TUNIS BUSINESS SCHOOL
UNIVERSITY OF TUNIS

Time Series Analysis Project Report

Presented By:
Ines Kamoun
Emna Ameur

(Senior Finance/BA)

Submission Date: 03/01/2024

I. Introduction

The report contains the implementation of the **Box-Jenkins Method** on the “CPILFESL” (*Consumer Price Index for All Urban Consumers: All Items Less Food and Energy in U.S. City Average*) dataset which includes a four-stage process as follows:

1. Model Specification
2. Parameter Estimation
3. Forecasting
4. Model Checking

Before testing and working on the dataset, we checked for missing values.

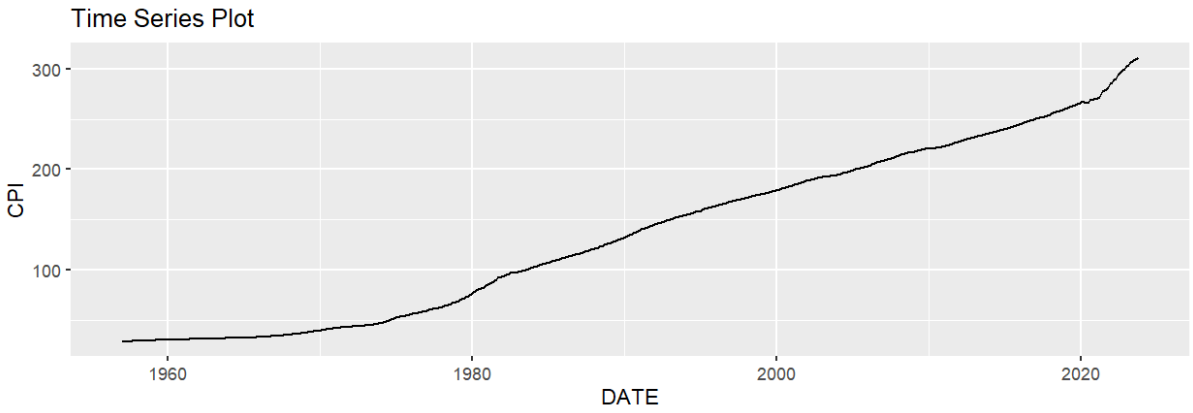
```
# Check for missing values in the "CPI" column
any(is.na(CPILFESL$CPI))
```

The results showed “FALSE” indicating that no missing values were found:

```
> # Check for missing values in the "CPI" column
> any(is.na(CPILFESL$CPI))
[1] FALSE
```

II. Model Identification

1. Checking Stationarity and Seasonality



◆ The initial data’s plot suggests non-constant mean with time-dependent variance:

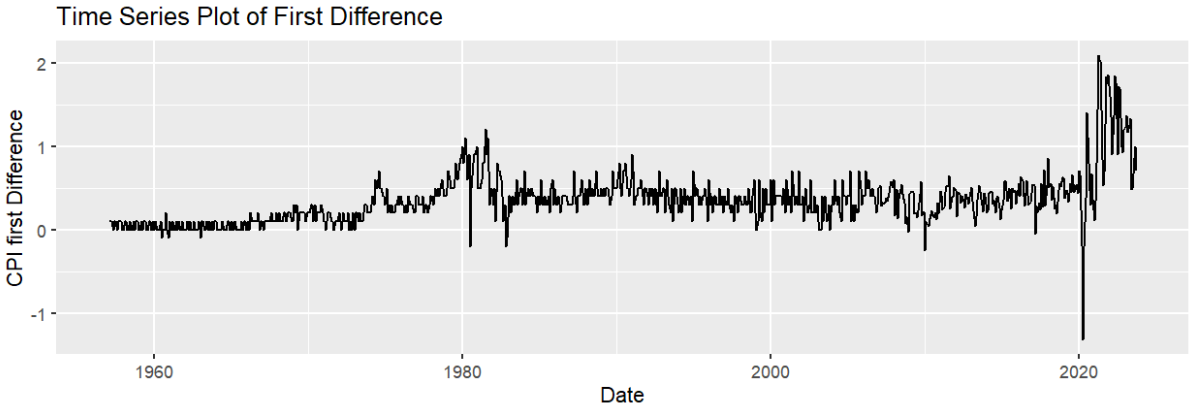
◆ The graph displays low CPI during the first period but starting from 1980 it shows a significant increase

➔ Our initial data is not stationary.

➔ It is essential to transform it into a stationary form prior to analysis using Differentiation method.

2) Differentiation:

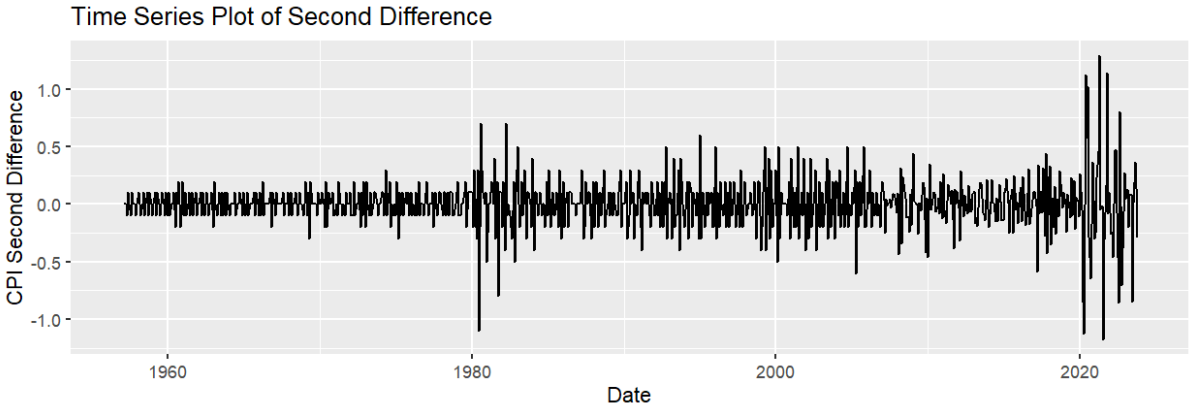
a) First difference:



The presence of increasing volatility in the data around and after 2020 could suggest non-stationarity. While the first difference operation can often help to make a time series stationary, it may not always be sufficient, especially if there are changes in volatility over time (a characteristic of heteroscedasticity) or if the mean isn't constant.

In the case of this time series, the increasing fluctuations around and after 2020 suggest that the variance is not constant over time, which is a violation of the stationarity assumption. This could be due to various factors such as changes in the underlying process generating the data or external factors impacting the variable.

b) Second difference:



this graph shows the second difference of the CPI changes over time. The plot suggests that the variable might be stationary for the period from 1960 to just before 2020, but there is increased volatility around and after 2020.
Implementing the necessary stationarity tests leads to a more sure conclusion about our time series' stationarity:

- **unit root tests:**

```
        Augmented Dickey-Fuller Test

data:  CPILFESL$CPI_diff2
Dickey-Fuller = -13.045, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

```
        Phillips-Perron Unit Root Test

data:  CPILFESL$CPI_diff2
Dickey-Fuller Z(alpha) = -758.51, Truncation lag
parameter = 6, p-value = 0.01
alternative hypothesis: stationary
```

For both tests, ADF and PP we have:

H0: One unit root (i.e. Non Stationarity)

H1: Zero unit root (i.e. Stationarity)

➢ The results show the two p-values= 0.01< 0.05

➢ We reject H0

⇒ the time series is **stationary**

-Stationarity test:

➢ KPSS, on the other hand, tests for stationarity around a deterministic trend. If it suggests non-stationarity, it means it failed to reject the null hypothesis that the series is trend-stationary as the p-value is < 0.05

```
data:  CPILFESL$CPI_diff2
KPSS Trend = 0.0091292, Truncation lag parameter = 6,
p-value = 0.1
```

H0: One unit root (i.e. Stationarity)

H1: Zero unit root (i.e. Non-Stationarity)

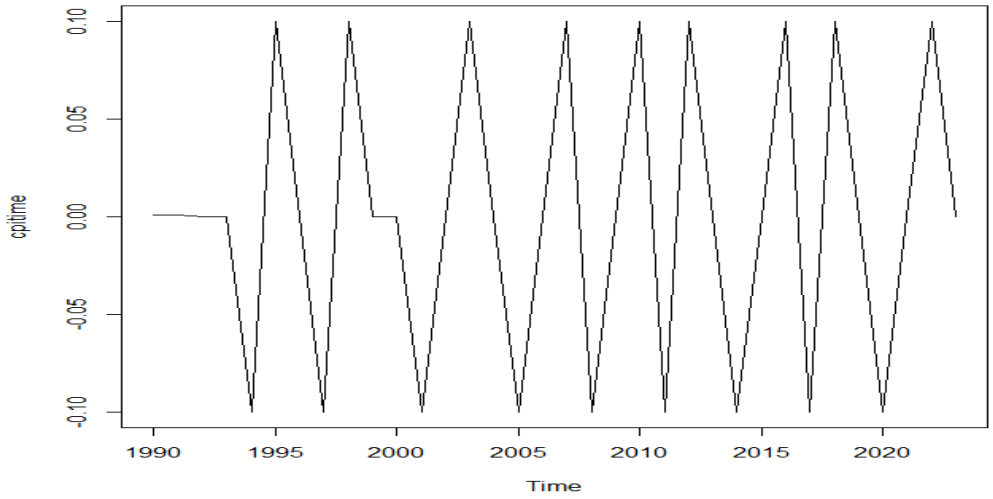
➢ The result shows that the p-value = 0.1 > 0.05

➢ We fail to reject H0

⇒ the time series is **stationary**

➢ Since ADF, PP and KPSS indicate stationarity, we conclude that the series is **stationary**, and we move forward with our analysis.

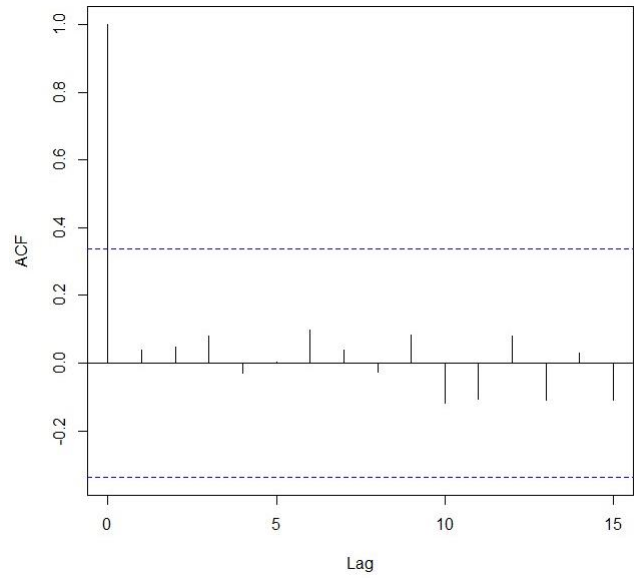
3. choosing model specification:



Graph: The plot of the new stationarity over time after differencing the data

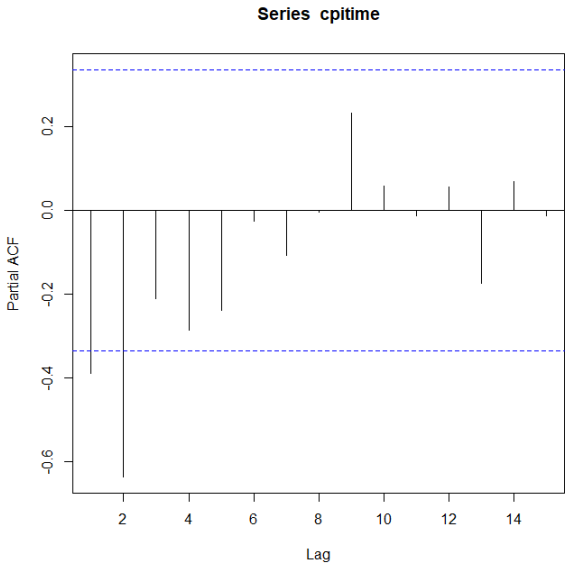
- Using the Confidence interval = [-1.96/√802;+1.96/√802] we plot:

1) ACF:



- The ACF plot cuts off after lag 1, the remaining autocorrelations fall within the interval (dashed lines).
  => We can assume that the model MA(1) is appropriate.
  ⇒ It's obvious from the ACF plot that the time series is stationary

2) PACF:



● The PACF plot shows two significant spikes at lag 1 , 2 and all the other spikes fall within the interval (dashed lines)

⇒ **We can assume that the model AR(2) is appropriate.**

⇒ It's obvious from the ACF plot that the time series is stationary

III. Parameter Estimation

Code:

```
#Parameter estimation
cpimodel=auto.arima(cpitime,ic="aic",trace = TRUE)
```

Output :

```
ARIMA(2,0,2) with non-zero mean : -102.8039
ARIMA(0,0,0) with non-zero mean : -77.71138
ARIMA(1,0,0) with non-zero mean : -81.12737
ARIMA(0,0,1) with non-zero mean : Inf
ARIMA(0,0,0) with zero mean     : -79.71137
ARIMA(1,0,2) with non-zero mean : Inf
ARIMA(2,0,1) with non-zero mean : -104.8019
ARIMA(1,0,1) with non-zero mean : Inf
ARIMA(2,0,0) with non-zero mean : -97.33322
ARIMA(3,0,1) with non-zero mean : -102.8024
ARIMA(3,0,0) with non-zero mean : -98.08039
ARIMA(3,0,2) with non-zero mean : -100.9527
ARIMA(2,0,1) with zero mean     : -105.5679
ARIMA(1,0,1) with zero mean     : -98.57845
ARIMA(2,0,0) with zero mean     : -99.2862
ARIMA(3,0,1) with zero mean     : -103.568
ARIMA(2,0,2) with zero mean     : -103.6526
ARIMA(1,0,0) with zero mean     : -83.12735
ARIMA(1,0,2) with zero mean     : -97.8171
ARIMA(3,0,0) with zero mean     : -99.91812
ARIMA(3,0,2) with zero mean     : -101.7062

Best model: ARIMA(2,0,1) with zero mean
```

⇨ The function auto.arima selected the best model ARIMA(2,0,1) with zero mean since it has the lowest AIC value -105.5679.

```
> cpimodel
Series: cpitime
ARIMA(2,0,1) with zero mean

Coefficients:
          ar1      ar2      ma1
      -0.3192  -0.5318  -0.7415
s.e.   0.1590   0.1508   0.1260

sigma^2 = 0.002123:  log likelihood = 56.78
AIC=-105.57   AICc=-104.19   BIC=-99.46
```

The results from the ARIMA model provide insights into the time series data that was analyzed. Here's what they tell us:

- **ARIMA Model**: The ARIMA(2,0,1) model suggests that the time series data has some form of autocorrelation, as it uses past values (order 2 autoregression) and past forecast errors (order 1 moving average) to model the data.
- **Coefficients**: The coefficients of the AR and MA terms (-0.3192 for AR1, -0.5318 for AR2, and -0.7415 for MA1) give us the weights of these terms in the model. The negative signs might suggest some form of oscillation in the data.
- **Sigma^2**: The low sigma^2 value (0.002123) suggests that the residuals from this model are quite small, indicating a good fit to the data.
- **Log Likelihood**: The log-likelihood value (56.78) is used for comparing different models. The higher the log-likelihood, the better the model fits the data.
- **AIC, AICc, BIC**: The AIC, AICc, and BIC are all measures of the relative quality of a statistical model. When comparing models, the model with the lower AIC, AICc, or BIC is typically the better. In this case, the values are -105.57 (AIC), -104.19 (AICc), and -99.46 (BIC).

=>, these results suggest that the ARIMA(2,0,1) model fits the time series data well and it is a good model for forecasting future values.

IV. Forecast

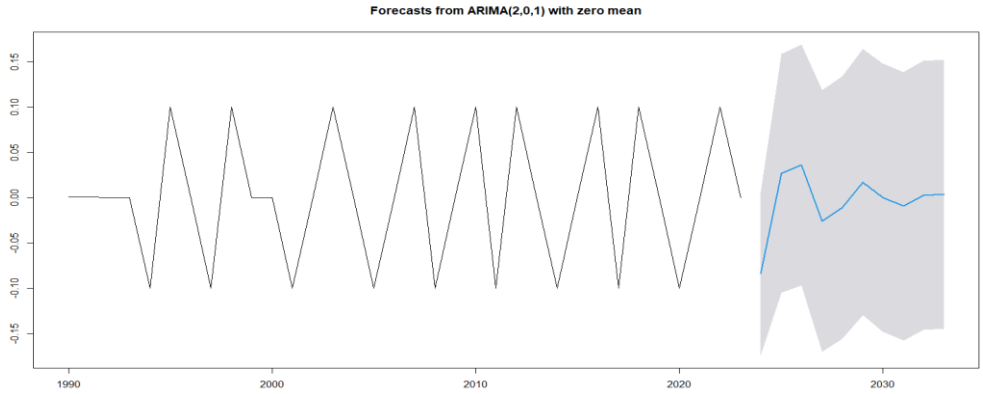
```
#forecast part
mycpiforecast=forecast(cpimodel,level = c(95),h=10*1)
mycpiforecast
plot(mycpiforecast)
```

```
> #forecast part
> mycpiforecast=forecast(cpimodel,level = c(95),h=10*1)
> mycpiforecast
     Point Forecast       Lo 95       Hi 95
2024  -0.0837742383 -0.17407613 0.006527652
2025   0.0267448309 -0.10489531 0.158384973
2026   0.0360118828 -0.09677879 0.168802553
2027  -0.0257193348 -0.17003233 0.118593663
2028  -0.0109398209 -0.15551867 0.133639030
2029   0.0171697607 -0.12995492 0.164294439
2030   0.0003362394 -0.14739382 0.148066296
2031  -0.0092380178 -0.15732162 0.138845582
2032   0.0027704186 -0.14567581 0.151216649
2033   0.0040282146 -0.14443327 0.152489694
```

- **Forecast**: The forecast is for the years 2024 to 2033. Each year has a point forecast and a 95% confidence interval (Lo 95 and Hi 95). The point forecasts are close to zero, indicating minimal change or growth expected in each respective year.
- Confidence Interval: The Lo 95 and Hi 95 columns represent the lower and upper bounds of the 95% confidence interval for each forecasted year. This interval represents the range in which we expect the actual value to fall 95% of the time. The wider the interval, the greater the uncertainty in the forecast.

⇨ These results suggest that the forecasted values for the next ten years are expected to be close to zero, with some degree of uncertainty as indicated by the 95% confidence intervals. The negative point forecasts might suggest a decrease, while the positive ones suggest an increase. However, the close proximity of these forecasts to zero might indicate that the changes are relatively small.

Here’s the forecast plot:



V. Model Checking

We proceed to validate this forecast using the box test

```
> #validate forecast using box test
> Box.test(mycpiforecast$resid, lag=5, type= "Ljung-Box")

        Box-Ljung test

data:  mycpiforecast$resid
X-squared = 1.4836, df = 5, p-value = 0.915

> Box.test(mycpiforecast$resid, lag=15, type= "Ljung-Box")

        Box-Ljung test

data:  mycpiforecast$resid
X-squared = 6.6334, df = 15, p-value = 0.967

> Box.test(mycpiforecast$resid, lag=25, type= "Ljung-Box")

        Box-Ljung test

data:  mycpiforecast$resid
X-squared = 16.804, df = 25, p-value = 0.8888
```

Test Results: The test was performed three times with different lag values (5, 15, and 25). The p-values for all three tests are greater than 0.05 (0.915, 0.967, and 0.8888 respectively).

Interpretation: A p-value greater than 0.05 indicates that we fail to reject the null hypothesis of the Box-Ljung test, which states that the residuals are independently distributed. In other words, these results suggest that there is no significant autocorrelation in the residuals at lags 5, 15, and 25.

=> These results suggest that the residuals from our forecast model do not exhibit significant autocorrelation, which is a good sign as it indicates that the model has adequately captured the underlying patterns in the time series data.