MATH1318 Time Series Analysis / MATH2204 Time Series and Forecasting

Final Project Report

Declaration of contributions:

No	Name of Team Member	Contribution to the project
1	Huu Thien Ngoc Phung	1
2		
3		
4		
5		
6		
	Sum:	1 must be 1

RMIT University,

School of Science

2022

Table of Contents

INTRODUCTION

The goal of this report is to analyse the Fish recruitment data for Northern Pike from Lake Windermere – North basin (1944-1978) using descriptive statistics and forecast the series for the next 10 years which is from 1979 to 1988 using suitable predicting model.

METHODOLOGY

The data set used for this analysis is obtained from FSAdata package (R built-in data package), containing 35 observations of yearly fish recruitment number for Northern Pike from Lake Windermere – North basin between 1944 and 1978.

First, the report will introduce a thorough descriptive analysis of the time series using multiple tools such as time series plot, scatter plots, sample autocorrelation function (ACF) plot, partial autocorrelation function (PACF) plot and summary statistics. After having a general understanding of the data series, the report also conducts Augmented Dickey-Fuller unit root test (ADF test) to investigate the stationarity of the data set. If the series is non-stationary, necessary transformation such as Box-Cox transformation and differencing transformation will be applied to make the data stationary before proceeding to model building stage.

Next, in order to develop a suitable predicting model for the data series, we adopt the three-step model-building strategy including model specification, model fitting, and model diagnostics (Cryer and Chan, 2008).

In model specification, we apply a wide range of model specification tools including ACF-PACF, the extended autocorrelation function (EACF) and Bayesian Information Criterion (BIC) table to list out suitable set of ARIMA(p,d,q) models.

In model fitting, the report will utilise two statistical estimation methods which are least squares estimation and maximum likelihood estimation to determine the best possible estimates of a number of parameters involved in the model. A comparison of AIC (Akaike information criterion), BIC (Bayesian information criterion) and error measures among the possible models will be made to select the model with the lowest AIC, BIC and error measures.

In model diagnostics step, the report will evaluate the selected model's quality using residual analysis. If the model fit the data well and the model assumptions are well-satisfied, the residuals are expected to have white noise like pattern. A residual analysis will include a time series plot, histogram, Quantile-Quantile (QQ) plot, ACF plot and plot of p-values of Ljung-Box statistics.

A discussion of all above factors will be provided to select the best fitting model then this model will be checked for anomalies with overfitting. If the model is confirmed in terms of goodness of fit, the report will the data series to this model to forecast the fish recruit number for the next 10 years.

RESULTS

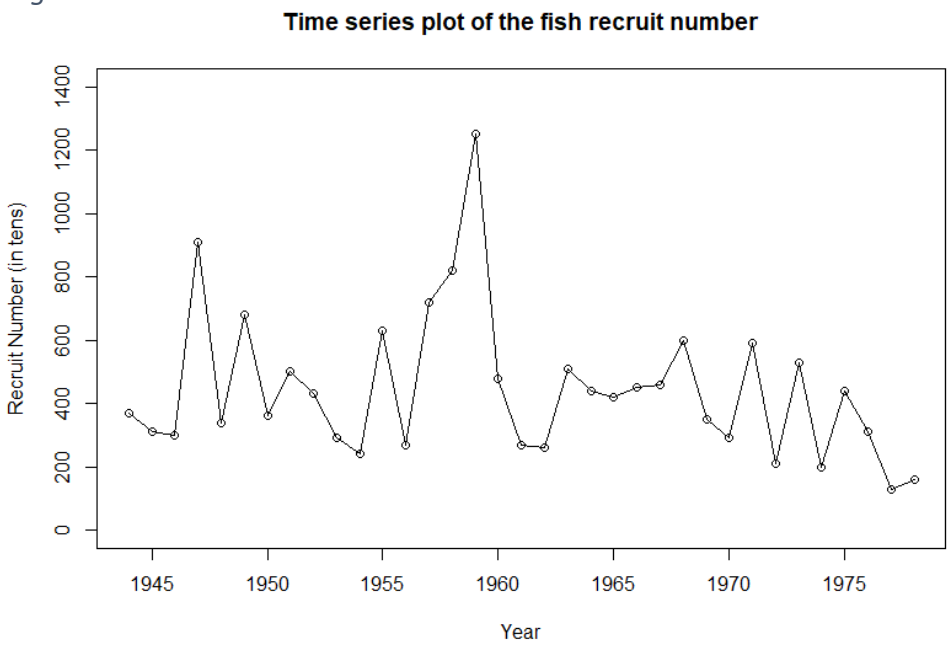
I. DESCRIPTIVE ANALYSIS

1. Time series plot

The **Figure 1** shows the time series plot of the raw fish recruit series:

- **Trend:** There is a slight linear downward tendency in the overall trend of the series. The fish recruit number (in tens) dropped from close to 400 in 1944 to only above 130 in 1978.
- **Seasonality**: there is no repeating pattern observed from the plot.
- **Changing variance**: There are changing variances throughout the entire period, in which the variance is largest between 1956 to 1961 comparing to other years.
- **Behaviour**: there are fluctuations in the entire period, and succeeding points from 1951 to 1954 and from 1956-1962, so it is likely to have both moving average behaviours and autoregressive behaviours.
- **Changing point**: no intervention point is observed from the plot.

Figure 1



2. Summary statistics

Figure 2 Table of Summary Statistics of the data series

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
130.0	290.0	420.0	443.4	520.0	1250.0

In **figure 2**, the summary statistics (measured in tens) of the data series shows a huge gap between minimum value of 130 and maximum value of 1,250, which indicates a large variation in the data series. 25% of the values are below 290 and 25% of the values are greater 520.

The median is 420 which is smaller than the mean of 443.4, indicating a right skewed distribution with some extremely large values. This is further supported by a small gap

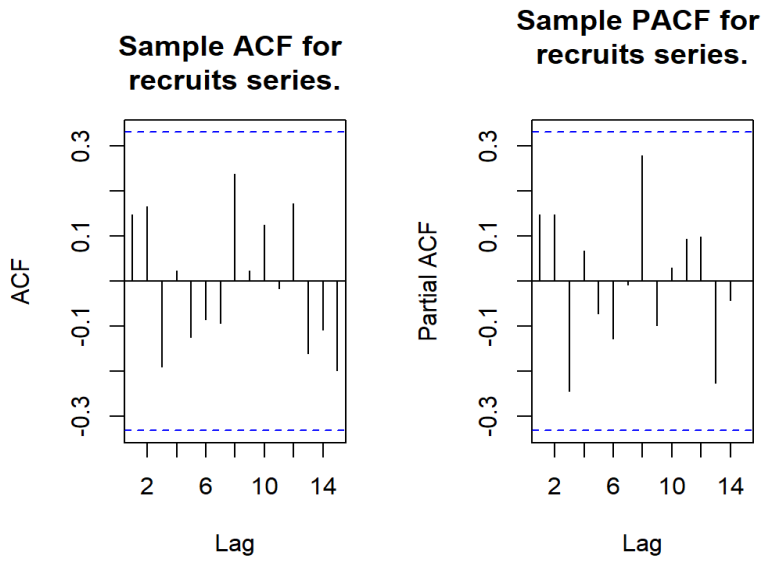
between the minimum (130) and first quartile (290) and much greater gap between the third quartile (520) and the maximum (1,250).

3. Autocorrelation

Figure 3 Scatter plots of recruit number and its first lag and second lag

From the scatter plots in **Figure 3**, it is unlikely to see the correlation of each data point and its previous year value as well as its second lag value.

Figure 4 ACF and PACF for fish recruits series



From **Figure 4**, there is no decaying pattern in ACF plot and no significant first lag in PACF to show proof of non-stationarity. However, considering the small size of the data set (only 35 observations), we need to confirm this observation with a unit root test.

```
        Augmented Dickey-Fuller Test

data:  raw
Dickey-Fuller = -2.5711, Lag order = 3, p-value = 0.3513
alternative hypothesis: stationary
```

From the ADF unit-root test result, the p-value is 0.35 which is larger than alpha level of 0.05 indicating a strong proof for non-stationarity

Even though there is no autocorrelation in the original series, the series is non-stationary with trend and changing variances, indicating a high possibility of stochastic trend. Therefore, in this report, we will focus on fitting the series to a suitable ARIMA model for forecasting.

The report will attempt to transform the data series to deal with the non-stationarity of the series.

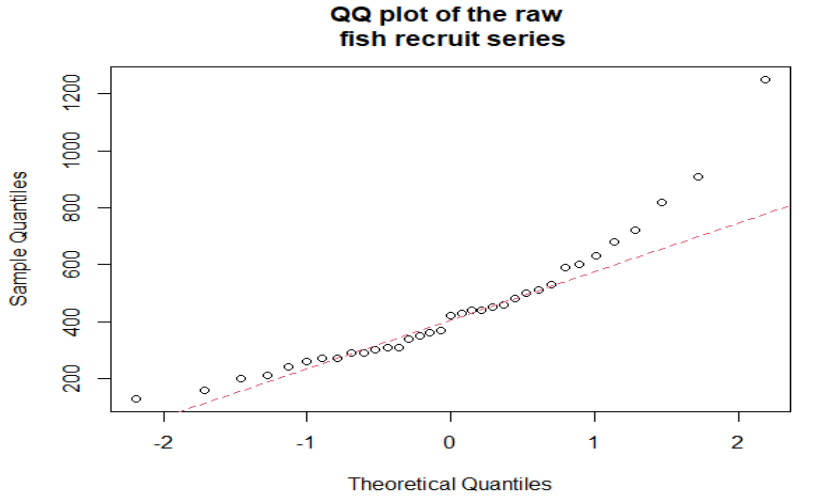
II. TRANSFORMATION

1. Box-Cox Transformation

As discussed in the descriptive analysis, there are changing variances throughout the entire period. Therefore, the report will apply Box-Cox transformation to stabilise the changing variances. In order to evaluate the Box-Cox transformation's effectiveness, we will inspect the normality of the series before and after transformation.

In **Figure 5**, the QQ plot of the fish recruit series shows that majority of the data points do not align with the reference line, indicating non-normality. This observation is confirmed by Shapiro-Wilk test p-value of 0.002 which is smaller than alpha level thus normality cannot be assumed.

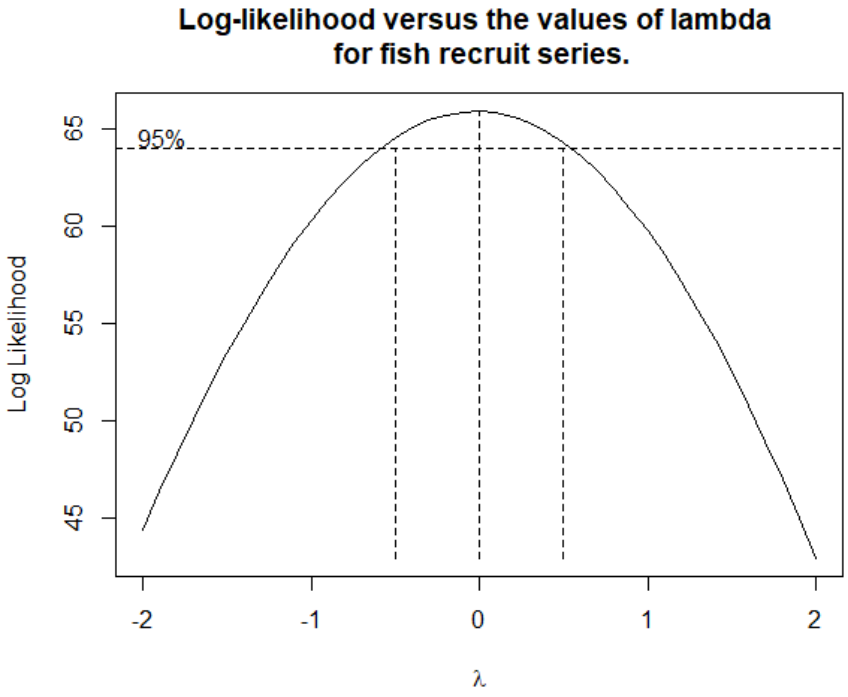
Figure 5




```
        Shapiro-Wilk normality test

data:  dataseries
W = 0.88715, p-value = 0.001817
```

Figure 6 illustrates the log-likelihood versus the values of lambda for fish recruit series. The 95% confidence interval for lambda is between -0.5 and 0.5. The lambda value is 0 which strongly suggests a logarithmic transformation.

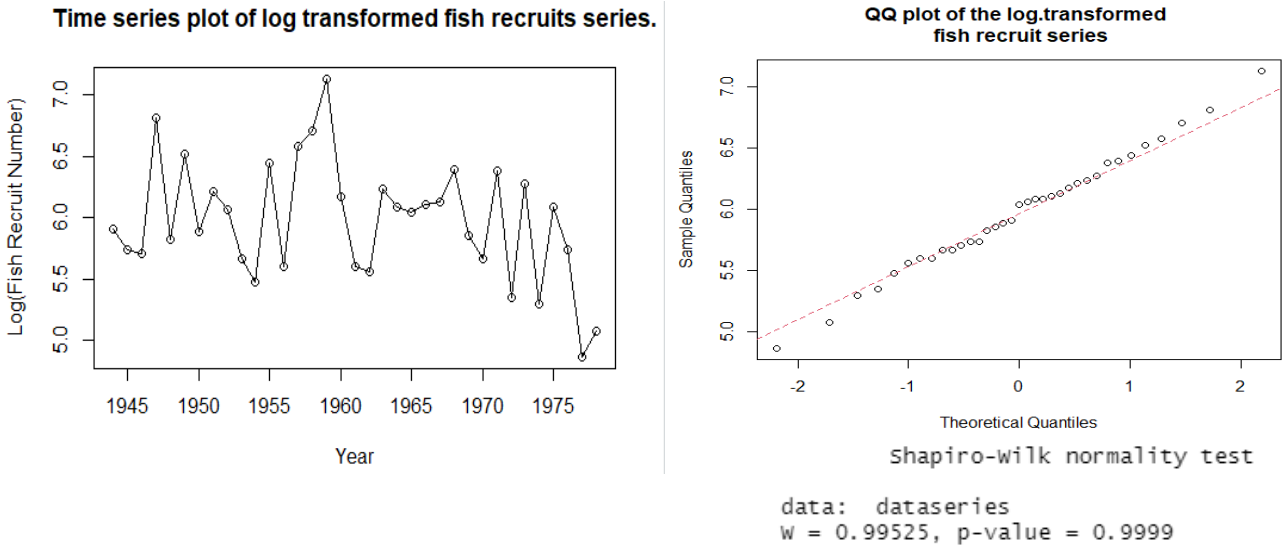
Figure 6




```
[1] "The lambda value is 0 ."
[1] "The 95% confidence interval for lambda is between -0.5 and 0.5"
```

Applying the log transformation to the raw series **(Figure 7),** the new time series plot showed more stable variances while the QQ plot shows most of the data points aligns with the reference line indicating normal distribution and nearly perfect p-value 0.99 in the Shapiro-Wilk test indicating normality can be assumed. Therefore, we will proceed with the log transformed series for further steps.

Figure 7



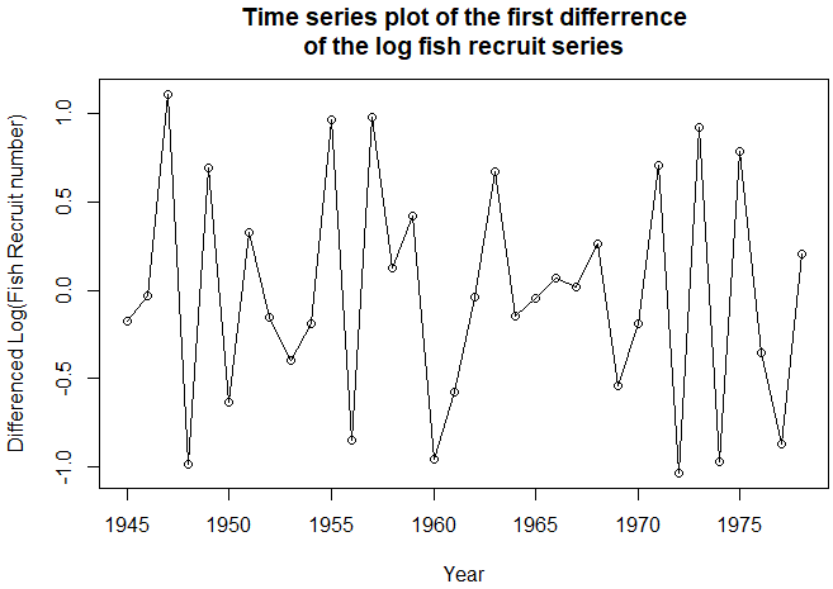
2. Differencing

It can be seen from **Figure 7** the time series ploot of log transformed fish recruit series that there is still existence of the downward trend. In order to detrend the series, we apply the first diffferencing to the log transformed series.

The time series plot of the first differrencing of log transformed fish recruit series in **Figure 8** shows no more trend. Meanwhile, the ADF test p-value is now improved to 0.02, which is smaller than alpha level 0.05 confirming stationarity can be assumed.

Hence we conclude that the first differencing make the series stationary (d=1), there is no need to go for a higher order of differencing and we can start with model specification.

Figure 8



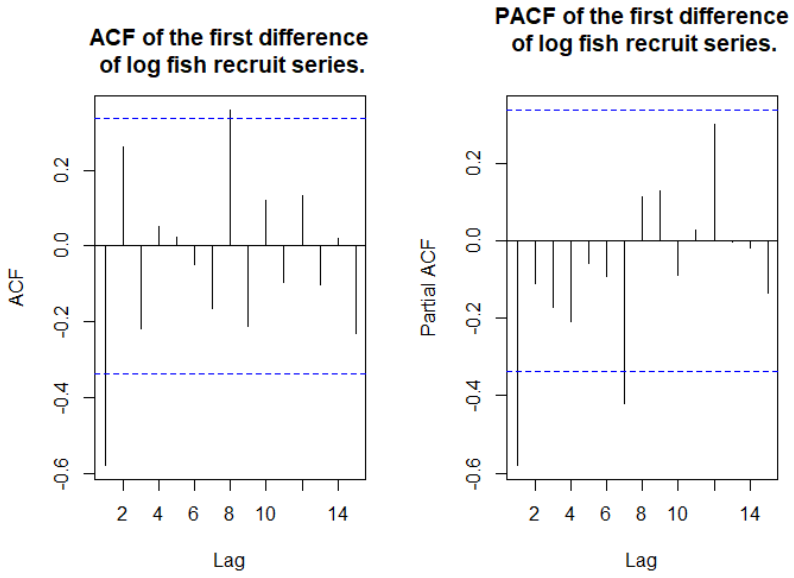

```
        Augmented Dickey-Fuller Test

data:  differenced.log
Dickey-Fuller = -4.1303, Lag order = 3, p-value = 0.01653
alternative hypothesis: stationary
```

III. MODEL-BUILDING

1. Model Specification

a) ACF and PACF

Figure 9



In Figure , there is significant sample autocorrelation at the first lag in ACF (q=1) and there is also significant partial autocorrelation at the first lag in PACF (p=1). At the same time, we ignore the significant autocorrelationn at lag 8 in ACF and at lag 7 in PACF because these are at the very late lags and should not affect the predicting model.

So from ACF and PACF, the possible model is {ARIMA(1,1,1)}.

b) EACF

From the EACF in **Figure 10**, we found the top left 'o' position corresponding to AR(1) and MA(0) that is p=1 and q=0. Its neiboring points are (p=1, q=1) and (p=2, q=1).

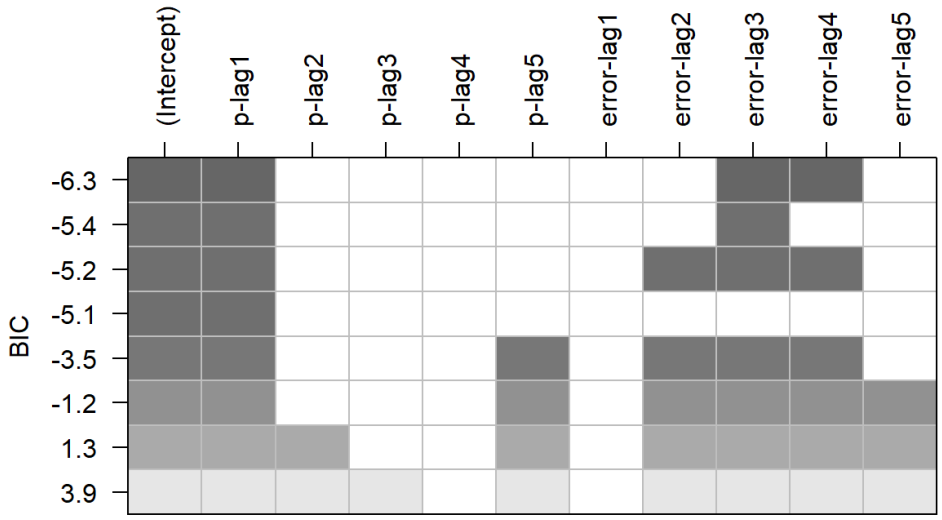
Figure 10: EACF of the first difference of the log transformed fish recruit series

AR/MA	0	1	2	3	4	5
0	x	o	o	o	o	o
1	o	o	o	o	o	o
2	x	o	o	o	o	o
3	x	o	o	o	o	o
4	o	o	o	o	o	o
5	o	o	o	o	o	o

Therefore, possible models from EACF are {ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(2,1,1)}

c) BIC

Figure 11 BIC table for the first difference of log transformed fish recruit series



The report inspects top 4 rows from the BIC table (**Figure 11**) for the top 4 best models in which the top row shows the best models.

- First row of the table:
  - p=1 (supported by all models)
  - q=3 (supported by top 3 models) and q=4 (also supported by third best model)

  Therefore, the best models are {ARIMA(1,1,3), ARIMA(1,1,4)}

- Second row of the table: supported the first row
- Third row of the table (only listed unspecified models from above):
  - p=1
  - q=2

  Therefore, the third best model is {ARIMA(1,1,2)}

- Fourth row of the table:
  - p=1
  - q=0

  Therefore, the fourth best model is {ARIMA(1,1,0)}

We do not inspect row that is further than top 4 rows in the BIC table because these rows support big models (p=5 and q all greater than 1) while there is no evidence from ACF_PACF and EACF for big models and our data set is small.

Hence, from BIC table, the set of possible models are { ARIMA(1,1,0), ARIMA(1,1,2), ARIMA(1,1,3), ARIMA(1,1,4)}.

From ACF-PACF, EACF and BIC table, we have the final set of possible models are { ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(1,1,2), ARIMA(2,1,1), ARIMA(1,1,3), ARIMA(1,1,4)} with underlined models are the one supported by two out of three methods.

2. Model Fitting & Model Diagnostics

a) Parameter estimations

In this section, we will fit each of specified models in Model Specification using Least Squares (Conditional SS - CSS) and Maximum Likelihood Estimation (MLE) methods. For the coefficients to be significant, its p-value need to be smaller than the alpha level 0.05.

```
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **ARIMA(1,1,0)**
  - CSS

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.57752    0.14010 -4.1222 3.753e-05 ***
---
```

  - MLE

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.56265    0.13684 -4.1118 3.926e-05 ***
---
```

For ARIMA(1,1,0), both CSS and MLE agreed that the coefficients (ar1) is highly significant.

- **ARIMA(1,1,1)**
  - CSS

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1 -0.045124   0.234603 -0.1923    0.8475
ma1 -0.752721   0.187820 -4.0077 6.132e-05 ***
---
```

  - MLE

```
z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ar1  0.0021514  0.2286256  0.0094    0.9925
ma1 -0.7907962  0.1782912 -4.4354 9.189e-06 ***
---
```

Next, adding MA(1) parameter to the model, we got ARIMA(1,1,1). Both CSS and MLE agreed that the coefficients of ma1 is significant however the coefficients (ar1) turned to insignificant. This result consistent with the descriptive analysis in which the raw series showed strong moving average behaviour, thus adding MA(1) to the model would make MA(1) parameter to take all the explaination power of AR(1).

- **ARIMA(1,1,2)**
  - CSS

```
z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ar1 -0.793896   1.226029 -0.6475   0.5173
ma1 -0.062418   1.194234 -0.0523   0.9583
ma2 -0.520010   1.057987 -0.4915   0.6231
```

  - MLE

```
z test of coefficients:

      Estimate Std. Error   z value  Pr(>|z|)
ar1 -0.9997207  0.0028536 -350.3388 < 2.2e-16 ***
ma1  0.2347850  0.1613072    1.4555    0.1455
ma2 -0.7501617  0.1589165   -4.7205 2.353e-06 ***
---
```

  - CSS-MLE

```
z test of coefficients:

      Estimate Std. Error   z value  Pr(>|z|)
ar1 -0.9996631  0.0034042 -293.6572 < 2.2e-16 ***
ma1  0.2329339  0.1616071    1.4414    0.1495
ma2 -0.7505598  0.1587607   -4.7276 2.272e-06 ***
---
```

Increasing the MA order by 1, for ARIMA(1,1,2), CSS method shows that all three coefficients are insignificant while MLE indicates that the coefficients ar1 and ma2 are highly significant while ma1 is insignificant.

Because CSS and MLE disagreed with each other, we use CSS-MLE method for parameter estimates and the result p-values of coefficients support MLE conclusion which shows coefficients ar1 and ma2 are significant and coefficients ma1 is insignificant.

- **ARIMA(2,1,1)**
  - CSS

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1  0.060335   0.247521  0.2438    0.8074
ar2  0.187293   0.215564  0.8689    0.3849
ma1 -0.824866   0.174420 -4.7292 2.254e-06 ***
---
```

  - MLE

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1  0.089109   0.235999  0.3776    0.7057
ar2  0.213025   0.216649  0.9833    0.3255
ma1 -0.869439   0.171624 -5.0660 4.063e-07 ***
---
```

Instead of increasing MA by 1, we increase AR by 1 which forms ARIMA(2,1,1), both CSS and MLE agreed that the coefficients (ar1 and ar2) are insignificant while the coefficients of ma1 is significant. Again, this is consistent with descriptive analysis observation in which moving average behaviour has stronger explaination power than autoregressive behaviour.

- **ARIMA(1,1,3)**
  - CSS

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1 -0.332757   0.182461 -1.8237  0.068195 .
ma1 -0.290134   0.090691 -3.1992  0.001378 **
ma2  0.188546   0.155766  1.2104  0.226108
ma3 -0.878994   0.182674 -4.8118 1.496e-06 ***
---
```

  - MLE

```
z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ar1 -0.35130    0.20301 -1.7305  0.08354 .
ma1 -0.25737    0.17305 -1.4872  0.13695
ma2  0.12376    0.16320  0.7583  0.44826
ma3 -0.67588    0.21924 -3.0829  0.00205 **
---
```

  - CSS-MLE

```
z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ar1 -0.35131    0.20300 -1.7306 0.083531 .
ma1 -0.25734    0.17304 -1.4871 0.136977
ma2  0.12372    0.16323  0.7579 0.448491
ma3 -0.67586    0.21926 -3.0824 0.002053 **
---
```

Next, we inspect ARIMA(1,1,3), CSS method indicates that the coefficients (ma1 and ma3) are significant and the coefficients (ar1 and ma2) are insignificant. Meanwhile, MLE shows that only ma3 coefficients is significant and other coefficients are insignificant.

CSS-MLE method supports MLE in which only ma3 coefficients is significant.

- **ARIMA(1,1,4)**

  - CSS

```
z test of coefficients:

     Estimate Std. Error  z value Pr(>|z|)
ar1  0.312630   0.188439   1.6591   0.0971 .
ma1 -1.088124   0.108431 -10.0352   <2e-16 ***
ma2  0.717193   0.050358  14.2420   <2e-16 ***
ma3 -1.358017   0.065871 -20.6163   <2e-16 ***
ma4  0.993639   0.102299   9.7131   <2e-16 ***
---
```

  - MLE

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1  0.35553    0.20958  1.6964   0.08982 .
ma1 -1.22364    0.18287 -6.6911 2.214e-11 ***
ma2  0.55357    0.17981  3.0786   0.00208 **
ma3 -0.91841    0.22625 -4.0594 4.921e-05 ***
ma4  0.83933    0.19025  4.4117 1.025e-05 ***
---
```

  - CSS-MLE

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1  0.35539    0.20957  1.6958  0.089921 .
ma1 -1.22365    0.18289 -6.6907 2.221e-11 ***
ma2  0.55363    0.17980  3.0791  0.002076 **
ma3 -0.91858    0.22622 -4.0606 4.896e-05 ***
ma4  0.83946    0.19027  4.4120 1.024e-05 ***
---
```

Increasing AR order to 4, both CSS and MLE agreed that the coefficients (ar1) of ARIMA(1,1,4) is insignificant while all other coefficients are significant. We conduct CSS-MLE method to see if coefficients ar1 is significant, however, the result is consistent with the conclusion from the other two methods.

To summarise, ARIMA(1,1,4) has no insignificant coefficient. ARIMA(1,1,0), ARIMA(1,1,1) and ARIMA(1,1,2) has an insignificant coefficient. ARIMA(2,1,1) has two insignificant coefficients. ARIMA(1,1,3) has three insignificant coefficients.

Figure 12 shows the AIC and BIC of each specified models. ARIMA(1,1,4) has the lowest AIC value of 52.84 and ARIMA(1,1,0) has the lowest BIC value of 58.23. Because AIC and BIC disagreed on the best model, we investigate on their error measures to further evaluate these two models.

Figure 12

```
$IC
  p d q      AIC     AICc      BIC
6 1 1 4 52.84247 55.95358 62.00063
3 1 1 2 54.32448 55.70379 60.42992
2 1 1 1 54.89704 55.69704 59.47613
1 1 1 0 55.17492 55.56201 58.22764
4 2 1 1 55.86537 57.24469 61.97082
5 1 1 3 55.87825 58.02110 63.51005
```

b) Error measures

Figure 13

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
ARIMA(1,1,0)	-0.005658137	0.7054287	0.5582477	50.73409	216.2661	0.5999140	-0.33676026
ARIMA(1,1,1)	-0.065788999	0.5093628	0.4088849	100.49900	130.8149	0.4394031	-0.08129586
ARIMA(1,1,2)	-0.081618439	0.4585132	0.3821614	124.72863	151.9752	0.4106850	-0.07863550
ARIMA(2,1,1)	-0.072387074	0.5068952	0.4113003	104.68286	136.8824	0.4419988	-0.05441623
ARIMA(1,1,3)	-0.084124951	0.4679613	0.3776588	107.00209	163.1803	0.4058464	0.01687320
ARIMA(1,1,4)	-0.049527941	0.4372422	0.3278492	91.16724	113.5342	0.3523191	-0.13092534

From **Figure 13**, ARIMA(1,1,4) gives the lowest Root mean square error(RMSE), Mean absolute error(MAE), Mean absolute percentage error (MAPE), and Mean absolute squared error (MASE).

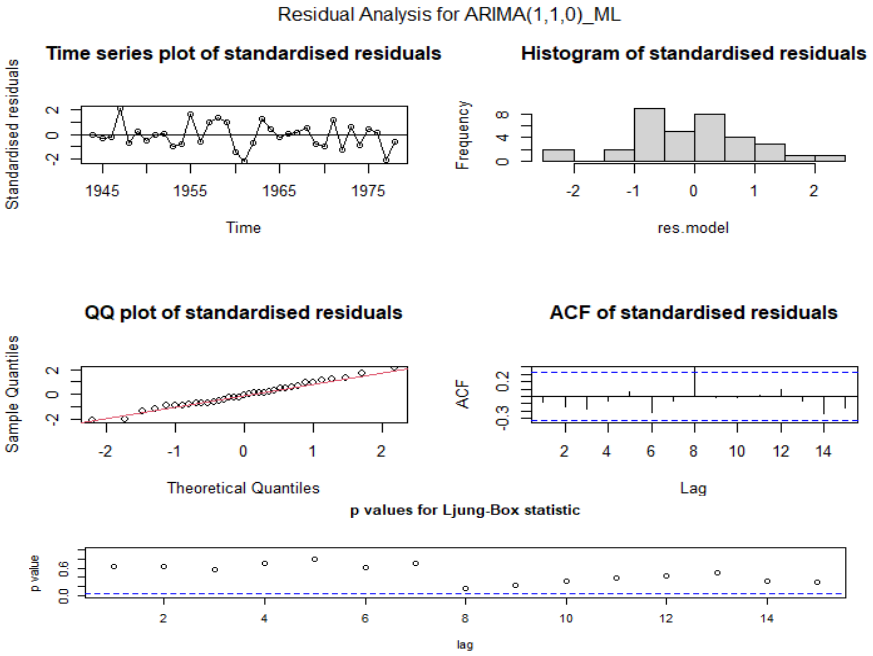
Meanwhile, ARIMA(1,1,0) though has the smallest Mean percentage error (MPE), it produces the highest values across all other error measures among the 6 specified models.

c) Residual analysis

- From **Figure 14**, the analysis of **ARIMA(1,1,0)** residual demonstrates:
  - Time series plot of standardised residual: points distributed randomly around the zero level. There is no clear trend, and no other information can be obtained from the plot.
  - Histogram: the histogram shape is closed to symmetry in range (-3,3) and most of the data points lie within (-1,1), indicating a good sign tthat we have small residuals.
  - QQ plot: all data points align with the reference line indicating strong normality of the residuals distribution. This is further confirmed with perfect p-value of (greater than alpha level 0.05) in Shapiro-Wilk test.

- ACF plot: there is only one significant autocorrelation at late lag (lag 8), which should not be impactful to the model (can be checked with Ljung-Box statistic), so there is no proof of trend or seasonality.
- Ljung-Box statistic: p-values of autocorrelation at all lags (including lag 8) are above alpha level 0.05, indicating there is no information about autocorrelation left in the residuals.

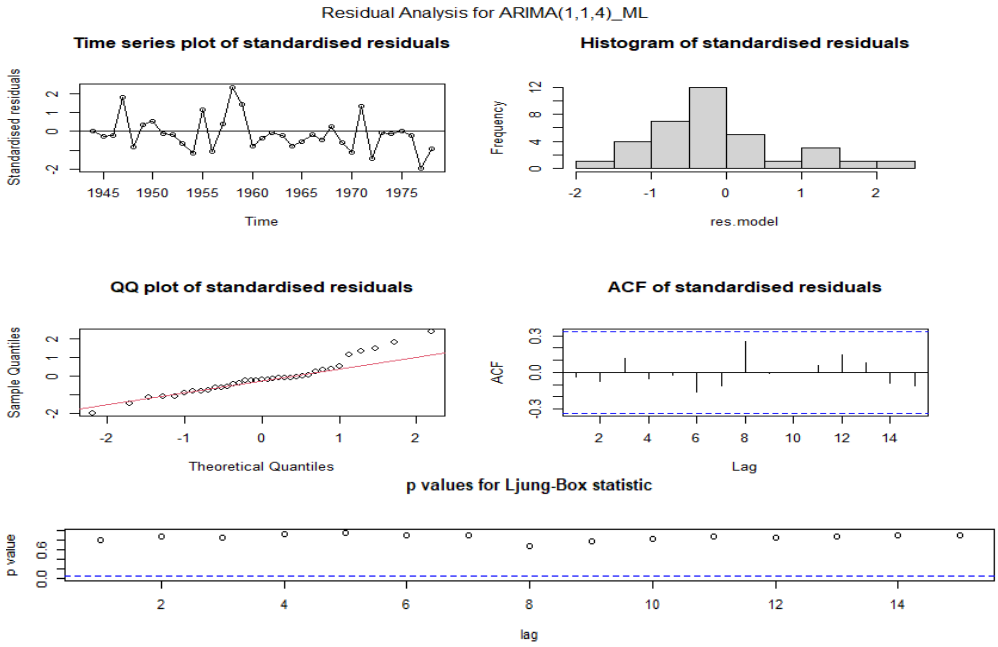
Figure 14




```
[1] "ARIMA(1,1,0)_ML"

        Shapiro-Wilk normality test

data:  res.model
W = 0.98788, p-value = 0.9599
```

- From **Figure 15**, the analysis of **ARIMA(1,1,4)** residual demonstrates:
  - Time series plot of standardised residual: points distributed randomly around the zero level. There is a very slight downward trend, beside that, there is no other information can be obtained from the plot.
  - Histogram: the histogram shape is not perfectly symmetry but all residuals are in range (-2,3) and most of the residuals lie within (-1,1), indicating a good sign that we have small residuals.
  - QQ plot: most residuals align with the red reference line indicating normal distribution, except some points at the right tail. This is further confirmed with p-value of 0.06 (greater than alpha level 0.05) in Shapiro-Wilk test.
  - ACF plot: there is only one significant autocorrelation at late lag (lag 8), which should not impact the model, so there is no proof of trend or seasonality.
  - Ljung-Box statistic: p-values of autocorrelation at all lags are way above alpha level 0.05, indicating there is no information about autocorrelation left in the residuals.

Figure 15




```
[1] "ARIMA(1,1,4)_ML"

        Shapiro-Wilk normality test

data:  res.model
W = 0.94206, p-value = 0.06468
```

In short, the residual analysis of ARIMA(1,1,0) and ARIMA(1,1,4) both satisfied the normality assumption of MLE, even though ARIMA(1,1,0) gave better normality results in QQ plot and Shapiro-Wilk test. Even though the time series plot of ARIMA(1,1,4) residuals might show a slight downward trend, there is no proof of trend and seasonality can be seen from the ACF plot.

DISCUSSION

In the 6 specified models, ARIMA(1,1,0) and ARIMA(1,1,4) are selected as top 2 promissing models for their lowest BIC value and AIC value respectively.
ARIMA(1,1,0) has no insignificant coefficient and it has the smallest BIC value, and MPE and good residual analysis, however, it has the worst values of other error measures and only ranked 4th in AIC. Its residual analysis shows satisfied result of white noise like pattern.
ARIMA(1,1,4) has the smallest AIC value and lowest RMSE, MAE, MAPE and MASE. However, there is an insignificant coefficient and it's only ranked 5th in BIC. This is likely because while both AIC and BIC consider the number of parameters, BIC also considers the length of the series. We have a short series (only 35 observations) and this model has more paramters than ARIMA(1,1,0). Its residual analysis also shows satisfied result of white noise like pattern even though cannot be as good as that of ARIMA(1,1,0) in terms of normality.
From the descriptive analysis, we can see that there is a strong moving average behaviour in the series, and from parameter estimation, [ar1] parameter lose its explaination power as soon as there is [ma] (moving average) component for most of the possible models. Therefore, considering all factors, we select ARIMA(1,1,4) as the most adequate model for predicting.

1. Overfitting

Next, we overfit the ARIMA(1,1,4) with ARIMA(2,1,4) and ARIMA(1,1,5) to see if there is anything overlooked during the analysis.

- ARIMA(2,1,4)
  - CSS

```
z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ar1 -0.5884536  0.4999911 -1.1769 0.2392242
ar2 -0.5203716  0.2255633 -2.3070 0.0210556 *
ma1 -0.0057961  0.5409135 -0.0107 0.9914505
ma2  0.5721646  0.1937003  2.9539 0.0031382 **
ma3 -0.9140766  0.2421269 -3.7752 0.0001599 ***
ma4 -0.1091434  0.5659112 -0.1929 0.8470662
---
```

  - MLE

```
z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ar1 -0.529274   0.672553 -0.7870  0.43130
ar2 -0.455217   0.236726 -1.9230  0.05448 .
ma1 -0.125021   0.722912 -0.1729  0.86270
ma2  0.369065   0.275629  1.3390  0.18057
ma3 -0.850701   0.264528 -3.2159  0.00130 **
ma4  0.054758   0.550996  0.0994  0.92084
---
```

  - CSS_MLE

```
z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ar1 -0.528024   0.670764 -0.7872 0.431166
ar2 -0.454952   0.236514 -1.9236 0.054408 .
ma1 -0.126284   0.720780 -0.1752 0.860918
ma2  0.369559   0.275324  1.3423 0.179509
ma3 -0.851068   0.264766 -3.2144 0.001307 **
ma4  0.055856   0.549446  0.1017 0.919027
---
```

From CSS, adding AR(2) to the model, ma1 and ma4 parameters now become insignificant. From MLE and CSS-MLE, only ma3 coefficient is significant.

- ARIMA(1,1,5)
  - CSS

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1 -0.185367   0.319146 -0.5808   0.56136
ma1 -0.546807   0.271947 -2.0107   0.04436 *
ma2  0.081747   0.293084  0.2789   0.78031
ma3 -0.975983   0.190145 -5.1328 2.854e-07 ***
ma4  0.276363   0.320777  0.8615   0.38894
ma5  0.582722   0.259427  2.2462   0.02469 *
---
```

  - MLE

```
z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ar1 -0.54849    0.57155 -0.9597 0.337222
ma1 -0.10048    0.64467 -0.1559 0.876137
ma2 -0.14742    0.32073 -0.4596 0.645773
ma3 -0.61993    0.20182 -3.0717 0.002128 **
ma4 -0.14442    0.56422 -0.2560 0.797982
ma5  0.31441    0.22451  1.4004 0.161396
---
```

  - CSS_MLE

```
z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ar1 -0.226899   0.482973 -0.4698 0.638500
ma1 -0.631910   0.469757 -1.3452 0.178565
ma2 -0.077829   0.445595 -0.1747 0.861344
ma3 -0.713500   0.266270 -2.6796 0.007371 **
ma4  0.326037   0.369205  0.8831 0.377193
ma5  0.523834   0.401187  1.3057 0.191652
---
```

From CSS, adding MA(5) to the model, even though ma5 is significant, coefficients of ma2 and ma4 are now insignificant. From MLE and CSS-MLE, only ma3 coefficient is significant.

Therefore, the result of overfitting ARIMA(1,1,4) has confirmed this model in terms of goodness of fit and we can use this model to forecast the fish recruit number number in the next 10 years.

2. Forecasting

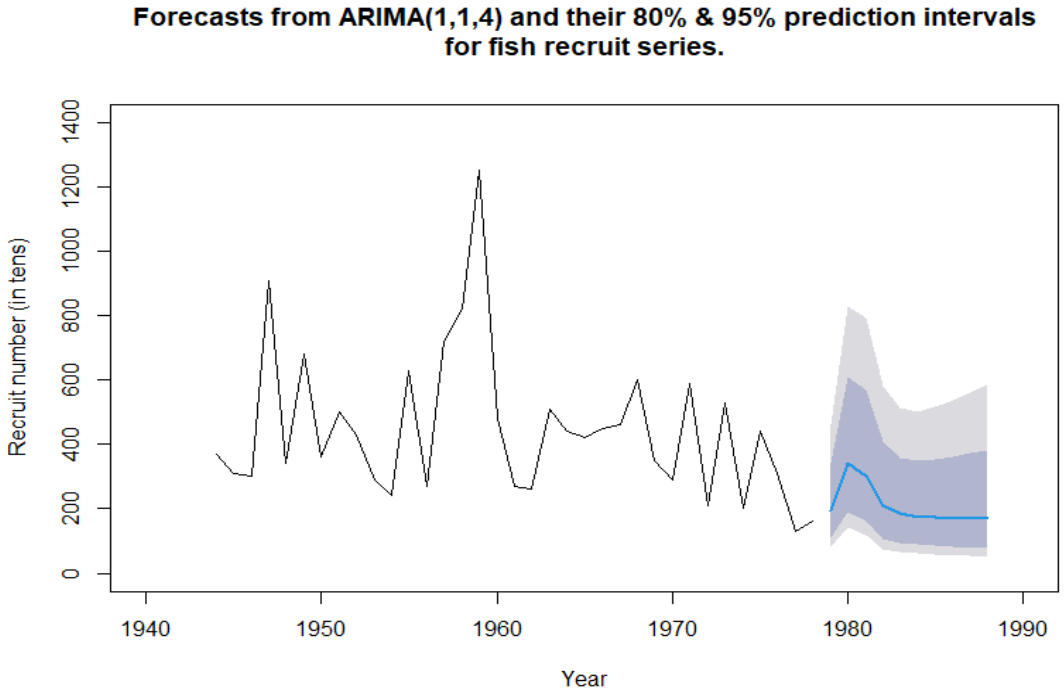
Fitting the data series to ARIMA(1,1,4), the forecasted values and its 80% and 95% confidence interval are listed in below table in **Figure 16**.

Figure 16 Table of forecasts value with 90% and 95% confidence intervals for 1979 – 1988 period.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1979	191.8130	108.32146	339.6578	80.04693	459.6333
1980	341.9967	191.55419	610.5935	140.93982	829.8701
1981	303.3571	161.95892	568.2029	116.17856	792.1043
1982	209.5746	107.52881	408.4629	75.52776	581.5283
1983	183.7627	94.06367	358.9987	65.98769	511.7429
1984	175.3764	88.09089	349.1494	61.18271	502.7054
1985	172.4891	84.28766	352.9874	57.69416	515.6932
1986	171.4744	81.31176	361.6141	54.77896	536.7660
1987	171.1153	78.73254	371.8976	52.20186	560.9080
1988	170.9878	76.39099	382.7262	49.86595	586.3086

The time series plot of fish recruit series and 10-year ahead forecast is illustrated in **Figure 17**. It can be seen that the forecasts successfully follow well the movement of the original series and is predicted to reach the next peak of fish recruit number of around 340 (in tens) in 1980.

Figure 17



CONCLUSION

From the Descriptive Analysis, the time series plot of the fish recruit series demonstrates a slight linear downward overall trend from 1944 to 1987. There is no seasonality, no changing point in the plot. There are clear changing variances with strong moving average behaviour and some autoregressive behaviour. Inspecting the series ACF and PACF plots, there are no significant autocorrelation in the series. The result of ADF unit-root test indicates non-stationary. Therefore, the series is highly likely to have stochastic trend and might be well-fit with an ARIMA model.

In order to deal with non-stationarity of the series, the report applied Box-Cox transformation to stabilise the variances and applied differencing transformation to de-trend the series. Box-Cox transformation indicated that a logarithmic transformation (lambda = 0) would optimise the transformation and improve the normality distribution of the data points. Meanwhile, the first differencing of the log transformed data is sufficient to make the series stationary.

In model specification, using ACF-PACF, EACF and BIC, the set of possible models are {ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(1,1,2), ARIMA(2,1,1), ARIMA(1,1,3), ARIMA(1,1,4)}.

In model fitting, the report performed parameter estimation utilising least square method and MLE, then chose the top 2 promising models which are ARIMA(1,1,0) with the lowest BIC and no insignificant coefficients and ARIMA(1,1,4) with the lowest AIC and a insignificant coefficient [ar1]. The report also produced the error measures for all 6 possible models. ARIMA(1,1,4) is found to have the lowest RMSE, MAE, MAPE and MASE while ARIMA(1,1,0) has the lowest MPE but ranked last in every other error measures (large values of errors).

Thus, for model diagnostics, the report conducted residual analysis for both ARIMA(1,1,0) and ARIMA(1,1,4). ARIMA(1,1,0) proved to provide better residuals (more random and higher normality) than ARIMA(1,1,4). However, ARIMA(1,1,4) residuals still statisfied the normality assumption of MLE and there is no proof of trend in ACF plot despite an observation of slight downward trend present in time series plot of standardised residuals.

Considering all above factors and the fact that there is strong moving average behaviour in the original series as well as [ar1] parameter lose its explaination power as soon as there is [ma] (moving average) component, we selected ARIMA(1,1,4) as the most adequate model for predicting and confirmed this choice with overfitting.

Finally, the report produced the 10-year ahead forecasts of fish recruit number at Northern Pike from Lake Windermere – North basin with a time series plot and a table of correspoding values including 80% and 95% confidence level. The forecasts successfully follows the movement of the original data.

APPENDICES

R script

```
library(FSAdata)
library(TSA)
library(fUnitRoots)
library(lmtest)
library(tseries)
library(forecast)

#User-defined function
# Normality analysis
norm <- function(dataseries){
  qqnorm(y=dataseries, main=paste('QQ plot of the', deparse(substitute(dataseries)),
                                  '\n fish recruit series'))
  qqline(y=dataseries, col = 2, lwd = 1, lty = 2)
  shapiro.test(dataseries)
  swt <- shapiro.test(dataseries)
  return(swt)
}
# Residual analysis
residual.analysis <- function(model,i, methodType) {
  res.model = rstandard(model)
  modelname = paste0("ARIMA(", orderList[[i]][1], ",",
                     orderList[[i]][2], ",", orderList[[i]][3], ")_", methodType)
  par(mfrow=c(2,2))
  par(mar=c(5.1, 4.1, 6, 2.1))
  plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of standardised
residuals")
  abline(h=0)
  hist(res.model,main="Histogram of standardised residuals")
  qqnorm(res.model,main="QQ plot of standardised residuals")
```

```
 qqline(res.model, col = 2)
 acf(res.model,main="ACF of standardised residuals")
 mtext(paste("Residual Analysis for", modelname), side = 3,
      line = -1.5, outer = TRUE)
 par(mfrow=c(1,1))
 tsdiag(model,gof=15,omit.initial=F)
 print(modelname)
 print(shapiro.test(res.model))
 par(mfrow=c(1,1))
}
# Function for model fitting
myCandidate <- function(timeSeries, orderList,
                    methodType = c("CSS-ML", "ML", "CSS")[1],
                    fixedList = NULL, icSortBy = c("AIC", "AICc","BIC")[1],
                    ...){

 # timeSeries = the time series (a ts object)
 # orderList = a list object of c(p, d, q)
 # methodType = estimation method; default = "CSS-ML"
 # fixedList = a list object of free/fixed coefficient
 # icSortBy = information criterion (IC) to be used to sort the IC table
 #        : default value: by AIC
 # ... Additional arguments to be passed to Arima
 myCandidateEst <- list()
 n <- length(orderList)
 for(i in 1:n){
  order_i <- sapply(orderList,function(x) unlist(x))[,i]
  myCandidateEst[[i]] <- Arima(y = timeSeries, order = order_i, method = methodType)
 }
 significanceTest <- list()            # a list for significance tests
 resana <- list()
 ICTable <- matrix(NA, nrow = n, ncol = 6) # create a matrix to store IC
```

```
 for(i in 1:n){
  for(j in 1:3){
   ICTable[i,j] <- orderList[[i]][j]       # return the ARIMA orders
  }
  ICTable[i,4] <- myCandidateEst[[i]]$aic #
  ICTable[i,5] <- myCandidateEst[[i]]$aicc
  ICTable[i,6] <- myCandidateEst[[i]]$bic

  significanceTest[[i]]<- coeftest(myCandidateEst[[i]])
  resana[[i]] <- residual.analysis(myCandidateEst[[i]],i,methodType)
 }

 ICTable <- data.frame(ICTable)
 names(ICTable) <- c('p', 'd', 'q', 'AIC', 'AICc', 'BIC')

 if(icSortBy == "AIC"){
  ICTable <- ICTable[order(ICTable$AIC),] # sort the table by AIC
 }else if(icSortBy == "AICc"){
  ICTable <- ICTable[order(ICTable$AICc),] # sort the table by AICc
 }else if(icSortBy == "BIC"){
  ICTable <- ICTable[order(ICTable$BIC),]
 }else{
  stop("Incorrect Information Criterion")
 }

 myCandidateEst <- list(model = myCandidateEst, IC = ICTable,
                        significanceTest = significanceTest,
                        residualanalysis = resana,
                        orderList = orderList)
 return(myCandidateEst)
}
```

```
# Load data
data(PikeWindermere)

raw <- ts(PikeWindermere[PikeWindermere$basin=='North' &
                         PikeWindermere$year == c(1944:1978),'recruits'],
          start = 1944)


# Descriptive analysis

# Time Series plot
par(mfrow=c(1,1))
plot(raw,ylab='Recruit Number (in tens)',xlab='Year',type='o',
    main = "Time series plot of the fish recruit number", ylim=c(0,1400))
summary(raw)

# Scatter plots
y = raw
x = zlag(raw)

index = 2:length(x)
paste("Correlation between recruit number and its first lag values:",
     cor(y[index],x[index]))
plot(y= raw,x=zlag(raw),ylab='Recruit number (in tens)',
    xlab='Recruit number of the previous year (in tens)',
    main= "Scatter plot of neighboring recruit number.")

x = zlag(zlag(raw))
index = 3:length(x)
paste("Correlation between recruit number and its second lag values:",
     cor(y[index],x[index]))
```

```
plot(y= raw,x=zlag(zlag(raw)), ylab='Recruit number (in tens)',
    xlab='Recruit number with 2 years of lag (in tens)',
    main= "Scatter plot of recruit number and its second lag values.")

# Autocorrelation
par(mfrow=c(1,2))
acf(raw, main="Sample ACF for \nrecruits series.")
pacf(raw, main="Sample PACF for \nrecruits series.")
par(mfrow=c(1,1))

# Unit-root test
adf.test(raw)

# Normality test
norm(raw)

# Box-Cox Transformation
BC = BoxCox.ar(raw)
title(main = "Log-likelihood versus the values of lambda \n for fish recruit series.")
BC$ci
lambda = BC$lambda[which(max(BC$loglike) == BC$loglike)]

#Values of the first and third vertical lines
print(paste('The 95% confidence interval for lambda is between',
        BC$ci[1], 'and',BC$ci[2]))
# To find the lambda value of the middle vertical line
print(paste('The lambda value is', lambda, '.'))

# Log transformation
log.transformed = log(raw)
plot(log.transformed,type='o', ylab='Log(Fish Recruit Number)', xlab='Year',
    main='Time series plot of log transformed fish recruits series.')
```

```
norm(log.transformed)
adf.test(log.transformed)

# normality improved, go with log transformed data
differenced.log = diff(log.transformed, differences = 1)
plot(differenced.log,ylab='Differenced Log(Fish Recruit number)',xlab='Year',type='o',
    main = "Time series plot of the first differrence\nof the log fish recruit series")

adf.test(differenced.log)



# Model specification

# ACF_PACF

par(mfrow=c(1,2))
acf(differenced.log, main = "ACF of the first difference
 of log fish recruit series.")
pacf(differenced.log, main = "PACF of the first difference
 of log fish recruit series.")
par(mfrow=c(1,1))
# {ARIMA(1,1,1)}

# EACF
eacf(differenced.log, ar.max = 5, ma.max = 5)
# {ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(2,1,1)}

# BIC
res = armasubsets(y=differenced.log,nar=5,nma=5,y.name='p',ar.method='ols')
plot(res)
# Best models: p=1 (supported by all below models), q=3 (supported by 2 below models),4
```

```
# Third best: p=1, q=2
# {ARIMA(1,1,3), ARIMA(1,1,4), ARIMA(1,1,2)}

# Model fitting & residual analysis
timeSeries = log.transformed
orderList =  list(
  c(1,1,0), c(1,1,1), c(1,1,2), c(2,1,1), c(1,1,3), c(1,1,4))

# CSS
myCandidate(timeSeries, orderList, methodType = c("CSS-ML", "ML", "CSS")[3],
          fixedList = NULL, icSortBy = c("AIC", "AICc","BIC")[1])

# MLE
myCandidate(timeSeries, orderList, methodType = c("CSS-ML", "ML", "CSS")[2],
          fixedList = NULL, icSortBy = c("AIC", "AICc","BIC")[1])

#CSS-ML
myCandidate(timeSeries, orderList, methodType = c("CSS-ML", "ML", "CSS")[1],
          fixedList = NULL, icSortBy = c("AIC", "AICc","BIC")[3])

# Error Measures
df.smodels=data.frame()
for(i in orderList) {
  modelfit = Arima(y = differenced.log, order = i, method = "ML")
  smodel = data.frame(t(accuracy(modelfit)[1:7]))
  df.smodels = rbind(df.smodels,smodel)
}
rownames(df.smodels)= c("ARIMA(1,1,0)", "ARIMA(1,1,1)", "ARIMA(1,1,2)",
                        "ARIMA(2,1,1)", "ARIMA(1,1,3)", "ARIMA(1,1,4)")
colnames(df.smodels)=c("ME", "RMSE", "MAE", "MPE", "MAPE",
                       "MASE", "ACF1")
df.smodels
```

```
# Overfitting

timeSeries = log.transformed

orderList =  list(c(2,1,4), c(1,1,5))


myCandidate(timeSeries, orderList, methodType = c("CSS-ML", "ML", "CSS")[3],
        fixedList = NULL, icSortBy = c("AIC", "AICc","BIC")[1])
myCandidate(timeSeries, orderList, methodType = c("CSS-ML", "ML", "CSS")[2],
        fixedList = NULL, icSortBy = c("AIC", "AICc","BIC")[1])
myCandidate(timeSeries, orderList, methodType = c("CSS-ML", "ML", "CSS")[1],
        fixedList = NULL, icSortBy = c("AIC", "AICc","BIC")[3])


# Forecasting

fit = Arima(raw,c(1,1,4), lambda = 0)

fitFrc = forecast(fit,lambda=0, h=10)

fitFrc

par(mfrow=c(1,1))

plot(fitFrc,  ylab='Recruit number (in tens)',xlab='Year', main ="Forecasts from ARIMA(1,1,4) and
their 80% & 95% prediction intervals\nfor fish recruit series.", xlim=c(1940,1990), ylim=c(0,1400))
```

REFERENCES

Cryer, J, & Chan, K, 2008, *Time Series Analysis with Applications in R,* Springer