**Name: Aashik Bharath Egnya Varahan**

**Student No.: R00182866**

***TIME SERIES PROJECT***

**Introduction:**

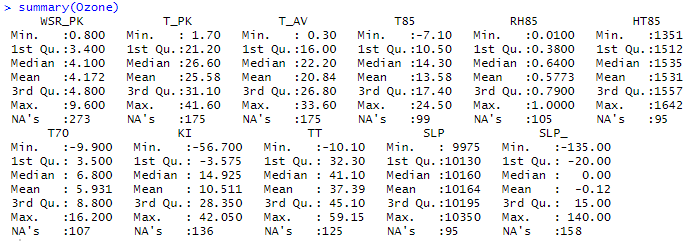
In this project, we will be focusing on Forecast of Ozone Level on a daily basis using Time Series models. The dataset used in this forecasting is obtained from UCI Repository and has data of 73 different attributes and observations for 2534 days from 1st January 1998.

Initially, Exploratory Data Analysis (EDA) will be carried out on the dataset dealing with the missing values, outliers and producing a descriptive summary of the dataset. The dataset is reduced to 11 key variables necessary for the forecasting. The time series model is built for these 11 variables using Exponential Smoothing and ARIMA modelling techniques.

**Exploratory Data Analysis:**

The dataset is imported into R for further processing. The unnecessary variables are removed from the dataset using Indexing. The columns are assigned with names as given in the dataset description. Finally, the datatype of each variable is mutated from Character type to Numeric type. Below is a table of summary statistics of the final dataset to be forecasted.

**DESCRIPTIVE SUMMARY:**



From the above table, it is evident that the dataset has many missing values and few outliers to be dealt with.

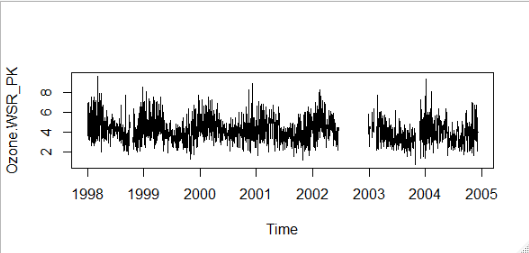
To remove the outliers and to replace NAs with values, the dataset is first converted into a time series object. Each and every key variable in the dataset is converted into a time series object separately for performing individual forecasting.

Let us now forecast each key variable one after the other.

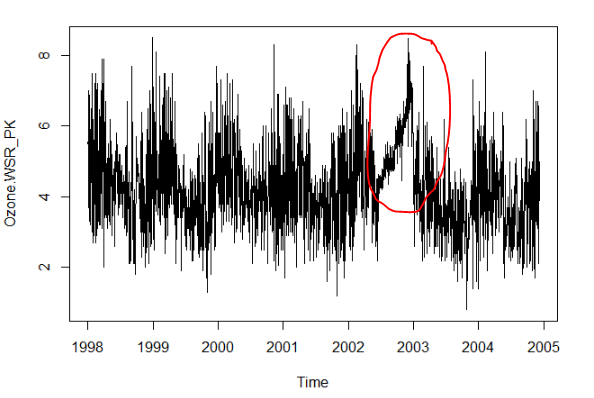
* ***WSR\_PK***

**tsclean()** function is used to remove the outliers and the missing values are replaced using linear interpolation.

***Before Removing Outliers & Dealing with NA values***



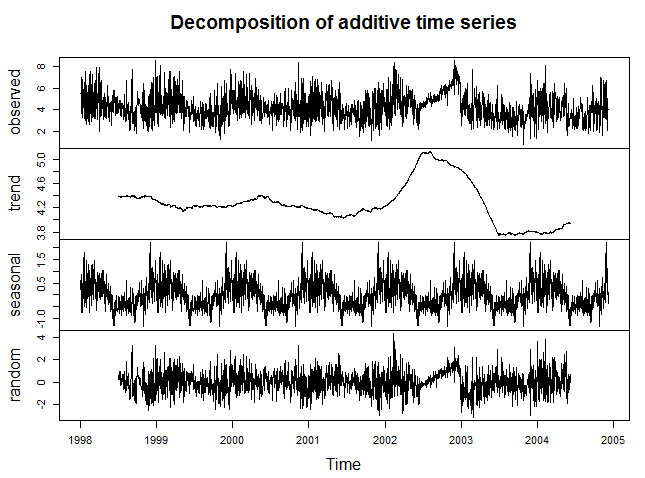
***After Removing Outliers & Dealing NA values***



From the above image, it can be observed that the missing values in this key attribute has been replaced by using **Linear Interpolation**, a method of fitting the curve using linear polynomials to construct new data points within the range of a discrete set of known data points.

**Decomposing Time Series Data**

In this, the data is decomposed or separated into its constituent components, which are trend, seasonality and an irregular component.



From the above graphs, we can observe a trend and a seasonality for this variable. There is a sinusoidal seasonality present in this variable. There is no trend till the year 2002, whereas in 2002, there is a sudden jump in the trend which is present till mid of the year 2002 and slowly decrease to the bottom in the mid of 2003.

**Test for Stationarity**

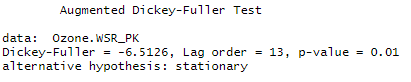
**Stationarity** is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

The hypothesis for Augmented Dicky-Fuller Test is,

H0 = Null Hypothesis -> Not Stationary

HA = Alternate Hypothesis -> Stationary

Significance Level = 0.05

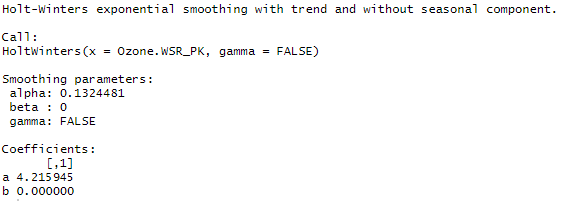


From the above statistics, we can conclude that this time series object is **stationary**.

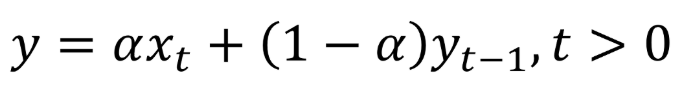
**Exponential Smoothing Model**

Since the time series for WSR\_PK variable can be described using an additive model with a sudden jump in trend and seasonality, **Holt-Winters Exponential Smoothing** is used for this variable.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.



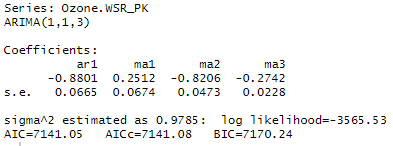
The above statistics show that the Smoothing parameters for the model has values such as alpha = 0.132, beta = 0 and gamma = 0. These parameters convey that the model has no trend and no seasonality. The alpha value is close to zero, which tells that the forecasts are based on recent previous observations.



Here, alpha is a **smoothing factor** that takes value between 0 and 1, It determines how fast the weight decreases for previous observations.

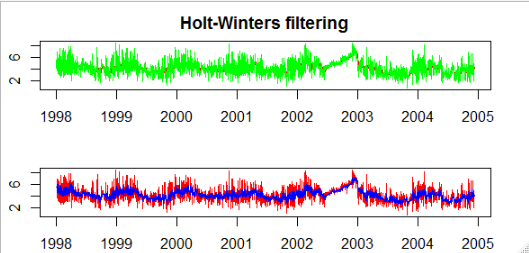
**ARIMA Model**

Since ARIMA (Autoregressive Integrated Moving Average) models are defined for stationary time series, a differencing in time series will be implemented to obtain a stationary time series. This model includes an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.



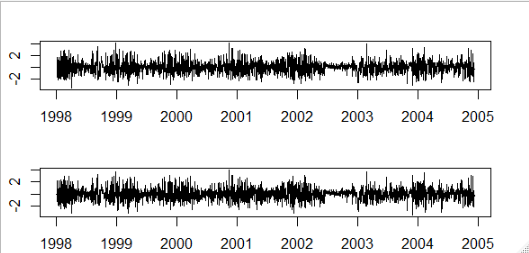
ARIMA (1,1,3) model with p=1 (partial correlogram), d=1 (differencing) and q=3 (autocorrelogram). The Goodness of Fit criteria for ARIMA models, **AIC** (Alkaline Information Criterion) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. **BIC** is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

**Observed vs Predicted Values**



From the graphs above, we can observe that the fitted values are more well defined in the ARIMA model (2nd graph) compared to the Exponential Smoothing model (1st graph). It is because, a differencing factor, partial correlogram and autocorrelogram are defined for the below ARIMA model.

**Residuals of Fitted Data**



The residuals in a time series model are the leftovers after fitting a model. In many times series models, the residuals are equal to the difference between the observations and corresponding fitted values.



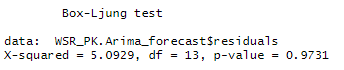
The graphs showing about residuals concludes that the residuals in this variable fulfils both the conditions, ,i.e., the residuals are not correlated, and they have zero mean. Since, both these conditions are satisfied, it can be said that the forecasting methods are proper and good.

**Fit and Measure Statistics**

The Ljung-Box test is a tool to test the lack of fit of a time series model. The test is applied to the residuals of a time series after fitting an ARMA(p,q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

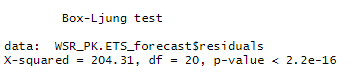
The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) of the Box Ljung Test, H0, is that our model *does not* show lack of fit (or in simple terms—the model is just fine). The [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/), Ha, is just that the model *does*show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series *isn’t*autocorrelated.

Let us now perform this test for the ARIMA model:



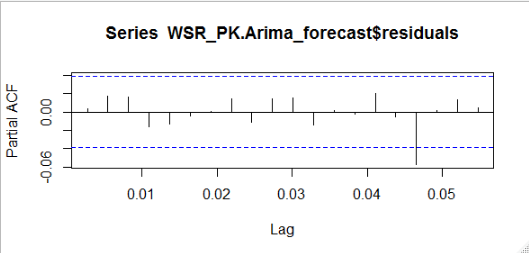
From the above test, we can find the Box-Ljung test statistic to be 5.0929, lags between 1-13 and p-value as 0.9731. This shows that the model does not show any lack of fit and its good.

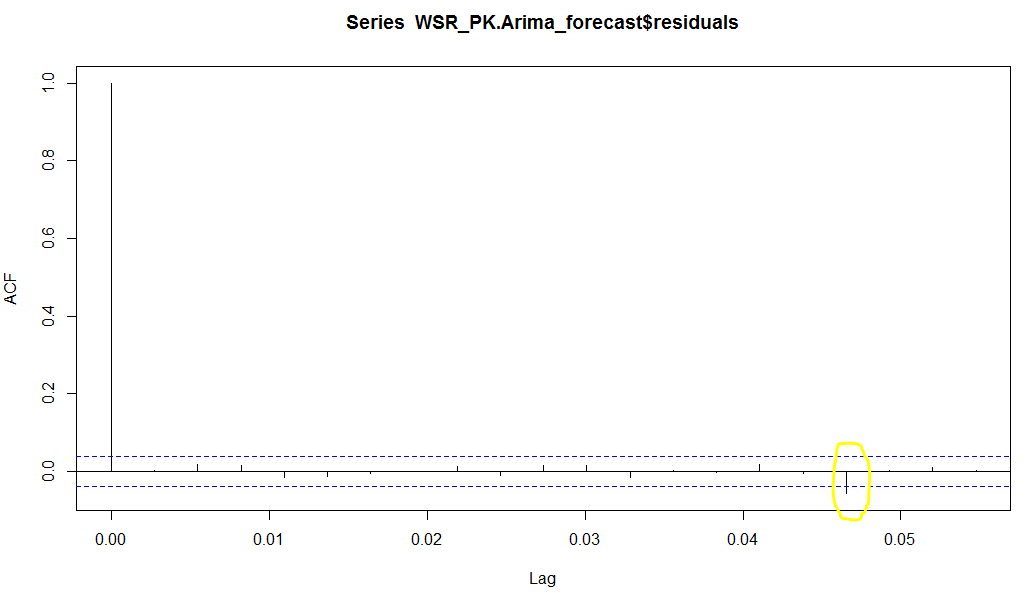
For Exponential Smoothing model:



From the above test, we can observe the p-value to be less than the significant value, which concluded that this model shows a lack of fit.

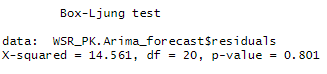
**Partial Correlation & Autocorrelogram**





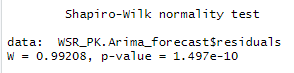
From both the correlogram plots above, we can observe that the sample autocorrelation and partial correlation for the in-sample forecast errors at lag 17 exceeds the significance bounds. However, we could expect one in 20 of the autocorrelations and partial autocorrelations for the first twenty lags to exceed the 95% significance bounds.

Let us now carry out the Ljung-Box test.



It is evident that the p-value of 0.801 indicates that there is very little evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

**Normality of Residuals**



From the above Shapiro-Wilk test for normality, we can observe that the p-value to be far less than the significance level, rejecting the null hypothesis H0 which states that the residual distribution is normal. Therefore, we can conclude that the residuals are **not normally distributed** here.

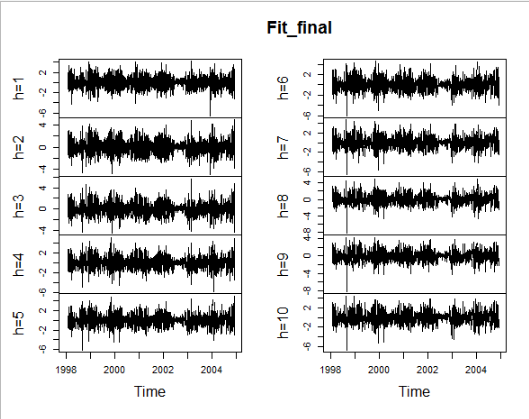
**Best Model**

From the two time-series models we used for forecasting the Ozone Level for the subsequent days, the model using ARIMA is better compared to that of the Exponential Smoothing model since the fit of ARIMA is better than Exponential Smoothing model. Also, for forecasting, the correlation between successive values of time series is important, which is measured only in ARIMA model.

The best candidate ARIMA model chosen for this variable is (1,1,3) which shows a differencing factor of 1, partial autocorrelogram tails off to 0 after lag 1 and autocorrelogram tails off to zero after lag 3.

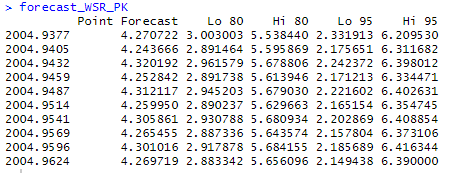
**Cross-Validation:**

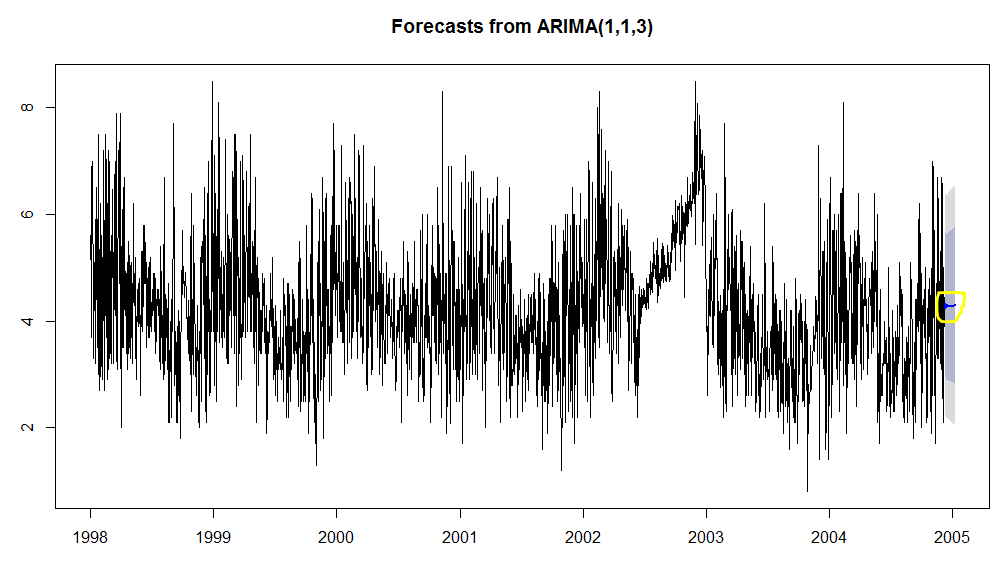
A Cross-Validation for the above ARIMA model with parameter values (p,d,q) = (1,1,3) is computed and plotted below for h= 10 (Forecast Horizon) and window = 30.



**Forecasting using ARIMA Model:**

The forecast of WSR\_PK (Peak Wind Speed) for the subsequent 10 days are as below:



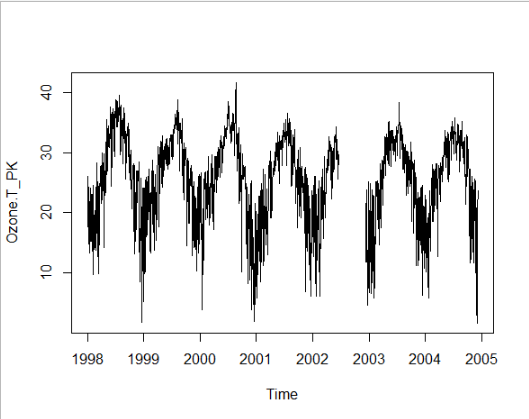


The above plot shows the forecast of the WSR\_PK variable for the next 30 subsequent days from the end of the day in dataset.

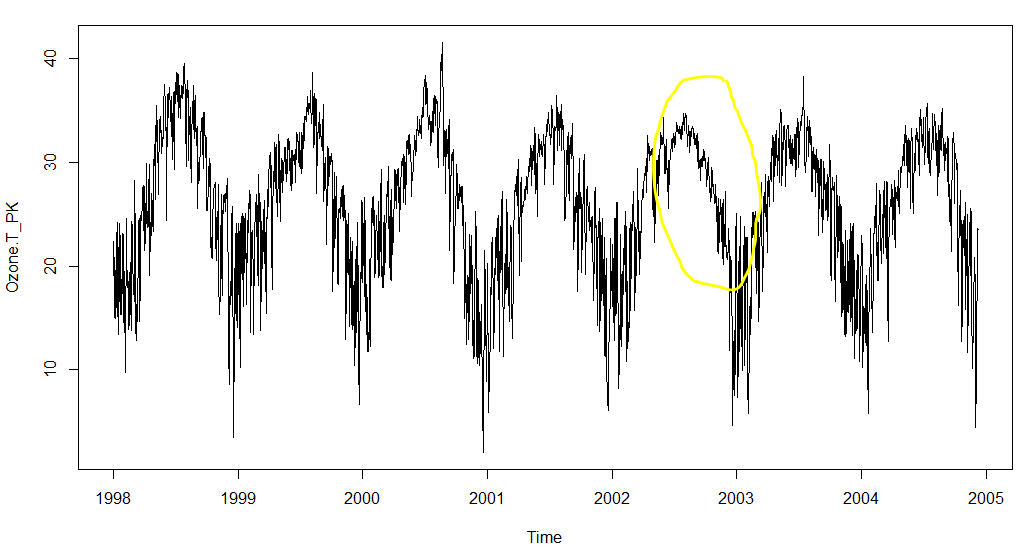
* ***T\_PK***

**tsclean()** function is used to remove the outliers and the missing values are replaced using linear interpolation.

***Before Removing Outliers & Dealing with NA values***



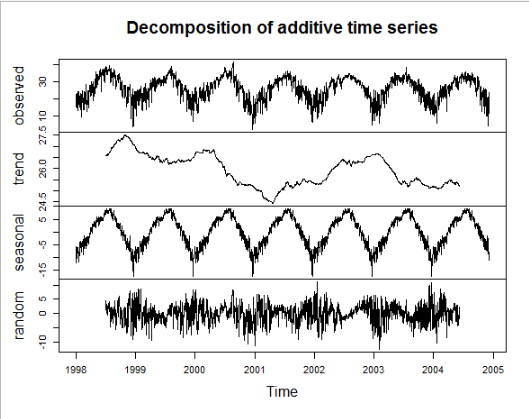
***After Removing Outliers & Dealing NA values***



From the above image, it can be observed that the missing values in this key attribute has been replaced by using **Linear Interpolation**, a method of fitting the curve using linear polynomials to construct new data points within the range of a discrete set of known data points.

**Decomposing Time Series Data**

In this, the data is decomposed or separated into its constituent components, which are trend, seasonality and an irregular component.



From the above graphs, we can observe both, a trend and a seasonality for this variable. A mountain like pattern of seasonality occurs every year. Also, we can see a decreasing trend from the year 1998 till mid of the year 2001 and then the trend increases, stabilizes after some time.

**Test for Stationarity**

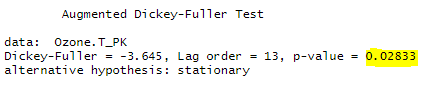
**Stationarity** is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

The hypothesis for Augmented Dicky-Fuller Test is,

H0 = Null Hypothesis

HA = Alternate Hypothesis

Significance Level = 0.05

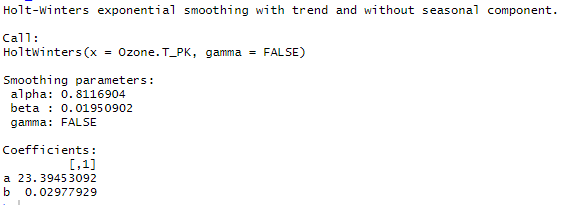


From the above statistics, we can conclude that this time series object, T\_PK is **stationary**.

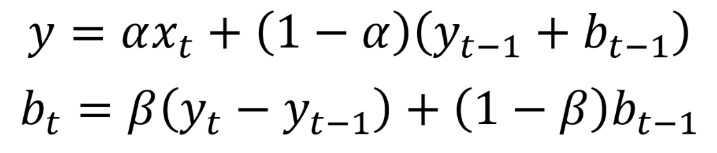
**Exponential Smoothing Model**

Since the time series for T\_PK variable can be described using an additive model with a sudden jump in trend and seasonality, **Holt-Winters Exponential Smoothing** is used for this variable.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.



The above statistics show that the Smoothing parameters for the model has values such as alpha = 0.811, beta = 0.02 and gamma = 0. These parameters convey that the model has a trend but no seasonality. The alpha value is close to one, which tells that the forecasts are based on current observation.

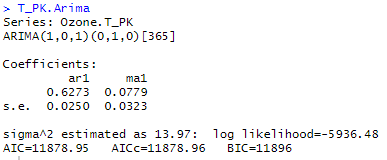


Here, alpha is a **smoothing factor** that takes value between 0 and 1, It determines how fast the weight decreases for previous observations.

The beta is the **trend smoothing factor**, and it takes values between 0 and 1.

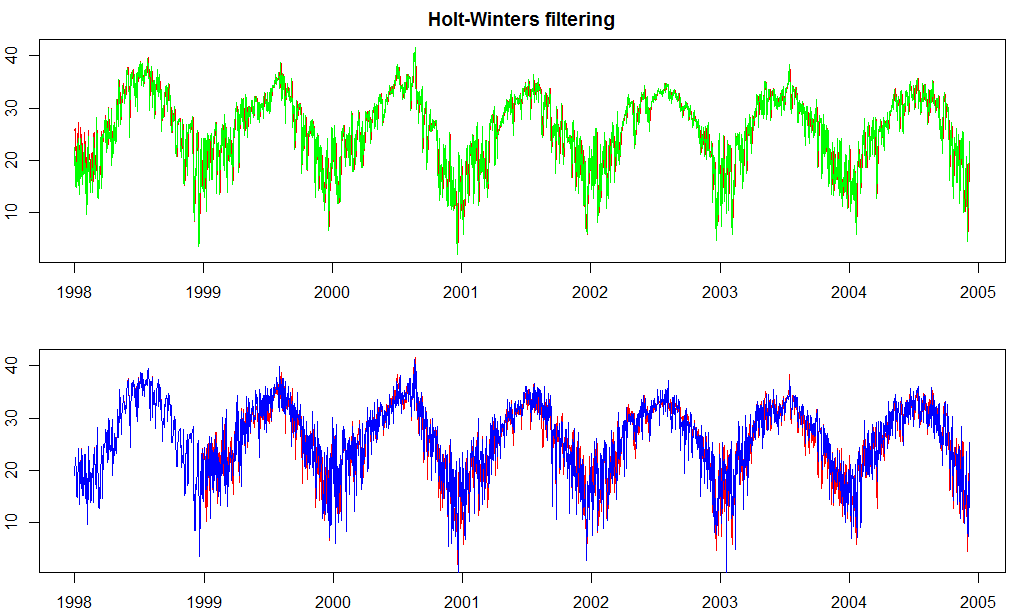
**ARIMA Model**

Since ARIMA (Autoregressive Integrated Moving Average) models are defined for stationary time series, a differencing in time series will be implemented to obtain a stationary time series. This model includes an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.



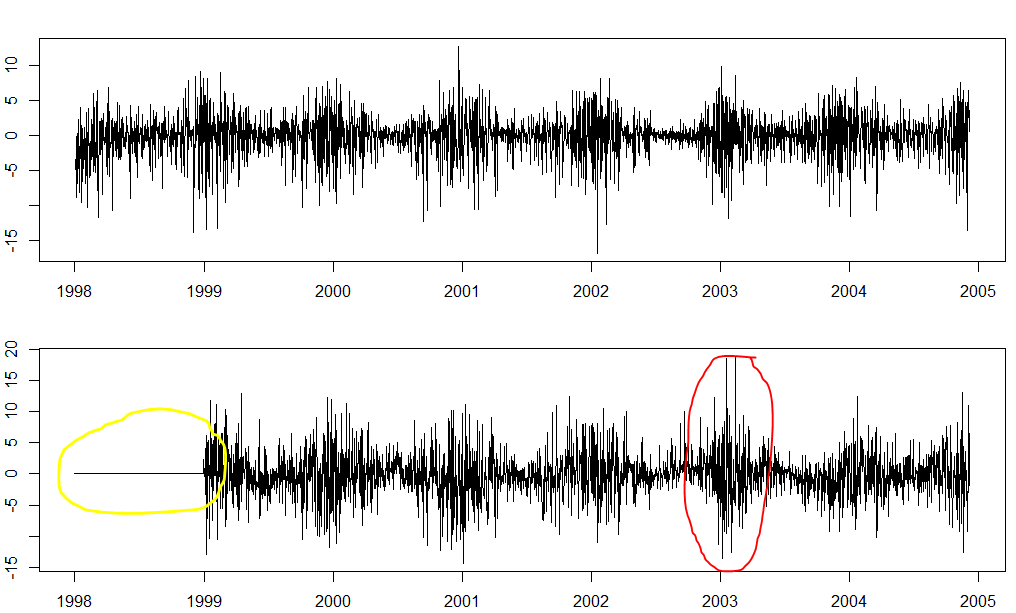
The Goodness of Fit criteria for ARIMA models, **AIC** (Alkaline Information Criterion) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. **BIC** is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

**Observed vs Predicted Values**



From the graphs above, we can observe that the fitted values are defined same in both the ARIMA model (2nd graph) and the Exponential Smoothing model (1st graph).

**Residuals of Fitted Data**



From the above graphs, we can observe that the residuals for ARIMA model has a mean of zero for the year 1998. But there is a sharp increase in the ARIMA model residuals for the year 2003.

Otherwise, Exponential Smoothing has a constant variance and mean.

The residuals in a time series model are the leftovers after fitting a model. In many times series models, the residuals are equal to the difference between the observations and corresponding fitted values.

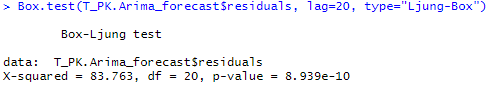


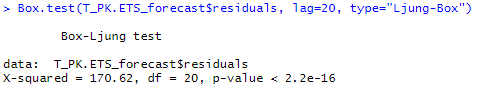
The graphs showing about residuals concludes that the residuals in this variable fulfils both the conditions, ,i.e., the residuals are not correlated, and they have zero mean. Since, both these conditions are satisfied, it can be said that the forecasting methods are proper and good.

**Fit and Measure Statistics**

The Ljung-Box test is a tool to test the lack of fit of a time series model. The test is applied to the residuals of a time series after fitting an ARMA(p,q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) of the Box Ljung Test, H0, is that our model *does not* show lack of fit (or in simple terms—the model is just fine). The [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/), Ha, is just that the model *does*show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series *isn’t*autocorrelated.

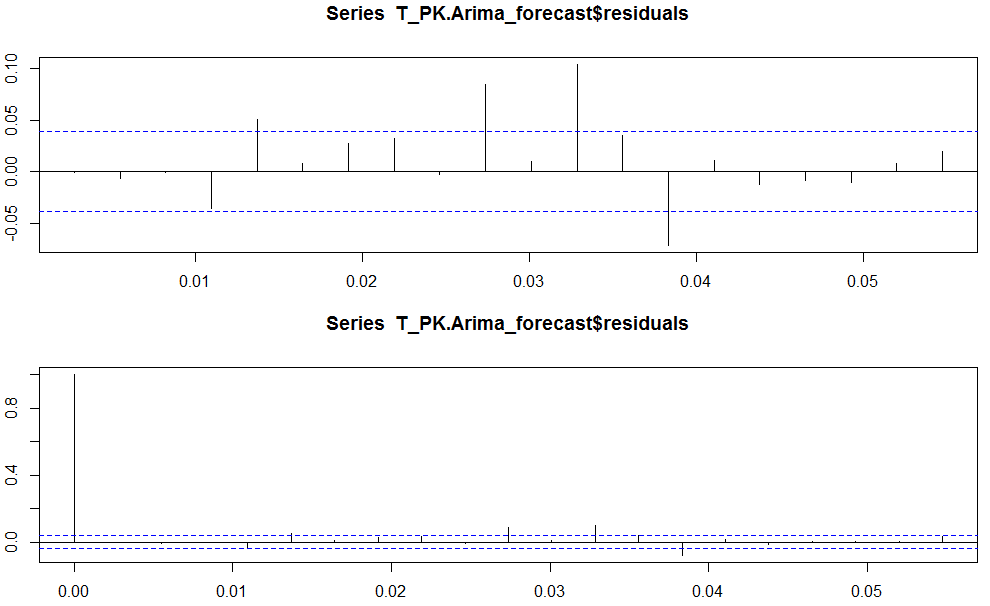




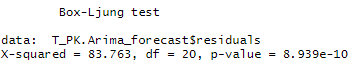
From the above statistics, it is evident that both the models (ARIMA and Exponential Smoothing) shows a lack of good fit.

However, the residual squared error is less for ARIMA model (83.763) comparatively to the Exponential Smoothing model (170.62). Therefore, we will focus on improving the fit of ARIMA model.

**Partial Correlation & Autocorrelogram**

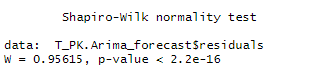


From the above plots, we can observe that the autocorrelation for lag 4 exceeds the significance bounds, and that the autocorrelation and partial autocorrelation tails off to zero after lag 4.



It is evident that the p-value of 8.939e-10 indicates that there is strong evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

**Normality of Residuals**



From the above Shapiro-Wilk test for normality, we can observe that the p-value to be far less than the significance level, rejecting the null hypothesis H0 which states that the residual distribution is normal. Therefore, we can conclude that the residuals are **not normally distributed** here.

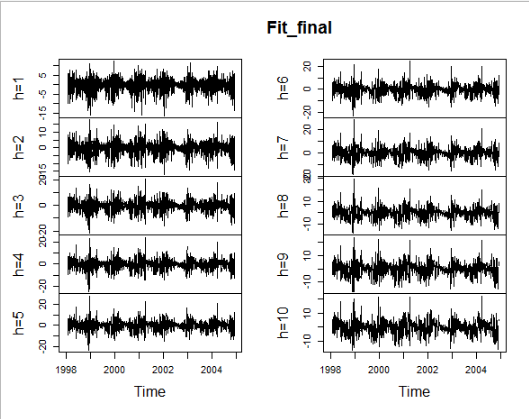
**Best Model**

From the two time-series models we used for forecasting the Ozone Level for the subsequent days, the model using ARIMA is better compared to that of the Exponential Smoothing model since the residual squared error mean is less for ARIMA than Exponential Smoothing model. Also, for forecasting, the correlation between successive values of time series is important, which is measured only in ARIMA model.

The best candidate ARIMA model chosen for this variable is (3,1,3) which shows a differencing factor of 1, partial autocorrelogram tails off to 0 after lag 1 and autocorrelogram tails off to zero after lag 3.

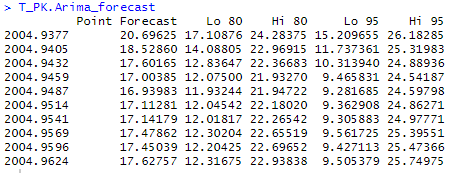
**Cross-Validation:**

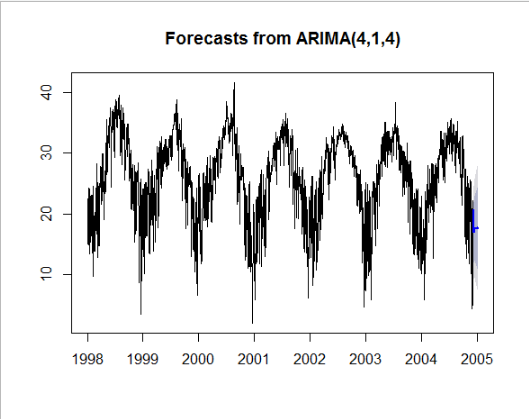
A Cross-Validation for the above ARIMA model with parameter values (p,d,q) = (3,1,3) is computed and plotted below for h= 10 (Forecast Horizon) and window = 30.



**Forecasting using ARIMA Model:**

The forecast of T\_PK (Peak Temperature) for the subsequent 10 days are as below:



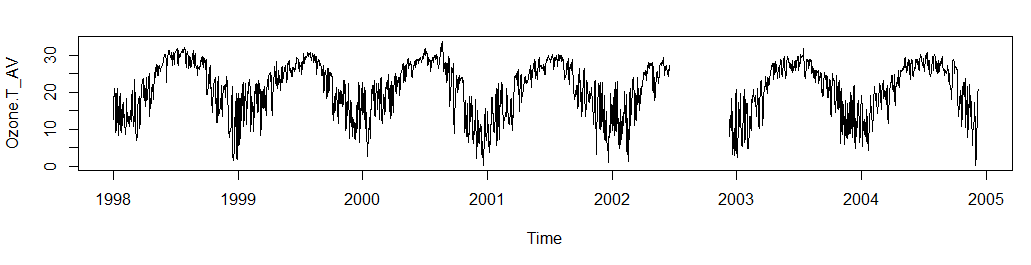


The above plot shows the forecast of the T\_PK variable for the next 30 subsequent days from the end of the day in dataset.

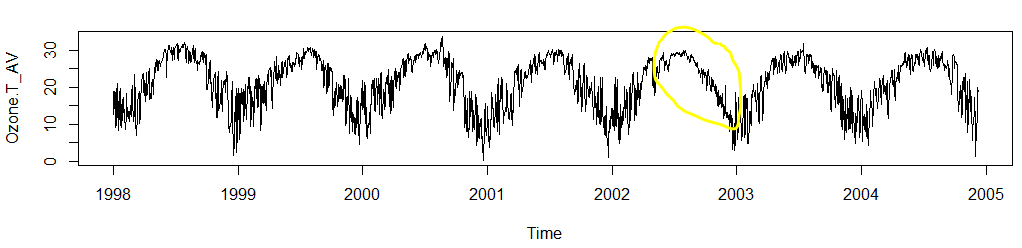
* ***T\_AV***

**tsclean()** function is used to remove the outliers and the missing values are replaced using linear interpolation.

***Before Removing Outliers & Dealing with NA values***



***After Removing Outliers & Dealing NA values***

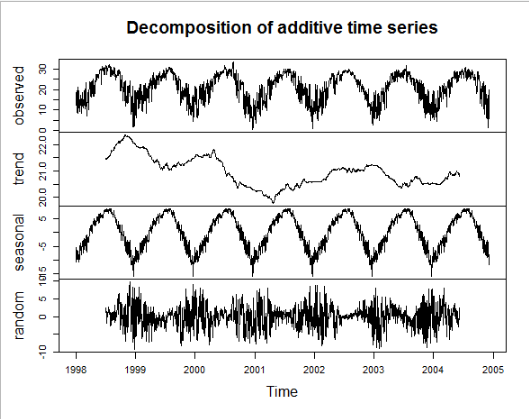


From the above image, it can be observed that the missing values in this key attribute has been replaced by using **Linear Interpolation**, a method of fitting the curve using linear polynomials to construct new data points within the range of a discrete set of known data points.

**Decomposing Time Series Data**

In this, the data is decomposed or separated into its constituent components, which are trend, seasonality and an irregular component.

From the below graphs, we can observe both, a trend and a seasonality for this variable. A mountain like pattern of seasonality occurs every year. Also, we can see a decreasing trend from the year 1999 till mid of the year 2001 and then the trend increases, stabilizes after some time.



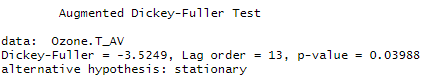
**Test for Stationarity**

**Stationarity** is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

The hypothesis for Augmented Dicky-Fuller Test is,

H0 = Null Hypothesis

HA = Alternate Hypothesis

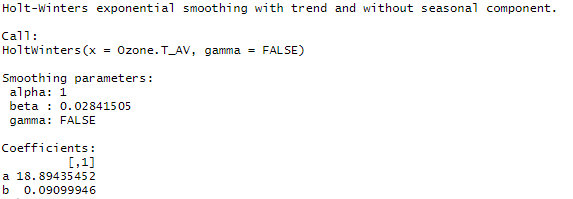


From the above statistics, we can conclude that this time series object is **stationary**.

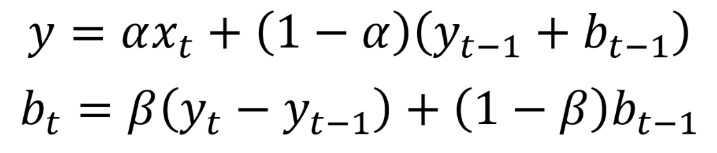
**Exponential Smoothing Model**

Since the time series for WSR\_PK variable can be described using an additive model with a sudden jump in trend and seasonality, **Holt-Winters Exponential Smoothing** is used for this variable.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.



The above statistics show that the Smoothing parameters for the model has values such as alpha = 1, beta = 0.03 and gamma = 0. These parameters convey that the model has a trend but no seasonality. The alpha value is one, which tells that the forecasts are based only on current observation.

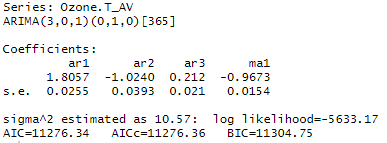


Here, alpha is a **smoothing factor** that takes value between 0 and 1, It determines how fast the weight decreases for previous observations.

The beta is the **trend smoothing factor**, and it takes values between 0 and 1.

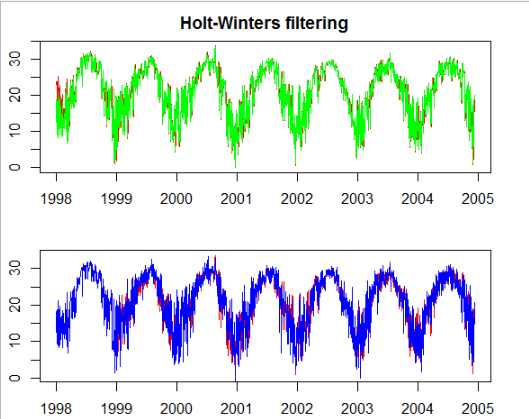
**ARIMA Model**

Since ARIMA (Autoregressive Integrated Moving Average) models are defined for stationary time series, a differencing in time series will be implemented to obtain a stationary time series. This model includes an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.



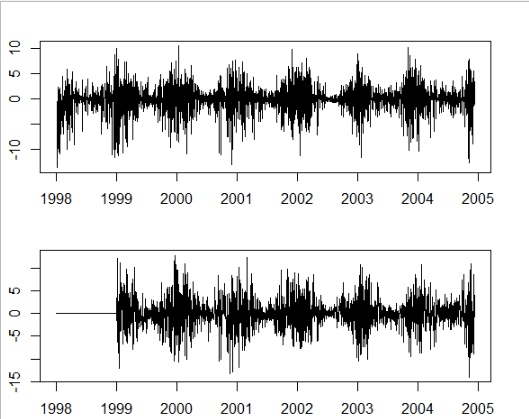
The Goodness of Fit criteria for ARIMA models, **AIC** (Alkaline Information Criterion) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. **BIC** is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

**Observed vs Predicted Values**



From the graphs above, we can observe that the fitted values are defined same in both the ARIMA model (2nd graph) and the Exponential Smoothing model (1st graph).

**Residuals of Fitted Data**



From the above graphs, we can observe that the residuals for ARIMA model has a mean of zero for the year 1998. But there is a sharp increase in the ARIMA model residuals for the year 2003.

Otherwise, Exponential Smoothing has a constant variance and mean.

The residuals in a time series model are the leftovers after fitting a model. In many times series models, the residuals are equal to the difference between the observations and corresponding fitted values.

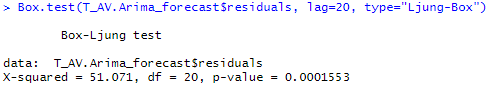


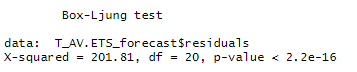
The graphs showing about residuals concludes that the residuals in this variable fulfils both the conditions, ,i.e., the residuals are not correlated, and they have zero mean. Since, both these conditions are satisfied, it can be said that the forecasting methods are proper and good.

**Fit and Measure Statistics**

The Ljung-Box test is a tool to test the lack of fit of a time series model. The test is applied to the residuals of a time series after fitting an ARMA(p,q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) of the Box Ljung Test, H0, is that our model *does not* show lack of fit (or in simple terms—the model is just fine). The [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/), Ha, is just that the model *does*show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series *isn’t*autocorrelated.

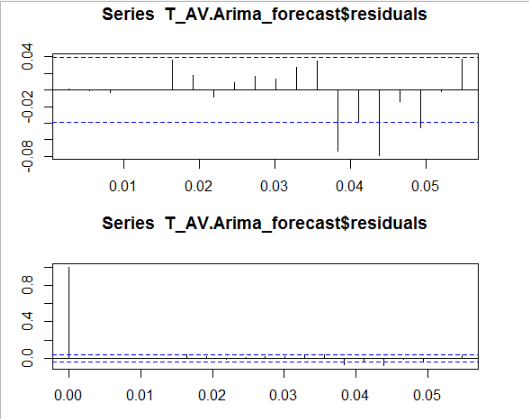




From the above statistics, it is evident that both the models (ARIMA and Exponential Smoothing) shows a lack of good fit.

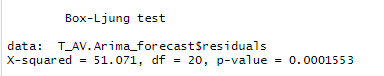
However, the residual squared error is less for ARIMA model (51.071) comparatively to the Exponential Smoothing model (201.81). Therefore, we will focus on improving the fit of ARIMA model.

**Partial Correlation & Autocorrelogram**



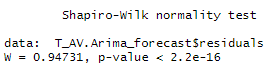
From both the correlogram plots above, we can observe that the sample autocorrelation and partial correlation for the in-sample forecast errors at lag 14 exceeds the significance bounds. However, we could expect one in 20 of the autocorrelations and partial autocorrelations for the first twenty lags to exceed the 95% significance bounds.

Let us now carry out the Ljung-Box test.



It is evident that the p-value of 0.0001553 indicates that there is strong evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

**Normality of Residuals**

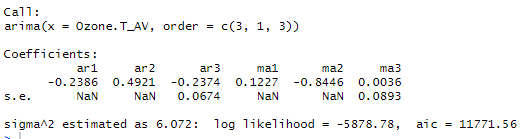


From the above Shapiro-Wilk test for normality, we can observe that the p-value to be far less than the significance level, rejecting the null hypothesis H0 which states that the residual distribution is normal. Therefore, we can conclude that the residuals are **not normally distributed** here.

**Best Model**

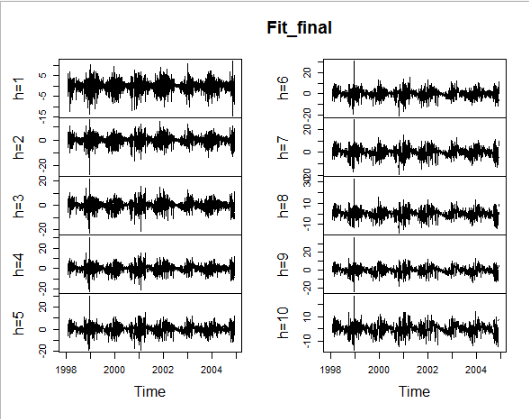
From the two time-series models we used for forecasting the Ozone Level for the subsequent days, the model using ARIMA is better compared to that of the Exponential Smoothing model since the residual squared error mean is less for ARIMA than Exponential Smoothing model. Also, for forecasting, the correlation between successive values of time series is important, which is measured only in ARIMA model.

The best candidate ARIMA model chosen for this variable is (3,1,3) which shows a differencing factor of 1, partial autocorrelogram tails off to 0 after lag 1 and autocorrelogram tails off to zero after lag 3.



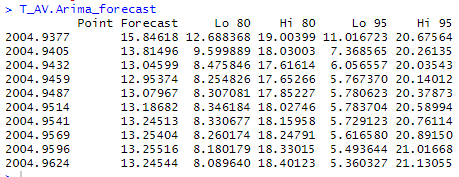
**Cross-Validation:**

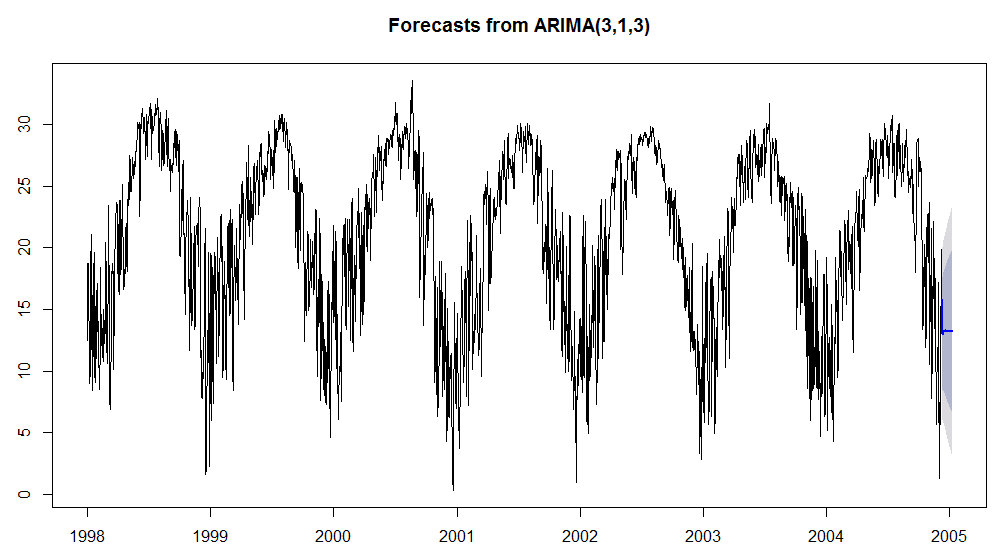
A Cross-Validation for the above ARIMA model with parameter values (p,d,q) = (3,1,3) is computed and plotted below for h= 10 (Forecast Horizon) and window = 30.



**Forecasting using ARIMA Model:**

The forecast of T\_AV (Average Temperature) for the subsequent 10 days are as below:



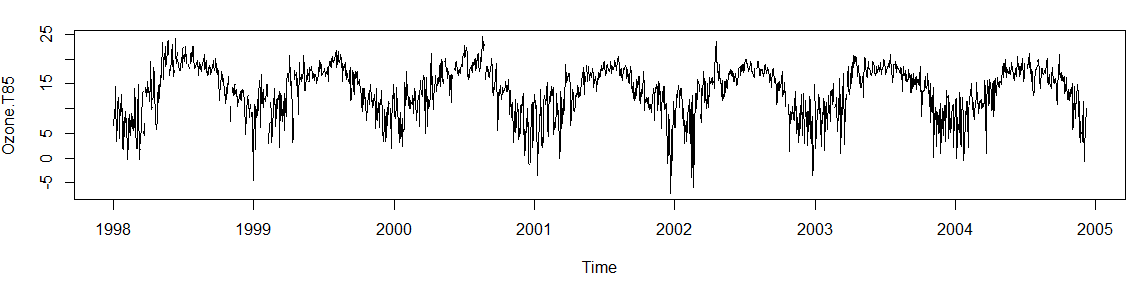


The above plot shows the forecast of the T\_AV variable for the next 30 subsequent days from the end of the day in dataset.

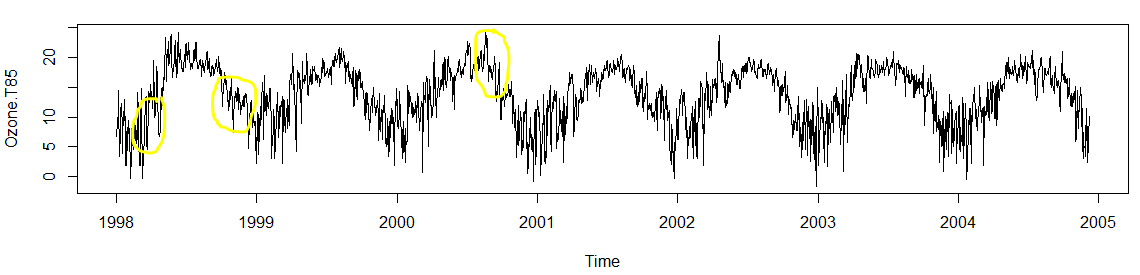
* ***T85***

**tsclean()** function is used to remove the outliers and the missing values are replaced using linear interpolation.

***Before Removing Outliers & Dealing with NA values***



***After Removing Outliers & Dealing NA values***

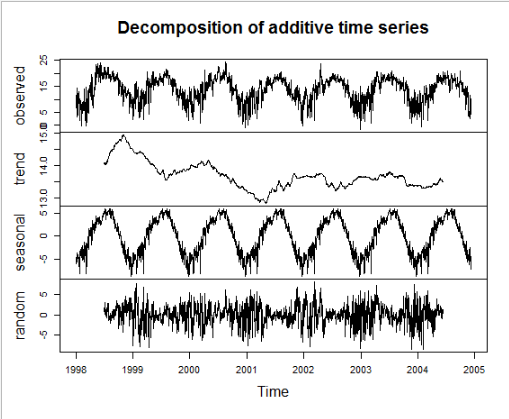


From the above image, it can be observed that the missing values in this key attribute has been replaced by using **Linear Interpolation**, a method of fitting the curve using linear polynomials to construct new data points within the range of a discrete set of known data points.

**Decomposing Time Series Data**

In this, the data is decomposed or separated into its constituent components, which are trend, seasonality and an irregular component.

From the below graphs, we can observe both, a trend and a seasonality for this variable. A mountain like pattern of seasonality occurs every year. Also, we can see a decreasing trend from the end of year 1998 till mid of the year 1999 and then the trend stabilizes, again decreases after the year 2000.



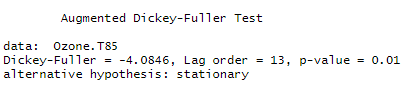
**Test for Stationarity**

**Stationarity** is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

The hypothesis for Augmented Dicky-Fuller Test is,

H0 = Null Hypothesis -> p-value equals to 0

HA = Alternate Hypothesis -> p-values greater than 0

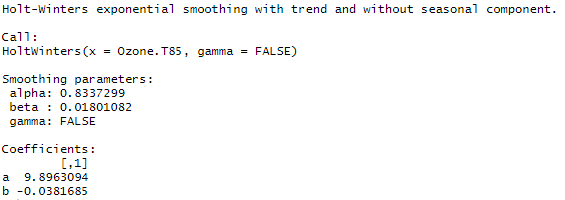


From the above statistics, we can conclude that this time series object is not **stationary**.

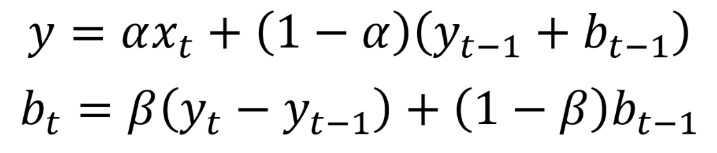
**Exponential Smoothing Model**

Since the time series for WSR\_PK variable can be described using an additive model with a sudden jump in trend and seasonality, **Holt-Winters Exponential Smoothing** is used for this variable.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.



The above statistics show that the Smoothing parameters for the model has values such as alpha = 0.834, beta = 0.018 and gamma = 0. These parameters convey that the model has a trend but no seasonality. The alpha value is close to one, which tells that the forecasts are based on current observations.

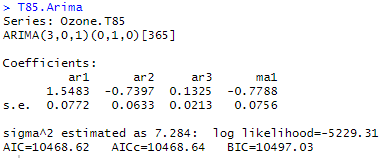


Here, alpha is a **smoothing factor** that takes value between 0 and 1, It determines how fast the weight decreases for previous observations.

The beta is the **trend smoothing factor**, and it takes values between 0 and 1.

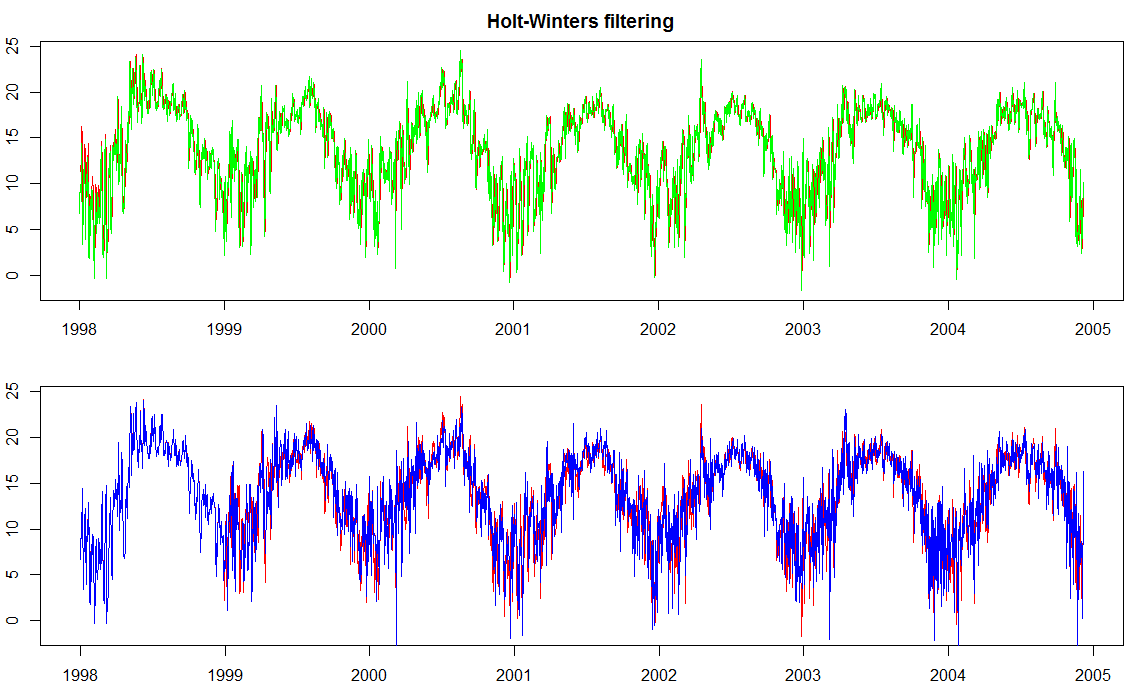
**ARIMA Model**

Since ARIMA (Autoregressive Integrated Moving Average) models are defined for stationary time series, a differencing in time series will be implemented to obtain a stationary time series. This model includes an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.



The Goodness of Fit criteria for ARIMA models, **AIC** (Alkaline Information Criterion) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. **BIC** is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

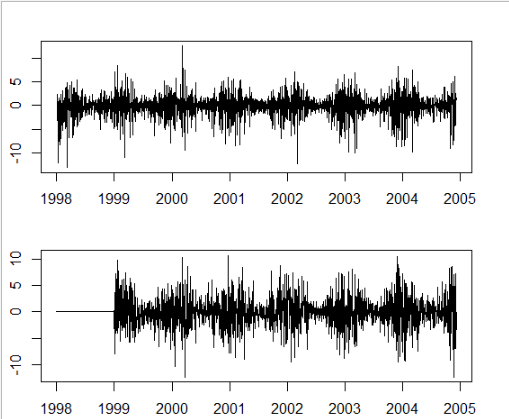
**Observed vs Predicted Values**



From the graphs above, we can observe that the fitted values are defined same in both the ARIMA model (2nd graph) and the Exponential Smoothing model (1st graph).

But, in ARIMA Model, we can see no observed values for the period between 1998 and 1999.

**Residuals of Fitted Data**



From the above graphs, we can observe that the residuals for ARIMA model has a mean of zero for the year 1998. But there is a sharp increase in the ARIMA model residuals for the year 2004.

Otherwise, Exponential Smoothing has a constant variance and mean.

The residuals in a time series model are the leftovers after fitting a model. In many times series models, the residuals are equal to the difference between the observations and corresponding fitted values.

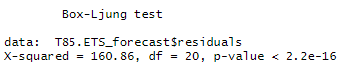


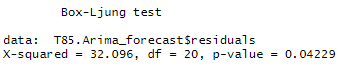
The graphs showing about residuals concludes that the residuals in this variable fulfils both the conditions, ,i.e., the residuals are not correlated, and they have zero mean. Since, both these conditions are satisfied, it can be said that the forecasting methods are proper and good.

**Fit and Measure Statistics**

The Ljung-Box test is a tool to test the lack of fit of a time series model. The test is applied to the residuals of a time series after fitting an ARMA(p,q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) of the Box Ljung Test, H0, is that our model *does not* show lack of fit (or in simple terms—the model is just fine). The [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/), Ha, is just that the model *does*show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series *isn’t*autocorrelated.

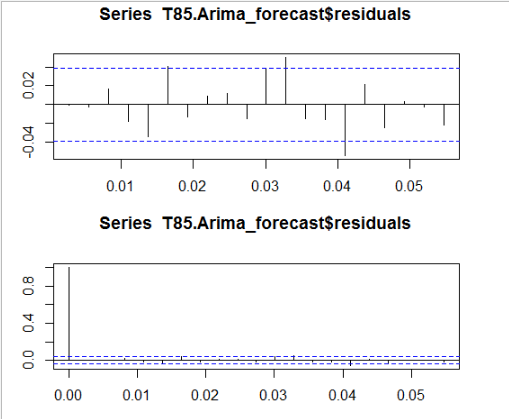




From the above statistics, it is evident that both the models (ARIMA and Exponential Smoothing) shows a lack of good fit.

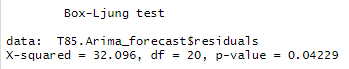
However, the residual squared error is less for ARIMA model (32.096) comparatively to the Exponential Smoothing model (160.86). Therefore, we will focus on improving the fit of ARIMA model.

**Partial Correlation & Autocorrelogram**



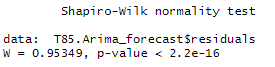
From both the correlogram plots above, we can observe that the sample autocorrelation and partial correlation for the in-sample forecast errors at lag 6 exceeds the significance bounds. However, we could expect one in 20 of the autocorrelations and partial autocorrelations for the first twenty lags to exceed the 95% significance bounds.

Let us now carry out the Ljung-Box test.



It is evident that the p-value of 0.04229 indicates that there is an evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

**Normality of Residuals**

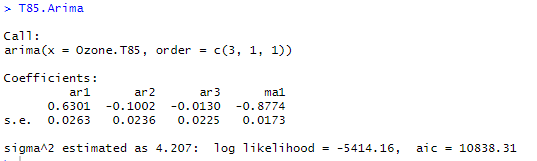


From the above Shapiro-Wilk test for normality, we can observe that the p-value to be far less than the significance level, rejecting the null hypothesis H0 which states that the residual distribution is normal. Therefore, we can conclude that the residuals are **not normally distributed** here.

**Best Model**

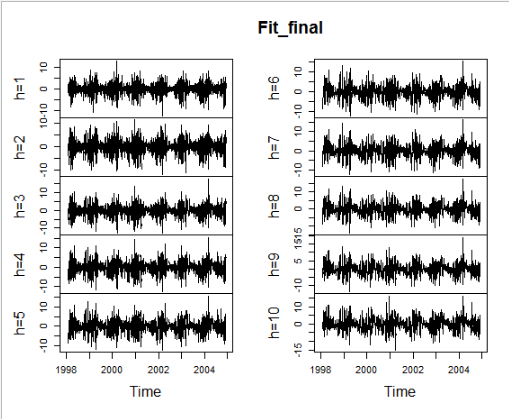
From the two time-series models we used for forecasting the Ozone Level for the subsequent days, the model using ARIMA is better compared to that of the Exponential Smoothing model since the residual squared error mean is less for ARIMA than Exponential Smoothing model. Also, for forecasting, the correlation between successive values of time series is important, which is measured only in ARIMA model.

The best candidate ARIMA model chosen for this variable is (3,1,1) which shows a differencing factor of 1, partial autocorrelogram tails off to 0 after lag 1 and autocorrelogram tails off to zero after lag 3.



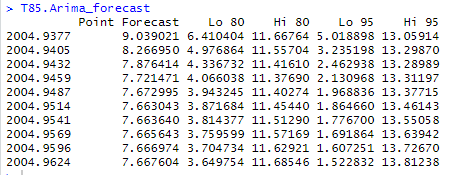
**Cross-Validation:**

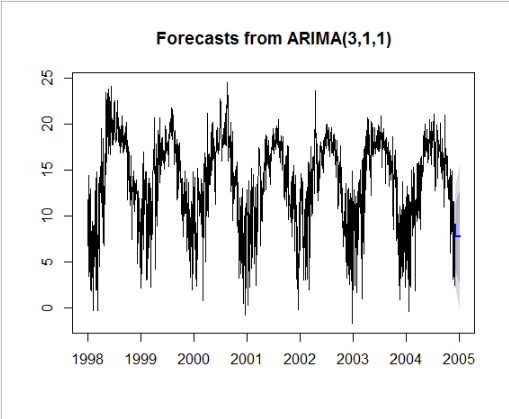
A Cross-Validation for the above ARIMA model with parameter values (p,d,q) = (3,1,1) is computed and plotted below for h= 10 (Forecast Horizon) and window = 30.



**Forecasting using ARIMA Model:**

The forecast of T85 (Temperature at 850 hpa level) for the subsequent 10 days are as below:



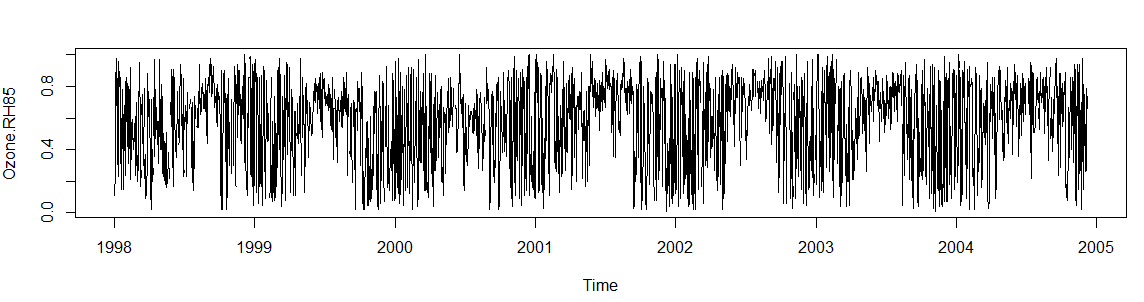


The above plot shows the forecast of the T85 variable for the next 30 subsequent days from the end of the day in dataset.

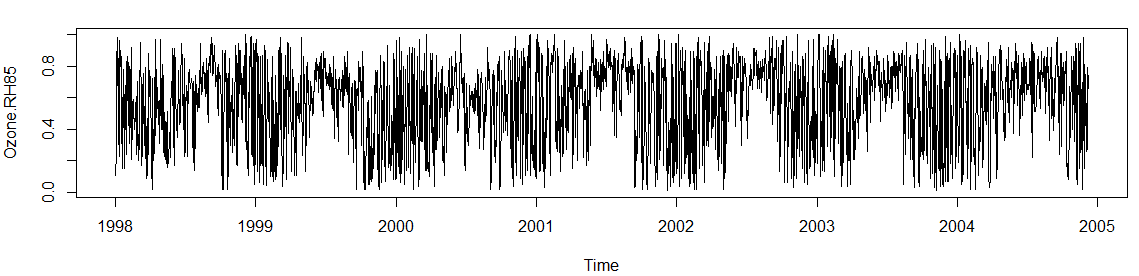
**RH85**

**tsclean()** function is used to remove the outliers and the missing values are replaced using linear interpolation.

***Before Removing Outliers & Dealing with NA values***



***After Removing Outliers & Dealing NA values***

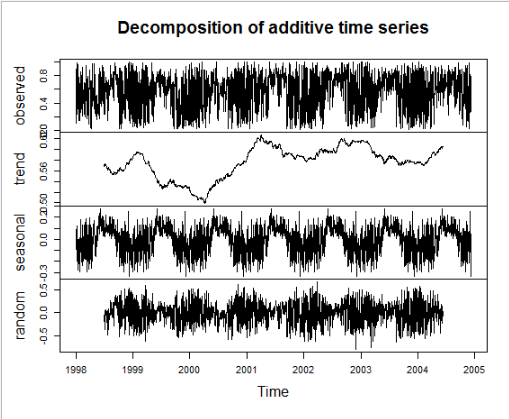


From the above image, it can be observed that the missing values in this key attribute has been replaced by using **Linear Interpolation**, a method of fitting the curve using linear polynomials to construct new data points within the range of a discrete set of known data points.

**Decomposing Time Series Data**

In this, the data is decomposed or separated into its constituent components, which are trend, seasonality and an irregular component.

From the below graphs, we can observe both, a trend and a seasonality for this variable. A unrecognizable pattern of seasonality occurs every year. Also, we can see a decreasing trend from the end of year 1999 till mid of the year 2000 and then the trend increases sharply, then stabilizes after middle of the year 2001.



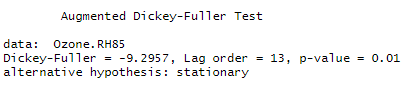
**Test for Stationarity**

**Stationarity** is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

The hypothesis for Augmented Dicky-Fuller Test is,

H0 = Null Hypothesis -> p-value equals to 0

HA = Alternate Hypothesis -> p-values greater than 0

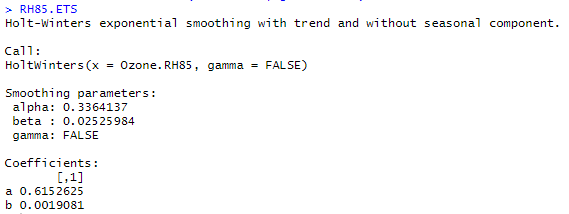


From the above statistics, we can conclude that this time series object is not **stationary**.

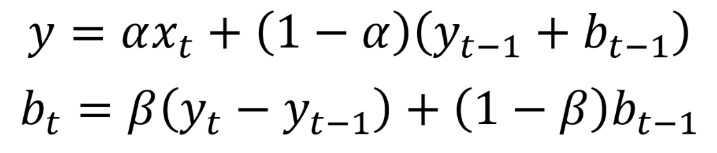
**Exponential Smoothing Model**

Since the time series for WSR\_PK variable can be described using an additive model with a sudden jump in trend and seasonality, **Holt-Winters Exponential Smoothing** is used for this variable.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.



The above statistics show that the Smoothing parameters for the model has values such as alpha = 0.615, beta = 0.02525 and gamma = 0. These parameters convey that the model has a trend but no seasonality. The alpha value is in middle, which tells that the forecasts are based on both previous and current observations.



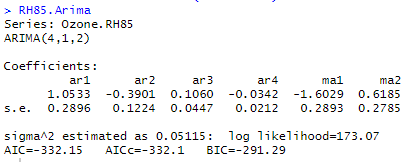
Here, alpha is a **smoothing factor** that takes value between 0 and 1, It determines how fast the weight decreases for previous observations.

The beta is the **trend smoothing factor**, and it takes values between 0 and 1.

**ARIMA Model**

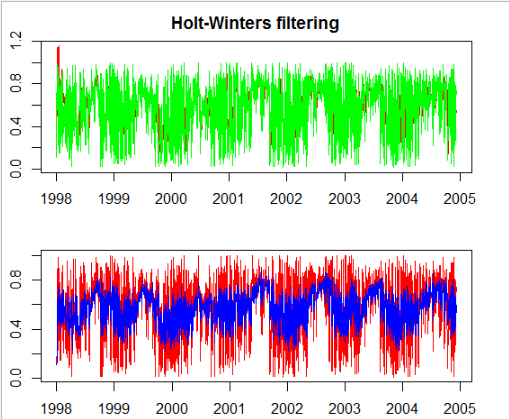
Since ARIMA (Autoregressive Integrated Moving Average) models are defined for stationary time series, a differencing in time series will be implemented to obtain a stationary time series.

This model includes an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.



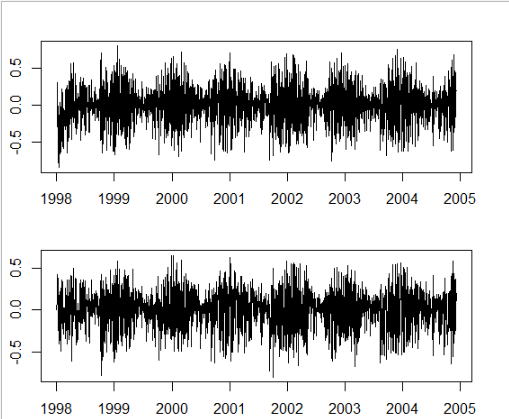
The Goodness of Fit criteria for ARIMA models, **AIC** (Alkaline Information Criterion) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. **BIC** is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

**Observed vs Predicted Values**



From the graphs above, we can observe that the fitted values are more well defined in the ARIMA model (2nd graph) compared to the Exponential Smoothing model (1st graph). It is because, a differencing factor, partial correlogram and autocorrelogram are defined for the below ARIMA model.

**Residuals of Fitted Data**



The residuals in a time series model are the leftovers after fitting a model. In many times series models, the residuals are equal to the difference between the observations and corresponding fitted values.



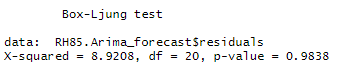
The graphs showing about residuals concludes that the residuals in this variable fulfils both the conditions, ,i.e., the residuals are not correlated, and they have zero mean. Since, both these conditions are satisfied, it can be said that the forecasting methods are proper and good.

**Fit and Measure Statistics**

The Ljung-Box test is a tool to test the lack of fit of a time series model. The test is applied to the residuals of a time series after fitting an ARMA(p,q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

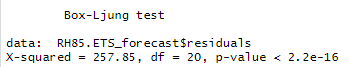
The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) of the Box Ljung Test, H0, is that our model *does not* show lack of fit (or in simple terms—the model is just fine). The [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/), Ha, is just that the model *does*show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series *isn’t*autocorrelated.

Let us now perform this test for the ARIMA model:



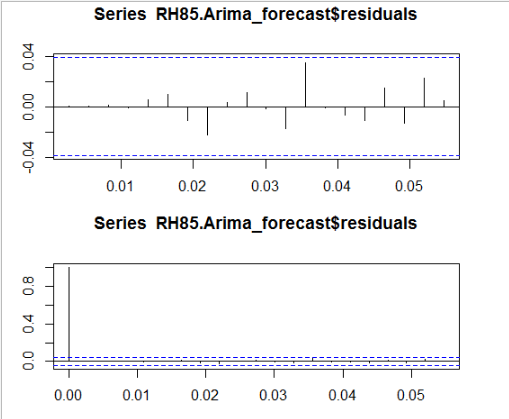
From the above test, we can find the Box-Ljung test statistic to be 8.9208, lags between 1-20 and p-value as 0.9838. This shows that the model does not show any lack of fit and its good.

For Exponential Smoothing model:



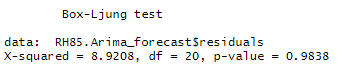
From the above test, we can observe the p-value to be less than the significant value, which concluded that this model shows a lack of fit.

**Partial Correlation & Autocorrelogram**



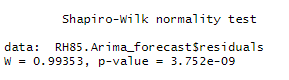
From both the correlogram plots above, we can observe that the sample autocorrelation and partial correlation for the in-sample forecast errors does not exceed the significance bounds. However, we could expect one in 20 of the autocorrelations and partial autocorrelations for the first twenty lags to exceed the 95% significance bounds.

Let us now carry out the Ljung-Box test.



It is evident that the p-value of 0.9838 indicates that there is very little evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

**Normality of Residuals**



From the above Shapiro-Wilk test for normality, we can observe that the p-value to be far less than the significance level, rejecting the null hypothesis H0 which states that the residual distribution is normal. Therefore, we can conclude that the residuals are **not normally distributed** here.

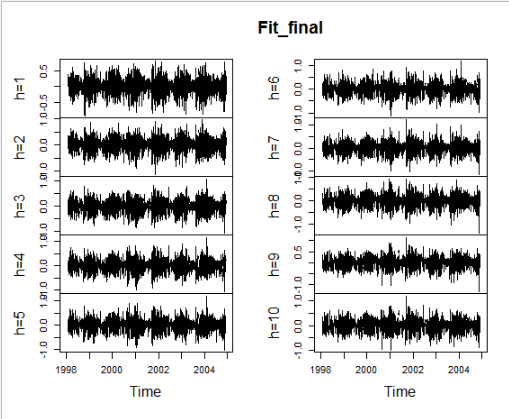
**Best Model**

From the two time-series models we used for forecasting the Ozone Level for the subsequent days, the model using ARIMA is better compared to that of the Exponential Smoothing model since the fit of ARIMA is better than Exponential Smoothing model. Also, for forecasting, the correlation between successive values of time series is important, which is measured only in ARIMA model.

The best candidate ARIMA model chosen for this variable is (4,1,2) which shows a differencing factor of 4, partial autocorrelogram tails off to 0 after lag 1 and autocorrelogram tails off to zero after lag 2.

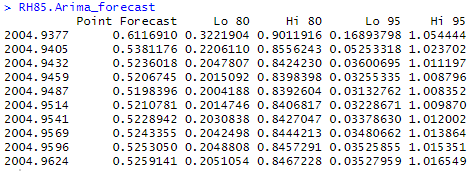
**Cross-Validation:**

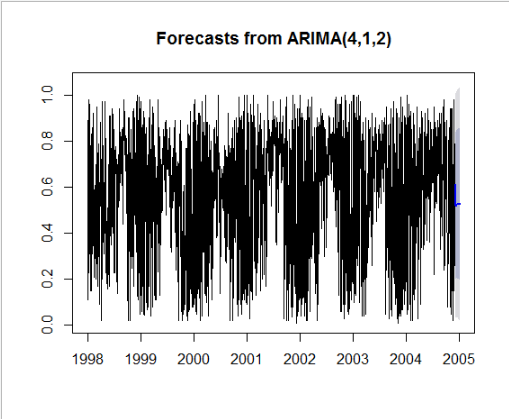
A Cross-Validation for the above ARIMA model with parameter values (p,d,q) = (4,1,2) is computed and plotted below for h= 10 (Forecast Horizon) and window = 30.



**Forecasting using ARIMA Model:**

The forecast of RH85 (Relative Humidity at 850 hpa) for the subsequent 10 days are as below:



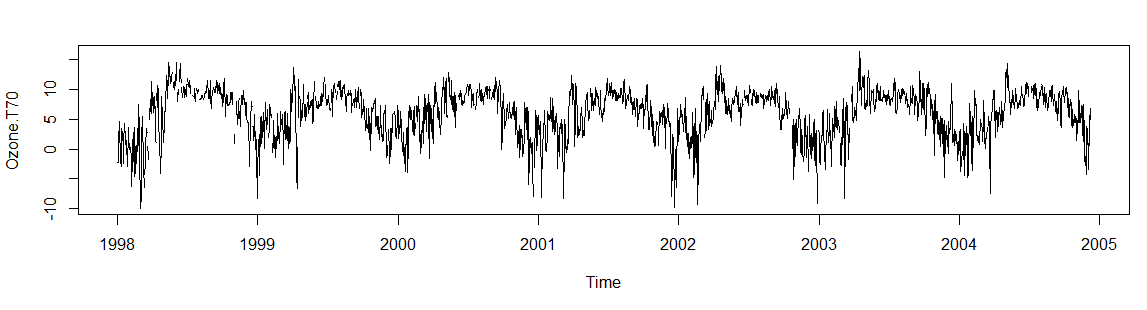


The above plot shows the forecast of the RH85 variable for the next 30 subsequent days from the end of the day in dataset.

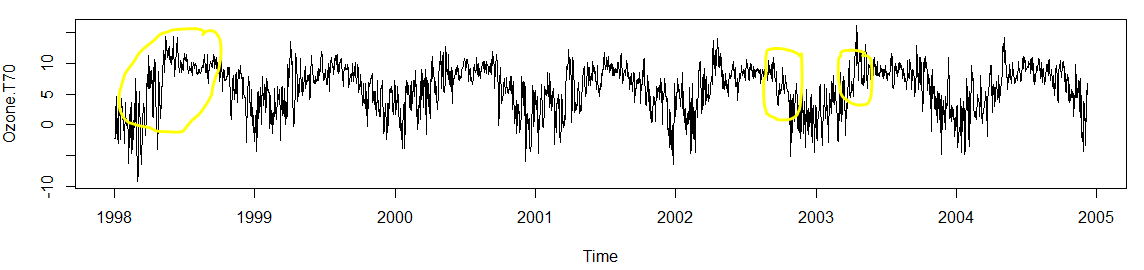
**T70**

**tsclean()** function is used to remove the outliers and the missing values are replaced using linear interpolation.

***Before Removing Outliers & Dealing with NA values***



***After Removing Outliers & Dealing NA values***

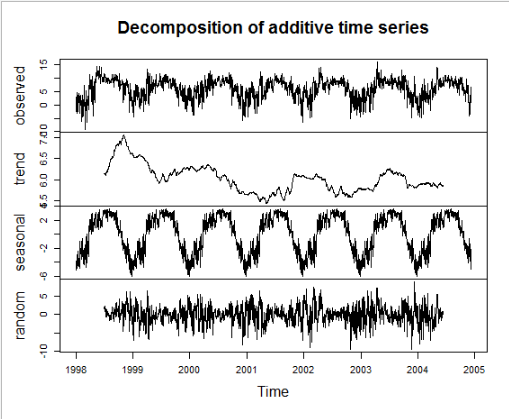


From the above image, it can be observed that the missing values in this key attribute has been replaced by using **Linear Interpolation**, a method of fitting the curve using linear polynomials to construct new data points within the range of a discrete set of known data points.

**Decomposing Time Series Data**

In this, the data is decomposed or separated into its constituent components, which are trend, seasonality and an irregular component.

From the below graphs, we can observe both, a trend and a seasonality for this variable. A mountain like pattern of seasonality occurs every year. Also, we can see a decreasing trend from the end of year 1998 till mid of the year 1999 and then the trend gradually decreases.



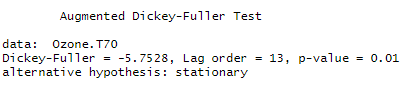
**Test for Stationarity**

**Stationarity** is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

The hypothesis for Augmented Dicky-Fuller Test is,

H0 = Null Hypothesis -> p-value equal or less than 0.05

HA = Alternate Hypothesis -> p-values greater than 0.05

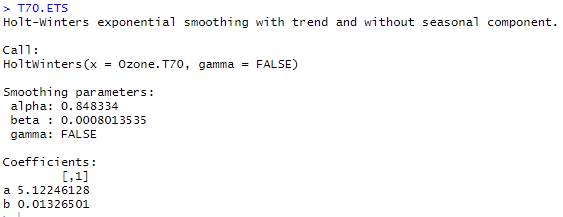


From the above statistics, we can conclude that this time series object is **stationary**.

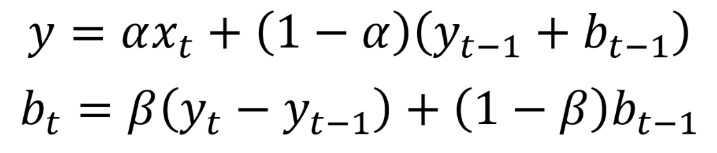
**Exponential Smoothing Model**

Since the time series for WSR\_PK variable can be described using an additive model with a sudden jump in trend and seasonality, **Holt-Winters Exponential Smoothing** is used for this variable.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.



The above statistics show that the Smoothing parameters for the model has values such as alpha = 0.8483, beta = 0.0008 and gamma = 0. These parameters convey that the model has a trend but no seasonality. The alpha value is close to one, which tells that the forecasts are based on current observations.

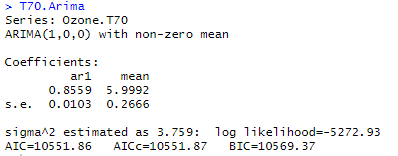


Here, alpha is a **smoothing factor** that takes value between 0 and 1, It determines how fast the weight decreases for previous observations.

The beta is the **trend smoothing factor**, and it takes values between 0 and 1.

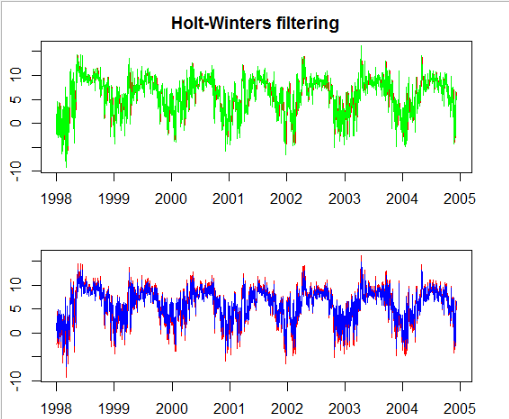
**ARIMA Model**

Since ARIMA (Autoregressive Integrated Moving Average) models are defined for stationary time series, a differencing in time series will be implemented to obtain a stationary time series. This model includes an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.



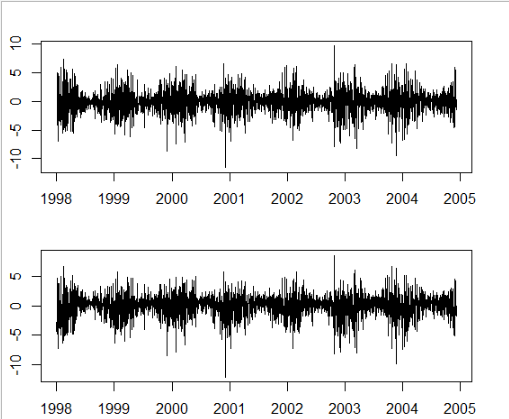
The Goodness of Fit criteria for ARIMA models, **AIC** (Alkaline Information Criterion) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. **BIC** is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

**Observed vs Predicted Values**



From the graphs above, we can observe that the fitted values are defined same in both the ARIMA model (2nd graph) and the Exponential Smoothing model (1st graph).

**Residuals of Fitted Data**



The residuals in a time series model are the leftovers after fitting a model. In many times series models, the residuals are equal to the difference between the observations and corresponding fitted values.



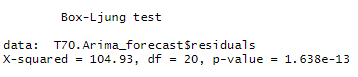
The graphs showing about residuals concludes that the residuals in this variable fulfils both the conditions, ,i.e., the residuals are not correlated, and they have zero mean. Since, both these conditions are satisfied, it can be said that the forecasting methods are proper and good.

**Fit and Measure Statistics**

The Ljung-Box test is a tool to test the lack of fit of a time series model. The test is applied to the residuals of a time series after fitting an ARMA(p,q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

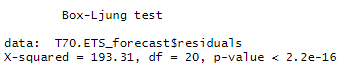
The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) of the Box Ljung Test, H0, is that our model *does not* show lack of fit (or in simple terms—the model is just fine). The [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/), Ha, is just that the model *does*show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series *isn’t*autocorrelated.

Let us now perform this test for the ARIMA model:



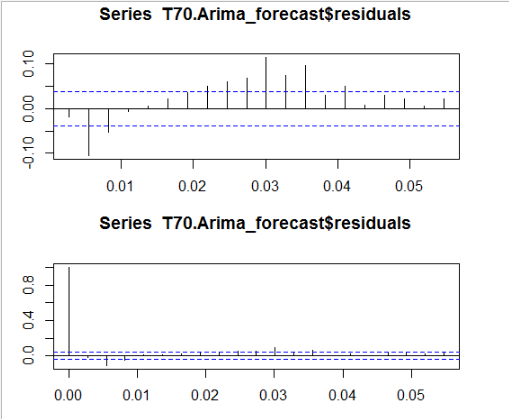
From the above test, we can find the Box-Ljung test statistic to be 104.93, lags between 1-20 and p-value as 1.638e-13. This shows that the model has lack of fit.

For Exponential Smoothing model:



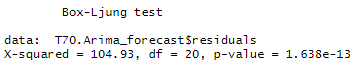
From the above test, we can observe the p-value to be less than the significant value, which concluded that this model shows a lack of fit.

**Partial Correlation & Autocorrelogram**



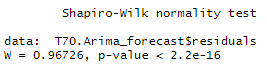
From both the correlogram plots above, we can observe that the sample autocorrelation and partial correlation for the in-sample forecast exceeds at lag 8 the significance bounds. However, we could expect one in 20 of the autocorrelations and partial autocorrelations for the first twenty lags to exceed the 95% significance bounds.

Let us now carry out the Ljung-Box test.



It is evident that the p-value of 1.638e-13 indicates that there is evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

**Normality of Residuals**



From the above Shapiro-Wilk test for normality, we can observe that the p-value to be far less than the significance level, rejecting the null hypothesis H0 which states that the residual distribution is normal. Therefore, we can conclude that the residuals are **not normally distributed** here.

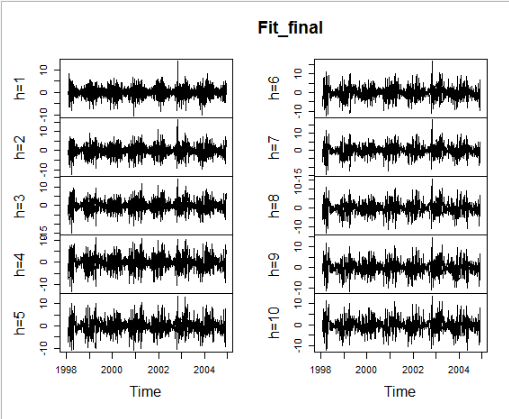
**Best Model**

From the two time-series models we used for forecasting the Ozone Level for the subsequent days, the model using ARIMA is better compared to that of the Exponential Smoothing model since the fit of ARIMA is better than Exponential Smoothing model. Also, for forecasting, the correlation between successive values of time series is important, which is measured only in ARIMA model.

The best candidate ARIMA model chosen for this variable is (1,0,0) which shows a differencing factor of 0, partial autocorrelogram tails off to 0 after lag 1 and autocorrelogram tails off to zero after lag 0.

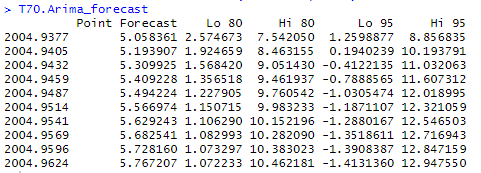
**Cross-Validation:**

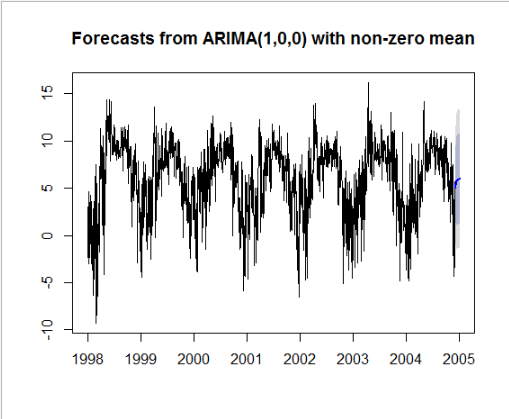
A Cross-Validation for the above ARIMA model with parameter values (p,d,q) = (1,0,0) is computed and plotted below for h= 10 (Forecast Horizon) and window = 30.



**Forecasting using ARIMA Model:**

The forecast of T70 (Temperature at 700 hpa) for the subsequent 10 days are as below:



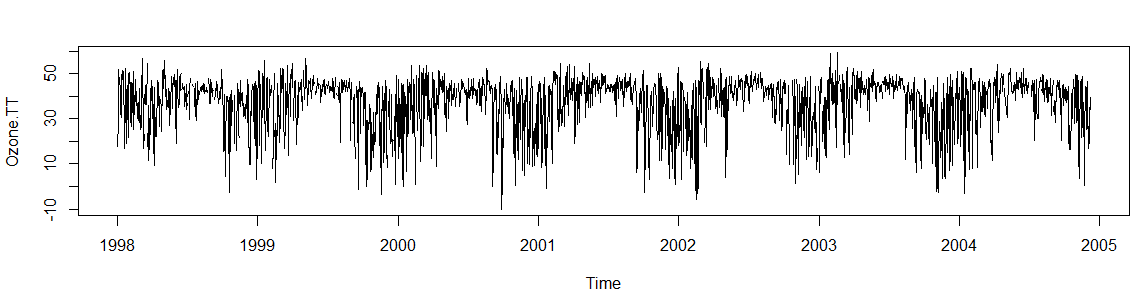


The above plot shows the forecast of the T70 variable for the next 30 subsequent days from the end of the day in dataset.

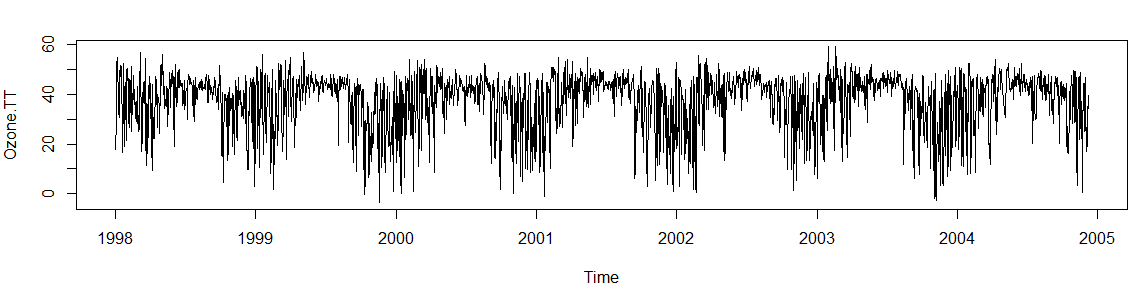
**TT**

**tsclean()** function is used to remove the outliers and the missing values are replaced using linear interpolation.

***Before Removing Outliers & Dealing with NA values***



***After Removing Outliers & Dealing NA values***

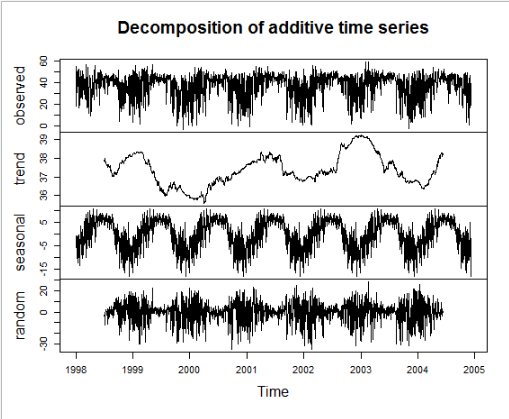


From the above image, it can be observed that the missing values in this key attribute has been replaced by using **Linear Interpolation**, a method of fitting the curve using linear polynomials to construct new data points within the range of a discrete set of known data points.

**Decomposing Time Series Data**

In this, the data is decomposed or separated into its constituent components, which are trend, seasonality and an irregular component.

From the below graphs, we can observe both, a trend and a seasonality for this variable. A mountain like pattern of seasonality occurs every year. Also, we can observe an uneven increasing and decreasing trend in the variable.



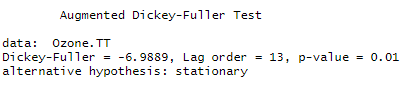
**Test for Stationarity**

**Stationarity** is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

The hypothesis for Augmented Dicky-Fuller Test is,

H0 = Null Hypothesis -> p-value equal or less than 0.05

HA = Alternate Hypothesis -> p-values greater than 0.05

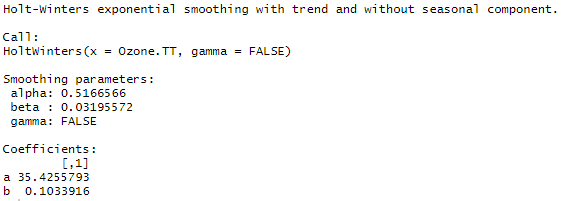


From the above statistics, we can conclude that this time series object is **stationary**.

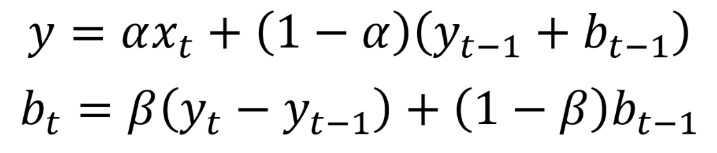
**Exponential Smoothing Model**

Since the time series for WSR\_PK variable can be described using an additive model with a sudden jump in trend and seasonality, **Holt-Winters Exponential Smoothing** is used for this variable.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.



The above statistics show that the Smoothing parameters for the model has values such as alpha = 0.5166566, beta = 0.032 and gamma = 0. These parameters convey that the model has a trend but no seasonality. The alpha value is in middle, which tells that the forecasts are based on both current and previous observations.

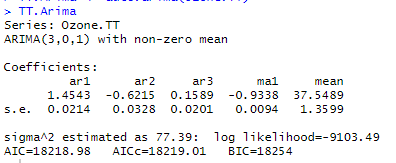


Here, alpha is a **smoothing factor** that takes value between 0 and 1, It determines how fast the weight decreases for previous observations.

The beta is the **trend smoothing factor**, and it takes values between 0 and 1.

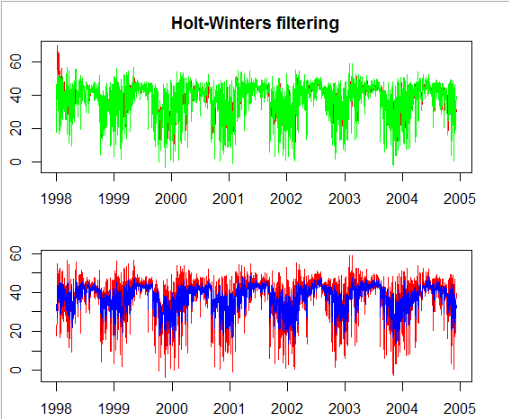
**ARIMA Model**

Since ARIMA (Autoregressive Integrated Moving Average) models are defined for stationary time series, a differencing in time series will be implemented to obtain a stationary time series. This model includes an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.



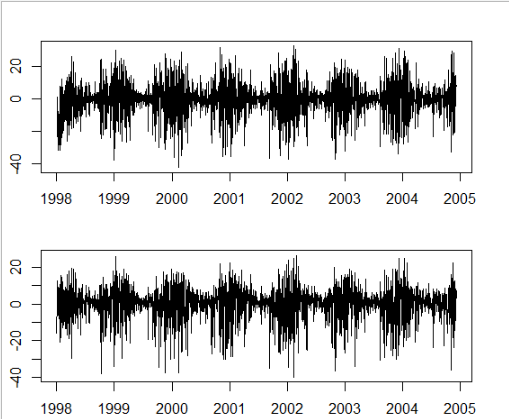
The Goodness of Fit criteria for ARIMA models, **AIC** (Alkaline Information Criterion) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. **BIC** is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

**Observed vs Predicted Values**



From the graphs above, we can observe that the fitted values are more well defined in the ARIMA model (2nd graph) compared to the Exponential Smoothing model (1st graph). It is because, a differencing factor, partial correlogram and autocorrelogram are defined for the below ARIMA model.

**Residuals of Fitted Data**



The residuals in a time series model are the leftovers after fitting a model. In many times series models, the residuals are equal to the difference between the observations and corresponding fitted values.



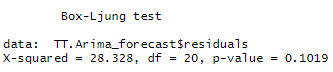
The graphs showing about residuals concludes that the residuals in this variable fulfils both the conditions, ,i.e., the residuals are not correlated, and they have zero mean. Since, both these conditions are satisfied, it can be said that the forecasting methods are proper and good.

**Fit and Measure Statistics**

The Ljung-Box test is a tool to test the lack of fit of a time series model. The test is applied to the residuals of a time series after fitting an ARMA(p,q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

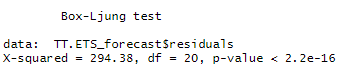
The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) of the Box Ljung Test, H0, is that our model *does not* show lack of fit (or in simple terms—the model is just fine). The [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/), Ha, is just that the model *does*show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series *isn’t*autocorrelated.

Let us now perform this test for the ARIMA model:



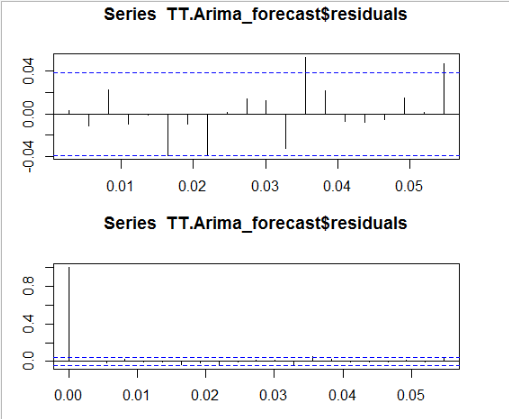
From the above test, we can find the Box-Ljung test statistic to be 28.328, lags between 1-20 and p-value as 0.1019. This shows that the model does not show any lack of fit and its good.

For Exponential Smoothing model:



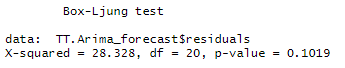
From the above test, we can observe the p-value to be less than the significant value, which concluded that this model shows a lack of fit.

**Partial Correlation & Autocorrelogram**



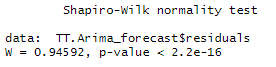
From both the correlogram plots above, we can observe that the sample autocorrelation and partial correlation for the in-sample forecast errors exceeds at lag=13 the significance bounds. However, we could expect one in 20 of the autocorrelations and partial autocorrelations for the first twenty lags to exceed the 95% significance bounds.

Let us now carry out the Ljung-Box test.



It is evident that the p-value of 0.1019 indicates that there is little evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

**Normality of Residuals**



From the above Shapiro-Wilk test for normality, we can observe that the p-value to be far less than the significance level, rejecting the null hypothesis H0 which states that the residual distribution is normal. Therefore, we can conclude that the residuals are **not normally distributed** here.

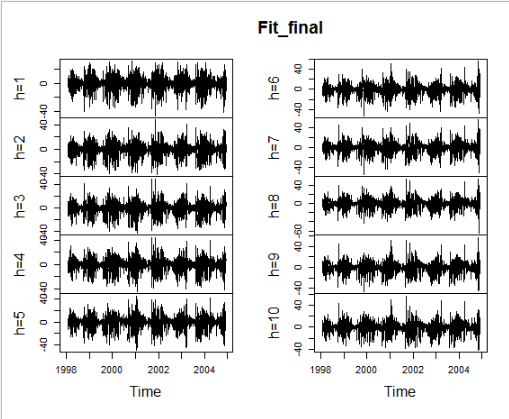
**Best Model**

From the two time-series models we used for forecasting the Ozone Level for the subsequent days, the model using ARIMA is better compared to that of the Exponential Smoothing model since the fit of ARIMA is better than Exponential Smoothing model. Also, for forecasting, the correlation between successive values of time series is important, which is measured only in ARIMA model.

The best candidate ARIMA model chosen for this variable is (3,0,1) which shows a differencing factor of 0, partial autocorrelogram tails off to 0 after lag 3 and autocorrelogram tails off to zero after lag 1.

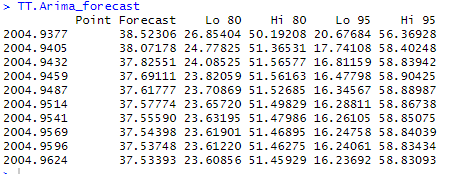
**Cross-Validation:**

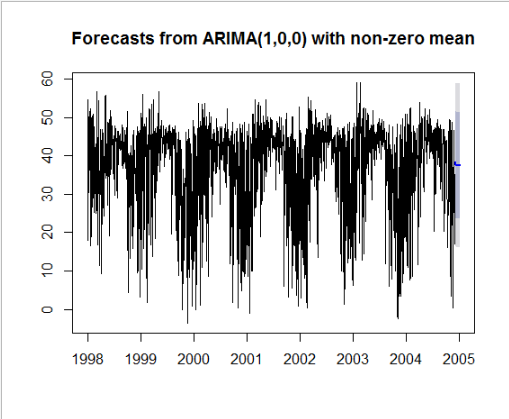
A Cross-Validation for the above ARIMA model with parameter values (p,d,q) = (3,0,1) is computed and plotted below for h= 10 (Forecast Horizon) and window = 30.



**Forecasting using ARIMA Model:**

The forecast of TT (T - Totals) for the subsequent 10 days are as below:



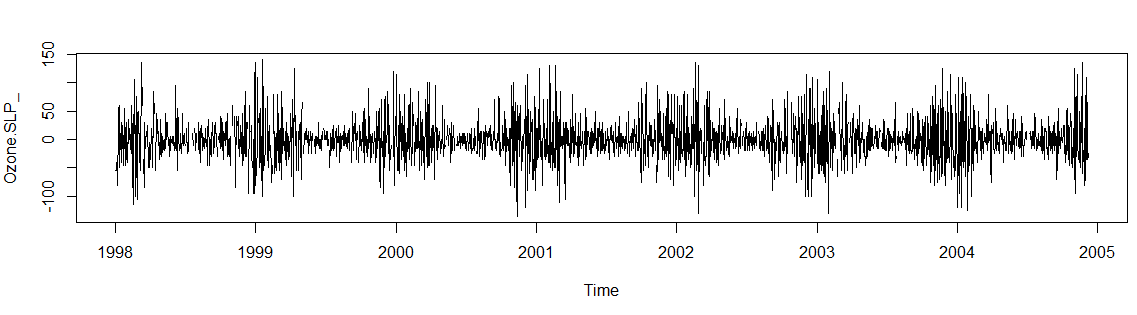


The above plot shows the forecast of the TT variable for the next 30 subsequent days from the end of the day in dataset.

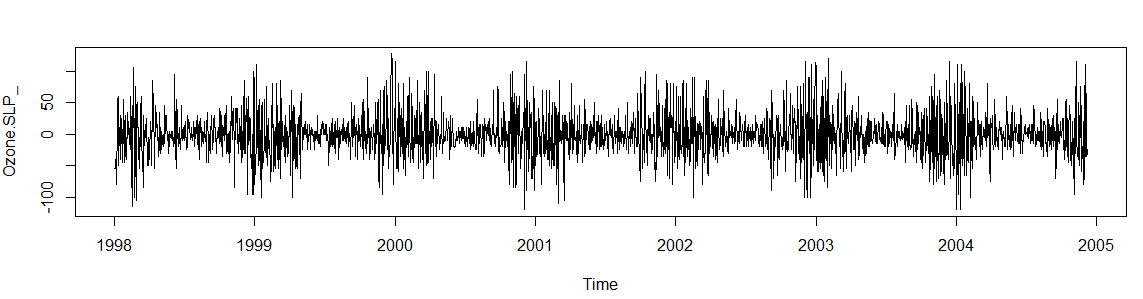
**SLP\_**

**tsclean()** function is used to remove the outliers and the missing values are replaced using linear interpolation.

***Before Removing Outliers & Dealing with NA values***



***After Removing Outliers & Dealing NA values***

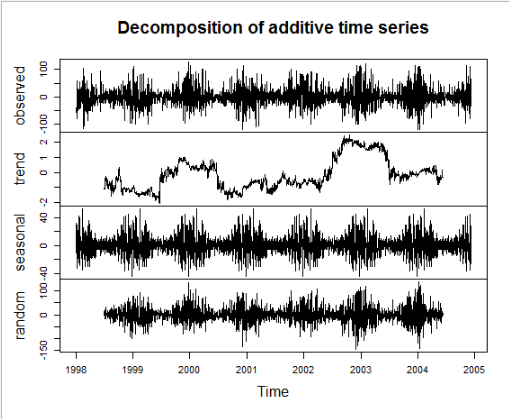


From the above image, it can be observed that the missing values in this key attribute has been replaced by using **Linear Interpolation**, a method of fitting the curve using linear polynomials to construct new data points within the range of a discrete set of known data points.

**Decomposing Time Series Data**

In this, the data is decomposed or separated into its constituent components, which are trend, seasonality and an irregular component.

From the below graphs, we can observe only a trend and no seasonality for this variable. An uneven pattern of trend is visible.



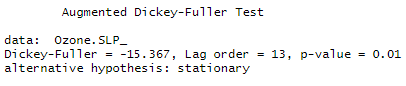
**Test for Stationarity**

**Stationarity** is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has **constant mean and variance**, and covariance is independent of time.

The hypothesis for Augmented Dicky-Fuller Test is,

H0 = Null Hypothesis -> p-value equal or less than 0.05

HA = Alternate Hypothesis -> p-values greater than 0.05

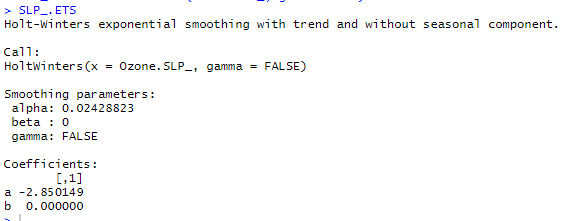


From the above statistics, we can conclude that this time series object is **stationary**.

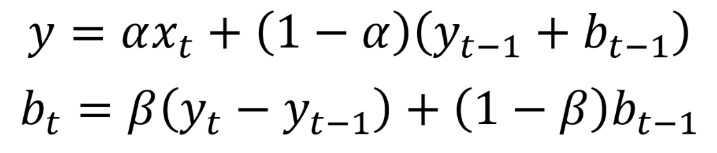
**Exponential Smoothing Model**

Since the time series for WSR\_PK variable can be described using an additive model with a sudden jump in trend and seasonality, **Holt-Winters Exponential Smoothing** is used for this variable.

Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.



The above statistics show that the Smoothing parameters for the model has values such as alpha = 0.0242, beta = 0 and gamma = 0. These parameters convey that the model has no trend and no seasonality. The alpha value is close to 0, which tells that the forecasts are based on previous observations.

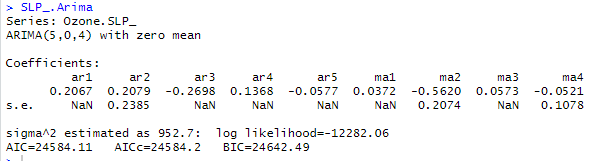


Here, alpha is a **smoothing factor** that takes value between 0 and 1, It determines how fast the weight decreases for previous observations.

The beta is the **trend smoothing factor**, and it takes values between 0 and 1.

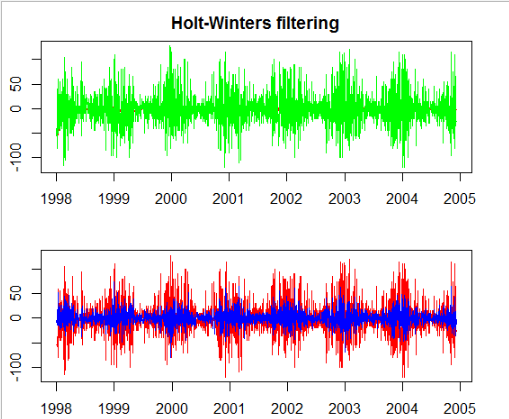
**ARIMA Model**

Since ARIMA (Autoregressive Integrated Moving Average) models are defined for stationary time series, a differencing in time series will be implemented to obtain a stationary time series. This model includes an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.



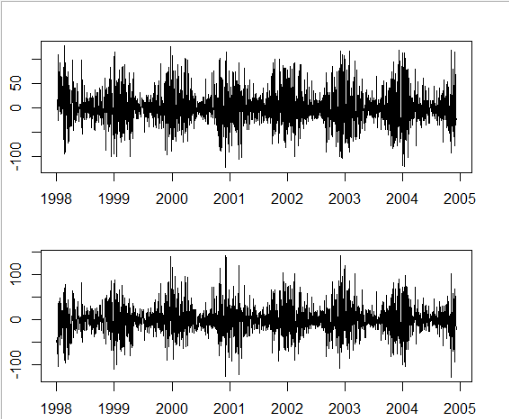
The Goodness of Fit criteria for ARIMA models, **AIC** (Alkaline Information Criterion) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. **BIC** is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

**Observed vs Predicted Values**



From the graphs above, we can observe that the fitted values are more well defined in the ARIMA model (2nd graph) compared to the Exponential Smoothing model (1st graph). It is because, a differencing factor, partial correlogram and autocorrelogram are defined for the below ARIMA model.

**Residuals of Fitted Data**



The residuals in a time series model are the leftovers after fitting a model. In many times series models, the residuals are equal to the difference between the observations and corresponding fitted values.



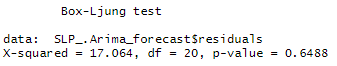
The graphs showing about residuals concludes that the residuals in this variable fulfils both the conditions, ,i.e., the residuals are not correlated, and they have zero mean. Since, both these conditions are satisfied, it can be said that the forecasting methods are proper and good.

**Fit and Measure Statistics**

The Ljung-Box test is a tool to test the lack of fit of a time series model. The test is applied to the residuals of a time series after fitting an ARMA(p,q) model to the data. The test examines m autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

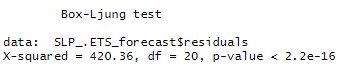
The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) of the Box Ljung Test, H0, is that our model *does not* show lack of fit (or in simple terms—the model is just fine). The [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/), Ha, is just that the model *does*show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series *isn’t*autocorrelated.

Let us now perform this test for the ARIMA model:



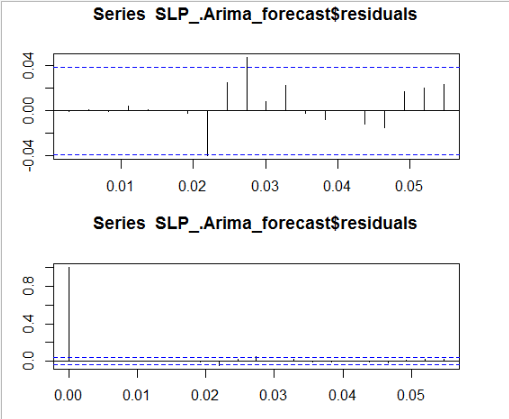
From the above test, we can find the Box-Ljung test statistic to be 17.064, lags between 1-20 and p-value as 0.6488. This shows that the model does not show any lack of fit and its good.

For Exponential Smoothing model:



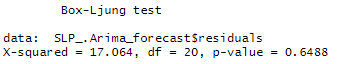
From the above test, we can observe the p-value to be less than the significant value, which concluded that this model shows a lack of fit.

**Partial Correlation & Autocorrelogram**



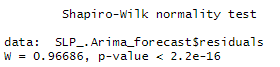
From both the correlogram plots above, we can observe that the sample autocorrelation and partial correlation for the in-sample forecast errors exceeds at lag=10 the significance bounds. However, we could expect one in 20 of the autocorrelations and partial autocorrelations for the first twenty lags to exceed the 95% significance bounds.

Let us now carry out the Ljung-Box test.



It is evident that the p-value of 0.6488 indicates that there is very little evidence of non-zero autocorrelations in the in-sample forecast errors at lags 1-20.

**Normality of Residuals**



From the above Shapiro-Wilk test for normality, we can observe that the p-value to be far less than the significance level, rejecting the null hypothesis H0 which states that the residual distribution is normal. Therefore, we can conclude that the residuals are **not normally distributed** here.

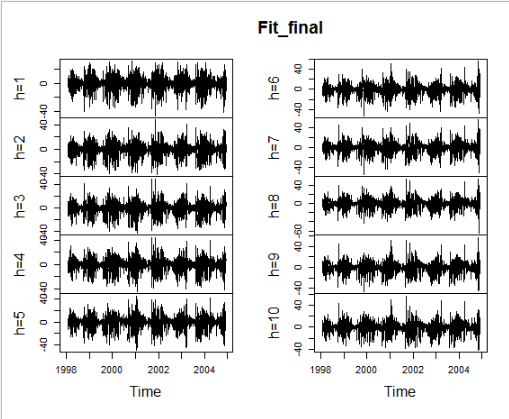
**Best Model**

From the two time-series models we used for forecasting the Ozone Level for the subsequent days, the model using ARIMA is better compared to that of the Exponential Smoothing model since the fit of ARIMA is better than Exponential Smoothing model. Also, for forecasting, the correlation between successive values of time series is important, which is measured only in ARIMA model.

The best candidate ARIMA model chosen for this variable is (5,0,4) which shows a differencing factor of 0, partial autocorrelogram tails off to 0 after lag 5 and autocorrelogram tails off to zero after lag 4.

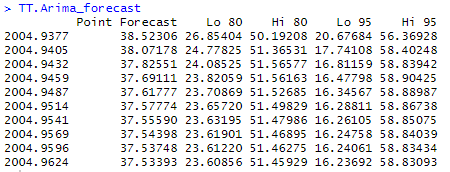
**Cross-Validation:**

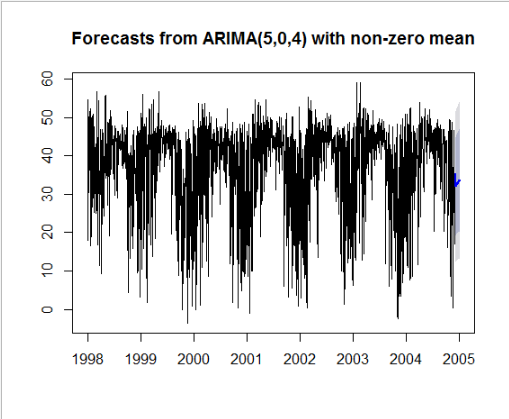
A Cross-Validation for the above ARIMA model with parameter values (p,d,q) = (5,0,4) is computed and plotted below for h= 10 (Forecast Horizon) and window = 30.



**Forecasting using ARIMA Model:**

The forecast of SLP\_ (SLP change from previous day ) for the subsequent 10 days are as below:





The above plot shows the forecast of the SLP\_ variable for the next 30 subsequent days from the end of the day in dataset.

**Conclusion:**

From the above Time Series analysis for the key variables from the Ozone Level Detection dataset, we infer that the ARIMA model shows a better fit for most of the models than the Exponential Smoothing model. It is evident from the graphs of Observed vs Fitted, from Box-Ljung Test for Goodness of fit, Residuals plot, Partial Correlogram and Auto Correlogram plots that ARIMA model stands out from the Exponential Smoothing model. [1]–[6]

**REFERENCES:**

[1] “UCI Machine Learning Repository: Ozone Level Detection Data Set.” http://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection (accessed May 25, 2020).

[2] M. Peixeiro, “Almost Everything You Need to Know About Time Series,” *Medium*, May 19, 2020. https://towardsdatascience.com/almost-everything-you-need-to-know-about-time-series-860241bdc578 (accessed May 25, 2020).

[3] J. Brownlee, “How to Develop a Probabilistic Forecasting Model to Predict Air Pollution Days,” *Machine Learning Mastery*, Sep. 06, 2018. https://machinelearningmastery.com/how-to-develop-a-probabilistic-forecasting-model-to-predict-air-pollution-days/ (accessed May 25, 2020).

[4] “Factor models.” https://psu-psychology.github.io/psy-597-SEM/06\_factor\_models/factor\_models.html (accessed May 25, 2020).

[5] “Using R for Time Series Analysis — Time Series 0.2 documentation.” https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html (accessed May 25, 2020).

[6] “6.4.4.8.1. Box-Ljung Test.” https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4481.htm (accessed May 25, 2020).