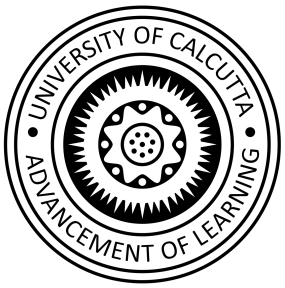
206: Time Series Analysis And National Development

Instructor: Prof. Sugata Sen Roy

Report Card
Time Series Data Analysis

Submitted by: Soumarya Basak

Submitted on: June 25, 2022



University of Calcutta

Department of Statistics

Contents

1 Problem 1 **2**
- 1.1 About the Practical . . . 2
- 1.2 Analysis . . . 2
  - 1.2.1 Correlogram . . . 2
  - 1.2.2 Fitting Alternative models . . . 3
  - 1.2.3 Test for Randomness of Residuals . . . 4
  - 1.2.4 Forecast . . . 5

2 Problem 2 **6**
- 2.1 About the Practical . . . 6
- 2.2 Analysis . . . 6
  - 2.2.1 Exponential Smoothing . . . 6
  - 2.2.2 Holt Winter Method . . . 8
  - 2.2.3 Box-Jenkins Model . . . 10

3 Problem 3 **13**
- 3.1 About the Practical . . . 13
- 3.2 Analysis . . . 13
  - 3.2.1 Fitting of SARIMA model . . . 14

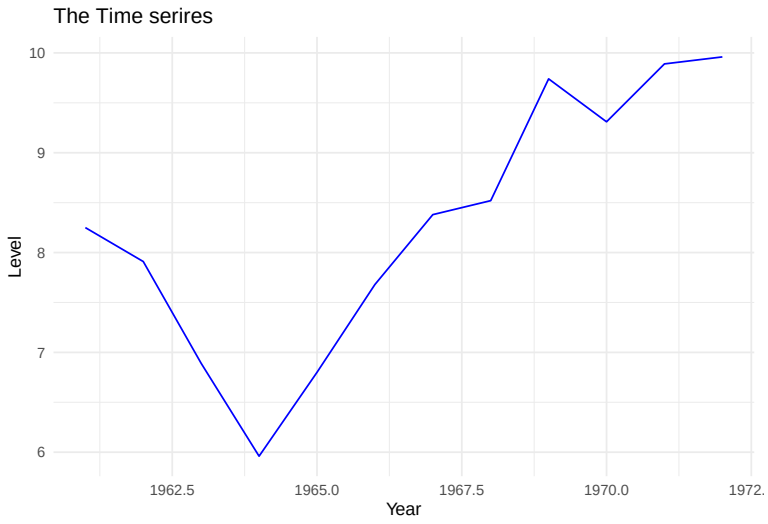
Problem 1

1.1 About the Practical

We have a data on the level of lake Huron for the year 1961 to 1972. Here we are asked to predict the level for the year 1974 and 1975

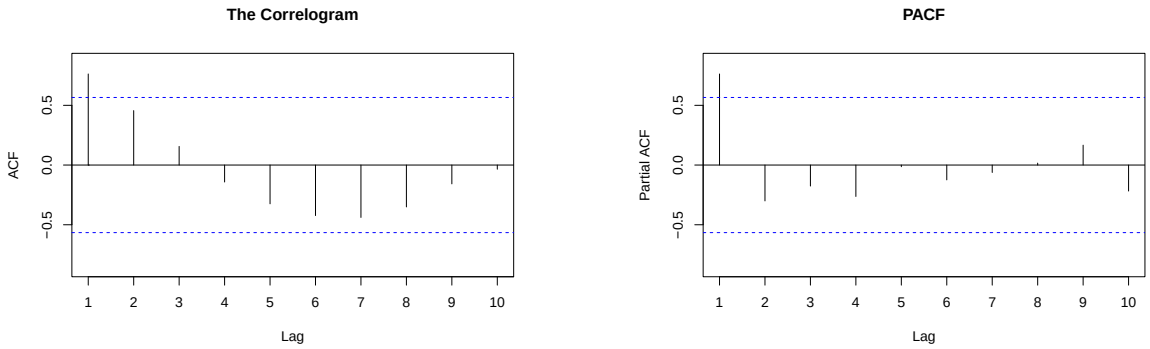
1.2 Analysis

For a time series analysis it's better to start with the visualization of the data. As per the given data, we have the following visual,



Its is a yearly data, so the presence of **seasonality** is avoidable here. Note that since it is a data of water level for lake, it shouldn't be show any **trend** which is also the case for this data.

1.2.1 Correlogram



The correlogram for the ACF and PACF of the data shows that there is only one significant spike at lag point 1, all the other spikes are insignificant, which tells us that there is hardly any trend in the data.

Also the Correlogram indicates that the data may be fitted by a **MA(1)** or a **AR(1)** process.

1.2.2 Fitting Alternative models

To predict for the future values it is necessary to fit a suitable model on our data set. Here we will use **Box-Jenkins** approach to fit a **ARIMA** model to our data.

ADF Test

To get an idea about the differencing order for the ARIMA model we perform use the Augmented Dickey Fuller Test on the data set.
On performing the test we have the result that after 2 differencing we getting a stationary series, so the values of $d = 2$ for the ARIMA model.

Estimates of *p* and *q*

From our correlogram we mainly get an idea about the order of MA and AR. The above above shows that are only 1 significant spike at lag point 1 on both the ACF and PACF plot. SO, we may assume that the MA and AR order will be 1 for both the cases. i.e. *p=1, q=1*

Model Fitting

Note that for a time series data, the models with close parameter estimates are almost fit similar the data.
So we will try different alternative models, and will choose that one which has the less AIC.

In the following table the values of AICs are shown for the different alternative models.

Sl. No.	Models	AIC values
1	ARIMA(1,0,0)	32.60485
2	ARIMA(0,0,1)	37.33643
3	ARIMA(0,1,1)	28.14554
4	ARIMA(1,1,0)	28.02052
5	ARIMA(0,2,1)	28.59037
6	ARIMA(1,2,0)	28.95715
7	ARIMA(1,1,1)	29.97675
8	ARIMA(1,2,1)	30.44888
9	ARIMA(1,0,1)	33.82283

Observation

From the table we observed that the 4th model i.e. **ARIMA(1,1,0)** has the less AIC, so we will prefer to choose that model for our forecasting purpose.

As we are asked to choose 3 alternative models we will consider those which have less AICs for the rest of the two model i.e. **ARIMA(0,1,1)**, **ARIMA(0,2,1)**

Model 1: ARIMA(1,1,0)

On fitting ARIMA(1,1,0) we have the the estimate of AR parameter as $\alpha_1 = 0.2179$, with the estimate of the σ^2 as 0.56 .

So the final model is now,

$$(Y_t - Y_{t-1}) = 0.2179 \times (Y_{t-1} - Y_{t-2})$$

Model 2: ARIMA(0,1,1)

On fitting ARIMA(0,1,1) we have the the estimate of MA parameter as $\beta_1 = 0.1687$, with the estimate of the σ^2 as 0.5769 .

Model 3: ARIMA(0,2,1)

On fitting ARIMA(0,2,1) we have the the estimate of MA parameter as $\beta_1 = -0.5903$, with the estimate of the σ^2 as 00.728.

1.2.3 Test for Randomness of Residuals

After fitting a model it is very much important to check whether the residuals are random or there may be any pattern within the residuals which can be further be modelled.
There are different methods to check whether the residuals are random or not but here we will use 2 statistical test namely *Portmanteau Test* and *Ljung Box Test* for testing the randomness of the residuals.
Since we will consider the ARIMA(1,1,0) model for our reference, so we will check the randomness of the residuals for this model only.

Portmanteau Test

On performing Portmanteau Test on the residuals we have a **p value of 0.998**, which is larger that 0.05, so at 5% level we have to accept the fact that the residuals are random.

Ljung-Box Test

Meanwhile if we also perform a Ljung Box test, we got a **p value of 0.97**, which is again larger than 0.05, so at 5% level we need to accept the same that we accept in the Portmanteau Test.

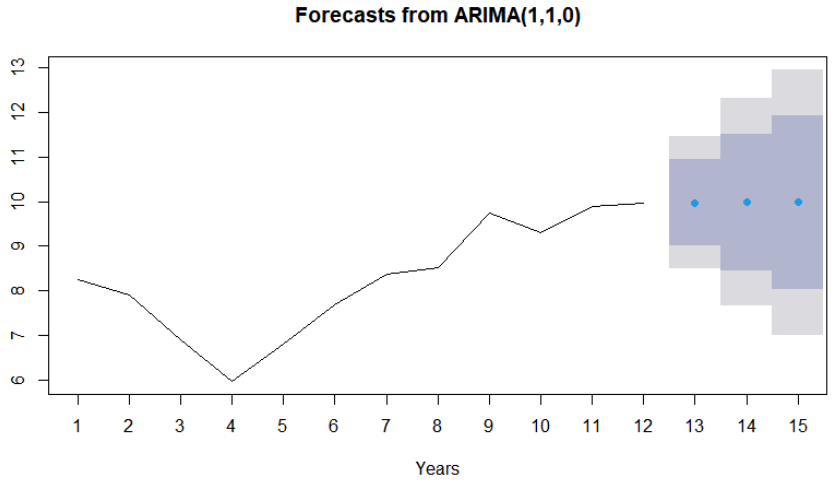
As the tests indicates that the residuals are totally random, so further they can't be modelled, so we will stick our original decision of choosing ARIMA(1,1,0) model for

forecasting.

1.2.4 Forecast

To predict the water level of the lake for the year 1974 and 1975, we will use ARIMA(1,1,0) model as mentioned above. Using One step ahead forecasting we have the following outcomes for the year 1974 and 1975,

Year	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1974	9.978573	8.454785	11.50236	7.648141	12.30901
1975	9.979297	8.025067	11.93353	6.990560	12.96803



The deep blue region indicates the 80% confidence interval for the forecasted value whereas the light blue region indicates the 95% confidence region for the forecast

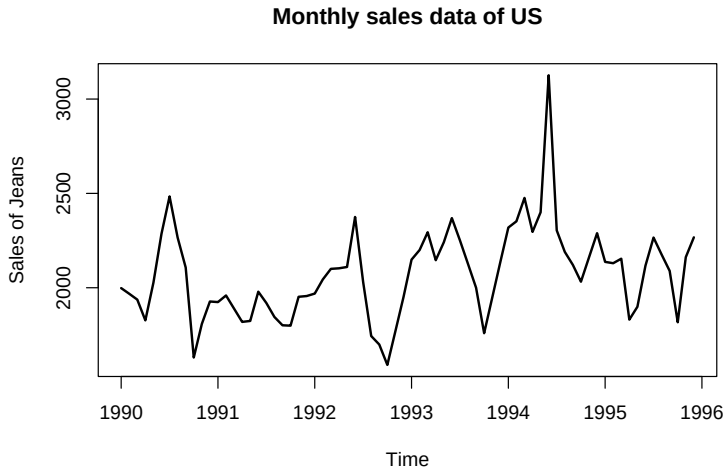
Problem 2

2.1 About the Practical

We have a monthly sales data of jeans in the US for 1990 to 1995. We need to use several forecasting model for forecasting the sales for the subsequent year.

2.2 Analysis

We have a monthly data on the sales of jeans in US. Before going in depth analysis we will first visualize the series.



From the plot it is clear that there is a seasonal effect in the sales of jeans in US , and in the span of this 6 years the sales got very small upward trend. Based on this data we will try to forecast sales for the very next year.
For the purpose of forecasting we will use 3 methods —

1. Exponential Smoothing
2. Holt Winter Forecasting
3. Box-Jenkins Model

2.2.1 Exponential Smoothing

Exponential Smoothing is a very basic method of forecasting in which the predicted values for the next time point is a convex combination of current original data value and current predicted value.

On fitting the data using simple Exponential Smoothing, we got the smoothing parameter

$$\widehat{\alpha} = 0.8695$$

The fitted values by the model is given is the following diagram.

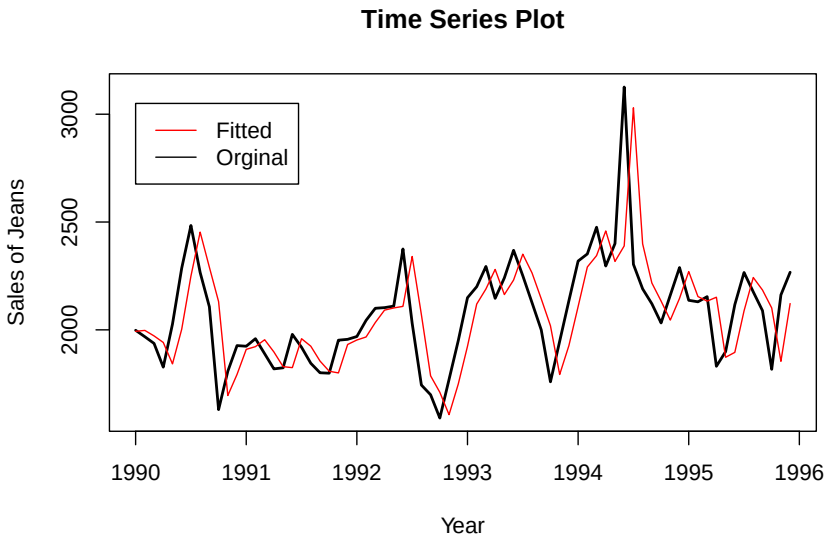
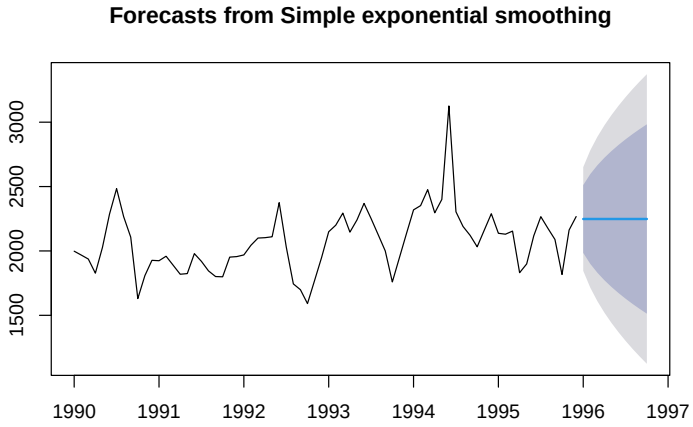


Figure 2.1: Fitted data by Exponential Smoothing

By using the above model the forecasted sales for the year 1996 is tabulated below with their 80% and 95% confidence intervals:

Forecasts:

Time Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 1996	2248.045	1984.800	2511.291	1845.447	2650.644
Feb 1996	2248.045	1899.214	2596.877	1714.553	2781.537
Mar 1996	2248.045	1830.830	2665.261	1609.969	2886.122
Apr 1996	2248.045	1772.173	2723.918	1520.261	2975.829
May 1996	2248.045	1719.992	2776.098	1440.458	3055.633
Jun 1996	2248.045	1672.524	2823.567	1367.861	3128.230
Jul 1996	2248.045	1628.682	2867.409	1300.811	3195.280
Aug 1996	2248.045	1587.745	2908.346	1238.203	3257.888
Sep 1996	2248.045	1549.202	2946.889	1179.257	3316.834
Oct 1996	2248.045	1512.677	2983.414	1123.396	3372.695



Explanation on this Forecast

From the forecast we can see that the Exponential Smoothing technique gives a very rough forecast for the future as it doesn't consider the trend and seasonal component for the purpose of prediction, only based the present data values, the seasonal effect is totally absent in the forecast and we get a flat line in the visual plot.

As the time point goes far from present the confidence level of the prediction increases very much which is quite obvious as we try to predict for far future our confidence will automatically becomes low as the circumstances at that time point will not remains the same. The light blue area is indicating the 95% confidence interval and the dark blue area is indicating the 80% of the confidence interval for the prediction.

2.2.2 Holt Winter Method

Somewhat a better method than Exponential Smoothing for forecasting a time series is **Holt Winter Method** which takes into the account of *trend* and *seasonality* for the prediction.

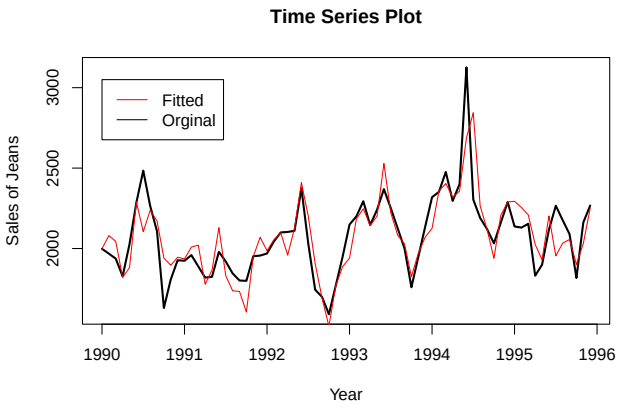
On fitting the Holt winter model in the sales data set, we come across the estimator of the model parameters as

$$\widehat{\alpha} = 0.7738$$

$$\widehat{\beta} = 1 \times 10^{-4}$$

$$\widehat{\gamma} = 1 \times 10^{-4}$$

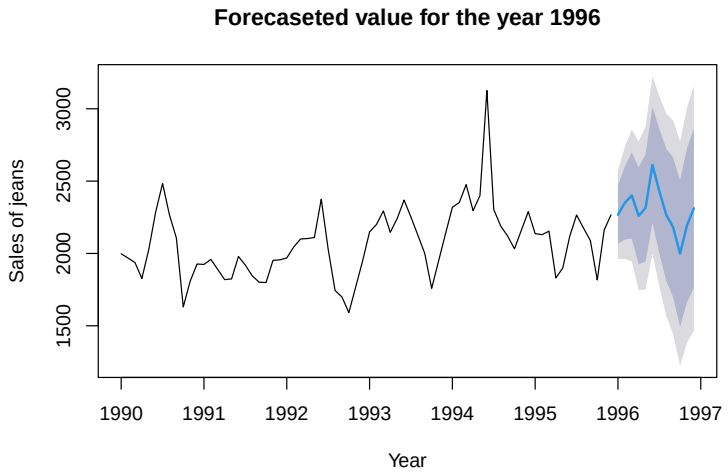
The plot compares the original series and the fitted time series by Holt Winter method.



Note that, the **RSS** for the Holt Winter Model is very much lesser than that of Exponential smoothing which also quite trivial as this method uses much information from the series.
Based on this method the forecasts of the year 1996 is given as:

Forecasts:

Time Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 1996	2267.796	2067.184	2468.408	1960.9858	2574.606
Feb 1996	2349.958	2096.282	2603.635	1961.9938	2737.923
Mar 1996	2400.562	2103.133	2697.992	1945.6832	2855.442
Apr 1996	2259.767	1924.233	2595.301	1746.6116	2772.922
May 1996	2315.181	1945.440	2684.921	1749.7110	2880.650
Jun 1996	2611.619	2210.571	3012.667	1998.2693	3224.969
Jul 1996	2429.313	1999.223	2859.402	1771.5475	3087.078
Aug 1996	2267.473	1810.176	2724.771	1568.0974	2966.849
Sep 1996	2179.787	1696.806	2662.768	1441.1312	2918.443
Oct 1996	1998.913	1491.540	2506.285	1222.9530	2774.872
Nov 1996	2194.277	1663.627	2724.927	1382.7175	3005.836
Dec 1996	2312.027	1759.072	2864.981	1466.3561	3157.697



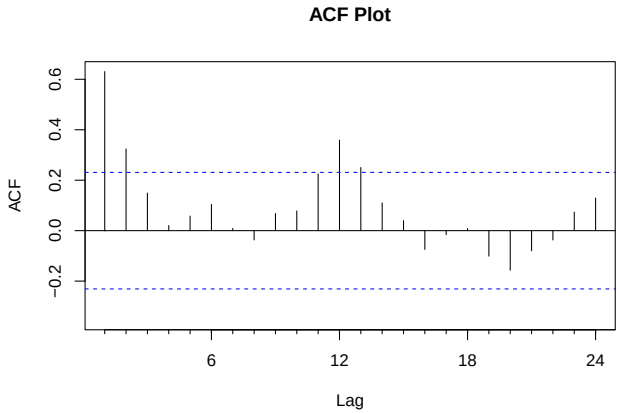
Explanation on this Forecast

Unlike Exponential Smoothing since Holt Winter takes into account of the seasonal component and trend of the series. That's why we can observe the seasonal pattern in the forecasts and helps the prediction to follow the trend of sales for the subsequent year.

As the method is much better and uses much more information than Exponential Smoothing, the confidence interval does not increases that much even for a far future. A Seasonal pattern is always helpful to make prediction for the next period with some better confidence.

2.2.3 Box-Jenkins Model

A much better approach to fit a time series for forecasting is to use a Box-Jenkins Model i.e. a *ARIMA* or *SARIMA* model.
As it is a monthly data, and has seasonality, its better to fit a *SARIMA*(*p,d,q,P,D,Q*) model. So to fit such model we need to find out the choice of p,d,q,P,D,Q . Mostly these choices are on our personal beliefs based on the ACF and PACF plot of the data.



From the ACF plot of the data shows that, the series is non stationary of seasonality, so before looking for MA or AR component we need to make it stationary.

Choice of d

The differencing order `d` is mainly useful to make the series trend stationary. By the virtue of ADF test we have **d=1**.

Choice of D

`D` is the seasonal differencing order, which makes the series seasonal stationary. Here even after `D=2`, the series comes out to be non-stationary, and due to small amount of data we can't further increase `D`. So we will use a grid search to find optimal `D`.

Choice of p,q,P,Q

From the ACF and PACF plots of the differenced series that we get an rought idea for remaining choices, such as **p=1, q=1 , P=1, Q=1 or 2** .

The SARIMA Model

For a time series analysis close alternative models some times gives a better result. That's why we fitted few alternative models with different permutation of these choices of (p,d,q,P,D,Q) and choose that one which produce a least AIC value.
In this respect we come with a SARIMA(1,1,1„1,0,1) as our final model.

Fitting of SARIMA(1,1,1„1,0,1):

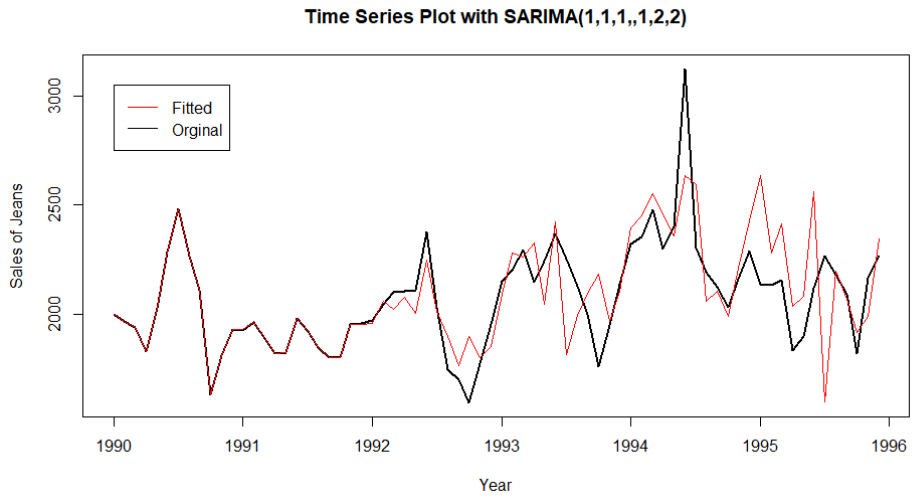
```
Coefficients:
```

```
          ar1      ma1     sar1     sma1    sma2
      ------------------------------------------
      0.7416  -0.9152  -0.6503  -0.5685  0.0938
s.e.  0.2069   0.1648   0.1896   0.3140  0.4439

AIC=677.08   AICc=679.18   BIC=688.18
```

Observation and Conclusion

The above model has least AIC among the alternative model. And the model of parameter is 5, which is not so high. Also the **Ljung Box test** for the model comes out in the favour of the null hypothesis of randomness of the residuals, so no more further modelling is needed upon this model. So finally the model is good enough to make forecasts. The fitted value for the above model is shown in the following diagram.

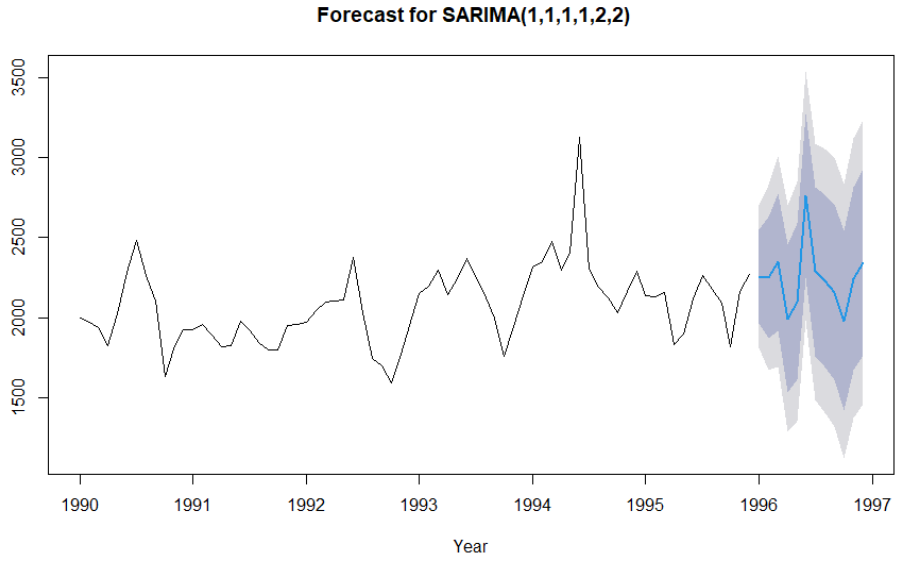


Forecast

To forecast the sales for the subsequent year we use the above SARIMA model, as we mentioned above.

```
 Forecasts:
------------
  Time Point       Forecast    Lo 80      Hi 80       Lo 95      Hi 95
 ---------------------------------------------------------------------------
   Jan 1996       2256.272   1964.707   2547.837   1810.363   2702.182
   Feb 1996       2249.248   1870.972   2627.524   1670.724   2827.772
   Mar 1996       2348.044   1918.400   2777.689   1690.960   3005.129
   Apr 1996       1994.013   1529.718   2458.308   1283.935   2704.091
   May 1996       2099.695   1610.089   2589.301   1350.908   2848.482
   Jun 1996       2761.867   2252.624   3271.110   1983.046   3540.688
   Jul 1996       2287.115   1761.869   2812.360   1483.821   3090.408
   Aug 1996       2232.514   1693.689   2771.340   1408.452   3056.577
   Sep 1996       2160.103   1609.361   2710.844   1317.816   3002.390
   Oct 1996       1979.333   1417.845   2540.822   1120.611   2838.056
```

```
Nov 1996       2246.873   1675.476   2818.269   1372.997   3120.748
Dec 1996       2345.474   1764.780   2926.168   1457.379   3233.569
-------------------------------------------------------------------
```



Explanation on this Forecast

The Box-Jenkins method is different from the other two methods we have used earlier, But still the forecasts are close to previous forecasts. Fitting a SARIMA model allows us to incorporate the seasonal effect which is very important in sales prospective.

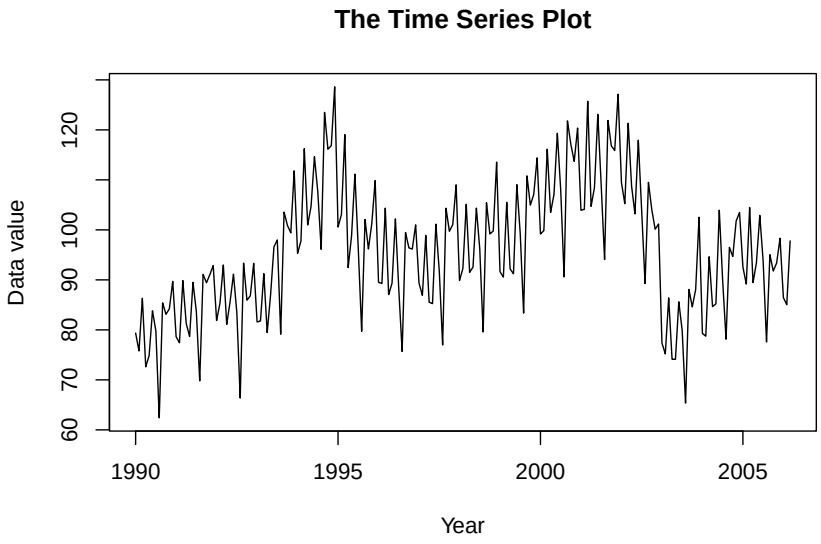
Problem 3

3.1 About the Practical

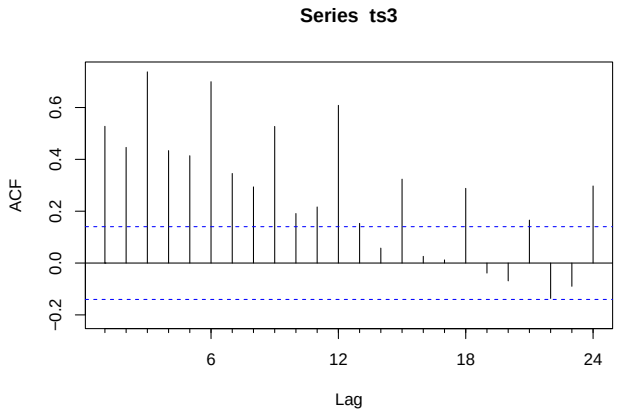
We have a time series data for the year 1990 to 2006 , we have to fit a suitable Box-Jenkins Model to the data.

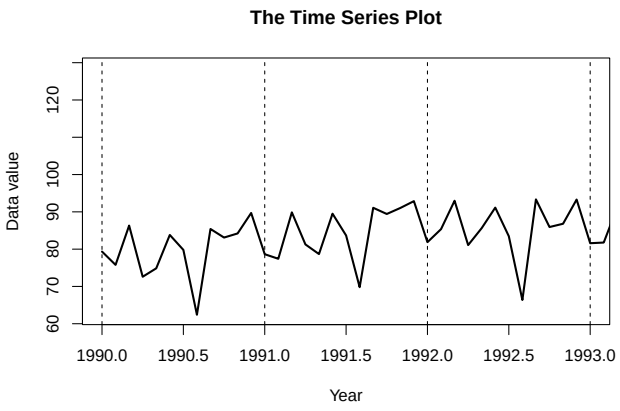
3.2 Analysis

From the tabulated data its very difficult to observe any pattern within the data so its better to have a visual idea about the plot.



So, their is a clear seasonal effect and even some trend in the time series. It will be more clear if we have a look at the ACF plot of the data and the series plot with a zoom view of the data.





Idea about the Series

From the ACF plot we can comment that the data is non-stationary in both trend and seasonality, which we can also assume from the raw time series plot and we can see a seasonal pattern in a period of 12 months.

So, to fit this series with a suitable Box-Jenkins Model, our first task is to make the series stationary and then to fit a SARIMA model.

3.2.1 Fitting of SARIMA model

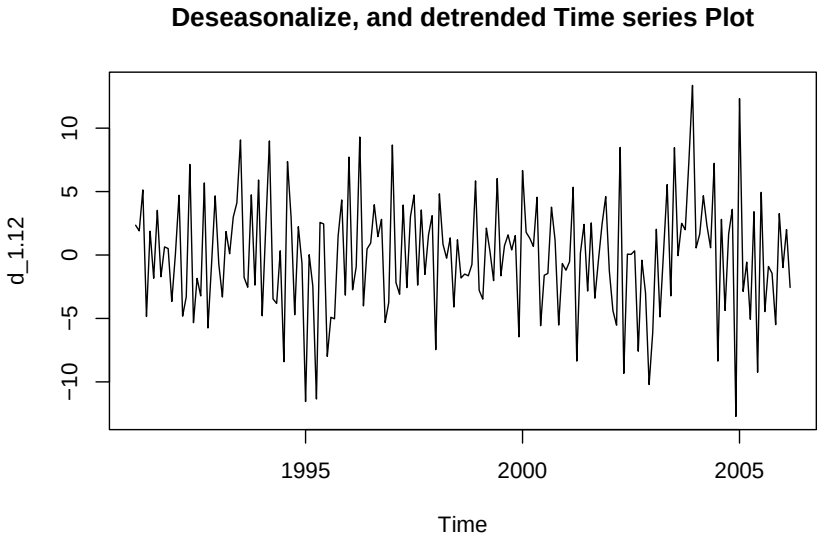
To fit a Seasonal ARIMA model, we need to have an idea about the order of the model p,d,q,P,D,Q from the series.

Idea about 'd'

The differencing order `d` is mainly useful to make the series trend stationary. By the virtue of ADF test we have **d=1**.

Idea about 'D'

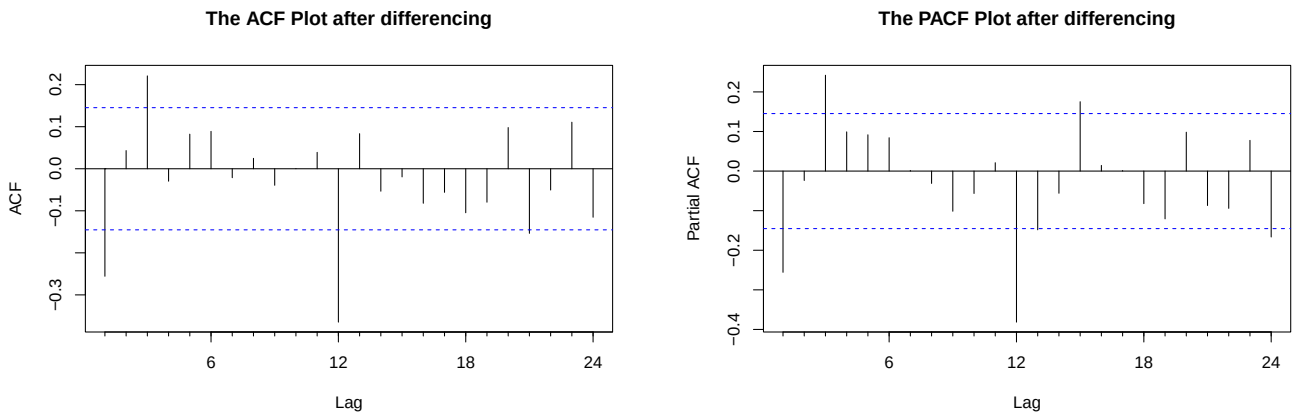
On taking the difference of order **s, the seasonal period**, the data becomes stationary on seasonality as per the ADF test. So here **D=1**.



Idea about 'p,q,P,Q'

After detrending and deseasonalizing the series we have an idea of the trend and seasonal MA order from the ACF plot and trend and seasonal order from the PACF plot.

We can have an idea that p may have choice 1/2/3 , q may be 1/2/3 , P may be 1/2/3/4 , Q may be 1



In time series analysis alternative models may result better that's why we fit several alternative models with different parmutation of p,d,q,P,D,Q and choose the model which has least AIC.

Preceding in this manner we have come with the **SARIMA(3,1,1,0,1,1)** model which has a least AIC.

Fitting SARIMA(3,1,1,0,1,1)

On fitting the SARIMA(3,1,1,0,1,1) model on the data set, we have the parameter estimates as :

Output:

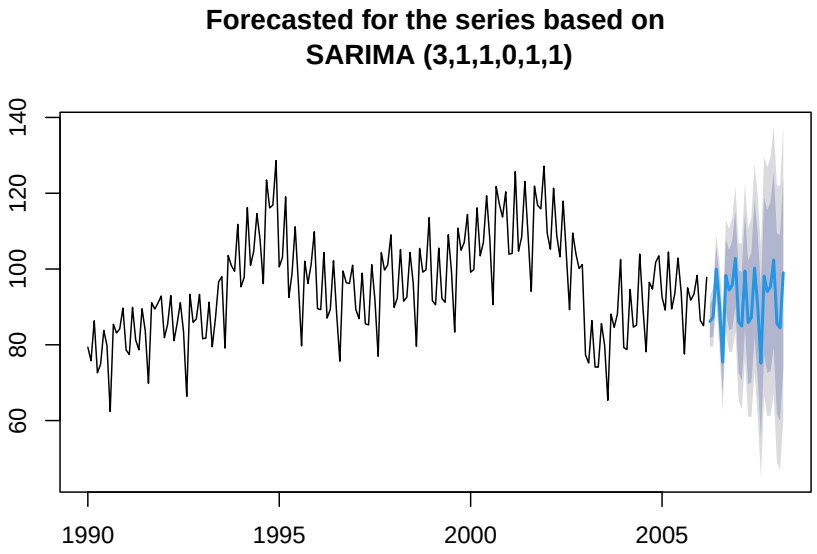
```
 Coefficients:
--------------
         ar1      ar2      ar3      ma1      sma1
       0.0653   0.1066   0.3595  -0.4381   -0.8425
s.e.   0.2062   0.0934   0.0701   0.2223    0.0753
      ----------------------------------------
AIC=981.6   AICc=982.08   BIC=1000.83
```

Observation and Conclusion

The above model has least AIC among the alternative model. And the model of parameter is 5, which is not so high. Also the **Ljung Box test** for the model comes out in the favour of the null hypothesis of randomness of the residuals, so no more further modelling is needed upon this model. So finally the model is good enough to make forecast.

Forecast

Using the above model we have the following forecast of the series. The Sky Blue line is the true prediction and the light blue area is the 95% confidence interval of the prediction.



Appendix

The overall analysis is performed in R software, and the necessary output is discussed here. The data set and the codes can be found in the following Github repository.

Source code : https://github.com/SoumaryaBasak/Time_series_analysis.git