Analysis and Forecasting of Australian Quarterly Electricity Production

Jingtang Sun(jingtang@ucsb.edu)

2024-06-14

Abstract

This report presents a comprehensive analysis and forecasting of Australia's quarterly electricity production from March 1956 to September 1994. Utilizing time series analysis techniques, the study aims to model the complex patterns within the data and generate accurate forecasts for future electricity production. Initially, the data undergoes a Box-Cox transformation to stabilize variance, followed by differencing to achieve stationarity. Two primary methodologies are employed: Seasonal Autoregressive Integrated Moving Average (SARIMA) modeling and Spectral Analysis.

The SARIMA model is tailored to capture both seasonal and non-seasonal components of the time series, effectively modeling the observed trends and cyclical patterns. Through exploratory data analysis, including Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, the optimal SARIMA parameters are identified. The model's adequacy is validated using diagnostic checks on the residuals, ensuring they exhibit white noise characteristics, thus confirming the model's robustness for forecasting purposes.

In addition to SARIMA, Spectral Analysis is conducted to investigate the frequency domain characteristics of the data. This approach identifies significant periodic components and assesses the residuals for any remaining cyclical patterns not captured by the time-domain model. The periodogram analysis reveals the dominant frequencies in the data, and tests like Fisher's test and the cumulative periodogram verify that the residuals are consistent with white noise.

The combination of these methodologies provides a robust framework for understanding and predicting the dynamics of electricity production in Australia. The results demonstrate that the chosen models effectively capture the underlying structure of the data, offering valuable insights for future energy planning and management.

1. Introduction

The analysis and forecasting of electricity production are critical components for managing and planning energy resources efficiently. In this project, we focus on Australia's quarterly electricity production data, spanning from March 1956 to September 1994. The primary aim of this study is to model the complex patterns within the time series data and to generate accurate forecasts that could be valuable for future energy planning and management.

Electricity is the backbone of modern civilization, playing a pivotal role in nearly every aspect of our daily lives. It powers our homes, enabling us to light our spaces, cook our meals, and stay connected through digital devices. In the workplace, electricity drives machinery, computers, and communication systems, making it indispensable for productivity and economic activity. Beyond personal and professional use, electricity is critical for public services, including healthcare, where it powers life-saving equipment, and transportation, where it fuels electric vehicles and supports infrastructure. The consistent and reliable supply of electricity is essential for the smooth functioning of contemporary society, making its production, distribution, and consumption a focal point of interest. Investigating electricity production trends helps us anticipate future demands, plan sustainable growth, and address the challenges posed by increasing energy needs, thus ensuring the continued operation of our daily lives and the advancement of technological and economic progress.

The motivation for selecting this particular dataset stems from the increasing importance of sustainable energy management and the need for precise forecasting methods to anticipate future energy demands. Australia, with its diverse energy portfolio and significant fluctuations in electricity production, provides a compelling case study for time series analysis. By understanding the historical trends and seasonal patterns in electricity production, we can gain insights that are critical for policy-making and operational strategies in the energy sector.

The dataset under consideration has been the subject of various studies aimed at exploring seasonal variations, long-term trends, and the impact of external factors on electricity production. Previous research has employed a range of statistical and machine learning techniques to analyze these data, including linear regression models, exponential smoothing methods, and classical ARIMA models. These studies have generally confirmed the presence of strong seasonal patterns and trends in the data, emphasizing the need for models that can adequately capture both seasonal and non-seasonal components.

2. Dataset Exploration

The dataset under analysis comprises the quarterly electricity production in Australia, measured in million kilowatt hours, spanning from March 1956 to September 1994. This dataset, covering nearly four decades, provides a comprehensive view of Australia's electricity production over time. The frequency of the dataset is quarterly, which aligns with the typical business and economic reporting periods, allowing for detailed examination of seasonal and cyclical patterns within the year.

```
##      Qtr1 Qtr2 Qtr3 Qtr4
## 1956 3923 4436 4806 4418
## 1957 4339 4811
```

The values in the dataset are recorded in million kilowatt hours, reflecting the amount of electricity produced each quarter. These values are non-negative, indicating the continuous production and availability of electricity throughout the observed period. The dataset consists of 155 observations, providing a robust sample size for time series analysis, which is essential for capturing the long-term trends and seasonal effects inherent in electricity production data.

The choice to study this dataset is motivated by the significant role electricity plays in both everyday life and the broader economic landscape. Understanding electricity production trends is crucial for energy planning, policy-making, and managing future demands. Australia's diverse energy portfolio and the historical context provided by this dataset make it an excellent candidate for analyzing how electricity production has evolved and how it may respond to future changes.

This dataset was sourced from the Time Series Data Library (TSDL), which is a well-regarded repository for time series data, created by Dr. Rob Hyndman, Professor of Statistics at Monash University in 2018 and maintained by Yangzhuoran Yang, PhD student at Monash University. It can be accessed via the following link: https://pkg.yangzhuoranyang.com/tsdl/. The dataset was collected and compiled to provide historical insights into electricity production trends and patterns in Australia. This historical perspective is invaluable for understanding the long-term dynamics and for developing forecasting models that can help in planning for sustainable energy management.

The dataset's significance lies in its ability to inform about the long-term trends and seasonal patterns in electricity production, which are critical for various stakeholders, including policymakers, energy companies, and researchers. By studying this dataset, the aim is to model these patterns accurately and generate forecasts that can aid in efficient energy planning and management. This project contributes to understanding how historical production patterns can guide future energy strategies and respond to evolving demands in the energy sector.

3. Methodology

3.1 SARIMA (p, d, q) x (P, D, Q) model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the traditional ARIMA model by incorporating seasonal components. This makes SARIMA particularly well-suited for time series data exhibiting seasonality, as it allows for both non-seasonal and seasonal differencing to remove trends and cyclical patterns. The model is characterized by seven parameters: non-seasonal autoregressive (p), differencing (d), and moving average (q) terms, along with their seasonal counterparts (P, D, Q), and the seasonal period (s).

In this project, we chose the SARIMA model to capture both the short-term dynamics and the quarterly seasonal effects in our dataset of Australian electricity production. To fit the SARIMA model, we followed the Box-Jenkins methodology, which involves identifying the appropriate model through the examination of time series plots and statistical tests. We began with exploratory data analysis, which included plotting the original time series data and transforming it using a Box-Cox transformation to stabilize the variance. Subsequently, we performed differencing to achieve stationarity, as indicated by the Augmented Dickey-Fuller (ADF) test.

The next step involved analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced data. These plots provided insights into the potential orders of the AR and MA components. We used these insights to estimate various SARIMA models and compared them using the Akaike Information Criterion (AIC), and the one with the lowest AIC value was selected as the optimal model.

Once the optimal model was identified, we conducted diagnostic checks on the residuals. These included the ACF and PACF plots of the residuals, Normal Q-Q plots, and the Ljung-Box test. These diagnostics confirmed that the residuals behaved like white noise, indicating that the model adequately captured the patterns in the data. The final SARIMA model was then used to generate forecasts, providing valuable insights into future electricity production trends.

3.2 Spectral Analysis

Spectral analysis is a powerful technique in time series analysis used to decompose a signal into its constituent frequencies. This method is particularly effective for identifying periodic or cyclical behaviors in data, such as those present in Australia's quarterly electricity production.

In my project, I started by generating a raw periodogram. This periodogram provided a preliminary estimate of the spectral density, revealing the distribution of variance across different frequencies. To enhance the clarity and interpretability of the periodogram, I applied a tapering window to reduce noise and smooth the spectral estimates. The tapering process helped emphasize significant frequencies more effectively by minimizing the impact of minor fluctuations.

To further refine the spectral density estimation, I used the Daniell Kernel for smoothing, which applied centered moving averages to provide a more refined spectrum. This smoothed periodogram highlighted the dominant frequencies in the data, indicating the presence of regular patterns or cycles in electricity production.

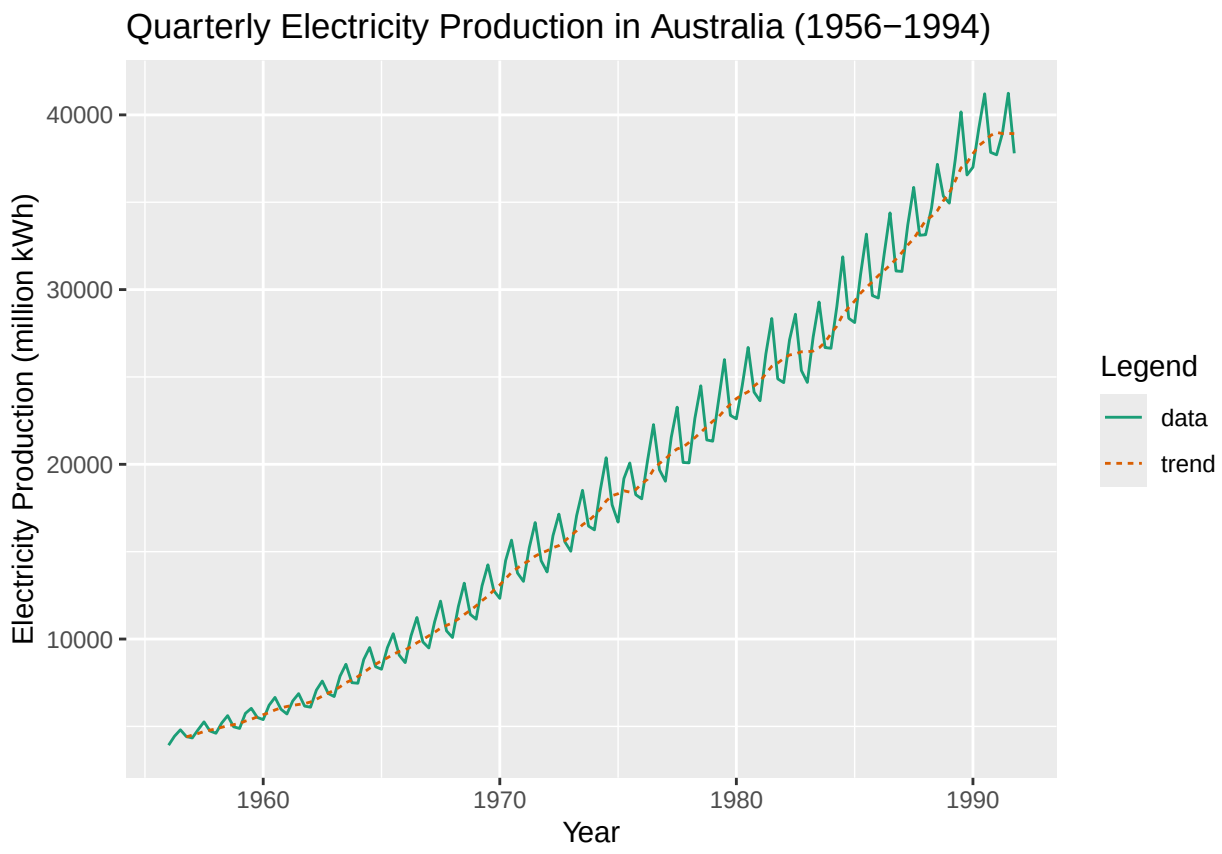
To validate our findings, we performed statistical tests such as Fisher's test for periodicity and the Kolmogorov-Smirnov test on the cumulative periodogram of the residuals from our SARIMA model. These tests assessed whether the residuals exhibited any remaining periodic behavior that the time-domain model might have missed.

The combination of time-domain analysis through SARIMA and frequency-domain analysis through spectral methods provided a comprehensive understanding of the dynamics in the electricity production data. Spectral analysis, in particular, offered a detailed view of the frequency components, complementing the insights obtained from the SARIMA model and enhancing our overall model's robustness and accuracy.

4. Results

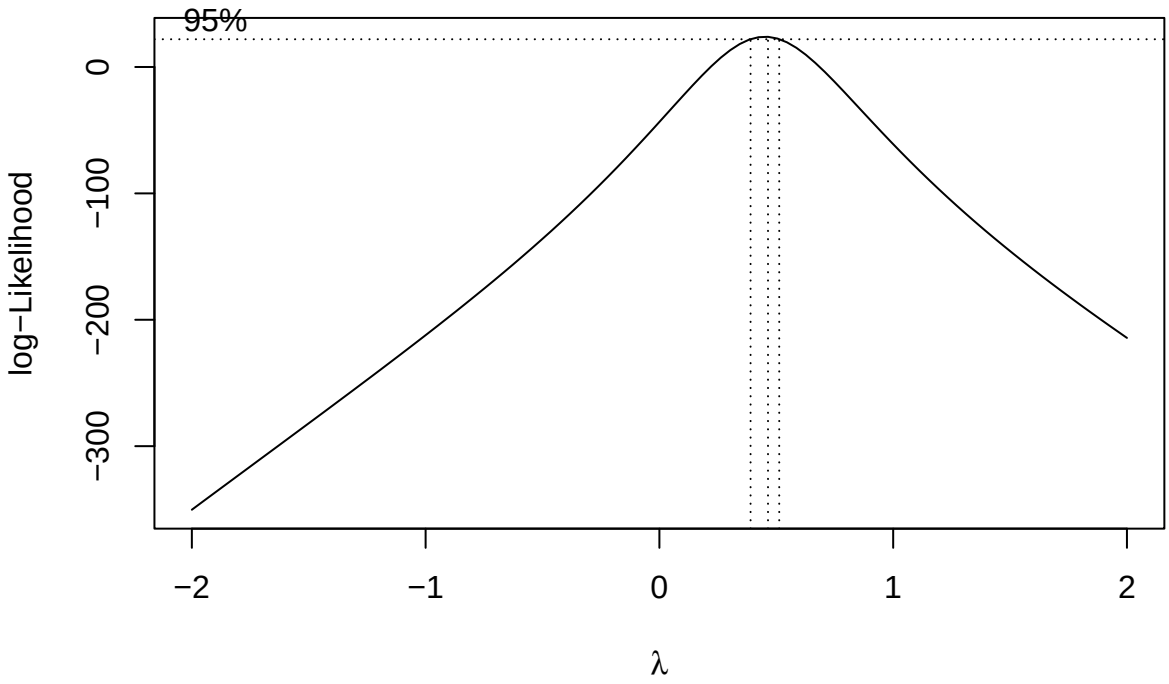
4.1 Results from SARIMA (p, d, q) x (P, D, Q)

Plot of original data



The plot shows that my time series is apparently non-stationary, and it does not have a constant variance over time. Therefore, I will first try the Boxcox analysis to find the optimal λ.

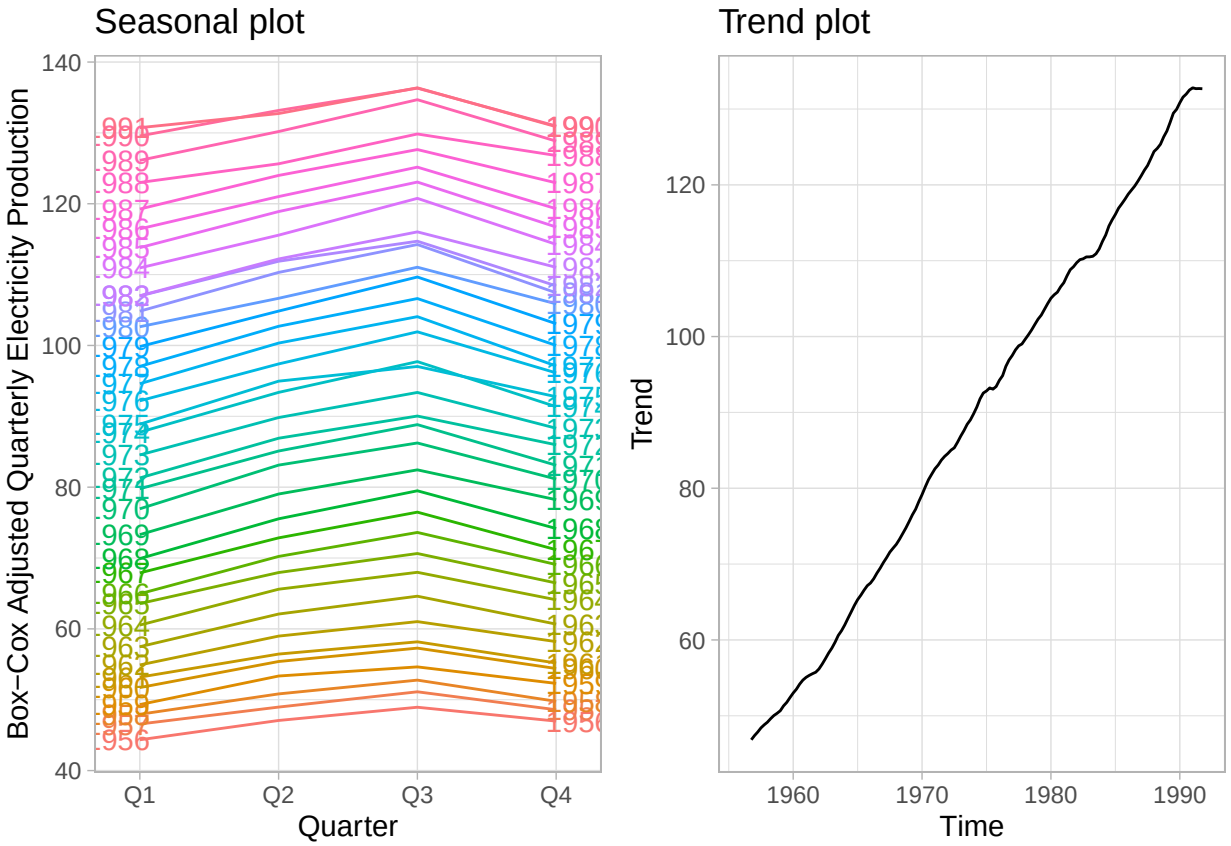
Boxcox analysis




```
## [1] 0.4646465
```

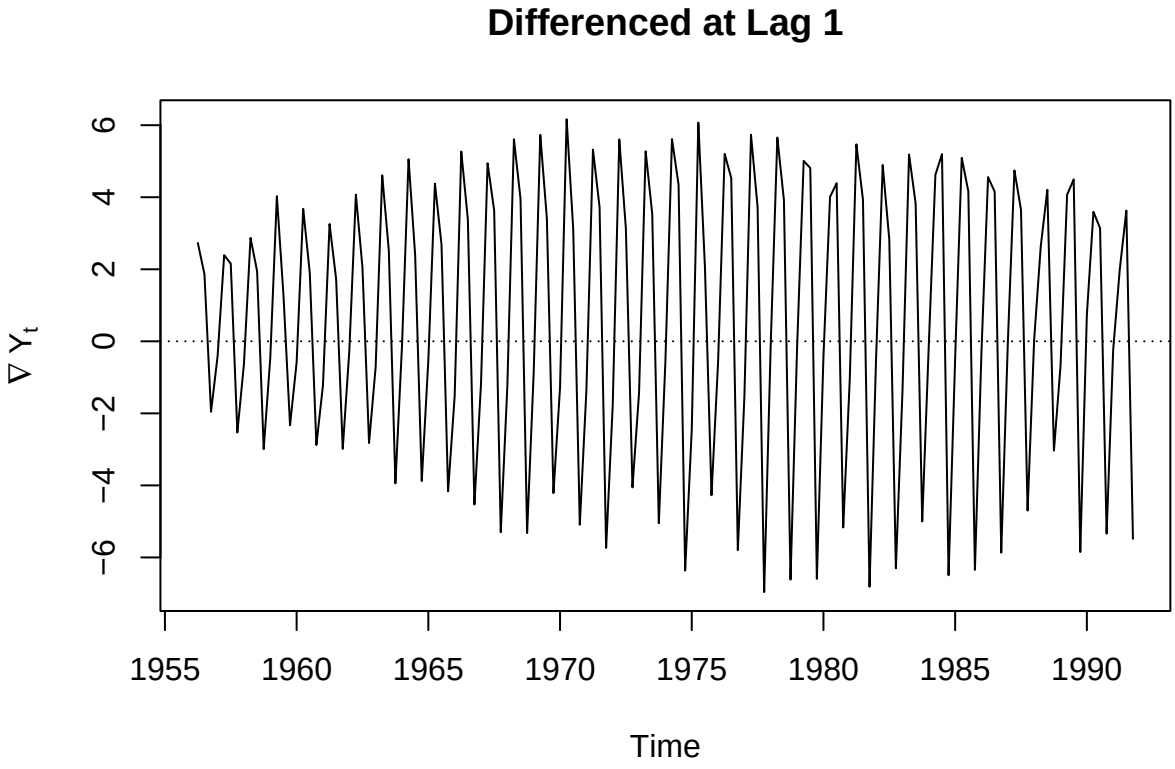
As we know, we will perform a log transformation if the λ is approximately 0, or a sqrt transformation if the λ is approximately 0.5. Since the optimal λ I get is around 0.464, the following Box-Cox transformation was applied to the data: $G(Y_t) = \frac{Y_t^{\lambda}-1}{\lambda}$

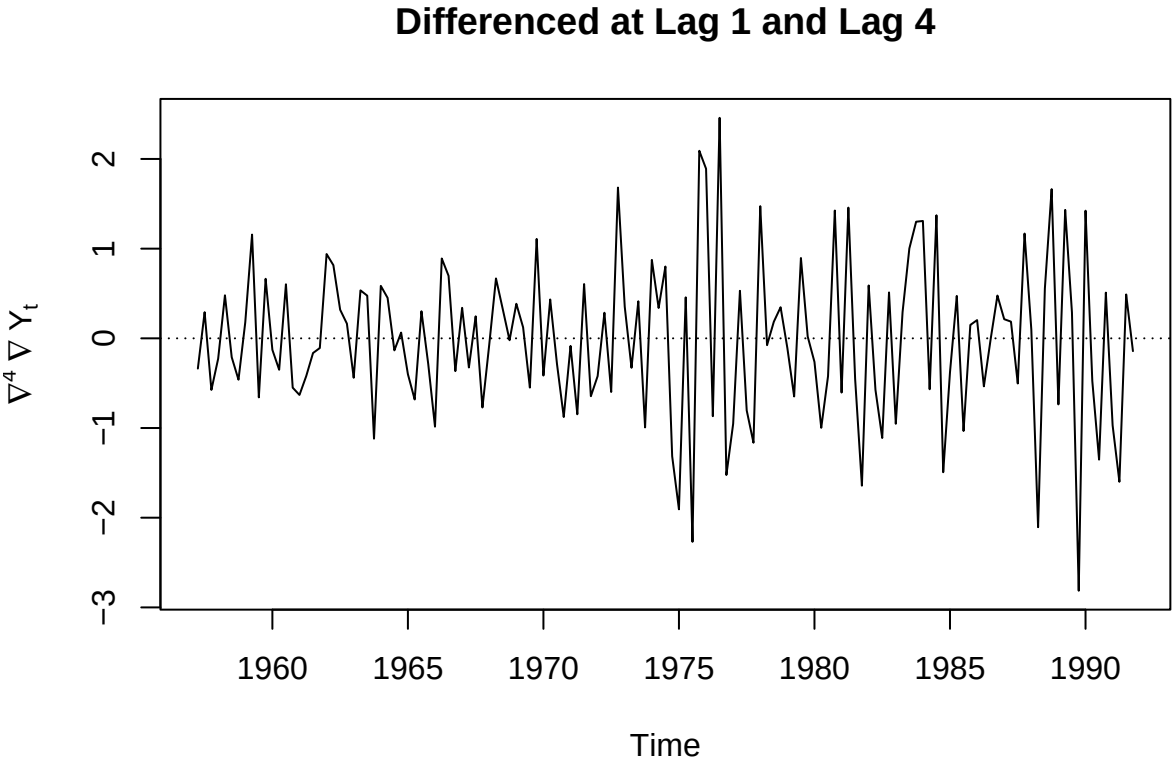
Seasonality and Trend



The Box-Cox transformed data shows a seasonality and trend, so we first differenced the data at lag 1 to remove the trend and again at lag 4 to remove the seasonality for the quarterly data, so that our series will be: $y_t^{2diff} = y_t - y_{t-1} = (1-B)y_t = (1-B)(1-B^4)x_t = \nabla_1\nabla_4 x_t$.

Differencing

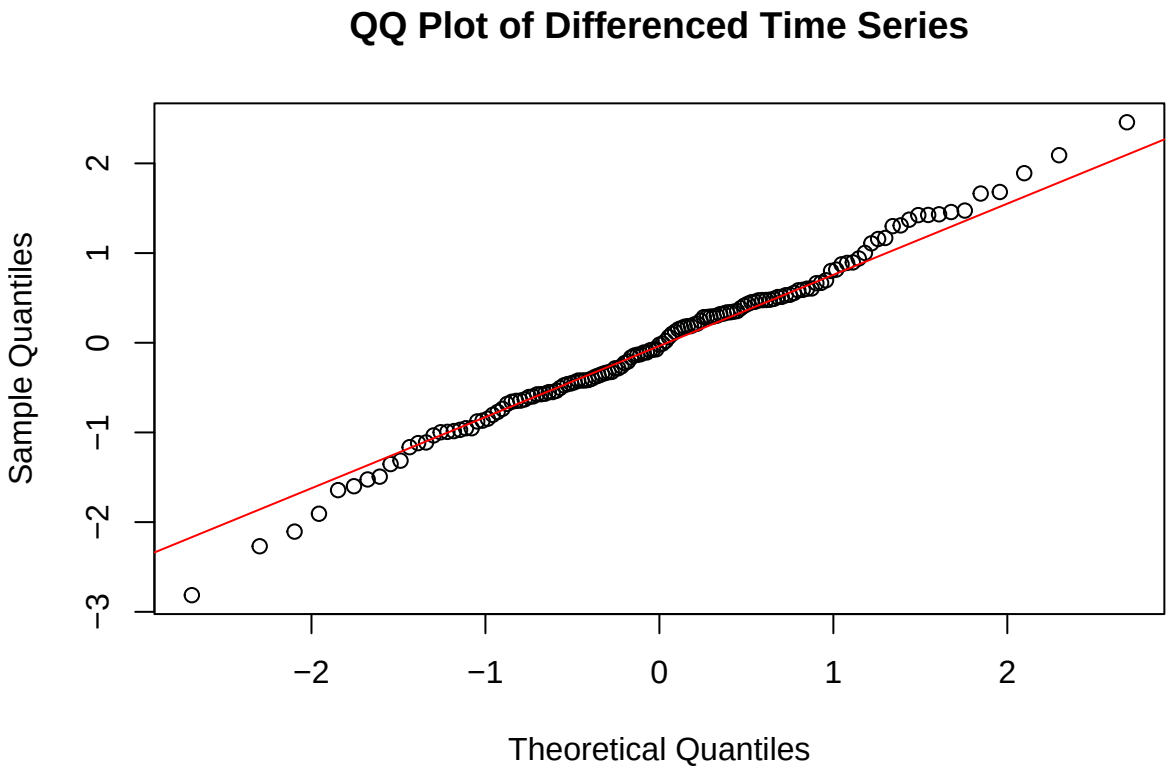




```
## [1] "Variance of transformed series = 714.529789626575"

## [1] "Variance of first differenced series = 14.8627852555925"

## [1] "Variance of second differenced series = 0.792945484759155"

## [1] "Variance of third differenced series = 2.34767134412705"
```

Since the variance of the third differencing increases comparing to the variance of the second differencing, we do not need to perform the third differencing.

Normality and Stationarity Inspection



In the QQ plot, most of the data points follow the red line, especially in the middle range. Nevertheless, the points in the tails (extreme values) diverge from the red line, especially in the right tail. Therefore, while the main body of my data appears to be normally distributed, the deviations in the tails indicate that we need to examine the normality further.

```
##
##  Shapiro-Wilk normality test
##
## data:  diff4
## W = 0.99295, p-value = 0.7239
```

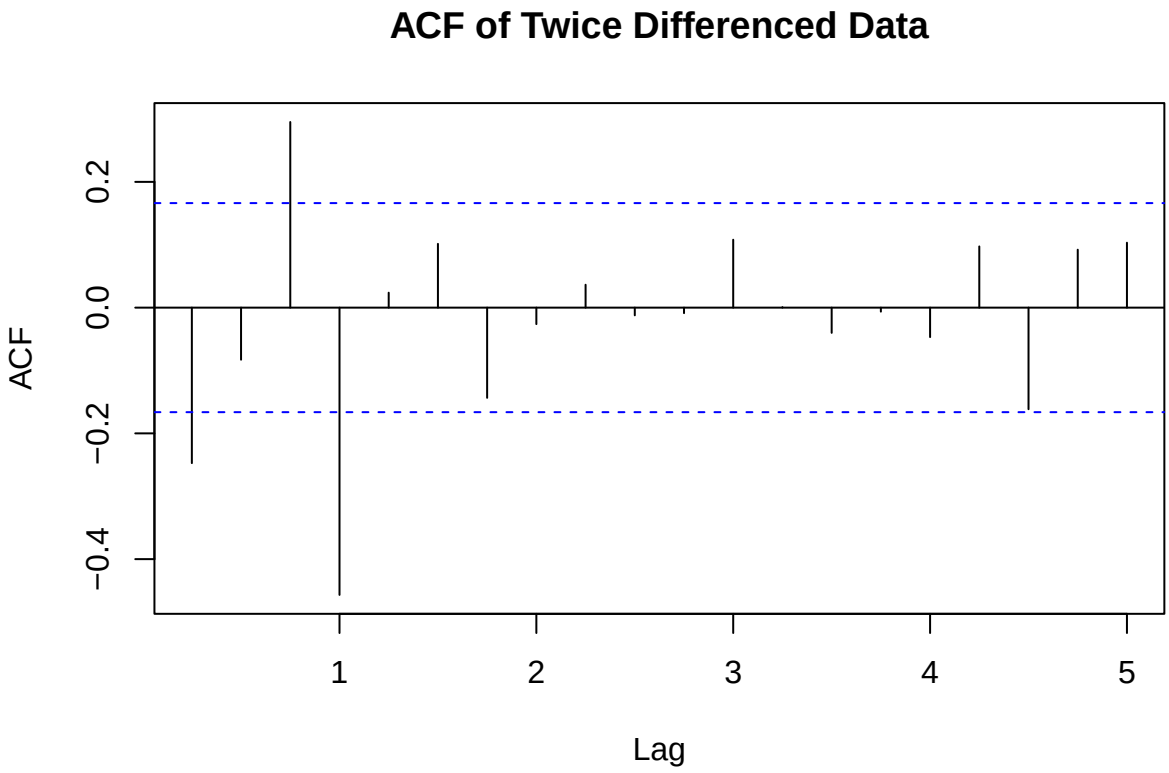
Shapiro-Wilk test is an alternative way to assess normality. Since the p-value(0.7239) is much greater than 0.05, we fail to reject the null hypothesis that the transformed time series data (`diff4`) is normally distributed. Therefore, we can conclude that the data is approximately **normal**.

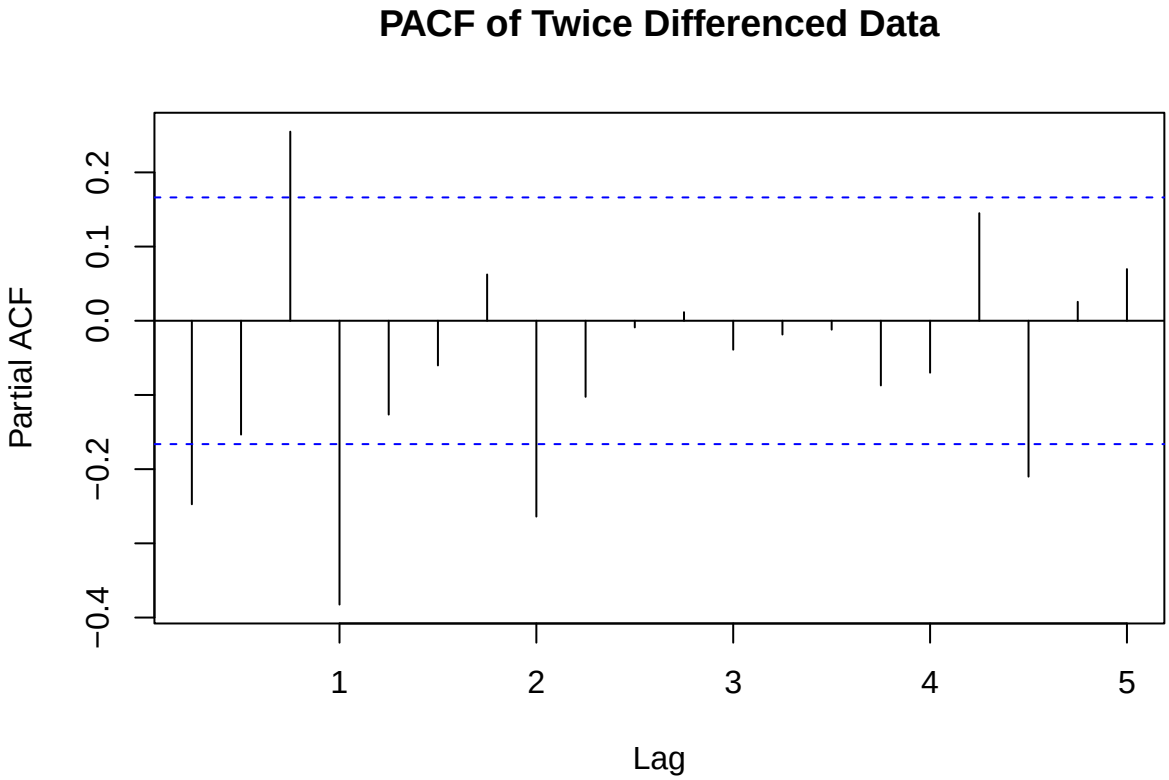
```
## Warning in adf.test(diff4): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff4
## Dickey-Fuller = -6.5209, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

Given the test statistic of -6.5209 and a p-value of 0.01, we reject the null hypothesis at the 1% significance level. This indicates that the differenced series (`diff4`) is generally **stationary**.

Model Selection





The dataset comprises quarterly data that exhibits both seasonality and a trend, suggesting the use of a SARIMA(p, d, q)(P, D, Q)s model structure. We performed differencing at lag 1 to eliminate the trend, which establishes $d = 1$. Additionally, we differenced at lag 4 to remove the seasonal component, resulting in $D = 1$.

Analyzing the ACF plot of the twice-differenced data, we observe significant spikes at lag 1 (which corresponds to lag 4 in the original data), indicating a strong non-seasonal moving average component. Thus, we consider $q = 4$ as a suitable choice for the non-seasonal moving average component. The presence of a significant spike at the first lag of the differenced series also suggests a seasonal moving average component $Q = 1$.

Turning to the PACF plot, we note significant spikes at the first four lags, which imply that the non-seasonal autoregressive component p could be 4. Furthermore, the PACF shows a notable spike at lag 4 (interpreted as lag 1 for the differenced series and lag 4 for the original data) and borderline significant spikes at lag 2 (corresponding to lag 8 in the original series). This suggests that the seasonal autoregressive component P could be 1 or 2.

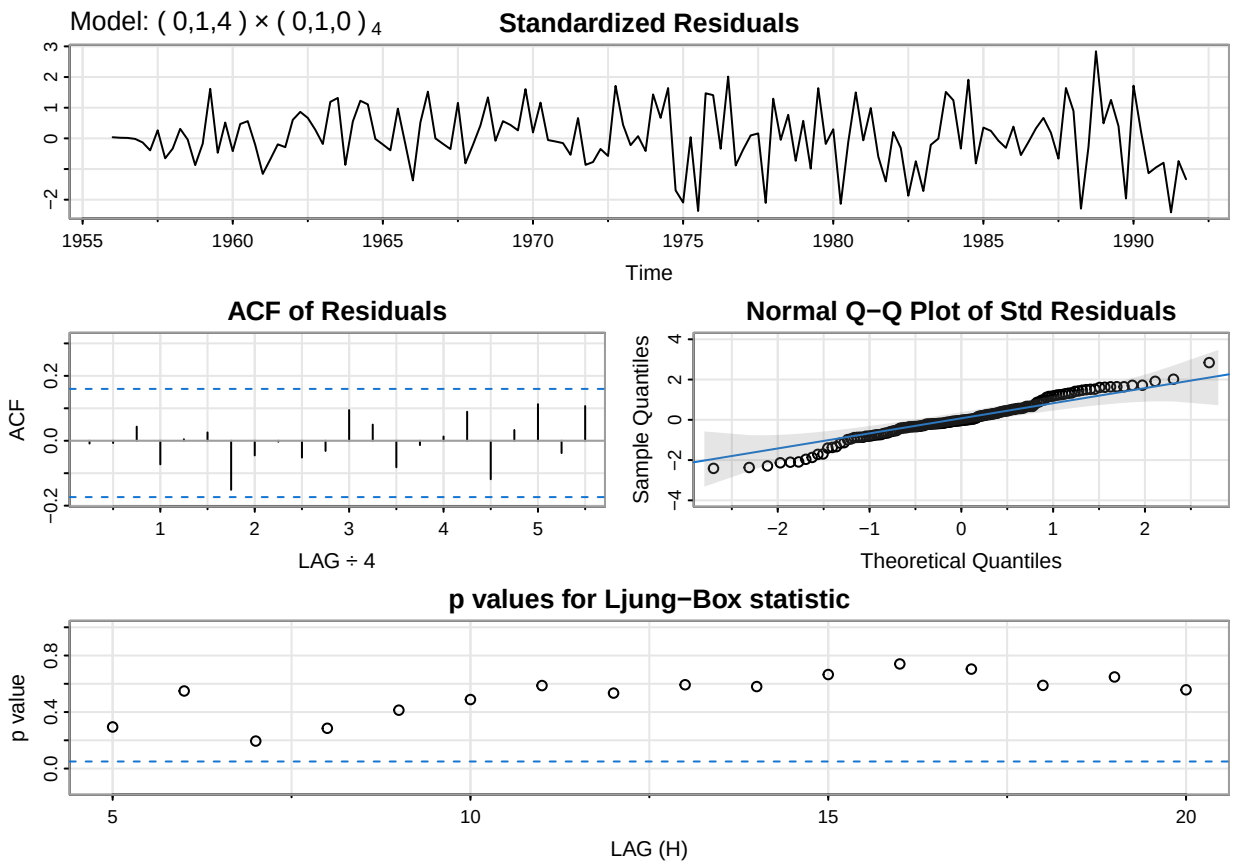
```
##    p d q P D Q      AIC     AICc      BIC
## 25 0 1 4 0 1 0 310.0023 310.4534 324.6747
```

```
## Series: bc_train
## ARIMA(0,1,4)(0,1,0)[4]
##
## Coefficients:
##           ma1      ma2     ma3      ma4
##       -0.2925  -0.1726  0.0920  -0.6270
## s.e.   0.0735   0.0694  0.0736   0.0637
##
## sigma^2 = 0.5045:  log likelihood = -150
## AIC=310   AICc=310.45   BIC=324.67
##
## Training set error measures:
##                       ME      RMSE       MAE       MPE      MAPE      MASE
## Training set 0.03147856 0.6876988 0.5239584 0.0481069 0.5925533 0.2121318
##                       ACF1
## Training set -0.008570307
```

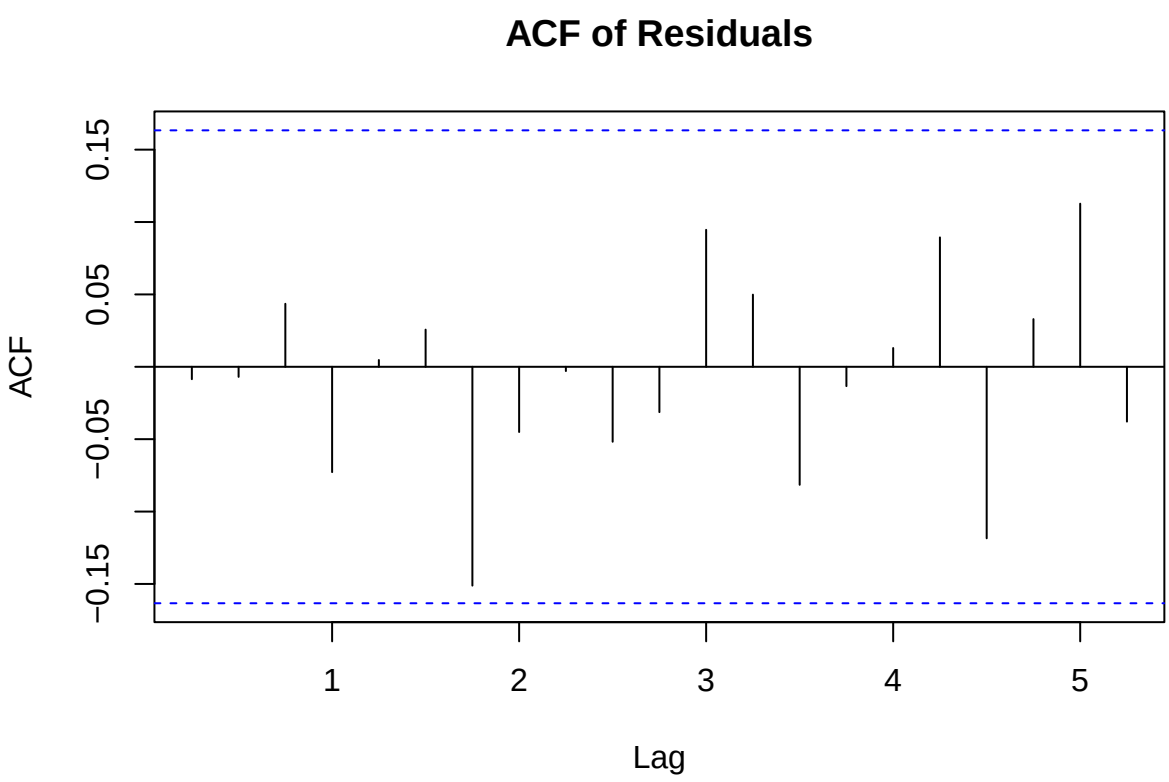
Model	Model Evaluation		
	AIC	AICc	BIC
SARIMA(0,1,4)(0,1,0)[4]	**310.0023**	**310.4534**	**324.6747**
SARIMA(4,1,1)(0,1,1)[4]	310.1994	311.0543	330.7407
SARIMA(4,1,1)(2,1,0)[4]	310.3112	311.4189	333.7870
SARIMA(0,1,4)(0,1,1)[4]	311.1276	311.7640	328.7345
SARIMA(0,1,4)(1,1,0)[4]	311.2849	311.9212	328.8917

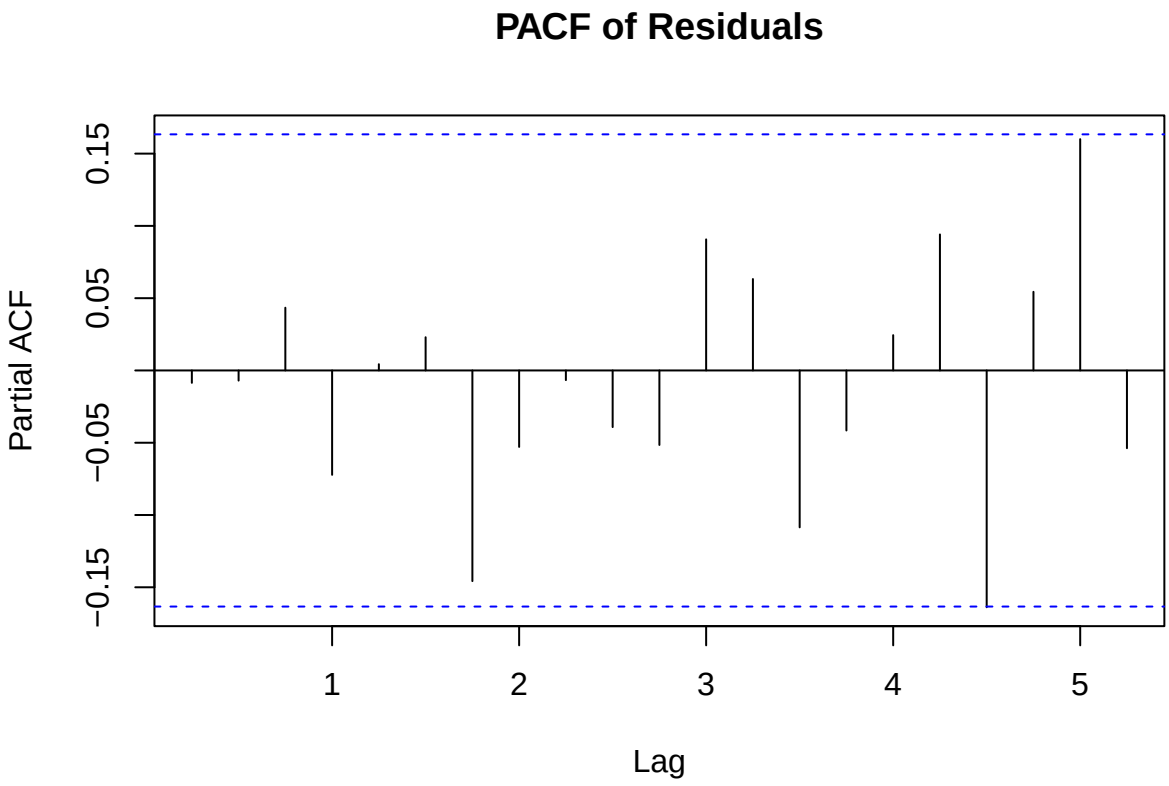
After attempting different combinations of (p,d,q)(P,D,Q), we find out that ARIMA(0,1,4)(0,1,0)[4] stands out with the lowest AIC, and AICc values among all the models.

Model Diagnostic



- The diagnostic plots for the ARIMA(0,1,4)(0,1,0)[4] model provide a comprehensive overview of the model's residuals to ensure that the model adequately captures the data's structure. The first plot shows the standardized residuals over time. Ideally, these residuals should oscillate around zero without any discernible pattern, indicating that they are white noise. In this case, the residuals do not exhibit any clear trend or seasonality, suggesting that the model has effectively captured the main patterns in the data.
- The Normal Q-Q plot evaluates whether the residuals are normally distributed. The points closely follow the diagonal line, which indicates that the residuals are approximately normally distributed, with only a few deviations at the tails. This is a positive sign as it suggests that the assumption of normality is reasonably satisfied.
- Finally, the Ljung-Box test p-values plot indicates that the residuals of the model do not exhibit significant autocorrelation at any lag up to 20. This absence of significant autocorrelation in the residuals supports the conclusion that the model has successfully captured the temporal dependencies in the data, and the residuals behave as white noise. Therefore, the ARIMA(0,1,4)(0,1,0)[4] model can be considered well-fitted and appropriate for the given time series data, providing confidence in its suitability for forecasting or further analysis.


ACF of Residuals
ACF
0.15
0.05
−0.05
−0.15
1
2
3
4
5
Lag

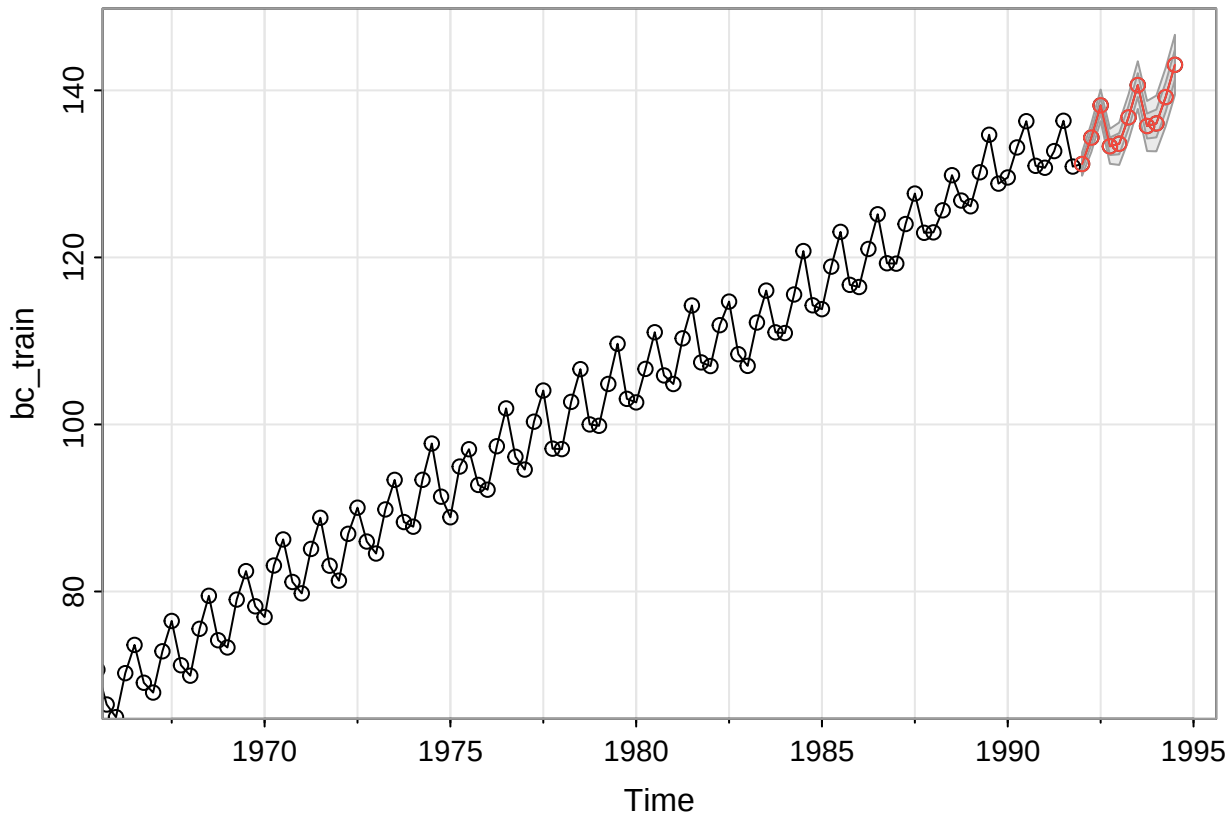


The ACF and PACF plots of the residuals provide a more detailed examination of the residuals:

- The ACF plot for the residuals displays the autocorrelations at various lags. Notably, most of the autocorrelations fall within the confidence intervals, and there are no significant spikes, suggesting that there is no substantial leftover autocorrelation that the model has failed to capture. This implies that the residuals behave like white noise, supporting the adequacy of the model.
- The PACF plot reveals the partial autocorrelation of the residuals. For a well-fitted model, these values should be close to zero, indicating no remaining pattern. In this plot, most of the partial autocorrelations are within the confidence bounds, with only one somewhat significant spikes between lag 4 and 5. It does not undermine the overall model fit, as the majority of the values remain within the expected range.

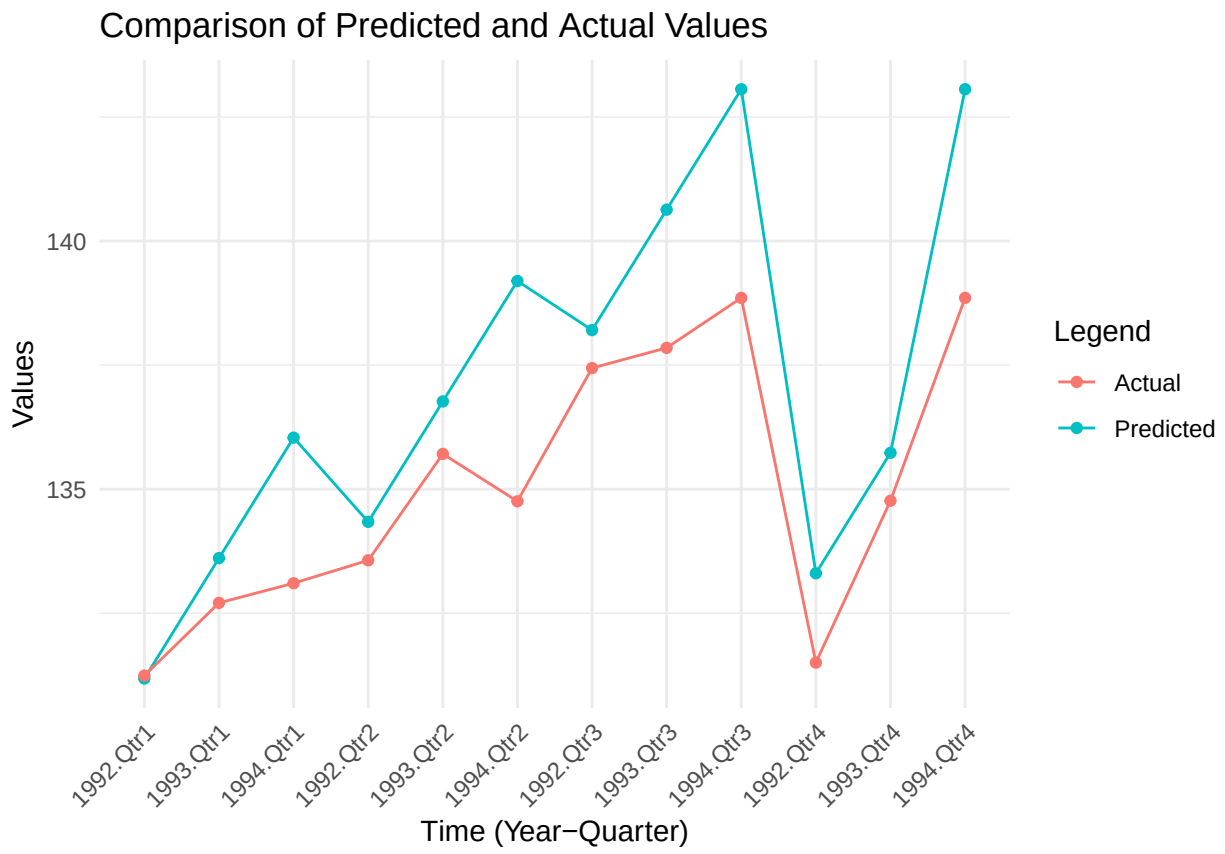
Together, the absence of significant autocorrelations in the ACF plot and the PACF plot confirm that the residuals of our ARIMA(0,1,4)(0,1,0)[4] model are effectively uncorrelated and do not exhibit any systematic pattern. This reaffirms the model's adequacy and stability for forecasting purposes, demonstrating that it has successfully captured the underlying temporal structure of the data.

Forecasting




```
## $pred
##          Qtr1     Qtr2     Qtr3     Qtr4
## 1992 131.1829 134.3402 138.2050 133.3043
## 1993 133.6103 136.7676 140.6324 135.7317
## 1994 136.0378 139.1950 143.0598
##
## $se
##           Qtr1      Qtr2      Qtr3      Qtr4
## 1992 0.7023587 0.8631967 0.9447535 1.0469701
## 1993 1.2685329 1.3692808 1.4258088 1.4996277
## 1994 1.6678232 1.7496691 1.7971902
```

The prediction plot illustrates the historical and forecasted quarterly electricity production in Australia, measured in million kilowatt-hours from March 1956 to September 1994. The black line represents the historical data, showing a clear upward trend with seasonal fluctuations. The model's forecast, depicted by the red line, extends this trend into the future for the next 11 quarters. The confidence intervals around these predictions widen over time, indicating increasing uncertainty as the forecast extends further into the future. Initially, the model's short-term predictions align closely with recent data, suggesting high reliability. Overall, the forecast anticipates continued growth in electricity production, capturing both the trend and seasonality observed in the past, providing essential insights for planning and managing Australia's energy supply.

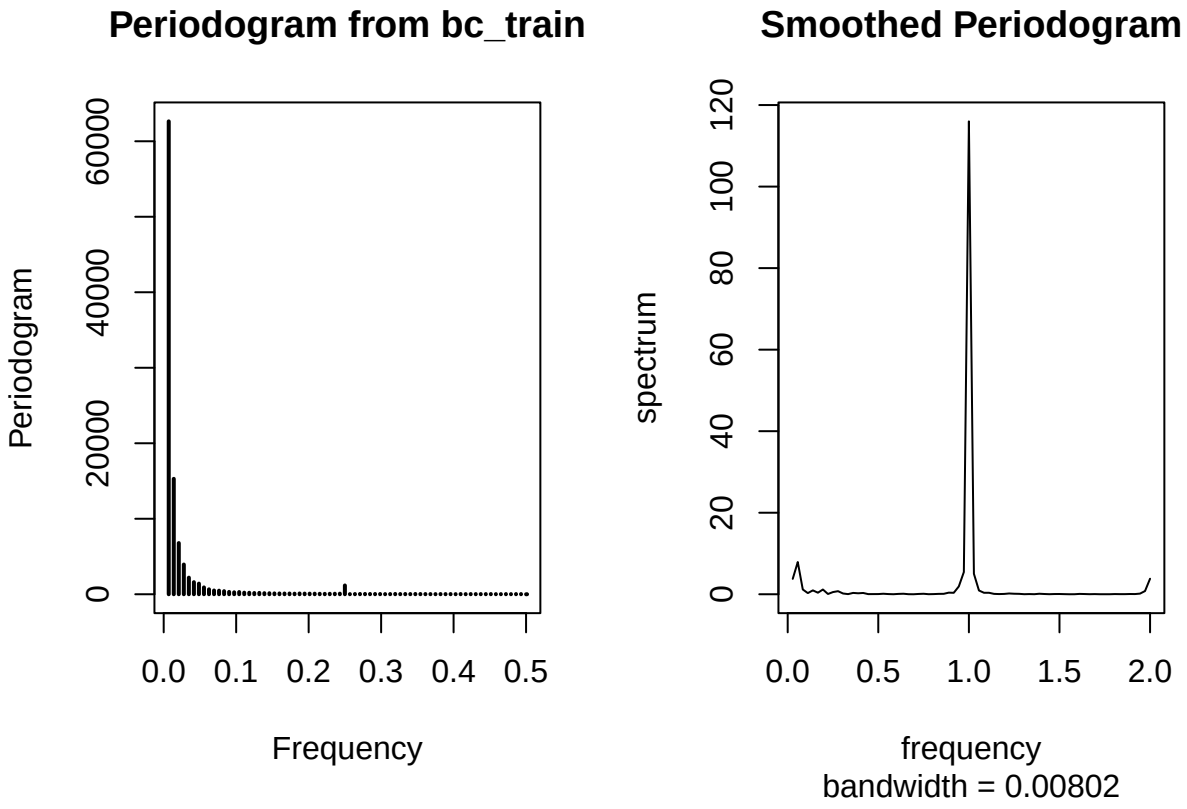


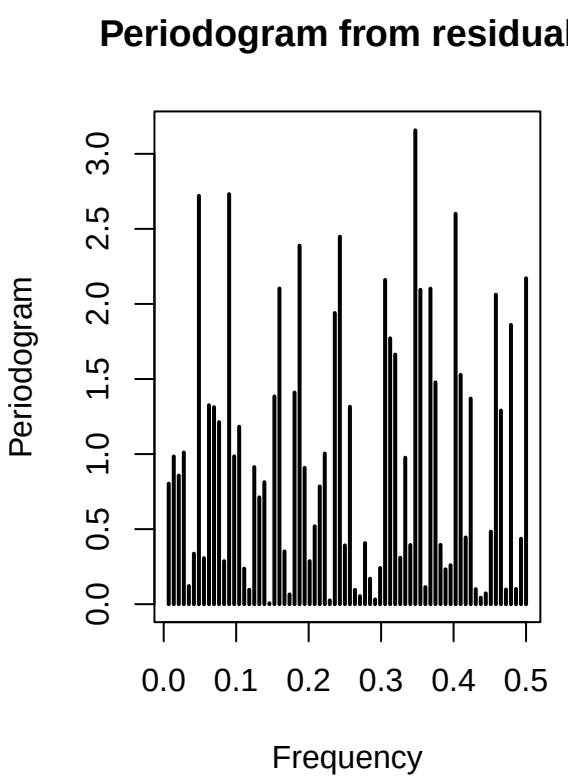
While the predicted values capture the general upward trend in electricity production, they tend to overestimate the peaks and underestimate the troughs. Besides, The model captures the seasonal fluctuations reasonably well but sometimes with amplified magnitude. Additionally, The differences between the actual and predicted values highlight areas where the model could be improved. Notably, the model's predictions are less accurate during periods of sharp changes in production.

Overall, this plot reveals that the optimal model effectively captures the overall trend and seasonal patterns in Australia's electricity production, but it tends to exaggerate the magnitude of fluctuations, especially during periods of rapid change. This indicates that while the model is robust for general trend prediction, further refinement may be necessary to improve its accuracy in capturing sudden shifts in production levels.

4.2 Spectral Analysis

Many time series show periodic behavior, like my time series. This periodic behavior can be very complex. Spectral analysis is a technique that allows us to discover underlying periodicities. The spectral analysis of our Box-Cox transformed time series data is represented through various plots and statistical tests.

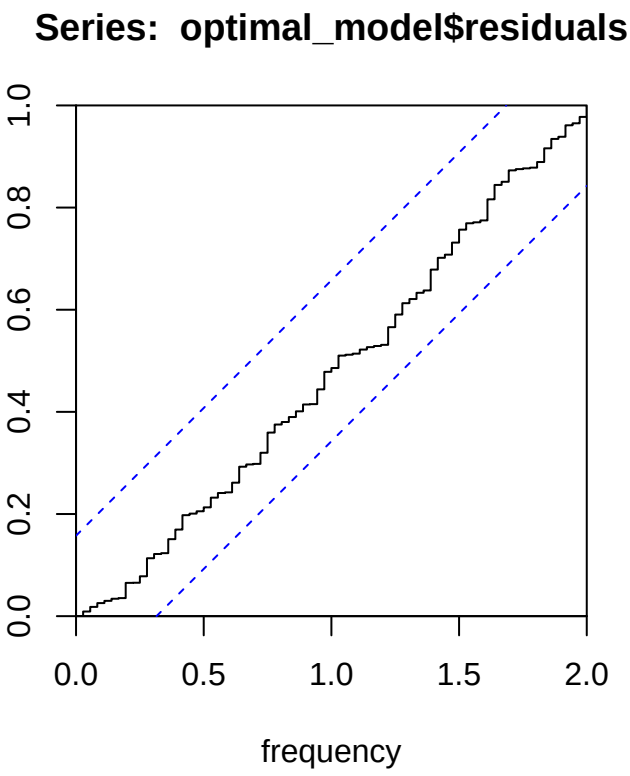

Periodogram from bc_train
Periodogram
0
20000
40000
60000
0.0
0.1
0.2
0.3
0.4
0.5
Frequency
Smoothed Periodogram
spectrum
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
frequency
bandwidth = 0.00802



The periodogram of the Box-Cox transformed series (`bc_train`) reveals a significant spike at low frequencies, particularly around 0.25, which corresponds to the quarterly seasonality (every 4 quarters) observed previously. This indicates the presence of a strong seasonal component in the data. The smoothed periodogram for the Box-Cox transformed time series reveals a pronounced peak at the frequency of around 1.0, indicating a strong quarterly seasonal cycle. This significant peak underscores the dominance of seasonal variations in the dataset, with a substantial portion of the series' variance explained by this regular pattern. The high spectral density at this frequency confirms the importance of quarterly cycles, aligning with the observed seasonality from the SARIMA model.

The periodogram of the residuals provides insight into the frequency domain characteristics of the residuals from our fitted model. In this plot, we observe that there are no significant dominant spikes across the range of frequencies. This absence of pronounced peaks indicates that there is no residual periodicity left in the data that the model has failed to capture. Essentially, the residuals appear to be distributed across a wide range of frequencies without any noticeable periodic structure. This suggests that the model has effectively accounted for the major cycles and seasonal components in the original time series data, leaving behind residuals that behave in a manner consistent with white noise. This supports the adequacy of the model in explaining the time series' behavior without leaving any unmodeled cyclical patterns.

A Fisher's test for periodicity was conducted and found to have a p-value of 0.7624137 so we fail to reject the null hypothesis that the model's residuals are Gaussian white noise, and conclude that our findings suggest that the residuals are in the form of white noise. This conclusion was checked against the findings of the Kolmogorov-Smirnov test, which showed that the cumulative periodogram was within the confidence interval for all of its values (Figure 13), verifying the results of the Fisher's test.


Series: optimal_model$residuals
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
frequency

5. Conclusion and Future Study

In this project, we conducted a comprehensive analysis and forecasting of Australia's quarterly electricity production from March 1956 to September 1994. Our primary goal was to model the intricate patterns in the time series data and generate accurate forecasts that could inform future energy planning and management. By leveraging two robust methodologies—SARIMA modeling and Spectral Analysis—we gained deep insights into the data's temporal dynamics and seasonal behavior.

The SARIMA model allowed us to effectively capture both the seasonal and non-seasonal components of the time series. Our analysis revealed strong quarterly cycles and a clear upward trend in electricity production over the study period. Diagnostic checks confirmed the model's adequacy, showing that the residuals exhibited white noise characteristics, indicating that the model had successfully captured the underlying structure of the data.

Complementing the SARIMA model, Spectral Analysis provided a frequency domain perspective of the time series. The periodogram identified a dominant peak corresponding to the quarterly seasonality, which was further corroborated by the SARIMA findings. This approach confirmed the significance of periodic components in the data and highlighted the strength of quarterly cycles in driving electricity production patterns.

Our study underscores the critical role of accurate modeling and forecasting in managing energy resources. The insights derived from our analysis are vital for anticipating future energy demands and planning sustainable growth in Australia's electricity sector.

For future research, several avenues could be explored to enhance the forecasting accuracy and depth of analysis. Integrating external factors, such as economic indicators or climate variables, into the models could provide a more holistic view of the influences on electricity production. Additionally, employing advanced machine learning techniques, such as deep learning or ensemble models, could capture complex patterns that traditional time series methods might overlook. Investigating the impact of policy changes and technological advancements on electricity production could also offer valuable insights for long-term energy planning.

Overall, our project lays a solid foundation for understanding and predicting electricity production trends, offering critical insights that can aid in effective energy management and planning.

6. Reference

Hyndman, R.J. “Time Series Data Library”, https://datamarket.com/data/list/?q=provider:tsdl.

ChatGPT4.0

7. Code appendix

```
electricity_p <- tsdl[[123]]
head(electricity_p)

# Split Train and TEST
train <- head(electricity_p, n=length(electricity_p) - 11)
test <- tail(electricity_p, n = 11)

# Make a trend line
trend <- rollmean(train, k=4, fill=NA, align='right')

plotdata <- data.frame(
  "Years" = seq(as.Date("1956/01/01"), by = "quarter", length.out = length(train)),
  "data" = train,
  "trend" = trend
)

# Plot the time series
p <- plotdata %>%
  tidyr::gather("id", "value", 2:3) %>%
  ggplot(aes(Years, value, col=id, linetype=id)) +
  geom_line() +
  labs(x = "Year", y = "Electricity Production (million kWh)",
       title = "Quarterly Electricity Production in Australia (1956-1994)",
       color = "Legend", linetype = "Legend") +
  scale_color_brewer(palette="Dark2")

print(p)

# Box Cox plot for optimal lambda
bcTransform = boxcox(train~as.numeric(1:length(train)), plotit = TRUE)
bcTransform$x[which(bcTransform$y == max(bcTransform$y))]

#apply box cox transformation
lambda <- 0.464

bc_train <- (train^lambda-1/lambda)
bc_test <- (test^lambda-1/lambda)
bc_fc <- (electricity_p^lambda-1/lambda)

library(grid)
# Seasonal plot
seasplot <- ggseasonplot(bc_train, year.labels=TRUE, year.labels.left=TRUE) +
  ylab("Box-Cox Adjusted Quarterly Electricity Production") +
  ggtitle("Seasonal plot") +
```

```r
  theme_light()

# Trend plot
Trend <- rollmean(bc_train, k=4, fill=NA, align='right')
trendplot <- autoplot(Trend, ts.color="red") +
  ggtitle("Trend plot") +
  theme_light()

# Combine plots and add bottom caption
combined_plot <- grid.arrange(
  seasplot, trendplot,
  layout_matrix = rbind(c(1, 2))
  # bottom = textGrob(
  #   "Figure 4: Seasonal and Trend Plot of BC transformed data",
  #   gp = gpar(fontsize = 10)
  # )
)

# Differencing the data to remove trend
diff1 <- diff(bc_train, 1)
plot(diff1,main = "Differenced at Lag 1",
  ylab = expression(nabla~Y[t]))
abline(h = 0,lty = 3)
# Differencing the data to remove seasonality
diff4 <- diff(diff1,4)
ts.plot(diff4, main = "Differenced at Lag 1 and Lag 4",
  ylab = expression(nabla^{4}~nabla~Y[t]))
abline(h = 0,lty = 3)

# Get the variances of the original and differenced data to check if the variance is continually decrea
v1 <- var(bc_train)
print(paste("Variance of transformed series =", v1))

v2 <- var(diff1)
print(paste("Variance of first differenced series =", v2))

v3 <- var(diff4)
print(paste("Variance of second differenced series =", v3))

# repeating the second differencing and check the variance
diff5 <- diff(diff4, 4) # 12 since the time series is monthly data

v4 <- var(diff5)
print(paste("Variance of third differenced series =", v4))

par(mfrow=c(1,1))
qqnorm(diff4, main="QQ Plot of Differenced Time Series")
qqline(diff4, col = "red")

#Shapiro Test for normality of the data
shapiro.test(diff4)

adf.test(diff4)
```

```r
# ACF and PACF graphs of the differenced data
par(mfrow = c(1,1))
acf(diff4, lag.max = 20, main = "ACF of Twice Differenced Data")
pacf(diff4, lag.max = 20, main = "PACF of Twice Differenced Data")


# Placeholder to store the results
results <- data.frame(
  p = integer(),
  d = integer(),
  q = integer(),
  P = integer(),
  D = integer(),
  Q = integer(),
  AIC = numeric()
)

# Define the ranges for parameters
p_values <- 0:4
d_values <- 1  # Fixed based on your analysis
q_values <- 0:4
P_values <- 0:2
D_values <- 1  # Fixed based on your analysis
Q_values <- 0:1

# Iterate through all combinations of parameters
for (p in p_values) {
  for (q in q_values) {
    for (P in P_values) {
      for (Q in Q_values) {
        # Try fitting the model
        try({
          fit <- Arima(bc_train, order = c(p, d_values, q),
                       seasonal = c(P, D_values, Q),
                       include.constant = FALSE,
                       method = "ML")

          # Store the AIC and parameters
          results <- rbind(results, data.frame(
            p = p, d = d_values, q = q,
            P = P, D = D_values, Q = Q,
            AIC = fit$aic,
            AICc = fit$aicc,
            BIC = fit$bic
          ))
        }, silent = TRUE) # Continue even if the model fitting fails
      }
    }
  }
}

# Find the model with the lowest AIC
best_model <- results[which.min(results$AIC),]
```

```r
# Print the best model parameters and its AIC
print(best_model)

# Fit the best model
optimal_model <- Arima(bc_train, order = c(best_model$p, best_model$d, best_model$q),
                       seasonal = c(best_model$P, best_model$D, best_model$Q),
                       include.constant = FALSE,
                       method = "ML")

# Print the summary of the best model
summary(optimal_model)

library(knitr)
library(kableExtra)

# Step 1: Prepare the data based on the new models and criteria from the screenshot
models <- c("SARIMA(0,1,4)(0,1,0)[4]",
            "SARIMA(4,1,1)(0,1,1)[4]",
            "SARIMA(4,1,1)(2,1,0)[4]",
            "SARIMA(0,1,4)(0,1,1)[4]",
            "SARIMA(0,1,4)(1,1,0)[4]")

aic <- c(310.0023, 310.1994, 310.3112, 311.1276, 311.2849)
aicc <- c(310.4534, 311.0543, 311.4189, 311.7640, 311.9212)
bic <- c(324.6747, 330.7407, 333.7870, 328.7345, 328.8917)

# Step 2: Create a data frame
results_df <- data.frame(
  Model = models,
  AIC = aic,
  AICc = aicc,
  BIC = bic
)

# Step 3: Generate the table using kable and kableExtra for LaTeX
kable(results_df, format = "latex", escape = FALSE, align = "c", booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position"), full_width = FALSE) %>%
  column_spec(1, bold = TRUE) %>%  # Make model names bold
  row_spec(which.min(results_df$AIC), bold = TRUE, color = "red", background = "#ffcccc") %>%  # Highli
  add_header_above(c(" " = 1, "Model Evaluation" = 3))


suppress_sarima <- function(data, p, d, q, P, D, Q, S) {
  invisible(capture.output(sarima(data, p, d, q, P, D, Q, S)))
}

# Fit the SARIMA model and generate diagnostic plots without verbose output
suppress_sarima(bc_train, p = 0, d = 1, q = 4,  # Non-seasonal parameters
                P = 0, D = 1, Q = 0, S = 4)    # Seasonal parameters with period S = 4

acf(residuals(optimal_model), main = "ACF of Residuals")
pacf(residuals(optimal_model), main = "PACF of Residuals")
```

```r
sarima.for(bc_train, n.ahead = 11,  p = 0, d = 1, q = 4,  # Non-seasonal parameters
               P = 0, D = 1, Q = 0, S = 4)

library(dplyr)

# Predicted values from the model (from the first screenshot)
predicted_values <- tibble::tibble(
  Year = rep(1992:1994, each = 4),
  Quarter = rep(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), times = 3),
  Predicted = c(131.1829, 134.3402, 138.2050, 133.3043,
                133.6103, 136.7676, 140.6324, 135.7317,
                136.0378, 139.1950, 143.0598, 143.0598)
)

# Actual validation values (from the second screenshot)
actual_values <- tibble::tibble(
  Year = rep(1992:1994, each = 4),
  Quarter = rep(c("Qtr1", "Qtr2", "Qtr3", "Qtr4"), times = 3),
  Actual = c(131.2421, 133.5660, 137.4401, 131.5038,
             132.7076, 135.7101, 137.8472, 134.7662,
             133.1039, 134.7551, 138.8544, 138.8544)
)

# Combine predicted and actual values
comparison <- left_join(predicted_values, actual_values, by = c("Year", "Quarter"))

# Calculate the difference
comparison <- comparison %>%
  mutate(Difference = Predicted - Actual)

# Create a plot to compare predicted vs actual values
ggplot(comparison, aes(x = interaction(Year, Quarter), group = 1)) +
  geom_line(aes(y = Predicted, color = "Predicted")) +
  geom_line(aes(y = Actual, color = "Actual")) +
  geom_point(aes(y = Predicted, color = "Predicted")) +
  geom_point(aes(y = Actual, color = "Actual")) +
  labs(title = "Comparison of Predicted and Actual Values",
       x = "Time (Year-Quarter)",
       y = "Values",
       color = "Legend") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Spectral Analysis
par(mfrow=c(1,2))
periodogram(bc_train, main="Periodogram from bc_train")

spectrum(bc_train, main = "Smoothed Periodogram", log = "no", taper = 0.1)

periodogram(optimal_model$residuals, main="Periodogram from residuals")

fisher.g.test(optimal_model$residuals)
```

```r
# cumulative periodogram
cpgram(optimal_model$residuals)
```