Athens University of Economics and Business

Master of Science (MSc) in Business Analytics

Advanced Topics in Statistics - FT

-Project -

Instructors: Dimitris Karlis

Student Name: Eleni Ralli

Student ID: f2822312

June, 2024

Contents

1. Introduction

The purpose of this study is to analyze historical electricity prices in Hungary and develop a predictive model to forecast prices for the first half of 2019. Using data from 2010 to 2018, the analysis involves cleaning and preparing the data, exploring its characteristics, estimating various predictive models, and validating the chosen model. The goal is to accurately forecast monthly average prices and provide insights into whether prices will increase or not. The analysis will be performed using the R statistical package.

2. Data preparation

The steps taken to prepare the data for analysis. This includes loading the data, handling missing values with linear interpolation (Linear interpolation is a way to fill in missing values in a dataset by estimating the values between two known points), merging the electricity price data with inflation data to obtain real prices from link from the World Bank, and converting daily prices to monthly averages.

3. Descriptive analysis and exploratory data analysis

Descriptive statistics of the dataset.

The prices vary, with some periods having more ups and downs than others (see Figure 1). There are also some extreme highs or lows, which might be due to unusual events that affected the prices a lot. Comparing the actual prices to the deflated prices (which have been adjusted for inflation) helps us see the true changes in value over time more clearly.

There is a slight upward trend in the prices over time, indicating a general increase in electricity costs. Also, there is evidence of seasonality, as indicated by recurring peaks at specific intervals. The descriptive statistics support these findings, showing a mean price of 46.52 with a standard deviation of 10.46 (see Table 1), indicating moderate variability around the mean. The skewness and kurtosis values suggest a distribution with a slight positive skew and relatively normal kurtosis, implying that most price values are clustered around the mean with some extreme values.

Observations (one each month)	Mean	SD (standard deviation)	1st Qu.	Median	3rd Qu.	Trimmed	MAD	Min	Max	Range	Skew	Kurtosis	SE
102	46.52	10.46	40.45	45.24	52.19	45.88	9.1	26.45	82.99	56.54	0.7	0.85	1.04

Table 1: Descriptive statistics of the dataset for the monthly deflated data

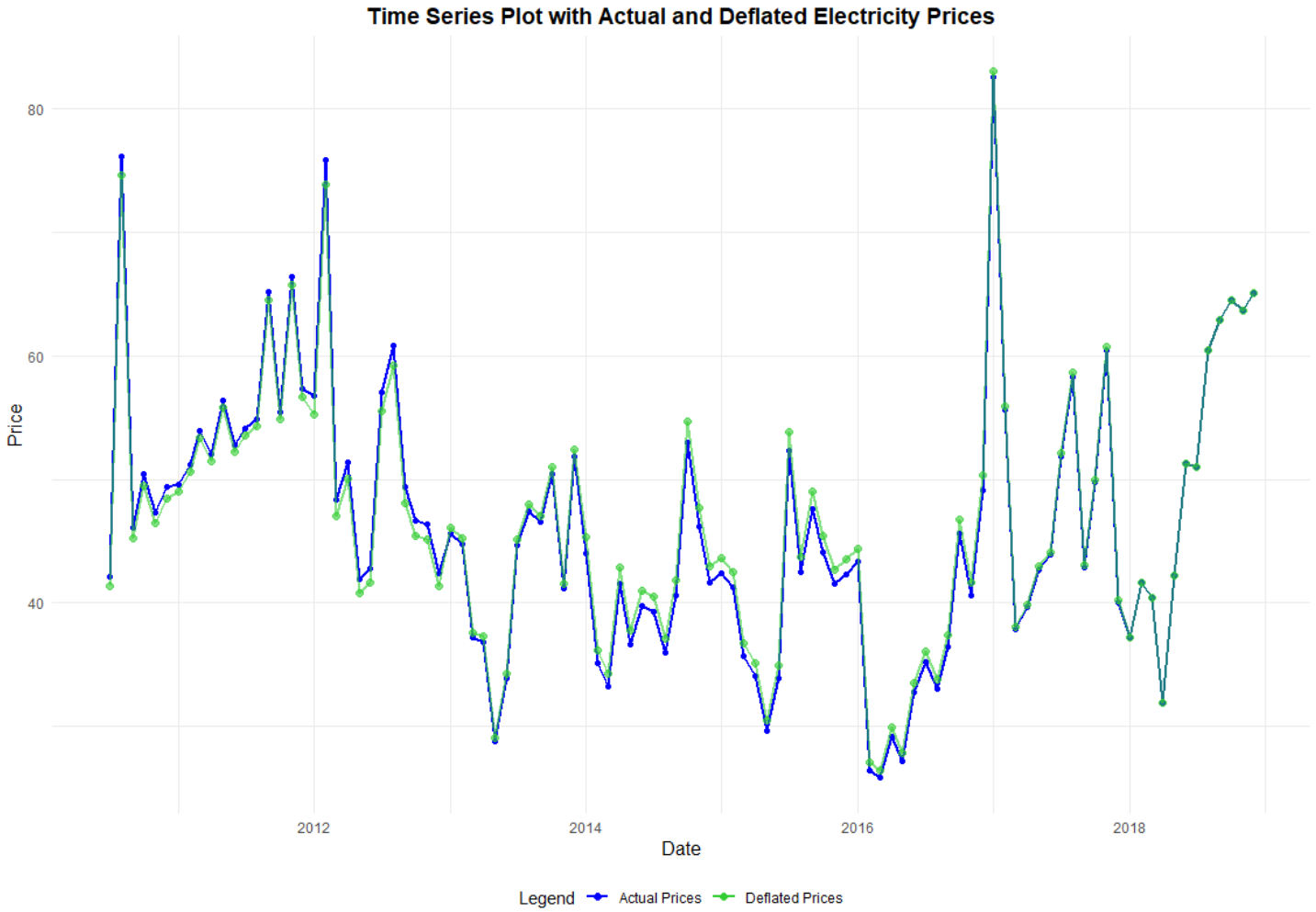

Figure 1: Time Series Plot with Actual and Deflated Electricity Prices

Visualization of the trends and seasonality in the data (Decomposition of the time series).

From the Figure 2 we see deflated prices with a trend line, shows that although prices go up and down, they generally increase over the observed period. From the Figure 3 we see the detrended deflated electricity prices, where the long-term trend has been removed. This plot highlights the variability and seasonality within the data more clearly, without the influence of the overall upward trend. The presence of high peaks and low troughs indicates periods of significant volatility, which might be associated with seasonal demand changes or external factors affecting the market.

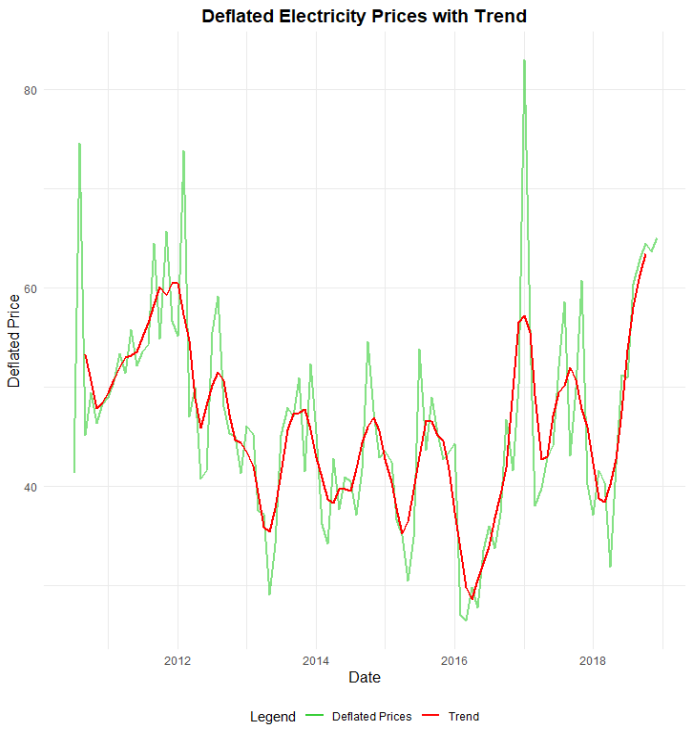

Figure 2: Deflated Electricity Prices with Trend

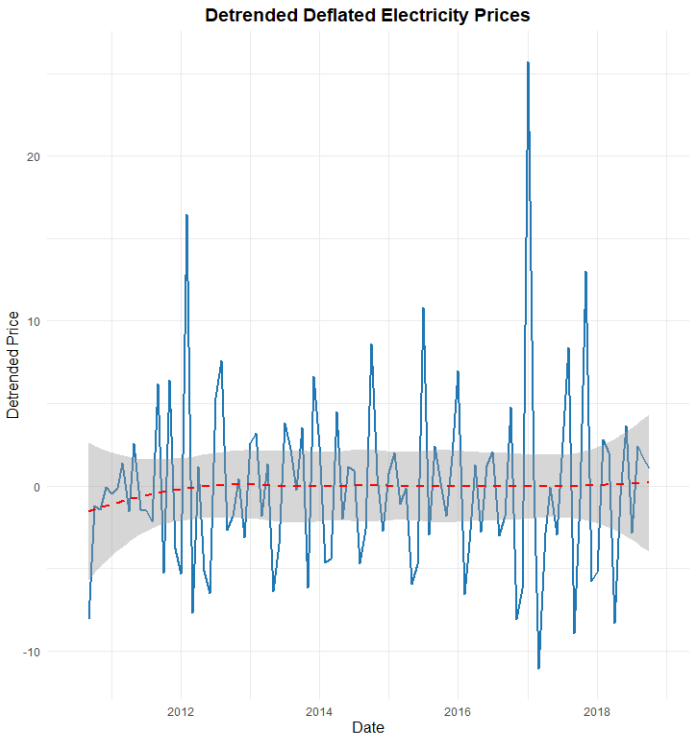

Figure 3: Detrend Deflated Electricity Prices

The seasonal component of deflated electricity prices reveals clear, repeating patterns on an annual basis. This means that the prices follow a specific trend each year, with regular peaks and troughs indicating predictable seasonal increases and decreases. These consistent periodic fluctuations suggest that seasonal changes, such as higher demand during winter for heating or summer for cooling, significantly impact electricity prices. Understanding this seasonal pattern allows for better forecasting of future prices, as these seasonal variations are predictable and repeat each year. Overall, the seasonality in electricity prices is well-defined and can be used to inform more accurate future price predictions.

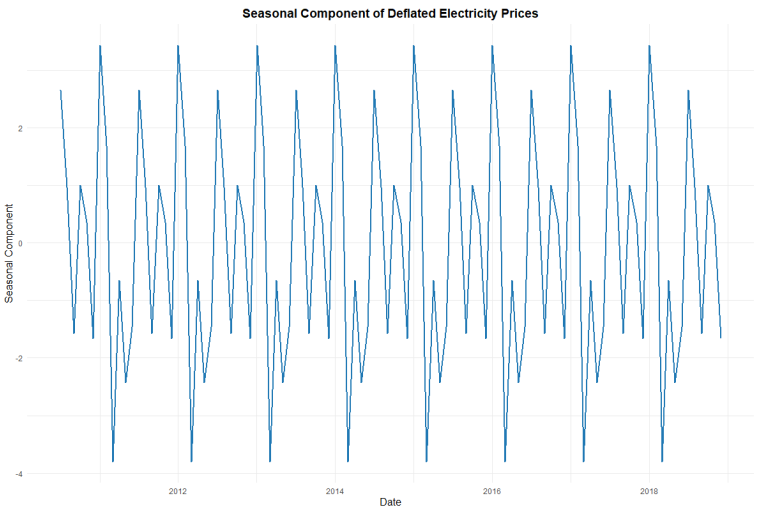

Figure 4: Seasonal Component of Deflated Electricity Prices

The random component represents the price changes that cannot be explained by trends or seasonal patterns (see Figure 5). These random fluctuations are important to analyze as they can reveal unpredictable events or external influences on electricity prices. There is no obvious pattern or trend in this plot, indicating that these fluctuations are random and do not follow any predictable pattern. There are some sharp peaks or dips, which might indicate unusual events or external factors that affected electricity prices.

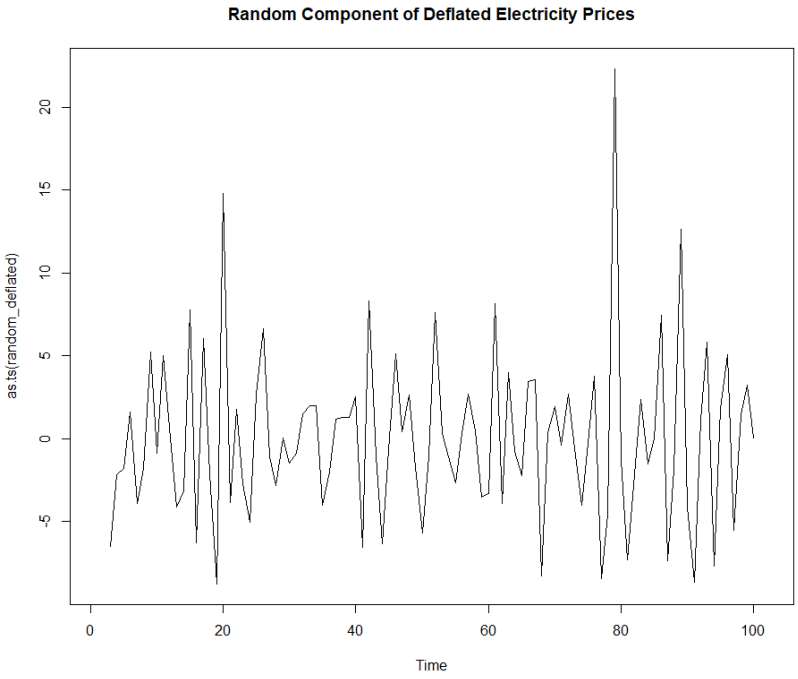

Figure 5: Random Component of Deflated Electricity Prices

4. Stationarity and Differencing

For achieving stationarity, which is essential for accurate time series forecasting we create **the first differences of the deflated prices** which demonstrates more stable fluctuations around a mean of zero, indicating that the series is stationary in its first differences. Also, we create **the log-transformed deflated prices** help to stabilize the variance, making it easier to observe trends and seasonality. Finally, we create and **the First Differences of Log Prices** that shows stable fluctuations around zero, confirming that the log-transformed series is stationary after differencing. The deflated prices and their log transformation show a right-skewed distribution, while the first differences (both in original and log-transformed form) show a more normal-like distribution centered around zero (see in Appendix Figure 8 and 9).

For the **Original, Differenced, Log-Transformed, and Log-Differenced Series** of deflated prices we create **ACF and PACF plots** and we conduct **Statistical tests** (ADF, KPSS, and PP tests) to check for stationarity (see Appendix Figure 10 and Table 2).

We check for the stationarity of deflated electricity prices using three tests: Augmented Dickey-Fuller (ADF), Kwiatkowski-Phillips-Schmidt-Shin (KPSS), and Phillips-Perron (PP). The results shows that the original deflated prices and their logarithms are likely non-stationary, as suggested by the ADF test (p-value 0.28 and 0.19 > 0.05 respectively). However, the first differences of both the deflated prices and the logarithms of deflated prices are stationary according to all three tests. However, it is also evident from the plots that the first differences have more values outside the confidence bands, suggesting potential issues with autocorrelation that need further attention.

Based on these findings, **we will use the first differences and the first differences of the logarithms of the prices for further analysis since these series are stationary.**

5. Model Estimation and Model Diagnostics

We tried out and estimate various models to find a good fit for the data. In more details, we estimate MA, AR, ARMA, ARIMA models, Seasonal ARIMA (SARIMA) models, GARCH models, Holt-Winters exponential smoothing.

Each model's performance is evaluated using criteria like AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) (The model with the lowest AIC ναι BIC value is considered the best), and Forecast Error Measures like MAE (Mean Absolute Error- The average of the absolute prediction errors) , MSE (Mean Squared Error -The average of the squared prediction errors), RMSE (Root Mean Squared Error -The square root of the MSE), MAPE (Mean Absolute Percentage Error -the average of the absolute percentage prediction errors) .

Diagnostic checks are performed on the residuals to ensure the models are appropriate for the data. We check for white noise (if are independent and normally distributed). This includes checking for autocorrelation (ACF and PACF plots to check for autocorrelation in the residuals), normality (Q-Q plots to check if the residuals follow a normal distribution), and using tests such as Ljung-Box and Shapiro-Wilk to validate the model assumptions.

The model that we finally choose is the SARIMA model for the first differences of the Deflated prices

$$y_t = \varphi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t = 0.4414\, y_{t-1} - 0.8981\, \epsilon_{t-1} + \epsilon_t$$

The current value y_t of the series is determined by the previous value y_{t-1}, the error term from the previous period ϵ_{t-1}, and the current error term ϵ_t

The SARIMA model can be expressed mathematically using its components: the AR (AutoRegressive) part, the MA (Moving Average) part, and the seasonal component if present. For the model we have, ARIMA(1,0,1) with zero mean.

y_t is the differenced series of the Deflated prices

ϕ_1 is the coefficient of the AR(1) term (= 0.4414).

θ_1 is the coefficient of the MA(1) term (= -0.8981).

ϵ_t is the white noise error term at time t.

ϵ_{t-1} is the white noise error term at time t-1.

We also perform an extensive search to identify the best SARIMA models for two time series: the first differences of prices and the first differences of the logarithms of prices. The process iterated through all possible combinations of SARIMA parameters (p, d, q) and seasonal parameters (P, D, Q, s), evaluating each model based on the Akaike Information Criterion (AIC), and selecting the models with the lowest AIC as the optimal models. Once the best models are identified, they are trained on the respective time series data. However, the algorithm trapped in a small neighborhood of this parameters and for this reason the automatic function that the R package has (auto. arima) find a better model with smaller AIC.

The diagnostic analysis of the SARIMA model

The diagnostic analysis of the SARIMA model for the first differences of deflated prices indicates that the residuals are mostly behaving like white noise, as shown by the Ljung-Box test p-values which are above 0.05 for lags 8,9,10, 20, and 30. This suggests that there is no significant autocorrelation left in the residuals, meaning the model has captured the autocorrelations present in the data well. However, the Shapiro-Wilk test for normality has a p-value significantly less than 0.05, show that the residuals are not normally distributed. The ACF and PACF plots of the residuals and their squared values further support the lack of significant autocorrelation. The Q-Q plot shows some deviation from the normal line, reflecting the non-normality of residuals. However, form the qq-plot we can see that this not normality is for few observations.

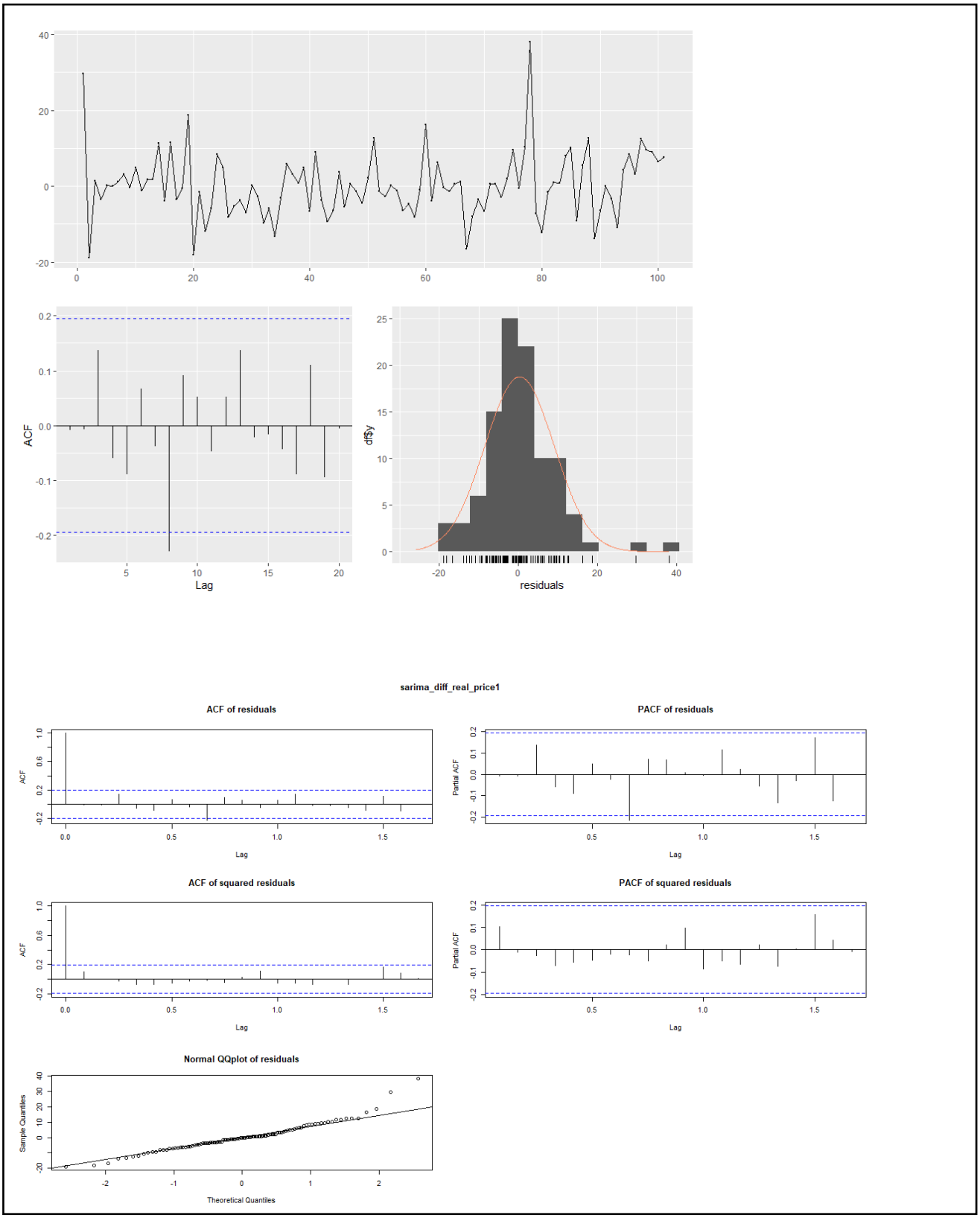

Figure 6: Diagnostics plots for SARIMA model

6. Forecasting

To forecast the electricity prices for the first half of 2019, we used the best-performing SARIMA model identified earlier: SARIMA(1,0,1) with zero mean. This model was applied to the first differences of the deflated electricity prices to predict future values.

The forecasts for the first six months of 2019, along with their confidence intervals, are shown in the following plot (figure 7). The plot shows the forecasted values as well as the 80% and 95% confidence intervals :

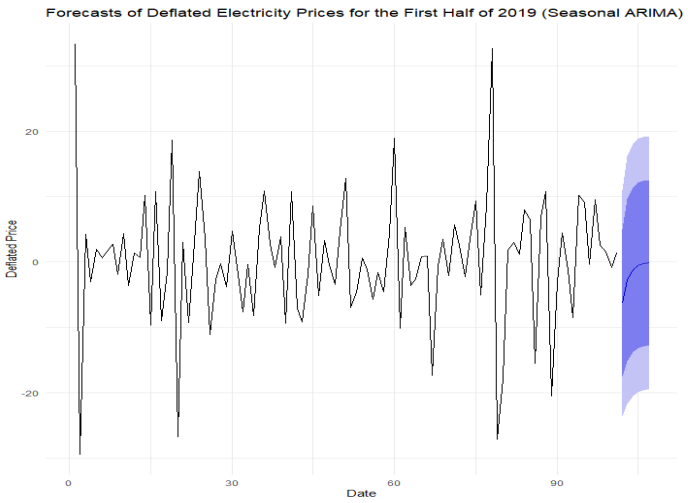

Figure 7: The forecasts of deflated electricity prices for the first six months of 2019

The forecasted values for the deflated electricity prices for the first half of 2019 are:

Date	Forecasted Price	80% Lower CI	80% Upper CI	95% Lower CI	95% Upper CI	Direction
2019-01-01	58.77534	47.497917	70.05277	41.5280092	76.02268	Decrease
2019-02-01	55.99866	32.323504	79.67382	19.7906329	92.20669	Decrease
2019-03-01	54.77291	18.493288	91.05253	-0.7119822	110.25780	Decrease
2019-04-01	54.23180	5.307829	103.15578	-20.5909580	129.05457	Decrease
2019-05-01	53.99294	-7.583153	115.56902	-40.1795651	148.16544	Decrease
2019-06-01	53.88749	-20.342226	128.11720	-59.6370635	167.41204	Decrease

The forecasted prices show a slight decreasing trend over the first six months of 2019. The forecast indicates that prices are expected to fall slightly, with the predicted values ranging from 58.77 in January to 53.88 in June. The confidence intervals widen over time, reflecting increased uncertainty further into the future. All forecasted months indicate a decrease in prices compared to the previous month. This trend might suggest seasonal effects or underlying factors leading to a consistent decrease over this period.

7. Rolling Forecast and Cross-Validation

To validate the predictive accuracy of the SARIMA model, a rolling forecast cross-validation approach is implemented. So, the initial dataset is divided into training and test sets and an initial training period is defined, and the model is trained on this data. The model is trained on the training data, and forecasts are generated for a specified horizon (in this case, 6 months). After that the forecasted values are compared with the actual test data. In each step the training set is then expanded by one time step, and the process is repeated until the entire dataset has been used for validation.

The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are calculated to measure the accuracy of the forecasts. The RMSE provides an indication of the average magnitude of the forecast errors, with larger errors being more heavily penalized. The MAE provides a straightforward measure of the average absolute forecast errors.

The rolling forecast cross-validation was performed using the first differences of the deflated electricity prices. The performance of the model was evaluated using RMSE and MAE:

Root Mean Squared Error: 10.99661

Mean Absolute Error: 8.142281

These metrics show the model's predictive power and accuracy in forecasting the deflated electricity prices. Lower values of RMSE and MAE suggest better model performance, as they represent smaller deviations between the forecasted and actual values.

8. Appendix

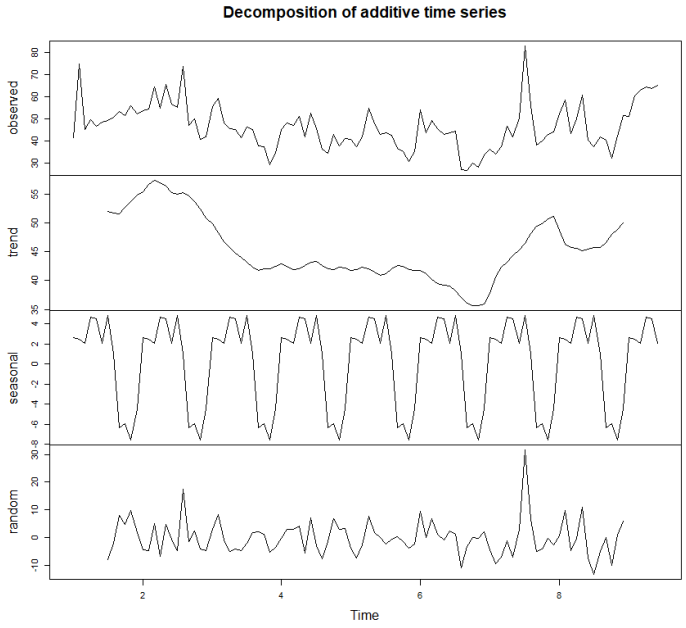

Figure 6: Decomposition of additive time series

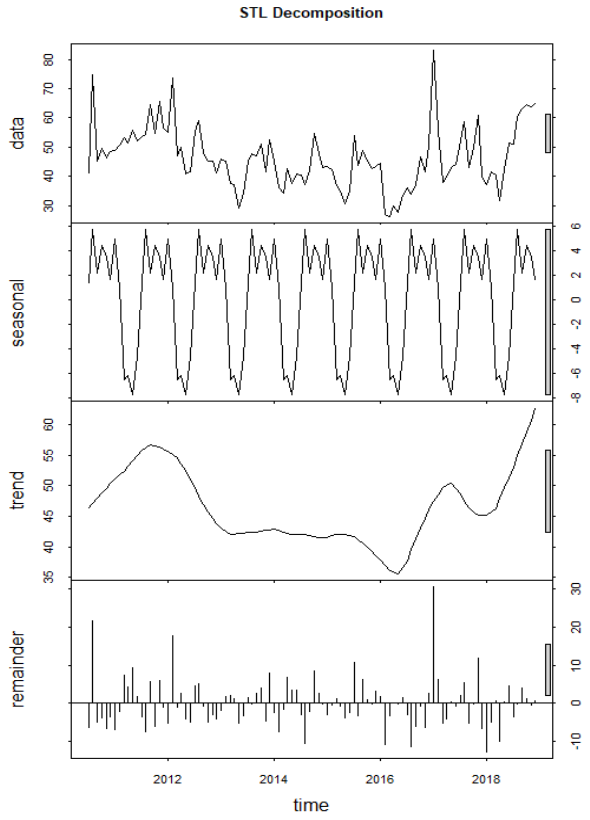

Figure 7: STL Decomposition of additive time series

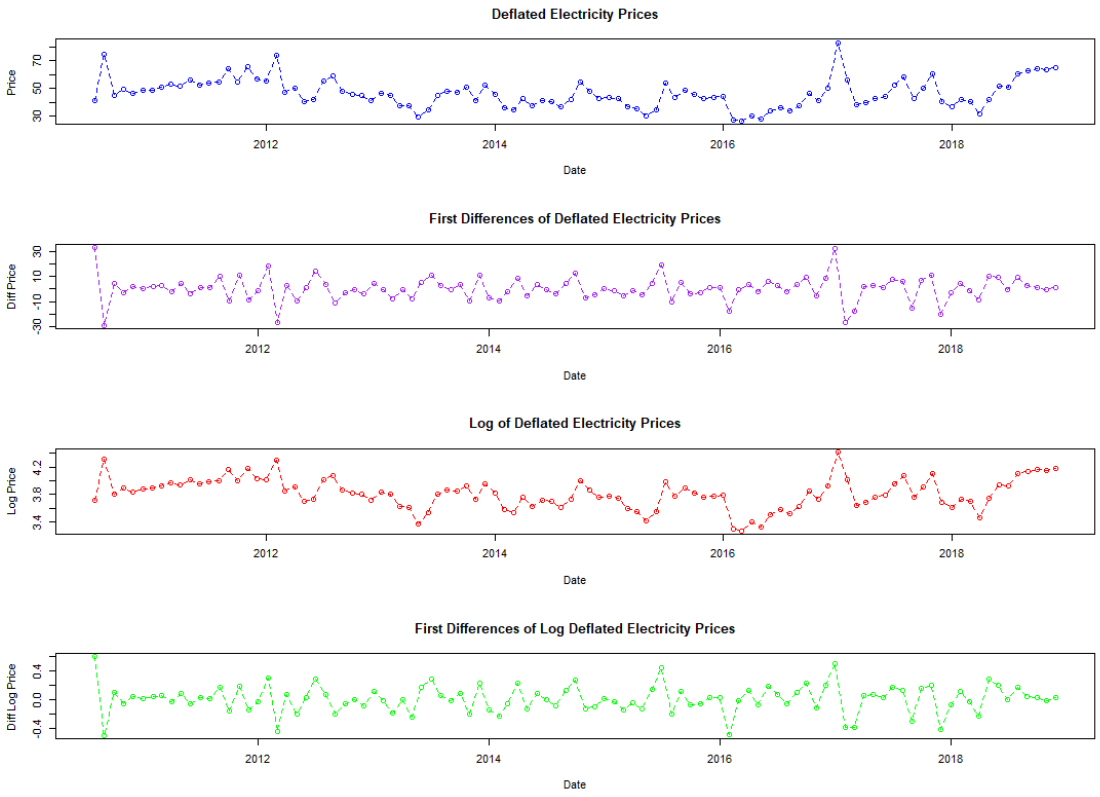

Figure 8: Deflated Electricity Prices: Original, Differenced, Log-Transformed, and Log-Differenced Series

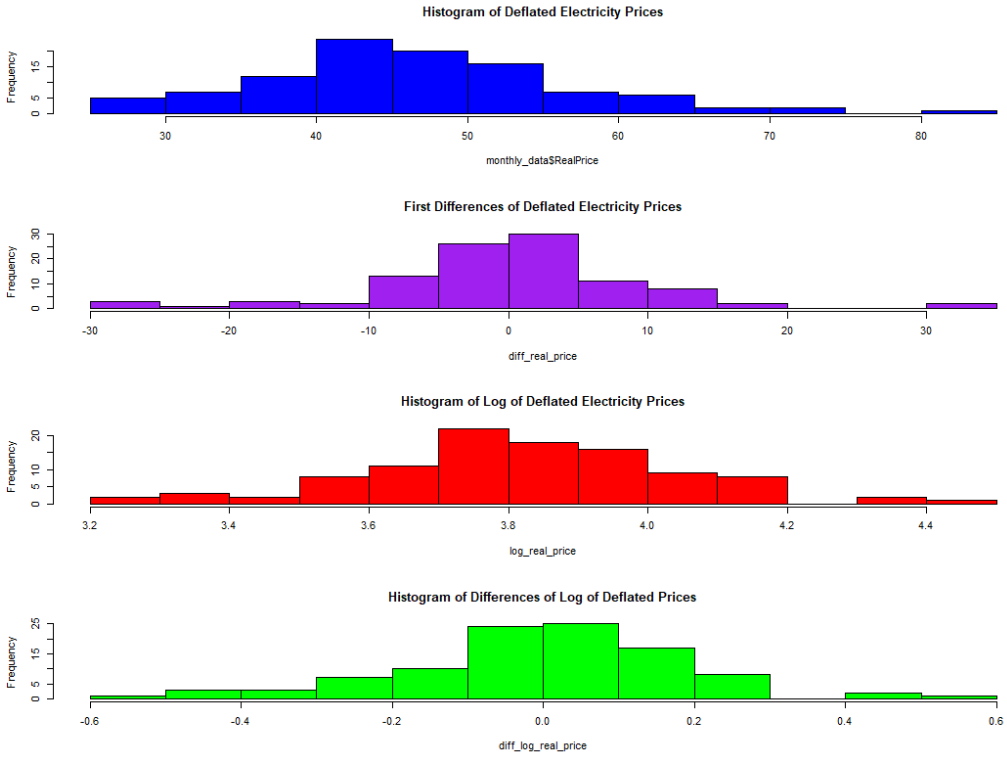

Figure 9: Histograms of Deflated Electricity Prices: Original, Differenced, Log-Transformed, and Log-Differenced Series

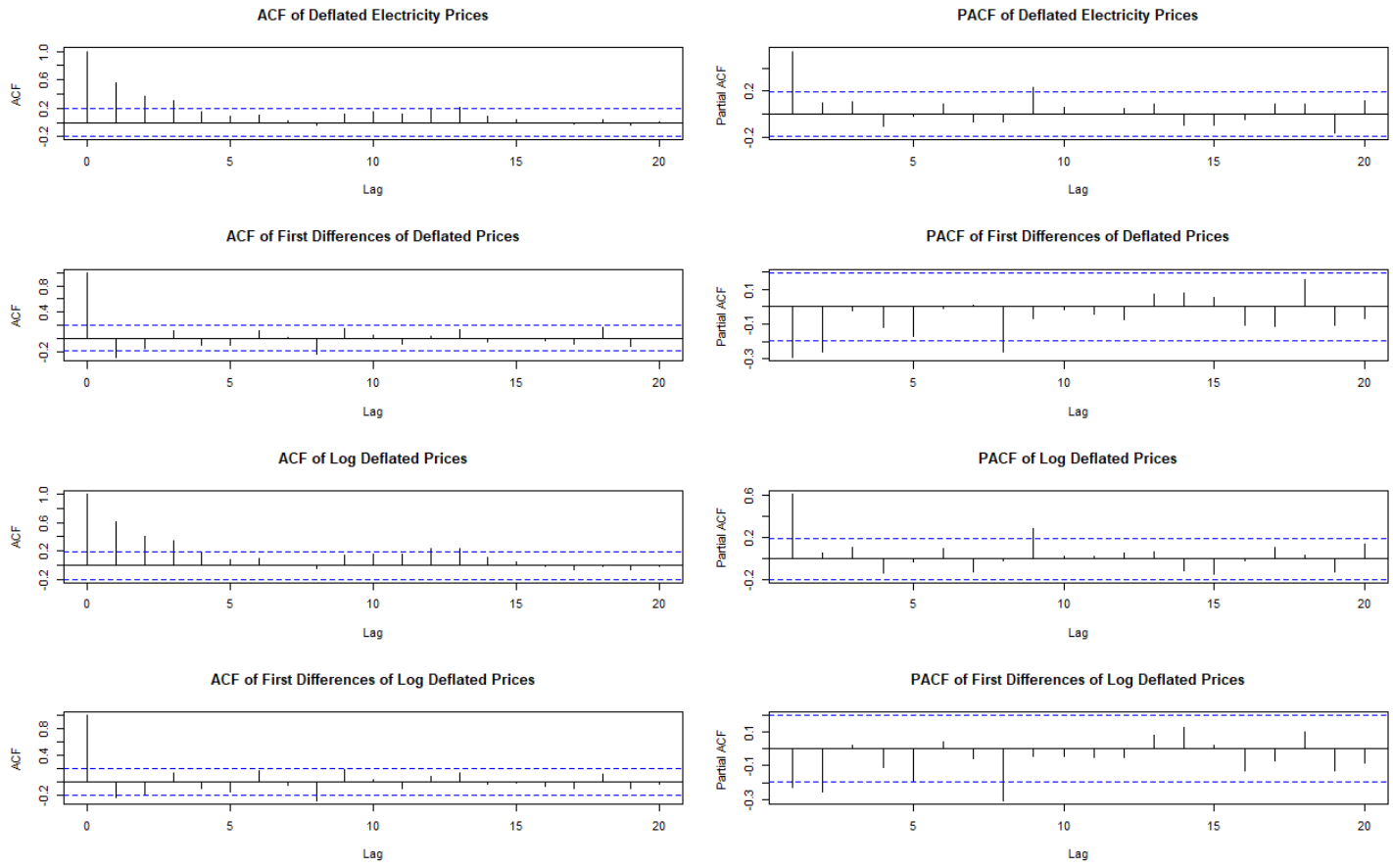

Figure 10: ACF and PACF Analysis of Deflated Electricity Prices and Their Transformations

Series	ADF_p_value	KPSS_p_value	PP_p_value
Original deflated	0.2787372	0.1	0.01
First Differences	0.0100000	0.1	0.01
Log Prices	0.1935298	0.1	0.01
First Differences of Log Prices	0.0100000	0.1	0.01

Table 2: P-Values from ADF, KPSS, and PP Tests for Various Time Series Transformations