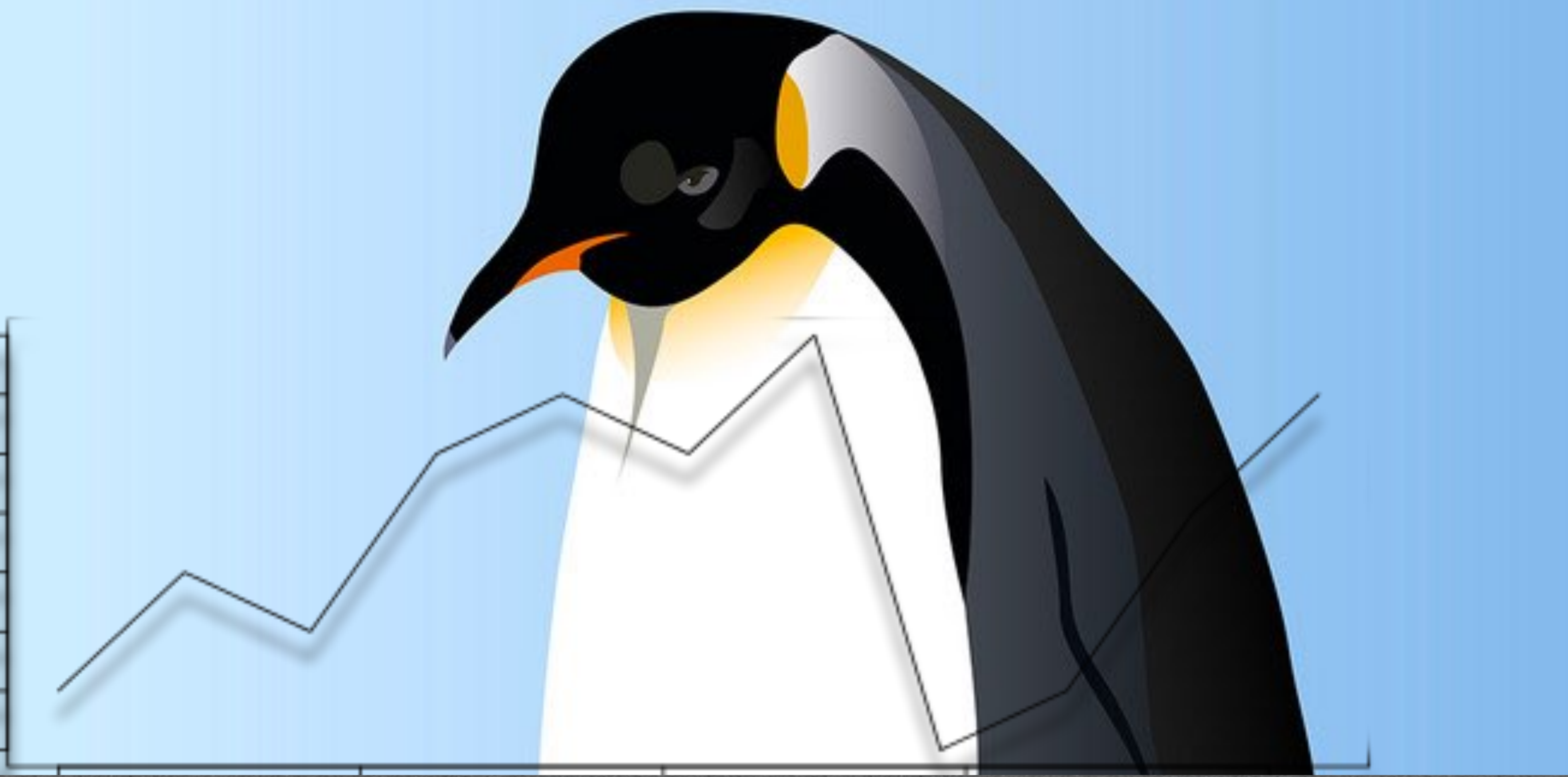
MATH1318 - Time Series Analysis Final Project Report



Time Series Analysis on Forecasting Penguin Counts

8 June 2020

By

- Jewel James

INDEX

ABSTRACT	Page 3
INTRODUCTION	Page 3
METHODOLOGY	Page 3
-GENERAL ANALYSIS	Page 3
-ACF & PACF	Page 5
FIRST DIFFERENCING	Page 5
SECOND DIFFERENCING	Page 6
MODEL FITTING	Page 6
COEFFICIENT TEST (ML)	Page 7
- SARIMA(0,1,1)x(1,1,1)	Page 7
- SARIMA(2,1,1)x(1,1,1)	Page 8
- SARIMA(2,1,2)x(1,1,1)	Page 8
SORTING BY AIC AND BIC	Page 9
MODEL DIAGNOSTICS	Page 10
FORECASTING	Page 11
CONCLUSION	Page 11
APPENDIX	Page 12
REFERENCES	Page 16

ABSTRACT

The report aims to find the best fitting model which can analyse and predict the penguin count by acquiring the trend in data and using various. approaches and hypothesis tests . Analysis showed the data to be non stationary therefore, appropriate steps were taken to coordinate them with stationarity .Various models are tested using EACF, AIC and BIC functions and the finest fitting ARIMA model is established. Finally forecasting is done to predict the penguin population from 2006 to 2016, using the selected ARIMA model.

INTRODUCTION

Penguins are aquatic birds, that inhabit mostly in the southern hemisphere. They have adapted flippers, which help them to swim under water, and are among they few species that can drink sea water. The penguin colony was reported to have declined 77% in last 50 years. An accurate prediction in the penguin count can provide us with numerous possibilities to design measures to conserve the ecosystem, create awareness among general public regarding the future conditions of the specie if current trend continues.Time Series Analysis techniques ideally provide us means to analyse and predict the penguin count for the coming 10 years from the chosen data employing statistical and descriptive tools.

METHODOLOGY

The dataset contains the penguin counts within the years 2000 and 2006 . Census at School website provided the count of data in from of monthly data. It contains two columns namely month and number. Initial analysis is done and the number column value is converted to time series object. Observations are made to determine seasonal, trend based on that further analysis is made.

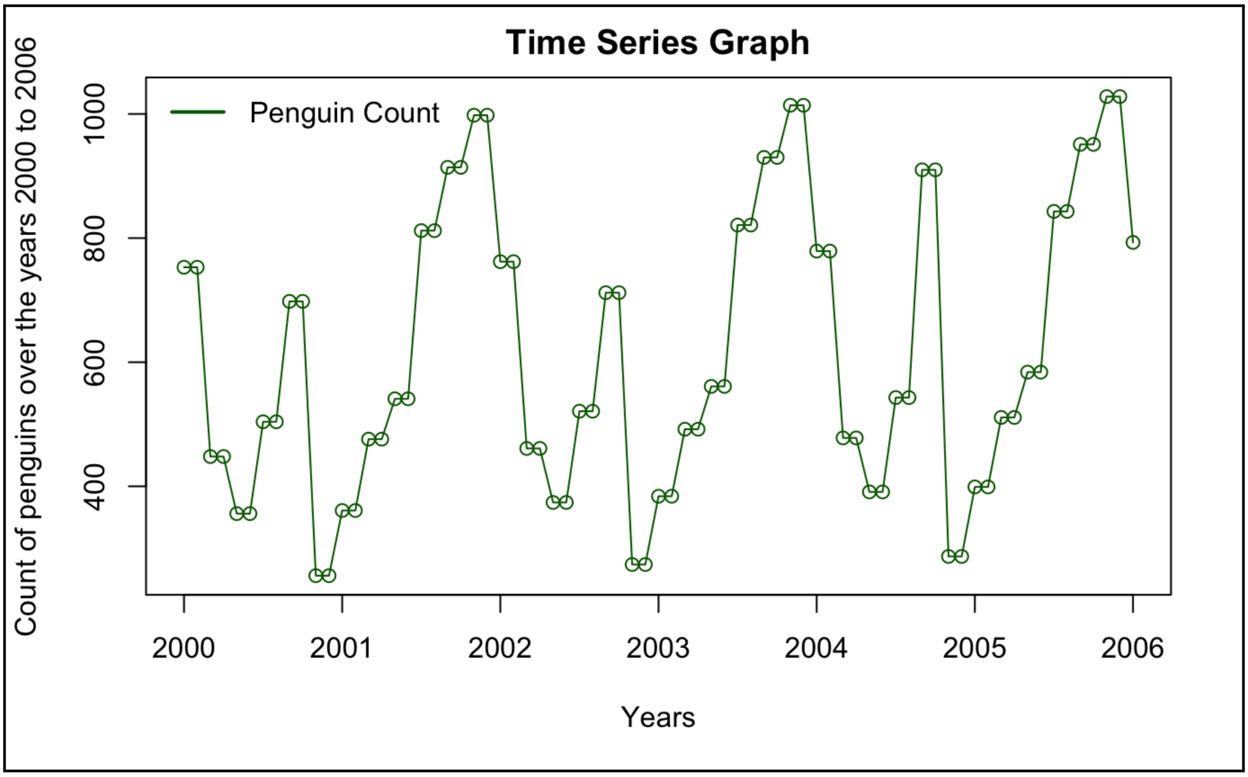


Fig 1.1

GENERAL ANALYSIS

From the figure (Fig: 1.1), we can observe there is seasonality in the data . Basic concepts we an infer from the above plot is. The behaviour of the series cannot be inferred clearly due to the seasonality present in the data.

I. Trend - There is a downward trend shown in the plot, which shows decreasing mean level.
II. Change in Variation - In primary analysis no obvious variation change can be inferred.
III. Seasonality - Repeating patterns denotes seasonality and its evidently present in our data, which is obvious since our data is monthly .
IV. Auto Correlation structure - This performance is seen by continuous points , there are series of continuous points depicted here,
V. Intervention - This is denoted by sudden movement changes. Here no clear intervention points are implied.

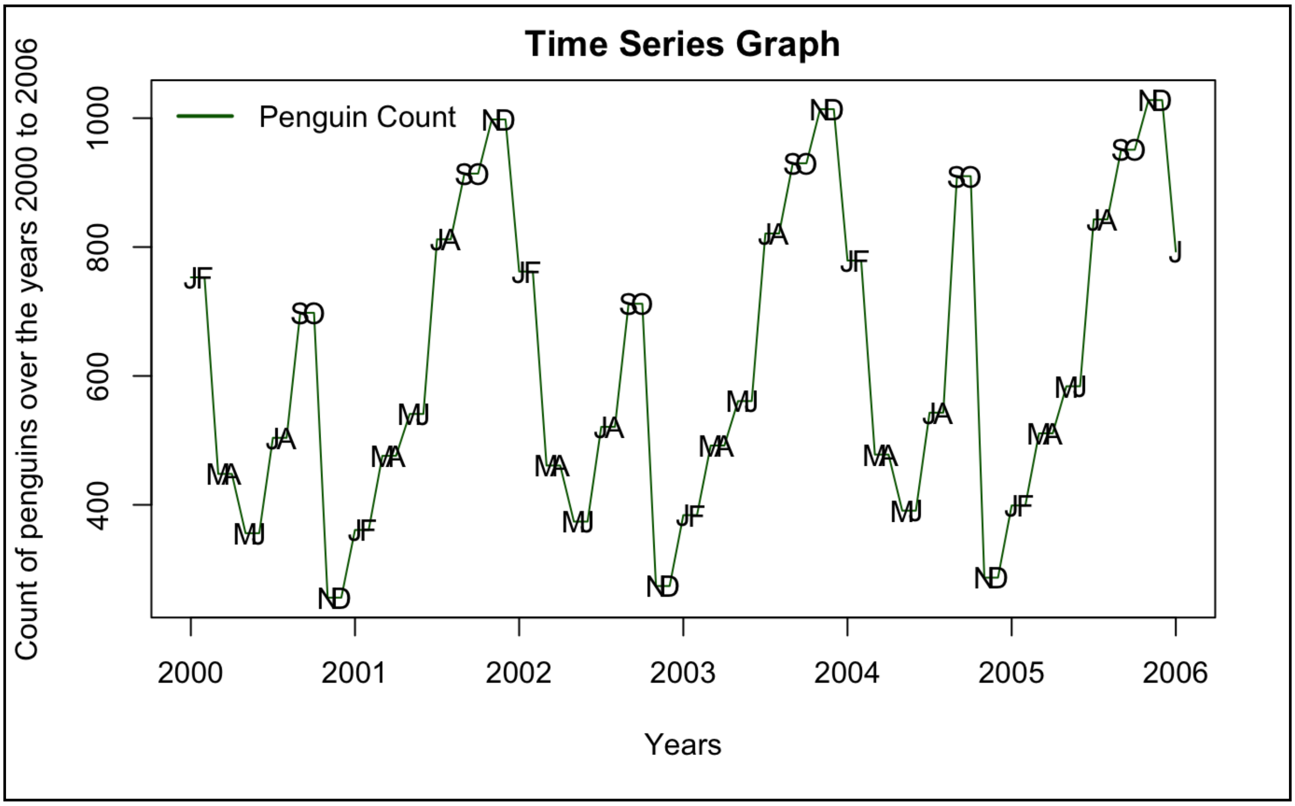
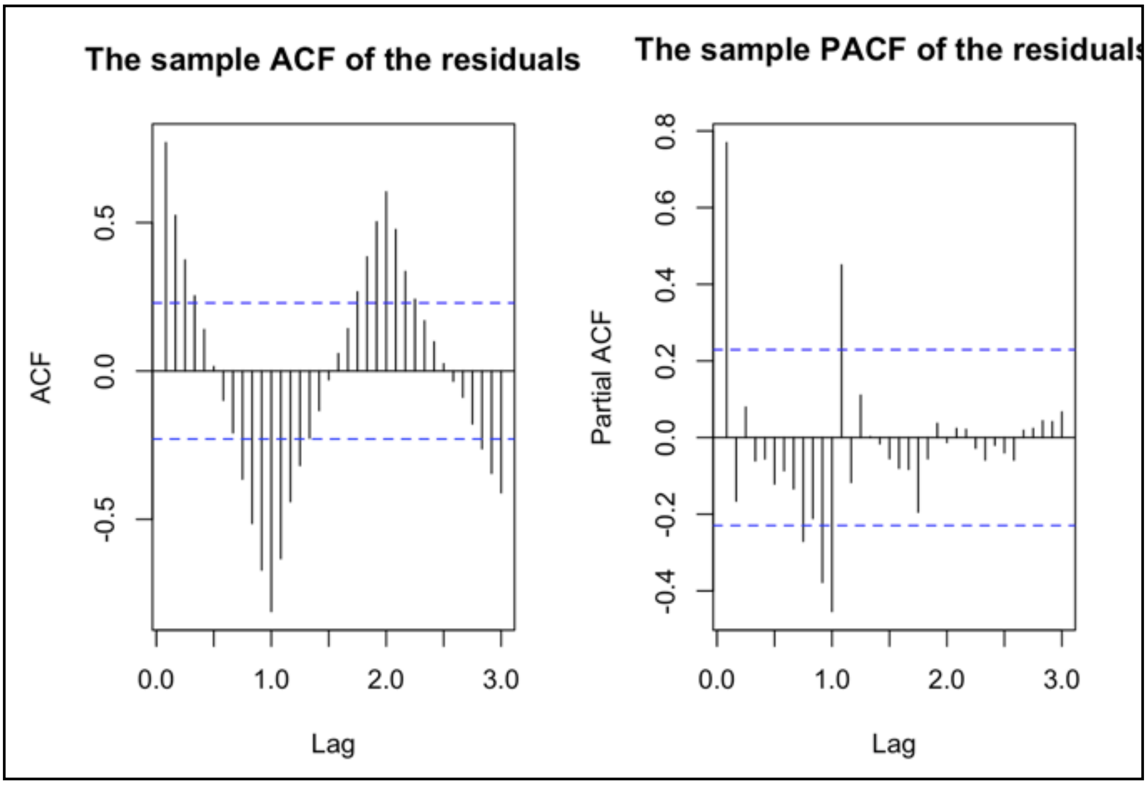


Fig 1.2

Figure 1.2 provides the labelled form of the time series graph to show seasonality with the months labels to show trend distribution within the months.



ACF & PACF

For further analysis on the behaviour of the given data we plot the ACF and PACF graphs.

From the above plots (Fig 2.1) we can clearly see that there is an existence of trend and seasonality. First fit a plain model with only the first seasonal difference with order D = 1 to observe if we can get rid of the seasonal trend effect.

FIRST DIFFERENCING

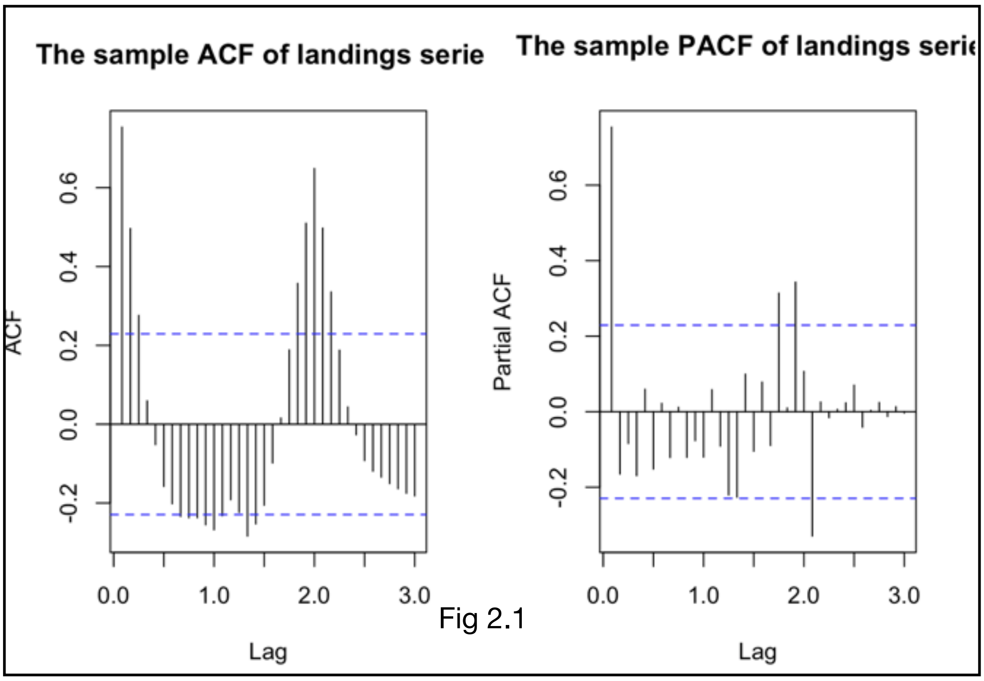
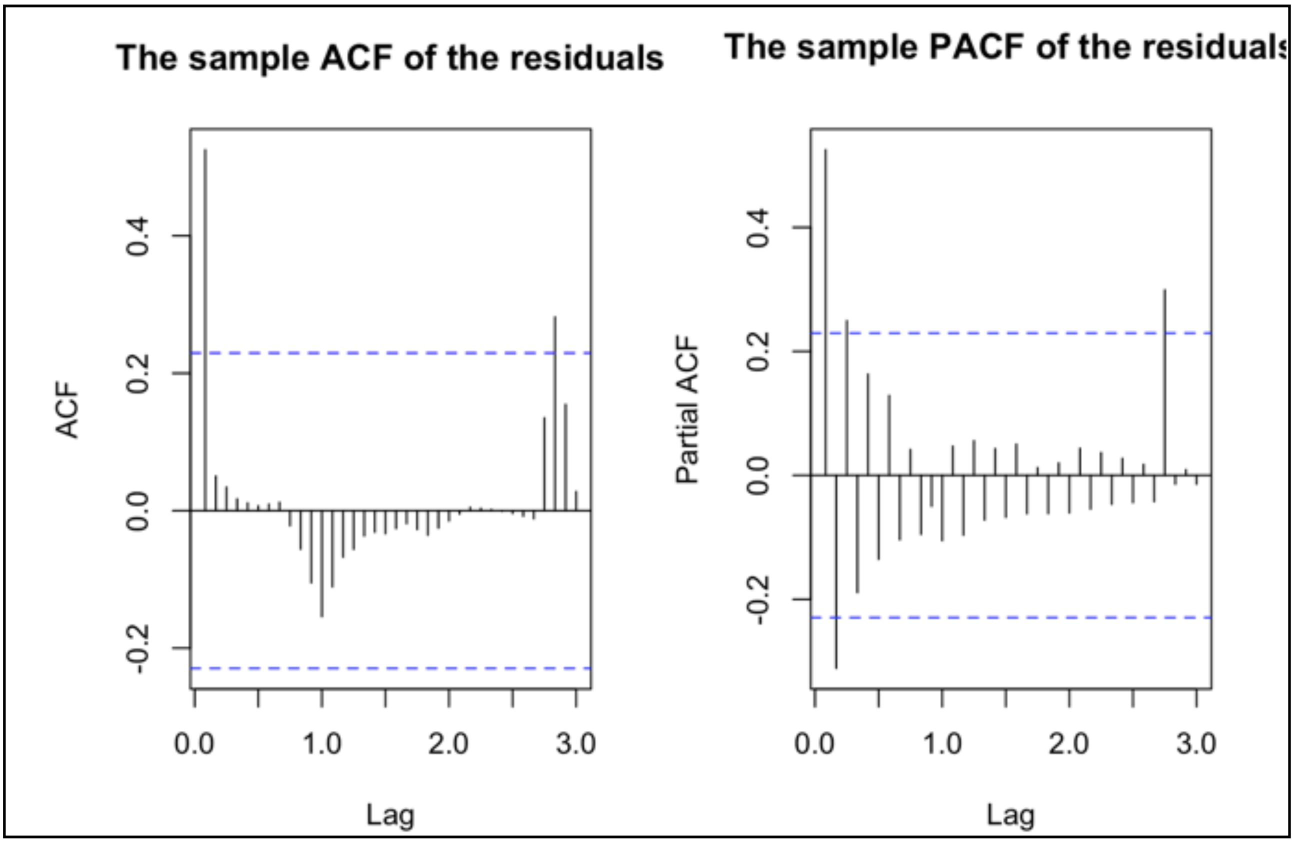


Fig 2.1

Fig 2.2



From the above plots (Fig 2.2) we can see ACF and PACF was constructed for this model by assuming D=1. From the ACF and PACF for the SARIMA model it was observed that the trend was De Trended. The ACF and PACF provide one significant lag and so we assume P =1 and Q = 1.

SECOND DIFFERENCING

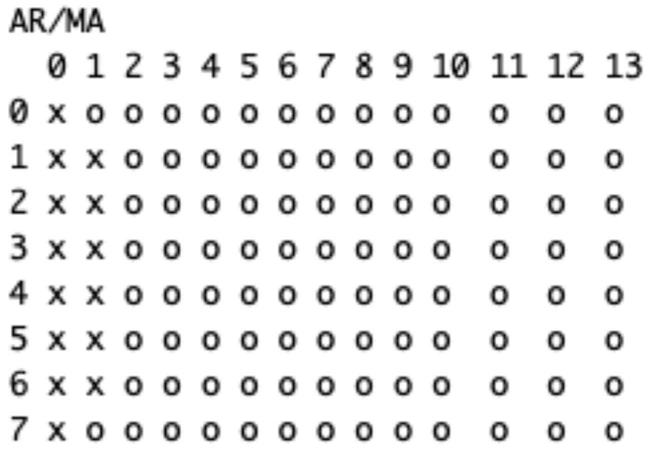
Now we will add the SARMA(1,1) component and see if we get rid of seasonal component. ACF and PACF was constructed for the model replacing P=1 and Q=1. Replacing P=1 and Q=1, another model was fitted. It was observed that the correlation in the model was reduced. ACF and PACF still provide 1 significant lag and so we have assumed p and q value to be 1.

MODEL FITTING

Now that we have got rid of some of the remaining trend and the correlations in ACF/PACF plots. We can observe from the plots that they provide one significant lag, in ACF q=1 and in PACF p=1.

Fig 2.3

EACF



```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  o  o  o
1 x x o o o o o o o o o  o  o  o
2 x x o o o o o o o o o  o  o  o
3 x x o o o o o o o o o  o  o  o
4 x x o o o o o o o o o  o  o  o
5 x x o o o o o o o o o  o  o  o
6 x x o o o o o o o o o  o  o  o
7 x o o o o o o o o o o  o  o  o
```

The tentative models are specified as

# SARIMA(1,0,1)x(1,1,1) by ACF and PACF

# SARIMA(0,1,1)x(1,1,1) by EACF

# SARIMA(1,1,2)x(1,1,1) by EACF

# SARIMA(2,1,2)x(1,1,1) by EACF

COEFFICIENT TEST (ML)

# SARIMA(0,1,1)x(1,1,1)

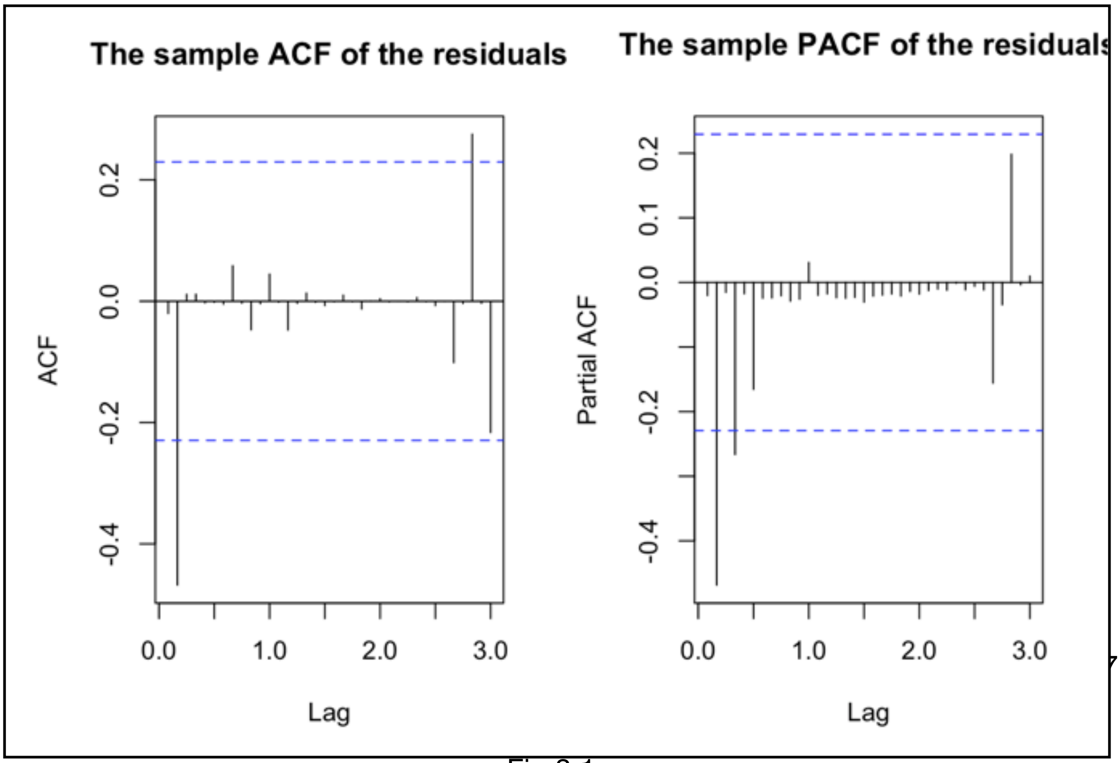


Fig 3.1

# SARIMA(2,1,1)x(1,1,1)

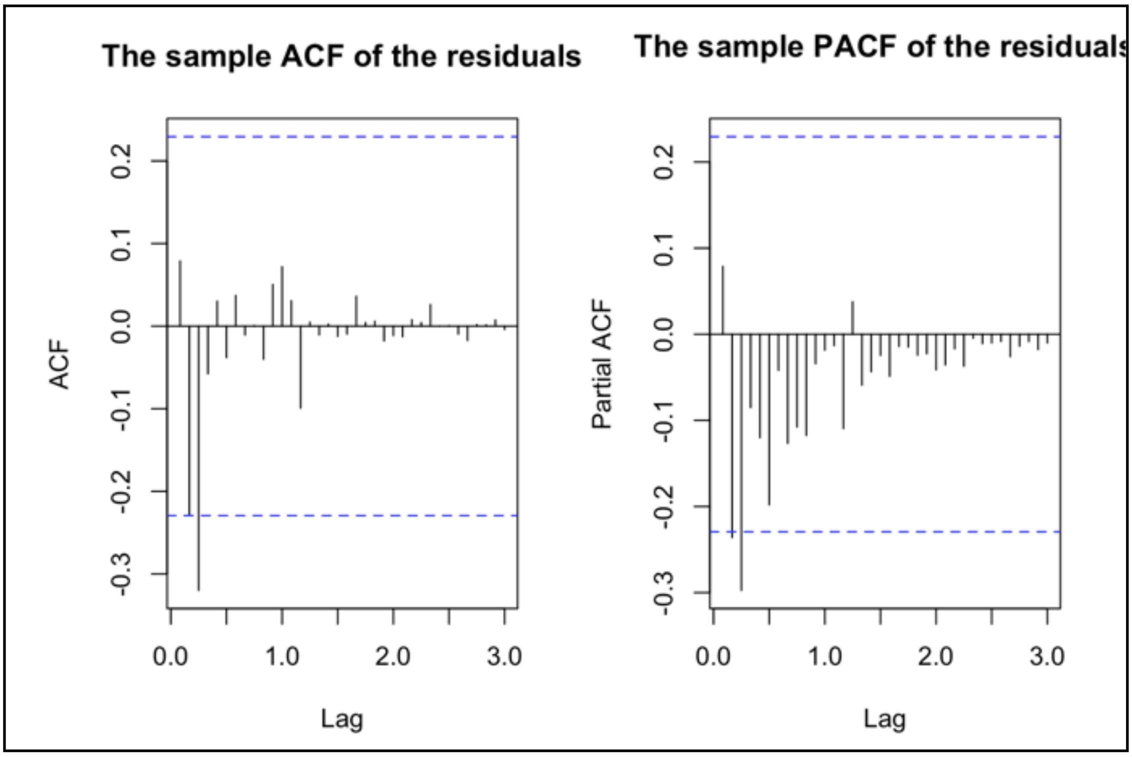


Fig 3.2

# SARIMA(2,1,2)x(1,1,1)

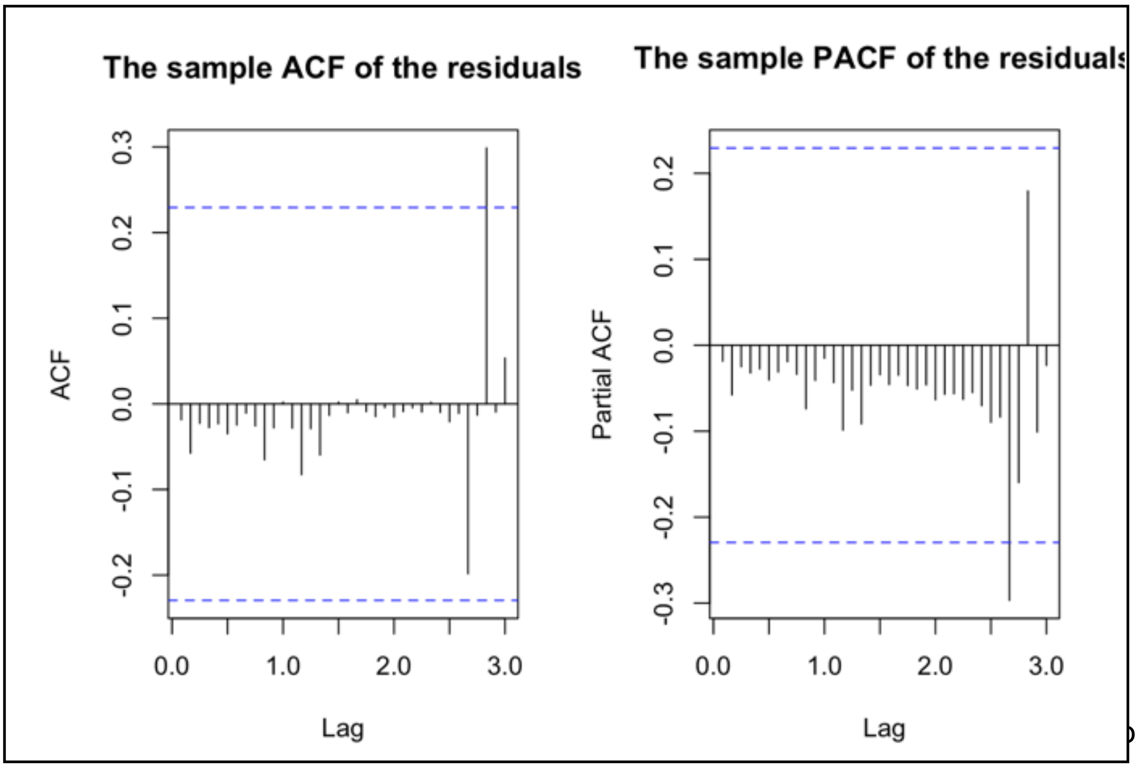
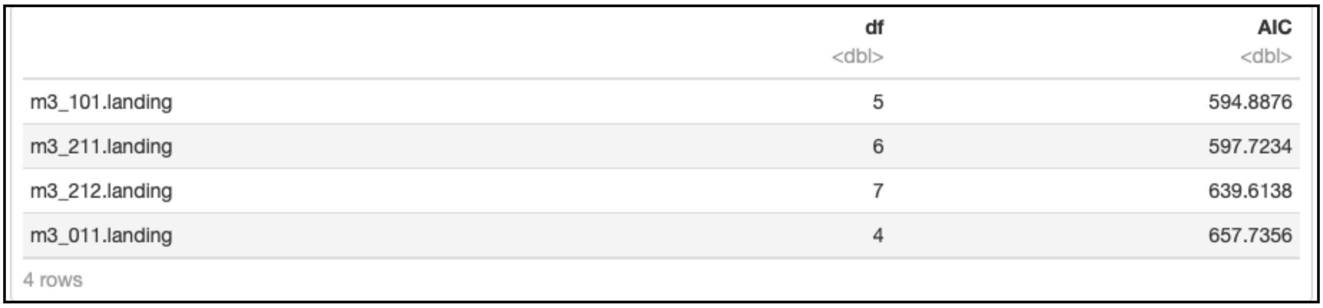


Fig 3.3

SORTING BY AIC AND BIC

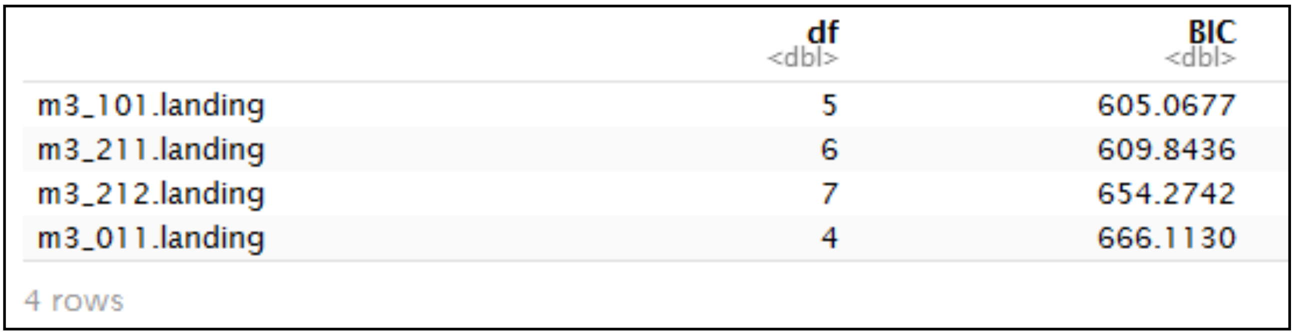
We use the sort function to find the best fitting model out of all the candidate models. The optimum model is selected by least value of AIC and BIC score level.



	df <dbl>	AIC <dbl>
m3_101.landing	5	594.8876
m3_211.landing	6	597.7234
m3_212.landing	7	639.6138
m3_011.landing	4	657.7356

4 rows

Fig 4.1



	df <dbl>	BIC <dbl>
m3_101.landing	5	605.0677
m3_211.landing	6	609.8436
m3_212.landing	7	654.2742
m3_011.landing	4	666.1130

4 rows

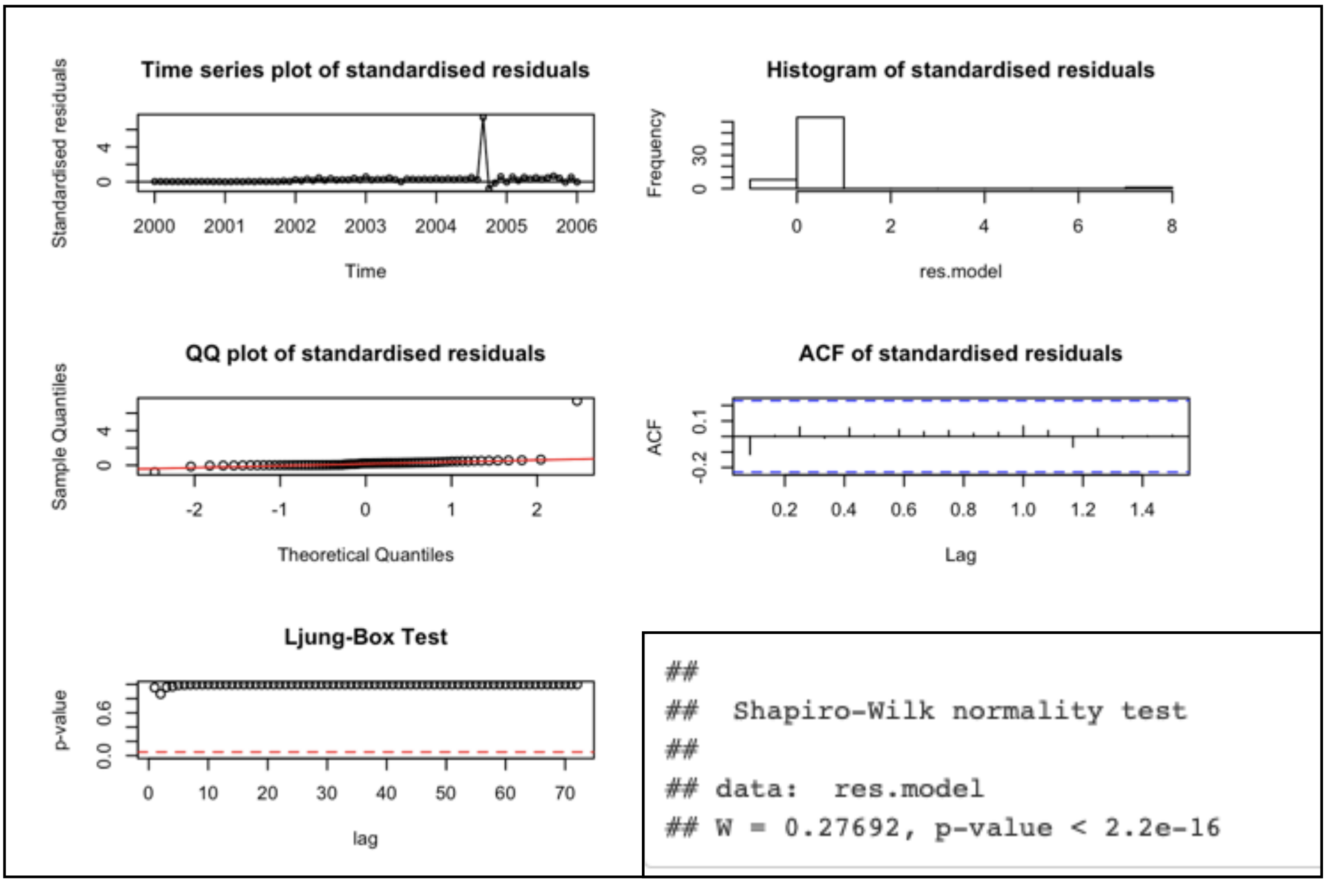
Fig 4.2

The AIC and BIC values are found for the following significant cases. ML gives better and most likely estimation when considering other parameters.

Using the sort score function we found the values of AIC and BIC.The AIC and BIC values are the least for SARIMA(1,0,1)x(1,1,1) model(ML) ,thus making it the best model.

MODEL DIAGNOSTICS

Model diagnostics is done via Residual analysis . Behaviour of standardised residuals is based on normality and autocorrelation. Here we will do residual analysis on SARIMA(1,0,1)x(1,1,1).



```
##
##   Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.27692, p-value < 2.2e-16
```

Fig 5

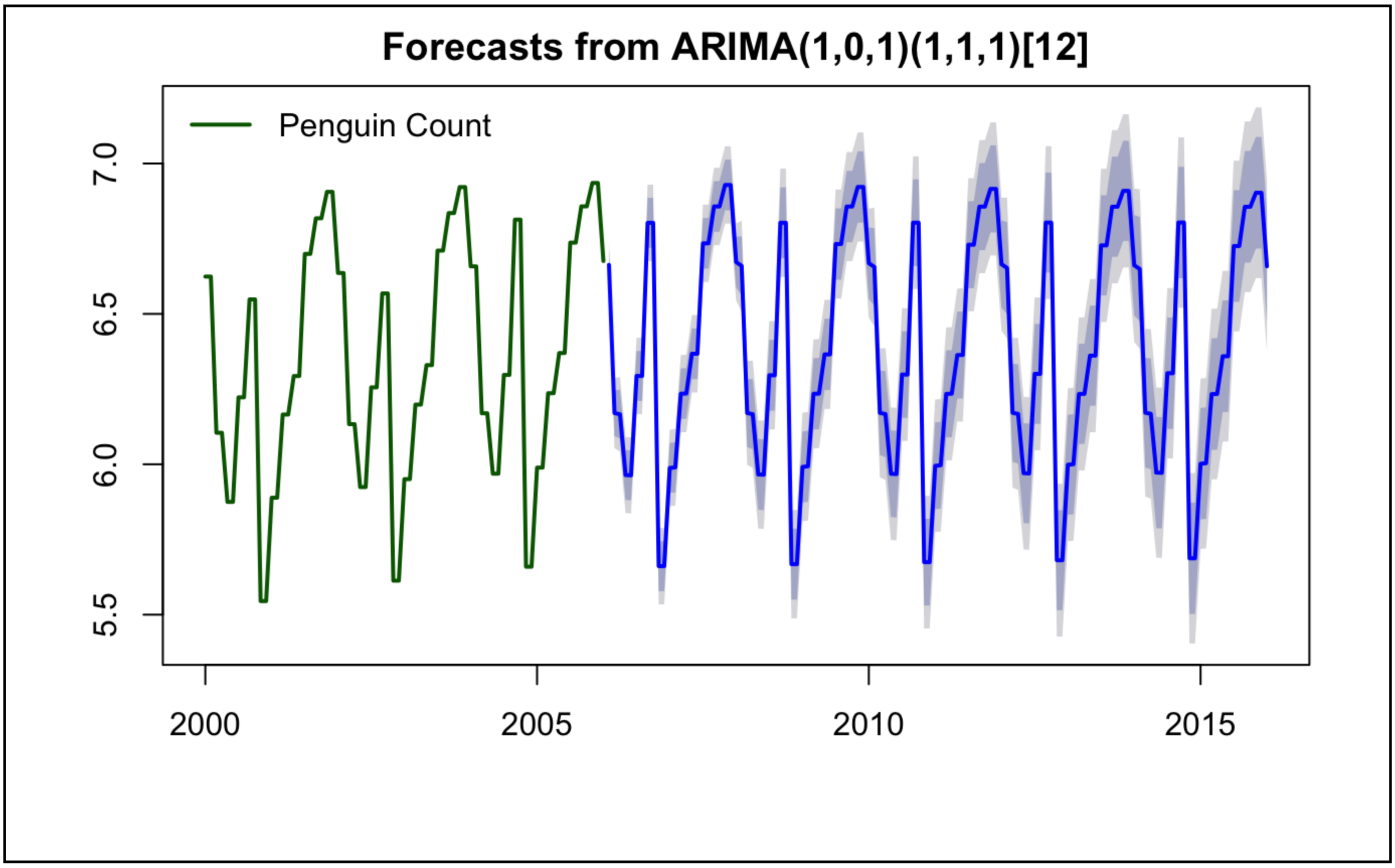
- The following inferences are made from the analysis
  - From the time series plot, it is observed that the residuals depict no obvious trends and no changing variance.
  - The Histogram of the residuals appears to be skewed to the right.
  - The Q-Q plot also looks normal, with all the residuals aligned with the normality curve, with the exception of a few values.
  - The ACF plot does not indicate any significant lags, indicating the presence of white noise.

- Assuming null hypothesis: the data to be non stationary ,while the alternate hypothesis: the data to be stationary. In the Sharpio-Wilk normality test the p-value is lesser than 0.05 and so we reject the null hypothesis thus proving the data to be stationary.
- The Ljung - Box test shows the size of the p-values relative to the 5% significance level, with most of the values above the threshold line.

- From these observations, we conclude that the SARIMA with fitted ARIMA(1,0,1)x(1,1,1) is the best model for the Penguin dataset.

FORECASTING

We use the predict() function for Penguin Population for the following 10 years using the SARIMA(1,0,1)x(1,1,1) model.



- With the best fitted model, we now predict the values for the Penguin population for the next 10 years, i.e; from 2006 to 2016.
- From the forecasts, we see that the trend of the Penguin population continues for the next 10 years, similar to the previous years.

CONCLUSION

From the various analysis and model fitting tools carried out on the Penguin population dataset for the years 2000to 2006, it was observed that the model SARIMA with fitted ARIMA(1,0,1)x(1,1,1) best fits the dataset. Upon forecasting the population values of Penguins for the next 10 years, it was fund that the forecasts tend to follow a similar trend as those of the previous years.

APPENDIX

```
#Required libraries

library(TSA)
library(forecast)
library(fUnitRoots)
library(ggplot2)
library(tseries)
library(tidyverse)
library(sandwich)
library(lmtest)
library(bestglm)
library(FitAR)
library(CombMSC)
library(fGarch)

#Accessing data present in package

penguin_data <- read.csv("~/Documents/penguin.csv", header = TRUE)
cat("Datatype of the penguin_data Dataset: ", class(penguin_data))
penguin_data

###### GENERAL ANALYSIS ######

# Convert to the TS object!

penguin_data.ts = matrix(penguin_data$Number, nrow = 84, ncol = 2)
penguin_data.ts = as.vector(t(penguin_data.ts))
penguin_data.ts = ts(penguin_data.ts,start= 2000, end=2006,freq=12)
class(penguin_data.ts)

# Plotting Time series graph

plot(penguin_data.ts,type='l',ylab='Count of penguins over the years 2000 to
2006',main="Time Series Graph",xlab='Years', col="darkgreen",lwd  = 1)
legend("topleft", bty = "n" ,col=c("darkgreen"),lwd=2, c("Penguin Count"))

# Plotting Time series graph with labels
```

```
plot(penguin_data.ts,type='l',ylab='Count of penguins over the years 2000 to
2006',main="Time Series Graph",xlab='Years', col="darkgreen",lwd  = 1)
legend("topleft", bty = "n" ,col=c("darkgreen"),lwd=2, c("Penguin Count”))

points(y=penguin_data.ts,x=time(penguin_data.ts),
pch=as.vector(season(penguin_data.ts)))


# ACF and PACF plots

par(mfrow=c(1,2))
acf(penguin_data.ts,  lag.max = 36,main="The sample ACF of landings series")
pacf(penguin_data.ts,  lag.max = 36,main="The sample PACF of landings series")
#Checking the p-value using adf test
adf.test(eggs.ts)


###### Differencing ######

#First Differencing

penguin1.landing = arima(penguin_data.ts,order=c(0,0,0),seasonal=list(order=c(0,1,0),
period=12))
res.m1 = residuals(penguin1.landing);
par(mfrow=c(1,1))
plot(res.m1,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
par(mfrow=c(1,2))
acf(res.m1,lag.max = 36,main = "The sample ACF of the residuals")
pacf(res.m1,lag.max = 36,main = "The sample PACF of the residuals")

#Second Differencing

penguin2.landing = arima(penguin_data.ts,order=c(0,0,0),seasonal=list(order=c(1,1,1),
period=12))
res.m2 = residuals(penguin2.landing);
par(mfrow=c(1,1))
plot(res.m2,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
par(mfrow=c(1,2))
acf(res.m2,lag.max = 36, main = "The sample ACF of the residuals")
pacf(res.m2,lag.max = 36, main = "The sample PACF of the residuals")


#EACF
eacf(res.m2)


#Coefficient test for SARIMA(0,1,1)x(1,1,1)

m3_011.landing = arima(penguin_data.ts,order=c(0,1,1),seasonal=list(order=c(1,1,1),
period=12),method = "ML")
coeftest(m3_011.landing)
par(mfrow=c(1,1))
```

```
plot(res.m4,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
par(mfrow=c(1,2))
acf(res.m4, lag.max = 36, main = "The sample ACF of the residuals")
pacf(res.m4, lag.max = 36, main = "The sample PACF of the residuals")

#Coefficient test for SARIMA(2,1,1)x(1,1,1)

m3_211.landing = arima(penguin_data.ts,order=c(2,1,1),seasonal=list(order=c(1,1,1),
period=12),method = "ML")
coeftest(m3_211.landing)
par(mfrow=c(1,1))
plot(res.m5,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
par(mfrow=c(1,2))
acf(res.m5, lag.max = 36, main = "The sample ACF of the residuals")
pacf(res.m5, lag.max = 36, main = "The sample PACF of the residuals")




#Coefficient test for SARIMA(2,1,2)x(1,1,1)

m3_212.landing = arima(penguin_data.ts,order=c(2,1,2),seasonal=list(order=c(1,1,1),
period=12),method = "ML")
coeftest(m3_212.landing)
par(mfrow=c(1,1))
plot(res.m9,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
par(mfrow=c(1,2))
acf(res.m9, lag.max = 36, main = "The sample ACF of the residuals")
pacf(res.m9, lag.max = 36, main = "The sample PACF of the residuals")

#Coefficient test for SARIMA(1,0,1)x(1,1,1)_12

m3_101.landing = arima(penguin_data.ts,order=c(1,0,1),seasonal=list(order=c(1,1,1),
period=12),method = "ML")
coeftest(m3_101.landing)
par(mfrow=c(1,1))
plot(res.m7,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
par(mfrow=c(1,2))
acf(res.m7, lag.max = 36, main = "The sample ACF of the residuals")
pacf(res.m7, lag.max = 36, main = "The sample PACF of the residuals")

#Sort score

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 ("aic","bic")')
  }
}

sort.score(AIC(m3_011.landing,m3_211.landing,m3_212.landing,m3_101.landingl), score
= "aic" )
sort.score(BIC(m3_011.landing,m3_211.landing,m3_212.landing,m3_101.landingl), score
= "bic" )

#residual analysis of SARIMA(1,0,1)x(1,1,1)
     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 = "Standard Residuals", xlab = "Time series plot of
     standard residuals")

     abline(h = 0)

     hist(res.model, main = "Histogram of Standard Residuals")

     qqnorm(res.model, main = "QQ plot of Standard Residuals")

     qqline(res.model, col = 2)

     acf(res.model, main = "ACF of Standard 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))

     }

     residual.analysis(model = m3_212.landing)

     par(mfrow = c(1,1))


# Forecast for next 10 years:

m3.landing = Arima(log.penguin_data.ts,order=c(1,0,1),seasonal=list(order=c(1,1,1),
period=12))
```

```
future = forecast(m3.landing, h = 120)

plot(future,col=c("darkgreen"),lwd = 2)

legend("topleft", bty = "n" ,col=c("darkgreen"),lwd=2, c("Penguin Count"))
```

REFERENCES

1.) Time Series. 2020. *Data Sets.*. [online] Available at: <https://. timeseries.weebly.com/data-sets.html> [Accessed 20 May 2020].

2.) En.wikipedia.org. 2020. *Penguin*. [online] Available at: <https://en.wikipedia.org/wiki/Penguin> [Accessed 2 June 2020].

3.) MATH1318 Time Series Analysis notes and tutorials by Dr. Haydar Demirhan

4.) Sermanet, P. and LeCun, Y., 2020. *Traffic Sign Recognition With Multi-Scale Convolutional Networks*. [online] Yann.lecun.com. Available at: <http://yann.lecun.com/exdb/publis/pdf/sermanet-ijcnn-11.pdf> [Accessed 1 June 2020].

5.) Medium. 2020. *An Overview Of Time Series Forecasting Models*. [online] Available at: <https://towardsdatascience.com/an-overview-of-time-series-forecasting-models-a2fa7a358fcb> [Accessed 3 June 2020].

6.) Abdallah, F., 2020. *Role Of Time Series Analysis In Forecasting Egg Production Depending On ARIMA Model*. [online] Article.sapub.org. Available at: <http://article.sapub.org/10.5923.j.am.20190901.01.html> [Accessed 3 June 2020].

7)Robjhyndman.com. 2020. *Forecasting Time Series With Multiple Seasonal Patterns*. [online] Available at: <https://robjhyndman.com/papers/multiseasonal.pdf> [Accessed 5 June 2020].

8) Vectors, R. and Vectors, M., 2020. *Melting Penguin Vector Image On Vectorstock*. [online] VectorStock. Available at: <https://www.vectorstock.com/royalty-free-vector/melting-penguin-vector-19750580> [Accessed 8 June 2020].

9)2020. *SARIMA (Seasonal ARIMA) Implementation On Time Series To Forecast The Number Of Malaria Incidence*. [online] Available at: <https://www.researchgate.net/publication/261307350_SARIMA_Seasonal_ARIMA_implementation_on_time_series_to_forecast_the_number_of_Malaria_incidence> [Accessed 7 June 2020].