Dissertation

Name: Subhasish Basak

Roll: 3-14-15-0446

Supervisor's name: Prof. Debjit Sengupta

Working Title: "Time Series analysis of gold price"

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Subhasish Basak

I. Introduction

Gold is one of the most valuable "precious metals" on earth and its price has remarkable importance on the global monetary system, as throughout history it has been used as a relative standard for currency equivalents. The price of gold is undoubtedly a significant index in the global economy and thus from economists to investors all requires a proper analysis of this valuable index to take right decisions in their respective fields. This study emphasizes on the application of various tools & methodologies of "Time series analysis" in analysing, modelling & forecasting gold price.

II. Data

This study is based on a secondary time series data of monthly gold price (in USD per ounce) ranging from January 1980 to December 2017. The data set containing a total of 456 data points, collected from the website of investing.com.

III. Statistical keywords & methodologies used

1.Time series data:

Data observed collected or recorded over time is called Time series data. For this dissertation the collected dataset on monthly gold price is an example of time series data.

2.Time series plot:

The simplest mode of diagrammatic representation of Time series data is the use of Time series plot. Taking two perpendicular axes of co-ordinates, the vertical one for the variable under study (here, monthly gold price) & the other for time points, each pair of values is plotted. The resulting set of points constitutes the Time series plot.

In my case the primary interest is to study the changes in gold price over the considered time period and to identify the relation between the components of the time series through this time series plot.

3. Stationary & Non-stationary component of a Time series:

Any time series is generally composed of a Non-stationary part N_t and a Stationary part S_t. If $S_t=0$ the time series is said to be purely Non-stationary. If $N_t=0$ then it is called a purely stationary series. Most time series observed in practice are Non-stationary & Stationary both superimposed.

Non-stationary as the name suggests, the changes in the values of a time series variable over a period of time is due to the Non-stationary component present in that time series. Analysis of Non-stationary in time series calls for:

- Identification of the components of which the time series is made up of.
- Isolation and measurement of the effects of these components separately and independently holding the other effects constant.

A time series Y_t is said to be strongly stationary if

$(Y_{t1}, Y_{t2}, Y_{t3}, \ldots\ldots, Y_{tk}) \xrightarrow{D} (Y_{t1+h}, Y_{t2+h}, Y_{t3+h}, \ldots\ldots, Y_{tk+h})$

i.e. $(Y_{t1}, Y_{t2}, Y_{t3}, \ldots\ldots, Y_{tk})$ converges in distribution to $(Y_{t1+h}, Y_{t2+h}, Y_{t3+h}, \ldots\ldots, Y_{tk+h})$

For any time points $t_1, t_2, \ldots, t_k$ and $\forall h$

A time series Y_t is said to be weakly stationary if its statistical properties do not change over time i.e.

- $E(Y_t) = E(Y_{t+h}) = \mu$ (say) $\forall h$
- $V(Y_t) = \sigma^2$ (say) is constant
- $Cov(Y_t, Y_{t+h}) = \gamma_h$ (say), i.e covariance depends only on lag.

4. Method of Classical decomposition:

To analyse the Non-stationary part of the time series in hand, this method is applied. In the Classical approach of Time series analysis it is assumed that any time series (Y_t) is composed of four major components viz.

- Trend (T_t)
- Seasonal component (S_t)
- Cyclical component (C_t)
- Irregular component (I_t)

$$Y_t = f(T_t, S_t, C_t, I_t)$$

The idea is to identify the functional relationship between these components by visual inspection & graphical analysis of the data and estimate each of these components separately using proper statistical tools. Then using the functional relationship, the estimated components are combined together to provide a estimate of the original time series.

4. a. **Trend**:

Trend is the smooth, regular, long term movement of the time series observed over time. It may be an upward, downward movement or remain at a constant level, but sudden of

frequent changes are incompatible with the idea of trend.

4.b. **Seasonal component**:

The seasonal component is the periodic movement in a time series which recurs or repeats at regular intervals of time and where the period is not longer than a year.

4.c. **Cyclical component**:

The Cyclical component is the oscillatory movement in time series, the period of oscillation being more than one. The length of a cycle and also the intensity of fluctuation may vary from one cycle to another.

4.d. **Irregular component**:

It is the component which is either wholly unaccountable or are caused by such unforseen events as was, floods, strikes etc.

5. Method of Simple Moving Averages(SMA):

The Method of Simple Moving Average is used to estimate the Trend component and also to average out the Seasonal or Cyclical components , if present in the time series. To compute a simple moving average (SMA) of a Time series of period k, the first k values are taken & their simple arithmetic mean is computed. Then the first value is left and the k+1 th value is included and again the mean is calculated. This process is repeats until the last k observations are arrived. Each computed mean corresponds to the middle most value of the k values. In case K is even an additional 2-pt SMA is required to centre the moving averages. In the estimation of trend, SMA is appropriate when the trend is linear but in case of a non- linear trend proper weights are used to construct weighted moving averages. In case monthly data is provided for several years (assuming there is no cyclical component present in the given time

series) a 12-pt SMA is appropriate for averaging out the seasonal component from the data, so that one is left with the trend component only.

6. Method of Monthly Averages:

This method is used to estimate the seasonal component from a time series comprising of "Seasonal" & "Irregular component" only. For monthly time series data provided for "T" years, the data is stored into a 12xT table. Now for an additive (multiplicative) model the column wise Arithmetic means (Geometric mean) , less (divided by) the adjustment factor i.e. the grand Arithmetic mean (Geometric mean) provides the estimates for the seasonal components, described as follows,

Consider a monthly data (y_{it}) for T years is provided, which is to be stored into a 12×T table. It is assumed that the given data comprises of only seasonal & irregular component, and then a multiplicative model will be as follows:

$$y_{it} = S_i \times I_{it}$$

Here,

y_{it} : Time series data corresponding to i th month & t th year. i=1(1)12 ; t=1(1)T

S_i : Seasonal index for i th season.

I_{it} : Irregular component corresponding to i th month & t th year.

Now for this multiplicative model column-wise Geometric mean is taken,

$$\left(\prod_{t=1}^{T} y_{it}\right)^{1/T} = \left(\prod_{t=1}^{T} S_i \times I_{it}\right)^{1/T}$$

$$\Rightarrow A_i = S_i \times \bar{I}_{io} \quad [\text{where, } \bar{I}_{io} = \left(\prod_{t=1}^{T} I_{it}\right)^{1/T} = 1 \ \forall i=1(1)12]$$

It is assumed that the average seasonal effect within a year should be 1, i.e. $\left(\prod_{i=1}^{12} S_i\right)^{1/12} = 1$

Hence the adjustment factor A= $\left(\prod_{i=1}^{12} A\,i\right)^{1/12} (\neq 1)$, is used to get the estimates of the seasonal indices as , $\hat{s}_i$ = Ai / A , ∀i=1(1)12

7. Stochastic process:

It is collection of random variables; more precisely it is a process which evolves in a random way where randomness can occur in either of the 2 ways:

- When the process will evolve is random in nature
- How the process will evolve is random in nature

8. Augmented Dicky-Fuller (ADF) test for stationarity:

The Augmented Dickey–Fuller test (ADF), tests the null hypothesis that a unit root is present in a time series sample. A unit root is a feature of some stochastic processes such that, a linear stochastic process has a unit root if 1 is a root of the process's characteristic equation. Such a process is non-stationary but does not always have a trend. Thus it is equivalent to state the Hypothesis of interest of an ADF test as,

H0: The series is non stationary.

H1: The series is stationary.

For a given time series y_t ,the ADF test considers the model,

$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \ldots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$

Where, α is a constant, β is the coefficient on a time trend and γ is the lag order of the autoregressive process. Imposing the constraints $\alpha = 0$ and $\beta = 0$ corresponds to modelling a random walk and using the constraint $\beta = 0$ corresponds to modelling a random walk with a drift. The unit root test is then equivalent to test,

H0: $\gamma = 0$

H1: $\gamma < 0$.

The appropriate test statistic is given by, $D = \hat{\gamma} / S.E(\hat{\gamma})$

Test rule: Reject H0 iff $D_{obs} <$ D-crit (obtained from Dickey–Fuller table) .

The intuition behind the test is that if the series is integrated then the lagged level of the series y_{t-1} provide no relevant information in predicting the change in y_t besides the one obtained in the lagged changes Δy_{t-k}. In this case the null hypothesis is not rejected.

10. Autocorrelation plot (ACF):

For a given time series Y_t the Autocorrelation function is given by

$$\rho_h = \frac{Cov(Yt, Yt+h)}{V(Yt)}$$

The graph of the autocorrelation function ρ_h against h gives the Autocorrelation plot. The graph expresses how correlation between any two values of the time series changes as the extent of separation in time changes. Observing the distinctive shape of the sample ACF one can identify the underline probability model.

Now for a given time series data $y_1, y_2, y_3, \ldots y_T$, the sample ACF is the estimate given by,

$$r_h = \hat{\rho}_h = \frac{Ch}{C0}$$

Where, $C_h = \sum_{t=1}^{T-h}$ ¿¿¿

$C_0 = \sum_{t=1}^{T}$ ¿¿¿ , $\bar{y} = \frac{\sum_{t=1}^{T} y_t}{T}$

11. Auto regressive process of order 2(AR(2)):

Let $\{X_t\}$be a purely random process i.e. X_t's are identically & independently distributed with mean 0 and variance σ^2 then $\{Y_t\}$is said to be an AR(2) process if,

$Yt = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + X_t$,

Here α_1 & α_2 are the 2 unknown parameters of the process.

with $|\alpha_1|<2$, $|\alpha_2|<1$, $\alpha_1+\alpha_2<1$, $\alpha_1-\alpha_2<1$ as restrictions for stationarity.

For an AR(2) process we have,

- $E(Y_t)=0$

No closed form for the Variance & auto covariance exists, but the ACF of an AR(2) has a damped oscillatory shape.

Now, an AR (2) process with mean μ is of the following form,

$Y_t-\mu = \alpha_1 (Y_{t-1}-\mu)+\alpha_2 (Y_{t-2}-\mu)+ X_t$

the parameters μ ,α_1,α_2 are estimated by using The Yule-Walker equation given by,

$\rho(h)= \alpha_1\rho(h-1) + \alpha_2\rho(h-2)$, where $\rho(h)$ is the ACF of the process.

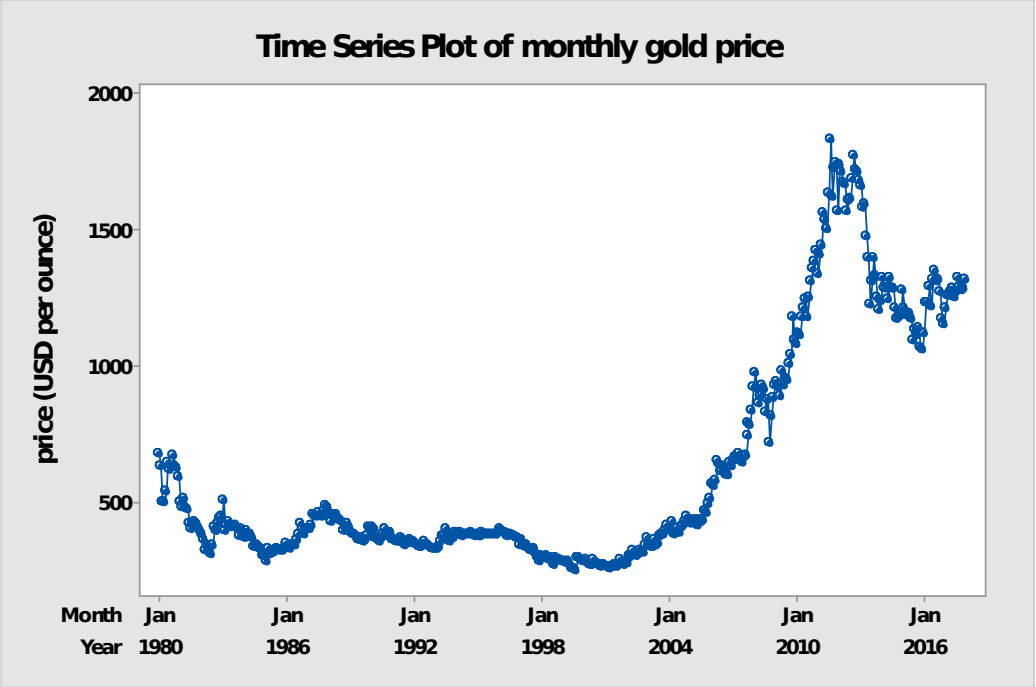
Putting k=1, 2 & simplifying we get the estimates of α_1 & α_2 as,

$\hat{\alpha}_1=\frac{r1(1-r2)}{1-r1^2}$, $\hat{\alpha}_2=\frac{r2-r1^2}{1-r1^2}$, $\hat{\mu} = \bar{y}$

Here, r_h denotes the sample ACF & using the above estimates an AR(2) process can be fitted to a stationary time series.

IV. Results & conclusions:

1. To study the changes in gold price over the considered time period and to identify the relation between the components of the time series, time series plot is constructed as follows:



Conclusions drawn;

The plot shows an up trend, concave upwards till 2012. The trend is reversed after attaining a peak later ward.

No identifiable cyclical fluctuation is observed in the data, hence in the later analysis I have dropped the cyclical component, assuming its absence.

The components of time series viz. Trend, seasonal & irregular seems to be interdependent as each of them is supposed to have significant effect on others. Hence a multiplicative model is considered as follows:

$$Y_t = Tt \times S_t \times I_t$$

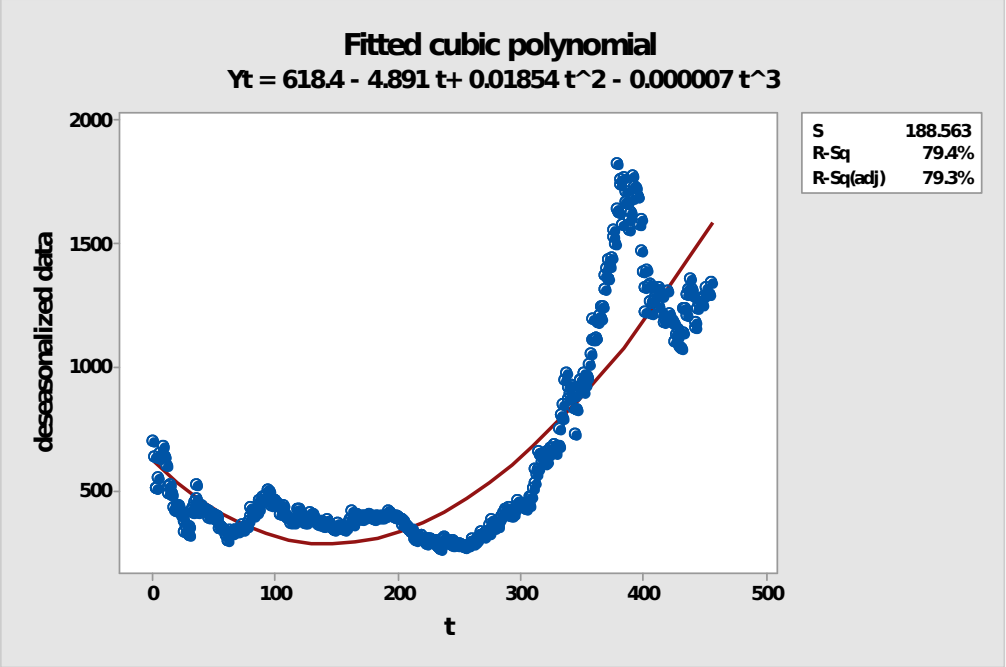
2. To smooth out the effects of seasonal components from the monthly data, a simple moving average of 12 points is used. Further to centre the moving averages an additional 2 point simple moving average is used. These averages obtained, give an estimate of the trend component. Then the time series is divided by the estimated values to obtain the detrended series, on which the method of Monthly averages is applied to get the following estimates of Seasonal indices:

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Seasonal Indices	0.98 499	1.00 469	1.00 399	1.00 310	1.00 757	1.00 648	0.99 600	1.00 612	1.00 149	0.99 592	0.99 443	0.99 520

3. Given the estimates of Seasonal component the original time series is deseasonalised to obtain a series containing only trend & irregular component. This series is fitted with a polynomial in time points (t) of degree 3. The parameters of the model are estimated by the method of Ordinary Least Squares (OLS), minimizing the error sum of squares. The fitted trend line is given by,

$$\hat{T}_t = 618.4 - 4.891 \times t + 0.01854 \times t\text{^}2 - 0.000007 \times t\text{^}3 \quad \text{.....(1)}$$

This fitted polynomial serves as a forecasting formula to forecast the trend values of the time series data on gold price.



From the graph it is observed that the trend line gives a moderate fit to the deseasonalized gold price.

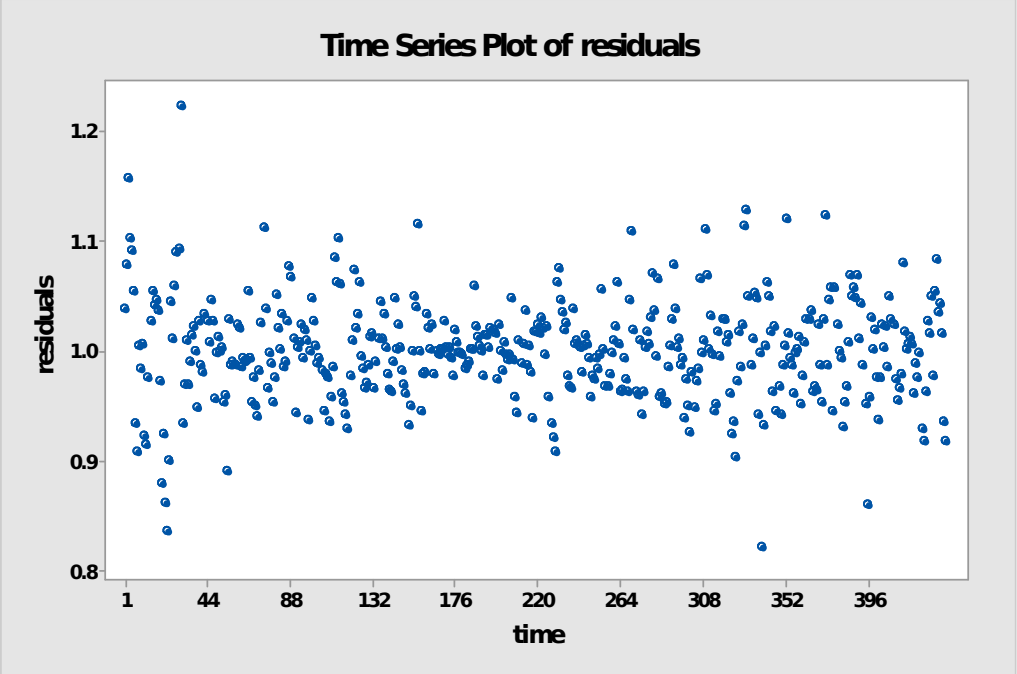
- The R-sq value for this fit is given by: 0.79

  which implies a moderate fit.

4. Having the Seasonal indices estimated by method of monthly averages & the trend component given by (1) a forecasting model for the Non-stationary component of the given time series can be obtained by using the multiplicative model as follows:

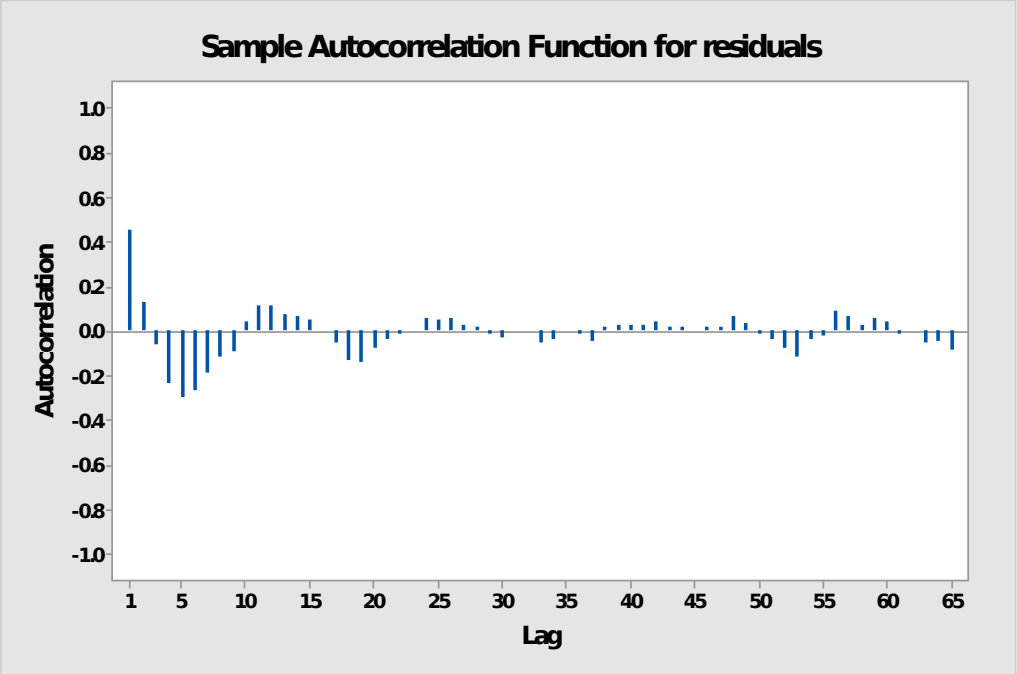
$\hat{Y}_t = \hat{T}_t \times \hat{S}_t$, for any time period t .....(*)

5. Given the estimates of trend & seasonal components the original time series is detrended & deseasonalised to obtain the residual series. The time series plot of the residual series is given as follows,



The time series plot of the residuals indicates stationarity since the residuals are randomly scattered around a central line.

The sample Autocorrelation plot (ACF) of the residual series is given as follows,



Clearly, the ACF shape is damped oscillatory type, which gives a clear indication of Stationarity of the residual series.

Apart from graphical investigation of ACF the following tests are performed to check the stationarity of the residual series. The ADF test yields the following results:

- Value of test statistic : -10.806
- p-value :0.01 < 0.05

Hence we reject H0 against H1 at 5% level of significance and conclude that the obtained residual series is stationary.

6. After transforming the given time series into a stationary one, the task is to model it with appropriate stochastic processes. Although from the shape of the ACF which is a damped oscillatory curve, clearly indicates that an AR(2) process will be appropriate here, I have fitted both AR(1) & AR(2) process to the stationary series and then calculated the Mean Absolute Error (MAE) & Mean Absolute Percentage Error (MAPE) to choose between the two fits of AR(1) & AR(2) process.

Where,

MAE=$\frac{1}{N}\sum_{t=1}^{N}$ ¿$et \vee$¿¿ , gives the mean absolute magnitude of error committed.

MAPE=¿ , gives the average percentage absolute deviation of the

forecast series from the actual series. For both the measures lower the values better the fit.

The following table shows the values of these statistics:

Process	MAE	MAPE
AR(1)	0.02993	0.030029
AR(2)	0.02974	0.029845

From the values of MAE & MAPE it is clear that an AR(2) process will be appropriate in this case. The fit of AR(2) process on the residual series is as follows:

$$Y_t = 0.4989 \times Y_{t-1} - .0913 \times Y_{t-2} + 0.3506 + X_t \quad \text{.....(**)}$$

The point forecasts using the above model is given by,(when value of the series is known upto Y_t)

$$\hat{Y}_{t+1} = 0.4989 \times Y_t - .0913 \times Y_{t-1} + 0.3506$$

$$\hat{Y}_{t+2} = 0.4989 \times \hat{Y}_{t+1} - .0913 \times Y_t + 0.3506$$

$$\hat{Y}_{t+3} = 0.4989 \times \hat{Y}_{t+2} - .0913 \times \hat{Y}_{t+1} + 0.3506$$

And so on.

7. The analysis till now provides isolation and separate estimation of the “Non-stationary” & “Stationary” component of the time series data. The estimation of the “Non-stationary” part is done using the “Method of Classical Decomposition” and the analysis of the “Stationary” part is done by the fitting of AR (2) process to the residual series. (*) & (**) as stated above gives forecasting formulae for the Non-stationary & Stationary part respectively. The formula combining (*) & (**) by the multiplicative model, provides a complete forecasting formula for the study variable Gold price. Using this the forecasts of gold price for the first 6 months of 2018 are computed as follows,

Month	**Jan**	**Feb**	**March**	**April**	**May**	**June**
Forecast price	1553.89	1592.64	1599.21	1605.48	1620.39	1626.39
Actual price	1343	1317	1322	-	-	-

Clearly the forecasts are not satisfactory since their deviations from the actual prices are significantly high. This might have been so due to some severe drawback of the analysing techniques adopted which can be investigated as follows;

- In estimating the seasonal indices by the method of Monthly Averages, the original time series data is detrended using a SMA of 12- pt, which requires the assumption of Linearity of the underline trend. But the time series plot clearly depicts the non linearity of the trend component. Thus the use of weighted Moving averages is recommended.
- The assumption of Multiplicative model ($Y_t = Tt \times S_t \times I_t$) among the components of the time series, does not have any logical or economic base related to this context. In practice where most of the time series components remain entangled together in "Mixed" type models (for e.g. $Y_t = (Tt \times S_t) + ¿ I_t$) , the choice of an pure Multiplicative model is not recommended.
- Moreover in estimating the Trend component, a cubic polynomial is used; the selection which also has no economic or logical base related to the context of gold price. The choice of such a polynomial is made since it has a high R^2 value compared to other mathematical models.
- One of the crucial assumptions made in the Classical Decomposition approach is: a time series is supposed to be composed of 4 major components only viz. Trend, Seasonal component, cyclical component and the Irregular component .Which may not be the case in practice. A time series is necessarily a stochastic process which is influenced by many more factors. Thus the use of the Method of Classical Decomposition is recommended only when the 4 above stated components show their sound presence in the data.

V. Future scope of this study

One may use advanced forecasting techniques like ARIMA, in which a time series is transformed by suitable methods (using Lag operator, Difference operator, Log transformation etc.) into a stationary series and then various probability models are fitted.

VI. References:

- investing.com (data collection purpose)
- https://www.rdocumentation.org/packages/aTSA/versions/3.1.2/topics/adf.test (for performing Augmented Dicky-Fuller (ADF) test for stationarity)
- Support.minitab.com
- Wikipedia
- Fundamentals of statistics : Gun,Gupta,Dasgupta (Vol II)