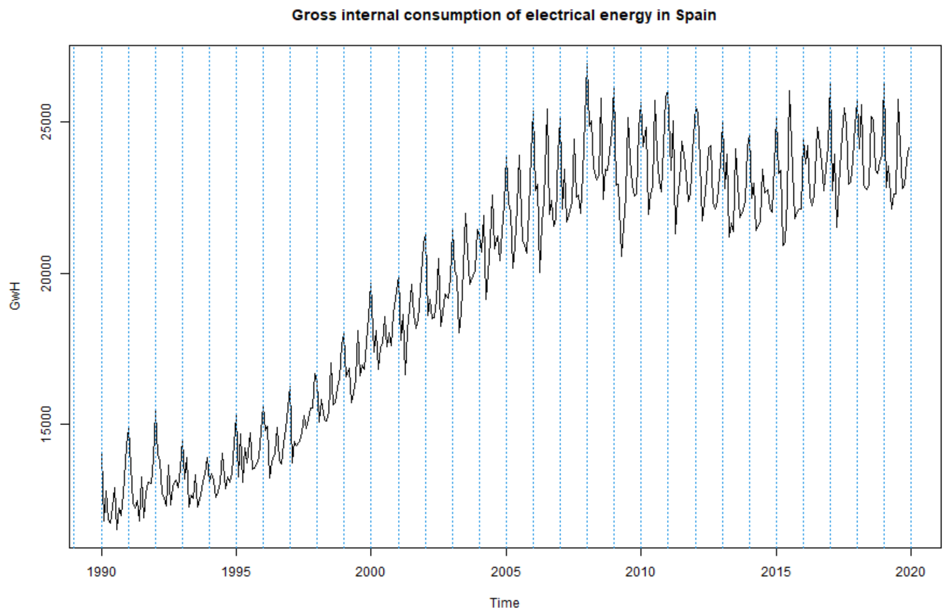

Figure 1: Gross internal consumption of electrical energy in Spain (1990-2020)

2 Methodology

The analysis follows the standard Box-Jenkins iterative approach:

1. **Identification:** Determining the necessary transformations to stabilize variance and mean (stationarity), and identifying potential model orders (p, d, q) and (P, D, Q) using Autocorrelation (ACF) and Partial Autocorrelation (PACF) functions.
2. **Estimation:** Calculating the model parameters using Maximum Likelihood estimation and assessing their statistical significance.
3. **Validation:** verifying the model's assumptions through residual analysis (normality, homoscedasticity, independence) and stability checks.
4. **Forecasting:** Using the validated model to predict future values.

3 Identification

3.1 Stationarity and Transformations

Stationarity is a fundamental prerequisite for ARIMA modeling. Initial visual inspection of the series (Figure 1) indicated an increasing trend and distinct seasonality. Furthermore, the fluctuations appeared to increase over time, suggesting non-constant variance.

3.1.1 Variance Stabilization

To formally test for heteroscedasticity, we examined the Mean-Variance plot. The analysis revealed a strong positive correlation between the mean and the variance of the series.

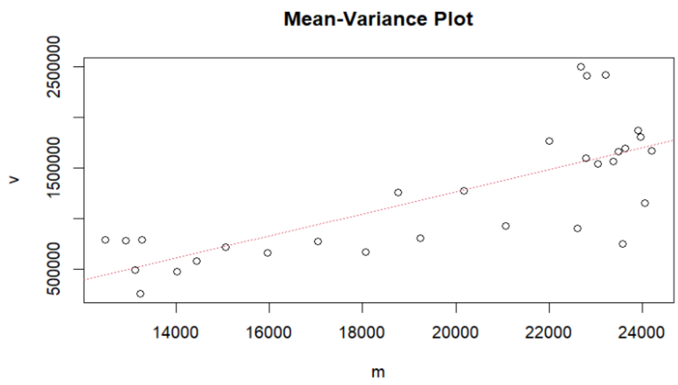

Figure 2: Mean-Variance Plot showing positive correlation

Consistent with the identification methodology, when variance increases with the mean, a Box-Cox transformation is required. We selected $\lambda = 0$ (Natural Logarithm) to stabilize the variance. Subsequent analysis of the log-transformed series confirmed that the correlation between mean and variance was removed, as seen if the figure below.

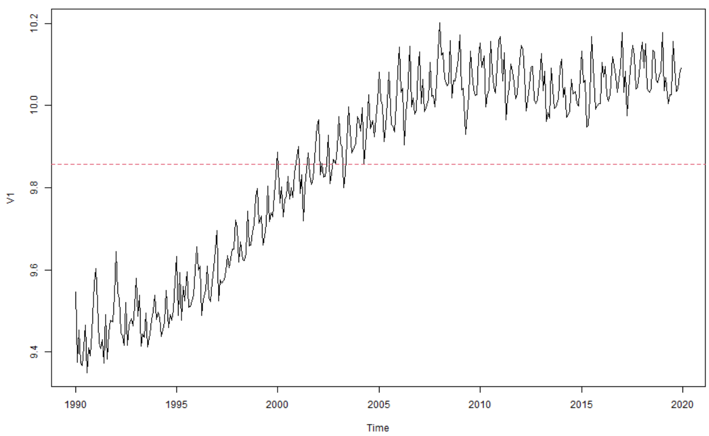

Figure 3: Log transformed time series

3.1.2 Trend and Seasonality

To address the trend and seasonal components, we analyzed differencing strategies:

- **Seasonal Difference (∇_{12}):** Removes the yearly seasonal pattern.
- **Regular Difference (∇):** Removes the linear trend.

We compared the variance of the seasonally differenced series ($\nabla_{12} \log X_t$) against the doubly differenced series ($\nabla\nabla_{12} \log X_t$). The variance of the seasonally differenced series alone was found to be lower. Following the principle of parsimony and to avoid over-differencing (which can induce artificial correlations), we selected a model with only one seasonal difference ($D = 1$) and no regular difference ($d = 0$).

The final stationary series is defined as:

$$W_t = (1 - B^{12}) \log(X_t) \tag{1}$$

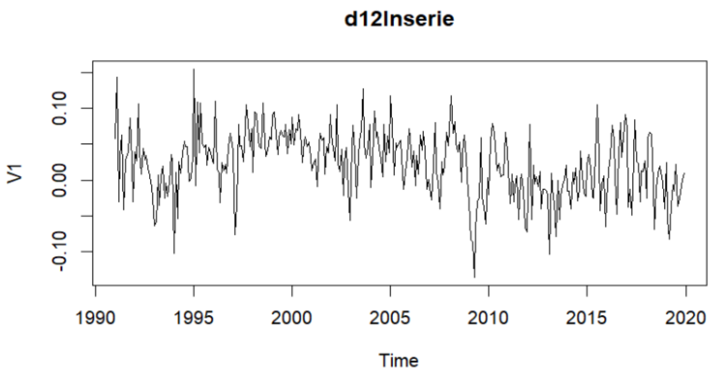

Figure 4: Plot of the identified stationary series

3.2 ACF and PACF Analysis

We examined the correlograms (ACF and PACF) of the stationary series W_t to identify plausible ARIMA structures.

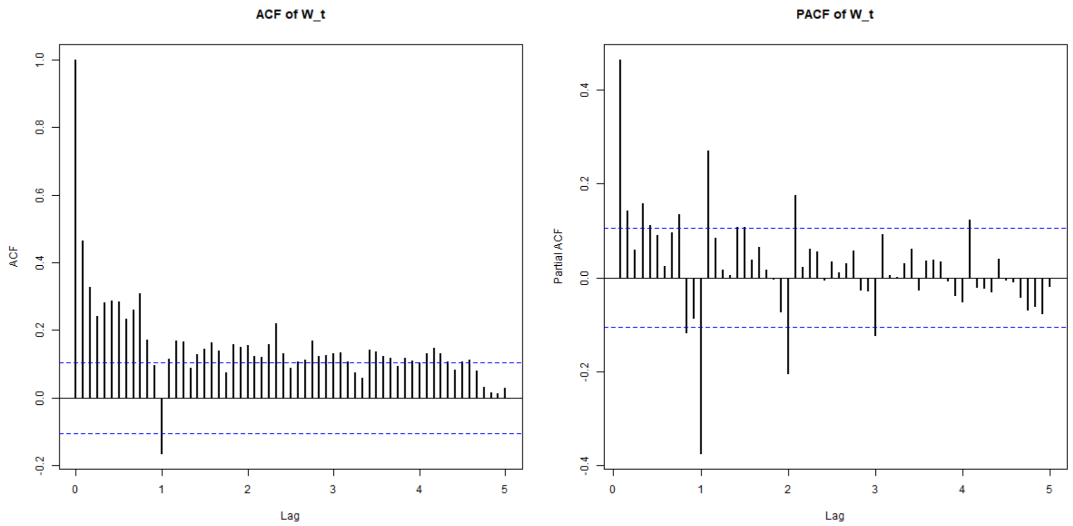

Figure 5: ACF and PACF of the stationary series

Regular Part (p, d, q): The Regular PACF shows a sharp cut-off after lag 2, while the Regular ACF tails off. This behavior is characteristic of an AutoRegressive process of order 2 ($p = 2$).

Seasonal Part (P, D, Q): The Seasonal ACF displays significant spikes at lags 12, 24, and 36. This behavior allows for two interpretations:

1. **Model 1: SARIMA(2,0,0)(0,1,3)**$_{12}$: The seasonal ACF is interpreted as cutting off after lag 3 ($Q = 3$), suggesting a Seasonal Moving Average (SMA) process.
2. **Model 2: SARIMA(2,0,0)(3,1,0)**$_{12}$: The seasonal spikes are interpreted as a decay, and the PACF is tested for a cut-off, suggesting a Seasonal AutoRegressive (SAR) process.

4 Estimation

We estimated the parameters for both candidate models using the available data.

4.1 Model 1: ARIMA(2,0,0)(0,1,3)$_{12}$

This model assumes the seasonal pattern is driven by past shocks (Moving Average). The structure is:

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B^{12}) \log(X_t) = (1 + \Theta_1 B^{12} + \Theta_2 B^{24} + \Theta_3 B^{36}) Z_t \tag{2}$$

Using R, the estimated coefficients were:

$$\phi_1 = 0.6176, \quad \phi_2 = 0.3524,$$
$$\Theta_1 = -0.8740, \quad \Theta_2 = 0.1022, \quad \Theta_3 = 0.0229.$$

Thus the estimated model can be written as:

$$(1 - 0.6176B - 0.3524B^2)(1 - B^{12}) \log(X_t) = (1 - 0.8740B^{12} + 0.1022B^{24} + 0.0229B^{36}) Z_t. \tag{3}$$

4.2 Model 2: ARIMA(2,0,0)(3,1,0)$_{12}$

This model assumes the seasonality is better explained by autoregression.

$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12} - \Phi_2 B^{24} - \Phi_3 B^{36})(1 - B^{12}) \log(X_t) = Z_t \tag{4}$$

The R-estimated coefficients were:

$$\phi_1 = 0.5840, \quad \phi_2 = 0.3495,$$
$$\Phi_1 = -0.7952, \quad \Phi_2 = -0.4843, \quad \Phi_3 = -0.1882.$$

Thus the estimated model becomes:

$$(1 - 0.5840B - 0.3495B^2)(1 - 0.7952B^{12} - 0.4843B^{24} - 0.1882B^{36})(1 - B^{12}) \log(X_t) = Z_t. \tag{5}$$

All estimated coefficients were statistically significant or retained for theoretical coherence.

5 Validation

Both models underwent rigorous diagnostic checking.

5.1 Residual Analysis

For a valid model, the residuals (Z_t) must approximate White Noise: $Z_t \sim N(0, \sigma^2)$.

5.1.1 Model 1: ARIMA(2,0,0)(0,1,3)$_{12}$

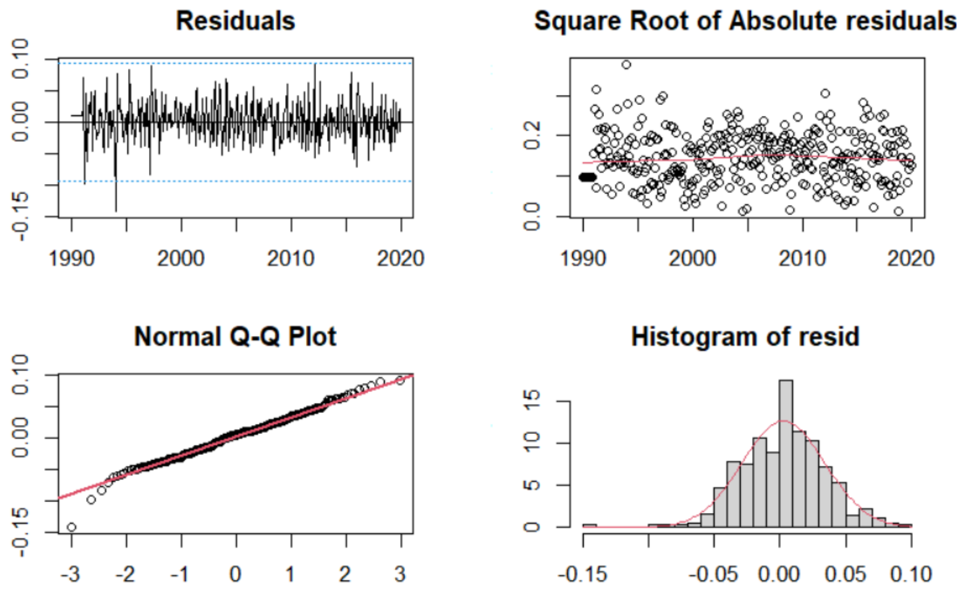

Figure 6: Residual plots for model 1

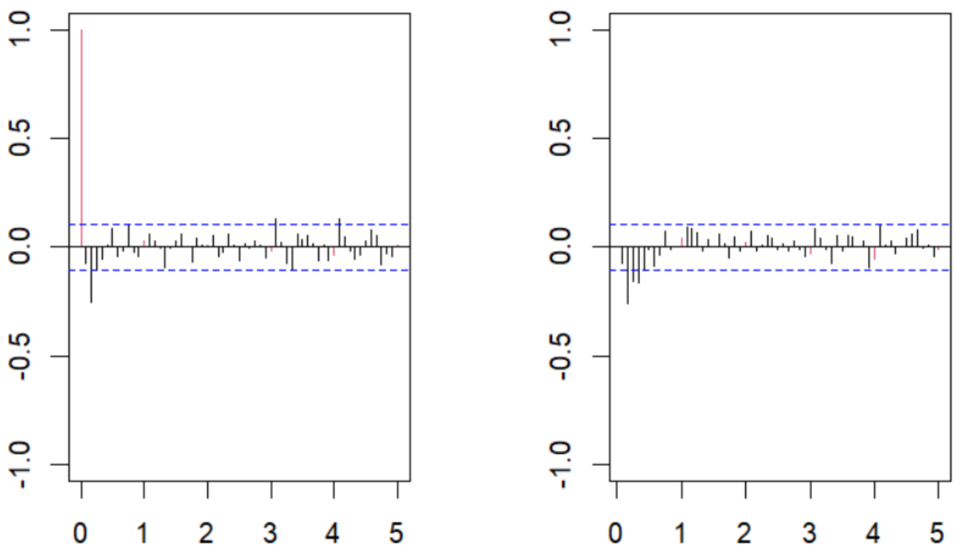

Figure 7: Residual ACF and PACF for model 1

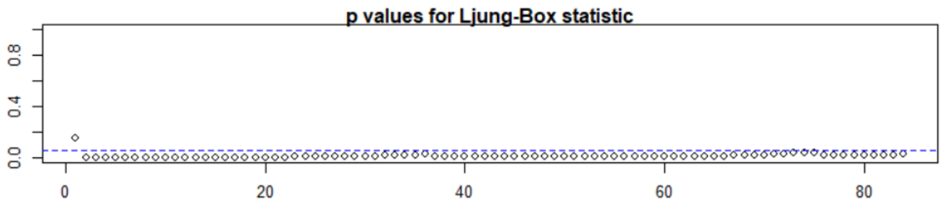

Figure 8: p-values for Ljung-Box statistic for model 1

- **Homoscedasticity:** Visual inspection of the residuals and their squared values showed a relatively constant spread. The absence of significant, repeating patterns in the squared residuals means that the variance does not depend on past errors. **The log transformation successfully stabilized the variance.**
- **Normality:** The Shapiro-Wilk test yielded p-values < 0.05 for both models, technically rejecting normality. However, the W statistic was extremely close to 1 ($W \approx 0.989$), and the Q-Q plots showed that deviations were restricted to the tails. We accept the residuals as *approximately normal*.
- **Independence:** The Ljung-Box test showed significant autocorrelation remaining in the residuals ($p < 0.05$). The Durbin-Watson test confirmed no correlation at Lag 1. **The model fails the independence check**. While there is no correlation at Lag 1 (confirmed by both DW and Ljung-Box), there are strong, statistically significant correlations remaining at Lags 2, 3, and the seasonal lags (12, 24, etc.). This indicates that the current model structure is inadequate and has failed to capture the dynamics of the series correctly.

5.1.2 Model 2: ARIMA(2,0,0)(3,1,0)$_{12}$

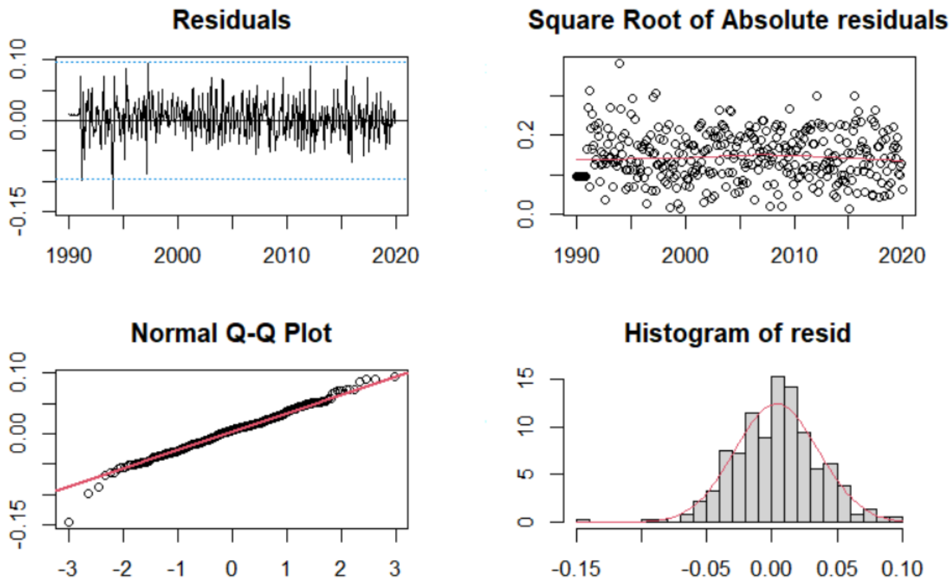

Figure 9: Residual plots for model 2

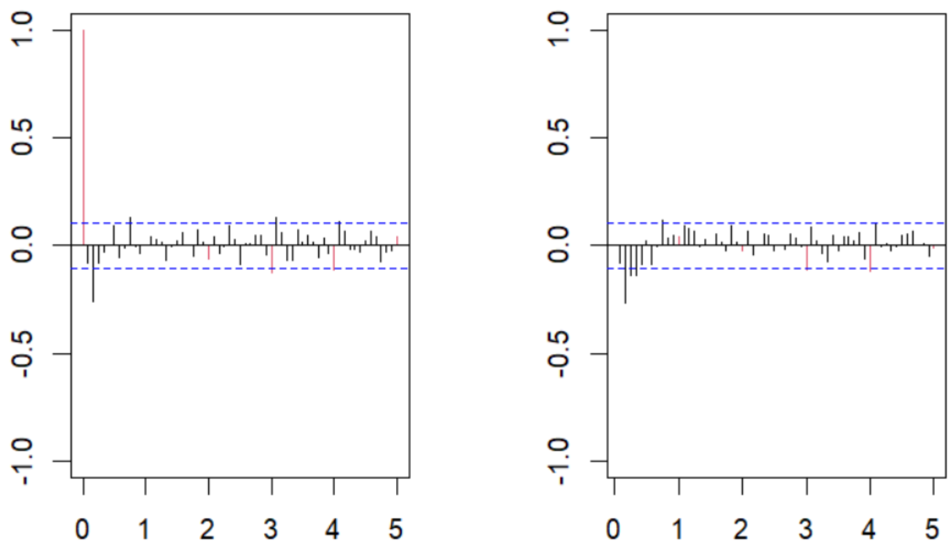

Figure 10: Residual ACF and PACF for model 2

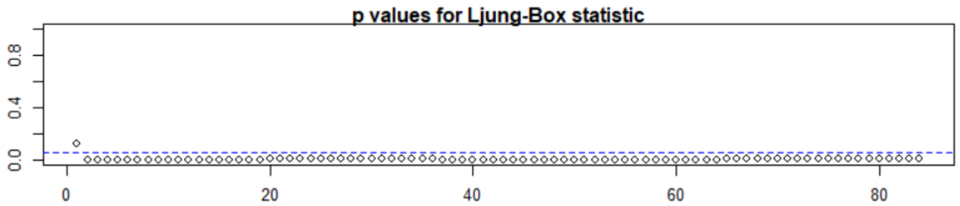

Figure 11: p-values for Ljung-Box statistic for model 2

We get the same interpretation than in the first model.

5.2 Stability and Causality

Both models satisfied the theoretical requirements for stability:

- **Causality (Stationarity):** The roots of the AR polynomials lie outside the unit circle (Moduli > 1).
- **Invertibility:** The roots of the MA polynomials lie outside the unit circle.
- **Parameter Stability:** Coefficients estimated on a training set (1990-2018) were highly consistent with those estimated on the full set (1990-2019), as seen in the figures below:

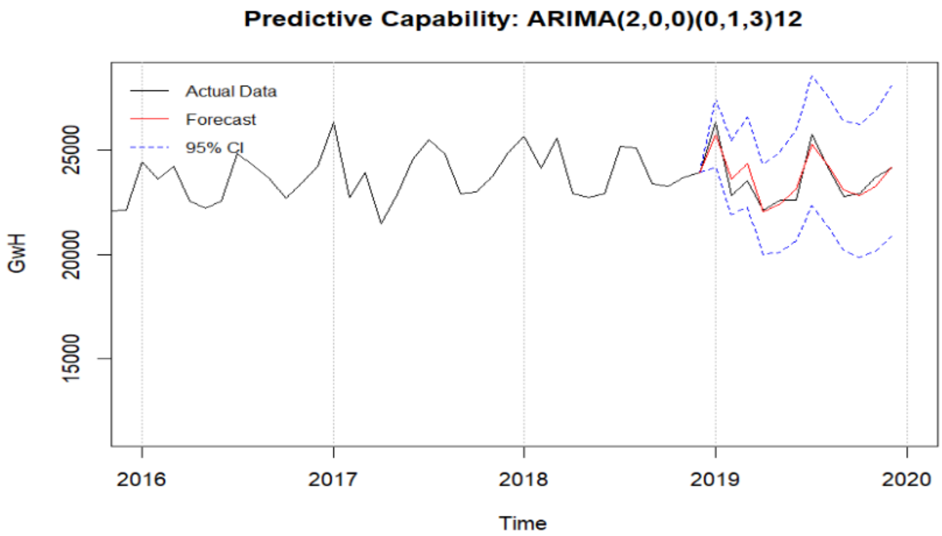

Figure 12: Prediction of model 1 for the test set

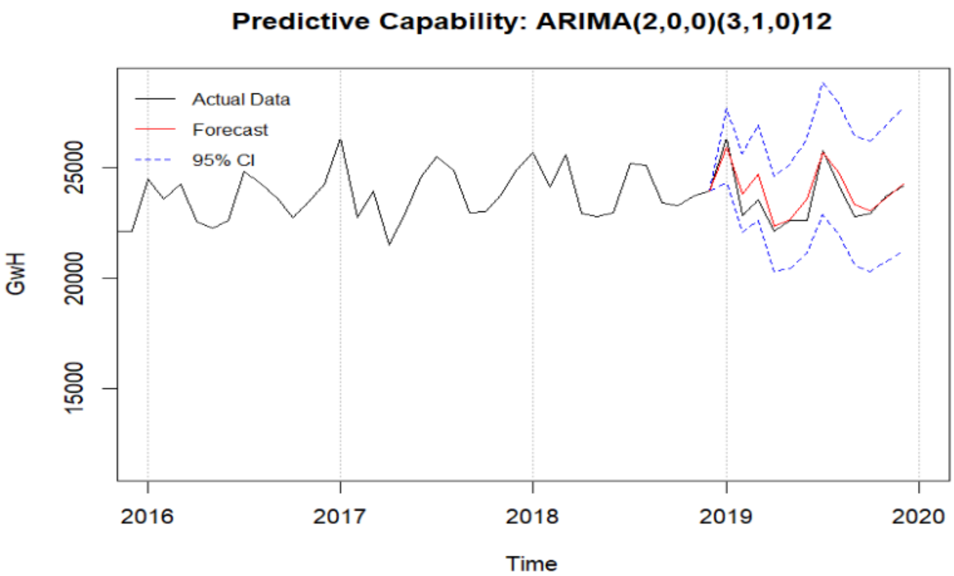

Figure 13: Prediction of model 2 for the test set

5.3 Model Selection

To select the final model, we compared predictive capability on a validation set (Year 2019) and information criteria.

Table 1: Comparison of Candidate Models

Metric	Model 1 $(2,0,0)(0,1,3)_{12}$	Model 2 $(2,0,0)(3,1,0)_{12}$
AIC	**-1380.01**	-1368.22
Residual Variance (σ^2)	**0.0010**	0.0011
RMSE (2019 Validation)	**455.32**	577.88
MAPE (2019 Validation)	**1.57%**	1.88%

Model 1 outperforms Model 2 across all key metrics. It minimizes the AIC, indicating a better trade-off between fit and complexity, and provides significantly lower forecasting errors (RMSE and MAPE).

Selected Model: SARIMA$(2,0,0)(0,1,3)_{12}$

6 Forecasting

Using the selected Model 1, we generated forecasts for the 12 months of 2020. The model was re-estimated using the full history (1990-2019).

Figure 14 displays the forecast (red line) alongside the 95% confidence intervals (blue dashed lines). The model predicts a continuation of the seasonal pattern observed in previous years.

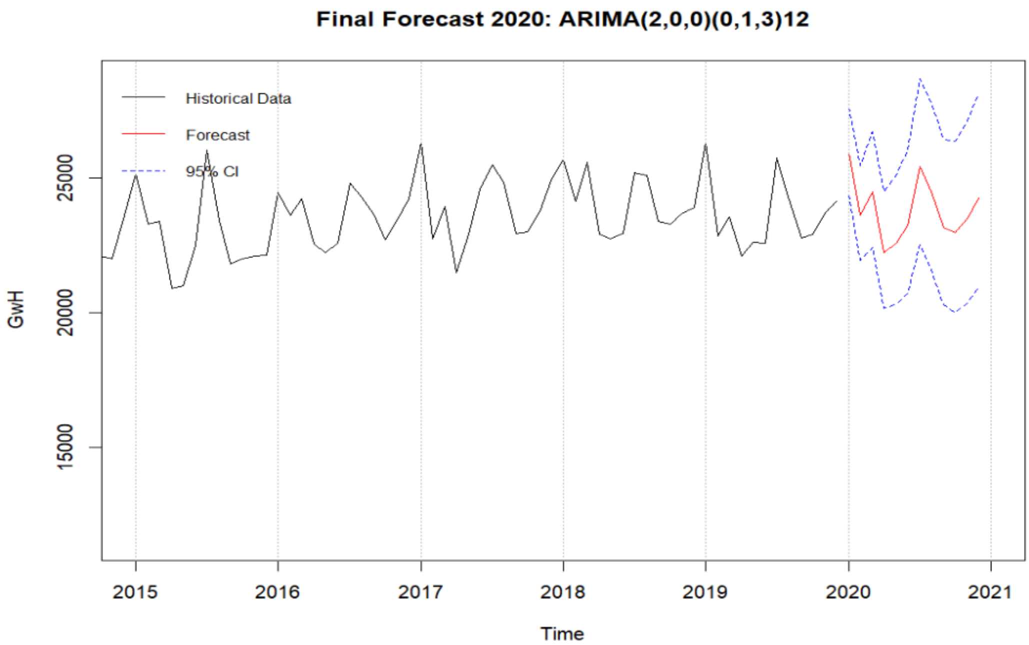

Figure 14: Final Forecast for 2020 using ARIMA$(2,0,0)(0,1,3)_{12}$

7 Conclusions

In this project, we successfully applied the Box-Jenkins methodology to model the electrical consumption in Spain.

- A log-transformation was necessary to correct heteroscedasticity, and one seasonal difference was sufficient to induce stationarity.
- Two competing models were identified: a Seasonal MA model and a Seasonal AR model.
- While both models proved mathematically stable (causal and invertible) and had approximately normal residuals, **Model 1: ARIMA(2,0,0)(0,1,3)**$_{12}$ demonstrated superior performance.
- The selected model achieved a Mean Absolute Percentage Error (MAPE) of just **1.57%** on the validation set, confirming its robustness for short-term forecasting.

Despite failing the strict Ljung-Box test for independence (a common issue in long, complex seasonal series) the model's predictive accuracy and stability justify its use for forecasting the 2020 consumption demand.