Lesson 1.1

Basic Terms in Time Series Analysis

Overview

- Social and economic conditions constantly change over time
- Assess and predict the effects of these changes in order to suggest the most appropriate actions to take

What is a time series?

A **time series** is a sequence of measurements of the same variable collected over time. The measurements are made at regular time intervals, say daily, weekly, monthly, quarterly, or annually. We can also think of a time series as a realization of a stochastic process $\{X_t\}$ Examples:

Examples:

1. Monthly Inflation rate from 2017 to 2021
2. Quarterly HH consumption expenditure 2000 to 2021
3. Quarterly swine and broiler production from 2007 to 2016

Time series analysis refers to the use of statistical methods to enhance understanding and prediction on any quantitative variable of interest (prices, measurements, social indicators, etc)

Objectives of time series analysis

1. describe the important features of time series pattern,
2. explain how the past affects the future or how two time series can *interact*,
3. forecast future values of the series, and
4. serve as a control standard for a variable that measures the quality of product in some manufacturing situations

Applications of time series analysis

- Forecasting of shipped parcels: *workforce planning*
- Forecasting of sales promotions: *optimizing warehouses*
- Claims prediction: *determining insurance policies*
- Predictive maintenance: *improving operational efficiency*
- Energy load forecasting: *better planning and strategies*

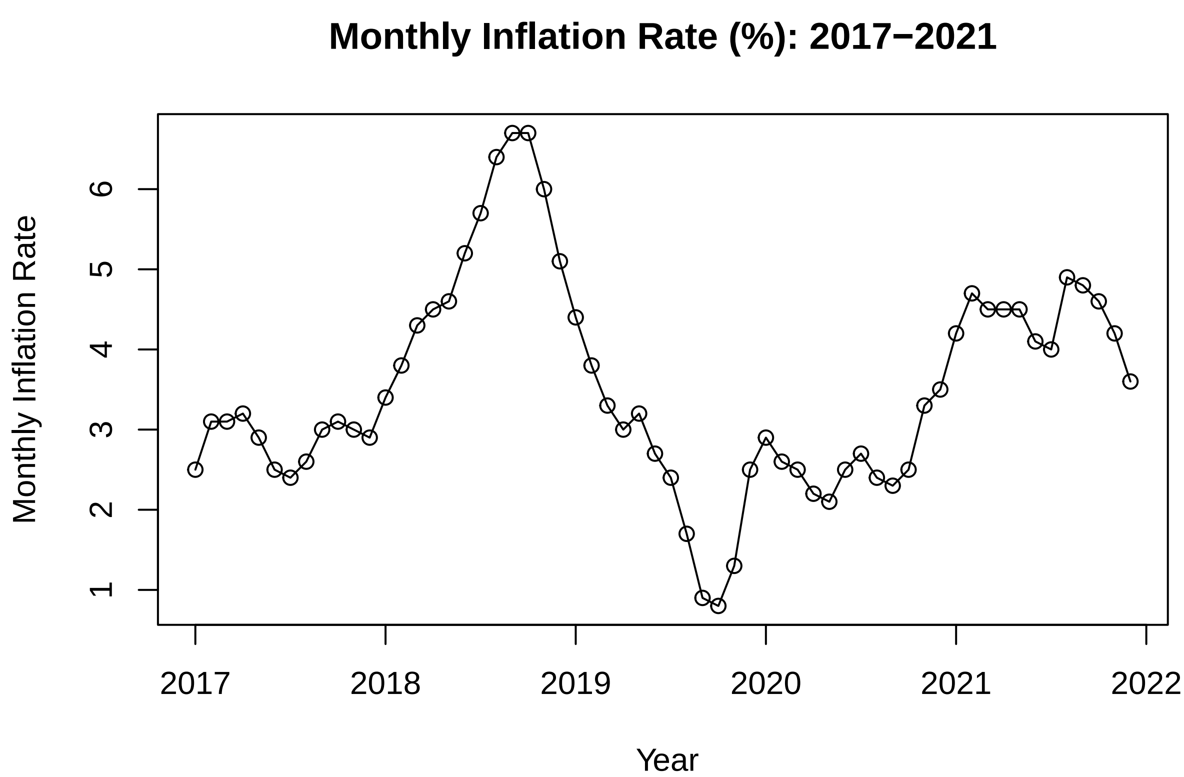
Time series plot

A **time series plot** displays patterns in the data over time, such as trends or seasonality and it helps one determine the appropriate time series model of the data.

Monthly Inflation Rate for Eastern Visayas: 2017-2021

Let us begin by loading required packages and importing the monthly inflation rate data for Eastern Visayas into R for 2017-2021.

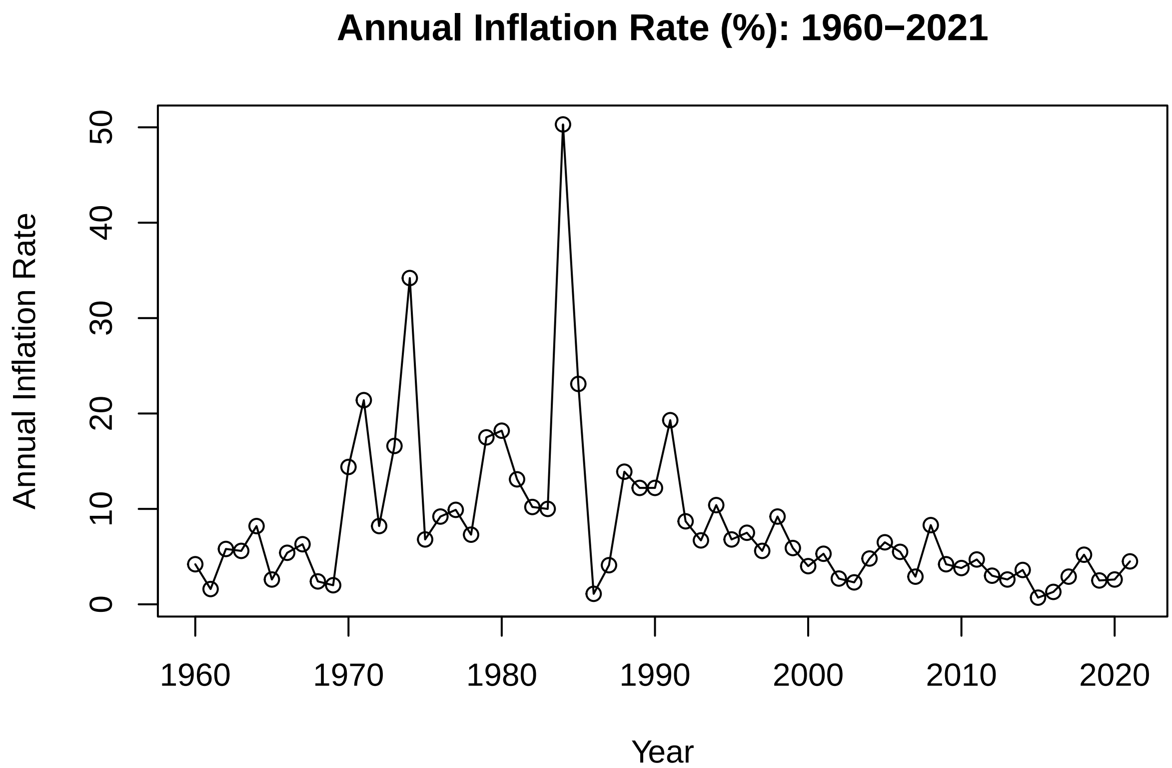
Next, we convert the **mir** data into a time series data using the *ts()* function and plot the resulting time series using the *plot()* function.



As shown in the plot, there is an increasding trend from January 2017 to around October or November of 2018, then a downward trend from then on until October 2019, and finally an upward trend again until the last quarter of 2021.

Annual Inflation Rate, Eastern Visayas: 1960-2021

Let us look at another data file, the annual inflation rate for Eastern Visayas.

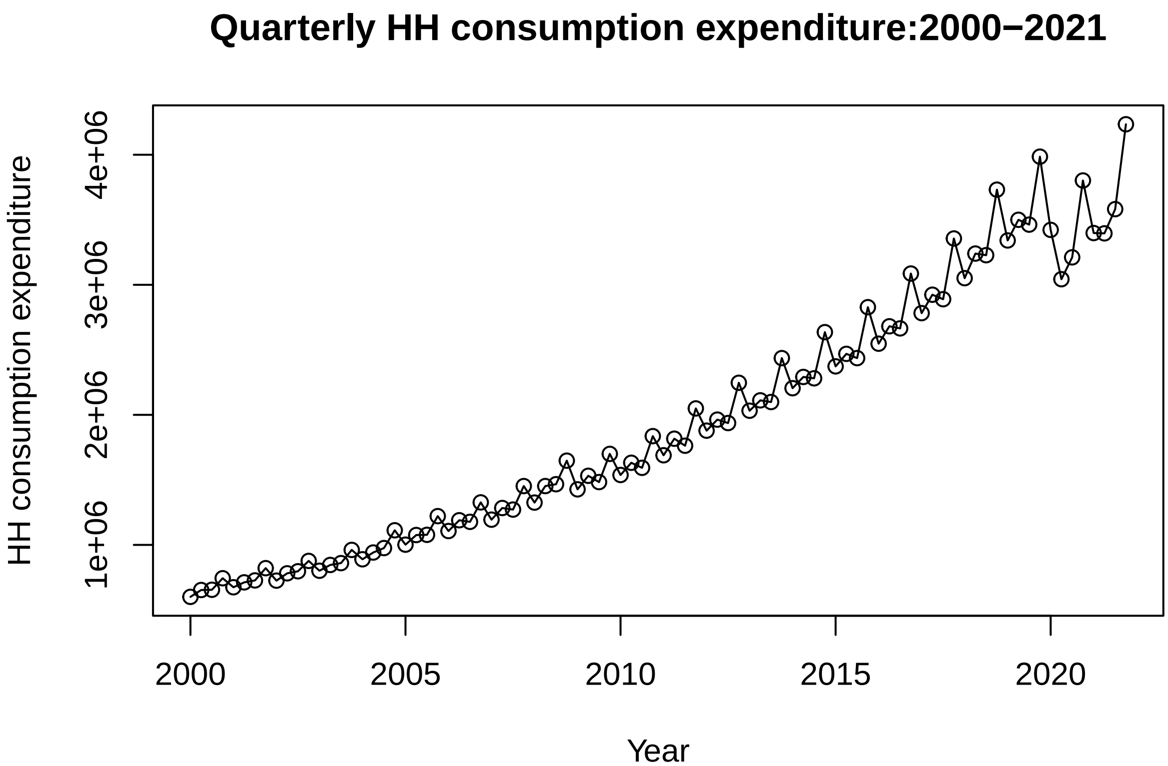


The plot show fairly flat annual pattern, except for the spikes in the early 1970s and mid-1980s.

Quarterly HH consumption expenditure

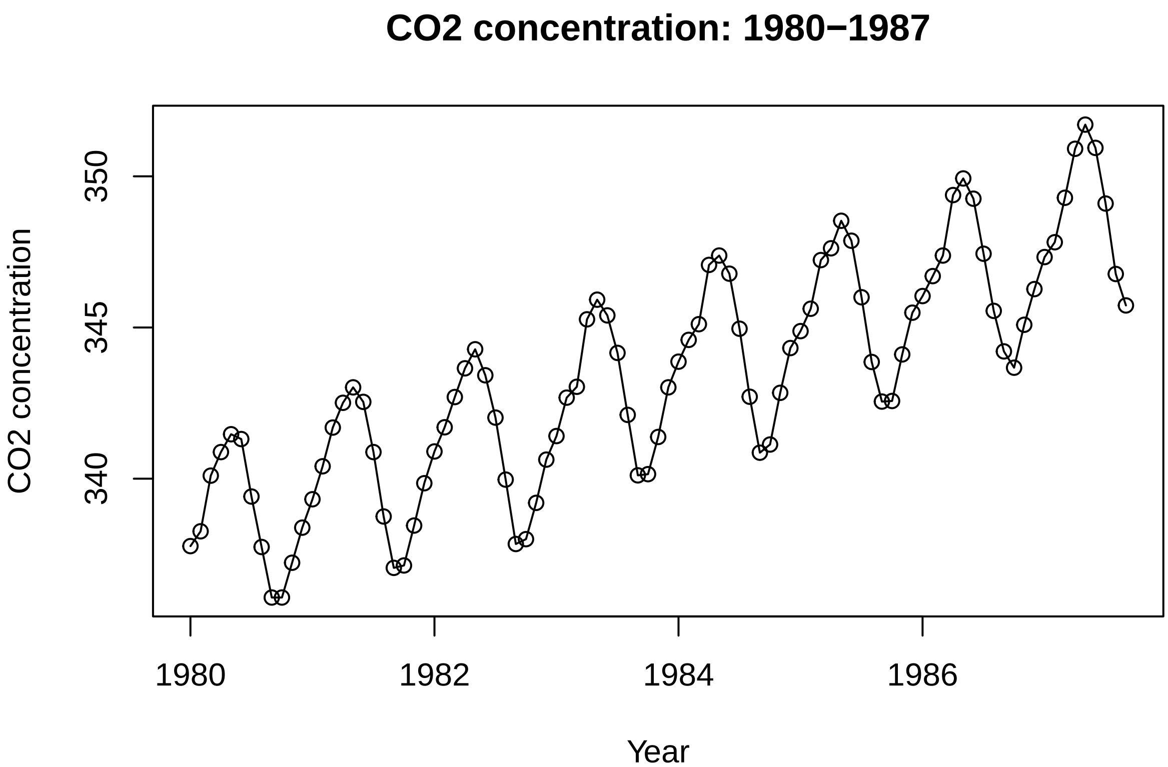
A third time series is the quarterly household consumption expenditure for the Philippines from 2000 to 2021. Let us import and plot the data.

```
hh.exp<-read_csv("Quarterly HH consumption expenditure.csv")

hh.exp_ts<-ts(hh.exp$HHCE,
              start = c(2000,1),
              end = c(2021,4),
              frequency = 4)

plot(hh.exp_ts,
     type = "o",
     xlab = "Year",
     ylab = "HH consumption expenditure",
     main = "Quarterly HH consumption expenditure:2000-2021")
```



The time series plot exhibits both (an increasing) trend and seasonality.

Global monthly carbon dioxide concentration in the atmosphere: 1980-1987



This time series also exhibits trend and seasonality patterns.

Time series operators

The lag operator

The lag operator or backward (shift) operator , denoted by L,is an operator that shifts the time index backward by one unit.

$$B(X_t) = X_{t-1}$$

That is applying the lag operator,L, on a time series $\{X_t\}$ produces $\{X_{t-1}\}$ series. Also, applying the lag operator on the series $\{X_{t-1}\}$ produces $\{X_{t-2}\}$.

This is illustrated in the following example.

Year	Quarter	X_t	X_{t-1}	X_{t-2}
2000	Quarter 1	1493		
	Quarter 2	2161	1493	
	Quarter 3	1841	2161	1493
	Quarter 4	2143	1841	2161
2001	Quarter 1	1500	2143	1841
	Quarter 2	2200	1500	2143
	Quarter 3	1913	2200	1500
	Quarter 4	1950	1913	2200
2002	Quarter 1	1492	1950	1913
	Quarter 2	2318	1492	1950
	Quarter 3	1817	2318	1492
	Quarter 4	2148	1817	2318
2003	Quarter 1	1466	2148	1817
	Quarter 2	2277	1466	2148
	Quarter 3	1544	2277	1466
	Quarter 4	2341	1544	2277

The lead operator

The lead operator or forward (shift) operator, denoted by F, shifts the time index forward by one unit

$$F(X_t) = X_{t+1}$$

This means that applying the forward operator on a time series X_t produces the X_{t+1} series.

See below for an example.

Year	Quarter	X_t	X_{t+1}	X_{t+2}
2000	Quarter 1	1493	2161	1841
	Quarter 2	2161	1841	2143
	Quarter 3	1841	2143	1500
	Quarter 4	2143	1500	2200
2001	Quarter 1	1500	2200	1913
	Quarter 2	2200	1913	1950
	Quarter 3	1913	1950	1492
	Quarter 4	1950	1492	2318
2002	Quarter 1	1492	2318	1817
	Quarter 2	2318	1817	2148
	Quarter 3	1817	2148	1466
	Quarter 4	2148	1466	2277
2003	Quarter 1	1466	2277	1544
	Quarter 2	2277	1544	2341
	Quarter 3	1544	2341	
	Quarter 4	2341		

Difference operator

The difference operator, denoted by Δ, is used to express the difference between two consecutive realizations of a time series. The first difference of a series is given by

$$\Delta X_t = X_t - X_{t-1}$$

Meanwhile, the second difference is given by

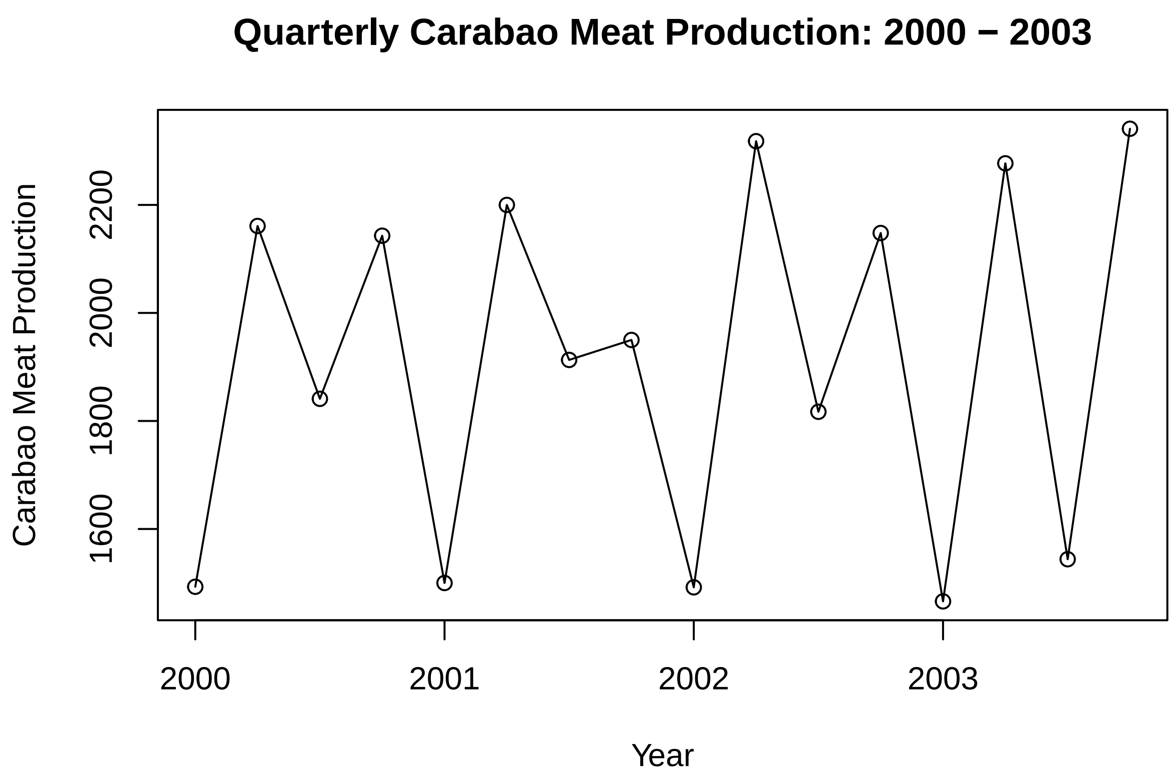
$$\begin{aligned} \Delta^2 X_t &= [X_t - X_{t-1}] - [X_{t-1} - X_{t-2}] \\ &= \Delta X_t - \Delta X_{t-1} \end{aligned}$$

Referring to the same time series, we have in the last 2 columns the first and second differences, respectively.

Year	Quarter	X_t	X_{t-1}	X_{t-2}	ΔX_t	$\Delta^2 X_t$
2000	Quarter 1	1493				
	Quarter 2	2161	1493		668	
	Quarter 3	1841	2161	1493	-320	-988
	Quarter 4	2143	1841	2161	302	622
2001	Quarter 1	1500	2143	1841	-643	-945
	Quarter 2	2200	1500	2143	700	1343
	Quarter 3	1913	2200	1500	-287	-987
	Quarter 4	1950	1913	2200	37	324
2002	Quarter 1	1492	1950	1913	-458	-495
	Quarter 2	2318	1492	1950	826	1284
	Quarter 3	1817	2318	1492	-501	-1327
	Quarter 4	2148	1817	2318	331	832
2003	Quarter 1	1466	2148	1817	-682	-1013
	Quarter 2	2277	1466	2148	811	1493
	Quarter 3	1544	2277	1466	-733	-1544
	Quarter 4	2341	1544	2277	797	1530

To illustrate *lag*, *lead*, and *difference* operators in R, let us use the quarterly carabao meat production for Eastern Visayas from 2000 to 2003.

```
cara<-read.csv("Carabao meat production.csv")

cara.ts<-ts(cara$Xt,
            frequency = 4,
            start = c(2000,1),
            end = c(2003,4))

plot.ts(cara.ts,
     type ="o",
     xlab = "Year",
     ylab = "Carabao Meat Production",
     main = "Quarterly Carabao Meat Production: 2000 - 2003")
```



Lag and Lead operators

The *lag()* function from the **stats** package can be used to create lags or leads for *ts* objects. This function has one argument, k, which defines the number of lags or leads to be created for a given input series. The n lag of the series is defined by $k = -n$ and, similarly, the n lead of the series is defined by $k = n$. The resulting objects are also *ts* objects.

```
cara.lag1 <- stats::lag(cara.ts, k = -1) # Creates lag-1 series
cara.lag2 <- stats::lag(cara.ts, k = -2) # Creates lag-1 series
cara.lead1 <- stats::lag(cara.ts, k = 1) # Creates lead-1 series
cara.lead2 <- stats::lag(cara.ts, k = 2) # Creates lead-2 series
```

You can view the resuting *ts* objects as a *tibble* object using the *tk_tbl()* function from the **timetk** package. For example,

```
tk_tbl(cara.lag1)
```

```
## # A tibble: 16 x 2
##    index     value
##    <yearqtr> <int>
```

```
## 1 2000 Q2 1493
## 2 2000 Q3 2161
## 3 2000 Q4 1841
## 4 2001 Q1 2143
## 5 2001 Q2 1500
## 6 2001 Q3 2200
## 7 2001 Q4 1913
## 8 2002 Q1 1950
## 9 2002 Q2 1492
## 10 2002 Q3 2318
## 11 2002 Q4 1817
## 12 2003 Q1 2148
## 13 2003 Q2 1466
## 14 2003 Q3 2277
## 15 2003 Q4 1544
## 16 2004 Q1 2341
```

```r
tk_tbl(cara.lead1)
```

```
## # A tibble: 16 x 2
##      index     value
##      <yearqtr> <int>
##    1 1999 Q4    1493
##    2 2000 Q1    2161
##    3 2000 Q2    1841
##    4 2000 Q3    2143
##    5 2000 Q4    1500
##    6 2001 Q1    2200
##    7 2001 Q2    1913
##    8 2001 Q3    1950
##    9 2001 Q4    1492
##   10 2002 Q1    2318
##   11 2002 Q2    1817
##   12 2002 Q3    2148
##   13 2002 Q4    1466
##   14 2003 Q1    2277
##   15 2003 Q2    1544
##   16 2003 Q3    2341
```

Alternatively, we can use the *lag()* and *lead()* operators in the **dplyr** package to compute lag and lead of a time series. Note, however, that the *lag()* and *lead()* operators in the **dplyr** package works only on vectors, not on *ts()* data.

```r
cara.lag1a <- dplyr::lag(cara$Xt,n=1) #computes lag-1 series
cara.lag2a <- dplyr::lag(cara$Xt,n=2) #computes lag-2 series
cara.lead1a <- dplyr::lead(cara$Xt,n=1) #computes lead-1 series
cara.lead2a <- dplyr::lead(cara$Xt,n=2) #computes lead-2 series
cara1a <- cbind(cara$Xt, cara.lag1a, cara.lag2a, cara.lead1a, cara.lead2a)
head(cara1a, 10)
```

```
##            cara.lag1a cara.lag2a cara.lead1a cara.lead2a
##  [1,] 1493         NA         NA        2161        1841
##  [2,] 2161       1493         NA        1841        2143
##  [3,] 1841       2161       1493        2143        1500
##  [4,] 2143       1841       2161        1500        2200
##  [5,] 1500       2143       1841        2200        1913
##  [6,] 2200       1500       2143        1913        1950
##  [7,] 1913       2200       1500        1950        1492
##  [8,] 1950       1913       2200        1492        2318
##  [9,] 1492       1950       1913        2318        1817
## [10,] 2318       1492       1950        1817        2148
```

```r
tail(cara1a,10)
```

```
##            cara.lag1a cara.lag2a cara.lead1a cara.lead2a
##  [7,] 1913       2200       1500        1950        1492
##  [8,] 1950       1913       2200        1492        2318
##  [9,] 1492       1950       1913        2318        1817
## [10,] 2318       1492       1950        1817        2148
## [11,] 1817       2318       1492        2148        1466
## [12,] 2148       1817       2318        1466        2277
## [13,] 1466       2148       1817        2277        1544
## [14,] 2277       1466       2148        1544        2341
## [15,] 1544       2277       1466        2341          NA
## [16,] 2341       1544       2277          NA          NA
```

Difference operator

We can use the *diff()* function in the **stats** package to generate "differenced" series. This function works on *ts()* data.

```r
dif1<-diff(cara.ts,lag = 1)#to compute the first difference
dif2<-diff(dif1,lag = 1)#to compute the 2nd difference
caranew1<-cbind(cara.ts, dif1, dif2)
print(caranew1)
```

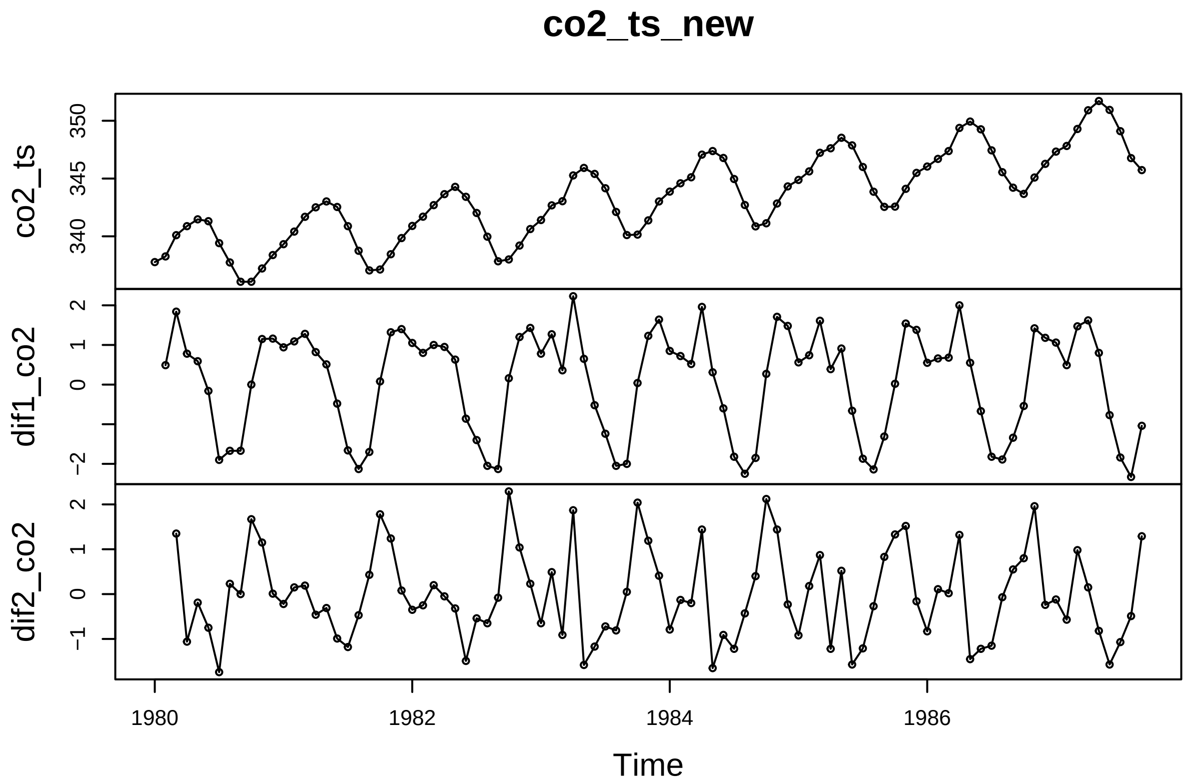
```
##         cara.ts dif1  dif2
## 2000 Q1    1493   NA    NA
## 2000 Q2    2161  668    NA
## 2000 Q3    1841 -320  -988
## 2000 Q4    2143  302   622
## 2001 Q1    1500 -643  -945
## 2001 Q2    2200  700  1343
## 2001 Q3    1913 -287  -987
## 2001 Q4    1950   37   324
## 2002 Q1    1492 -458  -495
## 2002 Q2    2318  826  1284
## 2002 Q3    1817 -501 -1327
## 2002 Q4    2148  331   832
## 2003 Q1    1466 -682 -1013
## 2003 Q2    2277  811  1493
## 2003 Q3    1544 -733 -1544
## 2003 Q4    2341  797  1530
```

Notice that to get the second "differenced" series, we applied the *diff()* function on the first "differences" contained in the *dif1* object.

Remarks

We have seen above the effect of a moving average in *smoothing* a series. Next I will illustrate, for curiosity sake, the effect of *differencing* to a time series with trend and seasonality. Recall that the CO_2 series exhibits both trend and seasonality, hence, we will use that data and observe the effect of *differencing* on the series.

```r
dif1_co2<-diff(co2_ts,lag=1)
dif2_co2<-diff(dif1_co2,lag=1)
co2_ts_new=cbind(co2_ts,dif1_co2,dif2_co2)
plot.ts(co2_ts_new, type="o")
```



The above plot shows that *differencing* removes the trend and seasonality in the series, and produces a series with purely random pattern.

Autocorrelation

Autocorrelation is the correlation between two values in a time series The number of intervals between the two observations is the lag. For example, the lag between the current and previous observation is one. In general, the observations at y_t and y_{t-k} are separated by k (lag) time units. When k=1, you're assessing adjacent observations.

The **autocorrelation function (ACF)** assesses the correlation between observations in a time series for a set of lags. The ACF is used to identify which lags have significant correlations, understand the patterns (trends and seasonality) of the time series, and then use that information to model the time series data. The ACF for time series $y_t$is given by

$$Corr(y_t, y_{t-k}),\ \ k = 1, 2 \ldots$$

Year	Quarter	X_t	X_{t-1}	X_{t-2}	X_{t-3}	X_{t-4}	X_{t-5}	X_{t-6}
2000	Quarter 1	1493						
	Quarter 2	2161	1493					
	Quarter 3	1841	2161	1493				
	Quarter 4	2143	1841	2161	1493			
2001	Quarter 1	1500	2143	1841	2161	1493		
	Quarter 2	2200	1500	2143	1841	2161	1493	
	Quarter 3	1913	2200	1500	2143	1841	2161	1493
	Quarter 4	1950	1913	2200	1500	2143	1841	2161
2002	Quarter 1	1492	1950	1913	2200	1500	2143	1841
	Quarter 2	2318	1492	1950	1913	2200	1500	2143
	Quarter 3	1817	2318	1492	1950	1913	2200	1500
	Quarter 4	2148	1817	2318	1492	1950	1913	2200
2003	Quarter 1	1466	2148	1817	2318	1492	1950	1913
	Quarter 2	2277	1466	2148	1817	2318	1492	1950
	Quarter 3	1544	2277	1466	2148	1817	2318	1492
	Quarter 4	2341	1544	2277	1466	2148	1817	2318
		r=	**-0.81056**	**0.59834**	**-0.75784**	**0.91406**	**-0.76366**	**0.601594**
		k=	**1**	**2**	**3**	**4**	**5**	**6**