MATH 1318- Time Series Analysis

Analysis and Forecast of Egg depositions (in millions) of age-3 Lake Huron Bloaters (Coregonus hoyi) 1981- 2001

By Avinash Matani

S3692165

Table of Contents

List of Figures 3
List of Exhibits 4
Introduction – 5
Methods – 5
Data Interpretation – 5
Time Series of the dataset – 5
Normality Test 7
Normality Test – Shapiro Wilk Test 8
Box Cox Transformation 8
Stationarity Test - Augmented Dickey-Fuller Test 8
Handling Non-Stationarity – Differencing 9
Order 9
of Differencing 9
Time Series Plot 9
ADF Test Results after differencing 9
Analysis of Non-Stationary Differenced Series 10
ACF and PACF Plot fourth differenced series 10
EACF Plot 10
BIC Table 11
Model Selection – Parameter Estimation 11
ARIMA (0,4,1) 11
ARIMA (1,4,1) 12
ARIMA (1,4,0) 12
ARIMA (2, 4, 0) 12
ARIMA (2, 4, 1) 13
ARIMA (0, 4, 2) 13
Model Selection – Residual Analysis 14
ARIMA (0, 4, 1) 15
ARIMA (1, 4, 1) 16
ARIMA (1, 4, 0) 17
ARIMA (2, 4, 0) 19
ARIMA (2, 4, 1) 20
ARIMA (0, 4, 2) 22
Model Selection Using AIC and BIC Scores 23
Model Selection - Overfitting Analysis of Selected Models 23

Forecast Analysis 24
Conclusion 25
References 25
Appendix: 26
Load Package, Viewing File and Series Conversion 26
ACF, PACF Plots 26
Normality Test 26
Box Cox Transformation 26
Differencing and ADF Test 27
ACF and PACF Plot 28
EACF Plot 28
BIC Table 28
Residual Analysis 28
Model and Coeff. Test 29
AIC and BIC Score 31
Overfitting Analysis 31
Forecasting 32

List of Figures

Figure 1 Time Series Plot of egg deposition (in millions) of age-3 Lake Huron from 1981 - 1996 5
Figure 2: Scatter Plot (Plot of Yt and Y(t-1)) 6
Figure 3: ACF and PACF plots of original series 7
Figure 4 QQ Plot of Original series 7
Figure 5 ACF and PACF of four time differenced series 10
Figure 6 Residual Analysis plots of ARIMA (0, 4, 1) 15
Figure 7 Residual Analysis plots of ARIMA (1, 4, 1) 16
Figure 8 Residual Analysis plots of ARIMA (1, 4, 0) 17
Figure 9 Residual Analysis plots of ARIMA (2, 4, 0) 19
Figure 10 Residual Analysis plots of ARIMA (2, 4, 1) 20
Figure 11 Residual Analysis plots of ARIMA (0, 4, 2) 22
Figure 12: Prediction plot of Egg Deposition (in millions) from year 1997 to 2001 24

List of Exhibits

Exhibit 1: Normality Test Results of Original Series 8
Exhibit 2: Stationarity Test Results of Original Series 8
Exhibit 3: EACF Plot of four times differenced series 10
Exhibit 4: BIC Table four times differenced series 11
Exhibit 5: Parameter Estimation based on CSS and Exhibit 6: Parameter Estimation based on ML 11
Exhibit 7: Parameter Estimation based on CSS and Exhibit 8: Parameter Estimation based on ML 12
Exhibit 9: Parameter Estimation based on CSS and Exhibit 10: Parameter Estimation based on ML .. 12
Exhibit 11: Parameter Estimation based on CSS and Exhibit 12: Parameter Estimation based on ML 12
Exhibit 13: Parameter Estimation based on CSS and Exhibit 14: Parameter Estimation based on ML 13
Exhibit 15: Parameter Estimation based on CSS and Exhibit 16: Parameter Estimation based on ML 13
Exhibit 17: AIC Score Model Comparison and Exhibit 18: BIC Score Model Comparison 23
Exhibit 19: Comparison of coefficient estimates of ARIMA (3, 4, 0) with ARIMA (2, 4, 0) 23
Exhibit 20: AIC Score Model Comparison and Exhibit 21: BIC Score Model Comparison 24
Exhibit 22: Egg Deposition (in millions) Forecast Results from year 1997 to 2001 24

Introduction –

The report is mainly concerned with describing, modelling and estimating stochastic trend model for time series data available in BloaterLH dataset of FSAdata package. The dataset consists of egg deposition (in millions) of age-3 Lake Huron Bloaters (Coregonus hoyi) between years 1981 and 1996. To model this series, we used an integrated autoregressive moving average model (ARIMA) and their properties are thoroughly explored. With a few iterations of the model building strategy, we arrived at the best possible models for the series. Further, various residual plots and error terms for constant variance, normality and independence were used for model diagnostics and check for model adequacy. At last, the original series is forecasted for the next five years (1997-2001) using the parameter estimated of the best model selected.

Methods –

We will be using R software to identify and select the appropriate models. The model selection and specification mainly employed the use of stochastic trend analysis, moving average and autoregressive process analysis, ACF and PACF plots, series transformation and differencing. Parameter estimation using maximum likelihood and the conditional sum of squares is also employed to check for model adequacy. Further, model diagnostic using residual analysis, overfitting and parameter redundancy methods are used to finalize models for forecasting application.

Data Interpretation –

The dataset represents egg deposition (in millions) of age-3 Lake Huron Bloaters (Coregonus hoyi) between years 1981 and 1996. This is a very short series which may not conform to the model selection and diagnostic framework used for modelling a large observation time series. However, an attempt is made to model and forecast this series with explanation and limitations of the model building concepts.

Time Series of the dataset –

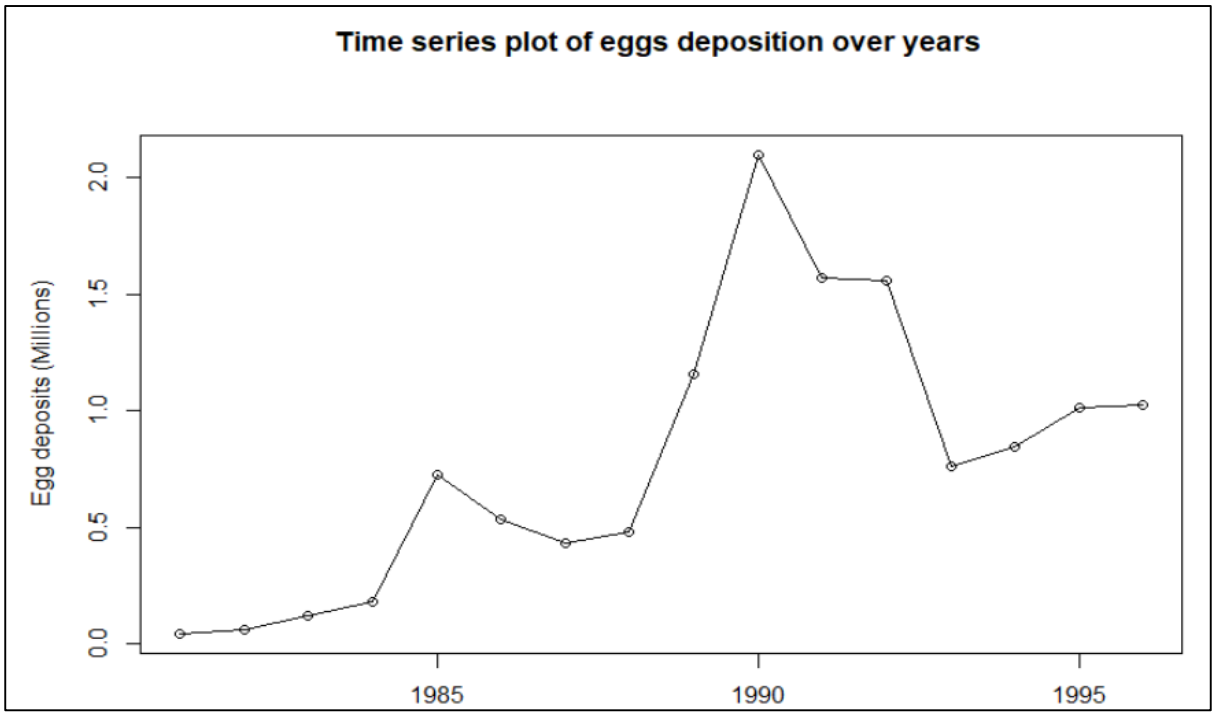


Figure 1 Time Series Plot of egg deposition (in millions) of age-3 Lake Huron from 1981 - 1996

Figure 1: shows the time series plot of the egg deposition (in millions) from the year 1981 - 1996. Here the values that are neighbours in time, tend to relate with one another. This shows the presence of the AR component in the series. Further, visual representation also suggests the presence of trend-making it non – stationary series.

The series has no seasonality and intervention points with a minor of changing variance. The AR component in the series can be better represented by constructing the scatter plot of values at consecutive time points. Minor presence of variation and trend in the series suggests that the stationary model is not reasonable. Thus, no deterministic model will work well for the series but one of the non-stationary models containing stochastic trend seems reasonable.

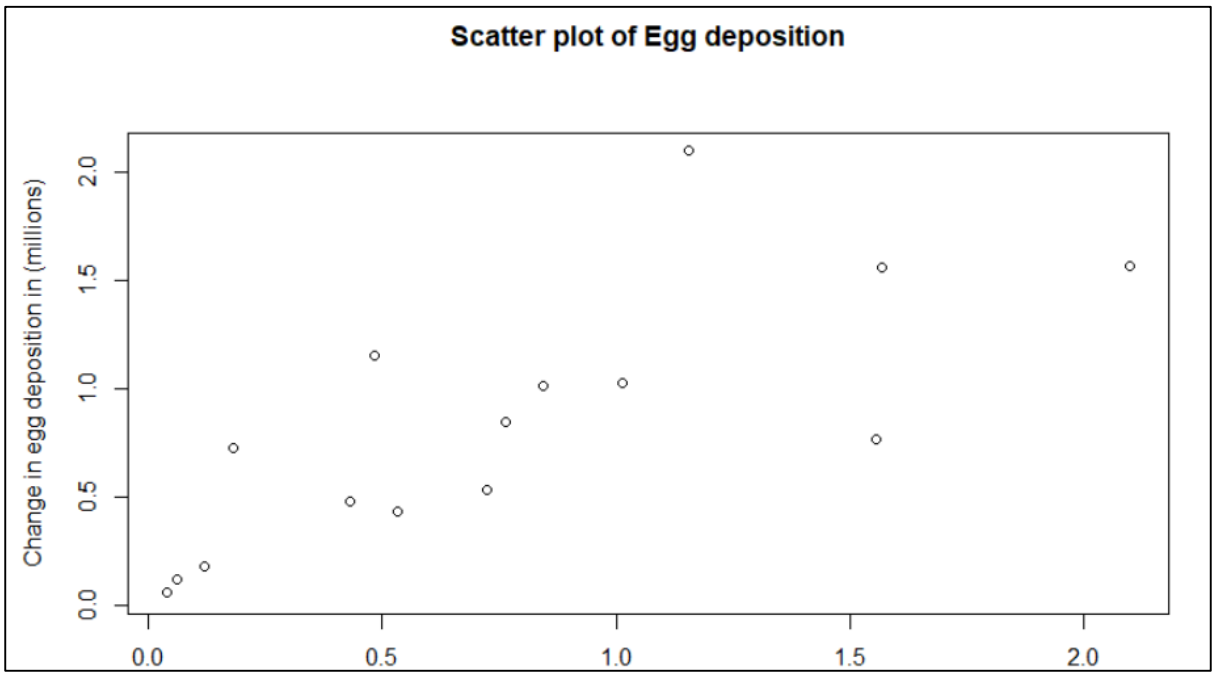


Figure 2: Scatter Plot (Plot of Yt and Y(t-1))

Figure 2: displays the scatter plot of the neighbouring pair of egg deposition (in million) values. It shows a slight upward trend – higher values of egg deposition tend to be followed in the next batch by higher values in the series. The upward trend is apparent and the scatter plot represents some correlation between the previous year and current year values.

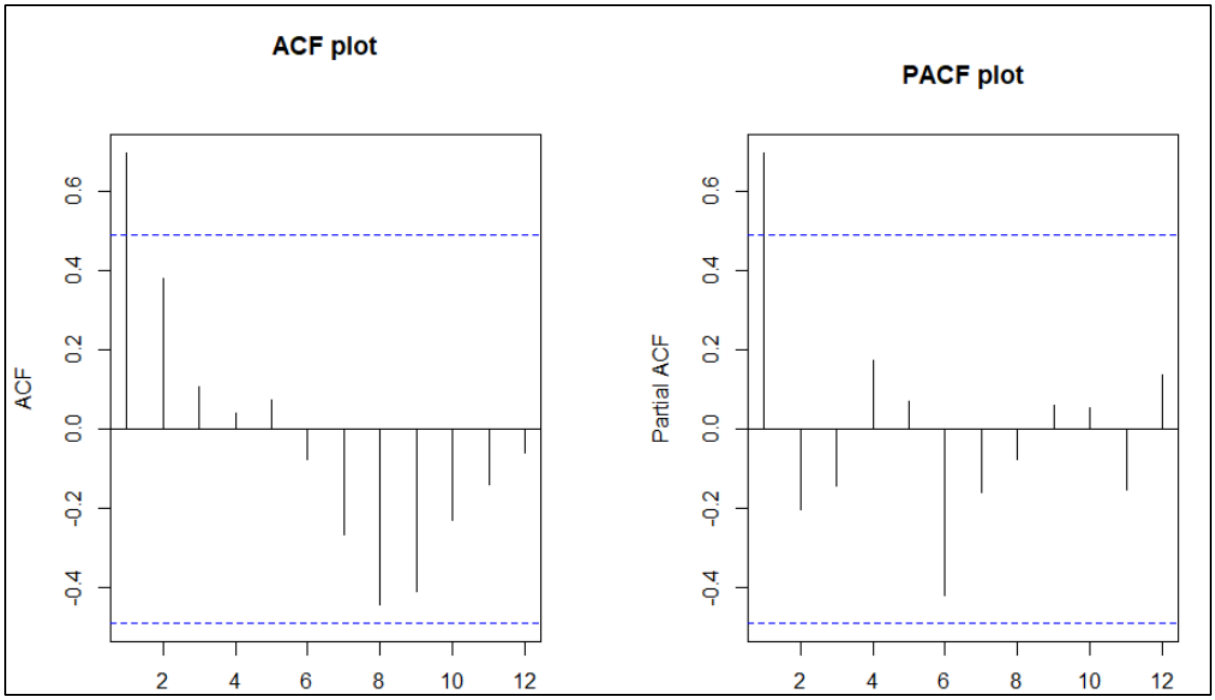


Figure 3: ACF and PACF plots of original series

Figure 3 represents the ACF and PACF plot of the original egg deposition series. The lag 1 autocorrelation in ACF and PACF plot is highly significant. However, owing to the presence of non-stationarity, no stationary model can be considered reasonable at this stage.

Normality Test

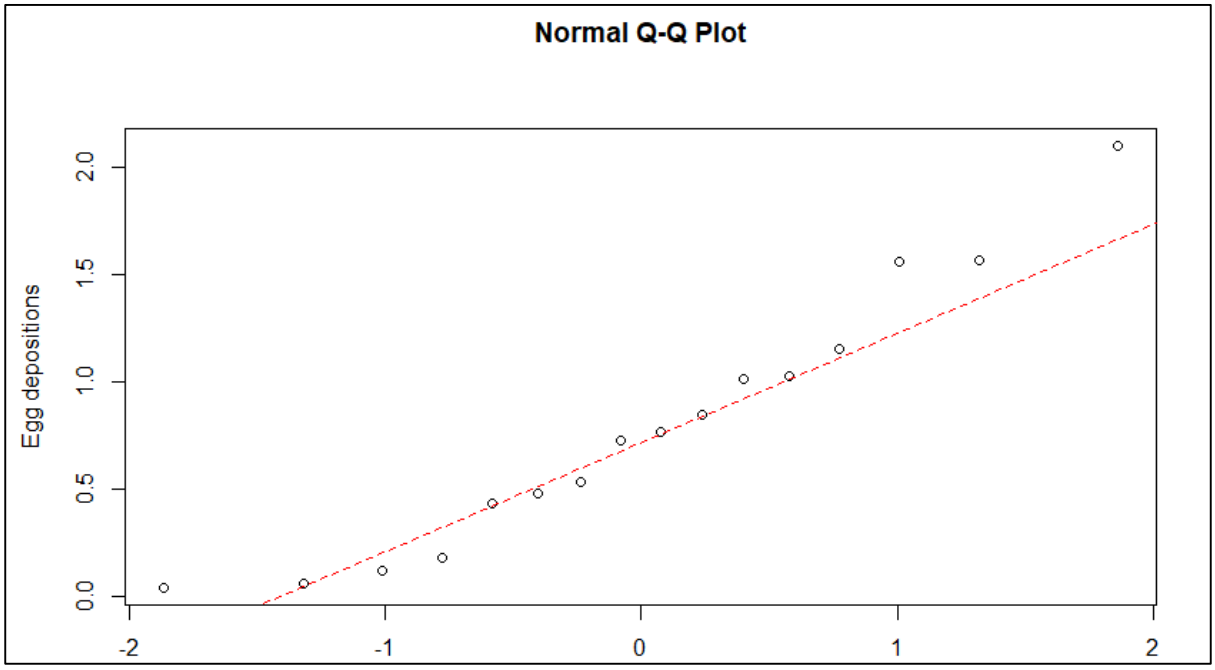


Figure 4 QQ Plot of Original series

Apart from non-stationary behaviour in the time series, ACF and PACF plots in Figure 1 and 3 respectively, the QQ plot of the series in figure 4 represents a slight deviation of data from normality. However, normality test can better reveal whether the data is normal.

Normality Test – Shapiro Wilk Test

```
        Shapiro-Wilk normality test

data:  egg_dep_ts
W = 0.94201, p-value = 0.3744
```

Exhibit 1: Normality Test Results of Original Series

In exhibit 1: the p-value of the Shapiro Wilk normality test 1 is 0.37 at 5% significance level. This indicates that we fail to reject the null hypothesis that the series is normality distributed. However, it should be noted that the test cannot be considered much powerful in case of short sample series.

Box Cox Transformation

Although the series is normally distributed there is a slight changing variance observed in time series in figure 1. To handle changing variance, we tried using box-cox transformation, however, the results of the transformation show a slight change in the variance over the original series. Due to this, it does not seem reasonable to use this complex transformation for such a small change. Thus we expect that the differencing will take care of this small changing variance along with non-stationarity.

Stationarity Test - Augmented Dickey-Fuller Test

This section deals with a check for stationarity in the original time series which is the most important characteristic required for fitting stochastic trend model.

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 0
  STATISTIC:
    Dickey-Fuller: -0.4911
  P VALUE:
    0.452

Description:
 Thu May 14 19:52:39 2020 by user: Avinash Matani
```

Exhibit 2: Stationarity Test Results of Original Series

The DK- Fuller test in exhibit 2 shows p-value = 0.452 at 5% significance level. This shows that we cannot reject the null hypothesis stating that the series is non-stationary. Presence of non-stationarity in the series suggests that the stationary model is inappropriate.

Stationarity is necessary because most of our time series theory applies only to stationary variables and not to non-stationary variables. In the next section, we will see how to convert non-stationary series into stationary for model fitting.

Handling Non-Stationarity – Differencing

This section primarily discuss the conversion of non-stationary series into stationary with the use of a differencing technique.

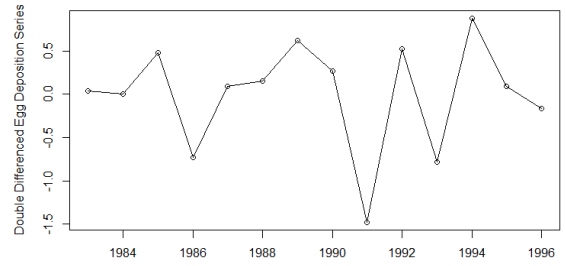
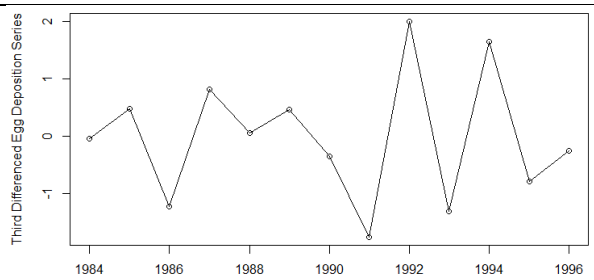
Order of Differencing	Time Series Plot	ADF Test Results after differencing
1		Title: Augmented Dickey-Fuller Test Test Results: PARAMETER: Lag Order: 4 STATISTIC: Dickey-Fuller: -0.7808 P VALUE: 0.3601
2		Title: Augmented Dickey-Fuller Test Test Results: PARAMETER: Lag Order: 4 STATISTIC: Dickey-Fuller: -1.3974 P VALUE: 0.1643
3		Title: Augmented Dickey-Fuller Test Test Results: PARAMETER: Lag Order: 4 STATISTIC: Dickey-Fuller: -0.7284 P VALUE: 0.3767
4		Title: Augmented Dickey-Fuller Test Test Results: PARAMETER: Lag Order: 2 STATISTIC: Dickey-Fuller: -2.1524 P VALUE: 0.03368

Table 1: Differenced Series Plot and ADF Test Results

Table 1: represents the order of differencing, time series plot and ADF test for stationarity. Although, the test series plot of the first differenced series indicates stationarity, the ADF test on the right side shows that the series is non-stationary. Only, the ADF test results of 4 times differenced series suggest that the series is stationary. This allows us to proceed with the 4 times differenced series for model fitting and forecasting applications.

At this point, it is important to note that 4 order differencing for such small series is not much reasonable. It is more reasonable to proceed with four order differencing for a series with large number of observations.

Analysis of Non-Stationary Differenced Series

This section discusses the tentative model selection from the ACF, PACF, EACF and BIC concepts.

ACF and PACF Plot fourth differenced series

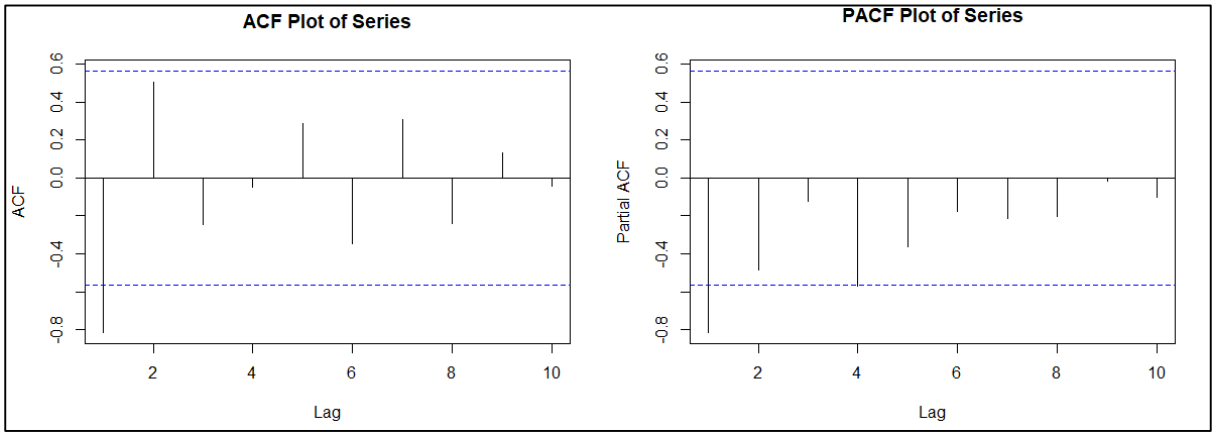


Figure 5 ACF and PACF of four time differenced series

ACF and PACF plots of four differenced series consist of significance at lag 1. From ACF plot it appears that MA (1) is applicable and PACF plot shows that presence of AR (1). Thus, the tentative model can be **ARIMA (1, 4, 1)**.

EACF Plot

AR/MA	0	1	2
0	x	o	o
1	o	o	o
2	o	o	o

Exhibit 3: EACF Plot of four times differenced series

Exhibit 3 shows the EACF plot, we will take the row corresponding to row 1, column 1, row 2, column 2 and row 3 and column 3: i.e. p, q = (0,1), p, q = (1,0), p,q =(2, 1), p,q = (0, 2) as the vertex and include **ARIMA(0, 4, 1), ARIMA(1, 4 , 0), ARIMA(2, 4, 0), ARIMA (2, 4, 1)** and **ARIMA(0, 4, 2)** into the set of possible models.

BIC Table

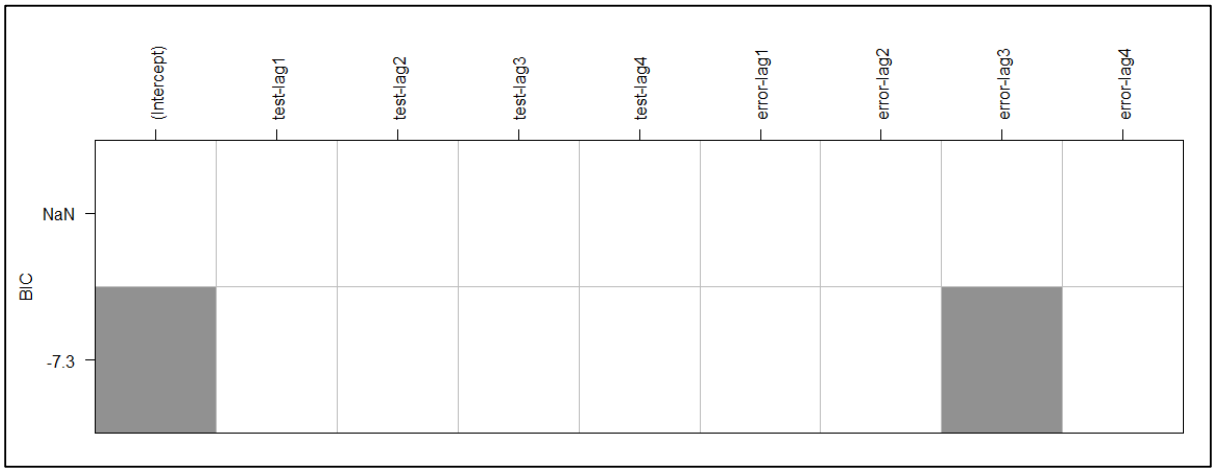


Exhibit 4: BIC Table four times differenced series

The BIC table in exhibit 4 suggest model **ARIMA (0, 4, 3)**. Considering such small order series, we will reject such high order model.

Model Selection – Parameter Estimation

This section deals with estimating the parameters of an ARIMA model based on the observed time series. At this stage, the models selected are already specified. Since the 4 order differenced series is assumed to be stationary, we will only estimate the parameters of these stationary models. Here we will mainly focus on the conditional sum of square estimators and full maximum likelihood estimators.

Although we have included the estimates based on both CSS and ML, we will only base our analysis on ML estimates as in further sections we will be considering AIC and BIC score to finalize the model for forecasting applications.

ARIMA (0,4,1)

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ma1 -0.83608    0.17372 -4.8129 1.487e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Exhibit 5: Parameter Estimation based on CSS

```
z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ma1 -0.94552    0.32317 -2.9257 0.003436 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
```

Exhibit 6: Parameter Estimation based on ML

Exhibit 6 shows that maximum likelihood estimates with standard error and p-value based significance for ARIMA (0, 4, 1). The conditional sum-of-squares estimate is −0.83608 and the maximum likelihood estimate is −0.94552. The standard error for the ML estimate is 0.32317. Considering the magnitude of this standard error and p-value significance, maximum likelihood method estimates reasonably well for ARIMA (0, 4, 1) model.

ARIMA (1,4,1)

```
      Estimate  Std. Error  z value  Pr(>|z|)
ar1 -0.76463313  0.00085452 -894.808 < 2.2e-16 ***
ma1 -2.87773487  0.05071835  -56.739 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.
```

Exhibit 7: Parameter Estimation based on CSS

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.91600    0.10375 -8.8286 < 2.2e-16 ***
ma1 -0.93775    0.34246 -2.7383  0.006176 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.
```

Exhibit 8: Parameter Estimation based on ML

Exhibit 8 shows that maximum likelihood estimates with standard error and p-value based significance for ARIMA (1, 4, 1). The conditional sum-of-squares estimates for AR1 and MA1 are −0.7646 and -0.2877 respectively. The maximum likelihood estimates for AR1 and MA1 are −0.916 and -0.937 with standard error estimated for ML estimates being 0.1037 and 0.342 respectively. Considering the magnitude of these standard errors and p-value significance, maximum likelihood method estimates reasonably well for ARIMA (1, 4, 1) model.

ARIMA (1,4,0)

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -1.02860    0.15699 -6.5521 5.673e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.
```

Exhibit 9: Parameter Estimation based on CSS

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.934594   0.088927  -10.51 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Exhibit 10: Parameter Estimation based on ML

Exhibit 10 shows that maximum likelihood estimates with standard error and p-value based significance for ARIMA (1, 4, 0). The conditional sum-of-squares estimate is −1.028 and the maximum likelihood estimate is −0.934594. The standard error for the ML estimate is 0.8892. Considering the magnitude of this standard error and p-value significance, maximum likelihood method estimates reasonably well for ARIMA (1, 4, 0) model.

ARIMA (2, 4, 0)

```
z test of coefficients:

    Estimate Std. Error  z value  Pr(>|z|)
ar1 -1.91871    0.11410 -16.8165 < 2.2e-16 ***
ar2 -1.22142    0.14469  -8.4415 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Exhibit 11: Parameter Estimation based on CSS

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -1.715202   0.079631 -21.540 < 2.2e-16 ***
ar2 -0.937529   0.070577 -13.284 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Exhibit 12: Parameter Estimation based on ML

Exhibit 12 shows that maximum likelihood estimates with standard error and p-value based significance for ARIMA (2, 4, 0). The conditional sum-of-squares estimates for AR1 and AR2 are −1.918 and -0.2.877 respectively. The maximum likelihood estimates for AR1 and AR2 are −1.715 and -0.937 with standard error estimated for ML estimates being 0.079 and 0.0705 respectively. Considering the magnitude of these standard errors and p-value significance, maximum likelihood method estimates reasonably well for ARIMA (2, 4, 0) model.

ARIMA (2, 4, 1)

```
z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)
ar1 -1.731066   0.024996 -69.2544 < 2.2e-16 ***
ar2 -0.850386   0.014173 -60.0004 < 2.2e-16 ***
ma1 -2.699361   0.434755  -6.2089 5.335e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Exhibit 13: Parameter Estimation based on CSS

```
z test of coefficients:

    Estimate Std. Error  z value  Pr(>|z|)
ar1 -1.67485    0.11165 -15.0014 < 2.2e-16 ***
ar2 -0.89117    0.11445  -7.7865 6.889e-15 ***
ma1 -0.88126    0.45078  -1.9550   0.05058 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Exhibit 14: Parameter Estimation based on ML

Exhibit 14 shows the maximum likelihood estimates with standard error and p-value based significance for ARIMA (2, 4, 1). The conditional sum-of-squares estimates for AR1, AR2, and MA1 are −1.731, -0.850, and -2.699 respectively. The maximum likelihood estimates for AR1, AR2 and MA1 are −1.674, 0.8911 and -0.8812 with standard error estimated for ML estimates being 0.111, 0.114 and 0.450 respectively. Considering the magnitude of these standard errors and p-value significance, maximum likelihood method estimates for MA1 is not reasonably well for ARIMA (2, 4, 1) model. Therefore, we consider it a slight weak model for selection and will explore further during residual analysis and AIC and BIC score estimation.

ARIMA (0, 4, 2)

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ma1 -1.22909    0.28728 -4.2784 1.882e-05 ***
ma2  0.61252    0.25067  2.4435   0.01455 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.(
```

Exhibit 15: Parameter Estimation based on CSS

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ma1 -1.71829    0.41396 -4.1509 3.312e-05 ***
ma2  0.91330    0.38729  2.3582   0.01836 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Exhibit 16: Parameter Estimation based on ML

Exhibit 16 shows that maximum likelihood estimates with standard error and p-value based significance for ARIMA (0, 4, 2). The conditional sum-of-squares estimates for MA1 and MA2 are −1.229 and -0.612 respectively. The maximum likelihood estimates for MA1 and MA2 are −1.718 and -0.913 with standard error estimated for ML estimates being 0.413 and 0.387 respectively. Considering the magnitude of these standard errors and p-value significance, maximum likelihood method estimates reasonably well for ARIMA (0, 4, 2) model.

Model Selection – Residual Analysis

This section deals with the residual analysis of all the above models selected. We will explore various plots of the residuals, check the error terms for constant variance, normality, and independence. The properties of residuals autocorrelation play a significant role in these diagnostics. The Ljung-Box statistic portmanteau plot is also discussed as a summary of autocorrelation in the residuals. In the next section, we will present overfitting and parameter redundancy to check for the selected model.

This section explores the following plots of standardized residuals for all the above models. Standardization allows us to see residuals of unusual size much more easily.

- Time series plot of the standardized residuals
- Histogram plot of Standardized residuals
- Quantile-Quantile plot of standardized residuals
- ACF plot of standardized residuals
- Ljung- Box test plot of standardized residuals

ARIMA (0, 4, 1)

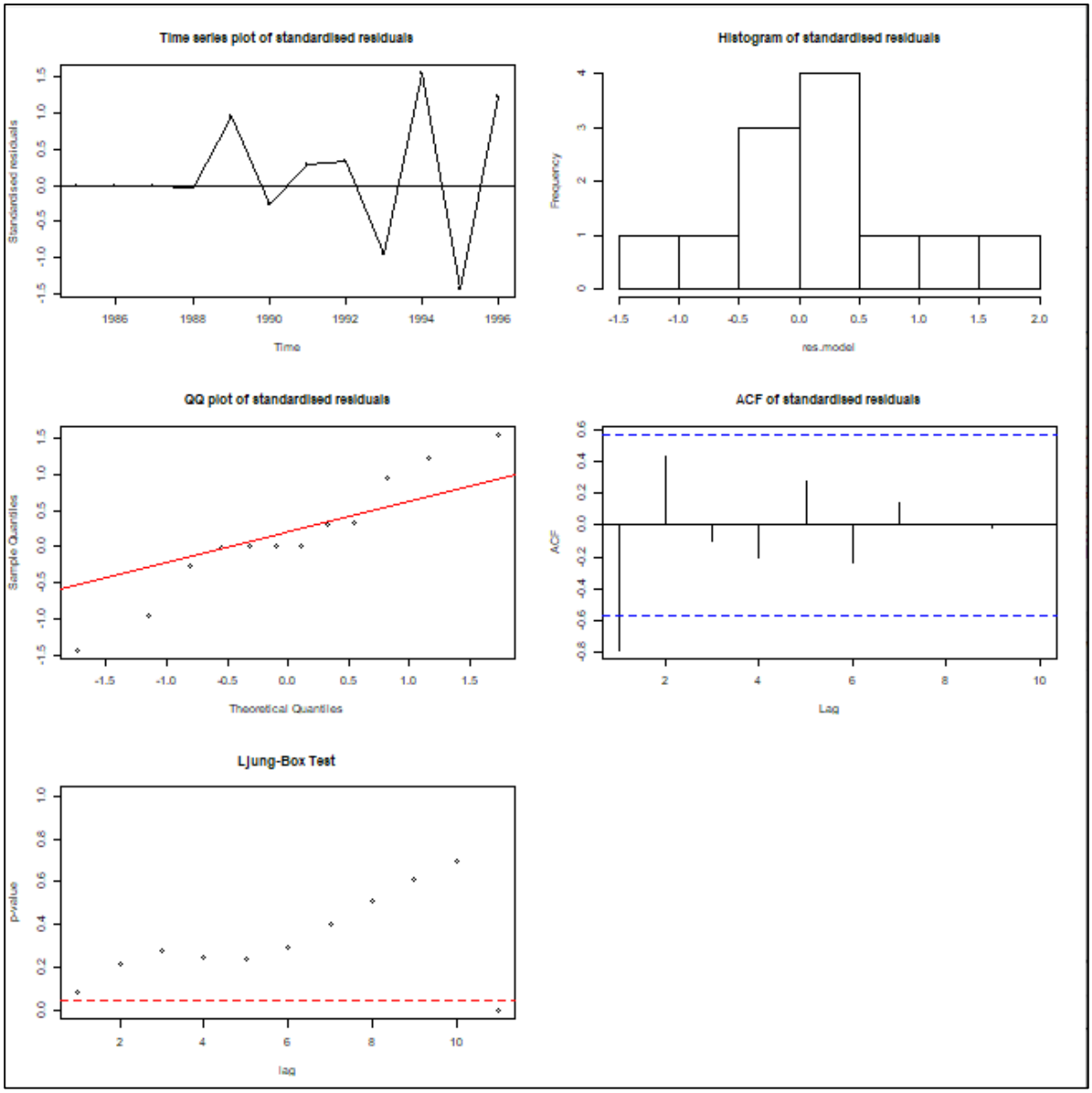


Figure 6 Residual Analysis plots of ARIMA (0, 4, 1)

The time series plot of the residual analysis above suggests a rectangular scatter around a zero horizontal level with no trends whatsoever. This suggests that the model is adequate.

Further, the points in the quantile-quantile plot seems to follow the straight line fairly closely—with some points lying outside the straight line. This graph could lead us to reject normality of the error terms in this model. Here the extreme values look suspect. However, the sample is small (n = 12) and, the Bonferroni criteria for outliers do not indicate a cause for alarm.

A graph of the sample ACF of these residuals is also shown above. The dashed horizontal lines plotted are based on the large lag standard error of ± 2 sqrt(n). There is some evidence of autocorrelation in the residuals of this model as there is a significant correlation at lag 1. Except for significance at lag 1, the model seems to have captured the essence of dependence in the series.

The Ljung – Box test suggest that the p-value for a whole range of lag 1 to 11 seems insignificant, so we do not have statistically significant evidence against the independence of the error terms in this model.

The overall analysis of the plots suggest that the model can be considered for further analysis check i.e. AIC and BIC score analysis.

ARIMA (1, 4, 1)

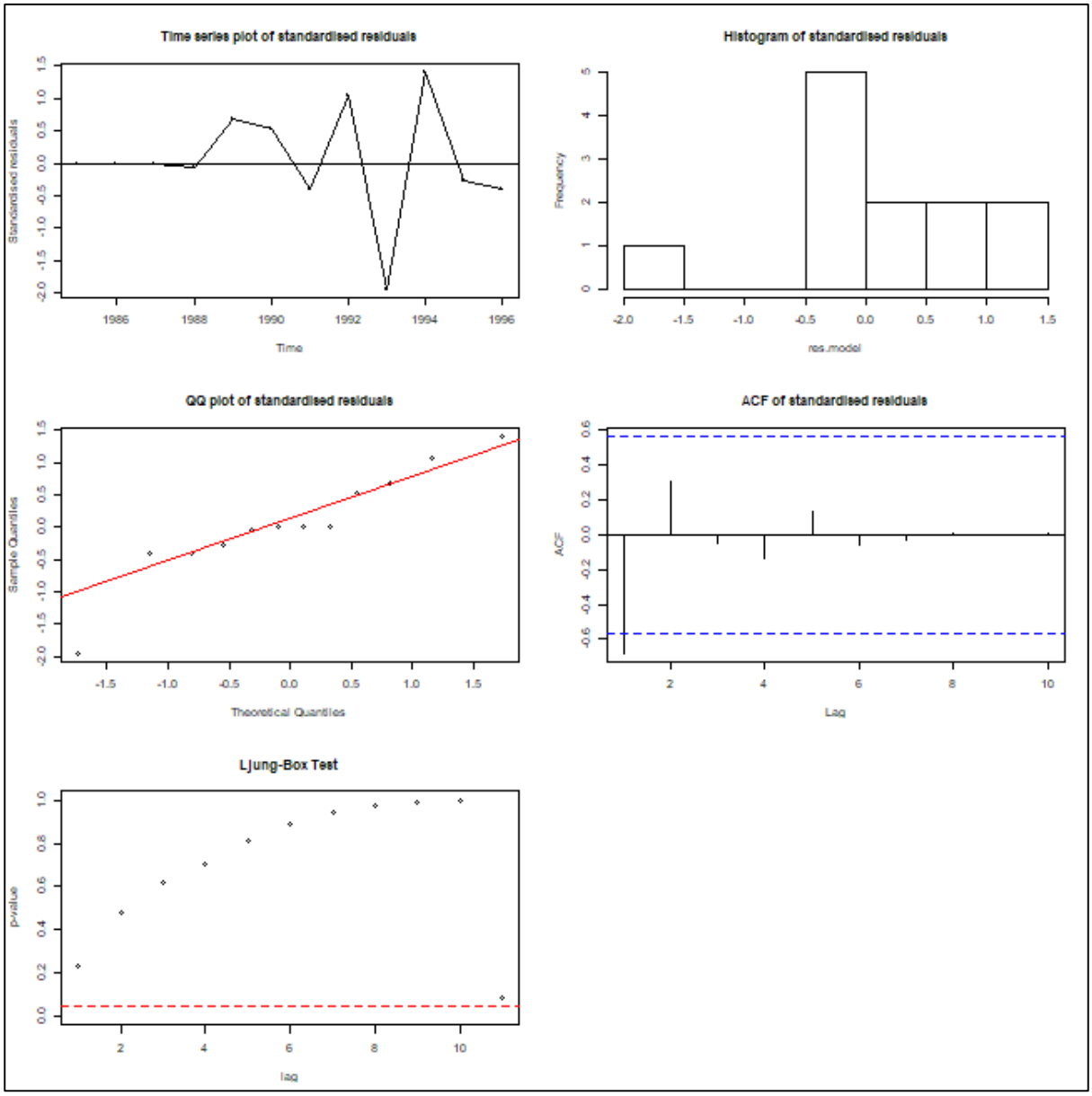


Figure 7 Residual Analysis plots of ARIMA (1, 4, 1)

The time series plot of the residual analysis in Figure 7 suggest a rectangular scatter around a zero horizontal level with no trends whatsoever. This suggests that the model is adequate.

Further, the points in the quantile-quantile plot seems to follow the straight line fairly closely. This graph suggests the normality of the error terms in this model. Here the extreme values look suspect.

However, the sample of the differenced series is small (n = 12) and, the Bonferroni criteria for outliers do not indicate a cause for alarm.

A graph of the sample ACF of these residuals is also shown above. The dashed horizontal lines plotted are based on the large lag standard error of ± 2 sqrt(n). There is some evidence of autocorrelation in the residuals of this model as there is a significant correlation at lag 1. Except for significance at lag 1, the model seems to have captured the essence of dependence in the series.

The Ljung – Box test suggest that the p-value for a whole range of lag 1 to 11 seems insignificant, so we do not have statistically significant evidence against the independence of the error terms in this model.

The overall analysis of the plots suggest that this model can also be considered for further analysis check i.e. AIC and BIC score analysis.

ARIMA (1, 4, 0)

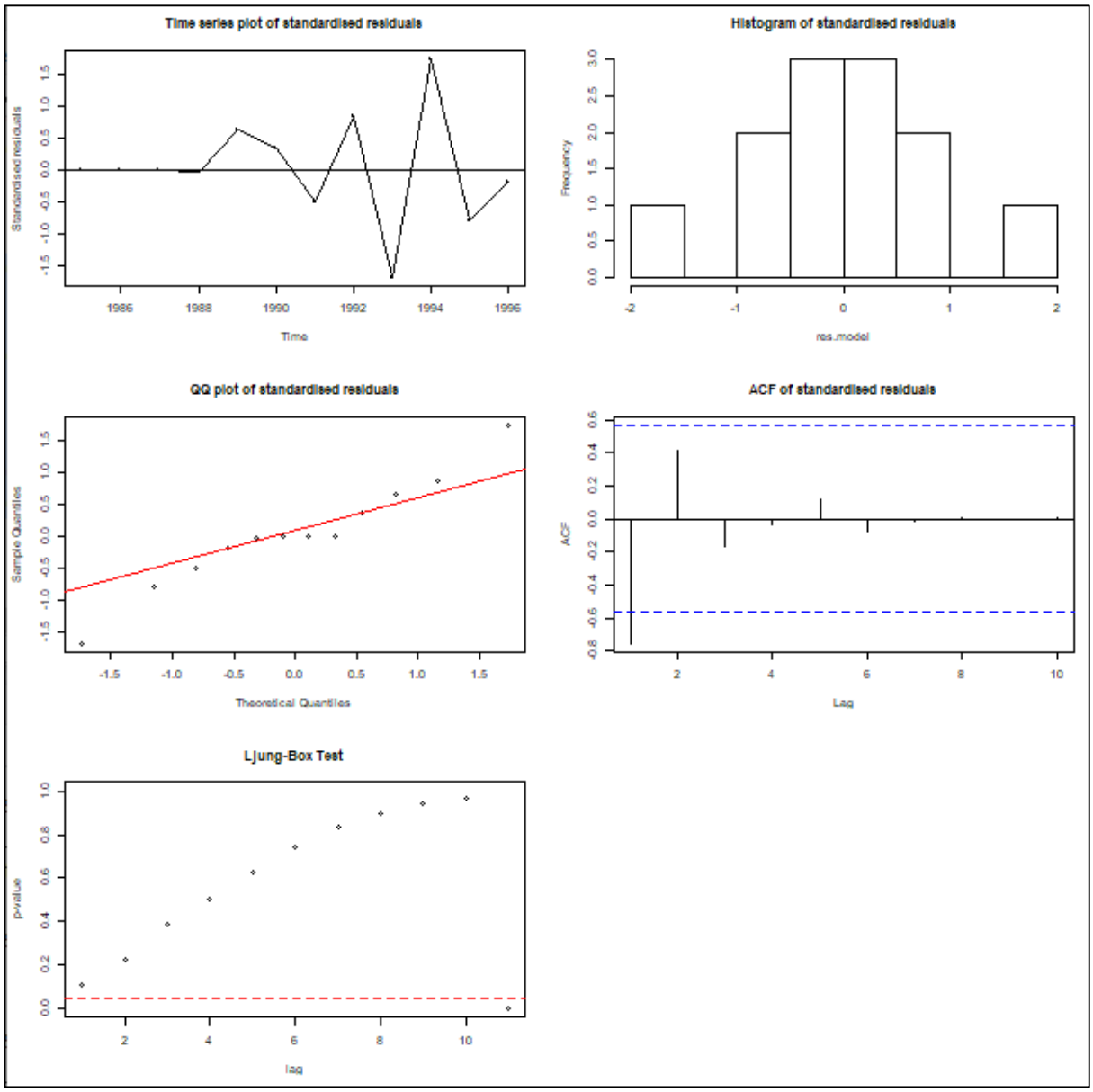


Figure 8 Residual Analysis plots of ARIMA (1, 4, 0)

The time series plot of the residual analysis in Figure 8 suggests a rectangular scatter around a zero horizontal level with no trends whatsoever. This suggests that the model is adequate.

Further, the points in the quantile-quantile plot seems to follow the straight line fairly closely—with some points lying outside the straight line. This graph could lead us to reject normality of the error terms in this model. Here the extreme values look suspect. However, the sample is small (n = 12) and, the Bonferroni criteria for outliers do not indicate a cause for alarm.

A graph of the sample ACF of these residuals is also shown above. The dashed horizontal lines plotted are based on the large lag standard error of ± 2 sqrt(n). There is some evidence of autocorrelation in the residuals of this model as there is a significant correlation at lag 1. Except for significance at lag 1, the model seems to have captured the essence of dependence in the series. This suggests that the model is weak, but we will consider this in the AIC, BIC score analysis.

The Ljung – Box test suggest that the p-value for a whole range of lag 1 to 11 seems insignificant, so we do not have statistically significant evidence against the independence of the error terms in the model.

The overall analysis of the plots suggest that the model can be considered for further analysis check i.e. AIC and BIC score analysis.

ARIMA (2, 4, 0)

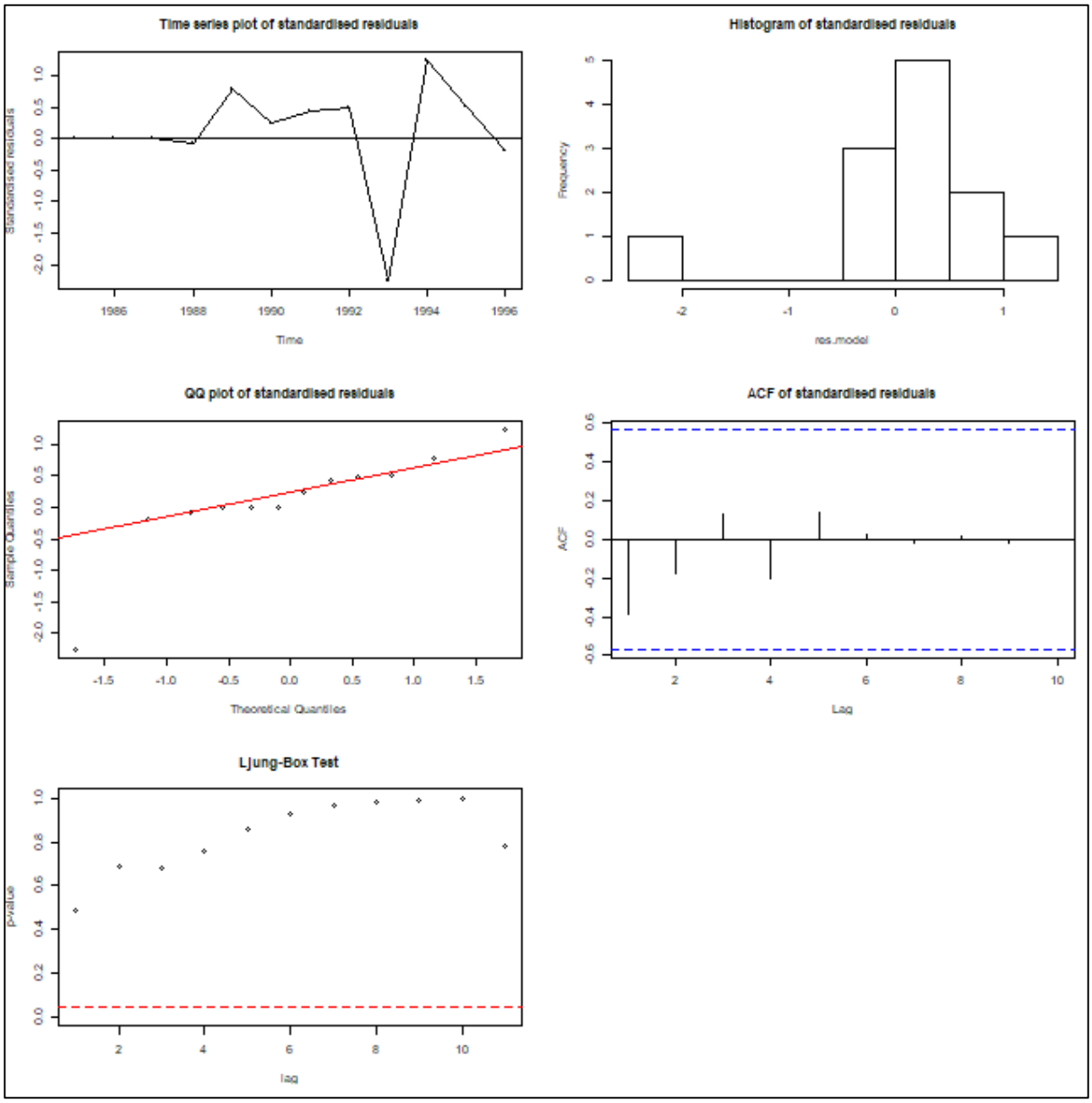


Figure 9 Residual Analysis plots of ARIMA (2, 4, 0)

The time series plot of the residual analysis above suggests a rectangular scatter around a zero horizontal level with no trends whatsoever. This suggests that the model is adequate.

Further, the points in the quantile-quantile plot seems to follow the straight line fairly closely—with one or two points lying outside the straight line. This graph could lead us to reject normality of the error terms in this model. Here the extreme values look suspect. However, the sample is small (n = 12) and, the Bonferroni criteria for outliers do not indicate a cause for alarm.

A graph of the sample ACF of these residuals is also shown above. The dashed horizontal lines plotted are based on the large lag standard error of ± 2 sqrt(n). There is no evidence of autocorrelation in the residuals of this model as it represents a white noise series.

The Ljung – Box test suggest that the p-value for a whole range of lag 1 to 11 seems insignificant, so we do not have statistically significant evidence against the independence of the error terms in this model.

The overall analysis of the plots suggests that the model can be considered for further analysis check i.e. AIC and BIC score analysis. Also, the coefficient estimates of the model as presented in exhibit 12 suggest that the model is best so far for forecast application. The model will be further analysed in the AIC, BIC score analysis to investigate its relevance over other models.

ARIMA (2, 4, 1)

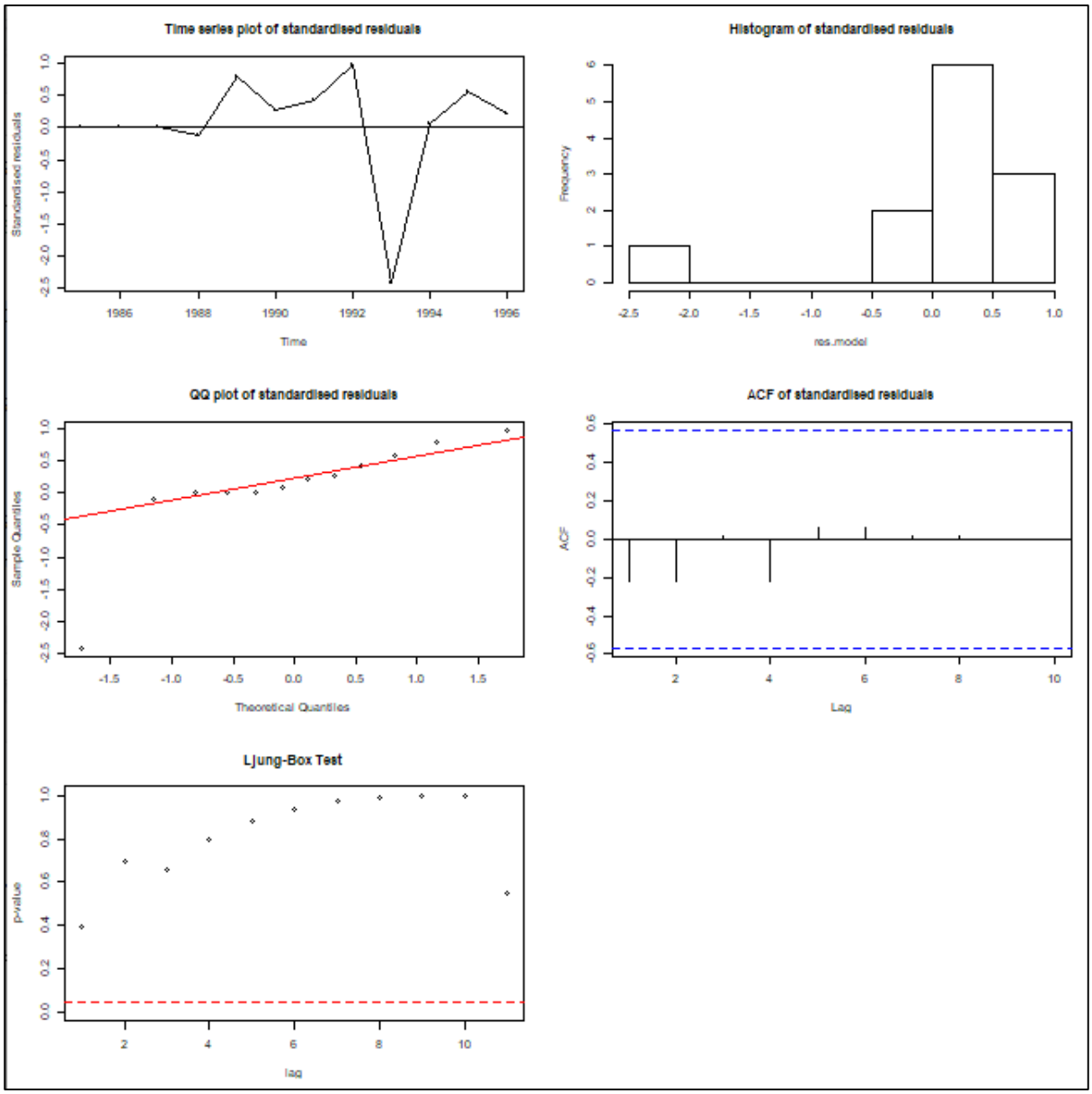


Figure 10 Residual Analysis plots of ARIMA (2, 4, 1)

The time series plot of the residual analysis in figure 10 does not represent an adequate rectangular scatter around a zero horizontal level with a sign of some trends in the series. This suggests that the model is inadequate.

However, the points in the quantile-quantile plot seem to follow a straight line fairly closely—with some points lying outside the straight line. This graph could lead us to reject normality of the error terms in this model. Here the extreme values look suspect. However, the sample is small (n = 12) and, the Bonferroni criteria for outliers do not indicate a cause for alarm.

A graph of the sample ACF of these residuals is also shown above. The dashed horizontal lines plotted are based on the large lag standard error of ± 2 sqrt(n). There is no evidence of autocorrelation in the residuals of this model as it represents a white noise series.

The Ljung – Box test suggest that the p-value for a whole range of lag 1 to 11 seems insignificant, so we do not have statistically significant evidence against the independence of the error terms in this model.

The overall analysis of the plots suggests that the model cannot be considered for further analysis check i.e. AIC and BIC score analysis. Also, the coefficient estimates of the model as presented in exhibit 14 suggest that MA1 coefficient is insignificant.

ARIMA (0, 4, 2)

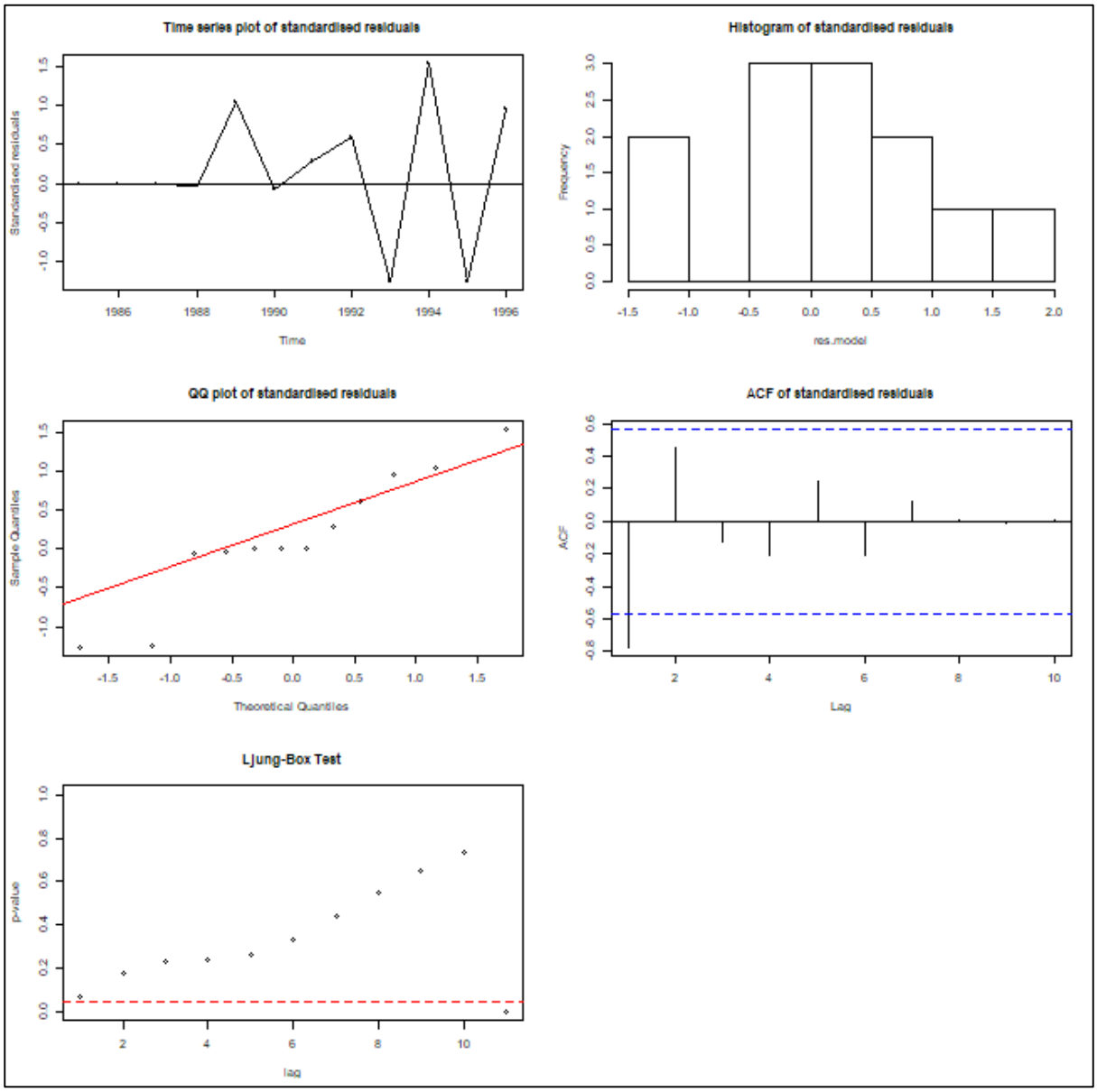


Figure 11 Residual Analysis plots of ARIMA (0, 4, 2)

The time series plot of the residual analysis in figure 11 suggest a rectangular scatter around a zero horizontal level with no trends whatsoever. This suggests that the model is adequate.

Further, the majority of points in the quantile - quantile plot does not seem to follow the straight line—with many points lying outside the straight line. This graph leads us to reject normality of the error terms in this model. Here the extreme values look suspect.

A graph of the sample ACF of these residuals is also shown above. The dashed horizontal lines plotted are based on the large lag standard error of ± 2 sqrt(n). There is some evidence of autocorrelation in the residuals of this model as there is a significant correlation at lag 1. Except for significance at lag 1, the model seems to have captured the essence of dependence in the series. This suggests that the model is weak, but we will consider this in the AIC, BIC score analysis.

The Ljung – Box test suggest that the p-value for a whole range of lag 1 to 11 seems insignificant, so we do not have statistically significant evidence against the independence of the error terms in this model. However, at lag 1 we seem to have a significant result.

Although the overall analysis of the plots suggest that the model is inadequate but owing to the short nature of the series we can consider this for further analysis check i.e. AIC and BIC score analysis.

Model Selection Using AIC and BIC Scores

This section compares the AIC and BIC score for the selected models from the above methods. The AIC and BIC is a probabilistic model selection criterion which provides an analytical technique for scoring and choosing among several candidate models.

	df	AIC
model_240_ml	3	64.54505
model_141_ml	3	69.99093
model_140_ml	2	74.89835
model_042_ml	3	75.10802
model_041_ml	2	79.93560

Exhibit 17: AIC Score Model Comparison

	df	BIC
model_240_ml	3	64.78338
model_141_ml	3	70.22926
model_140_ml	2	75.05724
model_042_ml	3	75.34635
model_041_ml	2	80.09448

Exhibit 18: BIC Score Model Comparison

Both AIC and BIC score analysis in exhibit 17 and 18 respectively suggest that ARIMA (2, 4, 0) is the most adequate model for forecast application of egg deposition series. This is because both AIC and BIC score is lowest for ARIMA (2, 4, 0). The overfitting analysis below can also be used to provide support for the selected model.

Model Selection - Overfitting Analysis of Selected Models

This section deals with specifying and fitting a more general model, a model "close by" that contains the ARIMA (2, 4, 0) as a special case. The comparison is made with ARIMA (3, 4, 0).

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -2.04185    0.33443 -6.1055 1.025e-09 ***
ar2 -1.52926    0.59662 -2.5632   0.01037 *
ar3 -0.34491    0.35295 -0.9772   0.32845
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Exhibit 19: Comparison of coefficient estimates of ARIMA (3, 4, 0) with ARIMA (2, 4, 0)

The estimates for the additional parameter in exhibit 19 is not significantly different from zero.

	df	AIC
model_240_ml	3	64.54505
model_340_ml	4	65.68846

Exhibit 20: AIC Score Model Comparison

	df	BIC
model_240_ml	3	64.78338
model_340_ml	4	66.00622

Exhibit 21: BIC Score Model Comparison

Exhibit 20 and 21 shows the model comparison based on AIC and BIC score. As expected the AIC and BIC score for ARIMA (2, 4, 0) is much lower. The above two points confirm the selection of ARIMA (2, 4, 0) and we can further use this for forecast applications.

Forecast Analysis

```
$pred
Time Series:
Start = 1997
End = 2001
Frequency = 1
[1]  2.027925  1.257079  9.912927  9.784411 22.247739

$se
Time Series:
Start = 1997
End = 2001
Frequency = 1
[1]  4.753668 11.855903 27.173663 50.826434 86.330530
```

Exhibit 22: Egg Deposition (in millions) Forecast Results from year 1997 to 2001

Exhibit 22 shows the standard error and prediction estimation from year 1997 – 2001.

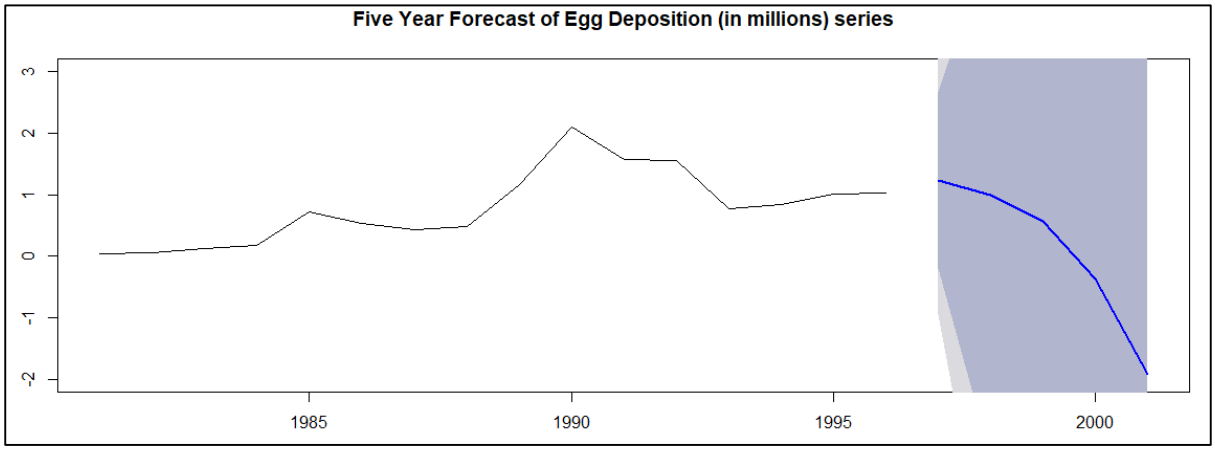


Figure 12: Prediction plot of Egg Deposition (in millions) from year 1997 to 2001

Figure 12: shows the forecast and prediction plot of the egg deposition (in millions) respectively from year 1997-2001. Shaded region indicates the confidence interval of the predicted values.

Conclusion

We successfully described, modelled and estimated the stochastic trend models for time series dataset which represent egg deposition (in million) from the year 1981 to 1996. As a part stochastic trend model analysis, we successfully demonstrated the use of identification tools including ACF, PACF, EACF and BIC plots and differencing and series transformation concepts to suggest **ARIMA (0, 4, 1), ARIMA (1, 4, 1), ARIMA (1, 4, 0), ARIMA (2, 4, 1), ARIMA (2, 4, 0)** and **ARIMA (0, 4, 2)** as a set of possible models. Further, based on model diagnostic, parameter estimation, residual analysis, overfitting and parameter redundancy analysis we finally select ARIMA **(2, 4, 0)** for forecast application.

References

D.Cryer, J., & Chan, K.-S. (2008). Time Series Analysis with applications in R. Springer.

Dr. Kalaylioglu. Module 1 - Basic Plots, Examples, and Fundamental Concepts, RMIT University, School of Science

Dr. Kalaylioglu. Module 2 - Analysis of Trends, RMIT University, School of Science

Dr. Kalaylioglu. Module 3 - Models for Stationary Time series, RMIT University, School of Science

Dr. Kalaylioglu. Module 4 - Models for Nonstationary Time Series, RMIT University, School of Science

Dr. Kalaylioglu. Module 5 - Model Specification, RMIT University, School of Science

Dr. Kalaylioglu. Module 6 -Parameter Estimation, RMIT University, School of Science

Dr. Kalaylioglu. Module 7 - Model Diagnostics, RMIT University, School of Science

Appendix:

This sections contains R software for the above analysis

Load Package, Viewing File and Series Conversion

```
library(TSA)

library(fUnitRoots)

library(lmtest)

egg_dep <- read.csv("eggs.csv")

View(egg_dep)

egg_dep_ts = ts(egg_dep[2], start = 1981)

class(egg_dep_ts)

plot(egg_dep_ts,type="o",main = "Time series plot of eggs deposition over years ",xlab="Year",

    ylab = "Egg deposits (Millions)")

plot(y=egg_dep_ts,x=zlag(egg_dep_ts),ylab="Change in egg deposition in (millions)",

    xlab="Previous Year egg deposition in Millions",

    main = "Scatter plot of Egg deposition")
```

ACF, PACF Plots

```
acf(egg_dep_ts,main="ACF plot")

pacf(egg_dep_ts,main="PACF plot")
```

Normality Test

```
qqnorm(egg_dep_ts, ylab="Egg depositions", xlab="Normal Scores")

qqline(egg_dep_ts, col = 2, lwd = 1, lty = 2)

shapiro.test(egg_dep_ts)
```

Box Cox Transformation

```
egg_dep.transform = BoxCox.ar(egg_dep_ts)

egg_dep.transform = BoxCox.ar(egg_dep_ts, method = "yule-walker")

egg_dep.transform$ci

lambda = 0.45

BC.egg_dep_ts = (egg_dep_ts^lambda-1)/lambda

qqnorm(BC.egg_dep_ts)

qqline(BC.egg_dep_ts, col = 2)

BC.egg_dep_ts
```

```
plot(BC.egg_dep_ts,type='o',ylab='Time series plot of transformed Egg Deposition Series')
```

Differencing and ADF Test

```
order = ar(diff(egg_dep_ts))$order

adfTest(egg_dep_ts, lags = order,  title = NULL,description = NULL)

# non-stationary (and becomes stationary with differencing)

# The output shows that the data is non stationary.


diff.egg_dep_ts = diff(egg_dep_ts,differences = 1)

plot(diff.egg_dep_ts,type='o',ylab='Single Differenced Egg Deposition Series')


order = ar(diff(diff.egg_dep_ts))$order

adfTest(diff.egg_dep_ts, lags = order,  title = NULL,description = NULL)

# The test shows that the data is non stationary


diff.egg_dep_ts = diff(egg_dep_ts,differences = 2)

plot(diff.egg_dep_ts,type='o',ylab='Double Differenced Egg Deposition Series')


order = ar(diff(diff.egg_dep_ts))$order

adfTest(diff.egg_dep_ts, lags = order,  title = NULL,description = NULL)

# Non Stationary after second differencing as well


diff.egg_dep_ts = diff(egg_dep_ts,differences = 3)

plot(diff.egg_dep_ts,type='o',ylab='Third Differenced Egg Deposition Series')


order = ar(diff(diff.egg_dep_ts))$order

adfTest(diff.egg_dep_ts, lags = order,  title = NULL,description = NULL)

# Non Stationary after 3 differecing as well


diff.egg_dep_ts = diff(egg_dep_ts,differences = 4)

plot(diff.egg_dep_ts,type='o',ylab='Fourth Differenced Egg Deposition Series')
```

```
order = ar(diff(diff.egg_dep_ts))$order

adfTest(diff.egg_dep_ts, lags = order,  title = NULL,description = NULL)

# Stationary after 4th differecing
```

ACF and PACF Plot

```
par(mfrow=c(1,2))

acf(diff.egg_dep_ts, main = 'ACF Plot of Series')

pacf(diff.egg_dep_ts, main = 'PACF Plot of Series')

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

EACF Plot

```
eacf(diff.egg_dep_ts,ar.max=3,ma.max=3)
```

BIC Table

```
res = armasubsets(y=diff.egg_dep_ts,nar=4,nma=4,y.name='test',ar.method='ols')

plot(res)
```

Residual Analysis

```
# Residual Analysis

residual.analysis <- function(model, std = TRUE){

 library(TSA)

 library(FitAR)

 if (std == TRUE){

  res.model = rstandard(model)

 }else{

  res.model = residuals(model)

 }

 par(mfrow=c(3,2))

 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")
```

```
  print(shapiro.test(res.model))
  k=0
  LBQPlot(res.model, lag.max = length(model$residuals)-1 , StartLag = k + 1, k = 0, SquaredQ = FALSE)
  par(mfrow=c(1,1))
}
```

Model and Coeff. Test

```
# ARIMA(1,4,1)

ARIMA(0,4,1)
model_041_css = arima(diff.egg_dep_ts,order=c(0,4,1),method='CSS')
coeftest(model_041_css)

model_041_ml = arima(diff.egg_dep_ts,order=c(0,4,1),method='ML')
coeftest(model_041_ml)

par(mfrow=c(1,1))
win.graph(width=10, height=10,pointsize=8)
residual.analysis(model = model_041_ml)
# Rejected: Significant correlation at lag 1

#ARIMA(1,4,1)
model_141_css = arima(diff.egg_dep_ts,order=c(1,4,1),method='CSS')
coeftest(model_141_css)

model_141_ml = arima(diff.egg_dep_ts,order=c(1,4,1),method='ML')
coeftest(model_141_ml)

par(mfrow=c(1,1))
win.graph(width=10, height=10,pointsize=8)
residual.analysis(model = model_141_ml)
```

```
#ARIMA(1,4,0)
model_140_css = arima(diff.egg_dep_ts,order=c(1,4,0),method='CSS')
coeftest(model_140_css)

model_140_ml = arima(diff.egg_dep_ts,order=c(1,4,0),method='ML')
coeftest(model_140_ml)

par(mfrow=c(1,1))
win.graph(width=10, height=10,pointsize=8)
residual.analysis(model = model_140_ml)
# Rejected as significant correlationa at lag 1

#ARIMA(2,4,0)
model_240_css = arima(diff.egg_dep_ts,order=c(2,4,0),method='CSS')
coeftest(model_240_css)

model_240_ml = arima(diff.egg_dep_ts,order=c(2,4,0),method='ML')
coeftest(model_240_ml)

par(mfrow=c(1,1))
win.graph(width=10, height=10,pointsize=8)
residual.analysis(model = model_240_ml)

#ARIMA(2,4,1)
model_241_css = arima(diff.egg_dep_ts,order=c(2,4,1),method='CSS')
coeftest(model_241_css)

model_241_ml = arima(diff.egg_dep_ts,order=c(2,4,1),method='ML')
coeftest(model_241_ml)
par(mfrow=c(1,1))
win.graph(width=10, height=10,pointsize=8)
```

```
residual.analysis(model = model_241_ml)
```

AIC and BIC Score

```
# AIC and BIC values Function

# you need to source the sort.score() function, which is available in Canvas shell

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 arguements ("aic","bic")')


  }

}

## Checking model score

sort.score(AIC(model_041_ml,model_141_ml,model_140_ml,model_240_ml, model_042_ml), score = "aic")

sort.score(BIC(model_041_ml,model_141_ml,model_140_ml,model_240_ml, model_042_ml), score = "bic" )
```

Overfitting Analysis

```
# Overfitting: To further assess the selected model ARIMA(2,4,0) by overfitting

# ARIMA(3,4,0) and ARIMA(2,4,1)


# ARIMA(3,4,0)

model_340_ml = arima(diff.egg_dep_ts,order=c(3,4,0),method='ML')

coeftest(model_340_ml)

# Comparing its AIC and BIC with  the selected model ARIMA(240)

sort.score(AIC(model_240_ml,model_340_ml), score = "aic")

sort.score(BIC(model_240_ml,model_340_ml), score = "bic" )
```

Forecasting

```
forecasting = Arima(egg_dep_ts,c(2,4,0))

plot(forecast(forecasting, h=5), ylim =c(-2,3), main = 'Five Year Forecast of Egg Deposition (in millions) series')
```