Report

TIME SERIES AND SPECTRAL ANALYSIS(5802)

Himanshu Jaiswal

Student id: 201577162

Introduction

This report is regarding studying, analyzing, and drawing the conclusions of the pond data which is given to us which consists of the water levels measured on the monthly basis from 1966 to 2015 in a small pond in Hampshire. Based on studying the data we have tried to find out different constituents of the time series to eventually get rid of them (like trend & seasonality).

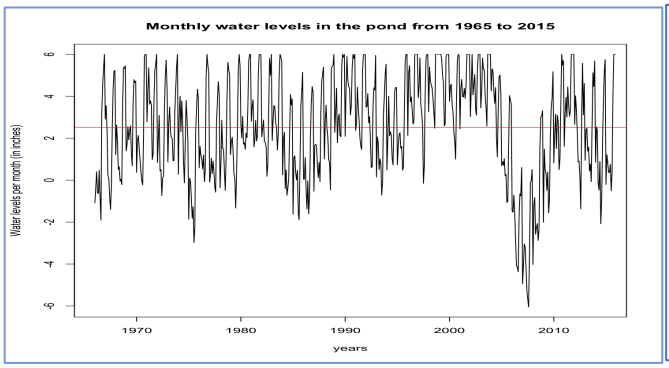
Furthermore, we went on to study which model would be the most appropriate fit for our data. Based on these studies, various patterns & graphs we went on to check whether the moving average model, white noise or the auto regression model would be the best fit. These studies, plotting of several graphs & inferring from them took us to our conclusions about the best model for our data. In order to further analyze the data, periodograms are being made of the raw data as well as of the residuals once the different components are removed such as trend and seasonality and fitting of the appropriate model based on our studies & inference.

The complete model would then be obtained for our time series X after our all of the analysis is done which will incur the estimated coefficients calculated when the models were being fit. This model would be useful to predict the water levels in pond in the upcoming years.

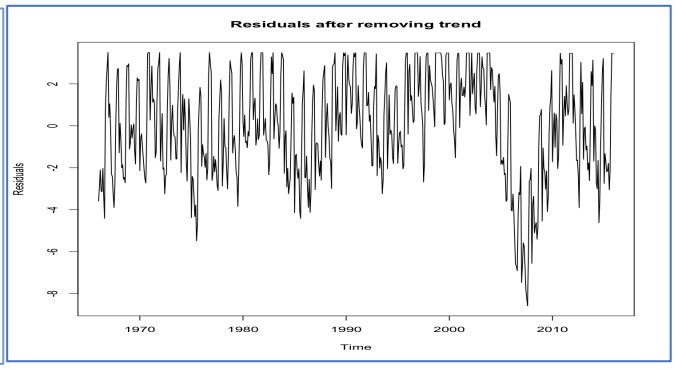
Findings and Reasonings

Summary of pond-data X

As we load the pond.R data in X & try and analyze it, we can observe that the pond water levels for each month from 1966 to 2015 is shown. The mean of the water level being 2.51 over this entire period ranging from -6.04 to 6.00(in inches), the variance being 6.14 & the standard deviation being 2.47. The data is clearly fluctuating over the mean water-level in the pond both above & below as we can see from the graph the red line representing the mean (fig 1)



(fig 1)

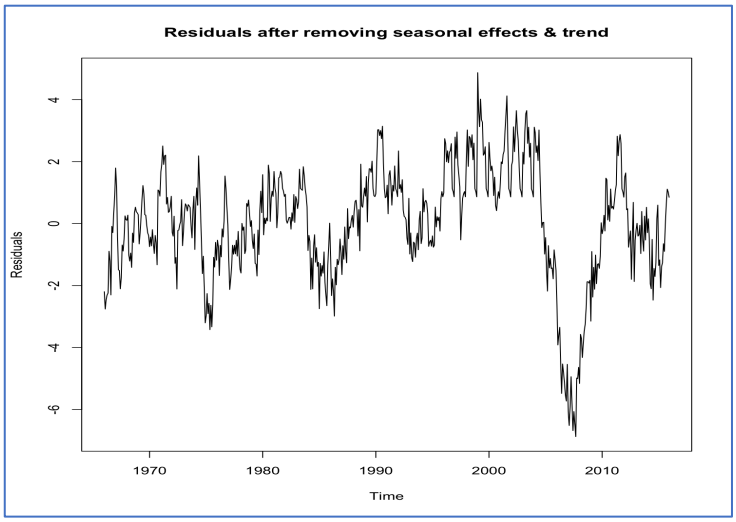


(fig 2)

Trend and Seasonality

As we move forward the application of Linear Regression is used to find the effects of trend and seasonality on our data. Trend is calculated and we're fitting the data after removing this trend i.e., residual Y is plotted (fig2) from which it is obvious that there is no trend in the given data as it doesn't have any effects on the data when compared to (fig1).

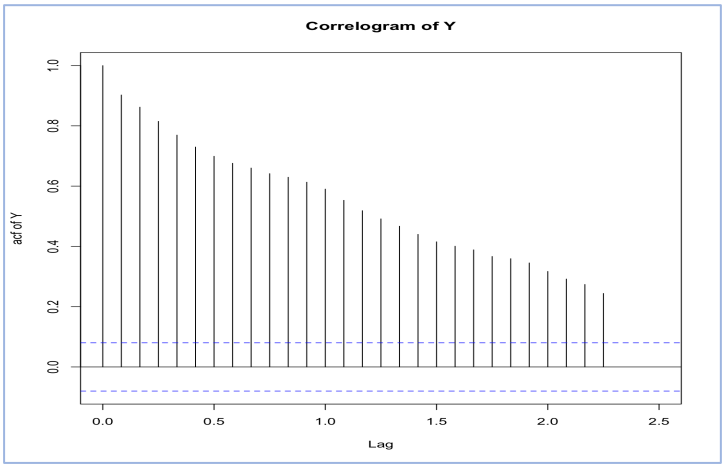
Now we will be analyzing the seasonal effects on the data which can be done with the help of creating 12 indicators for each month, fitting this into Linear model and calculating the seasonal effects in R. Further we will be plotting the Residuals(Y) without the calculated trend and seasonal effects. As from the (fig3) we can observe the residuals which clearly indicates that there were seasonal effects on our monthly pond water-levels data.



(fig 3)

Finding Suitable Model for Y

Now as we move further, we try and analyze whether the process which is the residual (Y) without the trend and seasonal effect is White Noise, Moving Average process or Auto Regressive with the help of a.c.f. function. We can do this by using correlograms which is used for plotting a.c.f's at different lags and also shows the bounds of Statistical Inference, blue lines as shown in (fig 4).



(fig 4)

Now, if the process (Y) is to be White Noise, then the a.c.f's at each lag has to be zero or very close to zero i.e. the residuals at different time lags should not be correlated to each other but from the (fig 4) which shows a.c.f of residual (Y) it is clear that the correlation between the data at different lags is very high so it cannot be a White Noise. Another reason being as we can see in (fig 4) there are large number of spikes outside of the bounds (blue-lines) hence the process Y is not a White Noise.

As we go on to check whether the process Y is Moving Average (MA) we must observe whether the residuals (Y) cuts-off at some lag q after which the values start decreasing extensively from the correlogram. As a.c.f of Y (fig 4) does not show any such sharp drop even after some lag q so it can be concluded that the process Y is not MA.

On the other hand, the autocorrelation between our residuals at different lags is quite significant as well as a.c.f of Y shows a gradual decreasing pattern from which we can say that the Auto Regressive (AR) model would be more appropriate for the Process Y.

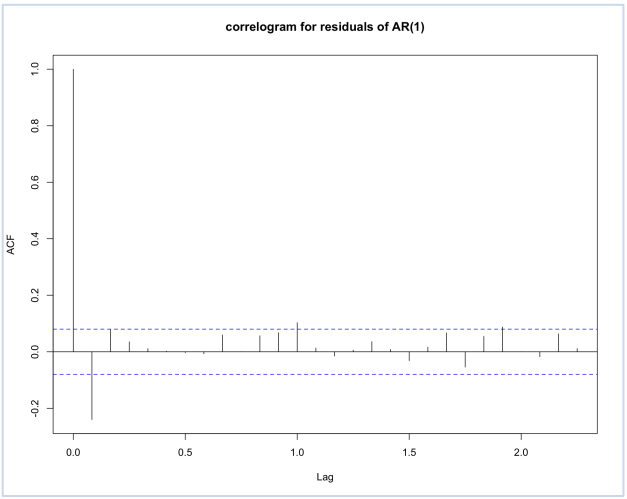
Fitting AR(p) Model

Now that we have analyzed and verified that the AR model would be the best fit for modelling our data, we will now try and fit AR(p) model to the residuals Y which we have obtained after removing the trend and the seasonal effects from our data X. The first thing to do would be to obtain the estimates of p's i.e. (p=1,2,3) & then solve the Yule Walker's equations to generate our coefficients α's for AR(1), AR(2) & AR(3) respectively. This can be done using the following r commands:

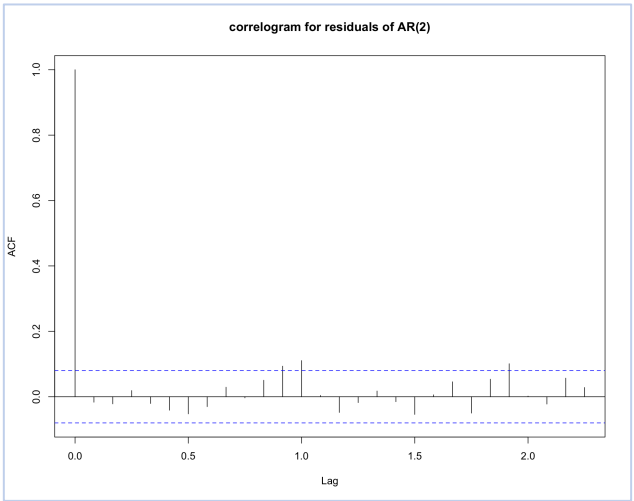
```
(Y.ar1=ar(Y, order=1,aic = FALSE))
Coefficients: 1 =  0.9025

(Y.ar2=ar(Y, order=2,aic = FALSE))
Coefficients: 1 =0.6709 , 2=0.2568

(Y.ar3=ar(Y, order=3,aic = FALSE))
Coefficients: 1 = 0.6634, 2=0.2374, 3=0.0288
```

Finding residuals of AR(p)

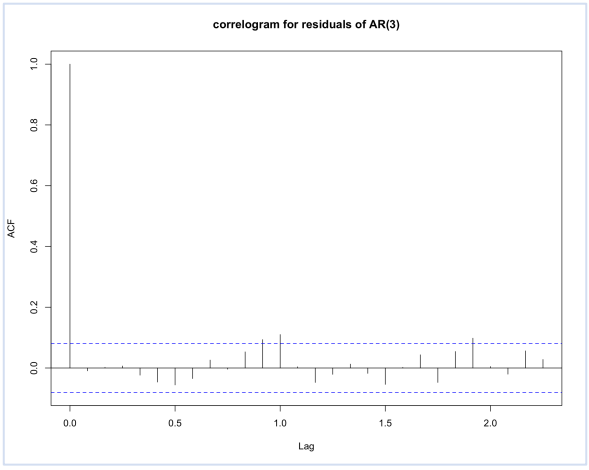
As we have considered all three models i.e. AR(1), AR(2) & AR(3) & calculated their respective coefficients i.e. α's, we will now go forward and plot the correlograms of their residuals for each one of them. All of these correlograms can be seen below:



(fig 5)



(fig 6)



(fig 7)

From the above correlograms of the residuals of AR(1) in (fig 6), AR(2) in (fig 7) & AR(3) in (fig 8) we further go on to analyse and draw the conclusions as follows:

- As we observe the correlogram of residuals for AR(1) model, we can see that there are number of spikes of ACF function outside the bounds of the statistical inference which are the permissible limits. Based on this observation we can say that AR(1) model would not be suitable for us.

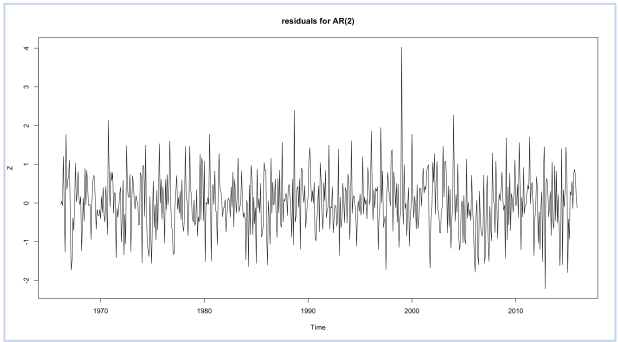
- Now as the AR(1) model is rejected we go on to observe & analyse the correlograms of the residuals for AR(2) & AR(3) models which appears to be a good choice from (fig 6) & fig (7).

- As we can observe from the above r functions the coefficients $\alpha 1$, $\alpha 2$ are almost same for both AR(2) model ($\alpha 1=0.6709$, $\alpha 2=0.2568$) & AR(3) model ($\alpha 1=0.6634$, $\alpha 2=0.2374$), estimating $\alpha 3 = 0.0288$ for AR(3) as we can see is very small value which is statistically insignificant.

- It just wouldn't be a good choice considering another coefficient without it being significant as such, this would only make our model complicated and less efficient as well. So it would be better to go with the simpler model which in this case is clearly AR(2) without considering another insignificant coefficient.

From the above reasonings & findings we can crisply justify that the AR(2) model would be the best fit.
We now refer the residuals of AR(2) as 'Z'.

As shown below we're plotting the residuals for AR(2) model for Y:

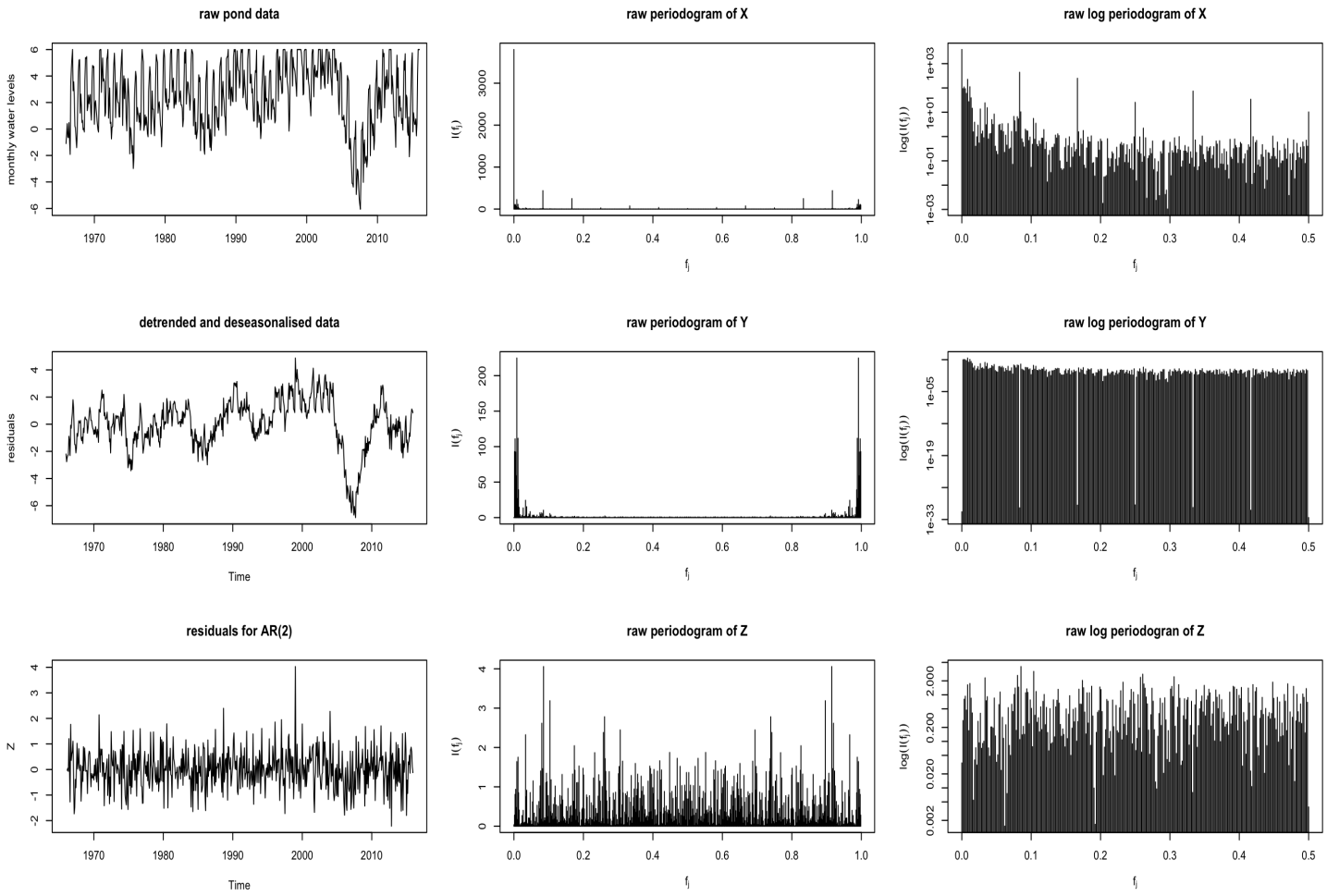


(fig 8)

Now for Y(AR(2)), as we can see the residual and correlogram plot (ACF functions) most of our data i.e. almost 95 % of sample autocorrelations lie between the statistical bounds (blue lines) from which we can say that residuals for AR(2) model looks like White Noise.

Plotting periodograms

For all of our X (raw pond data), Y (residuals of data after removing the trend & seasonal-effects) and Z (residuals after fitting AR(2) model)



(fig 9)

From analysing the graphs above, the following information can be inferred :

1. Periodograms are basically used to identify the dominant frequencies of our time series and also calculate their significance. 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 also have plotted their raw periodograms and the raw log periodograms.

2. 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 signifies that this is a hugely important frequency in our Time Series. The presence of small spikes at other frequencies signifies very clearly that there is some seasonal effect in the time series.

3. Once the seasonality is removed from the data, we can see from the log periodogram of Y that there is a fall in frequency at 0 but still there is the presence of some spikes from which we can infer that the models like AR & MA would be a suitable fit for this time series.

4. 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 particular frequency is absent.

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

6. As we can see from the raw log periodogram of X, there's no very large values present around 0 from which we can say that there's no trend present as such.

7. Also we can observe that there are some leaks in the data from the last graph of (fig 9) which are present for all the frequencies.

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:

1. We checked for the trend and seasonality in our data & found that there were seasonality-effects on our data which was then removed. There was no trend in our data.
2. After the removal of seasonality, residuals were plotted & studied to find the best suitable model with the help of plotting correlograms using a.c.f function.
3. After studying these correlograms we found out that the best fit would be Auto Regressive model. After further analysis AR(2) model was then selected.
4. Now after plotting the residuals of AR(2) model, we observed that these are very close to White Noise.
5. As we go on to build the final model for Y (detrended & deseasonalized), this will have the coefficients of seasonal effects & the AR(2) model.
6. The representation of AR(2) model for the residuals Y is:

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

- Therefore, our final model is:

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

Conclusion

The analysis in this report is being done on water levels in the pond on monthly basis for 50 years. Initially the check on the stationarity of our data was done by plotting X. We then found that there was no trend but the seasonality was present in the data. The residuals Y were considered after fitting our model for the seasonality & calculating the monthly coefficients. Ideally if the errors were to be white noise, we would stop but as it wasn't we went on further modelling our data & found that it was highly correlated. We then went on to plot the correlograms to find the most suitable model which in this case was to be Auto Regressive model for fitting our data. Using the Yule Walker equations, the fitting of AR model was done for (ρ=1, 2, 3). We went on to plot the Auto Correlation function for their residuals studying which we found that the AR(2) would here be the ideal choice. Once the AR(2) model was fitted on our data & the coefficients were calculated, we plotted the residuals studying them we reached the conclusion that it is white noise. The complete model for our data X comprises of the seasonal component, auto regression model of order 2 as well as the error term. We concluded this report by plotting the periodograms for all of our three variables X, Y & Z to check on the dominance of frequencies in our data.