1. Summary

The data we are observing is the pond data (pond.RData) of local pond in rural Hampshire from 1966 to 2015. The dataset gives the level of pond water in feet for every month in our timeline from January to December, observed on the first day of every month. We have loaded the data in X for analysing purpose. The water level **ranges** from -6.045747 to 6.000000 feet where the **mean** is 2.517537, **variance** is 6.141367 and **Standard Deviation** is 2.478178. The graph (Fig. 1) shown below clearly dictates that the mean is constant over the years as you can see the purple line(mean) which is been plotted over the graph.

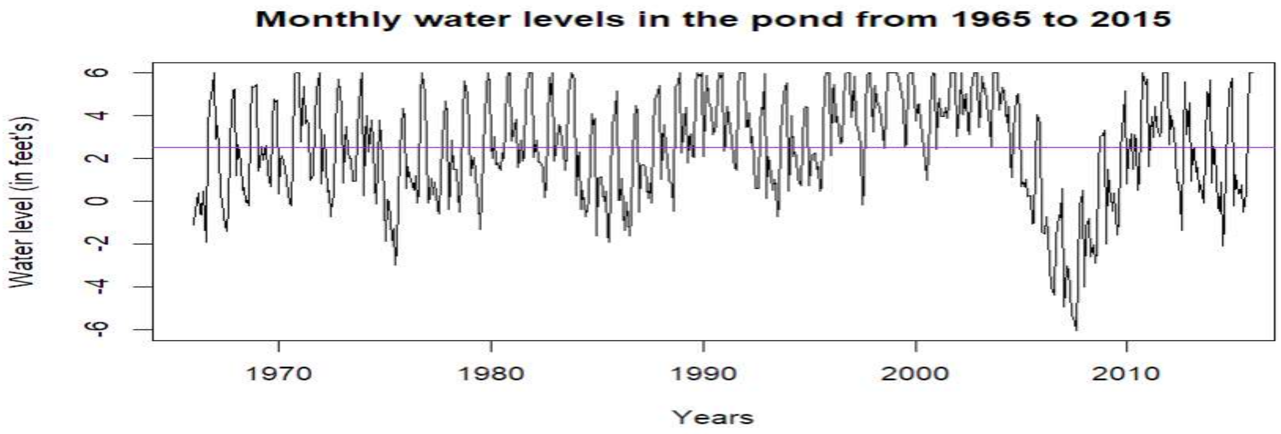

Fig. 1

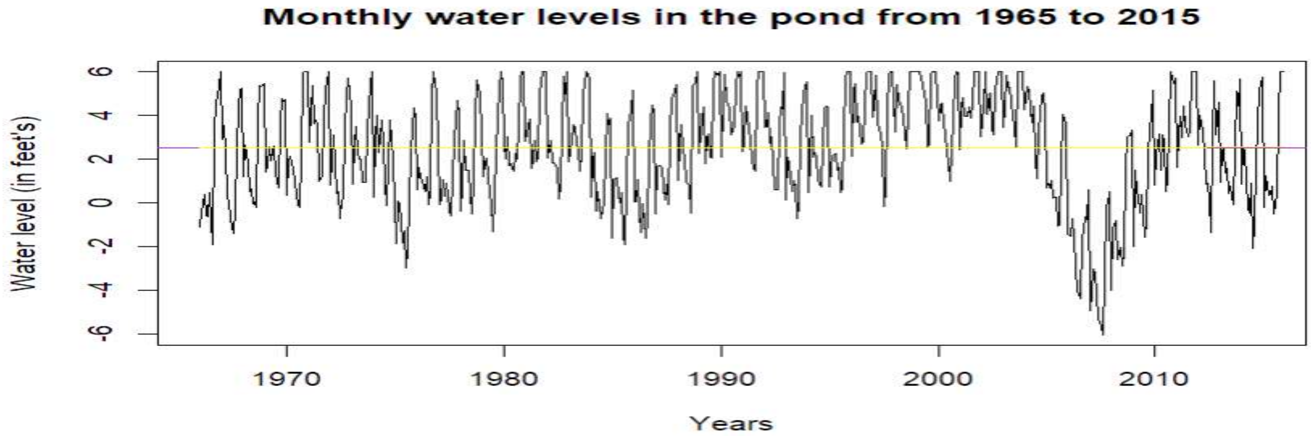

Fig. 2

2. Detecting Trend and Seasonality

Now for detecting that is there any trend or not in our dataset X, we are applying Linear model to our dataset. After fitting the trend on our data, we are seeing no change. As you can observer from the graph that there is not any upward or downward trend. And after fitting the trend the yellow line is passing over mean as you can see in Fig 2. Even after removing the trend, we see no change, therefore there is no trend in our data as we can see in below graph Fig.3

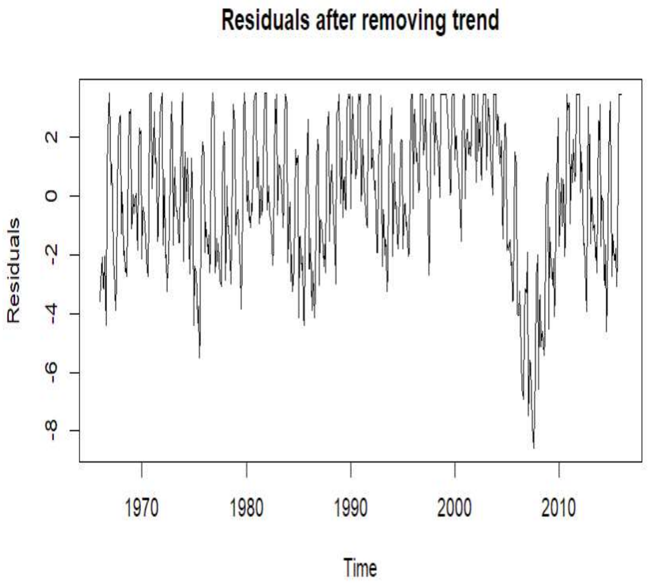

Fig. 3

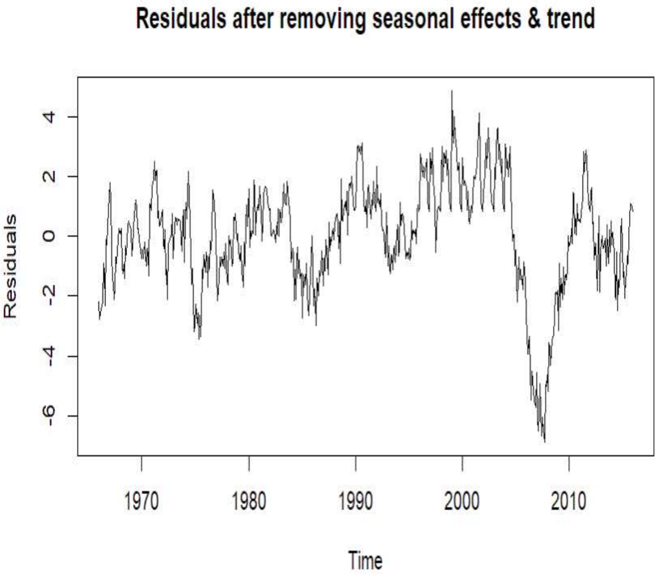

Fig.4

Now we will be analysing the seasonal effect in our data. For that we will be observing the changes in our data for each month every year. In short, we search that does the same change occur in pond's water level every year of that particular month to find seasonality by creating the datapoints every month. Now we will fit the datapoints in our linear model detecting seasonality. After that we are finding residuals which is removing the trend and seasonality from our linear model to find that the effects of seasonality as we already know there is no trend.

Now after observing the residuals(Y) Fig. 4 we can say that there were seasonal effects on our dataset.

3. *Finding the Suitable Model*

Now we will analyse which model is more suitable for our process Y which are White noise, Autoregression or Moving Average models. We can fit them by using auto-corelation function (A.C.F.). The correlogram is a commonly used tool for checking randomness in a data set and even widely used for model identification. If random, autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero. Statistical inference with correlograms is useful for us while observing the process Y.

The property of White Noise is that the Autocorrelation at each lag must be approximately zero. Therefore, for our process to be white noise the residuals at different time lags must be approximately zero. But by observing the Correlogram of our process Y we cannot say the same as the residuals of process Y at different time lags are correlated with each other. Also, we can observe the large no. of spikes outside of the bounds in Fig. 5 Hence, we can say firmly that our process Y is not White noise.

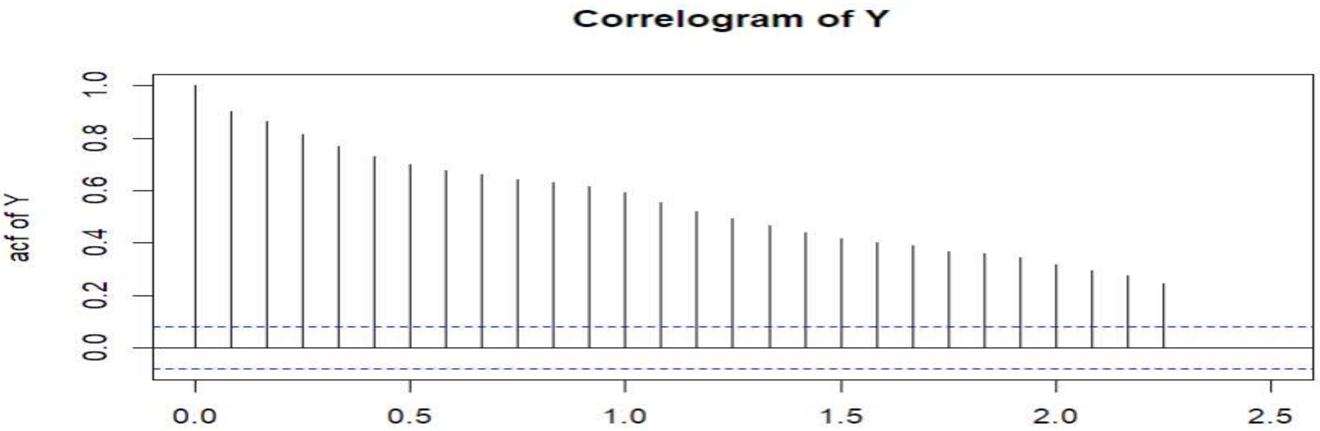

Fig. 5

Hence moving further, we are checking that our process appropriate to fit in moving Average Model or not. So, for that the ACF should show a sharp drop after a certain q number of lags. But our Correlogram does not show any sharp decrease in the residuals of process Y. Hence, we can conclude that our process Y does not fits the Moving Average model.

So, moving further we are checking the Correlogram for Autoregressive model. For our model to be Autoregressive the plot of Autocorrelation should gradually decrease. We can observe the same by looking at our plot Fig. 5 i.e. residuals of our process Y shows a gradual decrease over the lags. Therefore, we can conclude that our process is appropriate for Autoregressive model.

4. *Fitting an AR(p)*

Now that we have concluded that AR model would be the best fit for our process Y. We will now try fitting our process Y into AR model. Going further we need to find the order of Autoregressive model which fits best. First, we will consider up to 3 orders. Therefore we need to estimate the p's that's is p=(1,2,3) and also we will be needing the coefficients α's which we will be generating through Yule Walker's equation for AR(1), AR(2) and AR(3) .

The coefficients we got after running the command in R are as below:

Coefficients(p1): 1 = 0.9025

Coefficients(p2): 1 =0.6709 , 2=0.2568

Coefficients(p3): 1 = 0.6634, 2=0.2374, 3=0.0288

As you can see the coefficients of p2 and p3 are almost similar so that we should first observe till order 3 and then decide of going further.

5. Residuals of AR(p)

For finding the best fit of 3, we will find residuals for all 3 orders of our process Y and use Autocorrelation function to plot correlograms for all 3 orders of Autoregressive model, so that we can observe and find the best fit.

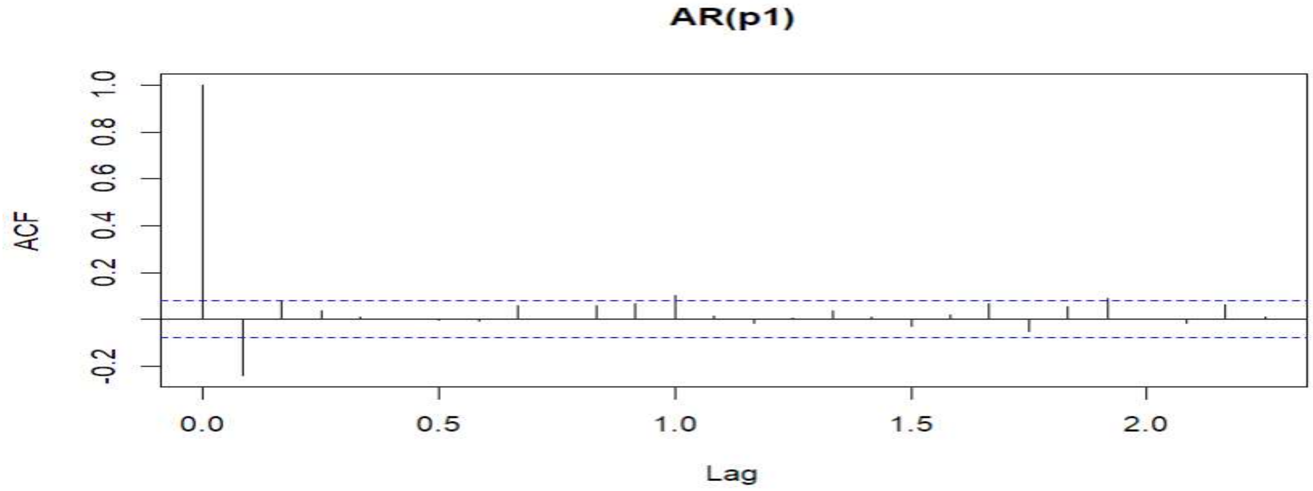

Fig. 6

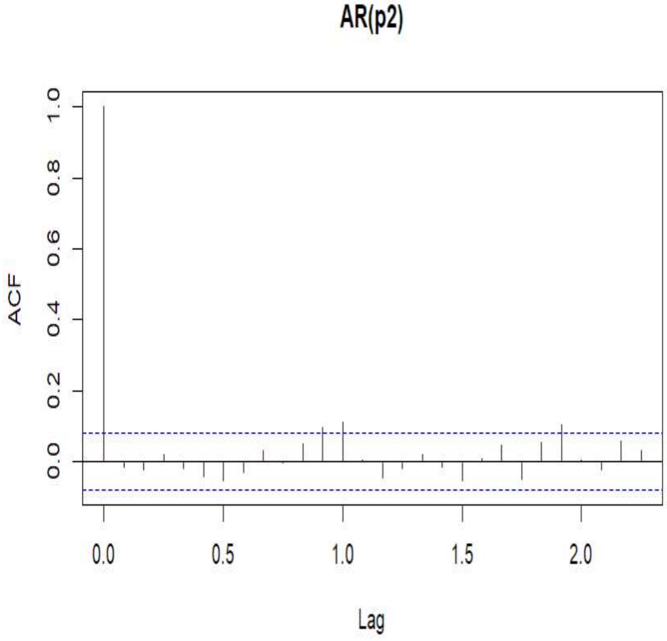

Fig. 7

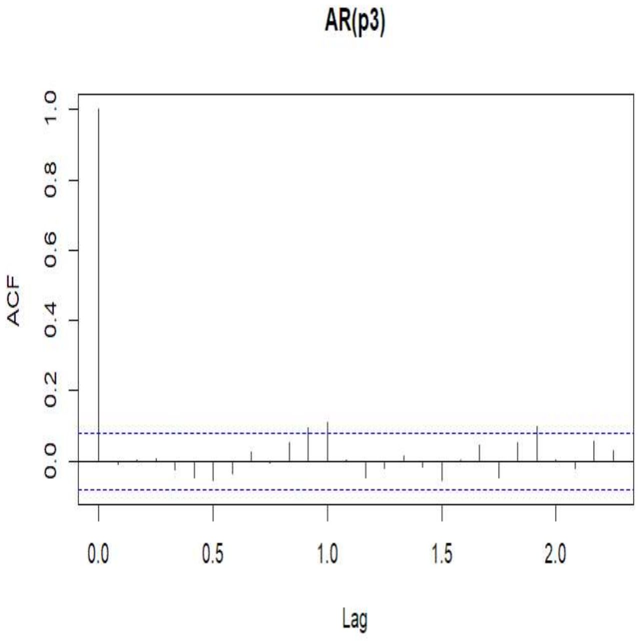

Fig. 8

The above correlograms of AR (1), AR (2), AR (3) are Fig. 6, Fig. 7 and Fig. 8 respectively. Now we will analyse and conclude which correlogram suits the best by stating below points:

1. After observing all the 3 correlograms we can say the AR (1) is the least suitable model as we can see more no. of spikes of ACF outside the bounds of Statistical inference. Therefore, on this basis we can reject AR (1)

2. As we can see that correlograms of AR (2) ($\alpha1=0.6709$, $\alpha2=0.2568$) and AR (3) ($\alpha1=0.6634$, $\alpha2=0.2374$) are almost same, and even the coefficients are almost same but are of different order, as $\alpha3 = 0.0288$ for AR (3) is very small which is insignificant.

3. Therefore, we can say that choosing AR (2) would be a good choice because by choosing AR (3) we are making our model complicated and less efficient by increasing computation as $\alpha3$ for AR (3) is insignificant we will get similar outputs by choosing AR (3). Therefore, we are choosing AR (2) which is simple model and less complicated with similar results.

As we have confirmed that AR (2) model best fits our process. So going forward we will call the residuals of AR (2) as Z.

Below graph is the residual of AR (2) model for Y:

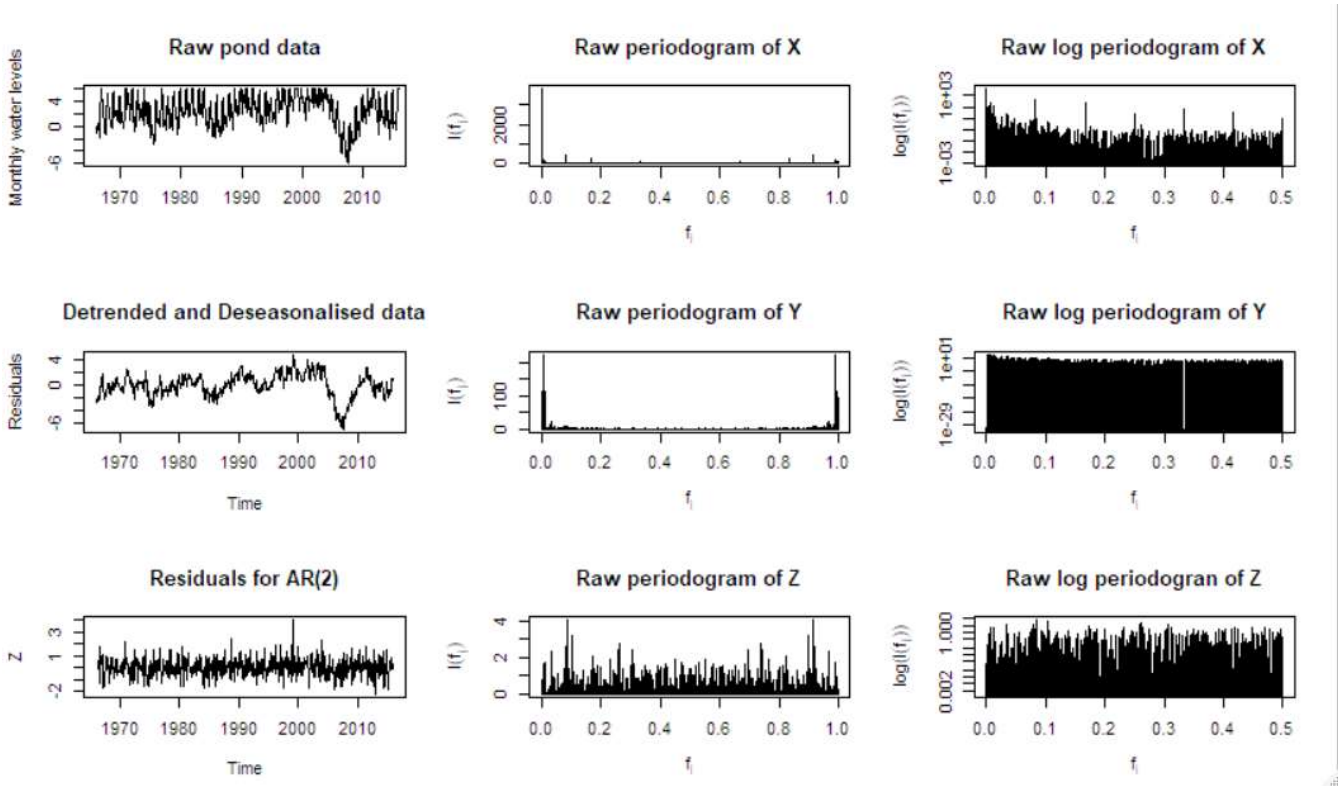

Fig. 9

We can conclude below statements by observing above graphs (Fig. 9)

- A periodogram is used to identify the dominant periods (or frequencies) of a time series. This can be a helpful tool for identifying the dominant cyclical behaviour in a series, particularly when the cycles are not related to the commonly encountered monthly or quarterly seasonality.

- In the above fig we have shown the data plots for X (raw pond data), Y (residuals of data after removing the trend & seasonal effects) and Z (residuals after fitting AR (2) model). We have plotted their raw periodograms and the raw log periodograms.

- After the removal of seasonality from our dataset, we can observe that log periodogram of Y that there is fall in frequency at 0 but still it shows presence of some spikes from which we can say that model like AR and MA would be a suitable fit.

- We can observe that the raw periodograms of Y (detrended & de-seasonalized data) and Z (residuals after fitting AR (2) model) show the symmetricity from the range of [0,1]. The raw periodogram for the raw pond data at 0 is high which shows that this is important frequency in our Time-series. And the small spike at different frequencies shows there is seasonal effect.

- As we can see from the raw periodogram of Z (residuals after fitting AR (2) model) the spikes have been removed & the dominance of any frequency is absent.

- The logarithms of spectral density are taken & log periodograms have been plotted in order to observe the frequencies closely for all our X, Y & Z. As the periodograms are symmetric due to aliasing, the interval from [0,5] has been considered.

- From raw log periodogram of X, we can say that there is no trend as there are very less large values present nearby 0.

6. Complete Model for Time Series X

After all of the analysis we have done on the pond data X, the following inferences can be made:

- We checked for the trend and seasonality in our dataset X and found that there was no trend, there was only seasonality effect.
- Then we removed seasonality from the linear model and plotted the residual. Then using ACF we plotted correlogram of residual.
- After observing the correlogram we found out that AR model would be the best fit for our residual of Y.
- By observing further, we choose AR (2) as it was simple and less complicated which was similar to AR (3)
- By plotting the residuals of AR (2) we got to know Its residuals are close to white noise.
- The representation of AR(2) model for the residuals Y is:

$$Y(t) = 0.6708Y_{t-1} + 0.2568Y_{t-2} + \varepsilon_t$$

- Our final model is:

X_t = - 1.3954jan - 0.1479feb + 0.3502 march - 0.7914apr - 0.9047may - 1.1336June - 2.1533july - 2.1083aug + 0.8768sep + 2.3493oct + 2.4439nov + 2.6145dec + Y(t)