University of West Florida

Time Series Analysis STA 6856 | Professor Dr.Tharindu De Alwis

Edwin Quijano

Saturday, November 22, 2025

Create rainfall index

1. (25 points) The Egyptian Ministry of Water Resources is concerned about the variability from year-to- year in the annual water flow of the Nile River. They believe that much of the variation is driven by rainfall in upstream regions, which is not directly observed in the Nile dataset, but can be approximated using satellite-based rainfall indices.

To evaluate this idea, a hydrology researcher constructs the following dataset:

Y_t = annual Nile river flow (built-in R dataset Nile)

X_t = rainfall index for the same year, created as:

```r
data(Nile)
rain <- jitter(Nile, amount = 30)
```

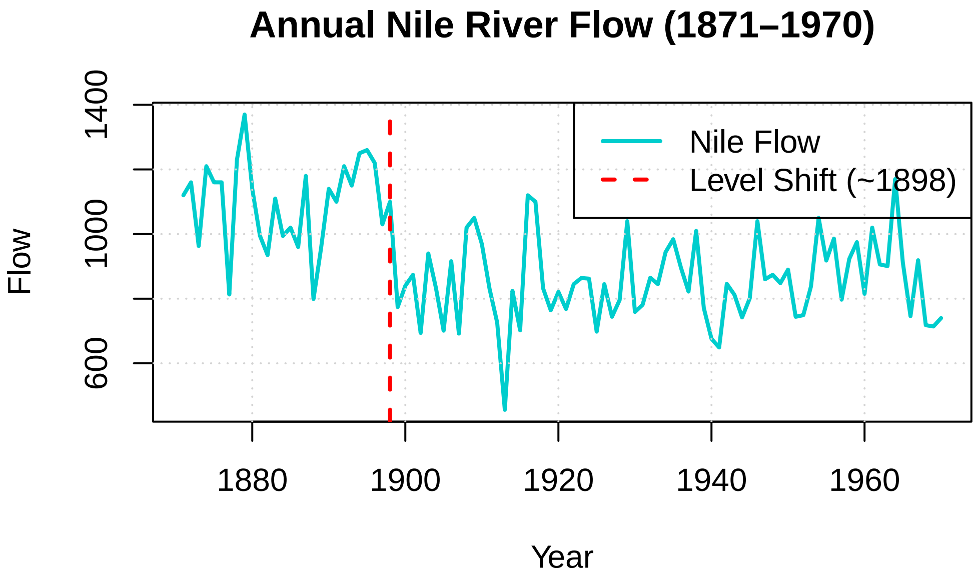
```r
head(Nile,10)
```

```
 [1] 1120 1160  963 1210 1160 1160  813 1230 1370 1140
```

(higher values indicate more rainfall in upstream areas) The Ministry wants to determine whether including the rainfall index improves forecasting of future water availability.

(a) (3 points) Plot the annual Nile flow series. Interpret the plot.

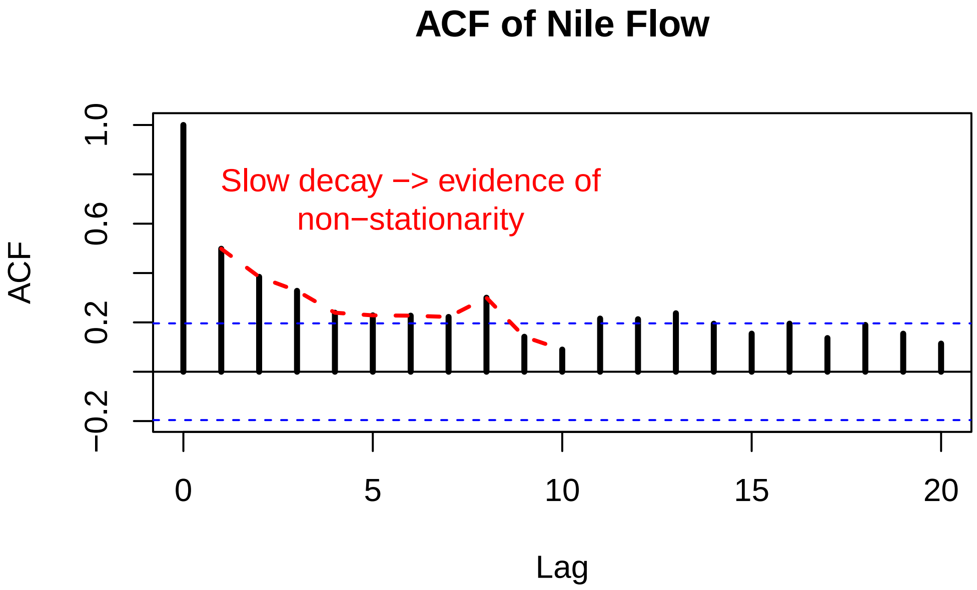
```r
plot(Nile,
     type = "l",
     main = "Annual Nile River Flow (1871-1970)",
     col ="cyan3",
     ylab = "Flow",
     xlab = "Year",
     lwd = 2)
grid()

abline(v = 1898, col = "red", lwd = 2, lty = 2)

legend("topright",
       legend = c("Nile Flow", "Level Shift (~1898)"),
       col = c("cyan3", "red"),
       lwd = 2,
       lty = c(1, 2))
```



The graph indicates a drop in the late 1890s, this is a level shift, and a strong indication of a non-stationary.

(b) (4 points) Is the Nile flow series stationary? Use plots and appropriate test for your argument.

```
acf_obj <- acf(Nile, plot = FALSE)
acf(Nile, main = "ACF of Nile Flow", lwd = 3)
lines(1:10, acf_obj$acf[2:11], col = "red", lwd = 2, lty=2)
text(x = 6, y = 0.7, "Slow decay → evidence of\nnon-stationarity", col = "red")
```



This slow decay pattern is characteristic level-shift non-stationarity, confirming that the Nile flow series is not stationary.

```
library(tseries)

adf_test <- adf.test(Nile)
adf_test
```

```

    Augmented Dickey-Fuller Test

data:  Nile
Dickey-Fuller = -3.3657, Lag order = 4, p-value = 0.0642
alternative hypothesis: stationary
```

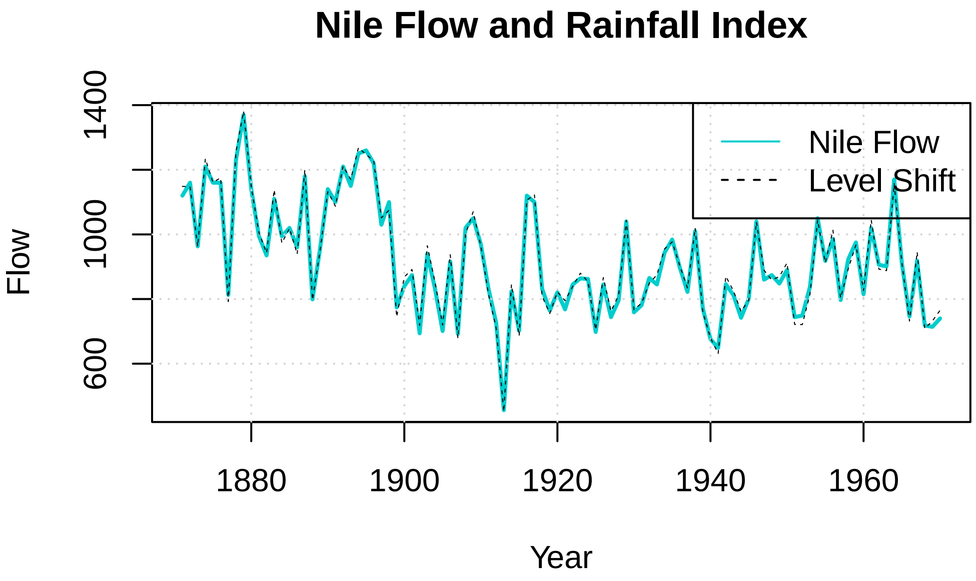
Since the p-value > 0.05, we fail to reject the null that the series has a unit root.

Evidence from the time-series plot (a clear level shift), the ACF plot (slow decay), and the ADF test (high p-value) all indicate that the Nile series is non-stationary.therefore, *the Nile series is NOT stationary* *(c) (5 points) Fit an ARIMAX model using the rainfall index X_t as an external regressor. Write down the estimated ARIMA (p, d, q) structure and the estimated regression coefficient for rainfall.*

```r
library(forecast)

rain <- jitter(Nile, amount = 30)
fit_arimax <- auto.arima(Nile, xreg = rain)
fit_arimax
```

```
Series: Nile
Regression with ARIMA(0,0,0) errors

Coefficients:
      intercept    xreg
        16.1635  0.9787
s.e.     9.0520  0.0096

sigma^2 = 278.2:  log likelihood = -422.3
AIC=850.61   AICc=850.86   BIC=858.42
```

```r
plot(Nile,
     type = "l",
     main = "Nile Flow and Rainfall Index",
     col ="cyan3",
     ylab = "Flow",
     xlab = "Year",
     lwd = 2)
grid()

lines(rain, col = "black", lwd = .5, lty=2)

legend("topright",
       legend = c("Nile Flow", "Level Shift"),
       col = c("cyan3", "black"),
       lwd = 1,
       lty = c(1, 2))
```



The estimated ARIMA structure, ARIMA(0,0,0), the rainfall index correlates with the Nile flow, therefore there is no autocorrelation to model.

Regression Coefficient : β=0.9787, which indicates a strong positive linear relationship between rainfall and river flow.

(d) (5 points) Fit a pure ARIMA model with no exogenous variables. Compare with the ARIMAX model from part (c). Which model is preferred for forecasting, and why?

```
fit_arima <- auto.arima(Nile)
fit_arima
```

```
Series: Nile
ARIMA(1,1,1)

Coefficients:
         ar1      ma1
      0.2544  -0.8741
s.e.  0.1194   0.0605

sigma^2 = 20177:  log likelihood = -630.63
AIC=1267.25   AICc=1267.51   BIC=1275.04
```

```
AIC(fit_arima); BIC(fit_arima)
```

```
[1] 1267.255
```

```
[1] 1275.04
```

```
AIC(fit_arimax); BIC(fit_arimax)
```

```
[1] 850.609
```

```
[1] 858.4245
```

ARIMAX with External regressor achieve a lower AIC/BIC and smaller residual variance than the pure ARIMA model. We need to point out that rainfall index is part of the Nile itself and should not be considered and exogenous variable.

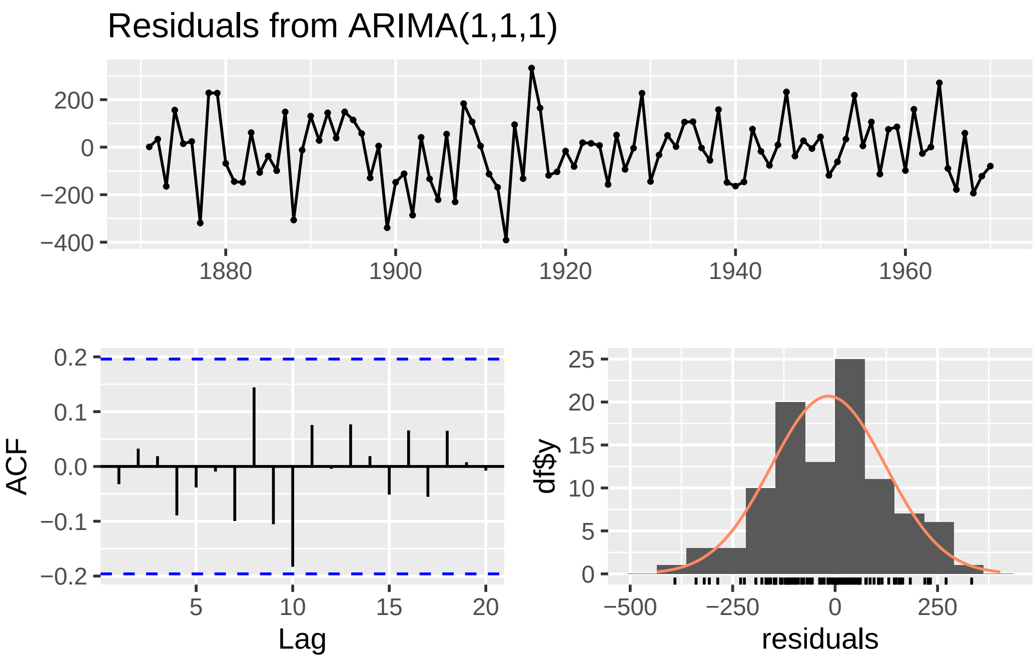
For forecasting purposes, pure ARIMA models could be used as a forecasting tool since it doesn't require rainfall values.

(e) (4 points) Interpret the rainfall coefficient. Does higher rainfall appear to increase Nile flow? Is the effect statistically significant?

The rainfall coefficient is estimated at 0.9787, indicating that a one-unit increase in the rainfall index is associated with about a one-unit increase in annual Nile flow.The higher rainfall leads to higher river flow. The effect is highly statistically significant (t   104), providing strong evidence that rainfall is an important predictor of Nile flow.

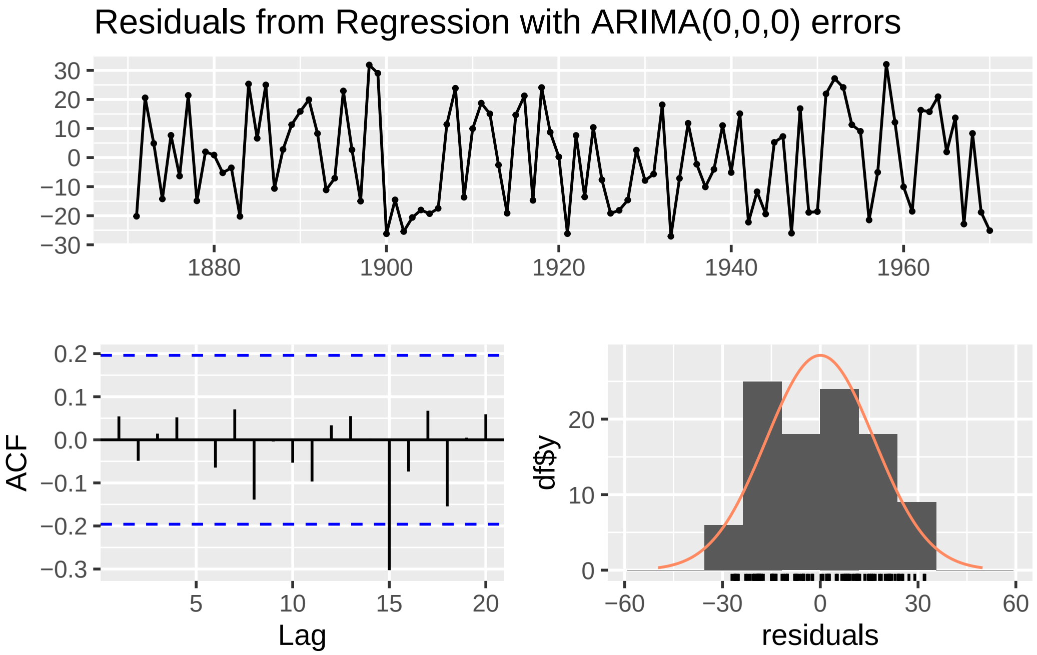
(f) (4 points) Based on residual diagnostics discuss whether including rainfall improves model adequacy. Should the Ministry use the ARIMAX model for forecasting next year's Nile flow? Explain your reasoning.

```r
checkresiduals(fit_arima)
```



```
    Ljung-Box test

data:  Residuals from ARIMA(1,1,1)
Q* = 9.7056, df = 8, p-value = 0.2863

Model df: 2.   Total lags used: 10
```

```r
checkresiduals(fit_arimax)
```



```
    Ljung-Box test

data:  Residuals from Regression with ARIMA(0,0,0) errors
Q* = 4.3203, df = 10, p-value = 0.9317

Model df: 0.   Total lags used: 10
```

ARIMAX model with rainfall provides a better in-sample fit than the pure ARIMA mode, the residuals have lower variance and show weaker autocorrelation. For forecasting next year's Nile flow, the Ministry would need future rainfall values. But Those are not known in advance, and the rainfall index cannot be use to forecast independently since it's based on Nile itself. therfore, the Ministry should not use this ARIMAX model for forecasting next year's Nile flow and instead rely on a pure ARIMA model or a model that uses genuinely forecastable exogenous variables.

2. (25 points) The following questions require analysis of the Canadian economic indicators: Gross Domestic Product (GDP), Interest rate (INT), Consumer Price Index (CPI), and Production (PRO). Please use the provided datasets to answers the followings.

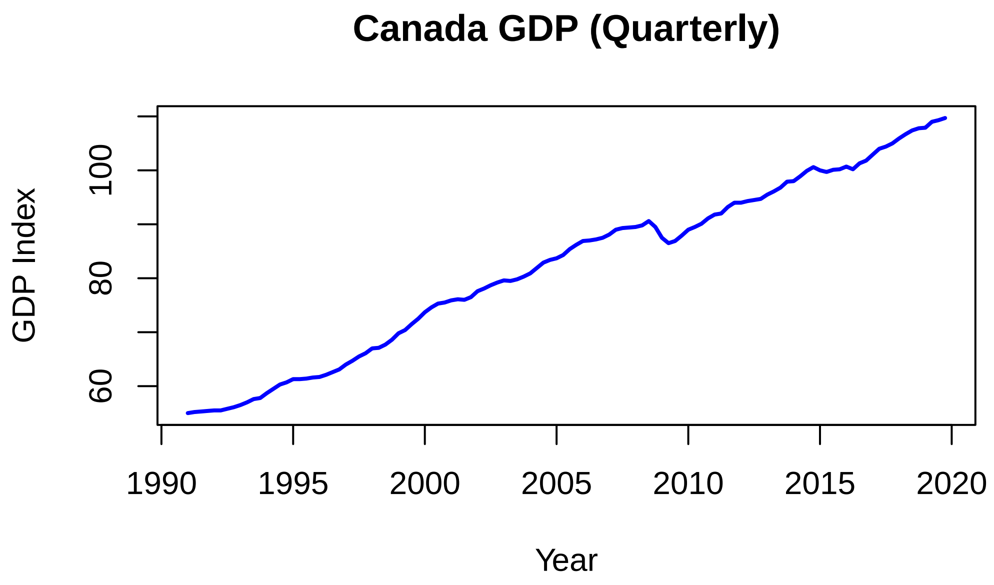
```r
library(tidyverse)
library(lubridate)
library(forecast)
library(tseries)
library(zoo)
```

(a) (2 points) Compute and interpret the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for each time series. Assess the stationarity of each series based on these analyses.

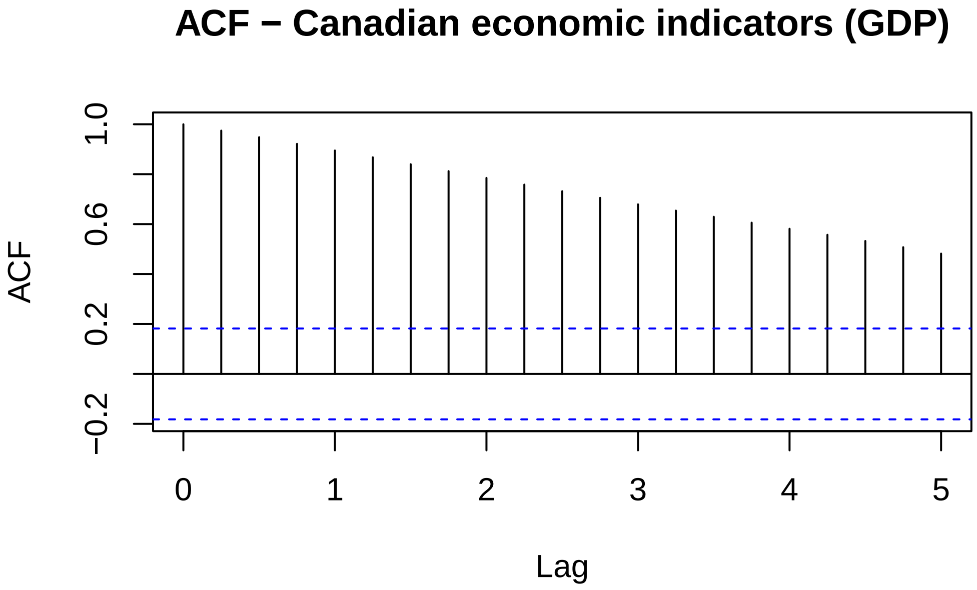
```
GDP <- read.csv("CAN_GDP.csv")
```

```
GDP$TIME <- as.yearqtr(GDP$TIME, format = "%Y-Q%q")
GDP$DATE <- as.Date(GDP$TIME)

gdp_ts <- ts(GDP$Value, start = c(1991, 1), frequency = 4)
```

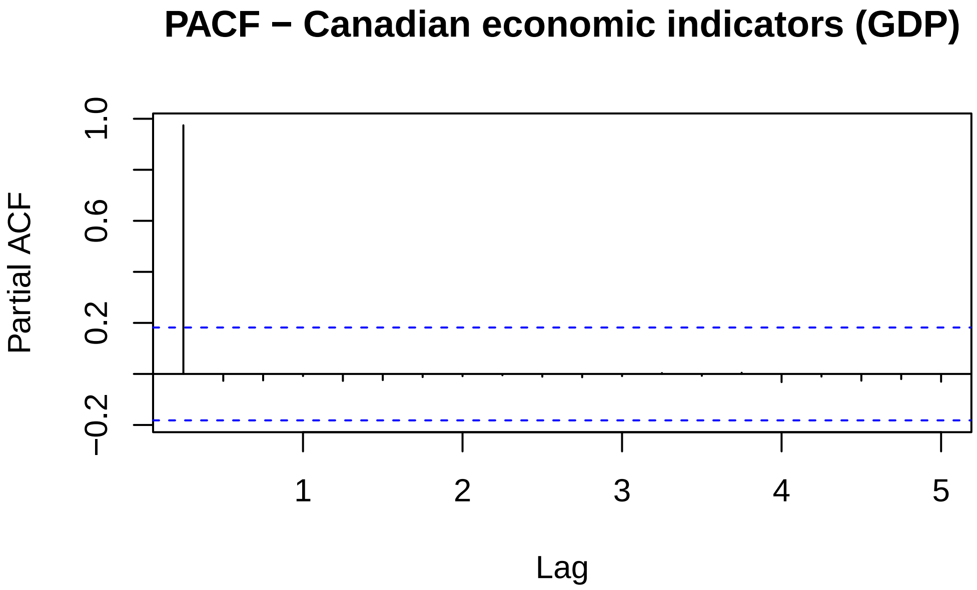
```
plot(gdp_ts,
     main = "Canada GDP (Quarterly)",
     ylab = "GDP Index",
     xlab = "Year",
     col = "blue",
     lwd = 2)
```



```
acf(gdp_ts, main = "ACF - Canadian economic indicators (GDP)")
```



```r
pacf(gdp_ts, main = "PACF - Canadian economic indicators (GDP)")
```



```r
adf_gdp <- adf.test(gdp_ts)

print(adf_gdp)
```

```

	Augmented Dickey-Fuller Test

data:  gdp_ts
Dickey-Fuller = -2.2565, Lag order = 4, p-value = 0.4701
alternative hypothesis: stationary
```

The ACF plot shows high and slowly decay autocorrelations sign of a non-stationary time series.

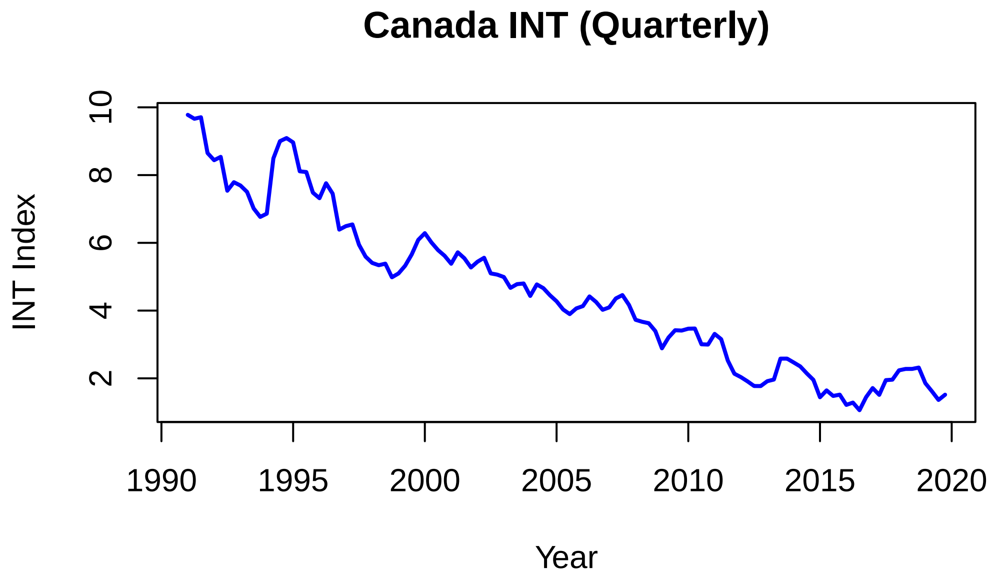
A slow decay in ACF suggests that past values have long memory and the level of the series is drifting over time.

The PACF shows A strong spike at lag 1. After lag 1, all other lags fall within the confidence interval.

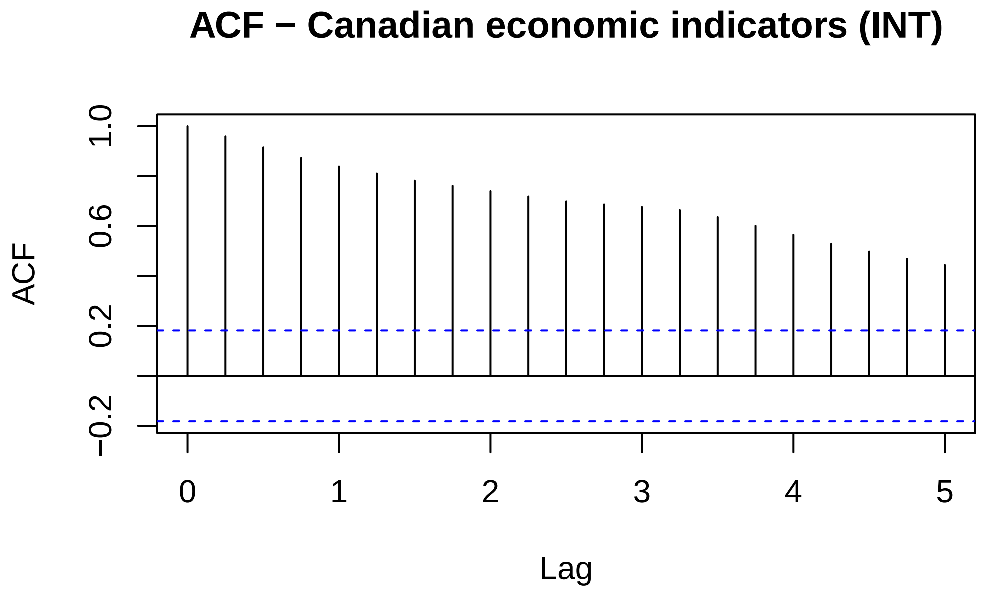
```r
INT <- read.csv("CAN_INT.csv")
```

```r
INT$TIME <- as.yearqtr(INT$TIME, format = "%Y-Q%q")
INT$DATE <- as.Date(INT$TIME)

int_ts <- ts(INT$Value, start = c(1991, 1), frequency = 4)
```

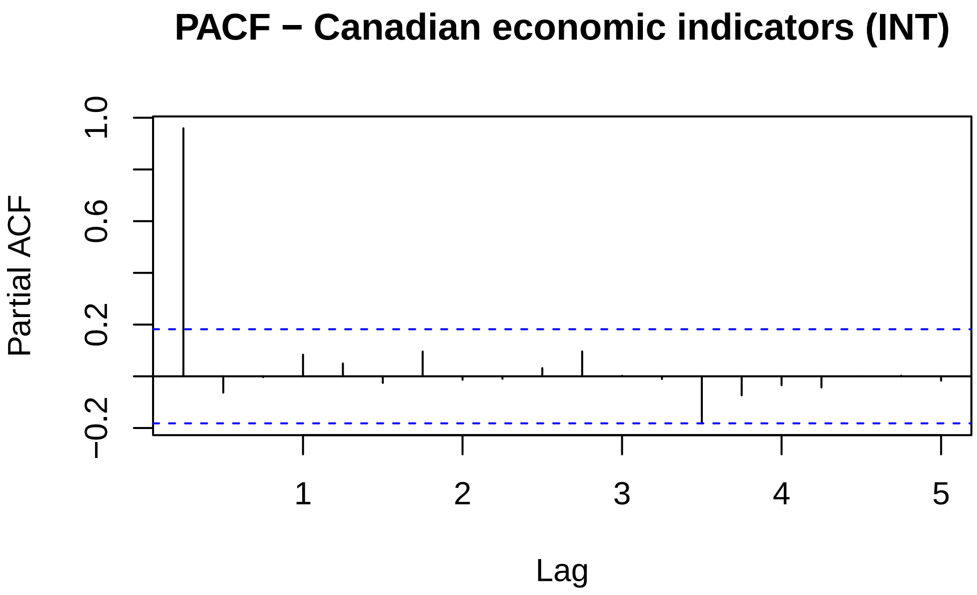
```r
plot(int_ts,
     main = "Canada INT (Quarterly)",
     ylab = "INT Index",
     xlab = "Year",
     col = "blue",
     lwd = 2)
```



```r
acf(int_ts, main = "ACF - Canadian economic indicators (INT)")
```



```r
pacf(int_ts, main = "PACF - Canadian economic indicators (INT)")
```



```r
adf_int <- adf.test(int_ts)

print(adf_int)
```

```
	Augmented Dickey-Fuller Test

data:  int_ts
Dickey-Fuller = -3.2941, Lag order = 4, p-value = 0.07565
alternative hypothesis: stationary
```

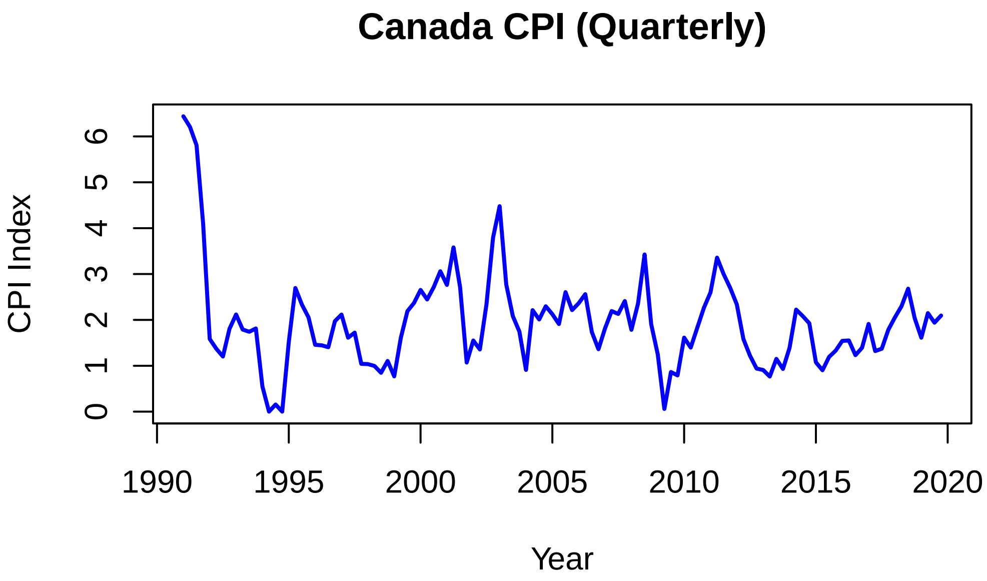
The ACF plot shows high first-lag autocorrelation and Slow, gradual decay across the next 15+ lags. The PACF plot shows a large spike at lag 1 and No significant spikes after lag 1 (lags fall inside the confidence bands)

This is the classic pattern of a non-stationary time series, most commonly, the ACF clearly indicates non-stationarity.

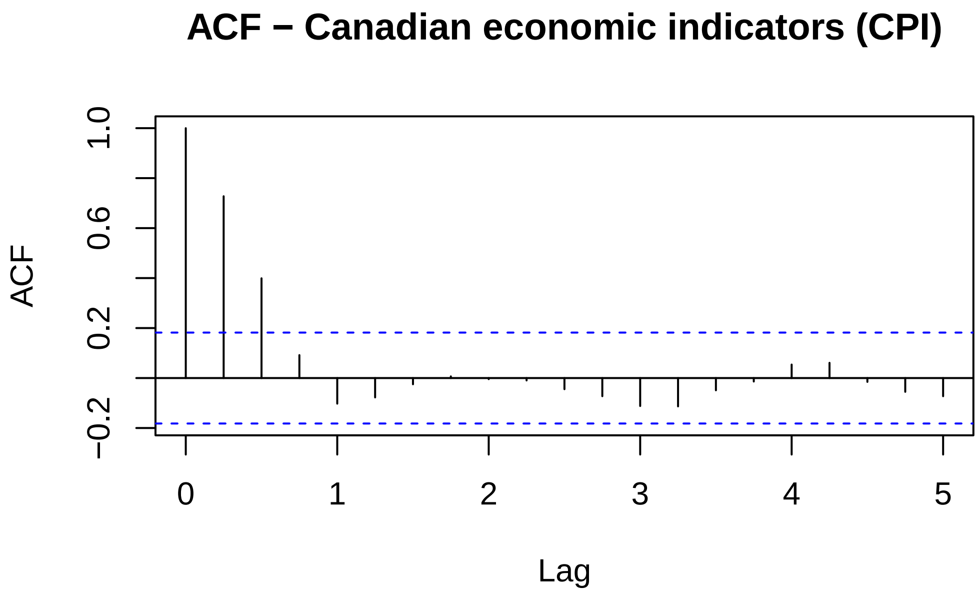
```
CPI <- read.csv("CAN_CPI.csv")
```

```
CPI$TIME <- as.yearqtr(CPI$TIME, format = "%Y-Q%q")
CPI$DATE <- as.Date(CPI$TIME)

cpi_ts <- ts(CPI$Value, start = c(1991, 1), frequency = 4)
```

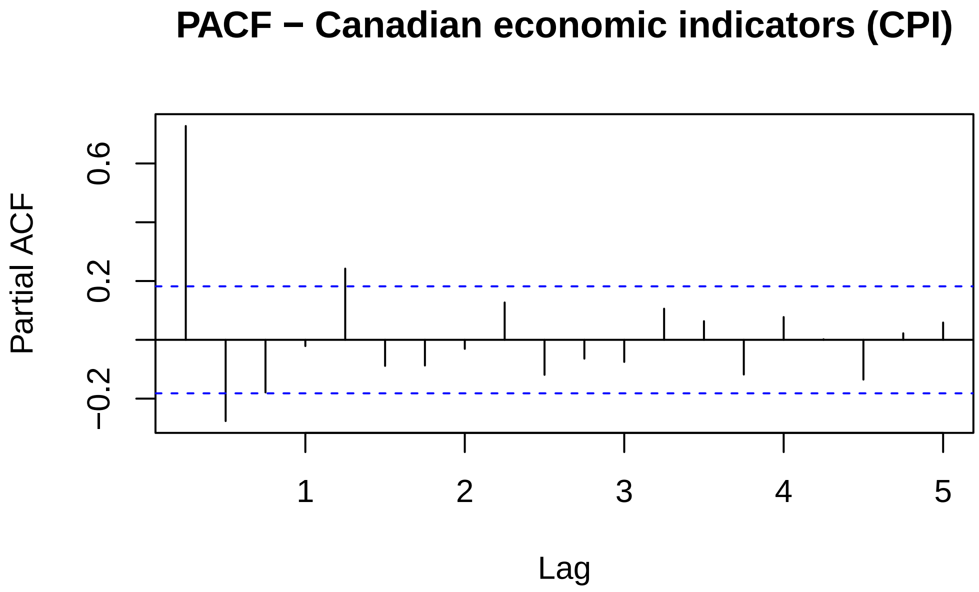
```
plot(cpi_ts,
     main = "Canada CPI (Quarterly)",
     ylab = "CPI Index",
     xlab = "Year",
     col = "blue",
     lwd = 2)
```



```r
acf(cpi_ts, main = "ACF - Canadian economic indicators (CPI)")
```



```r
pacf(cpi_ts, main = "PACF - Canadian economic indicators (CPI)")
```



```
adf_cpi <- adf.test(cpi_ts)

print(adf_cpi)
```

```

	Augmented Dickey-Fuller Test

data:  cpi_ts
Dickey-Fuller = -3.8826, Lag order = 4, p-value = 0.0174
alternative hypothesis: stationary
```

The ACF plot for CPI shows high autocorrelation at lag 1, Moderately high at lag 2 and persistent positive correlations The pattern indicates The CPI series has a trend. The series is likely non-stationary,

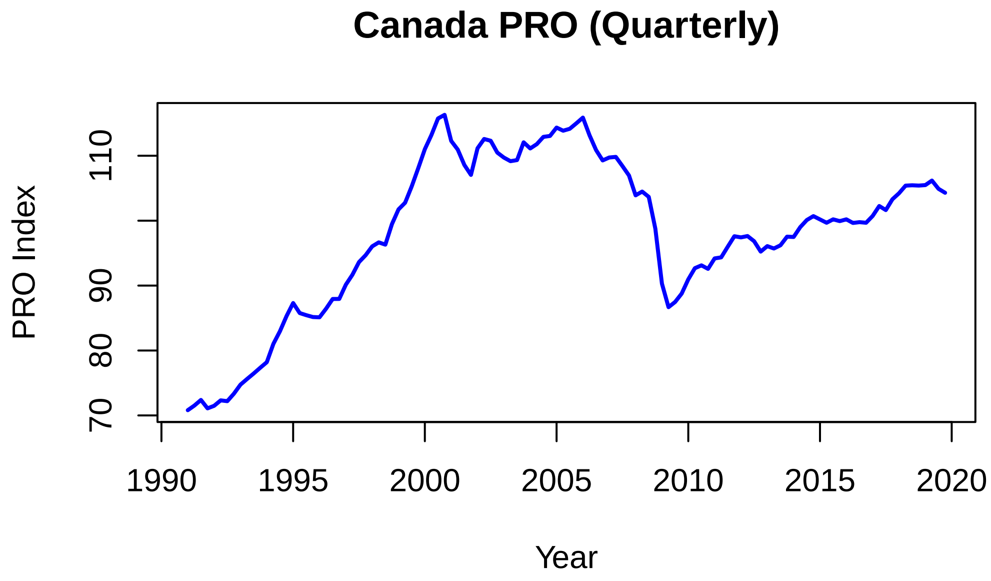
The PACF plot shows strong spike at lag 1 with Significant spike at lag 2. The spike at lag 5 may indicate seasonality. Strong spikes at lags 1 and 2 Indicates AR-type structure.

ACF and PACF suggest non-stationary but ADF test shows a P value > 0.05, therefore, the series is stationary.

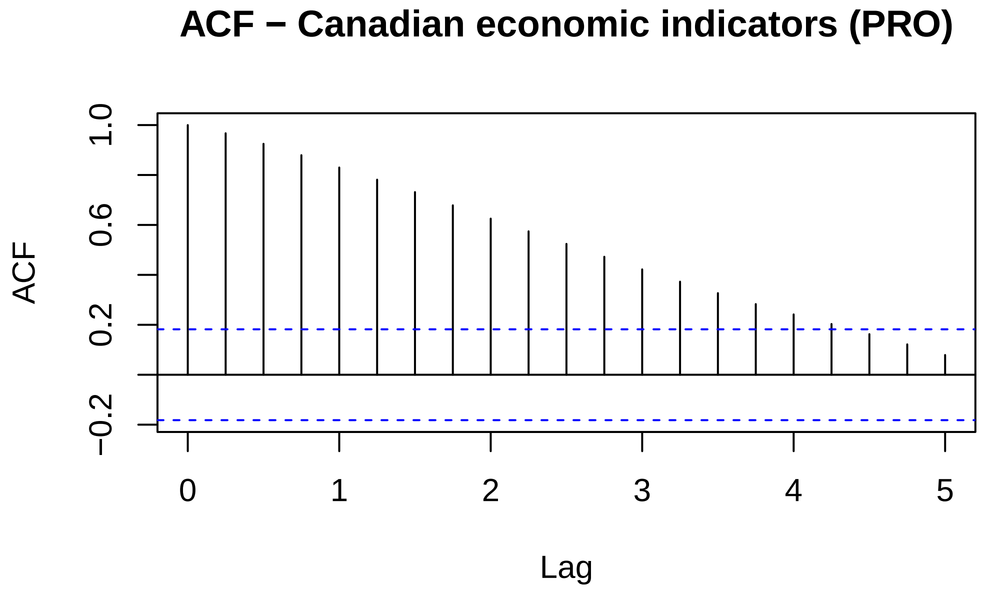
```
PRO <- read.csv("CAN_PRO.csv")
```

```
PRO$TIME <- as.yearqtr(PRO$TIME, format = "%Y-Q%q")
PRO$DATE <- as.Date(PRO$TIME)

pro_ts <- ts(PRO$Value, start = c(1991, 1), frequency = 4)
```

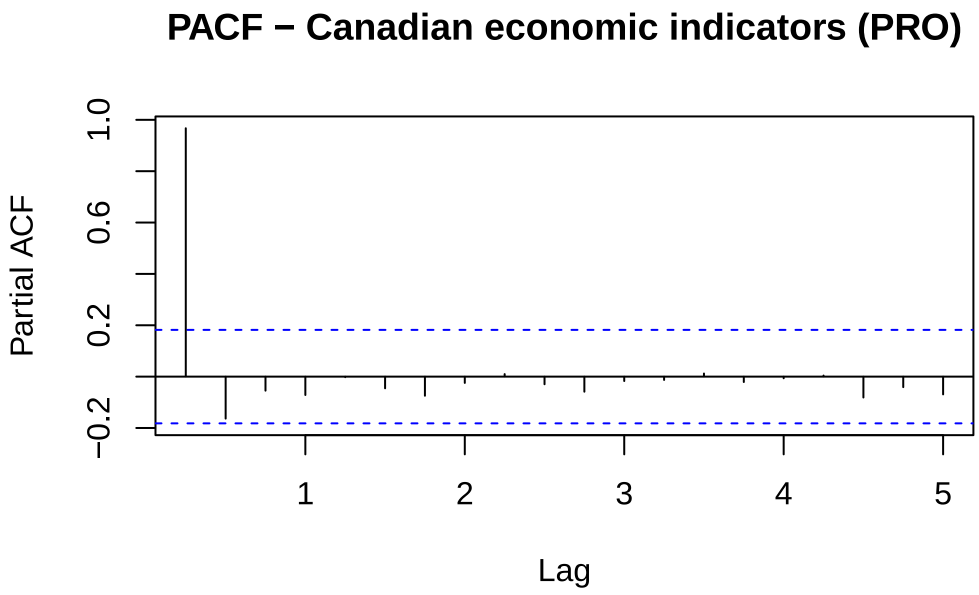
```
plot(pro_ts,
     main = "Canada PRO (Quarterly)",
     ylab = "PRO Index",
     xlab = "Year",
     col = "blue",
     lwd = 2)
```



```r
acf(pro_ts, main = "ACF - Canadian economic indicators (PRO)")
```



```
pacf(pro_ts, main = "PACF - Canadian economic indicators (PRO)")
```



```
adf_pro <- adf.test(pro_ts)

print(adf_pro)
```

```

	Augmented Dickey-Fuller Test

data:  pro_ts
Dickey-Fuller = -1.9537, Lag order = 4, p-value = 0.5958
alternative hypothesis: stationary
```

The ACF plot shows high autocorrelation at lag 1 with slow, gradual decay across many lags and No sharp cut-off and the ACF strongly suggests the series is non-stationary.

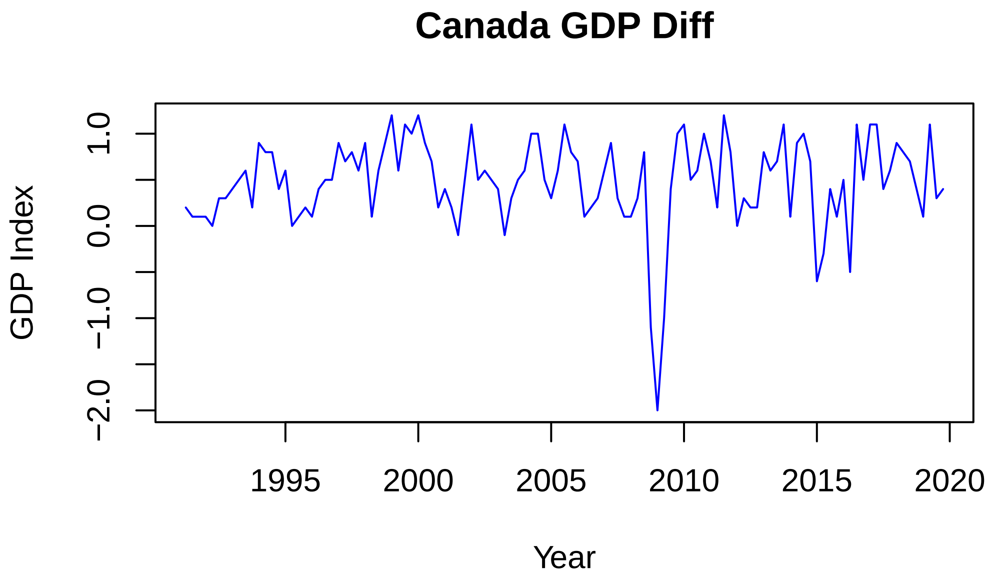
The PACF plot shows A large spike at lag 1, a small spike at lag 2

All remaining lags fall inside the confidence bands

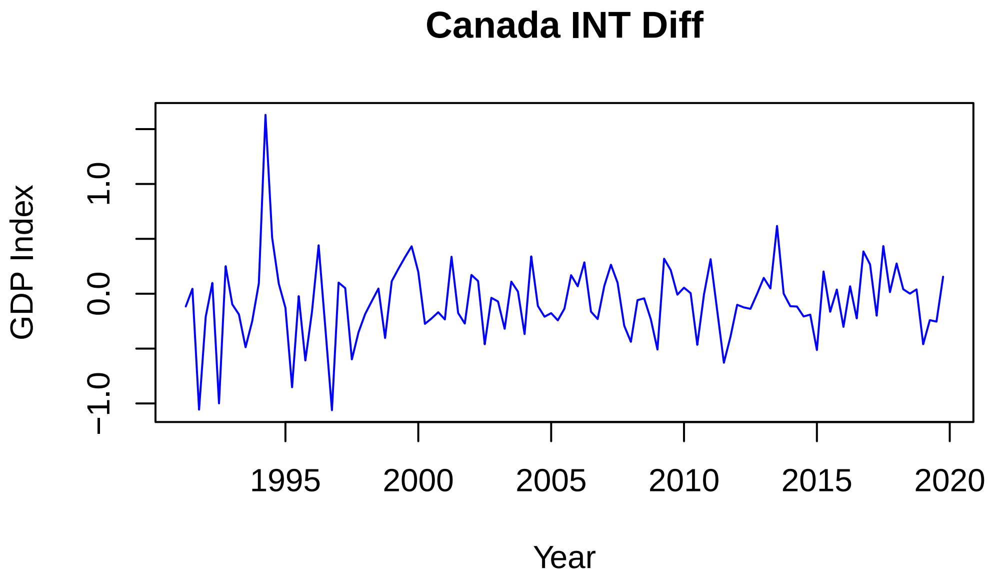
The Canadian Productivity Index PRO series is non-stationary.

(b) (3 points) For any non-stationary series, apply a log transformation, simple differencing, or a combination of both to achieve stationarity.

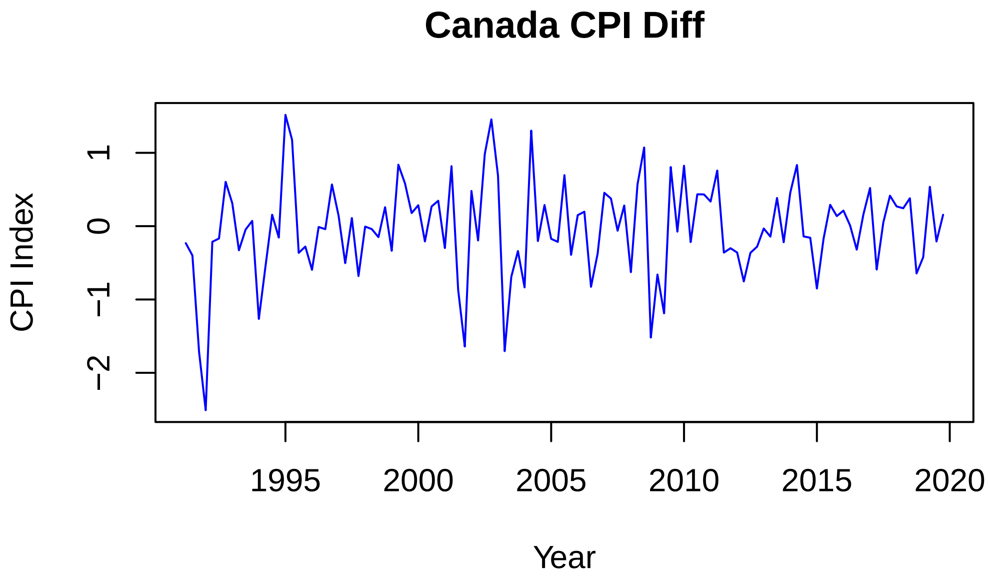
```
gdp_diff <- diff(gdp_ts)

plot(gdp_diff,
     main = "Canada GDP Diff",
     ylab = "GDP Index",
     xlab = "Year",
     col = "blue",
     lwd = 1)
```



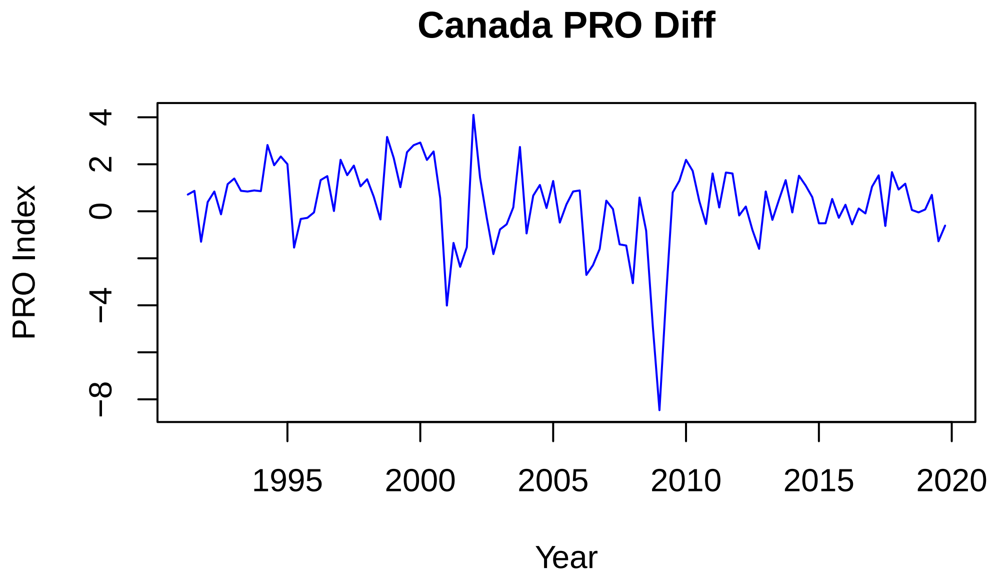
```
int_diff <- diff(int_ts)

plot(int_diff,
     main = "Canada INT Diff",
     ylab = "GDP Index",
     xlab = "Year",
     col = "blue",
     lwd = 1)
```



```
cpi_diff <- diff(cpi_ts)

plot(cpi_diff,
     main = "Canada CPI Diff",
     ylab = "CPI Index",
     xlab = "Year",
     col = "blue",
     lwd = 1)
```



```
pro_diff <- diff(pro_ts)

plot(pro_diff,
     main = "Canada PRO Diff",
     ylab = "PRO Index",
     xlab = "Year",
     col = "blue",
     lwd = 1)
```



```r
adf_gdp_diff <- adf.test(gdp_diff)
adf_gdp_diff
```

```
    Augmented Dickey-Fuller Test

data:  gdp_diff
Dickey-Fuller = -4.2954, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

```r
adf_int_diff <- adf.test(int_diff)
adf_int_diff
```

```
    Augmented Dickey-Fuller Test

data:  int_diff
Dickey-Fuller = -5.2043, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

```r
adf_cpi_diff <- adf.test(cpi_diff)
adf_cpi_diff
```

```

	Augmented Dickey-Fuller Test

data:  cpi_diff
Dickey-Fuller = -7.0443, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

```r
adf_pro_diff <- adf.test(pro_diff)
adf_pro_diff
```

```

	Augmented Dickey-Fuller Test

data:  pro_diff
Dickey-Fuller = -3.6775, Lag order = 4, p-value = 0.02938
alternative hypothesis: stationary
```

(c) (2 points) Propose candidate orders for a VAR model for the four series.

```r
library(vars)
```

```r
df_all <- merge(
  merge(
    merge(GDP[, c("TIME","Value")], INT[, c("TIME","Value")], by="TIME"),
    CPI[, c("TIME","Value")], by="TIME"
  ),
  PRO[, c("TIME","Value")], by="TIME"
)

names(df_all) <- c("TIME","GDP","INT","CPI","PRO")
```

```r
ts_data <- ts(df_all[, -1], start = c(1991,1), frequency = 4)
```

```r
ts_diff <- diff(ts_data)
```

```r
var_selection <- VARselect(ts_diff, lag.max = 8, type = "const")
```

```r
print(var_selection)
```

```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     1      1      1      1

$criteria
                    1            2            3           4           5
AIC(n) -4.756530058 -4.625689623 -4.622224112 -4.65572943 -4.52125775
HQ(n)  -4.554001223 -4.261137720 -4.095649140 -3.96713139 -3.67063664
SC(n)  -4.256935884 -3.726420109 -3.323279258 -2.95710923 -2.42296222
FPE(n)  0.008597725  0.009812522  0.009878458  0.00961078  0.01110212
                 6           7           8
AIC(n) -4.40111646 -4.27199664 -4.32147548
HQ(n)  -3.38847229 -3.09732940 -2.98478517
SC(n)  -1.90314559 -1.37435043 -1.02415393
FPE(n)  0.01270265  0.01475111  0.01442986
```

```
select.p=VARselect(ts_diff)
select.p$selection
```

```
AIC(n)  HQ(n)  SC(n) FPE(n) 
     1      1      1      1 
```

Candidate models VAR1 AND VAR2

(d) (3 points) Estimate VAR models for the data based on the order specified in part (c).

```
var1 <- VAR(ts_diff, p = 1, type = "const")
var1
```

```

VAR Estimation Results:
=======================

Estimated coefficients for equation GDP:
========================================
Call:
GDP = GDP.l1 + INT.l1 + CPI.l1 + PRO.l1 + const

    GDP.l1     INT.l1     CPI.l1     PRO.l1      const
0.24497752 0.13769199 0.01552869 0.06805885 0.35180542


Estimated coefficients for equation INT:
========================================
Call:
INT = GDP.l1 + INT.l1 + CPI.l1 + PRO.l1 + const

     GDP.l1      INT.l1      CPI.l1      PRO.l1       const
 0.32616959  0.16707951 -0.11341562 -0.07880796 -0.19537443


Estimated coefficients for equation CPI:
========================================
Call:
CPI = GDP.l1 + INT.l1 + CPI.l1 + PRO.l1 + const

      GDP.l1       INT.l1       CPI.l1       PRO.l1        const
 0.200124853  0.164577770  0.092132292  0.005304653 -0.117197164


Estimated coefficients for equation PRO:
========================================
Call:
PRO = GDP.l1 + INT.l1 + CPI.l1 + PRO.l1 + const

    GDP.l1     INT.l1     CPI.l1     PRO.l1      const
 0.7290371  0.2966722 -0.5144435  0.3904537 -0.1749231
```

```r
var2 <- VAR(ts_diff, p = 2, type = "const")
var2
```

```
VAR Estimation Results:
=======================

Estimated coefficients for equation GDP:
========================================
Call:
GDP = GDP.l1 + INT.l1 + CPI.l1 + PRO.l1 + GDP.l2 + INT.l2 + CPI.l2 + PRO.l2 + const

      GDP.l1       INT.l1       CPI.l1       PRO.l1       GDP.l2       INT.l2 
 0.283534734  0.151216496  0.024517868  0.068062569 -0.159567824 -0.001910313 
      CPI.l2       PRO.l2        const 
-0.035869284  0.020671819  0.405876701 


Estimated coefficients for equation INT:
========================================
Call:
INT = GDP.l1 + INT.l1 + CPI.l1 + PRO.l1 + GDP.l2 + INT.l2 + CPI.l2 + PRO.l2 + const

      GDP.l1       INT.l1       CPI.l1       PRO.l1       GDP.l2       INT.l2 
 0.375879340  0.143405124 -0.102744794 -0.080439490  0.004619741 -0.072603779 
      CPI.l2       PRO.l2        const 
-0.053705276 -0.014183032 -0.227062418 


Estimated coefficients for equation CPI:
========================================
Call:
CPI = GDP.l1 + INT.l1 + CPI.l1 + PRO.l1 + GDP.l2 + INT.l2 + CPI.l2 + PRO.l2 + const

     GDP.l1      INT.l1      CPI.l1      PRO.l1      GDP.l2      INT.l2 
 0.21765155  0.23888080  0.10028156 -0.02815615 -0.14803243 -0.01065461 
     CPI.l2      PRO.l2       const 
-0.07445333  0.08750820 -0.06765746 


Estimated coefficients for equation PRO:
========================================
Call:
PRO = GDP.l1 + INT.l1 + CPI.l1 + PRO.l1 + GDP.l2 + INT.l2 + CPI.l2 + PRO.l2 + const

    GDP.l1     INT.l1     CPI.l1     PRO.l1     GDP.l2     INT.l2     CPI.l2 
 1.0413307  0.3562165 -0.4491569  0.3581434 -0.6999875 -0.2340087 -0.2854555 
    PRO.l2      const 
 0.1000591 -0.0373458 
```

(e) (5 points) Estimate AR models for each of the four series using the order specified in part (c).

```
arima(gdp_diff, order=c(1,0,0))
```

```


Call:
arima(x = gdp_diff, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.4442     0.4732
s.e.  0.0830     0.0724

sigma^2 estimated as 0.1886:  log likelihood = -67.36,  aic = 140.72
```

```
arima(int_diff, order=c(1,0,0))
```

```


Call:
arima(x = int_diff, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.1208    -0.0716
s.e.  0.0923     0.0362

sigma^2 estimated as 0.1169:  log likelihood = -39.75,  aic = 85.51
```

```
arima(cpi_diff, order=c(1,0,0))
```

```


Call:
arima(x = cpi_diff, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.1630    -0.0378
s.e.  0.0916     0.0714

sigma^2 estimated as 0.4121:  log likelihood = -112.22,  aic = 230.45
```

```
arima(pro_diff, order=c(1,0,0))
```

```
Call:
arima(x = pro_diff, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.5003     0.2871
s.e.  0.0800     0.2729

sigma^2 estimated as 2.176:  log likelihood = -208.02,  aic = 422.05
```

(f) (5 points) Evaluate the VAR and individual AR models by comparing their mean squared errors and the number of parameters.

(g) (5 points) What model(s) do you recommend for this data?