An Analysis of Egg Deposits in the late 1900s of Freshwater Fish.

Written by Maxwell Repin

S3822965

22/09/2021

Introduction

Human beings have been trying to understand the animal kingdom for centuries. This dynamic and complex world is constantly being monitored to best understand how humans, and other species, influence ecosystems around the world. The study of fish egg deposits is one such area of research that has taken many data scientists' interests as levels of reproduction usually directly correlate to a species wellbeing in a certain environment. Fish will deposit differing amounts of eggs depending on many factors such as levels of pollution in the water or amount of fish present in a certain environment (e.g. a certain lake, stream, etc). Therefore, it is easy to summarise the overall prosperity of a population through monitoring the levels of eggs produced.

This quantitative investigation aims to identify and discuss the trends happening between 1981 and 1996 for the yearly amount of egg depositions of Lake Huron Bloaters (Coregonus hoyi) and forecast their numbers in the next 5 years after 1996.

This report utilises RStudio with the 'TSA', 'forecast, 'lmtest' and 'tseries' packages and a provided dataset for bloater egg depositions (in millions) that contains data from the years 1981 and 1996.

Results

We began by changing the data provided from raw data into a time series and creating a plot to explore any visible trends or behaviour.

Figure 1

Figure 1 reveals a time series plot with a slight upwards trend, no seasonality and a change point in 1990. There is also clear auto regressive behaviour (AR) and changing variance. This plot is non-stationary.

Due to the changing variance this series must be transformed before being made stationary to allow us to fit any models.

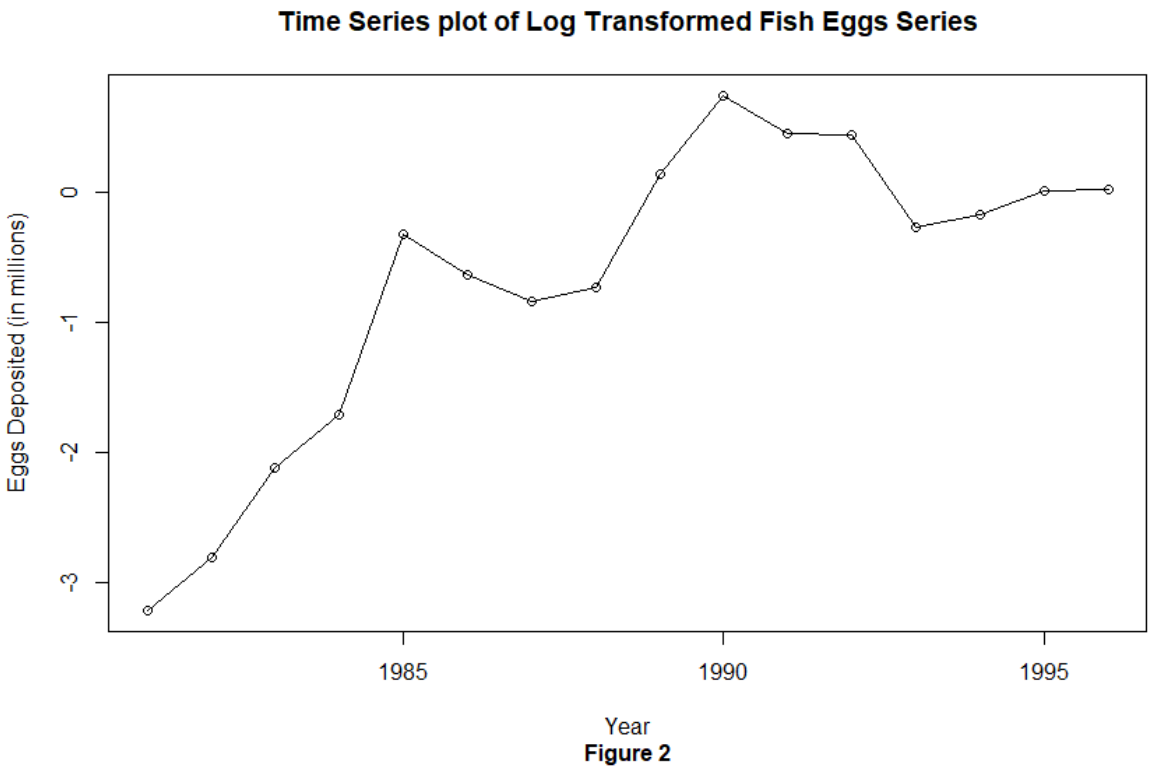

Figure 2

Figure 2 now depicts this series after undergoing a log transformation and now visually has no changing variance. Now moving on to making this plot stationary.

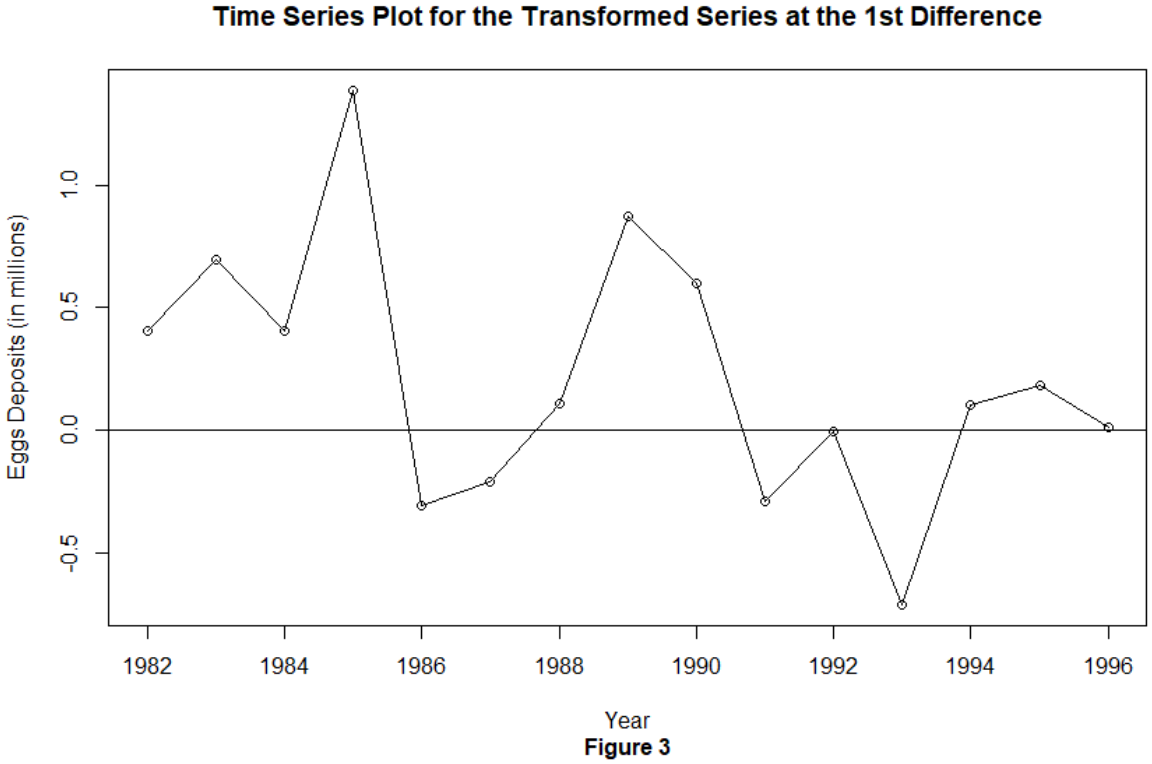

Figure 3

Here in figure 3, we can see that the plot for the series at the first differencing has a slight downward trend but is difficult to tell from just a visual analysis and we are unsure if it is stationary or not. Therefore, we must explore the tests for stationarity to come to conclusion.

```
> adf.test(eggsTSDiff, alternative = c("stationary"))

        Augmented Dickey-Fuller Test

data:  eggsTSDiff
Dickey-Fuller = -4.0304, Lag order = 2, p-value = 0.02219
alternative hypothesis: stationary

> pp.test(eggsTSDiff, alternative = c("stationary"))

        Phillips-Perron Unit Root Test

data:  eggsTSDiff
Dickey-Fuller Z(alpha) = -13.564, Truncation lag parameter = 2, p-value = 0.2364
alternative hypothesis: stationary
```

Figure 4

Figure 4 depicts and ADF test with a p-value that is **less** that the alpha level of 0.05, therefore we have enough evidence to reject the null hypothesis and conclude that this series is now stationary from just the ADF test. However, the PP test has an alpha level **greater** than 0.05. Therefore, according to the PP test, we have insignificant evidence to reject the null hypothesis that the series is non-stationary.

After conducting some research on the Phillips-Perron test, it was discovered that this test does not perform well on finite samples when compared to the Augmented Dickey-Fuller test. After taking this into account, it was decided to conclude that this series was non-stationary still as the p-value for the PP test was not close to the alpha level and visually the plot (Figure 3) does not look stationary.

Since we have concluded this 1st differenced series is non-stationary we must difference again.

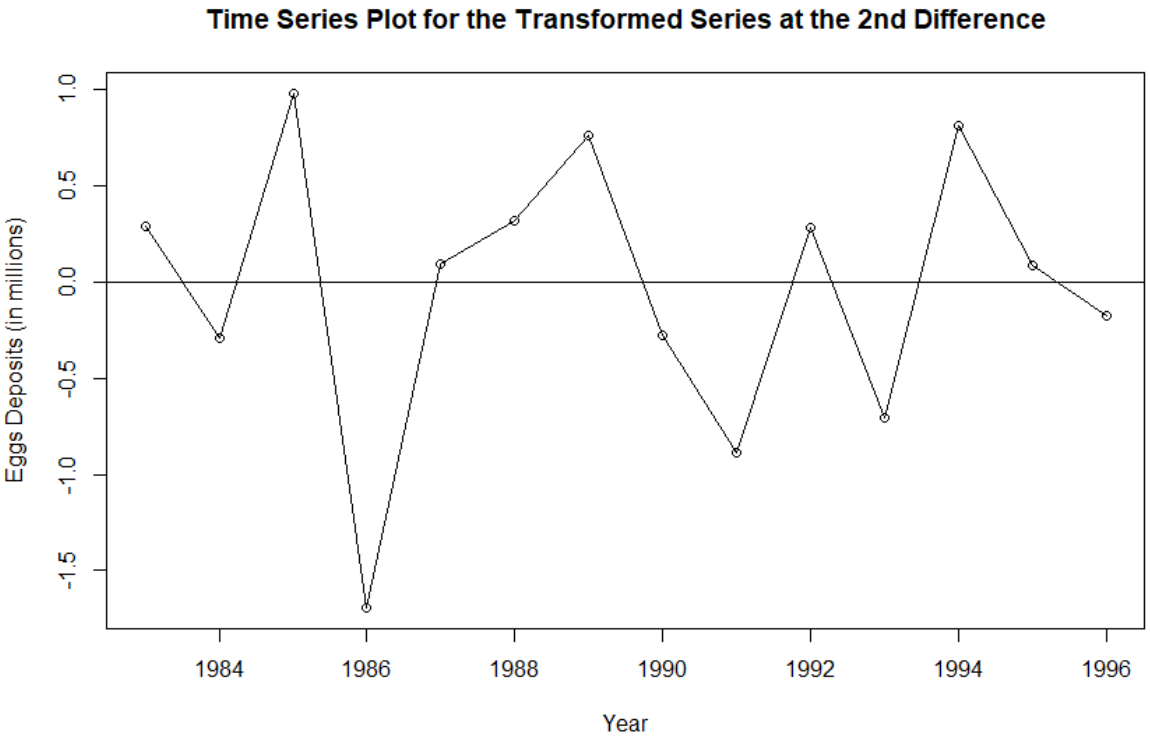

Figure 5

This time series plot (Figure 5) reveals a plot that looks stationary at first glance with no major changes in variance. To further confirm this, we will again consider the tests for stationarity.

```
> adf.test(eggsTSDiff2, alternative = c("stationary"))

        Augmented Dickey-Fuller Test

data:  eggsTSDiff2
Dickey-Fuller = -3.8026, Lag order = 2, p-value = 0.03553
alternative hypothesis: stationary

> pp.test(eggsTSDiff2, alternative = c('stationary'))

        Phillips-Perron Unit Root Test

data:  eggsTSDiff2
Dickey-Fuller Z(alpha) = -16.559, Truncation lag parameter = 2, p-value = 0.07916
alternative hypothesis: stationary
```

Figure 6

Here in Figure 6, the ADF test has a p-value that is **less** than the alpha level which means there is significant evidence to reject the null hypothesis that the series is non-stationary and conclude that this series is stationary. The PP test has a p-value that is very slightly over the alpha level (0.05) and, after considering that the PP test is not as accurate on finite samples when compared to the ADF test, we can conclude that we will reject the null hypothesis for this test as well. Therefore, this series is now stationary after performing a log transformation and taking it to the 2nd difference.

Now we must check for any suitable models to fit to the series.

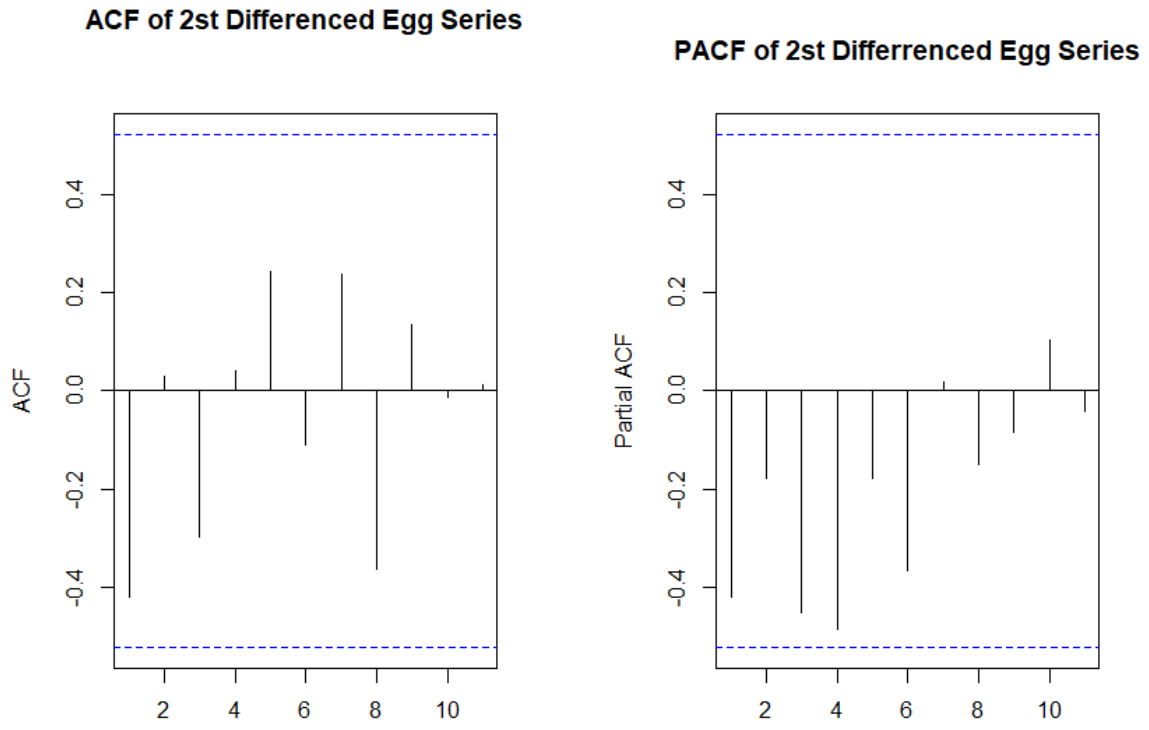

Figure 7

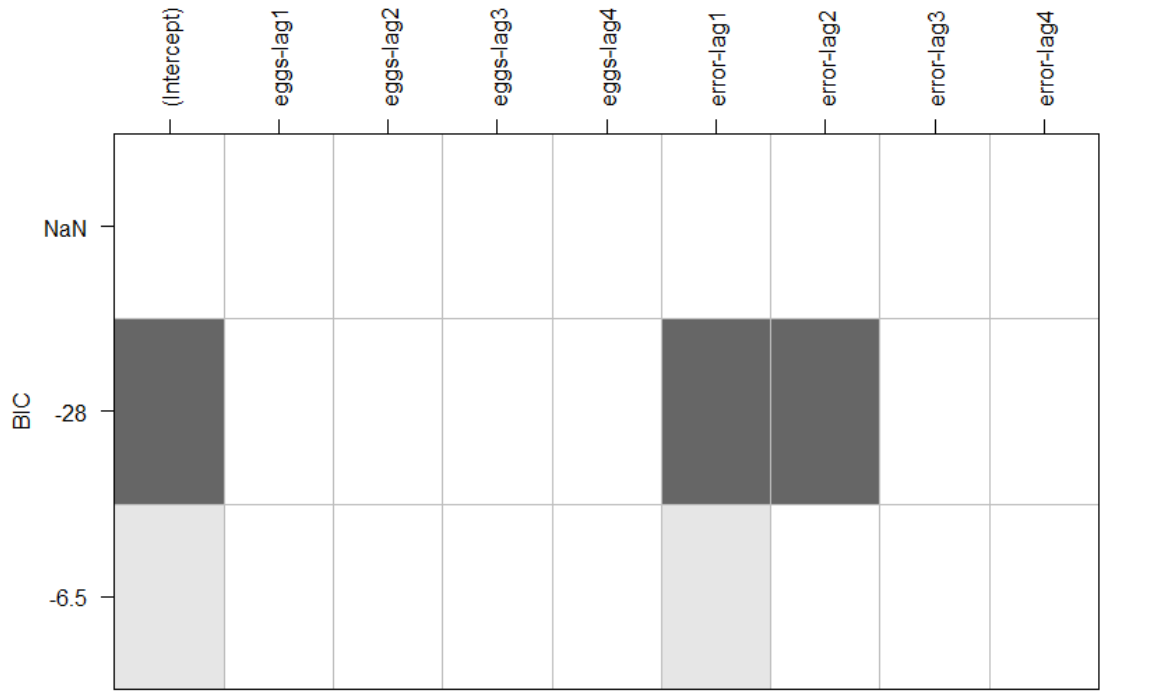

Figure 8

```
> eacf(eggsTSDiff2, ar.max = 3, ma.max = 3)
AR/MA
  0 1 2 3
0 o o o o
1 o o o o
2 o o o o
3 o o o o
```

Figure 9

Figure 7 depicts both the ACF and PACF plots that do not have any significant lags therefore nothing to conclude out of them. Figure 8 however reveals that ARIMA(0,2,1) and ARIMA(0,2,2) would be good models to fit to this data. Furthermore, Figure 9 reveals an EACF function with no 'x' in the outputs. From this output, we will take the 3 top left models (highlighted in blue) in this vertex which are ARIMA(0,2,0), ARIMA(0,2,1) and ARIMA (1,2,1).

Our set of possible models is currently:

ARIMA(0,2,1), ARIMA(0,2,2), ARIMA(0,2,0)and ARIMA (1,2,1).

Giving us 4 possible models to fit to the data and check which will be the best model for forecasting.

Model Fitting

For each model, we fitted them to our log transformed, time series data and then performed z tests on the co-efficients to compute their significance.

```
> model.021 <- arima(eggsTSLog, order = c(0,2,1), method = 'ML')
> coeftest(model.021)

z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ma1 -0.78931    0.19625  -4.022 5.77e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.022 <- arima(eggsTSLog, order = c(0,2,2), method = 'ML')
> coeftest(model.022)

z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ma1 -0.728447   0.267762 -2.7205 0.006518 **
ma2 -0.082007   0.250783 -0.3270 0.743664
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.121 <- arima(eggsTSLog, order = c(1,2,1), method = 'ML')
> coeftest(model.121)

z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1  0.10190    0.32891  0.3098 0.7567118
ma1 -0.83177    0.24656 -3.3735 0.0007422 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10

```
> model.020 <- arima(eggsTSLog, order = c(0,2,0), method = 'ML')
> coeftest(model.020)
Error in dimnames(x) <- dn :
  length of 'dimnames' [2] not equal to array extent
> |
```

Figure 11

The outputs for the co-efficient tests (Figure 10) all display that the only significant co-efficient for every test is MA(1). Figure 11 displays the error message that outputs when using ARIMA(0,2,0) as the model does not cover the extent of the series and will therefore not be used.

Now we will sort the models according from most relevant to least using a custom function "sort.score" (created by Mr Yong Kai Wong) (see Appendix A for code).

```
> sort.score(AIC(model.021, model.022, model.121), score = "aic")
           df      AIC
model.021   2 28.09899
model.022   3 29.99404
model.121   3 29.99927
> sort.score(BIC(model.021, model.022, model.121), score = "bic")
           df      BIC
model.021   2 29.37710
model.022   3 31.91121
model.121   3 31.91644
```

Figure 12

As seen in the outputs for both the BIC and AIC scores (Figure 12), ARIMA(0,2,1) is the best model with ARIMA(0,2,2) being noted as a backup.

Moving forward, we will check the residuals of ARIMA(0,2,1) and compare them with the overfitted models of ARIMA(0,2,2) and ARIMA(1,2,1). We want to make certain that the residuals are normally distributed for this model to be a good fit for forecasting our series.

ARIMA(0,2,1)

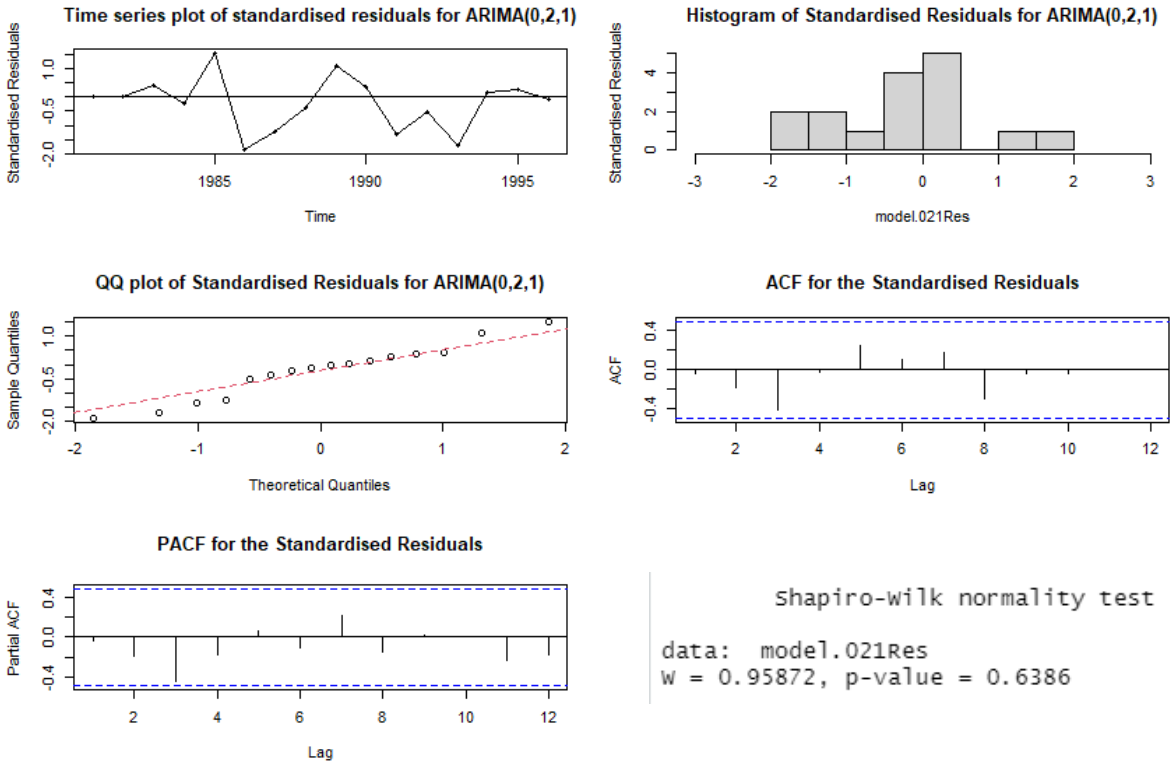

Figure 13

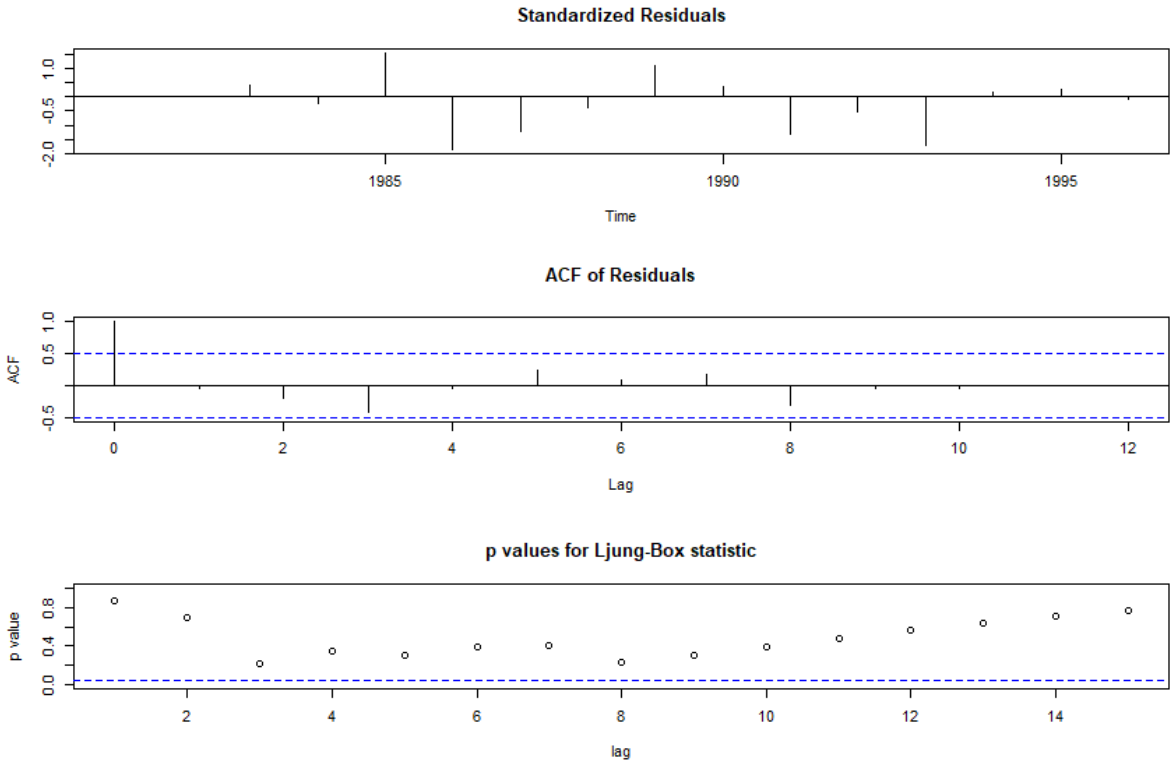

Figure 14

The residuals for ARIMA(0,2,1) (Figures 13) are normally distributed as shown by the Shapiro-Wilk test, with a p-value that is greater than the alpha level, so we do not have enough evidence to reject the null-hypothesis that they are normally distributed. In addition, there are no significant lags in the ACF or PACF plots, the time series plot looks stationary and the values in the QQ plot all follow the reference line closely with only a few straying at the top and bottom. The p-values in the Ljung-Box plot (Figure 14) are all above the reference line.

However, in the histogram, while all the values do exist between -3 and 3, the bars are relatively symmetrical but have many gaps between some of the values. Furthermore, there is 1 significant lag in the ACF plot for the residuals (Figure 14).

Therefore, the residuals for this model, while being normally distributed and a relatively good fit for this model, there is plenty of room for improvement so we shall explore the residuals for ARIMA(1,2,1) and ARIMA(0,2,2).

ARIMA(1,2,1)

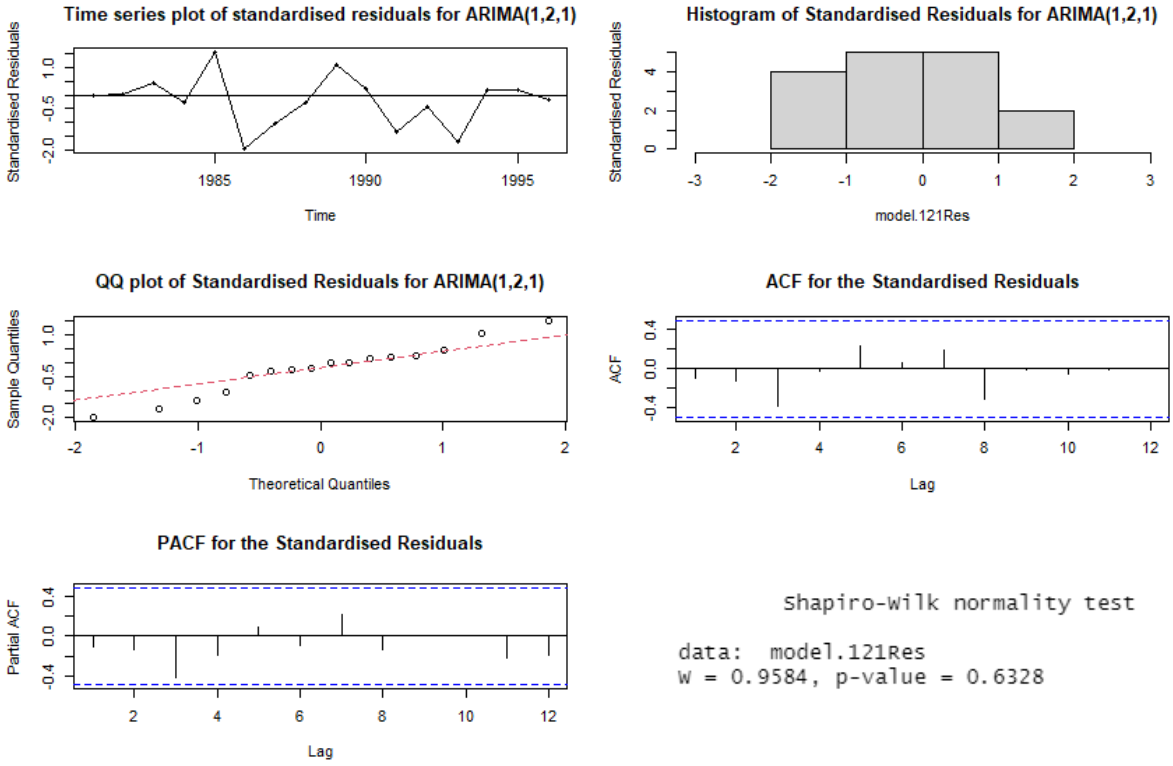

Figure 15

For ARIMA(1,2,1), we have a more normally distributed set of residuals as seen by Shapiro-Wilk test with a p-value **greater** than the alpha level which leads us to the conclusion that the residuals are normally distributed as we do not have enough evidence to reject the null hypothesis. The stationary time series plot and the lack of significant lags in the ACF or PACF plots are identical to our best model ARIMA(0,2,1). However, the QQ plot has values that stray further from the reference line in comparison to our original best model ARIMA(0,2,1) but the histogram is more symmetrical.

Therefore, while there is some improvement in the histogram, there are is more deviance in the QQ plot from the reference line.

ARIMA(0,2,2)

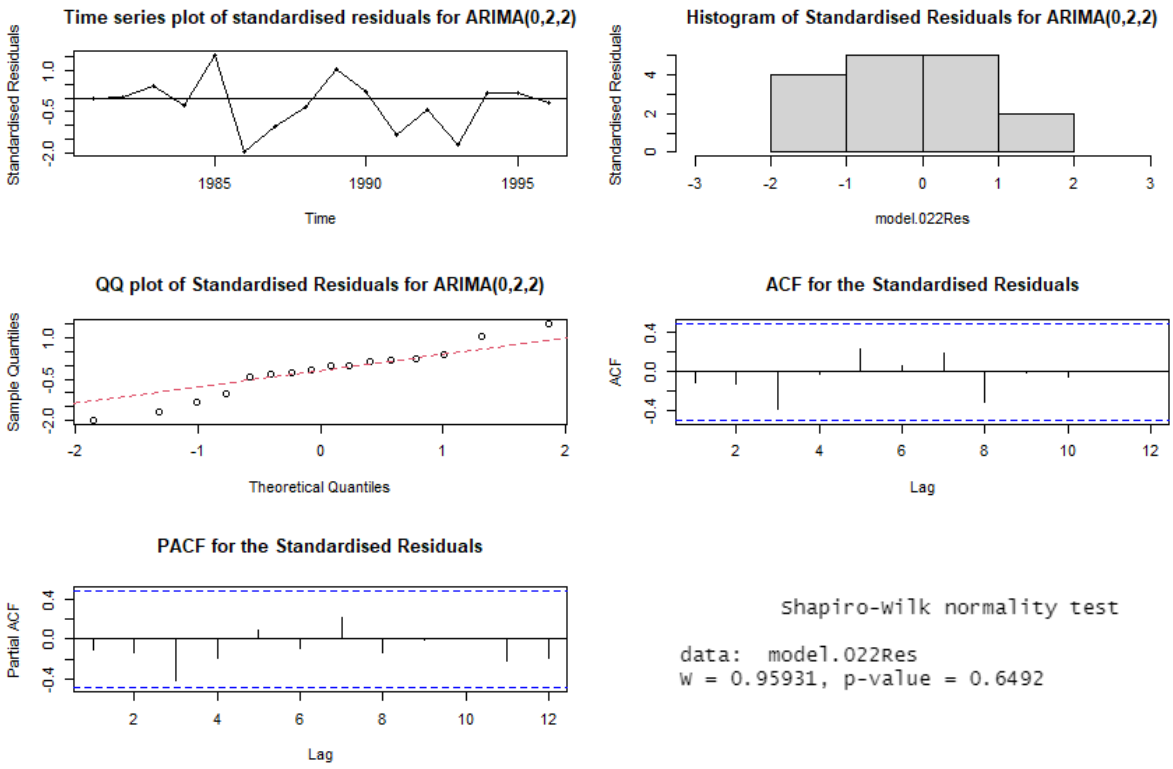

Figure 16

The residuals for ARIMA(0,2,2) (Figure 16) show almost identical outputs in comparison to ARIMA(1,2,1). The only difference is a very slightly higher p-value in the Shapiro-Wilk test. This would mean that the residuals are more normally distributed than in Figure 15 but on a small scale.

In conclusion, the best model to fit to our series would be ARIMA(0,2,1) as it may have a semi-symmetrical histogram, the QQ plots lead us to believe that the residuals for ARIMA(0,2,1) are **more** normally distributed than either ARIMA(0,2,2) or ARIMA(1,2,1) in comparison.

Therefore, we will now commence with fitting ARIMA(0,2,1) to our data and forecasting the next 5 years of data.

Forecasting

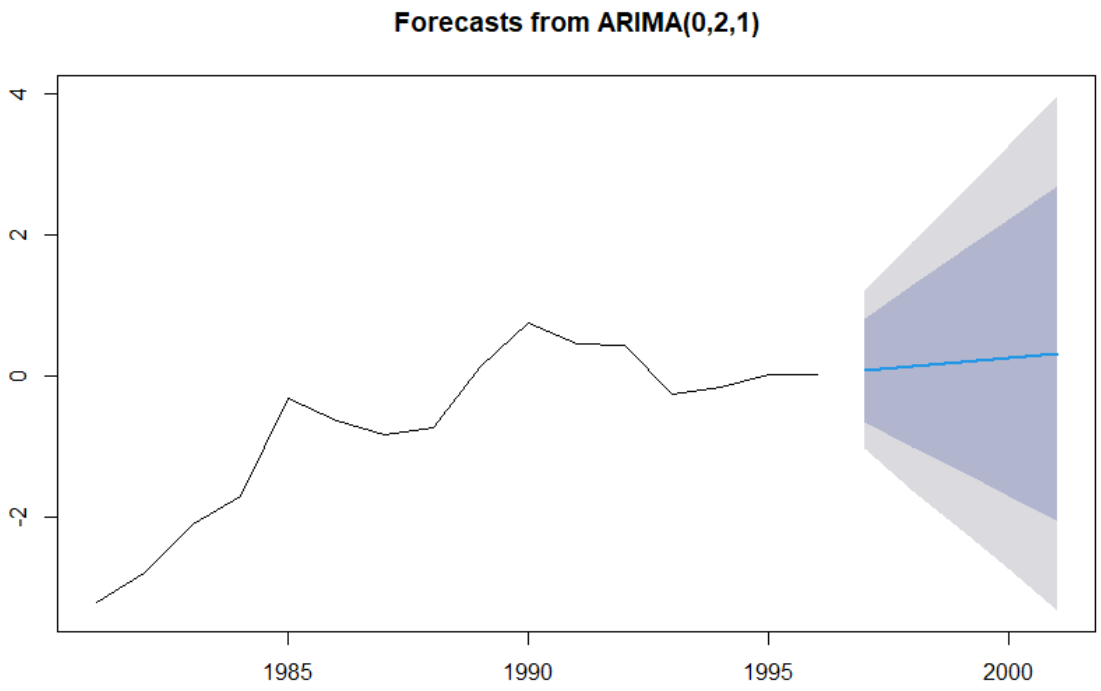

Figure 17

Figure 17 depicts the result of our model specification, testing and analysis. It reveals a visual representation of the forecast of the amount of Bloater egg depositions (in millions) for the years 1997-2001 (the 5 years following 1996). As you can see from the plot, there will be a continuous, slow upward trend for egg deposits in the next 5 years.

Conclusion

As noted in the introduction, there are many reasons to study the reproduction rates of a certain population of animals as it tends to give insight into a certain environment or ecosystem. After some research, the steep drop in egg deposits in the early 1990s was due to 'predation from other adult bloaters' and 'general higher survival rates of female bloaters which lead to a female predominance in their population' (Jeffrey S. Schaeffer, 2004). According to our forecasting, those levels were predicted to continue to slowly rise, but due to these factors in their population afore mentioned, their numbers did not reach levels seen before 1990.

Therefore, this exploration of Lake Huron Bloater egg deposits gave us greater insight into the environment that their population lived in and allowed us to accurately forecast trends in their population through our analysis.

References

Schaeffer, J. S. 2004: 'Population dynamics of bloaters Coregonus hoyi in Lake Huron, 1980–1998' — Ann. Zool. Fennici 41: 271–279.

Appendix

A. Sort Function

```
sort.score <- function(x, score = c("bic", "aic")){
        if (score == "aic"){
                x[with(x, order(AIC)),]
        } else if (score == "bic") {
                x[with(x, order(BIC)),]
        } else {
                warning('score = "x" only accepts valid arguments ("bic", "aic)')
        }
}
```

B. Total R Code

```
rm(list = ls())
library(TSA)
library(forecast)
library(tseries)
library(lmtest)

# reading csv file into workspace

eggs <- read.csv("D:/uni/R/math2204/eggs.csv", header = TRUE)
eggs
summary(eggs)

# creating time series

eggsTS <- ts(eggs$eggs, start = 1981, freq = 1, end = 1996)
eggsTS
summary(eggsTS)

plot(eggsTS,
     type = 'o',
     sub = "Figure 1",
     font.sub = 2,
     main = "Time Series Plot for the Fish Egg Series",
     ylab = "Eggs Deposits (in millions)",
     xlab = "Year")

# changes in variance so log transformation

eggsTSLog <- log(eggsTS)
plot(eggsTSLog,
     main = "Time Series plot of Log Transformed Fish Eggs Series",
     type = 'o',
     xlab = 'Year',
     ylab = 'Eggs Deposited (in millions)',
     sub = 'Figure 2',
     font.sub = 2)


# 1st Difference


eggsTSDiff <- diff(eggsTSLog, differences = 1)

plot(eggsTSDiff,
     type = 'o',
     sub = "Figure 3",
     font.sub = 2,
     main = "Time Series Plot for the Transformed Series at the 1st Difference",
     ylab = "Eggs Deposits (in millions)",
     xlab = "Year")
abline(h=0)
```

```r

# stationary tests for 1st difference

adf.test(eggsTSDiff, alternative = c("stationary"))
pp.test(eggsTSDiff, alternative = c("stationary"))


# 2nd Difference

eggsTSDiff2 <- diff(eggsTSLog, differences = 2)

plot(eggsTSDiff2,
     type = 'o',
     sub = "Figure 5",
     font.sub = 2,
     main = "Time Series Plot for the Transformed Series at the 2nd Difference",
     ylab = "Eggs Deposits (in millions)",
     xlab = "Year")
abline(h=0)

# stationary tests for 2nd difference

adf.test(eggsTSDiff2, alternative = c("stationary"))
pp.test(eggsTSDiff2, alternative = c('stationary'))



## MODEL SPECIFICATION

par(mfrow = c(1,2))

acf(eggsTSDiff2, main = "ACF of 2st Differenced Egg Series")
pacf(eggsTSDiff2, main = "PACF of 2st Differrenced Egg Series")

par(mfrow = c(1,1))

eacf(eggsTSDiff2, ar.max = 3, ma.max = 3)

res = armasubsets(y = eggsTSDiff2, nar = 4, nma = 4, y.name = 'eggs', ar.method = 'ols')
plot(res)
```

```r
## MODEL FITTING

model.021 <- arima(eggsTSLog, order = c(0,2,1), method = 'ML')
coeftest(model.021)

model.022 <- arima(eggsTSLog, order = c(0,2,2), method = 'ML')
coeftest(model.022)

model.121 <- arima(eggsTSLog, order = c(1,2,1), method = 'ML')
coeftest(model.121)

model.020 <- arima(eggsTSLog, order = c(0,2,0), method = 'ML')
coeftest(model.020)

# sort function

sort.score <- function(x, score = c("bic", "aic")){
        if (score == "aic"){
                x[with(x, order(AIC)),]
        } else if (score == "bic") {
                x[with(x, order(BIC)),]
        } else {
                warning('score = "x" only accepts valid arguments ("bic", "aic)')
        }
}

sort.score(AIC(model.021, model.022, model.121), score = "aic")
sort.score(BIC(model.021, model.022, model.121), score = "bic")
```

```r
## Residuals for ARIMA(0,2,1)

model.021Res = rstandard(model.021)

par(mfrow = c(3,2))

plot(model.021Res,
     xlab = 'Time',
     ylab = 'Standardised Residuals',
     type = 'o',
     main = 'Time series plot of standardised residuals for ARIMA(0,2,1)')
abline(h=0)

hist(model.021Res,
     ylab = 'Standardised Residuals',
     main = "Histogram of Standardised Residuals for ARIMA(0,2,1)",
     xlim = c(-3, 3))

qqnorm(model.021Res,
       main = 'QQ plot of Standardised Residuals for ARIMA(0,2,1)')
qqline(model.021Res, col = 2, lwd = 1, lty = 2)

shapiro.test(model.021Res)

acf(model.021Res, main = 'ACF for the Standardised Residuals')
pacf(model.021Res, main = 'PACF for the Standardised Residuals')

par(mfrow = c(1,1))

tsdiag(model.021, gof = 15, omit.initial = F)
```

```r
## Residuals for ARIMA(1,2,1)

model.121Res = rstandard(model.121)

par(mfrow = c(3,2))

plot(model.121Res,
     xlab = 'Time',
     ylab = 'Standardised Residuals',
     type = 'o',
     main = 'Time series plot of standardised residuals for ARIMA(1,2,1)')
abline(h=0)

hist(model.121Res,
     ylab = 'Standardised Residuals',
     main = "Histogram of Standardised Residuals for ARIMA(1,2,1)",
     xlim = c(-3, 3))

qqnorm(model.121Res,
       main = 'QQ plot of Standardised Residuals for ARIMA(1,2,1)')
qqline(model.121Res, col = 2, lwd = 1, lty = 2)

shapiro.test(model.121Res)

acf(model.121Res, main = 'ACF for the Standardised Residuals')
pacf(model.121Res, main = 'PACF for the Standardised Residuals')

par(mfrow = c(1,1))
```

```r
# Residuals for ARIMA(0,2,2)

model.022Res = rstandard(model.022)

par(mfrow = c(3,2))

plot(model.022Res,
     xlab = 'Time',
     ylab = 'Standardised Residuals',
     type = 'o',
     main = 'Time series plot of standardised residuals for ARIMA(0,2,2)')
abline(h=0)

hist(model.022Res,
     ylab = 'Standardised Residuals',
     main = "Histogram of Standardised Residuals for ARIMA(0,2,2)",
     xlim = c(-3, 3))

qqnorm(model.022Res,
       main = 'QQ plot of Standardised Residuals for ARIMA(0,2,2)')
qqline(model.022Res, col = 2, lwd = 1, lty = 2)

shapiro.test(model.022Res)

acf(model.022Res, main = 'ACF for the Standardised Residuals')
pacf(model.022Res, main = 'PACF for the Standardised Residuals')

par(mfrow = c(1,1))


## Forecasting

fit = Arima(eggsTSLog, c(0,2,1))
fitFrc = forecast(fit, h = 5)
fitFrc

plot(fitFrc,
     sub = "Figure 17",
     font.sub = 2)
```