TIME SERIES ANALYSIS, FORECASTING AND METHOD OPTIMIZATION OF OZONE LAYER DENSITY IN TEXAS, USA

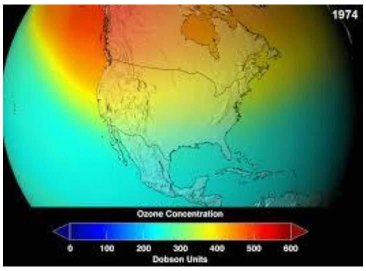
19

➢ ***INTRODUCTION AND OBJECTIVE:***

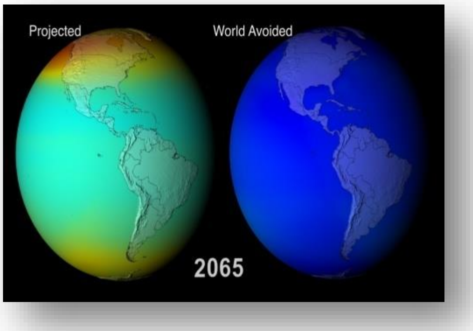
Ozone is a gas in the atmosphere that protects everything living on the Earth from risky ultraviolet (UV) rays from the Sun. Without the **layer** of **ozone** in the atmosphere, it would be very difficult for anything to survive on the surface.

The **ozone** layer acts as a filter for the shorter wavelength and highly hazardous ultraviolet **radiation** (UVR) from the sun, **protecting** life on Earth from its potentially harmful effects. When the sky is clear, there is an inverse relationship between stratospheric **ozone** and solar UVR measured at **the Earth's** surface. The key role that ozone, a major component of the stratosphere, plays in how climate change occurs, and the possible implications for predictions of global warming.

In addition to its role in protecting the Earth from the Sun's harmful Ultraviolet rays, ozone are also a greenhouse gas. The ozone layer is part of a vast chemical network, and changes in environmental conditions, such as changes in temperature or the atmospheric circulation, result in changes in ozone abundance. This process is known as an atmospheric



Chemical feedback. A reduction in global surface warming of approximately 20% – equating to 1° Celsius – when compared with most models after 75 years. Ozone plays an important role on earth's temperature so prediction of ozone density is important.



And for this prediction purpose we need some suitable statistical tools. Methods like "Exponential Smoothing", "Holt Winters" and "ARMA Model" are performing effectively for this purpose. We are interested to discuss how the ozone layer density around Texas in USA changes over time. As we have historical monthly data about Ozone density, we forecast the ozone layer density in 2006(monthly) by using the time-series data from 1994 to 2005 and secondly to compare the forecasted values from different forecasting methods and models, with the original value, in order to know which method fits the original data with higher accuracy.

➢ ***Summary of the Data Set:***

- source: http://arcive.ics.uci.edu/ml/about.html

A) data Structure: In the given dataset there are three columns an d 144 present. First column denotes the Years, the column denotes the months of the corresponding years and the third column provides us the ozone density values for that corresponding month and year. It can be noted that there are the values present from the Year 1994 to 2005 on monthly basis. The ozone density values are all numerical.

The data set at a glance-

Table no: 1

year	*month*	*ozone*
1994	jan	15.38106
	feb	20.19316
	mar	26.8192
	apr	29.68339
	may	28.50184
	jun	26.17017
	jul	27.06638
	aug	27.18806
	sep	28.42477
	oct	19.88523
	nov	19.28253
	dec	14.85843
1995	jan	13.79857
	feb	16.59344
	mar	26.11295
	apr	40.12648
	may	38.27489
	jun	38.4862
	jul	29.33316
	aug	31.91513
	sep	38.14529
	oct	29.36267
	nov	20.22389
	dec	19.49466
1996	jan	21.80466
	feb	25.71142
	mar	30.59906
	apr	34.9785
	may	29.00132
	jun	28.97738
	jul	24.50603
	aug	21.12368
	sep	25.62932
	oct	19.03345
	nov	11.53844
	dec	9.120702

We have provided ozone density values of first 3 years i.e. 1994 to 1996 on monthly basis.

- ***ANALYSIS:***

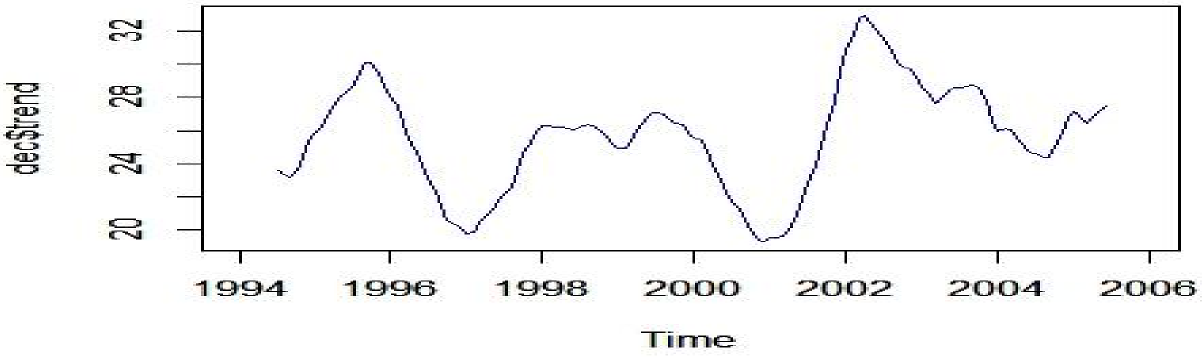
  We use three methods mainly for forecasting purpose of ozone densities. We will discuss the methods and their applications below:

A) In this section of analysis we discussed whether the natural properties of time-series such as trend, seasonality and stationary presented in our dataset or not.

- ***TREND***

Over a long period, a time series is very likely to show a definite tendency, such atendency may be a constant in direction may change direction of a constant rate, or may be characterized by shifts in direction. The fact should be emphasized that by trend is meant the smooth, regular, long-term movement of a statistical series; sudden or frequent changes either in absolute amounts or in rates of increase or decrease are quite inconsistence with idea of secular trend.

In our dataset we need to find out whether there is any trend present or not. For that purpose a graphical representation of the ozone densities is given below:



From this graph it is clear that there is no trend in the data set.

- ***SEASONALITY***

A periodic movement is one which recurs, with some degree of regularity, within a definite period. The most frequently studied periodic movement is that which occurs within a year and which is known as seasonal variation. Many time series, such as sales figures and temperature readings exhibit variations which is annual in period and the customs, habits which the people follow in different time, climate conditions, and product variations in time series data.

Here we are interested to check whether seasonality is present in dataset or not. Here we use R-programming on this purpose. The following R-codes are given below:

- <u>Seasonality R-code:</u>

```
d=read.csv("C:/Users/USER/Desktop/ozone.csv", header = F)
library(tseries)
data=ts(d)
data1=ts(as.vector(data),start=c(1994,1),end=c(2005,12), frequency =
12)
dec=decompose(data1,type = "additive")
season=dec$seasonal
```

The seasonal components obtained using R-Programming is given below in the following table no: 2

months	Seasonal values
jan	-8.400866357
feb	-4.256406432
mar	2.210327122
apr	7.928721561
may	8.088329176
june	4.135469307
july	-6.3314
aug	-3.054098622
sep	6.275493747
oct	-1.58227386
nov	-7.122350365
dec	-8.999003479

This table shows the seasonal values of the dataset.

- ***<u>STATIONATITY:</u>***

A very special class of stochastic process, called stationary process, is based on the assumption that the process is in a particular state of statistical equilibrium. A stochastic process is said to be strictly stationary if its properties are unaffected by a change of time origin. If the joint probability distribution associated with m observations $x_{t1},x_{t2},......x_{tm}$ is same as that associated with m observations x_{t1+h}, $x_{t2+h},.....$ x_{tm+h}.In statistics and econometrics, an augmented Dickey–Fuller test (ADF) tests the null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test is used, but is usually stationary or trend-stationary. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models. The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit roots at some level of confidence. The testing procedure for the ADF test is the same as for the Dickey–Fuller test but it is applied to the model

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

,where alpha is a constant, beta the coefficient on a time trend and p the lag order of the autoregressive process.

Imposing the constraints α =0 andβ=0 corresponds to modelling a random walk and using the constraint β= 0 corresponds to modelling a random walk with a drift.

Consequently, there are three main versions of the test, analogous to the ones discussed on Dickey–Fuller test (see that page for a discussion on dealing with uncertainty about including the intercept and deterministic time trend terms in the test equation.)

$$DF=\tilde{y}/ SE (\tilde{y})$$

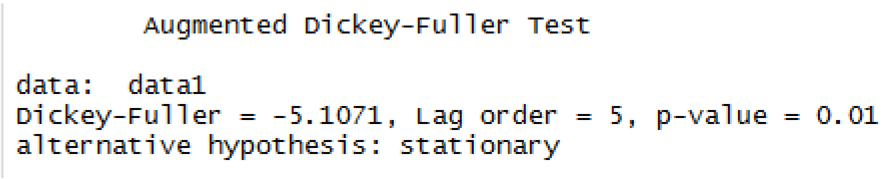
By including lags of the order p the ADF formulation allows for higher-order autoregressive processes. This means that the lag length p has to be determined when applying the test. One possible approach is to test down from high orders and examine the t-values on coefficients. An alternative approach is to examine information criteria such as the Akaike information criterion, Bayesian information criterion or the Hannan–Quinn information criterion.The unit root test is then carried out under the null hypothesis $\gamma=0$ against the alternative hypothesis of $\gamma<0$. Once a value for the test statisticis computed it can be compared to the relevant critical value for the Dickey–Fuller Test. If the test statistic is less (this test is non symmetrical so we do not consider an absolute value) than the (larger negative) critical value, then the null hypothesis of $\gamma=0$ is rejected and no unit root is present.

- *Application of augmented Dickey-Fuller test to our time-series data:*

  We have applied ADF test procedure to check if the stationarityis present or not in our dataset.The whole test procedure is performed in R-programming using the code given bellow:

```
d=read.csv("C:/Users/USER/Desktop/ozone.csv", header = F)
library(tseries)
data=ts(d)
data1=ts(as.vector(data),start=c(1994,1),end=c(2005,12), frequency = 12)
adf.test(data1)
```

The output of the above R code of the ADF test for stationarity checking looks like-



```
        Augmented Dickey-Fuller Test

data:  data1
Dickey-Fuller = -5.1071, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

From the above output result we can wee that the p-value is very low and this indicates the rejection of null hypothesis i.e. the time-series is not stationary.

As the time-series data is stationary hence there is no need to apply the difference method to make it stationary. Therefore for the ARIMA model the integrated part I is 0 and the ARIMA model automatically implies the ARMA model.

- We use three methods mainly for forecasting purpose of ozone densities. We will discuss the methods and their applications below:-

1: *Exponential Smoothing*: - This forecasting procedure, first suggested by C.C Holt in about 1958, should only be used in its basic form for non-seasonal time-series showing no systematic trend. Suppose given a non-seasonal time-series with no trend x_1, x_2 ...., x_N it is natural to take as an estimate of x_{N+1} a weighted sum of past observation and the equation is represented by:-

Smoothed value at time t = α (data at t) + (1 – α) (smoothed value at time t-1) where "α" is a smoothing factor.

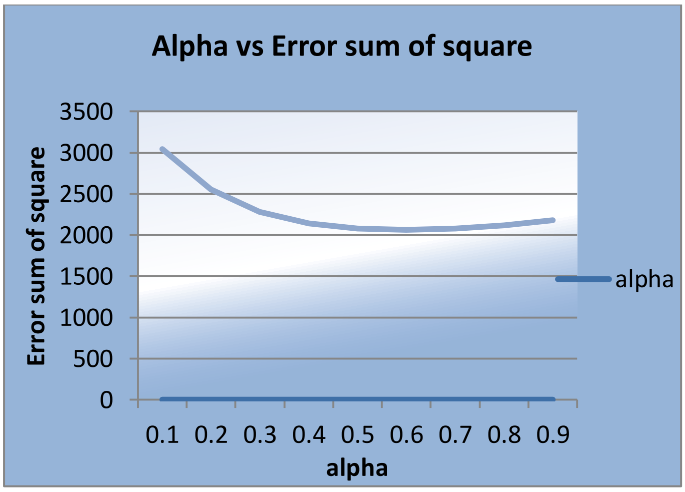
$\hat{X}$ (N, 1) = αx_N+ (1 – α) $\hat{x}$ (N-1, 1)value of α is estimated from data. The sum of squares prediction errors is computed from different values of alpha and the value is chosen which minimizes the sum of squares. With a given value of α calculate x (1, 1) =x_1 e_2=x_2-x (1, 1)

x(2,1)= α e_2+ x(1,1) e_3=x_3-x(2,1) ....

By repeating procedure for other values of alpha between 0 and 1, select value that minimize$\sum e^2$.

- <u>Application:</u>

From the previous part we have seen that there is no trend is present. But the seasonality is present in the given time-series data. As we are applying exponential smoothing method, we have to adjust the time-series by eliminating the seasonal components. Not only that, we have to make an optimum choice of $\alpha$for which the error sum of squares is minimum.



<u>Alpha</u>	***<u>Sum of error squares</u>***
0.1	3044
0.2	2555
0.3	2279
0.4	2142
0.5	2080
0.6	2065
0.7	2081
0.8	2120
0.9	2180

From the above table we have got an optimum α which has value 0.6.Hence the exponential smoothing equation will be as follows –

X (N, 1) = 0.6 x_N+ (1 – *0.6*) x (N-1, 1)

= 0.6 x_N+0.4x (N-1, 1) -------- (*)

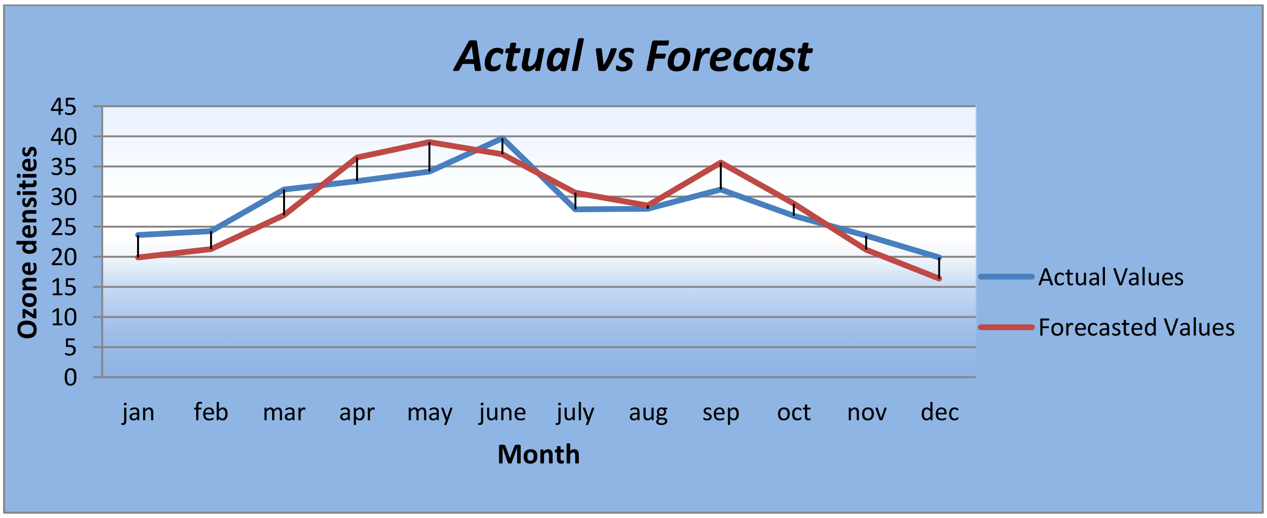
Forecast:

Using the equation marked as (*) we get the forecasted smoothed value of the ozone densities for the year 2006 on monthly basis. As we are interested to forecast the ozone densities for the year 2006, we add up the seasonal components with the smoothed values. It is noted that the seasonal components are tabulated from the table No(3). Hence the forecasted values are presented in the following table-

Table3: Forecast of Ozone densities on monthly basis in the year 2006

Months	*Actual Values*	*Seasonal components*	*Smoothed Values*	*Forecasted Values*
jan	23.6721611	-8.400866357	28.28698249	19.88611614
feb	24.22662114	-4.256406432	25.51808966	21.26168322
mar	31.24621658	2.210327122	24.74320854	26.95353567
apr	32.609397	7.928721561	28.64501337	36.57373493
may	34.16429212	8.088329176	31.02364354	39.11197272
june	39.73221865	4.135469307	32.90803269	37.043502
july	27.90559498	-6.3314	37.00254426	30.67114426
aug	27.97692382	-3.054098622	31.5443747	28.49027607
sep	31.18045038	6.275493747	29.40390417	35.67939792
oct	26.77703788	-1.58227386	30.46983189	28.88755803
nov	23.54195459	-7.122350365	28.25415548	21.13180512
dec	19.859539	-8.999003479	25.42683495	16.42783147

The following chart shows a graphical representation of actual and forecasted values of ozone densities of 2006-



The above table and graph represent the forecasted values are quite near to the actual values of ozone densities in 2006.

- Holt-Winters Forecasting :-

Exponential smoothing may be generalized to deal with time series containing trend and seasonal variation. The resulting procedure is referred to as Holt-Winters procedure. Trend and seasonal terms are introduced which are also updated by exponential smoothing. Suppose the observations are monthly and L_t, T_t, I_t denote the local level, trend and seasonal index at time t. Thus T_t is the expected increase or decrease per month in current level. And α, γ,δdenote three smoothing parameters for updating level, trend and seasonal index. The smoothing parameters are between 0 and 1.When a new observation x_i becomes available, the values of L_t, T_t, I_t are all updated. The equations are $L_t= \alpha\ (x_t/I_{t-12}) + (1- \alpha)\ (L_{t-1}+T_{t-1})$$I_t=\delta\ (x_t\ /\ L_t) + (1- \delta)\ I_{t-12}$

$T_t= \gamma(L_t-L_{t-1})+(1- \gamma)T_{t-1}$ and the forecasting model at "t" are then $x(t,k)=(L_t+kT_t)I_{t-12+k}$for k=1(1)12

- **Application:**
  In the given time-series dataset, we wish to forecast the ozone densities in 2006 by using the Holt-winters method. Since the Holt-Winters algorithm involves both the trend and seasonal components we need not to do any special treatment with trend and seasonal components.
  The whole Holt-Winters method is performed in the purpose of forecasting on the given dataset through R-Programming language. The R-codes used on this purpose are given below-

```
d=read.csv("C:/Users/USER/Desktop/ozone.csv", header = F)
library(tseries)
data=ts(d)
data1=ts(as.vector(data),start=c(1994,1),end=c(2005,12), frequency = 12)
h=HoltWinters(data1)
p=predict(h,n.ahead=12)
```

The corresponding values of parameters are evaluated by the programming language itself and their values are given below-

```
alpha: 0.5495342
beta : 0.009550782
gamma: 0.568049
```

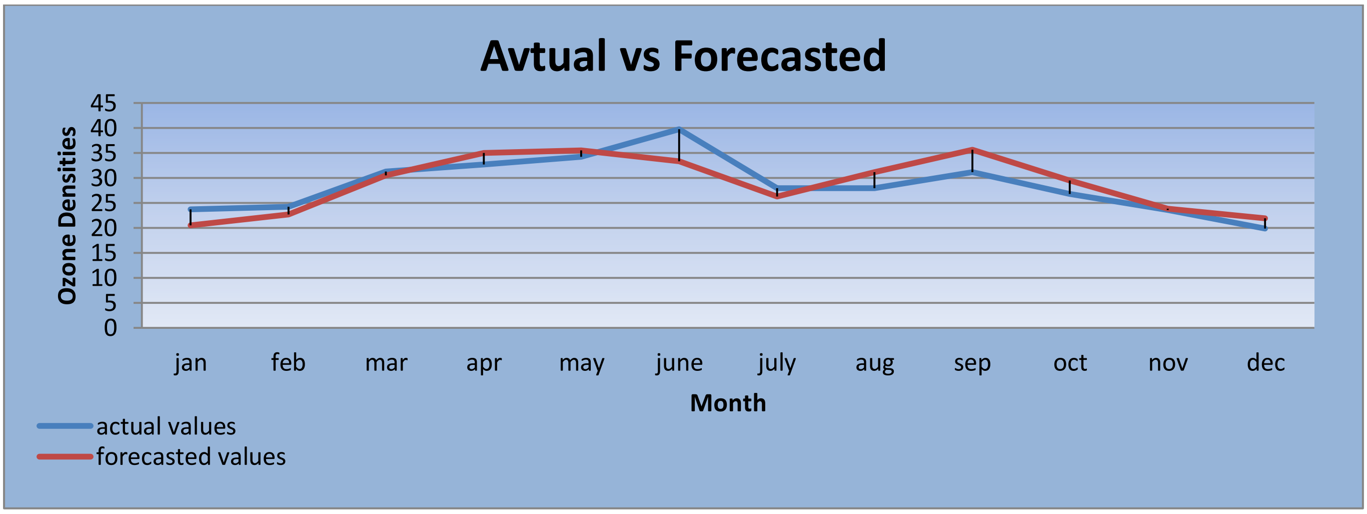
- Forecast:

Using the R-codes mentioned above we get the forecasted smoothed value of the ozone densities for the year 2006 on monthly basis. As Holt-Winters method includes seasonal components, adding seasonal components with smoothed valuesto obtain forecasted values like exponential soothing forecasting method, is not required.

*Table4:*Forecast of Ozone densities on monthly basis in the year 2006 using ***holt-winters*** forecasting method:

Months	*Actual Values*	Forecasted values
jan	23.6721611	20.43634
feb	24.22662114	22.71014
mar	31.24621658	30.44025
apr	32.609397	35.03593
may	34.16429212	35.44054
june	39.73221865	33.26808
july	27.90559498	26.26692
aug	27.97692382	31.16865
sep	31.18045038	35.64664
oct	26.77703788	29.46619
nov	23.54195459	23.8067
dec	19.859539	21.91653

The following chart shows a graphical representation of actual and forecasted values of ozone densities of 2006-



The above table and graph represent the forecasted values are quite near to the actual values of ozone densities in 2006.

➢ ***ARMA MODEL:***

the statistical analysis , autoregressive–moving-average (ARMA) models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the auto regression (AR) and the second for the moving average (MA). The general ARMA model was described in the 1951 thesis of Peter Whittle, Hypothesis testing in time series analysis, and it was popularized in the 1970 book by George E. P. Box and Gwilym Jenkins.

Given a time series of data Xt, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modelling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The model is usually referred to as the ARMA (p, q) model where p is the order of the AR part and q is the order of the MA part (as defined below).ARMA models can be estimated by using the Box–Jenkins method.

The process $\{X_t\}$ is an ARMA (p, q) process if

1. It is stationary.

2. The equation of X_t follows ARMA (p,q) process if it can be written as

$$X_t=\theta+\alpha_1X_{t-1}+\alpha_2X_{t-2}+........+\alpha_pX_{t-p}+\beta_0+\beta_1e_{t-1}+........+\beta_qe_{t-q}$$

As there is one auto-regressive, one MA series. In the above equation θ represents a constant term. In general in an ARMA (p, q) process, there will be p autoregressive and q MA terms.

➢ Application:

In the above we have discussed about s stationary of the given time-series dataset. We have noticed that the dickyfuller test shows us that there is no stationarity present. Therefore the integrated part of the ARIMA process becomes to 0. And the process is converted into a pure ARMA process. We have performed the ARMA process in our given dataset using the R programming language and the therefore the optimum choice of the orders of the ARMA process have been determined by the programming language itself. . The R-codes used on this purpose are given below-

```
d=read.csv("C:/Users/USER/Desktop/ozone.csv", header = F)
library(tseries)
data=ts(d)
data1=ts(as.vector(data),start=c(1994,1),end=c(2005,12),   frequency
= 12)
arma= auto.arima(data1)
pred_arma=predict(arma,n.ahead = 12)
```

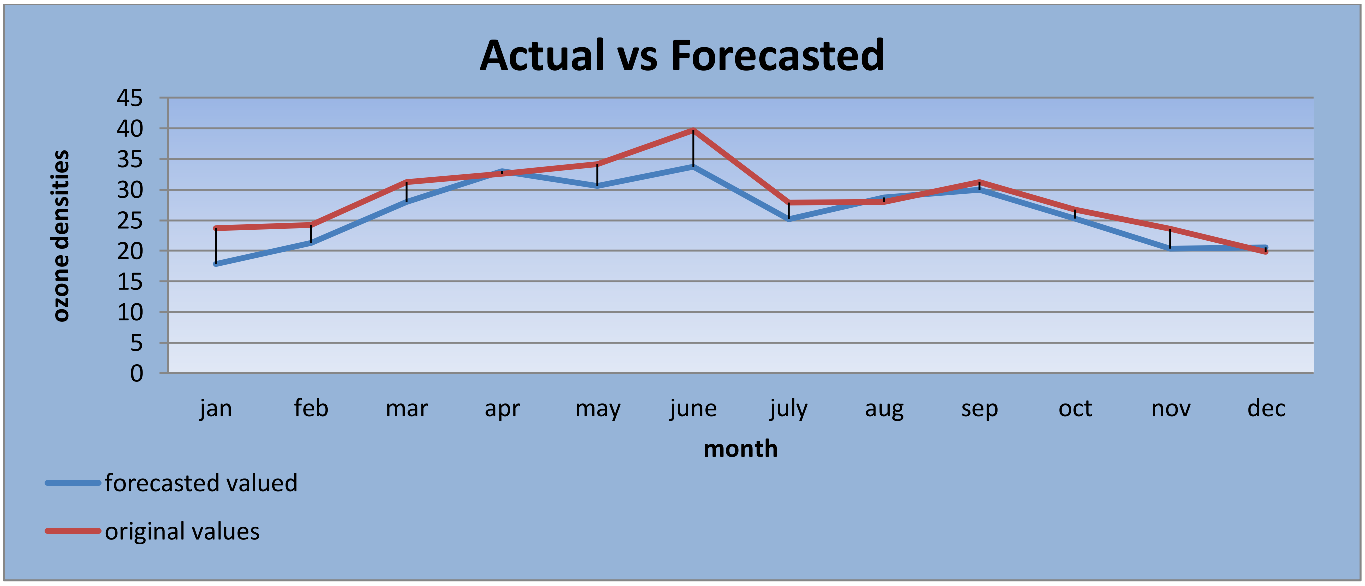
- Forecast:

Using the R-codes given above we get the forecasted smoothed value of the ozone densities for the year 2006 on monthly basis. The forecasted values of ozone densities of year 2006(monthly basis) is given below-

*Table5:*Forecast of Ozone densities on monthly basis in the year 2006 using ***ARMA*** model:

Months	*Actual Values*	Forecasted values
jan	23.6721611	17.86746
feb	24.22662114	21.31488
mar	31.24621658	27.93205
apr	32.609397	32.97895
may	34.16429212	30.61015
june	39.73221865	33.68709
july	27.90559498	25.17445
aug	27.97692382	28.75019
sep	31.18045038	29.92944
oct	26.77703788	25.27267
nov	23.54195459	20.33553
dec	19.859539	20.53406

The following chart shows a graphical representation of actual and forecasted values of ozone densities of 2006-



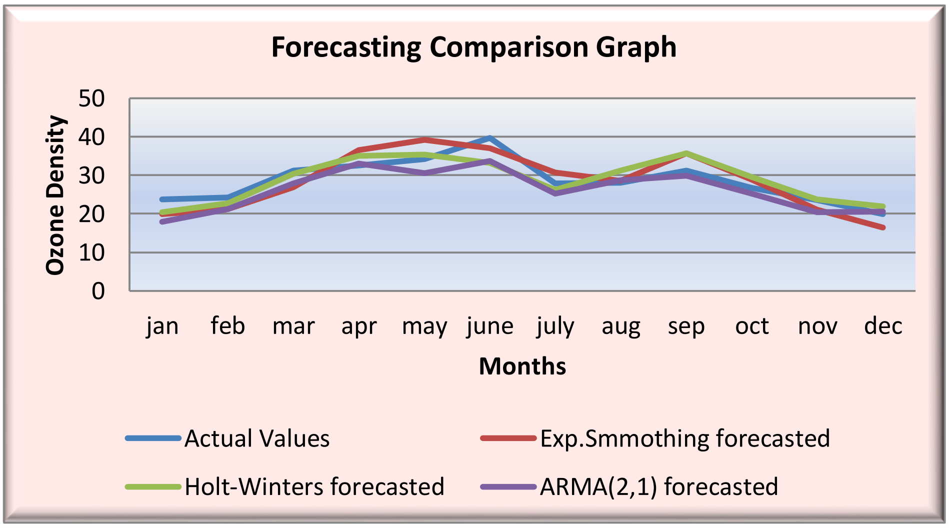
The above table and graph represent the forecasted values are quite near to the actual values of ozone densities in 2006.

- Comparative discussion of performances of above three discussed forecasting methods:

In this part we are interested to discuss the performance of three fore-casting methods viz. exponential smoothing, Holt-winters and ARMA . Here we have measured the performance of the forecasting method by calculating the Mean-square error with the original data. MSE=$\sqrt{\ }$ {1/n $\sum_{1}^{n}(y_0 - y_f)$ ^2 } Where y_0 =Original value in the given dataset. y_f =Corresponding forecasted value .Here we have provided a table in which we have discussed the comparison of the three forecasting methods:

Table No6 Performance of the three forecasting methods using MSE

Months	Actual Values	Exp.Smmothing forecasted	Holt-Winters forecasted	ARMA(2,1) forecasted	RMSE Exp. Smoothing	RMSE Holt-winters	RMSE ARMA Model
jan	23.6721611	19.88611614	20.4363389	17.86745722			
feb	24.22662114	21.26168322	22.71013844	21.3148793			
mar	31.24621658	26.95353567	30.44024967	27.93205329			
apr	32.609397	36.57373493	35.03592556	32.97895179			
may	34.16429212	39.11197272	35.44053693	30.61015122			
june	39.73221865	37.043502	33.26807685	33.68708784	3.4055	2.9871	3.2286
july	27.90559498	30.67114426	26.26691654	25.17445483			
aug	27.97692382	28.49027607	31.16865332	28.7501911			
sep	31.18045038	35.67939792	35.64663785	29.9294402			
oct	26.77703788	28.88755803	29.46619257	25.27267284			
nov	23.54195459	21.13180512	23.80669614	20.33552611			
dec	19.859539	16.42783147	21.9165282	20.53405554			



Conclusion:

We have thoroughly discussed about the time-series components in our dataset and performed three different forecasting methodologies viz. Exponential Smoothing, Holt-Winters and ARMA model to get the forecasted values of ozone densities in Texas, USA in the year 2006. We have also discussed about the performances of the above three methods by evaluating RMSE with the actual ozone density values and we see that among the three methods Holt-winters methodology shows least RMSE i.e. 2.9871.

Though there may be one or more forecasting methods useful to get the more accurate forecasted values, but in this context we come to the conclusion that Holt-Winters method is more efficient to forecast the ozone densities of Texas, USA in the year 2006 among the three forecasted methods.

Appendix:

1. Fundamental of statistics vol:2
2. Time-Series Analysis and Its applications ; Robert H.Shumway
3. Time series analysis, forecasting and control; George E.P.Box
4. http://m.youtube.com
5. http://www.wwoz.org
6. http://exceltable.com