Time Series Forecasting

Assignment 1

Taylor Spinks - 10417229

Contents

1 Task 1 - Plots and Descriptions of Data

The following are Time Series plots and Autocorrelation Function plots for the lake.data and aus.data datasets. An analysis is performed on the plots to determine the time series characteristics; Trend, Seasonality, Cyclic, and the Irregular components.

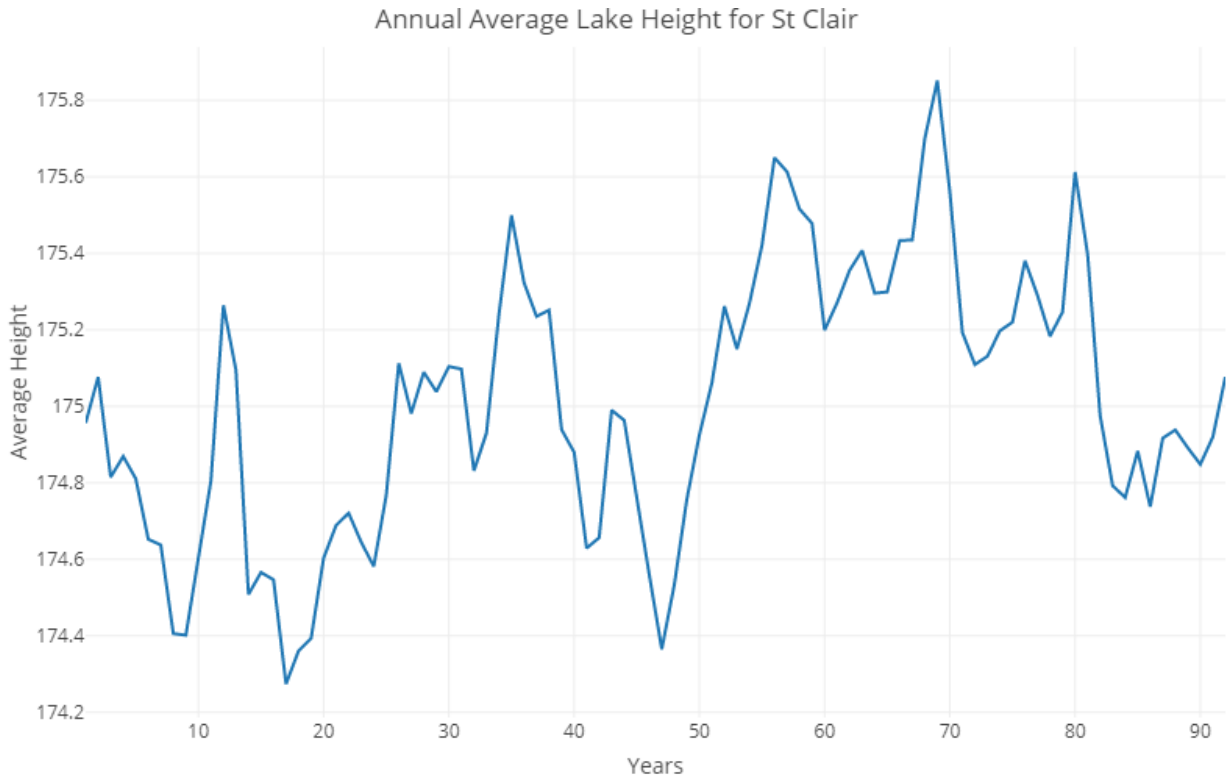
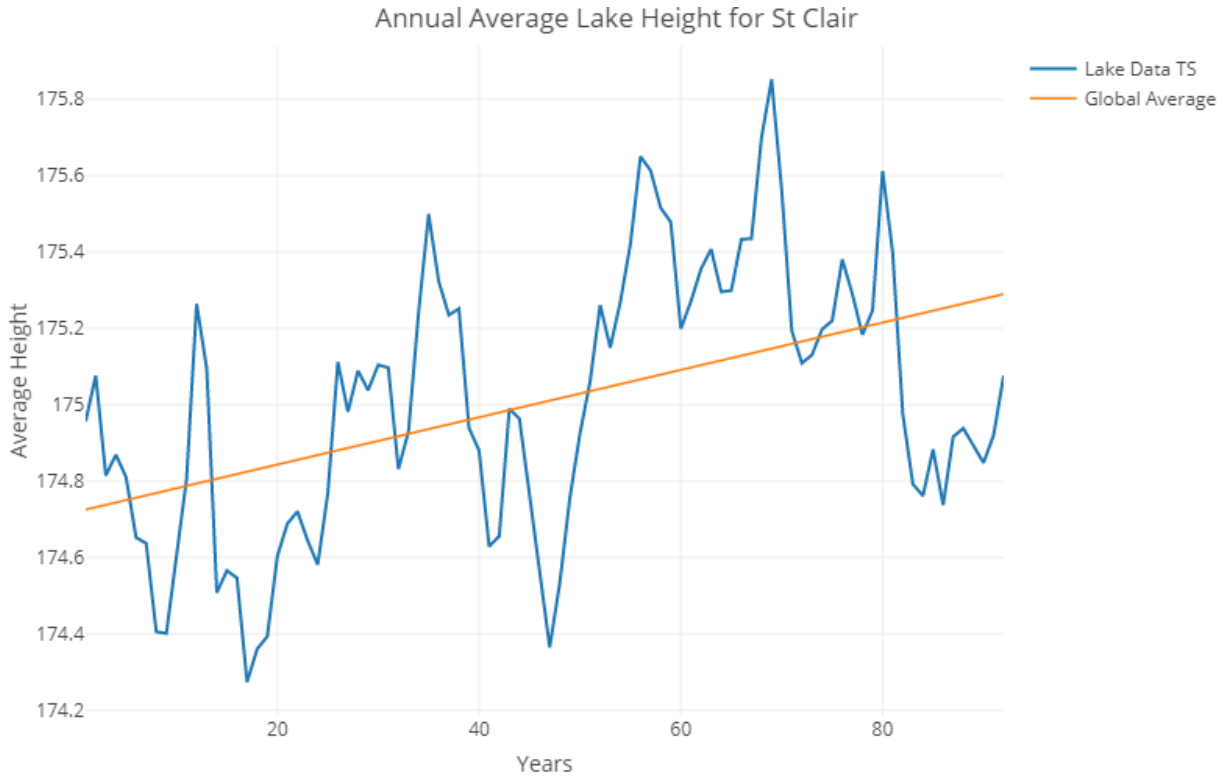


Figure 1: Lake data

This is a Time series plot of the St Clair average annual height data from the DAAG great lake's dataset.

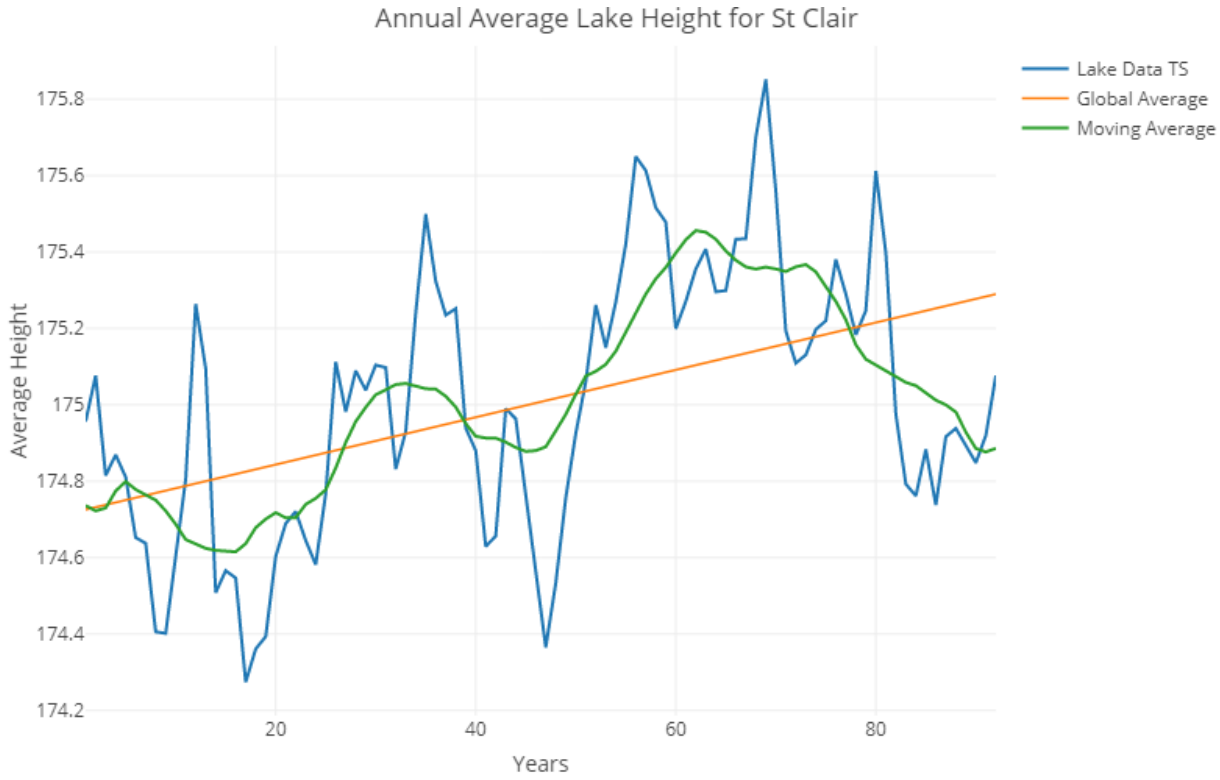
1.1 St Clair Time Series Annual Average Lake Data Initial Analysis

This time series shows 92 years of lake height data for St Clair. The graph shape appears to have a global slight positive trend, due to what looks like a change in the mean. Without doing a deeper analysis it is difficult to accurately determine if this is true, as it could be by chance that this positive trend occurs. The reason for this suspicion is how the trend seems to be comparable to Gaussian White Noise, which may indicate that there is minimal correlation with lake height over time to 2009. Or it may indicate that a larger time slice is required to better understand the relationship. It is however necessary to perform an initial analysis to see if this time series is worth investigating further.

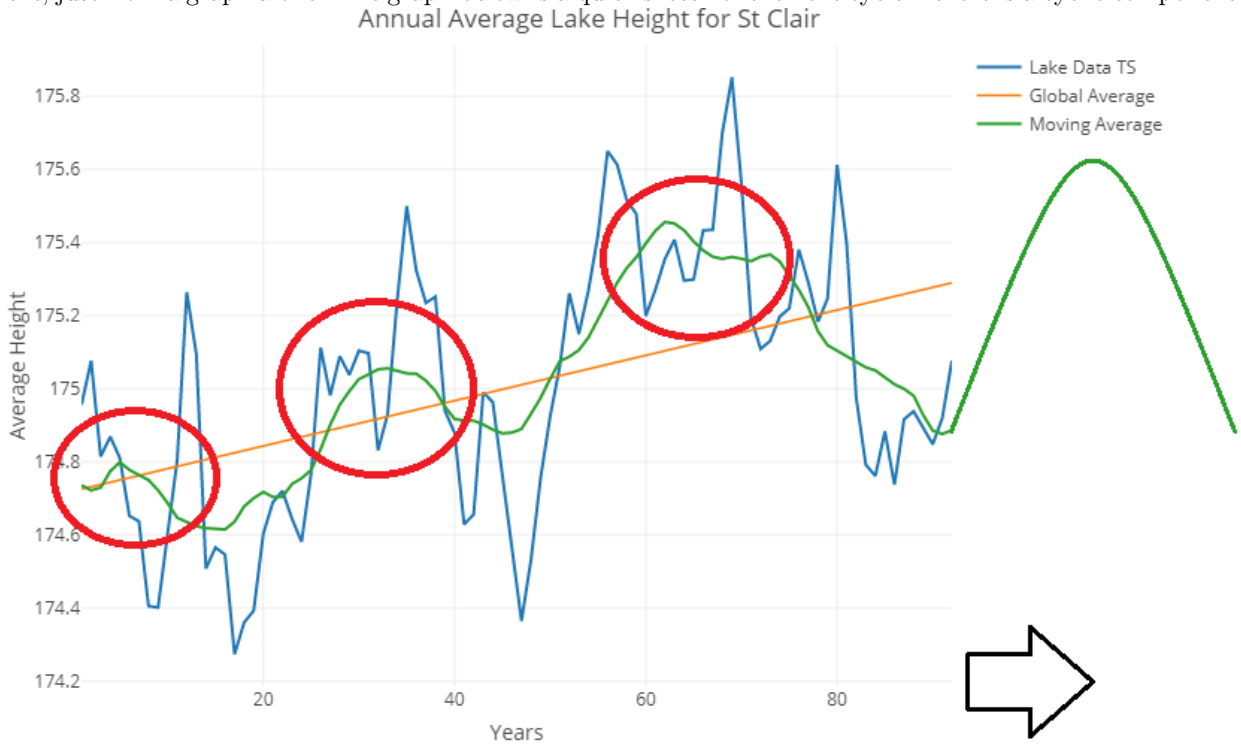


As suspected there is a slight average trend, however this perhaps is not that meaningful, as there may be some amount of irregularity in the Time Series indicated by the line plot. Irregularity is where a component is much larger than what is expected from the normal distribution (shown better by an AFC). Due to these peaks in the data, it may and in this case, likely does, increase the global average. Irregularity will be confirmed later with Autocorrelation.

At this scale, it is difficult to determine seasonality, as the data is recorded annually. Which makes sense as generally seasons only occur within a year, and not yearly. It is unlikely in this case that there is seasonality from year to year. Seasonality might exist however if the data was recorded semi-annually, for the Winter-Summer seasons, due to an increase in rainfall in a season.

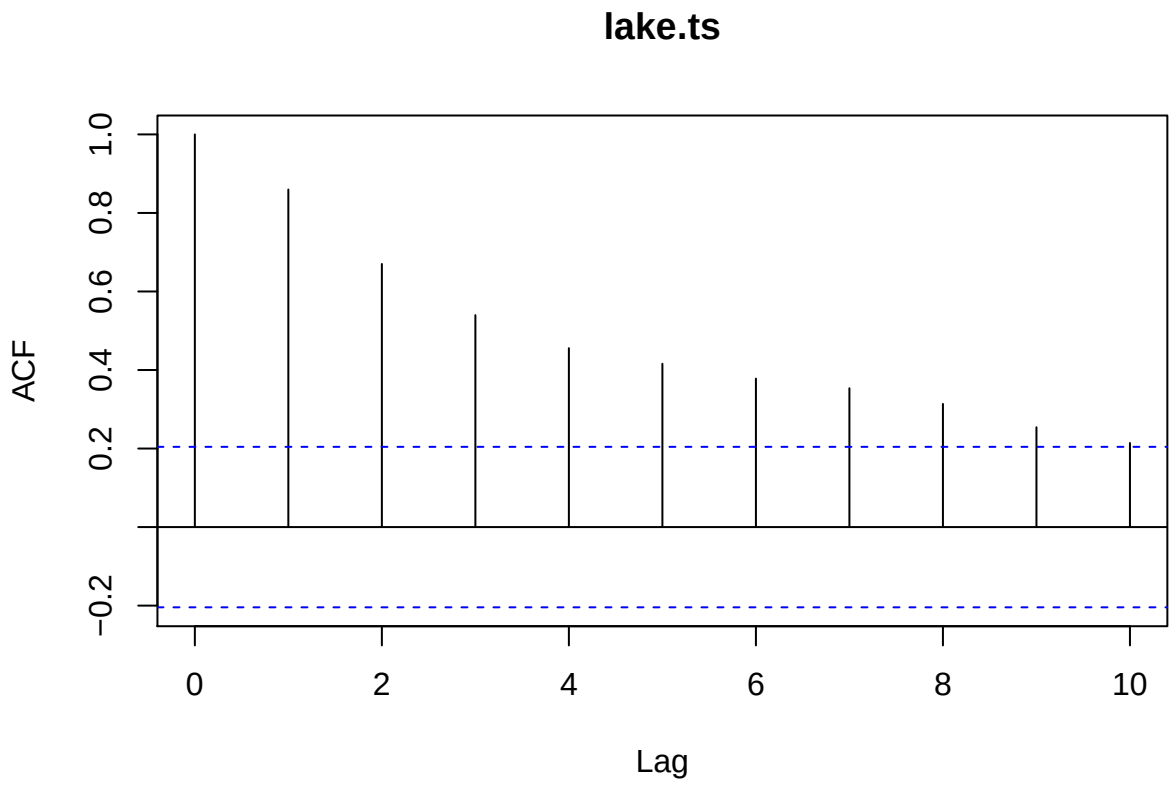


In the graph above there is possibly a slight cyclic component, as indicated by the moving average. There seems to be a new cycle approximately every twenty years. However on the last cycle it breaks down. So a much larger resolution for this data would be required to conclude that there is a cyclic component, just from a graph alone. The graph below is a quick sketch of the next cycle if there is a cyclic component.

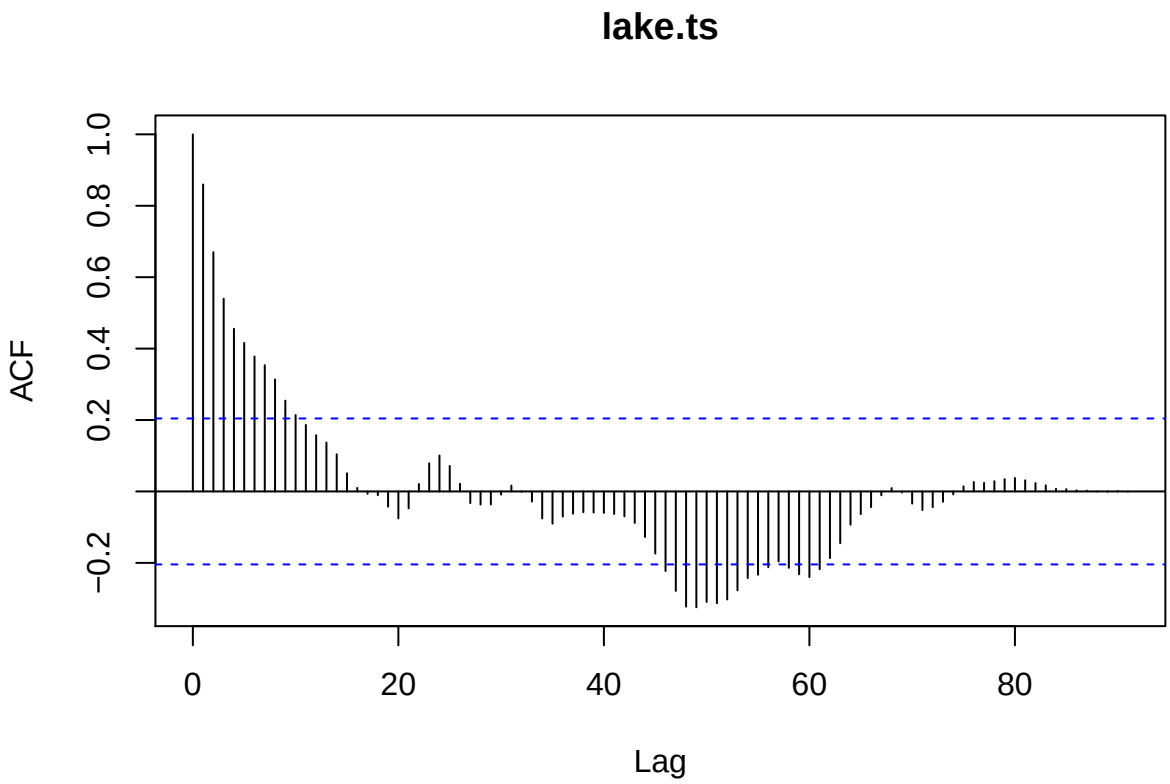


1.1.1 Auto Correlation for the St Clair Data

Auto Correlation looks at the relationship between current data and past data at specific lag intervals. The lag refers how far back to look, and then to compare that with the present. In R, to ACF function can be run to produce a correlogram, which shows this relationship. Patterns in this a correlogram can assist in identifying the components for the Time Series; Trend, Seasonality, Cycles or Irregularity.

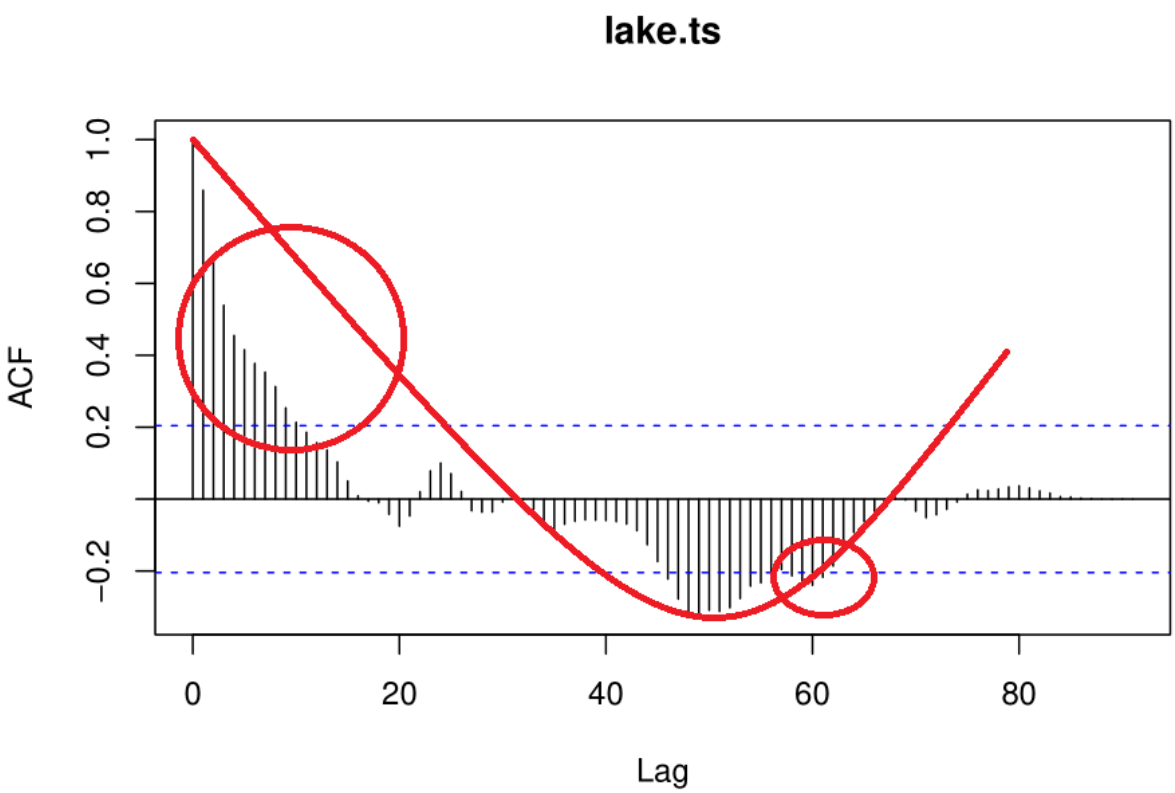


In the correlogram above with a lag of 10 and a confidence interval of 0.95, it seems there is a trend. However to confirm this, the lag needs to be expanded and to see exactly what kind of trend is the strongest.



Increasing the lag to 100 shows that there is a slight resemblance of a quadratic when focusing outside of the confidence interval. After cross-analyzing the global mean on the Time Series and the ACF correlogram, there is possibly some amount of linear positive trend. There is confirmed to be no seasonality present, as discussed in the Time Series analysis. At looking at the ACF, it seems there is no cyclic component to this Time Series. It is possible that with a larger resolution, this could change. The reason a cyclic was expected was due to prior knowledge that climate experiences cycles of variation. 92 years may not be long enough to observe this.

There is definitely a bit of irregularity, as it does slightly vary from a what would be a linear trend in some areas of this graph. As seen from this quick analysis below. A deeper analysis into how it fits the quadratic will likely be required in the future.



As can be seen from this quick sketch, while it vaguely follows the quadratic, there are locations where this isn't true that are sigificant. From the indicated areas, it possible to see that the bars drop off faster than they should than if it were perfectly linear. There is also an increase of the bars in another region that move outside of the approximated quadratic. These parts of irregularity shows there are some areas that vary from the linear trend. This isn't Gaussian White Noise as suggested before, however there is irregularity.

In conclusion there seems to be a mix of trend, irregularity and possibly some cycles, all of which are not that strong. It seems that the irregular component is one of the strongest as well as the trend component, which is positive.

1.2 Australian Eletricity Quarterly Production Time Series Initial Analysis

The following time series shows Australian electricity production from 1956, through to 1994 for a total of 38 years. The line plot Time Series shows a strongly positive linear trend, indicating that it is increasing over the 38 years. There is also a strong degree of seasonality, as the eletricity consumption will increase depending on the time of the year.

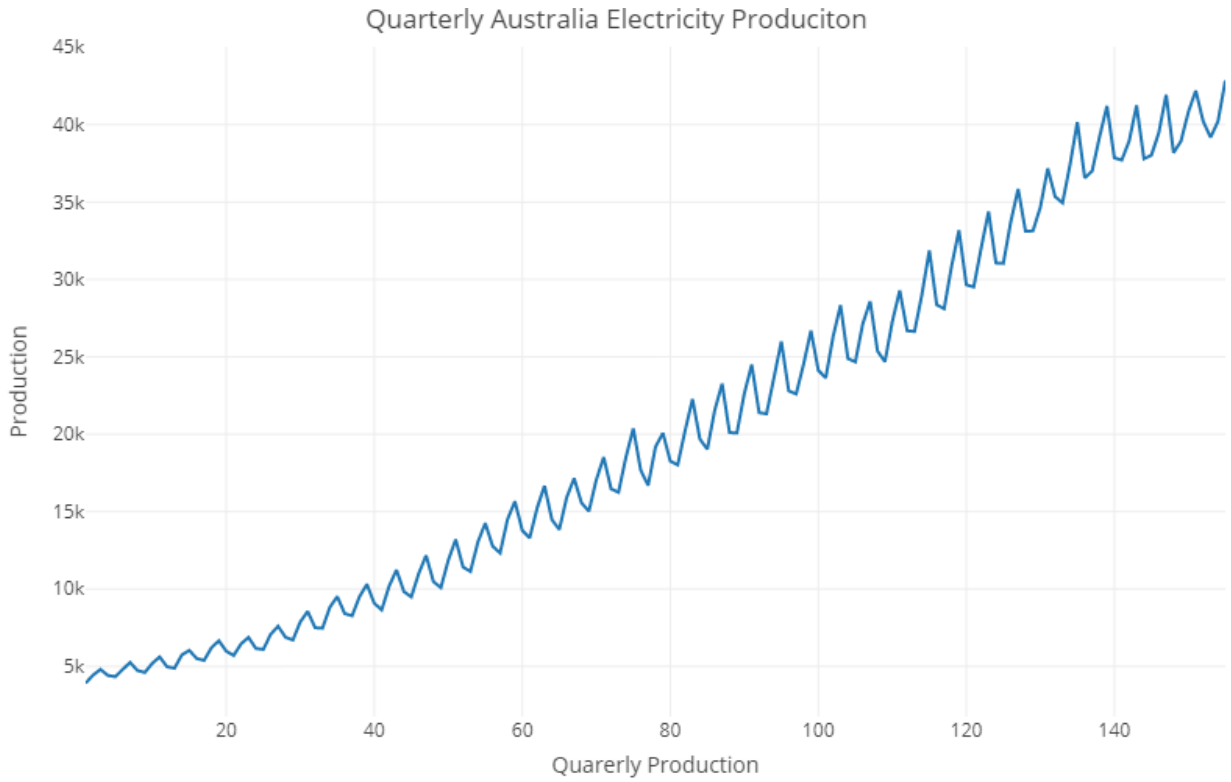
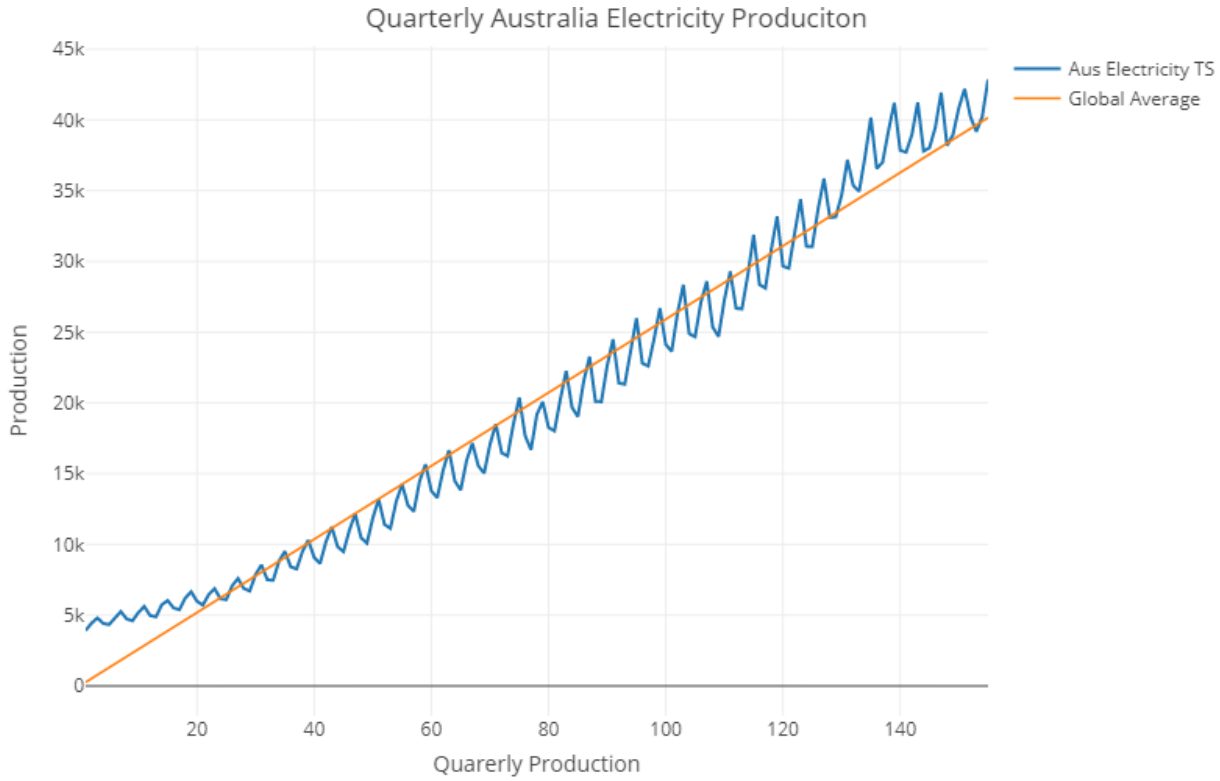
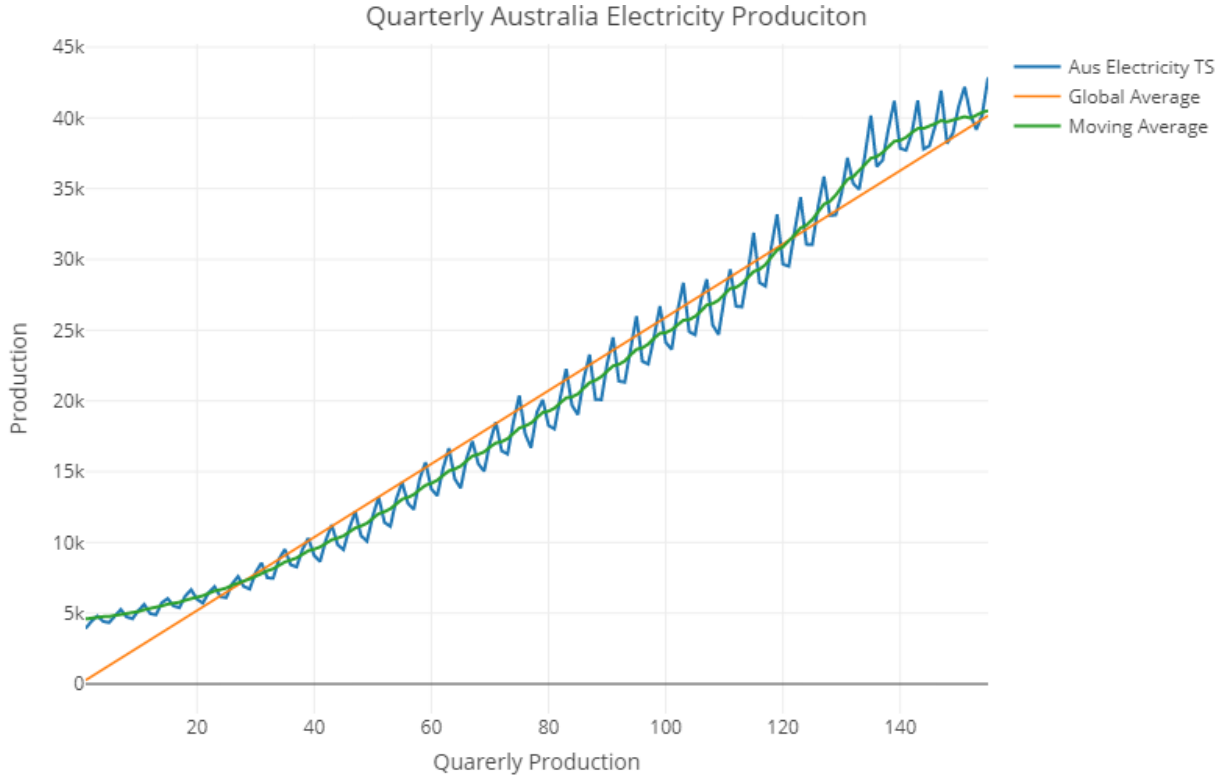


Figure 2: Australia Electric TS



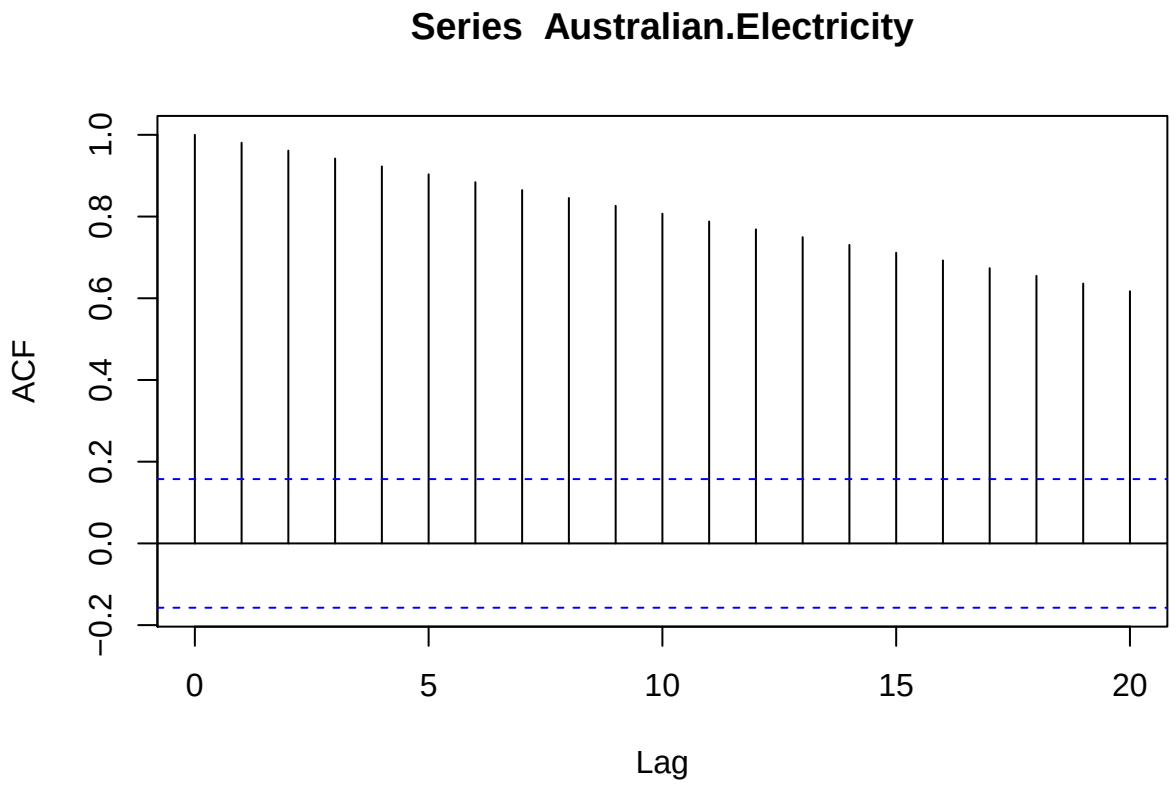
The graph above shows a Global Average for the Time Series data. It shows a strong positive correlation and it appears to be linear. This is evidenced by the trend line for the data closely following the linear function. This will be investigated further when performing the ACF.



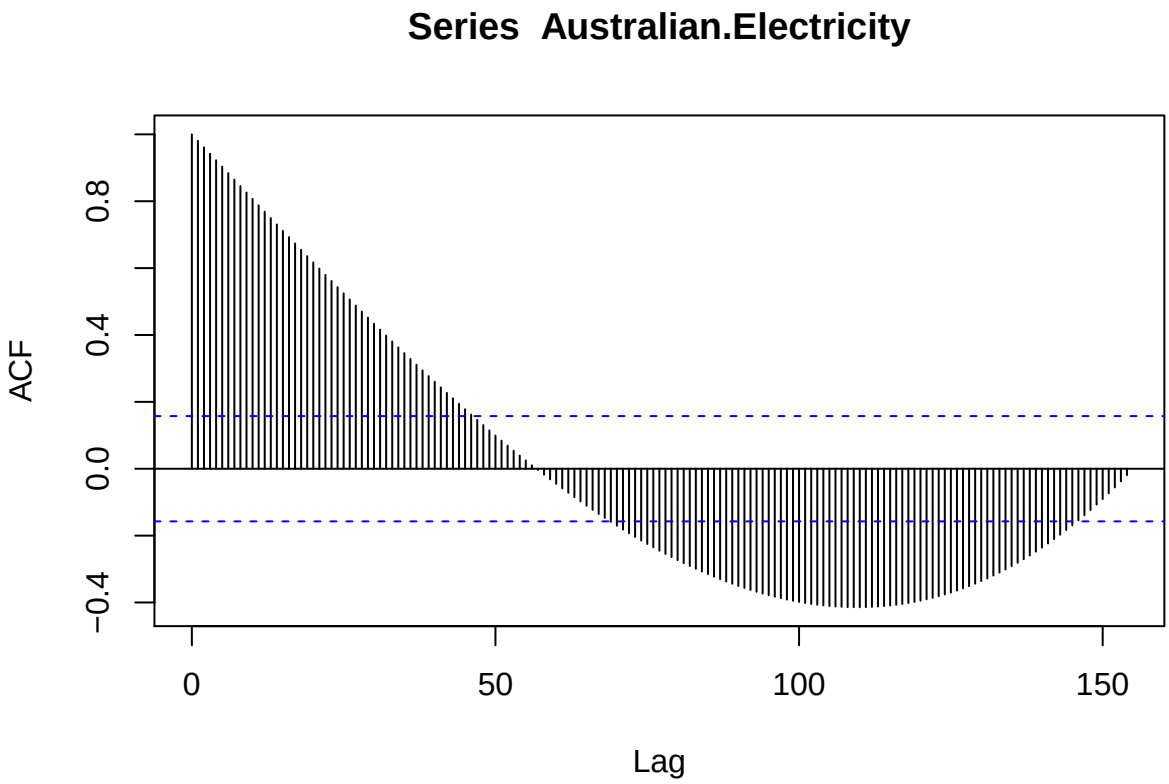
The Moving Average shown above indicates there is no cyclic component to this Time Series. But it also assists in evidencing that there is a linear trend to the data, as the Moving Average approximately follows the

Global Average. Due to the strong linear and seasonal trend, it is suspected that there is little-no irregular component. However more will be uncovered when the ACF is performed.

1.2.1 Auto Correlation for the Australian Eletricity Data



At a lag of 20 and a confidence interval of 0.95, there seems to be a strong trend, which shown by all bars being above the confidence interval.



With a lag of 200, it is very clear that the data follows the quadratic function, without drawing an approximation. Which confirms the theory from the time series that it is indeed a linear trend. The nearly perfect quadratic shape also evidences that it has little to no irregularity. It is however, difficult to see any sort of seasonality, as the linear trend is far greater strength than the seasonal component. Which shows a weakness of the ACF. Despite this, the Time Series shows clear indications of seasonality.

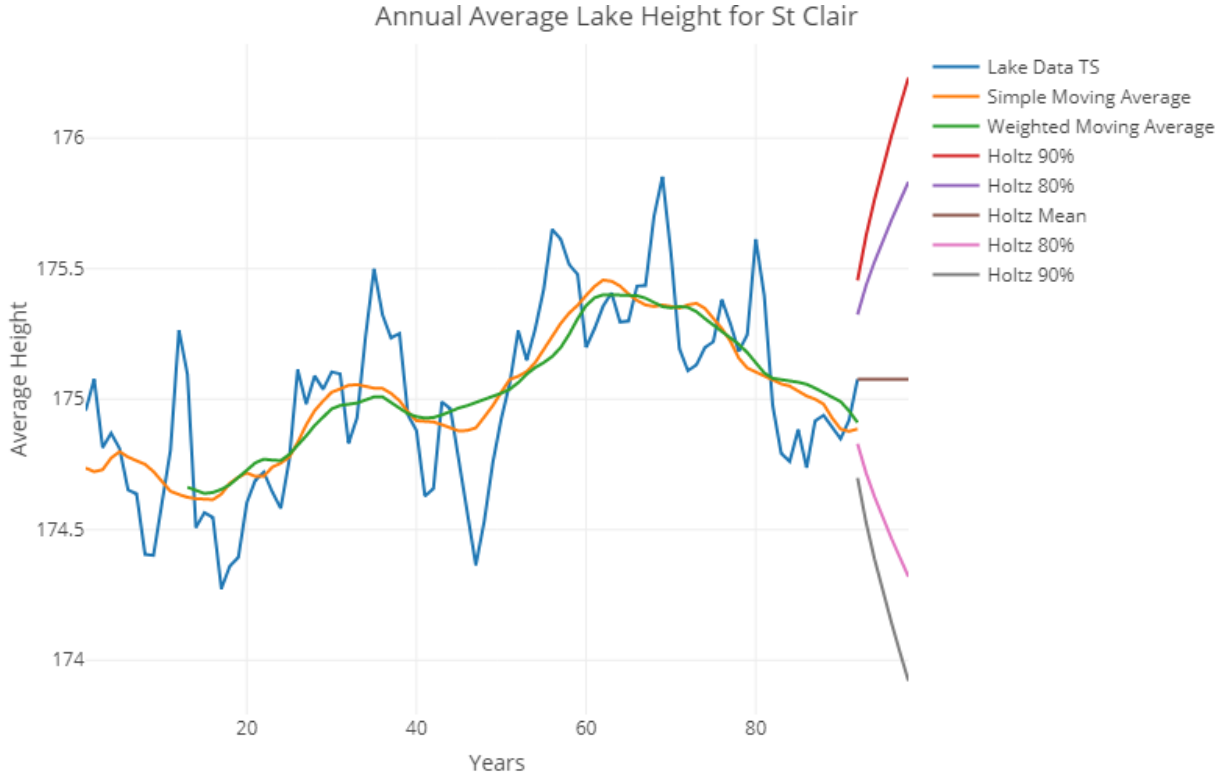
2 Task 2 - Simple Forcasting and Expotential Smoothing

2.1 St Clair Time Series Annual Average Lake Data Analysis and Forcast

Based on the initial analysis conducted in part 1, the lake data seems to be appropriate for an addictive model. The reasoning for this choice is that the maximums and minimums approximately follow a linear trend with an approximately constant variance between them. A multiplicative model is more appropriate where the seasonal oscillations are increasing with each occurrence along with the global trend, and with increasing or decreasing variance, a non-constant variance. When performing the initial analysis a "Moving Average" model was applied, which was the Simple Moving Average model, this showed some potential for cyclic behavior. However, 6 years is likely not enough to observe the expected cyclic component of this time series. So an exponential smoothing method is preferable for a relatively short time period.

The exponential smoothing that could be applied to this to make predictions are the Exponential Moving Average or Double Smoothing, specifically Holt's Method. Triple Smoothing Methods are not applicable for this time series, due to the lack of a seasonal component. A combination of both Moving Averages and Holt's Method are going to be used.

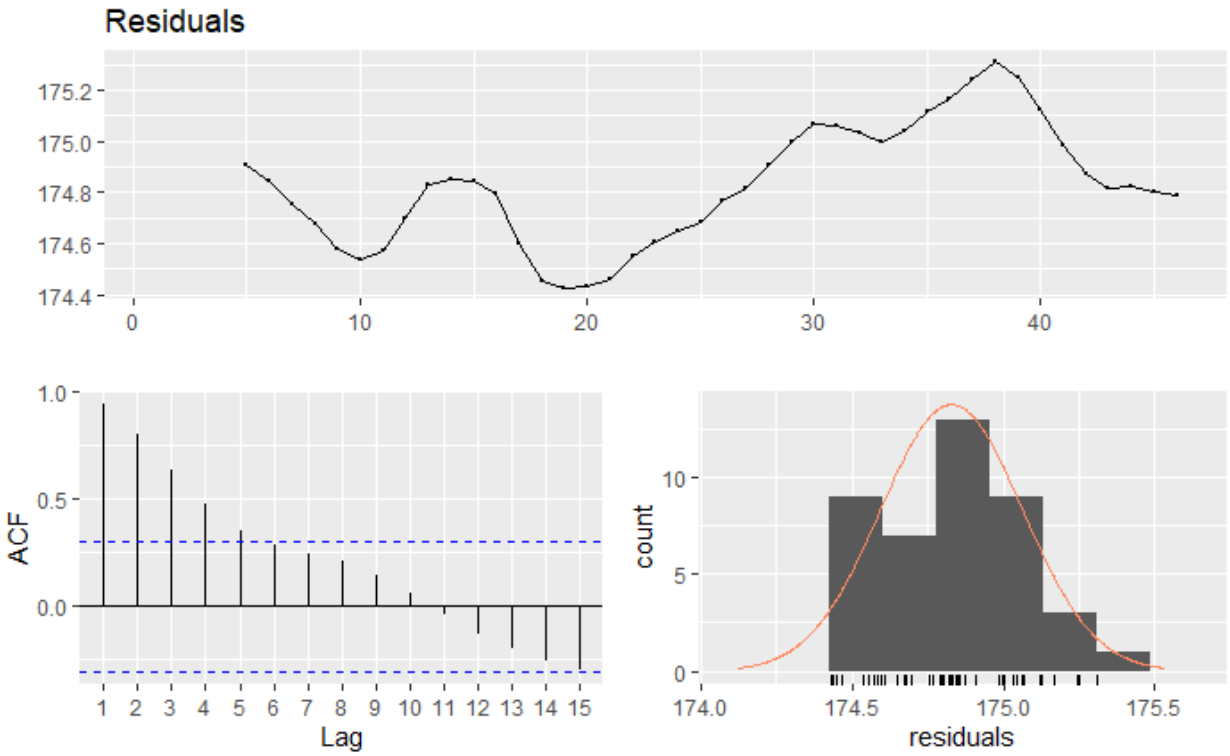
So that the analysis is not redundant, both Simple Moving Average and Weighted Moving Average (WMA) will be applied to smooth the data and analyzed as opposed to just using the Simple Moving Average. The reason for using the WMA, is it weights the (seemingly) irregular peaks, hopefully showing the cyclic component more clearly.

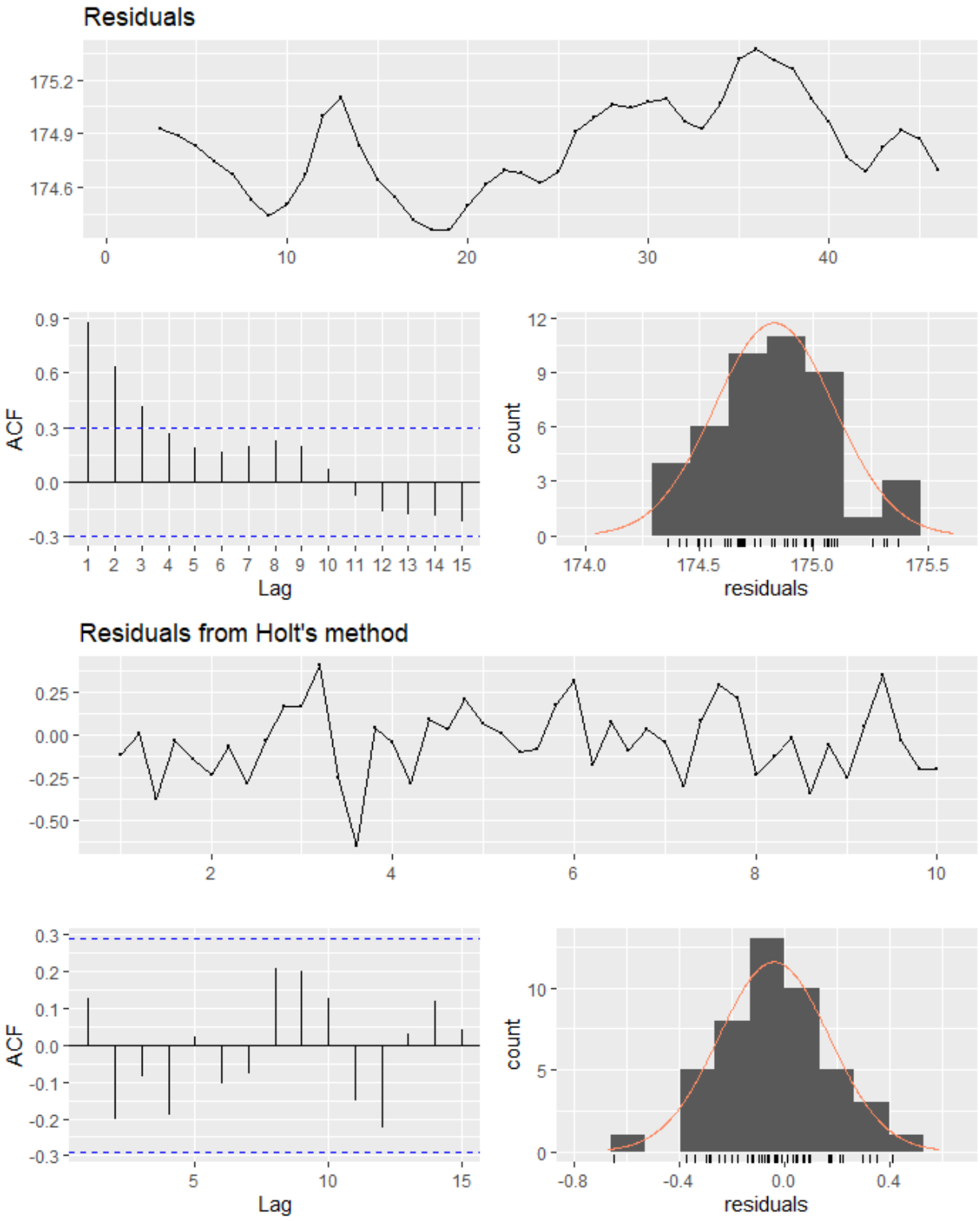


After adding the weighted moving average it it possible to see some potential for a cyclic component. However, from the current time series period, it is not possible to confirm this. More data would be required as suggested in task 1. The high irregularity is potentially causing this appearance of cyclic behavior. The forecast for the Holtz method shows a high amount of variability and a flat trend 6 years after the end of the time series. This shows a high a mount of irregularity in variance in the 80% and 90% prediction intervals See below for the accuracy of the SMA, WMA, and Holts

```
#sma
#          ME             RMSE         MAE           MPE           MAPE         ACF1         Theil's U
#Test set -0.007452381  0.2175315    0.1819206    -0.004395564  0.1040557    0.5894585    1.047623
#
#wma
#          ME             RMSE         MAE           MPE             MAPE           ACF1         Theil's U
#Test set -0.006578283  0.1101034    0.08239268    -0.003809767   0.04712543     0.3306031    0.5426181
#
#
#Holts
#               ME     RMSE         MAE          MPE             MAPE         MASE         ACF1         Theil's U
#Train -0.04206327  0.2111417    0.1642163    -0.02411758   0.09392461   0.5080132    0.1287464    NA
#Test   1.02580104  1.1298030    1.0395198    0.58473038    0.59259823   3.2158183    0.8001661    6.667926
```

While SMA and WMA can sufficiently smooth the plot. The methods do not have a great long range forecast for the time series. By looking at the WMA, it is possible to see slightly ahead of time where there might be an increase or decrease. However in this case, the SMA seems to better, slightly predict increases and decreases. To confirm this theory, looking at the results from the accuracy tests should give some indication. The results of the MAPE, show that the SMA has a higher results than the WMA, which could indicate that the SMA is more accurate at forecasting. And by looking the plot, it seems that this is the case. The MAE however from the SMA, seems to have a larger value than the WMA, which would indicate a larger difference between the predictions and the actual values. So the WMA is better at predicting the trend than the SMA. With this in mind, it is possible to use the two smoothing methods in conjunction to make slightly better estimations about what the time series might do. Although, these correlations may be caused by noise, however there is definitely some trend, as will be shown in the ACF for the residuals.






```
#   Ljung-Box test

#data:  Residuals from Holt's method
#Q* = 11.062, df = 5, p-value = 0.05017
```

```
#Model df: 4.   Total lags used: 9
```

As can be seen in the SMA residual analysis for the ACF, there is definitely some amount of trend as indicated by the autocorrelation. The residuals appear to show a normal distribution. And with the detrended, a increasing amount of variance. With the WMA, the story is similar, however the ACF shows sightly less trend and the residuals represent something that is better normally distributed. The result from this show correlated residuals and are thus are dependent. The holts residuals show what seems like uncorrelated data, with a normally distributed residual and a detrended graph with a variability that becomes constant. Due to the p-value being greater than 0.05, there is insufficient evidence to conclude that it is not normally distribution.

2.2 Australian Energy Production Time Series Quarterly Energy Production Analysis and Forcast

From the initial analysis in task one, it was identified that this time series has strong trend and strong seasonality. This makes the choice for a model easy. The model of choice is the Holts-Winter model, as this works well with trend and seasonality. The increasing variance with time also indicates a multiplicative model is more useful here. Seasonal Naive is another possibility, but it would not model the trend well. However a combination could be used to make accurate forecasts. Both will not be used in this case.

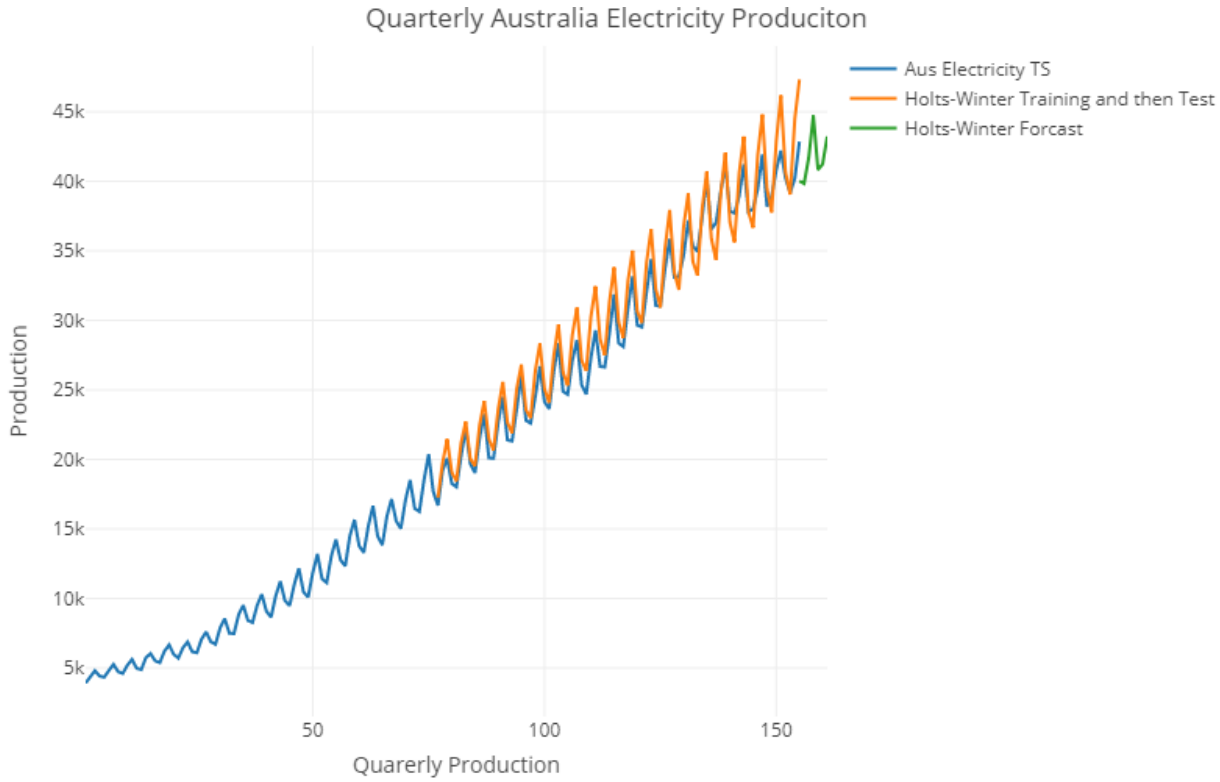
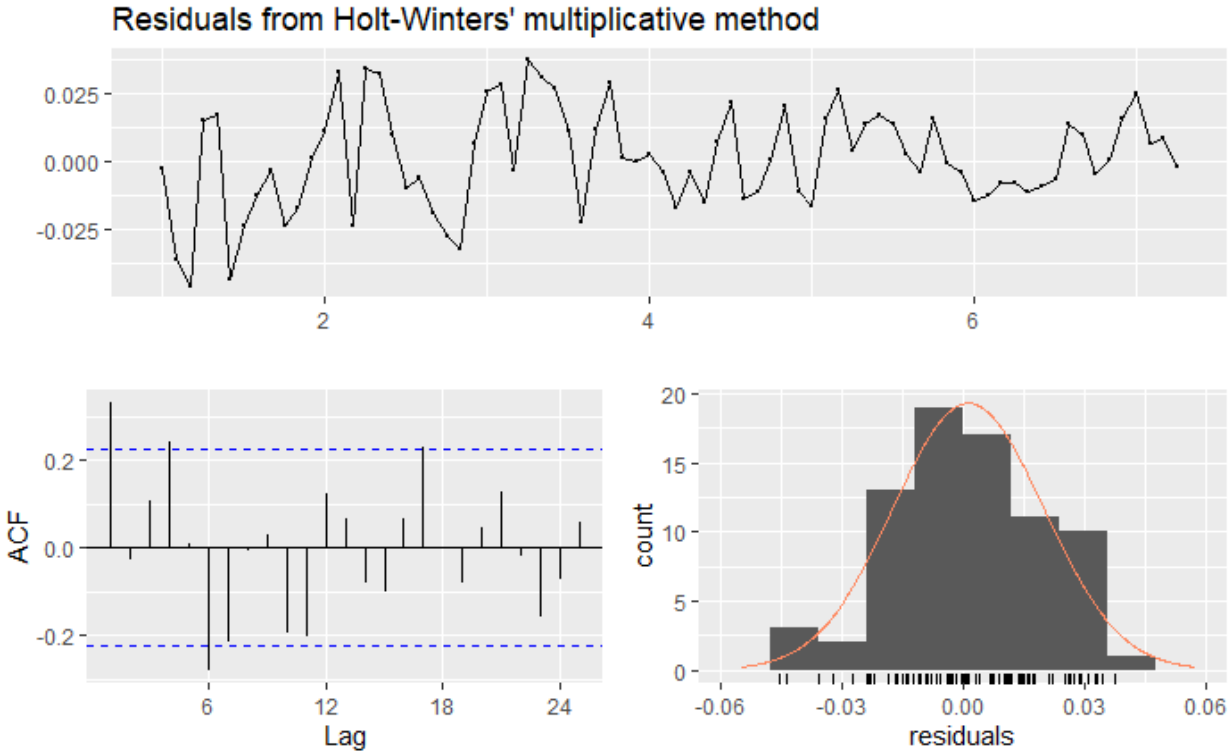


Figure 3: Australia Electric TS Moving Average

As can be seen, the Holts-Winter method performs well in predicting the next 6 time periods. The trained data that has been plotted along side the test data, which shows the accuracy of the Holts-Winter, visually method over a long period of time. It is to be pointed out that the further, the trained plot tries to forecast, the less accurate it is. And the greater the variance. This is due to the strong trend component of the time series and the use of a multiplicative model. The reason the accuracy breaks down, is that the time series's trend decreases at a point, but the Holts-Winter Model still has the affects from the strong trend that existed previously.

```
#wma
#      ME            RMSE           MAE             MPE             MAPE            ACF1          Theil's U
#      -1037.428       1141.302       1037.428       -4.680266       4.680266       0.2048267     0.5450349
```

This shows very high RMSE and MAE values, showing a very good prediction of the test data. This is evidenced by looking at both the visual plot prediction of the data along with these results.



From the ACF, it is clear that there is seasonality, as the bar plot is above the error, dips below and the dips above again. The residual graph shows what seems to be skewed slightly negatively. The detrended plot shows a high amount of seasonality that decreases towards the end.

```
#   Ljung-Box test

#data:  Residuals from Holt-Winters' multiplicative method
#Q* = 82.393, df = 32, p-value = 2.53e-06

#Model df: 16.   Total lags used: 48
```

The Ljung-Box test shows many degrees of freedom, with a very small p-value that is less than 0.05. Therefore we can reject that it is normally distributed and is independent. Therefore we will perform a t-test.

```
#One Sample t-test

#data:  residuals(training.hw)
#t = 0.54907, df = 75, p-value = 0.5846
#alternative hypothesis: true mean is not equal to 0
#95 percent confidence interval:
# -0.003085536  0.005433600
#sample estimates:
#  mean of x
#0.001174032
```

As the p-value is greater than 0.05, there is insufficient evidence that the mean is not zero.

3 Appendix: Source Code

3.1 Task 1

3.1.1 Aus Electric

The code for Aus Electric is nearly identical

```
library(plotly)

#loading in ausElectric
#ausElectric <- read.csv("aus-electricity.csv")
#ausElectric.ts <- ts(ausElectric)

lake.data <- DAAG::greatLakes[,4]
lake.ts <- ts(lake.data)
x <- c(1:length(lake.ts))
data <- data.frame(lake.ts)

linearLine = lm(lake.data~x)

data.fmt = list(color=rgb(0.8,0.8,0.8,0.8), width=4)
line.fmt = list(dash="solid", width = 1.5, color=NULL)

curve.fmt = list(dash="solid", width = 1.5, color=NULL)

movingAvg.fmt = ksmooth(x, lake.data, "box", 16, x.points=x)

fig <- plot_ly(data, x = ~x, y = ~lake.ts, name = 'Lake Data TS', type = 'scatter',
               mode = 'lines')
fig <- fig %>% layout(title = "Annual Average Lake Height for St Clair",
                      xaxis = list(title = "Years"),
                      yaxis = list(title = "Average Height"))
fig <- add_lines(fig, x=~x, y=fitted(linearLine), line=line.fmt,
                 name="Global Average")
fig <- add_lines(fig, x=movingAvg.fmt$x, y=movingAvg.fmt$y, line=movingAvg.fmt,
                 name="Moving Average")
fig
```

3.1.2 Lakedata AFC

```
afc.lakedata <- plot(acf(data, lag.max = 10),ci=0.95)
```

3.2 Task 2

3.2.1 Lake

```
library(plotly)
library(forecast)

lake.data <- DAAG::greatLakes[,4]
lake.ts <- ts(lake.data,frequency=5)
x <- c(1:length(lake.ts))
data <- data.frame(lake.ts)

trainingX <- c(1:46)
testX <- c(47:92)

training.data <- window(lake.ts,end=c(1,46))
test.data <- window(lake.ts,end=c(47,92))

data.fmt = list(color=rgb(0.8,0.8,0.8,0.8), width=4)
curve.fmt = list(dash="solid", width = 1.5, color=NULL)

lake.sma.all <- TTR::SMA(lake.ts,n=5)
lake.wma.all <- TTR::WMA(lake.ts,n=3)
lake.holt.all <- holt(lake.ts, h=25, initial="simple")

lake.sma.train <- TTR::SMA(training.data,n=5)
lake.wma.train <- TTR::WMA(training.data,n=3)
lake.holt.train <- holt(training.data, h=25, initial="simple")

accuracy(lake.sma.train,training.data)
accuracy(lake.wma.train,training.data)
accuracy(lake.holt.train,test.data)


checkresiduals(lake.sma.train)
checkresiduals(lake.wma.train)
checkresiduals(lake.holt.train)



#names(lake.holt)
#predictX <- c(length(lake.ts):lake.holt$upper[,0])


predictX <- c(length(lake.ts) : (length(lake.ts) + 6))


movingAvg.fmt = ksmooth(x, lake.data, "box", 16, x.points=x)
weightmovingAvg.fmt = ksmooth(x, lake.wma.all, "box", 20, x.points=x)
holtupperupper.fmt = ksmooth(predictX, lake.holt.all$upper[,2], "box", 1, x.points=predictX)
holtupper.fmt = ksmooth(predictX, lake.holt.all$upper[,1], "box", 1, x.points=predictX)
holtmean.fmt = ksmooth(predictX, lake.holt.all$mean, "box", 1, x.points=predictX)
holtlower.fmt = ksmooth(predictX, lake.holt.all$lower[,1], "box", 1, x.points=predictX)
holtlowerlower.fmt = ksmooth(predictX, lake.holt.all$lower[,2], "box", 1, x.points=predictX)
```

```
fig <- plot_ly(data, x = ~x, y = ~lake.ts, name = 'Lake Data TS', type = 'scatter', mode = 'lines')
fig <- fig %>% layout(title = "Annual Average Lake Height for St Clair", xaxis = list(title = "Years"),
fig <- add_lines(fig, x=movingAvg.fmt$x, y=movingAvg.fmt$y, line=movingAvg.fmt, name="Simple Moving Ave

fig <- add_lines(fig, x=weightmovingAvg.fmt$x, y=weightmovingAvg.fmt$y, line=weightmovingAvg.fmt, name=

fig <- add_lines(fig, x=holtupperupper.fmt$x, y=holtupperupper.fmt$y, line=holtupperupper.fmt, name="Ho
fig <- add_lines(fig, x=holtupper.fmt$x, y=holtupper.fmt$y, line=holtupper.fmt, name="Holtz 80%")
fig <- add_lines(fig, x=holtmean.fmt$x, y=holtmean.fmt$y, line=holtmean.fmt, name="Holtz Mean")
fig <- add_lines(fig, x=holtlower.fmt$x, y=holtlower.fmt$y, line=holtlower.fmt, name="Holtz 80%")
fig <- add_lines(fig, x=holtlowerlower.fmt$x, y=holtlowerlower.fmt$y, line=holtlowerlower.fmt, name="Ho
```

3.2.2 Aus electric

```
library(plotly)
ausElectric <- read.csv("aus-electricity.csv")
ausElectric.ts <- ts(ausElectric,frequency=12)
x <- c(1:length(ausElectric.ts[,'Quarter']))
trainingX <- c(1:76)
testX <- c(77:155)

data.fmt = list(color=rgb(0.8,0.8,0.8,0.8), width=4)

training.data <- window(ausElectric.ts[,'Production'],end=c(1,76))
test.data <- window(ausElectric.ts[,'Production'],end=c(77,155))

training.hw <- hw(training.data,seasonal = "multiplicative")

#ausElectric.snaive

#test for accuracy
print("-------")
accuracy(training.hw$mean,test.data)
print("-------")

checkresiduals(training.hw,48)

t.test(residuals(training.hw))

testForcast = forecast(training.data,h=100)
finalForcast = forecast(test.data,h=100)

predictX.hw.forcast <- c(155 : (155 + 6))


hwForcast.test.fmt = ksmooth(testX, testForcast$mean, "box", 1, x.points=testX)
hwForcast.forcast.fmt = ksmooth(predictX.hw.forcast, finalForcast$mean, "box", 1, x.points=predictX.hw.

#fix the h values for training forcast
#fix the seasonal niave to be at the appropriate location
#graph the residuals
```

```
fig <- plot_ly(ausElectric, x = ~x, y = ~ausElectric.ts[,'Production'], name = 'Aus Electricity TS', ty
fig <- fig %>% layout(title = "Quarterly Australia Electricity Produciton", xaxis = list(title = "Quare
fig <- add_lines(fig, x=hwForcast.test.fmt$x, y=hwForcast.test.fmt$y, line=hwForcast.test.fmt, name="Ho
fig <- add_lines(fig, x=hwForcast.forcast.fmt$x, y=hwForcast.forcast.fmt$y, line=hwForcast.forcast.fmt,
```