DSC 425 Project Submission

Sea Ice Extent

by

Sidhant Thakur (2020181)

Dhruv Chandulal Dobariya (2069889)

Saransh Thakur (2020070)

Parth Babubhai Patel (2068905)

Under the guidance of:

John McDonald

DePaul University

Index:

Non-technical Summary:

The purpose of this report is to analyze the time series data of sea ice levels. The increasing severity of the greenhouse effect has raised concerns about rising sea levels and the potential sinking of many countries in the future. Considering these alarming predictions, we have undertaken an exploration of the extent of sea ice to gain a better understanding of the situation.

Choosing this dataset was motivated by the belief that nature is interconnected with every one of us, and as humans, we have a responsibility to observe and comprehend the patterns and changes in the natural world. By studying the extent of sea ice, we hope to identify trends and potentially find solutions to the problems posed by climate change.

The extent of the sea ice serves as a significant indicator of how the sea level is changing. The dataset we have utilized records the sea ice extent data for the North and South poles from 1980 to 2019, spanning a period of 39 years. During this time, there have been substantial changes in the sea ice extent, with an overall decrease observed in the combined ice coverage from both poles.

Technical Summary

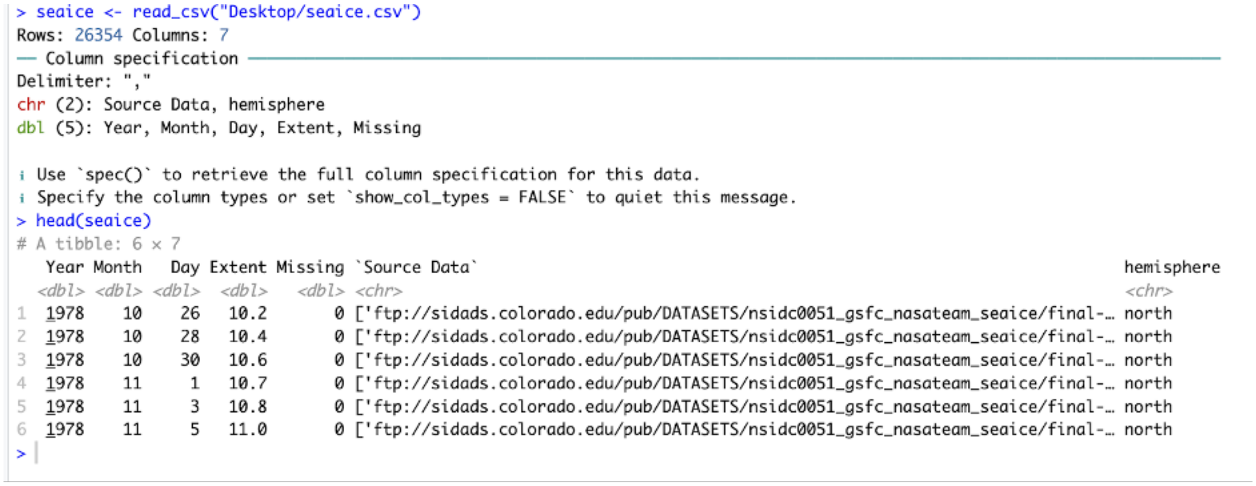
Dataset summary:

We used a dataset from Kaggle.

The dataset provides the total extent for each day for the entire time period (1978-2019) and here is the

There are 7 variables:

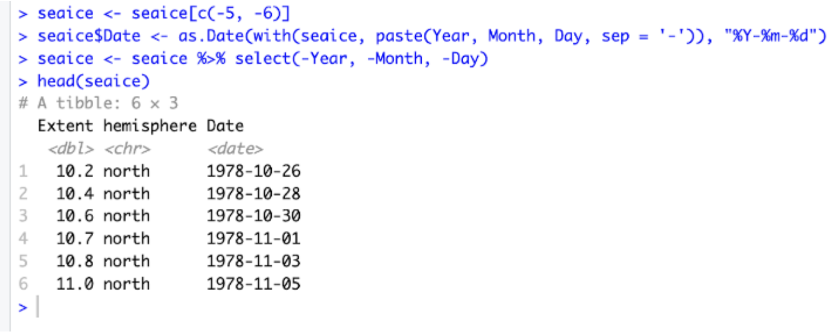
1. Year
2. Month
3. Day
4. Extent: unit is 10^6 sq km
5. Missing: unit is 10^6 sq km
6. Source: Source data product web site
7. hemisphere



```
> seaice <- read_csv("Desktop/seaice.csv")
Rows: 26354 Columns: 7
── Column specification ────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr (2): Source Data, hemisphere
dbl (5): Year, Month, Day, Extent, Missing

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> head(seaice)
# A tibble: 6 × 7
   Year Month   Day Extent Missing `Source Data`                                                  hemisphere
  <dbl> <dbl> <dbl>  <dbl>   <dbl> <chr>                                                          <chr>
1  1978    10    26   10.2       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
2  1978    10    28   10.4       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
3  1978    10    30   10.6       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
4  1978    11     1   10.7       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
5  1978    11     3   10.8       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
6  1978    11     5   11.0       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
> 
```

Data Processing:

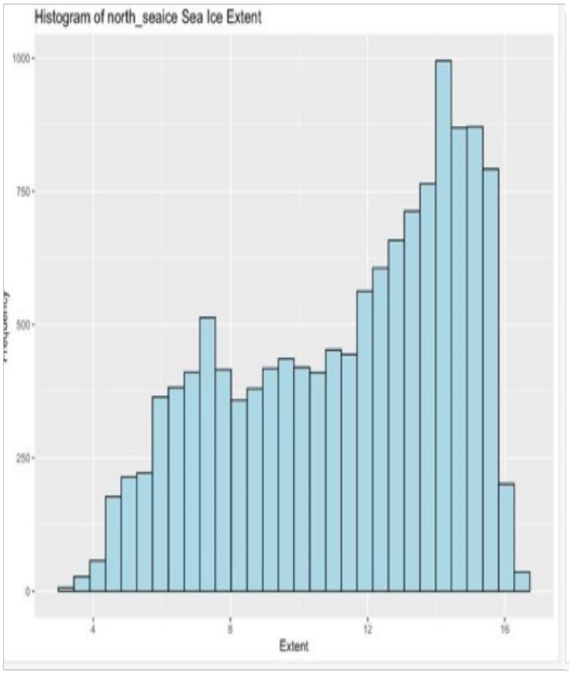
We eliminated the missing column and the source data from our data before combining the year, month, and day into a single column and renaming it "date." Moreover, we separated out the dataset based on the hemisphere into two parts which are North and South.

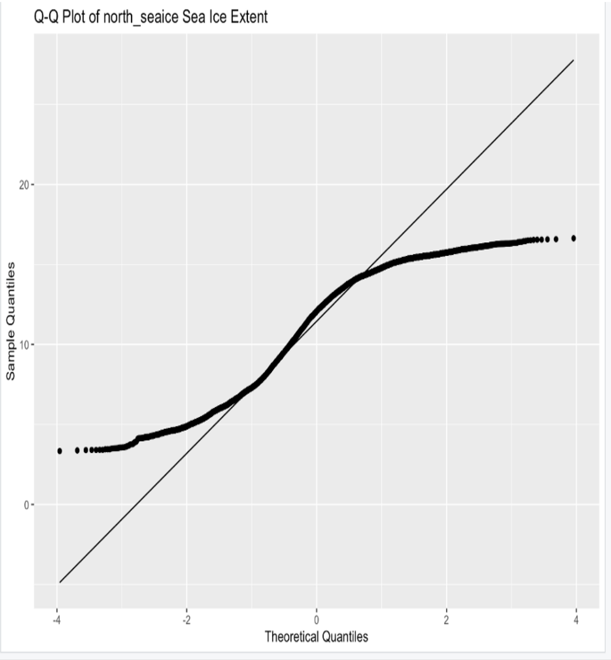


```
> seaice <- seaice[c(-5, -6)]
> seaice$Date <- as.Date(with(seaice, paste(Year, Month, Day, sep = '-')), "%Y-%m-%d")
> seaice <- seaice %>% select(-Year, -Month, -Day)
> head(seaice)
# A tibble: 6 × 3
  Extent hemisphere Date
   <dbl> <chr>      <date>
1   10.2 north      1978-10-26
2   10.4 north      1978-10-28
3   10.6 north      1978-10-30
4   10.7 north      1978-11-01
5   10.8 north      1978-11-03
6   11.0 north      1978-11-05
> 
```

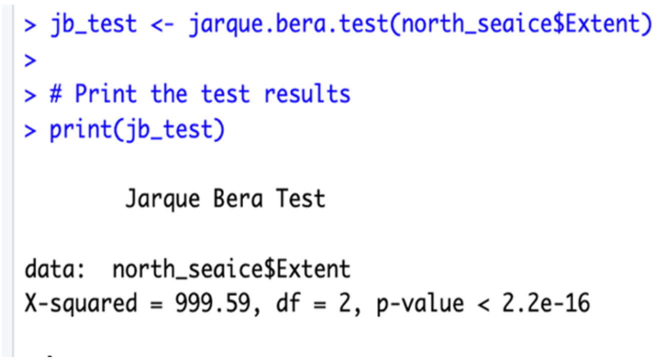
Exploratory Data Analysis

North Dataset:





Histogram **QQ Plot**



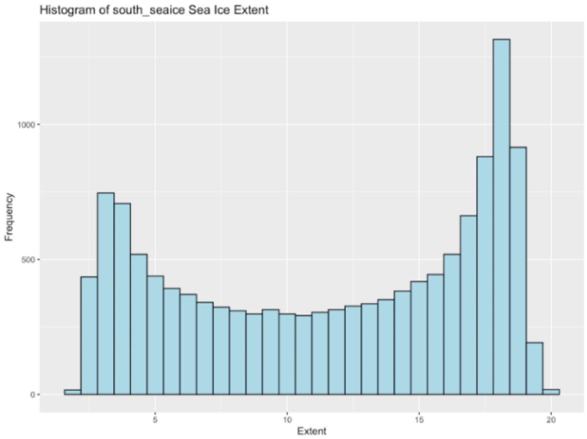
```
> jb_test <- jarque.bera.test(north_seaice$Extent)
>
> # Print the test results
> print(jb_test)

        Jarque Bera Test

data:  north_seaice$Extent
X-squared = 999.59, df = 2, p-value < 2.2e-16
```

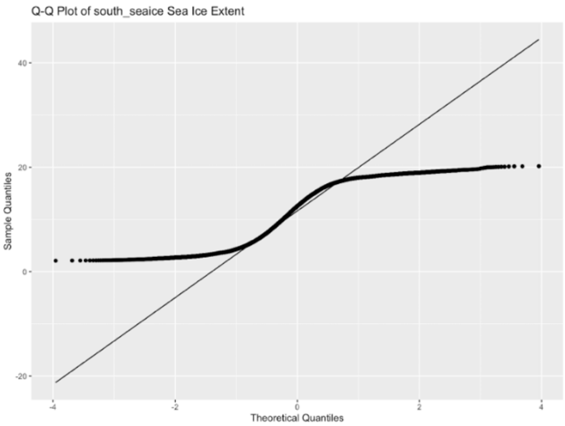
The results of the test statistic were 999.59 and a p-value less than 2.2e-16. Based on that we can say that the data in the north significantly deviates from a normal distribution. The extremely small p-value suggests strong evidence against the null hypothesis of normality.

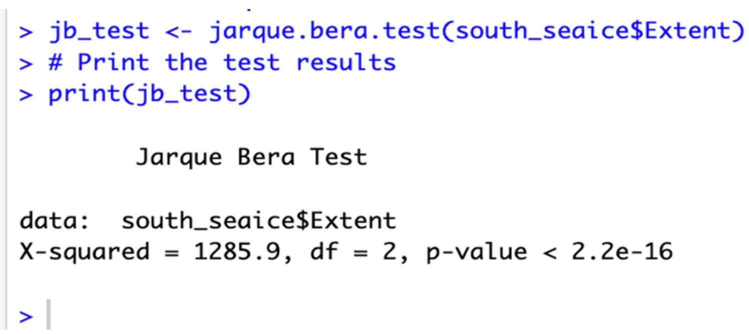
These results tell us that the data does not follow a normal distribution based on this test

Overall, the northern hemisphere is not normally distributed as we can see from plots and tests.

South Dataset:







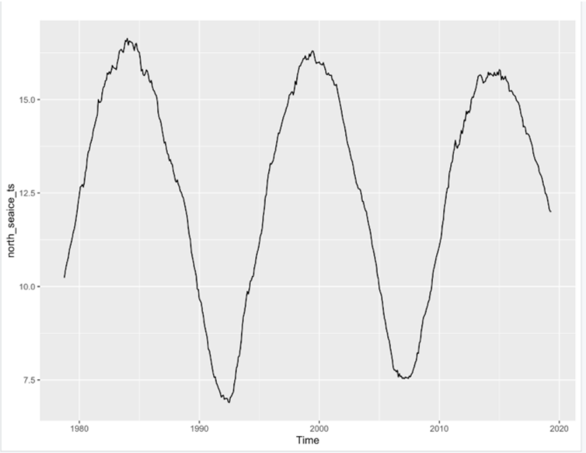
```
> jb_test <- jarque.bera.test(south_seaice$Extent)
> # Print the test results
> print(jb_test)

        Jarque Bera Test

data:  south_seaice$Extent
X-squared = 1285.9, df = 2, p-value < 2.2e-16

> 
```

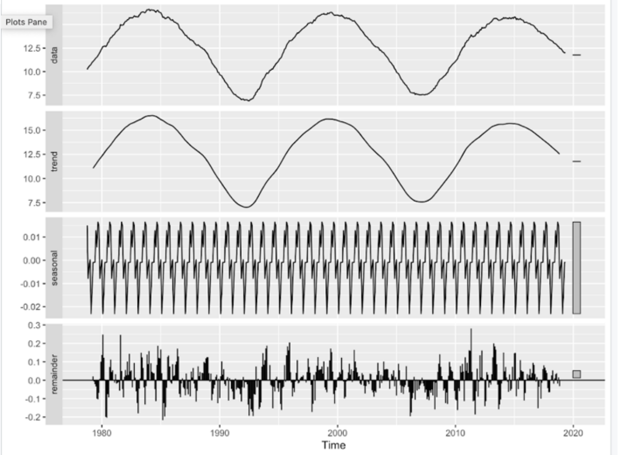
The degree of freedom for the South is 2, X-squared value is 1285.9 and p-value is less than significant value (0.05). The low p-value tells us that we can reject null hypothesis. This suggests to us that our data is not normally distributed.

Based on the above results we can say that the southern hemisphere is not normally distributed as we can see from plots and tests.

Creating Time Series:

For North:

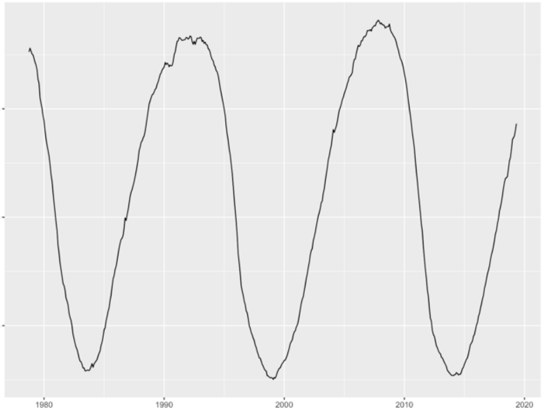


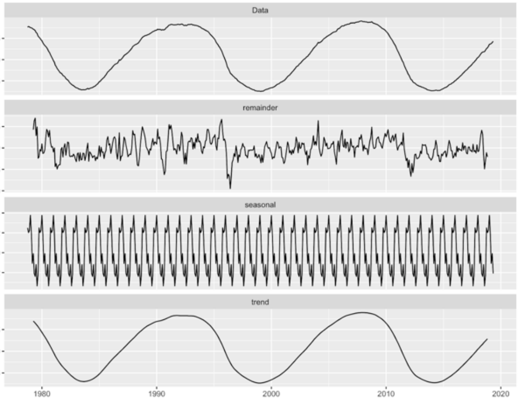


From the plots we can see that there might be a seasonality present in data and seasonal term is tiny compared to the "remainder".

We can also see that there is a periodic trend in the North hemisphere data.

For South Hemisphere:

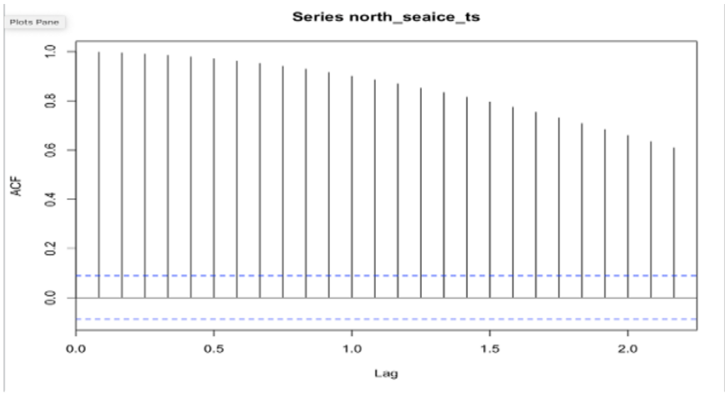


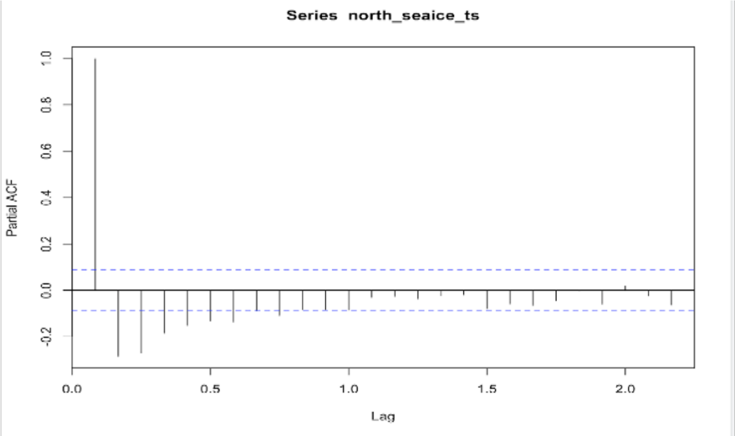


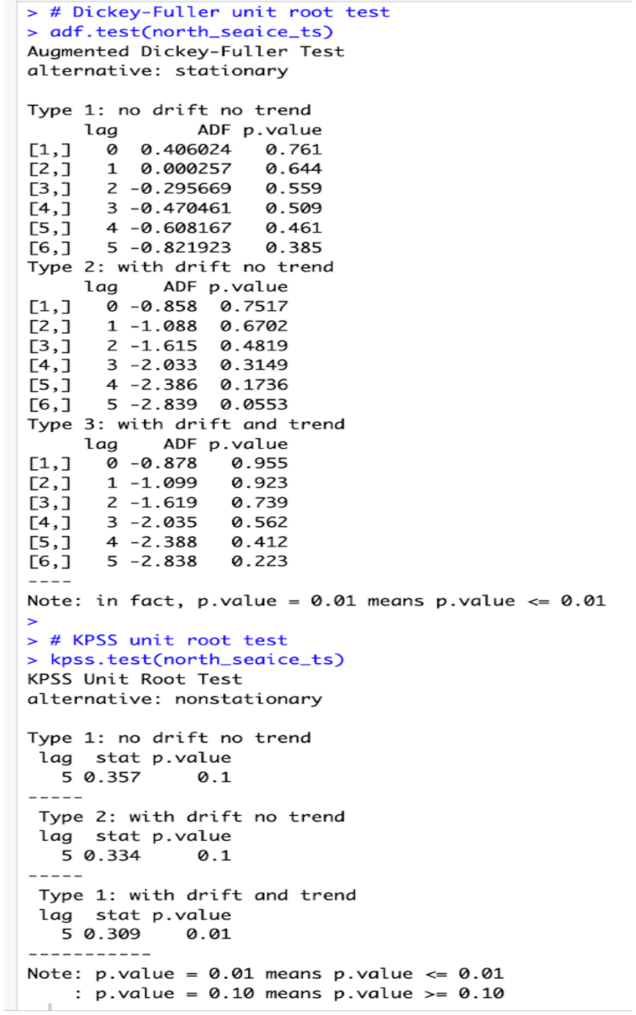
The above plot tells us about trend, seasonality and remainder of the South hemisphere data. The trend component tells us that there is some percentage of decline in sea ice extent. The seasonal component tells us there is some cyclic pattern in sea ice extent indicating some seasonality. The remainder plot tells us that there are random fluctuations in the sea ice extent.

Checking For Stationary of both Hemisphere:

North Hemisphere:



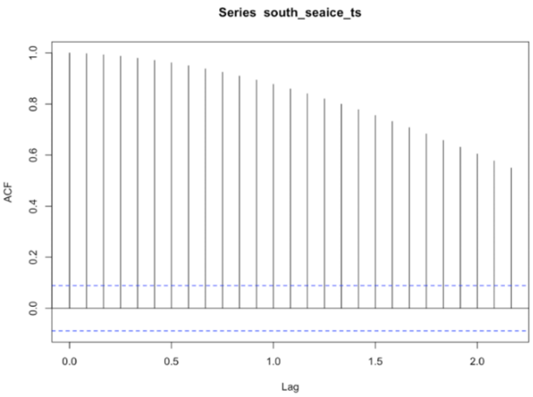


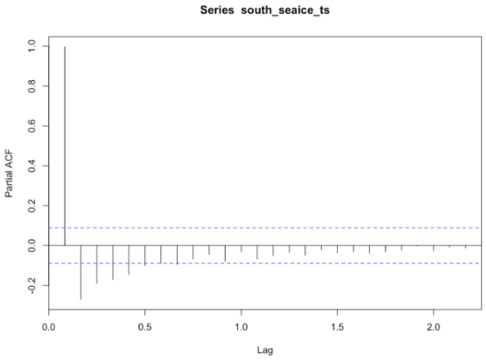


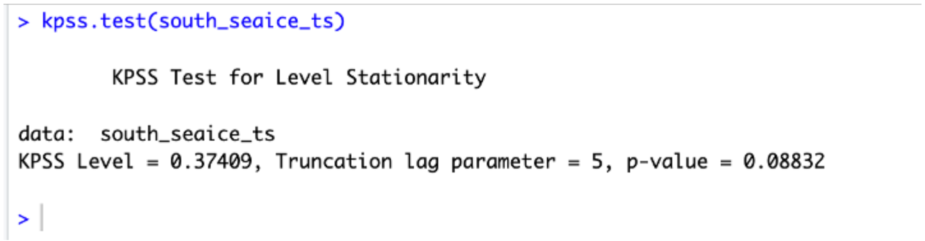
```
> # Dickey-Fuller unit root test
> adf.test(north_seaice_ts)
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag       ADF p.value
[1,]   0  0.406024   0.761
[2,]   1  0.000257   0.644
[3,]   2 -0.295669   0.559
[4,]   3 -0.470461   0.509
[5,]   4 -0.608167   0.461
[6,]   5 -0.821923   0.385
Type 2: with drift no trend
     lag    ADF p.value
[1,]   0 -0.858  0.7517
[2,]   1 -1.088  0.6702
[3,]   2 -1.615  0.4819
[4,]   3 -2.033  0.3149
[5,]   4 -2.386  0.1736
[6,]   5 -2.839  0.0553
Type 3: with drift and trend
     lag    ADF p.value
[1,]   0 -0.878   0.955
[2,]   1 -1.099   0.923
[3,]   2 -1.619   0.739
[4,]   3 -2.035   0.562
[5,]   4 -2.388   0.412
[6,]   5 -2.838   0.223
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
>
> # KPSS unit root test
> kpss.test(north_seaice_ts)
KPSS Unit Root Test
alternative: nonstationary

Type 1: no drift no trend
 lag  stat p.value
   5 0.357     0.1
-----
 Type 2: with drift no trend
 lag  stat p.value
   5 0.334     0.1
-----
 Type 1: with drift and trend
 lag  stat p.value
   5 0.309    0.01
-----------
Note: p.value = 0.01 means p.value <= 0.01
    : p.value = 0.10 means p.value >= 0.10
```

The plots and the test results tell us that the series is non-stationary. Based on those results we can say that the series does not have a constant meaning and variance over time.

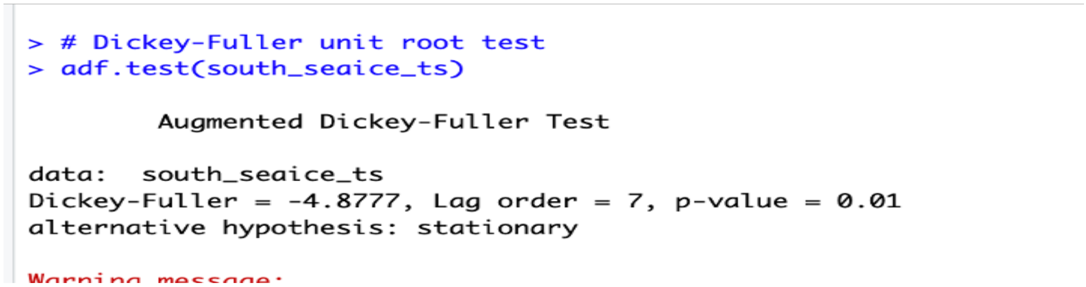
South Hemisphere:







```
> kpss.test(south_seaice_ts)

        KPSS Test for Level Stationarity

data:  south_seaice_ts
KPSS Level = 0.37409, Truncation lag parameter = 5, p-value = 0.08832

> 
```



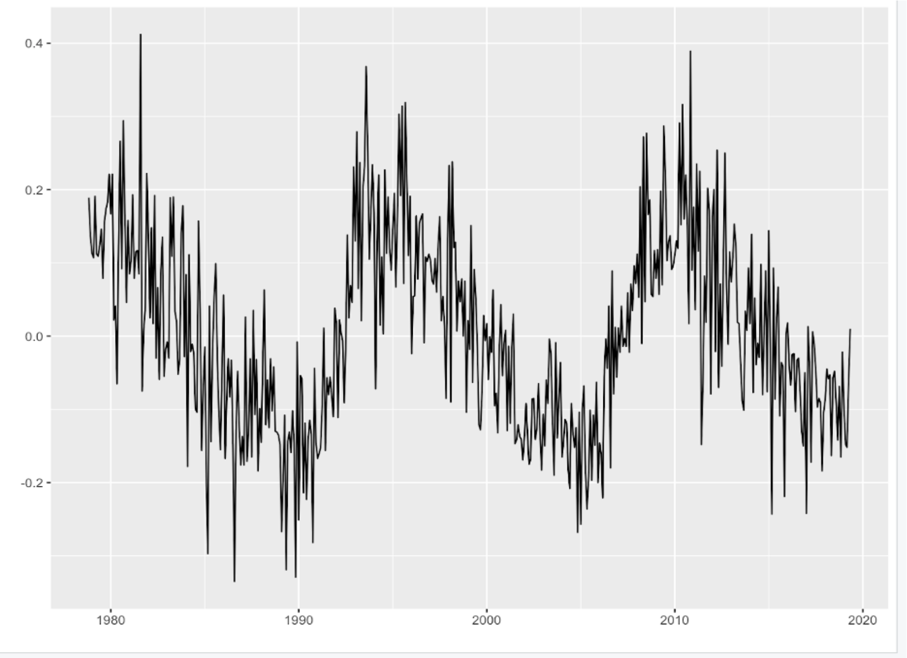
```
> # Dickey-Fuller unit root test
> adf.test(south_seaice_ts)

        Augmented Dickey-Fuller Test

data:  south_seaice_ts
Dickey-Fuller = -4.8777, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
```

Based on the above results of the Dickey-Fuller test, we see that the lag order is 7, Dickey-Fuller value is -4.8777 and p-value is less than significant value (0.05). The low p-value tells us that we can reject null hypothesis. The results show that the series is stationary, however there is a discrepancy when we look at ACF and PACF plots we can tell that the series is non-stationary.

Based on the KPSS test results, we see that KPSS level statistic is 0.37409, the truncation lag parameter used is 5. The p-value is 0.08832 which is greater than the significant value. Hence, we fail to reject the null hypothesis of level stationery.

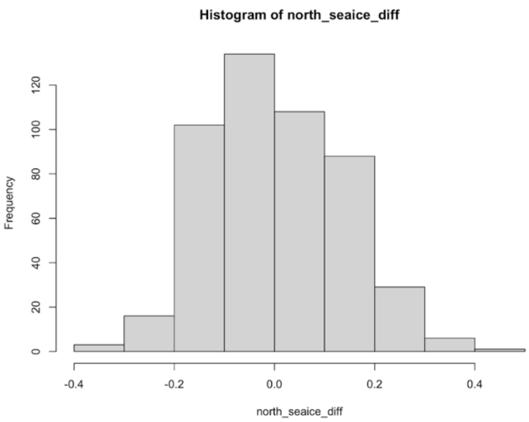
Taking difference of model for stationarity:

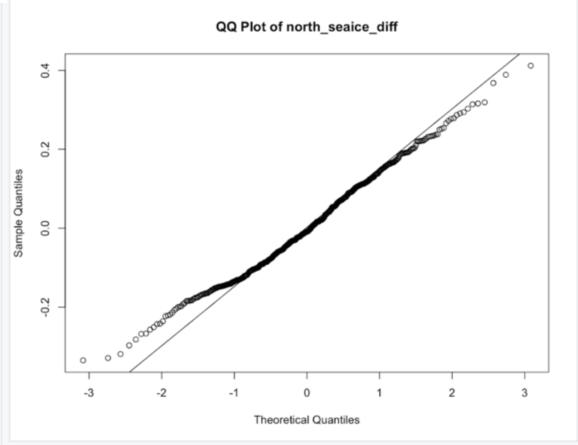
North Model:



The above plot shows the presence of a trend within the series indicates that the series exhibits a discernible pattern of overall increase or decline, as indicated by the plot. The plot also implies that the data has seasonality which means that the time series data exhibits repeating trends.

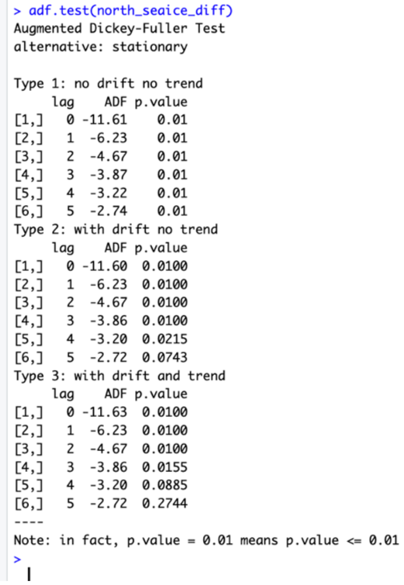
EDA of difference north model:



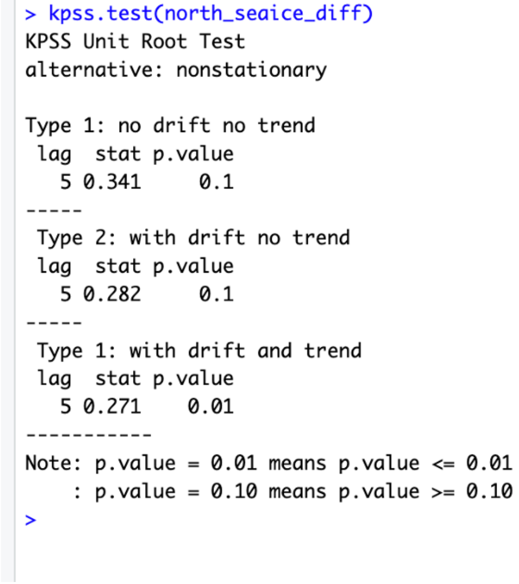


The normal quantile and histogram plot shows that data seems to be normally distributed.

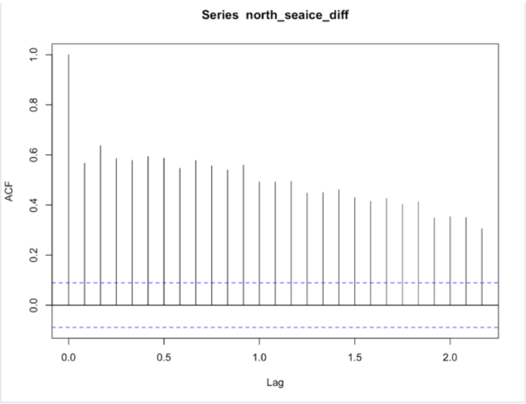
Checking Stationary.

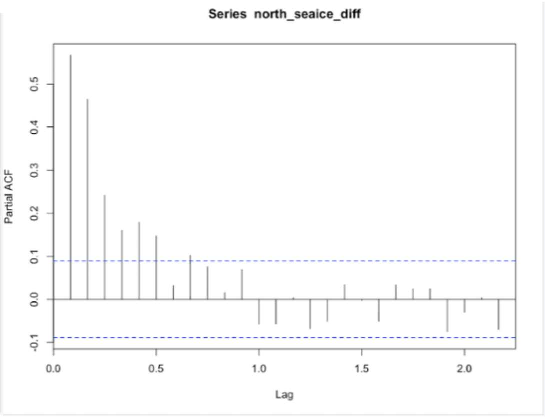


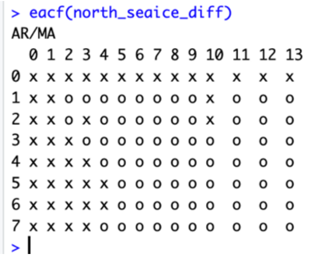
```
> adf.test(north_seaice_diff)
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag    ADF p.value
[1,]   0 -11.61    0.01
[2,]   1  -6.23    0.01
[3,]   2  -4.67    0.01
[4,]   3  -3.87    0.01
[5,]   4  -3.22    0.01
[6,]   5  -2.74    0.01
Type 2: with drift no trend
     lag    ADF p.value
[1,]   0 -11.60  0.0100
[2,]   1  -6.23  0.0100
[3,]   2  -4.67  0.0100
[4,]   3  -3.86  0.0100
[5,]   4  -3.20  0.0215
[6,]   5  -2.72  0.0743
Type 3: with drift and trend
     lag    ADF p.value
[1,]   0 -11.63  0.0100
[2,]   1  -6.23  0.0100
[3,]   2  -4.67  0.0100
[4,]   3  -3.86  0.0155
[5,]   4  -3.20  0.0885
[6,]   5  -2.72  0.2744
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
>
```



```
> kpss.test(north_seaice_diff)
KPSS Unit Root Test
alternative: nonstationary

Type 1: no drift no trend
 lag  stat p.value
   5 0.341     0.1
-----
 Type 2: with drift no trend
 lag  stat p.value
   5 0.282     0.1
-----
 Type 1: with drift and trend
 lag  stat p.value
   5 0.271    0.01
-----------
Note: p.value = 0.01 means p.value <= 0.01
    : p.value = 0.10 means p.value >= 0.10
>
```







```
> eacf(north_seaice_diff)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x x o o o o o o o o x  o  o  o
2 x x o x o o o o o o x  o  o  o
3 x x x o o o o o o o o  o  o  o
4 x x x x o o o o o o o  o  o  o
5 x x x x x o o o o o o  o  o  o
6 x x x x x o o o o o o  o  o  o
7 x x x x o o o o o o o  o  o  o
> |
```

The ACF, PACF and EACF plot suggested that model might be ARIMA (2,0,1)

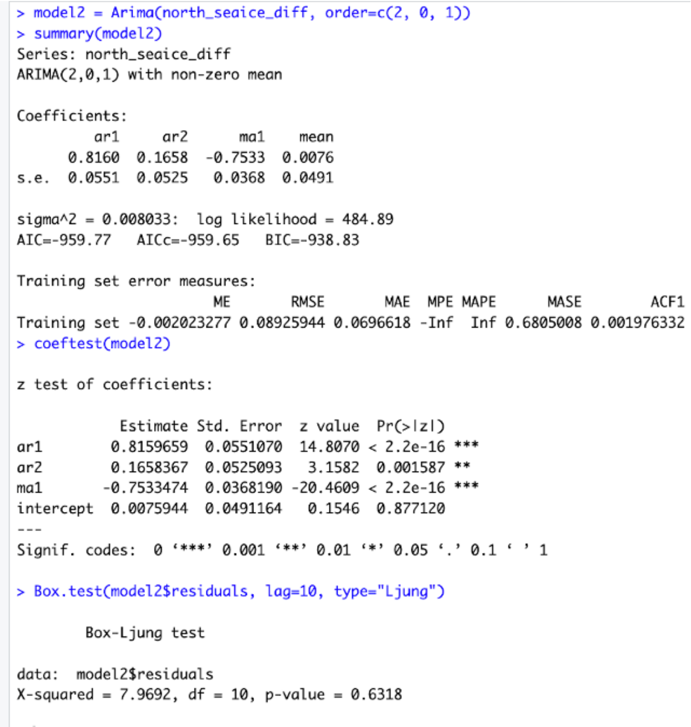
The p-values for all lags of ADF (Augmented Dickey Fuller) test are equal to 0.01, telling us that the null hypothesis of non-stationarity is rejected, and series is in favor of stationarity.

The p-values for all scenarios and lags are greater than the significance level of 0.01 based on the KPSS test results. This tells us that the evidence is insufficient to reject the null hypothesis of stationarity. As a result, we cannot reject the null hypothesis and conclude that the differenced series is stationary.

Models North Hemisphere:

Manual Model:

We choose ARIMA (2,0,1) to be our best model



```
> model2 = Arima(north_seaice_diff, order=c(2, 0, 1))
> summary(model2)
Series: north_seaice_diff
ARIMA(2,0,1) with non-zero mean

Coefficients:
         ar1     ar2      ma1    mean
      0.8160  0.1658  -0.7533  0.0076
s.e.  0.0551  0.0525   0.0368  0.0491

sigma^2 = 0.008033:  log likelihood = 484.89
AIC=-959.77   AICc=-959.65   BIC=-938.83

Training set error measures:
                       ME       RMSE       MAE  MPE MAPE      MASE        ACF1
Training set -0.002023277 0.08925944 0.0696618 -Inf  Inf 0.6805008 0.001976332
> coeftest(model2)

z test of coefficients:

           Estimate Std. Error  z value  Pr(>|z|)
ar1        0.8159659  0.0551070  14.8070 < 2.2e-16 ***
ar2        0.1658367  0.0525093   3.1582  0.001587 **
ma1       -0.7533474  0.0368190 -20.4609 < 2.2e-16 ***
intercept  0.0075944  0.0491164   0.1546  0.877120
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Box.test(model2$residuals, lag=10, type="Ljung")

        Box-Ljung test

data:  model2$residuals
X-squared = 7.9692, df = 10, p-value = 0.6318
```

The ARIMA (2,0,1) model with a non-zero mean was fitted to the differenced series "north_seaice_diff" of the North Sea ice data.

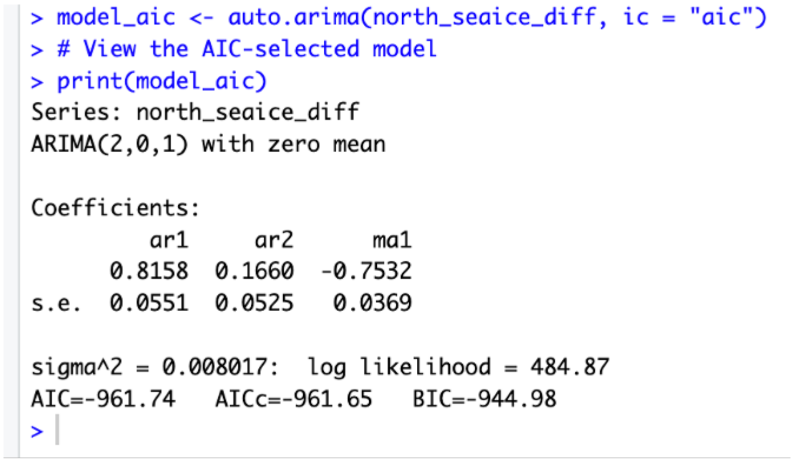
The estimated coefficients for the model are as follows: the first autoregressive term AR (1) is 0.8160, the second autoregressive term AR (2) is 0.1658, the moving average term MA (1) is -0.7533, and the mean is 0.0076. The standard errors for these coefficients were also calculated.

The model's estimated residual variance (sigma^2) is 0.008033, and the log likelihood is 484.89. The AIC is -959.77, the corrected AIC is -959.65, and the BIC is -938.83.

The model's performance on the training set shows low errors and good fit. The hypothesis tests on the coefficients indicate that all coefficients, except for the mean, are statistically significant.

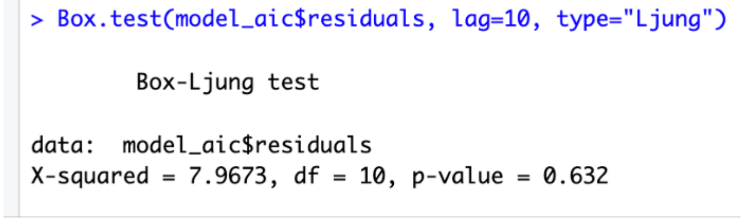
The Box-Ljung test on the residuals shows no significant autocorrelation.

Auto Arima Model:



```
> model_aic <- auto.arima(north_seaice_diff, ic = "aic")
> # View the AIC-selected model
> print(model_aic)
Series: north_seaice_diff
ARIMA(2,0,1) with zero mean

Coefficients:
         ar1     ar2      ma1
      0.8158  0.1660  -0.7532
s.e.  0.0551  0.0525   0.0369

sigma^2 = 0.008017:  log likelihood = 484.87
AIC=-961.74   AICc=-961.65   BIC=-944.98
> 
```

The AIC-selected model suggests that an ARIMA (2,0,1) model with the given coefficients and zero mean provides a good fit to the differencing this case, the test statistic is X-squared = 7.9673, and the degrees of freedom (df) is 10. The p-value associated with the test is 0.632.



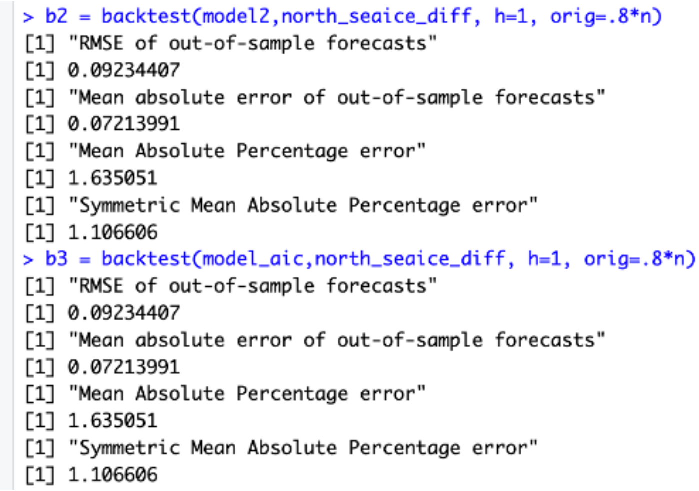
```
> Box.test(model_aic$residuals, lag=10, type="Ljung")

        Box-Ljung test

data:  model_aic$residuals
X-squared = 7.9673, df = 10, p-value = 0.632
```

Looking at the results we see that with a significance level of 0.05, since the p-value (0.632) is greater than the significance level, we fail to reject the null hypothesis. This suggests that there is no significant evidence of residual autocorrelation up to lag 10.

Overall, there is no strong indication of residual autocorrelation in the ARIMA model residuals selected using the AIC criterion

Both Manual and auto Arima model are same

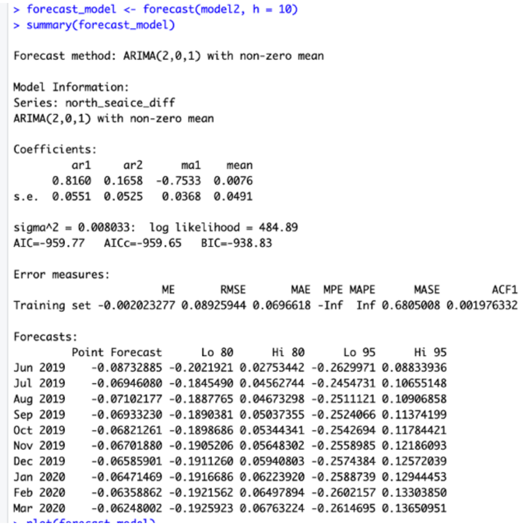
Back testing:



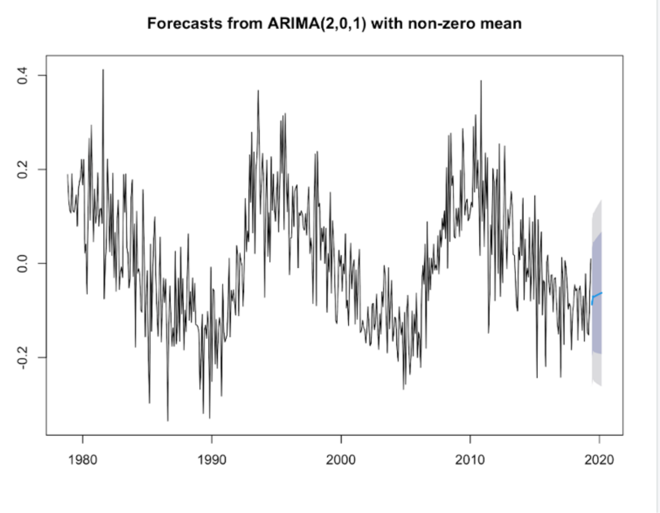
```
> b2 = backtest(model2,north_seaice_diff, h=1, orig=.8*n)
[1] "RMSE of out-of-sample forecasts"
[1] 0.09234407
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.07213991
[1] "Mean Absolute Percentage error"
[1] 1.635051
[1] "Symmetric Mean Absolute Percentage error"
[1] 1.106606
> b3 = backtest(model_aic,north_seaice_diff, h=1, orig=.8*n)
[1] "RMSE of out-of-sample forecasts"
[1] 0.09234407
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.07213991
[1] "Mean Absolute Percentage error"
[1] 1.635051
[1] "Symmetric Mean Absolute Percentage error"
[1] 1.106606
```

We can see from these metrics that all of the models perform similarly in terms of RMSE, mean absolute error, mean absolute percentage error, and symmetric mean absolute percentage error. The differences between the models are extremely minor.

Forecast:

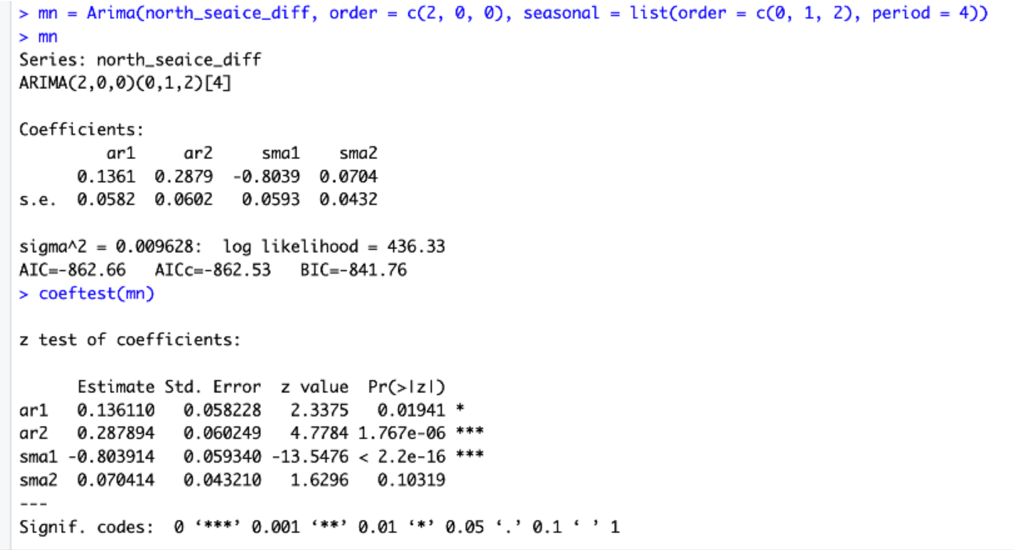


```
> forecast_model <- forecast(model2, h = 10)
> summary(forecast_model)

Forecast method: ARIMA(2,0,1) with non-zero mean

Model Information:
Series: north_seaice_diff
ARIMA(2,0,1) with non-zero mean

Coefficients:
         ar1     ar2      ma1    mean
      0.8160  0.1658  -0.7533  0.0076
s.e.  0.0551  0.0525   0.0368  0.0491

sigma^2 = 0.008033:  log likelihood = 484.89
AIC=-959.77   AICc=-959.65   BIC=-938.83

Error measures:
                       ME       RMSE       MAE  MPE MAPE      MASE        ACF1
Training set -0.002023277 0.08925944 0.0696618 -Inf  Inf 0.6805008 0.001976332

Forecasts:
         Point Forecast      Lo 80      Hi 80      Lo 95      Hi 95
Jun 2019    -0.08732885 -0.2021921 0.02753442 -0.2629971 0.08833936
Jul 2019    -0.06946080 -0.1845490 0.04562744 -0.2454731 0.10655148
Aug 2019    -0.07102177 -0.1887765 0.04673298 -0.2511121 0.10906858
Sep 2019    -0.06933230 -0.1890381 0.05037355 -0.2524066 0.11374199
Oct 2019    -0.06821261 -0.1898686 0.05344341 -0.2542694 0.11784421
Nov 2019    -0.06701880 -0.1905206 0.05648302 -0.2558985 0.12186093
Dec 2019    -0.06585901 -0.1911260 0.05940803 -0.2574384 0.12572039
Jan 2020    -0.06471469 -0.1916686 0.06223920 -0.2588739 0.12944453
Feb 2020    -0.06358862 -0.1921562 0.06497894 -0.2602157 0.13303850
Mar 2020    -0.06248002 -0.1925923 0.06763224 -0.2614695 0.13650951
> plot(forecast_model)
```



The model forecasts converge to a mean value of approximately 0.0076. This means that as the forecast horizon increases, the predicted values tend to approach this mean value.

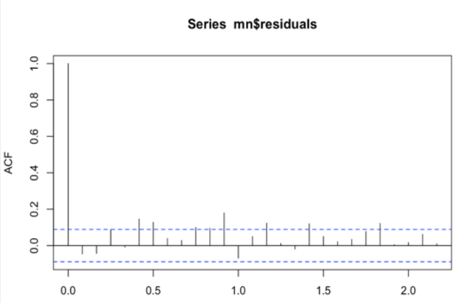
SARIMA Model:

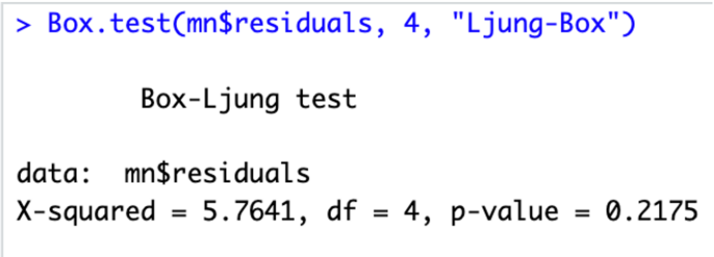


```
> mn = Arima(north_seaice_diff, order = c(2, 0, 0), seasonal = list(order = c(0, 1, 2), period = 4))
> mn
Series: north_seaice_diff
ARIMA(2,0,0)(0,1,2)[4]

Coefficients:
         ar1     ar2     sma1    sma2
      0.1361  0.2879  -0.8039  0.0704
s.e.  0.0582  0.0602   0.0593  0.0432

sigma^2 = 0.009628:  log likelihood = 436.33
AIC=-862.66   AICc=-862.53   BIC=-841.76
> coeftest(mn)

z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)    
ar1   0.136110   0.058228   2.3375   0.01941 *  
ar2   0.287894   0.060249   4.7784 1.767e-06 ***
sma1 -0.803914   0.059340 -13.5476 < 2.2e-16 ***
sma2  0.070414   0.043210   1.6296   0.10319    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ARIMA (2,0,0) (0,1,2) [4] model demonstrates significant coefficients for ar1, ar2, and sma1. These coefficients indicate the autoregressive (AR) and seasonal moving average (SMA) components of the model. The model's AIC and AICc values suggest a reasonably good fit.

ACF Residual and Box Test:

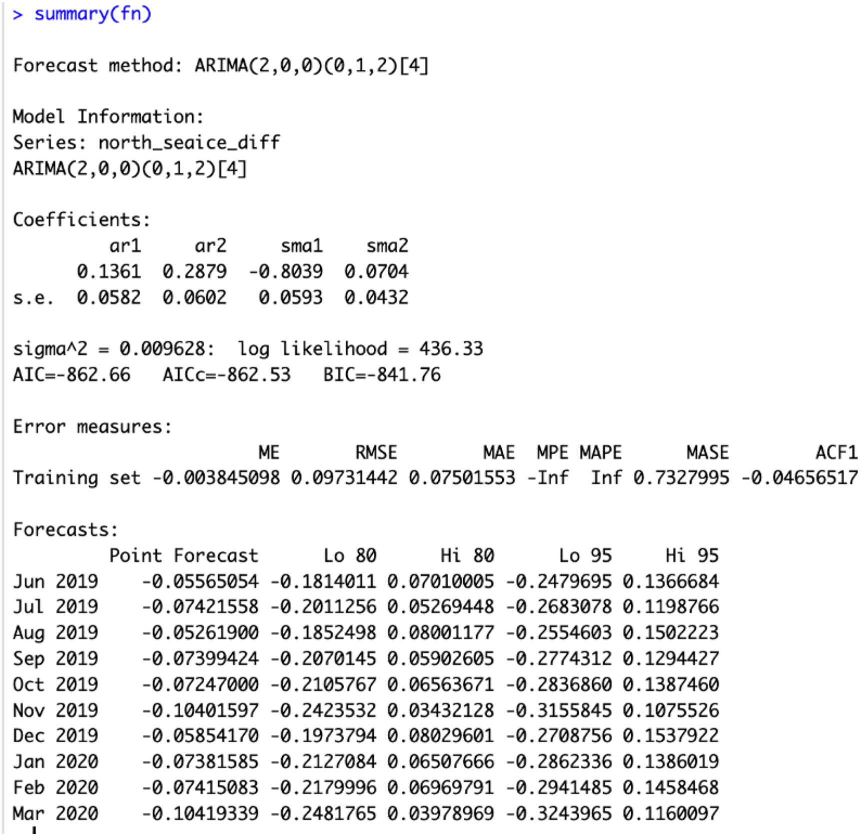




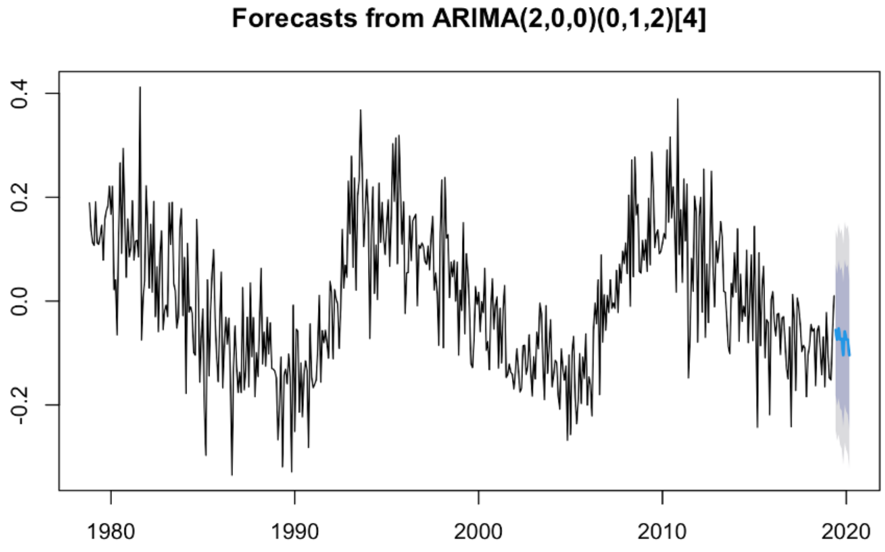
```
> Box.test(mn$residuals, 4, "Ljung-Box")

	Box-Ljung test

data:  mn$residuals
X-squared = 5.7641, df = 4, p-value = 0.2175
```

From the residuals ACF plot, the value decays very quickly close to zero and the p-value of Box-Ljung test are all larger than 0.05, indicating non-significate. Both the results of the ACF plot and Box-Ljung test are good, which means this model is good for forecasting.

Forecast:



```
> summary(fn)

Forecast method: ARIMA(2,0,0)(0,1,2)[4]

Model Information:
Series: north_seaice_diff
ARIMA(2,0,0)(0,1,2)[4]

Coefficients:
         ar1     ar2     sma1    sma2
      0.1361  0.2879  -0.8039  0.0704
s.e.  0.0582  0.0602   0.0593  0.0432

sigma^2 = 0.009628:  log likelihood = 436.33
AIC=-862.66   AICc=-862.53   BIC=-841.76

Error measures:
                       ME       RMSE        MAE  MPE MAPE      MASE        ACF1
Training set -0.003845098 0.09731442 0.07501553 -Inf  Inf 0.7327995 -0.04656517

Forecasts:
         Point Forecast      Lo 80      Hi 80      Lo 95     Hi 95
Jun 2019    -0.05565054 -0.1814011 0.07010005 -0.2479695 0.1366684
Jul 2019    -0.07421558 -0.2011256 0.05269448 -0.2683078 0.1198766
Aug 2019    -0.05261900 -0.1852498 0.08001177 -0.2554603 0.1502223
Sep 2019    -0.07399424 -0.2070145 0.05902605 -0.2774312 0.1294427
Oct 2019    -0.07247000 -0.2105767 0.06563671 -0.2836860 0.1387460
Nov 2019    -0.10401597 -0.2423532 0.03432128 -0.3155845 0.1075526
Dec 2019    -0.05854170 -0.1973794 0.08029601 -0.2708756 0.1537922
Jan 2020    -0.07381585 -0.2127084 0.06507666 -0.2862336 0.1386019
Feb 2020    -0.07415083 -0.2179996 0.06969791 -0.2941485 0.1458468
Mar 2020    -0.10419339 -0.2481765 0.03978969 -0.3243965 0.1160097
```

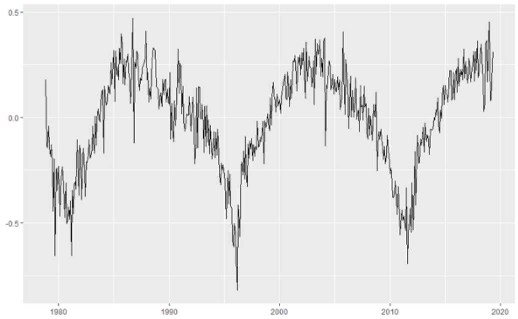


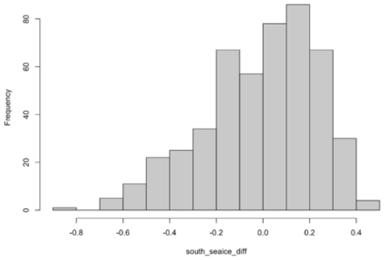
By using the SARIMA model, we also predict one year period after the last month of the original dataset.

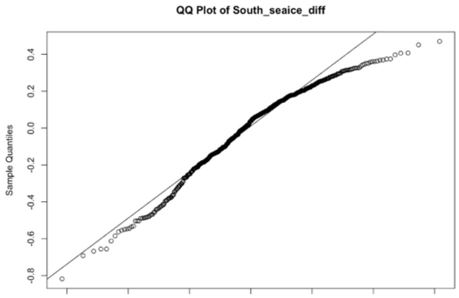
The model forecasts converge to a mean value of approximately 0.0076. This means that as the forecast horizon increases, the predicted values tend to approach this mean value.

Taking difference of model for stationarity:

South Model:

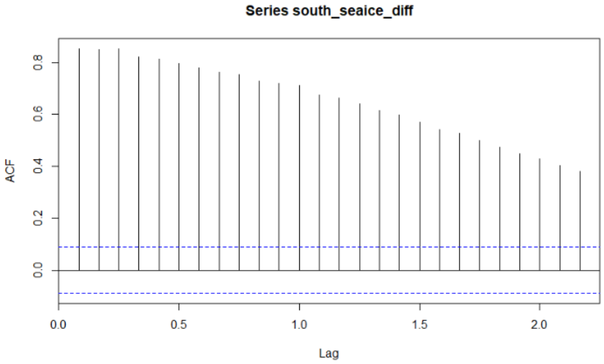


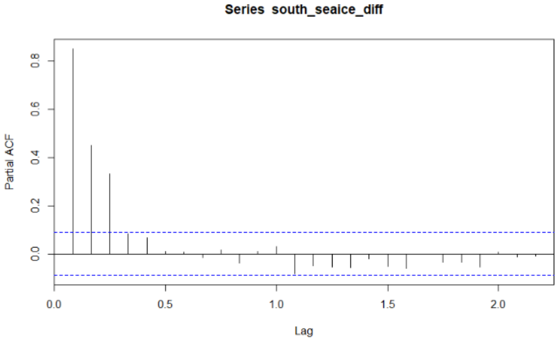




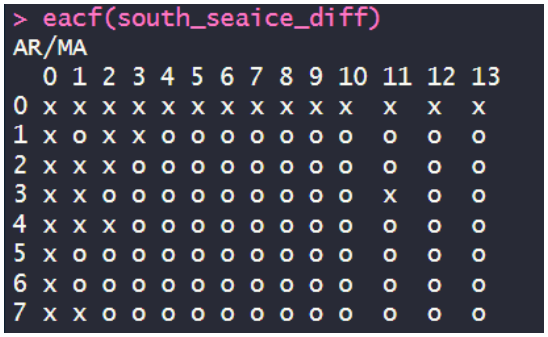
From the histogram we can see that the graph is slightly skewed on the right side and from the Q-Q plot we can see that some points are not on the line they have more variance compared to others which indicates it is not normally distributed

Checking Stationary for difference south model:





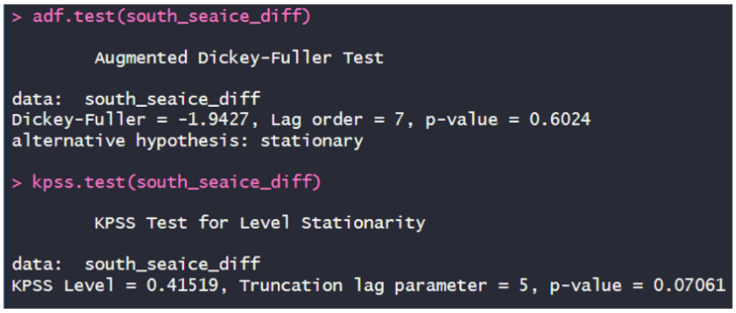
The autocorrelation coefficients are shown, and we can see that they typically decrease progressively as the latency grows. Even at longer delays, the coefficients are still rather high, demonstrating that the series still exhibits considerable autocorrelation. Which indicates the series is stationary.



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

The presence of "0" in the first column for such orders indicates that the AR order is most likely to be either 1 or 2.

The occurrence of "0" in the second, third, and fourth rows for those orders indicates that the MA order is probably either 1, 2, or 3.



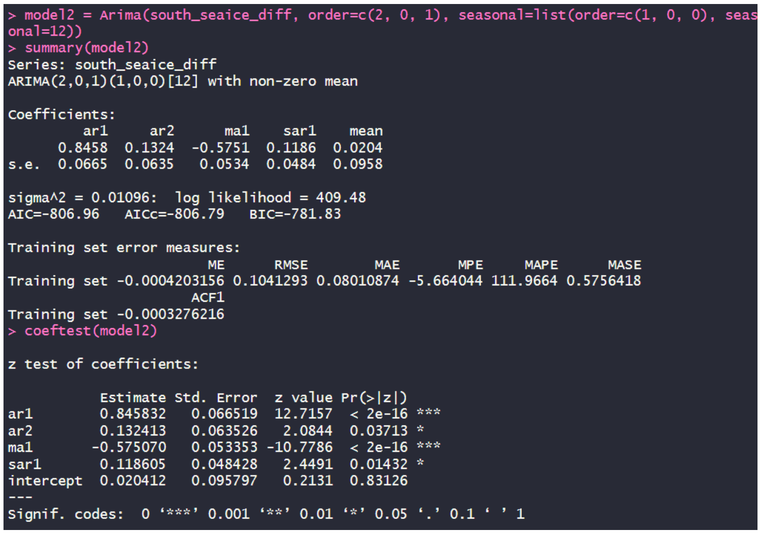
```
> adf.test(south_seaice_diff)

        Augmented Dickey-Fuller Test

data:  south_seaice_diff
Dickey-Fuller = -1.9427, Lag order = 7, p-value = 0.6024
alternative hypothesis: stationary

> kpss.test(south_seaice_diff)

        KPSS Test for Level Stationarity

data:  south_seaice_diff
KPSS Level = 0.41519, Truncation lag parameter = 5, p-value = 0.07061
```

The p-value in this instance (0.6024) exceeds the usually accepted significance level of 0.05, we lack sufficient data to reject the null hypothesis. Consequently, the'south_seaice_diff' series is probably non-stationary based on the findings of the ADF test.

For, p-value (0.07061) above the frequently accepted significance level of 0.05, the null hypothesis cannot be ruled out. The'south_seaice_diff' series is thus expected to be level or trend stagnant based on the KPSS test findings.

Model for Difference South Data:

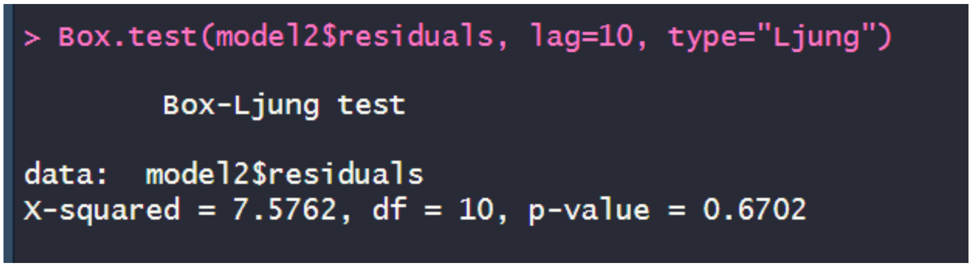
Manual Model:



```
> model2 = Arima(south_seaice_diff, order=c(2, 0, 1), seasonal=list(order=c(1, 0, 0), seas
onal=12))
> summary(model2)
Series: south_seaice_diff
ARIMA(2,0,1)(1,0,0)[12] with non-zero mean

Coefficients:
         ar1     ar2      ma1    sar1    mean
      0.8458  0.1324  -0.5751  0.1186  0.0204
s.e.  0.0665  0.0635   0.0534  0.0484  0.0958

sigma^2 = 0.01096:  log likelihood = 409.48
AIC=-806.96   AICc=-806.79   BIC=-781.83

Training set error measures:
                        ME      RMSE        MAE       MPE     MAPE      MASE
Training set -0.0004203156 0.1041293 0.08010874 -5.664044 111.9664 0.5756418
                      ACF1
Training set -0.0003276216
> coeftest(model2)

z test of coefficients:

           Estimate Std. Error  z value Pr(>|z|)    
ar1        0.845832   0.066519  12.7157  < 2e-16 ***
ar2        0.132413   0.063526   2.0844  0.03713 *  
ma1       -0.575070   0.053353 -10.7786  < 2e-16 ***
sar1       0.118605   0.048428   2.4491  0.01432 *  
intercept  0.020412   0.095797   0.2131  0.83126    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model has an AIC of -806.96, an AICc of -806.79, and a log likelihood of 409.48. The MAPE (Mean Absolute Percentage Error) is 111.97, while the RMSE is 0.1041.

The statistical significance of the coefficients for ar1, ar2, and ma1 highlights the significance of these variables in the model. Statistical significance is also seen for the sar1 coefficient.

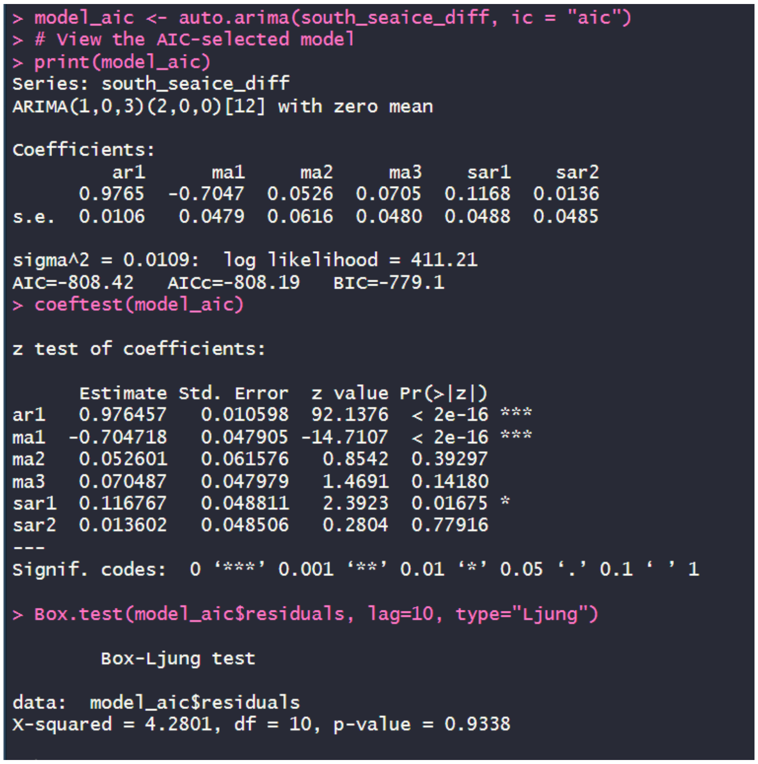


```
> Box.test(model2$residuals, lag=10, type="Ljung")

        Box-Ljung test

data:  model2$residuals
X-squared = 7.5762, df = 10, p-value = 0.6702
```

The test results show that the X-squared test statistic has 10 degrees of freedom and is 7.5762. The test statistic's corresponding p-value is 0.6702. We are unable to reject the null hypothesis since the p-value is higher than the customary significance level of 0.05.

As a result, model2's residual autocorrelation is not significantly supported by the Box-Ljung test. This may indicate that the model accurately depicts the data's autocorrelation structure.

Auto Arima Model:



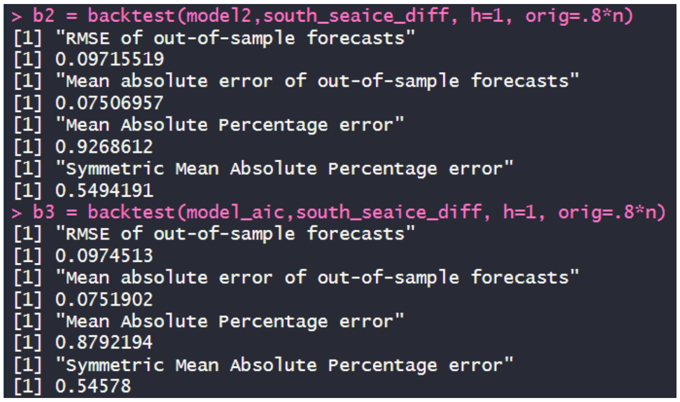
```
> model_aic <- auto.arima(south_seaice_diff, ic = "aic")
> # View the AIC-selected model
> print(model_aic)
Series: south_seaice_diff
ARIMA(1,0,3)(2,0,0)[12] with zero mean

Coefficients:
         ar1      ma1     ma2     ma3    sar1    sar2
      0.9765  -0.7047  0.0526  0.0705  0.1168  0.0136
s.e.  0.0106   0.0479  0.0616  0.0480  0.0488  0.0485

sigma^2 = 0.0109:  log likelihood = 411.21
AIC=-808.42   AICc=-808.19   BIC=-779.1
> coeftest(model_aic)

z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ar1   0.976457   0.010598  92.1376  < 2e-16 ***
ma1  -0.704718   0.047905 -14.7107  < 2e-16 ***
ma2   0.052601   0.061576   0.8542  0.39297
ma3   0.070487   0.047979   1.4691  0.14180
sar1  0.116767   0.048811   2.3923  0.01675 *
sar2  0.013602   0.048506   0.2804  0.77916
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Box.test(model_aic$residuals, lag=10, type="Ljung")

        Box-Ljung test

data:  model_aic$residuals
X-squared = 4.2801, df = 10, p-value = 0.9338
```

The selected model's coefficients and standard errors are displayed. The importance of each coefficient is shown by the p-values for the coefficient estimations. The other coefficients are not significant in this situation, but the AR (1) and MA (1) coefficients are very significant (p 2e-16).

Using the Ljung-Box test type and a lag setting of 10, the Box-Ljung test was run on the residuals of the chosen model. The test results show that the X-squared test statistic has a value of 4.2801 and ten degrees of freedom. The corresponding p-value is 0.9338, above the usual significant threshold of 0.05. Therefore, based on the Box-Ljung test, there is no meaningful evidence of residual autocorrelation in the chosen model.

These results imply that the model fits the series well.

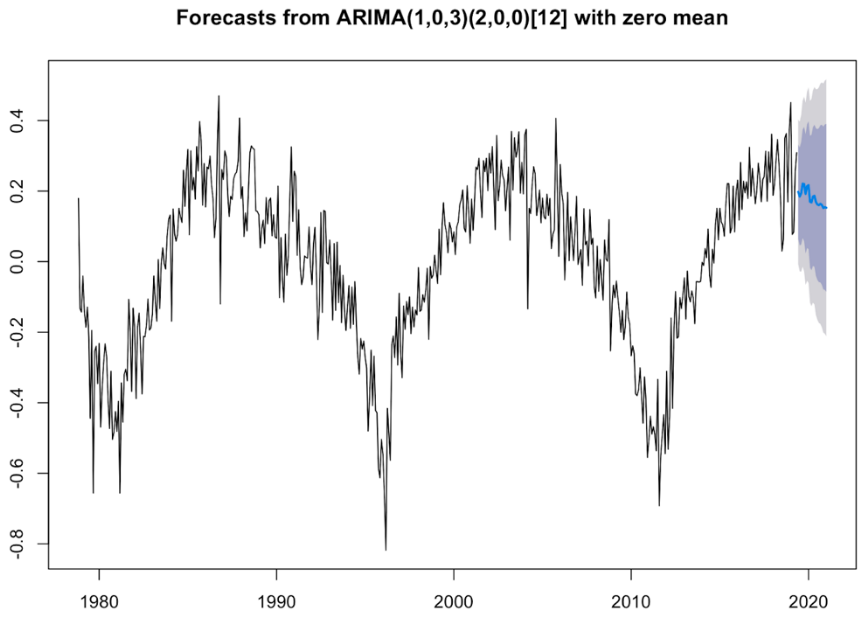
Back testing:



```
> b2 = backtest(model2,south_seaice_diff, h=1, orig=.8*n)
[1] "RMSE of out-of-sample forecasts"
[1] 0.09715519
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.07506957
[1] "Mean Absolute Percentage error"
[1] 0.9268612
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.5494191
> b3 = backtest(model_aic,south_seaice_diff, h=1, orig=.8*n)
[1] "RMSE of out-of-sample forecasts"
[1] 0.0974513
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.0751902
[1] "Mean Absolute Percentage error"
[1] 0.8792194
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.54578
```

The AIC-selected model (ARIMA (1,0,3) (2,0,0) [12]) can be thought of as good possibilities.

Forecast:



Our south hemisphere forecast got the mean error of 0.00034 which is closer to zero means our model captures the patterns and trends very accurately also the low MAE of 0.079 suggests that differences between actual value and forecasted values are very small that suggest the model may be bias but in our data has strong consistent and predictable trend and because of that it gives the accurate pattern in forecast.

Conclusion:

In conclusion, the analysis conducted using various models on the dataset has yielded highly promising results in forecasting sea ice extent changes. As expected, the findings have confirmed the widely acknowledged fact that the extent of sea ice is declining over time.

Throughout this project, our team has gained valuable knowledge and developed essential skills by leveraging the power of R for time series prediction with diverse models. Undoubtedly, these newly acquired competencies will prove immensely beneficial in our future endeavors.

Contributions:

Dhruv Dobariya

Milestone 2b: From the very beginning of milestone 1, I actively began exploring popular subjects for the project. I extensively searched the internet in search of inspiration and made concerted efforts to brainstorm a compelling topic. I am always fascinated by nature as I come from a farming background from India. As a son of farmer, I see nature closely. I know how the environment affects farming in India does. I saw Sidhant's post about Sea Ice Extent. I thought that this might be something which I can work on. I got in contact with him. At my first meeting with him, he presented the dataset. Although I had seen before our meeting. He helped me understand the dataset very well. I started brainstorming ideas to him to approach this dataset. Milestone 3: For this milestone, our group got finalized. We had our initial meetings. We had brainstorm ideas for the approach. I pitched an approach for the project. Later on, we had meeting with professor about our approach. Professor recommended some different approach than ours. Professor told us to split the group in 2 and also the dataset in 2 which were of North and South. After meeting with professor, we had internal meeting and finalized the approach. I along with Parth were going to work on South data and Sidhant along with Saransh were going to work on North data. I and Parth had a separate meeting where we did our part for Milestone 2. I along with Parth did some exploratory analysis on the South. Milestone 4: For this milestone, we had meeting with professor. After the meeting we decided to move further ahead. Parth started working on south dataset to generate some plots like Histogram, QQ plot. We made the data stationary by taking difference. I made the models for this milestone and compared it with auto Arima. Parth also made 1 model. After doing so I checked the all the models and compared the results. I took the model with best performance. I applied the backtest on the all the models. Parth checked which was giving the best performance. At end Parth applied residuals and evaluated final model. Parth also wrote some documentations.

Parth Patel

Milestone 2b :

I actively started looking into popular datasets for the project at the very start of milestone 1. I did a lot of research on kaggle and other websites, and I especially liked the information on https://www.ncei.noaa.gov/access/search/data-search/daily-summaries. I asked Sidhhant for these datasets in a conversation where he had previously posted the Sea Ice Extent. I felt that I might be able to work on this. We subsequently met and talked about the proposal.

Milestone 3:

We finalized our group for this milestone. We met and opted to take the average of the ice extent on the north and south poles after I gave some suggestions for what we could do with this set of data. We discussed our strategy with the professor in a meeting that followed. Professor suggested several alternative methods to our own. The professor instructed us to divide the group into two and the dataset, which included both North and South, into two. We had an internal discussion after our meeting with the professor to finalize our strategy. I was going to work on the southern hemisphere with Dhruv D, while Sidhant T and Saransh T were going to work on the northern hemisphere. In a separate meeting, Dhruv D and I completed our tasks for Milestone 2. I conducted some exploratory study on the South Pole with Dhruv D.

Milestone 4:

We had a meeting with the professor for this milestone. We made the decision to proceed after the meeting. I started generating charts like the histogram and QQ plot using the south dataset. I tried to segment the time series I had constructed in order to evaluate the mean variance difference and determine the type of series apart from some other testing. Next, we developed the models for the series and contrasted them with models generated by auto Arima. Next, we selected the model that performed the best. On all of the models, we ran the backtest. We opted not to construct a GARCH model and instead reviewed the final model that was chosen and completed the documentation after I performed various tests and graphs to see which was providing the greatest performance.

Sidhant Thakur

We started a project to analyze and predict changes in sea ice extent over time. I found a suitable dataset on Kaggle and shared it with my classmates. We formed a team with Dhruv, Parth, Saransh, and me to work on the project together. Our professor suggested that we split into two groups to handle the dataset. Saransh and I worked on the data for the North Hemisphere, while Dhruv and Parth focused on the South Hemisphere data. To work effectively as a team, we had a meeting where we discussed and finalized our approach. At the beginning of the project, my main task was to prepare the North Hemisphere data for analysis. I dealt with missing values and inconsistent date formats in the dataset. By merging the date columns into a single format, I made the data consistent for further analysis. Next, I explored the North Hemisphere Sea ice extent data to gain insights. I used visualizations like histograms to understand the data's distribution. I also used the QQ plot to check if the data followed a normal distribution and conducted the Jarque-Bera test for further confirmation. The autocorrelation and partial autocorrelation plots helped identify patterns that could guide our forecasting models. I also examined the stationarity of the time series data using statistical tests like the Ljung-Box test, Dickey-Fuller test, and KPSS test. These tests helped determine the appropriate differencing needed to achieve stationarity, which was important for our modeling.

Saransh Thakur:

Throughout the project, my contributions have been primarily focused on model evaluation and refinement, working collaboratively with Sidhant to improve the accuracy of our analysis.

One of my key responsibilities was assessing the performance of the selected models. To accomplish this, I employed various evaluation metrics, including root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (MAPE). By comparing the performance of different models using these metrics, I obtained a comprehensive understanding of their strengths and weaknesses.

To ensure the validity of our models, I meticulously examined the residuals using the Box-Ljung test, aiming to identify any significant autocorrelation. By validating the model assumptions, I ensured the reliability of our analysis.

Based on the initial evaluation, I collaborated closely with Sidhant to refine our chosen models. We explored alternative model configurations, considering different orders of autoregressive (AR) and moving average (MA) terms to capture the underlying patterns in the data more accurately. Additionally, I evaluated the potential inclusion of exogenous variables to enhance the predictive performance of our models. Through this iterative process, we continuously fine-tuned and optimized the models, significantly improving the overall accuracy of our forecasts.

Refinement of Models:

Building upon the initial evaluation and feedback received, Sidhant and I will continue refining the selected models. We will explore alternative model configurations, optimize hyperparameters, and assess the potential inclusion of exogenous variables. This iterative process aims to enhance the accuracy and reliability of our forecasting models, ensuring they capture the complex dynamics of sea ice extent accurately.

Final Forecasting:

Once our models have been refined, Sidhant and I will generate final forecasts for the North Hemisphere Sea ice extent. We will conduct a thorough analysis of the forecasted values, calculating prediction intervals to provide insights into the uncertainty associated with the forecasts. Incorporating visualizations such as time series plots and forecasted trajectories will aid in presenting a comprehensive understanding of the projected sea ice extent.

Project Report Preparation:

In parallel with the modeling and forecasting tasks, Sidhant and I will collaborate on preparing the final project report. We will consolidate our individual contributions, describe the methodology we employed, summarize our findings, and discuss the implications of our analysis. The report will adhere to the guidelines provided, effectively communicating the project's objectives, methods, and outcomes. By presenting our work cohesively, we aim to provide a comprehensive and insighful report.

Conclusion:

Sidhant and I have made significant contributions to the Sea Ice Extent Time Series Analysis and Forecasting project. Through our collaborative efforts, we have conducted exploratory data analysis, evaluated and refined models, and laid the foundation for accurate forecasting, focus on further refining the models, generating final forecasts, and preparing the project report. By leveraging our combined skills and expertise, our project aims to deliver accurate forecasts and valuable insights into the behavior of sea ice extent in the North Hemisphere.