**DSC 425 Time Series Analysis Forecasting**

**Project: Sea Ice Extent**

## Exploratory Data Analysis

```
> 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
>
```

We removed the missing column and source data from our data, and after that combine the year, month, day into a single column and gave that column new name called date.

```
> 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
>
```

Moreover, we separate out the hemisphere into two parts i.e., North, and South.

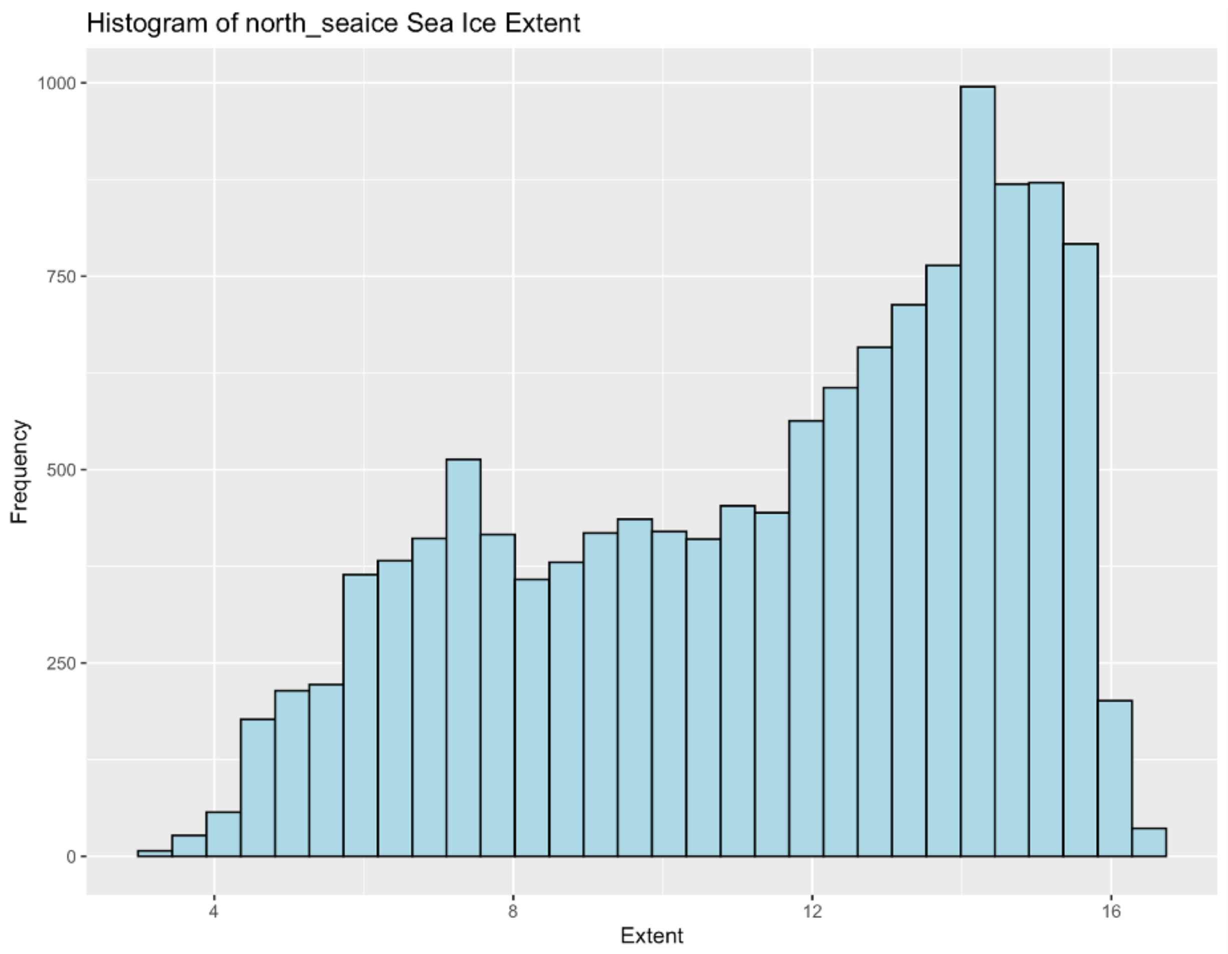
**North Hemisphere:**

```
> north_seaice <- filter(seaice, hemisphere == "north")
> head(north_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
```

Plotting Histogram, QQ plot and Jarque Bera Test to check whether data is normally distributed.

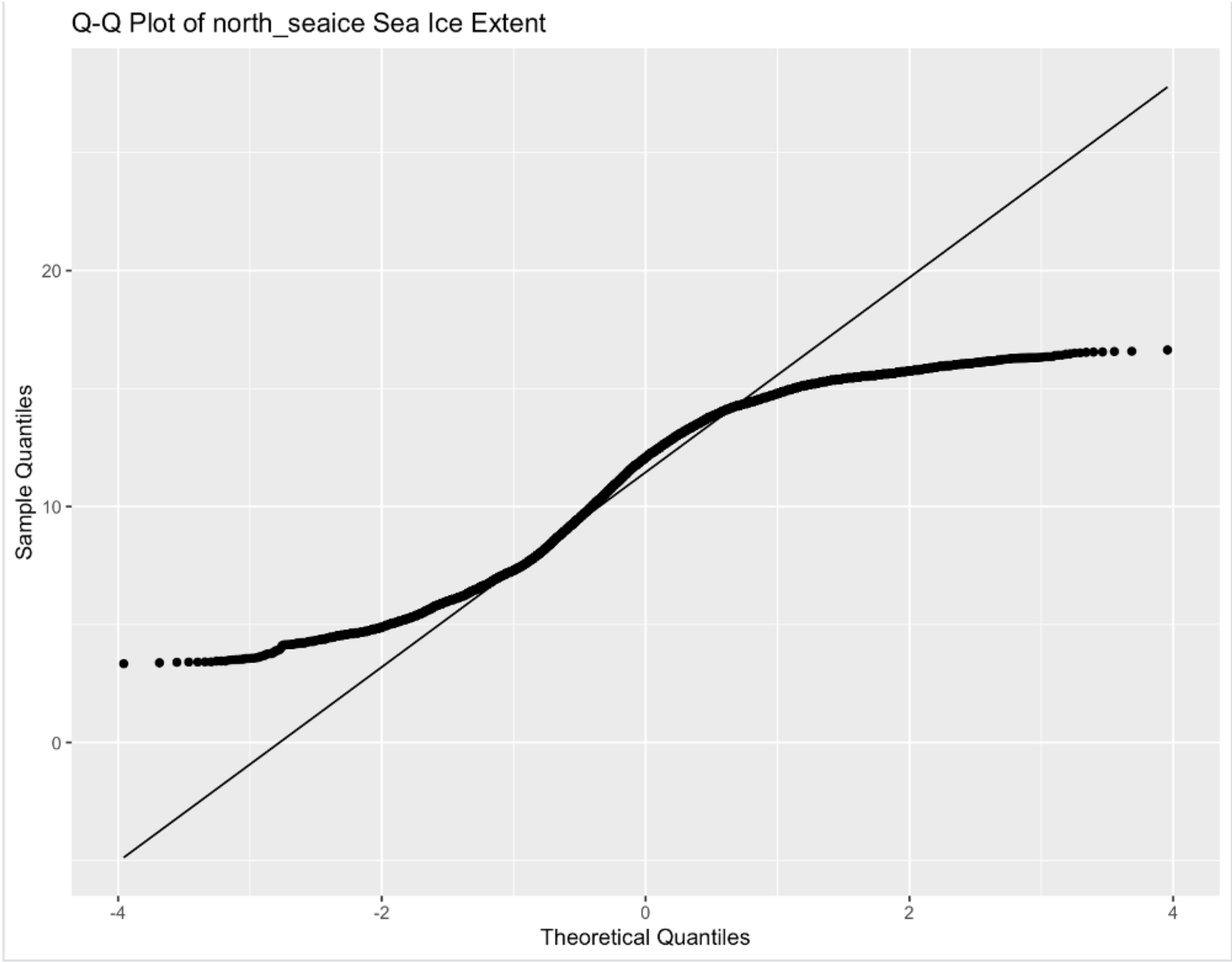
**Histogram:**

```
# Histogram north_seaice
ggplot(north_seaice, aes(x = Extent)) +
  geom_histogram( fill = "lightblue", color = "black") +
  labs(title = "Histogram of north_seaice Sea Ice Extent", x = "Extent", y = "Frequency")
```




**QQ plot:**

```
# Q-Q plot north_seaice
ggplot(north_seaice, aes(sample = Extent)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Q-Q Plot of north_seaice Sea Ice Extent", x = "Theoretical Quantiles", y = "Sample Quantiles")
```




From the above plots, it seems that data is not normally distributed.

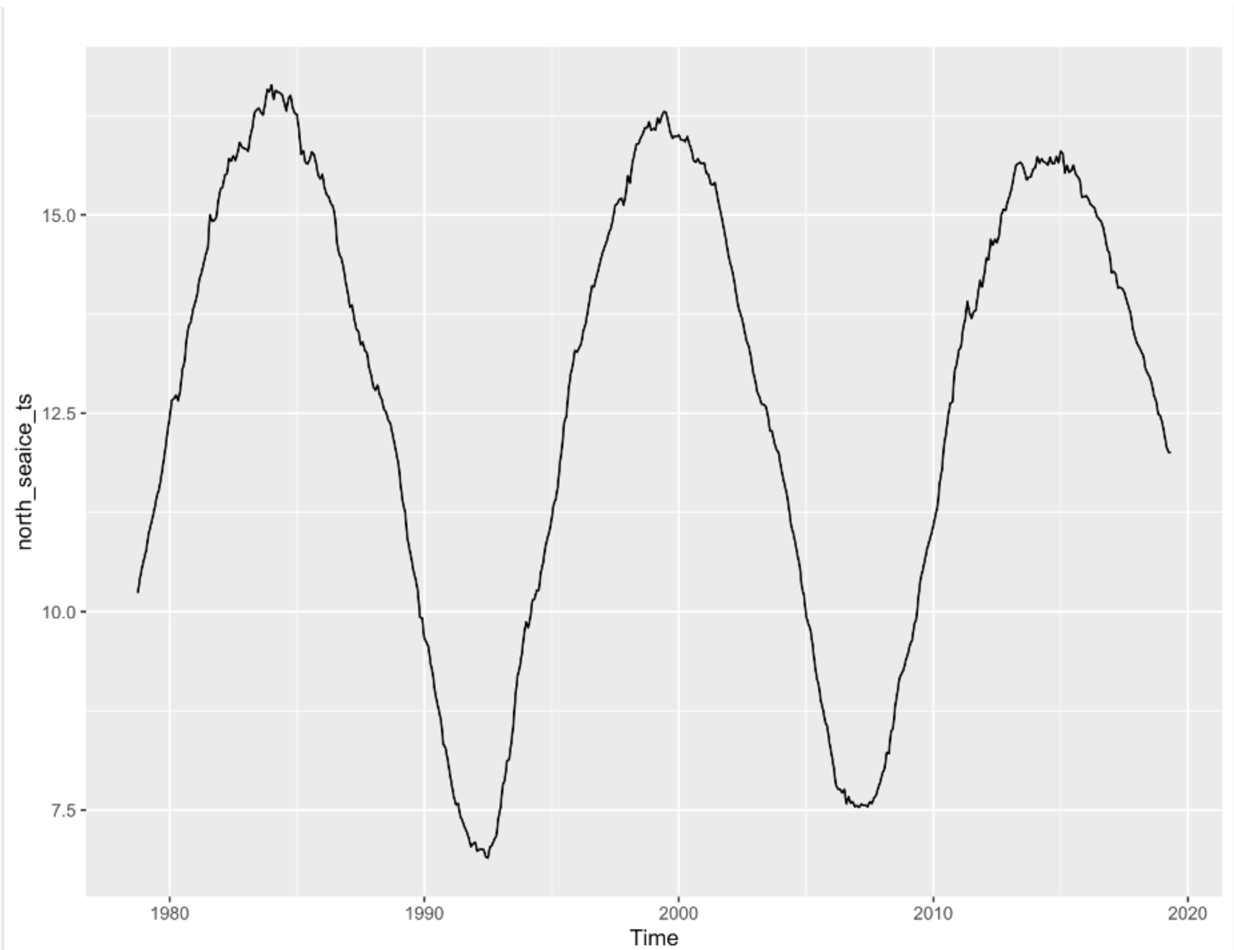
**Jb Test:**

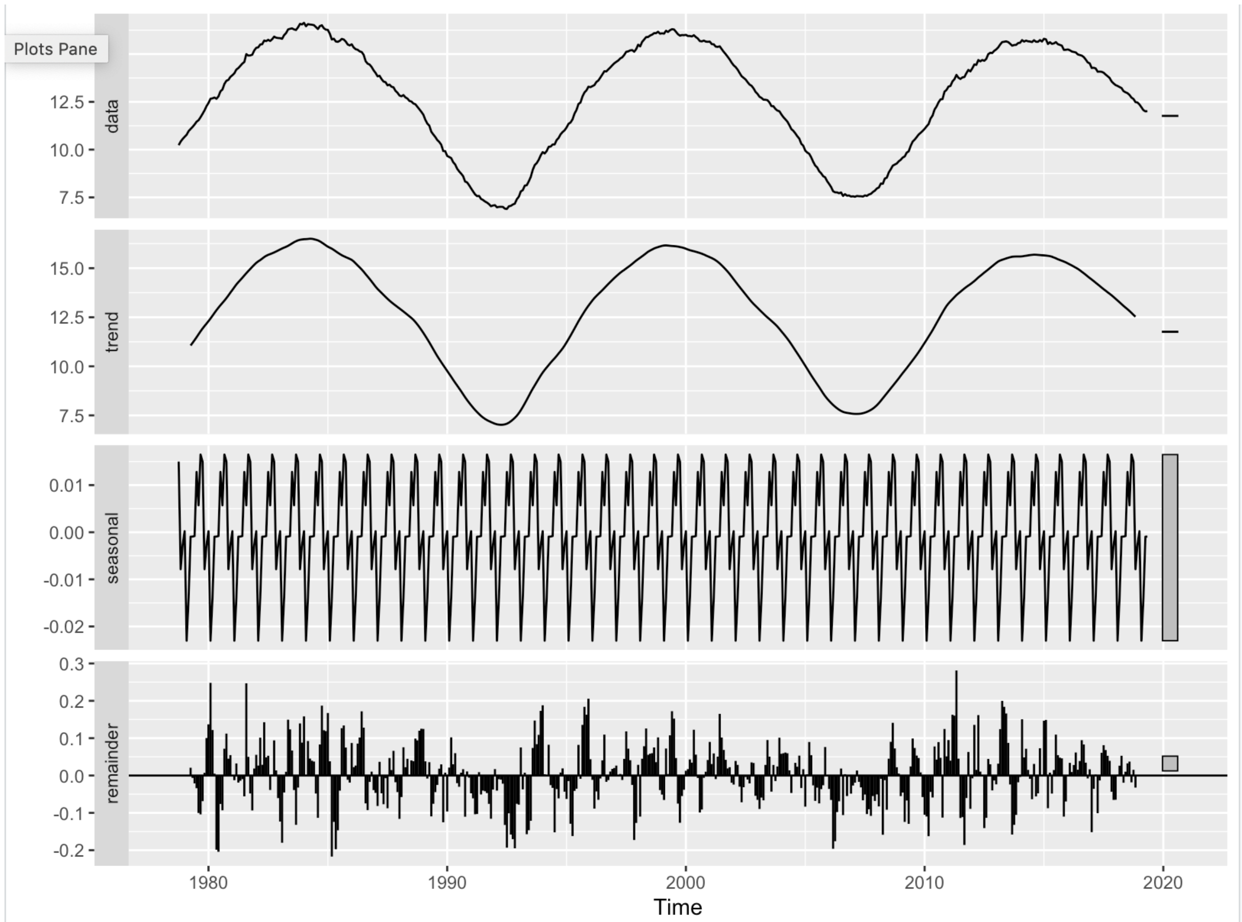
```
> jb_test <- jarque.bera.test(north_seaice$Extent)
>
> # Print the test results
> print(jb_test)

	Jarque Bera Test

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

Based on the Jarque-Bera test results you provided, with a test statistic of 999.59 and a p-value less than 2.2e-16, we can conclude that the data in north_seaice$Extent significantly deviates from a normal distribution. The extremely small p-value suggests strong evidence against the null hypothesis of normality.
Therefore, data does not follow a normal distribution based on this test.

**Creating Time series:**

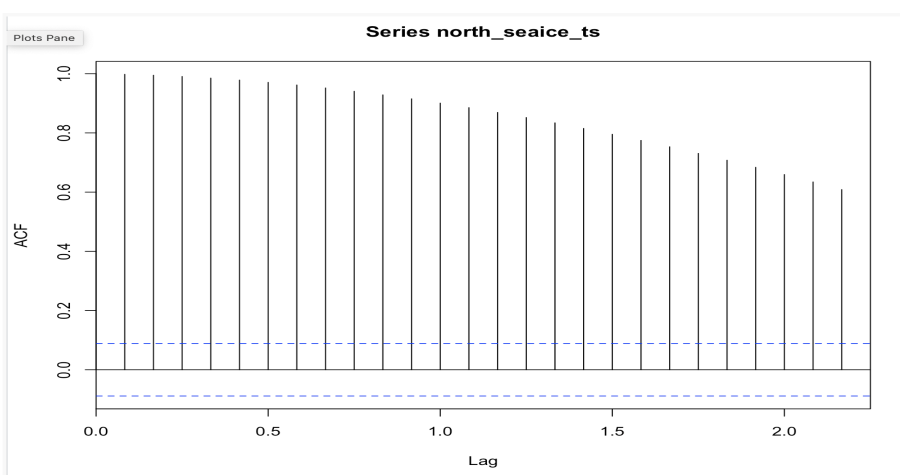
```
> north_seaice_ts = ts(north_seaice $Extent, start=c(1978, 10),end  = c(2019, 5), frequency=12)
> autoplot(north_seaice_ts)
>
```





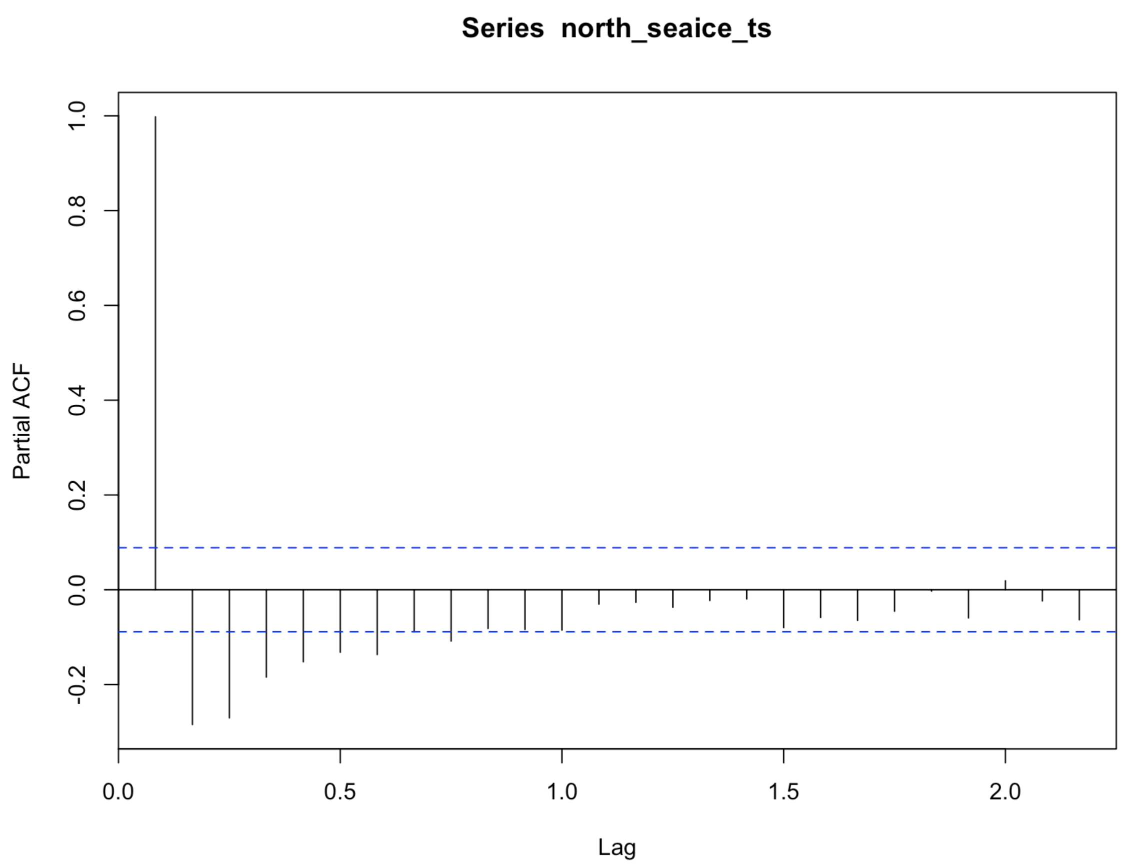

From the above plot we can see that there might be a seasonality present in data and seasonal term is tiny compared to the "remainder", and their appear to be periodic trend

**acf Plot:**



From the above plot we can see that there is a slow decrease, so it is non stationary.

**Pacf Plot:**




From the Pacf plot, there is a significant negative correlation at lag 2,3 and 4.

**Ljung-Box test:**

```
> Box.test(north_seaice_ts, type = "Ljung-Box")

	Box-Ljung test

data:  north_seaice_ts
X-squared = 489.1, df = 1, p-value < 2.2e-16
```

Based on these results, we can conclude that there is strong evidence against the null hypothesis of no autocorrelation in the **north_seaice_ts** data. The extremely small p-value (< 2.2e-16) suggests that there is significant autocorrelation present in the time series.

In summary, the Box-Ljung test indicates that the residuals of the **north_seaice_ts** data exhibit significant autocorrelation at different lags.

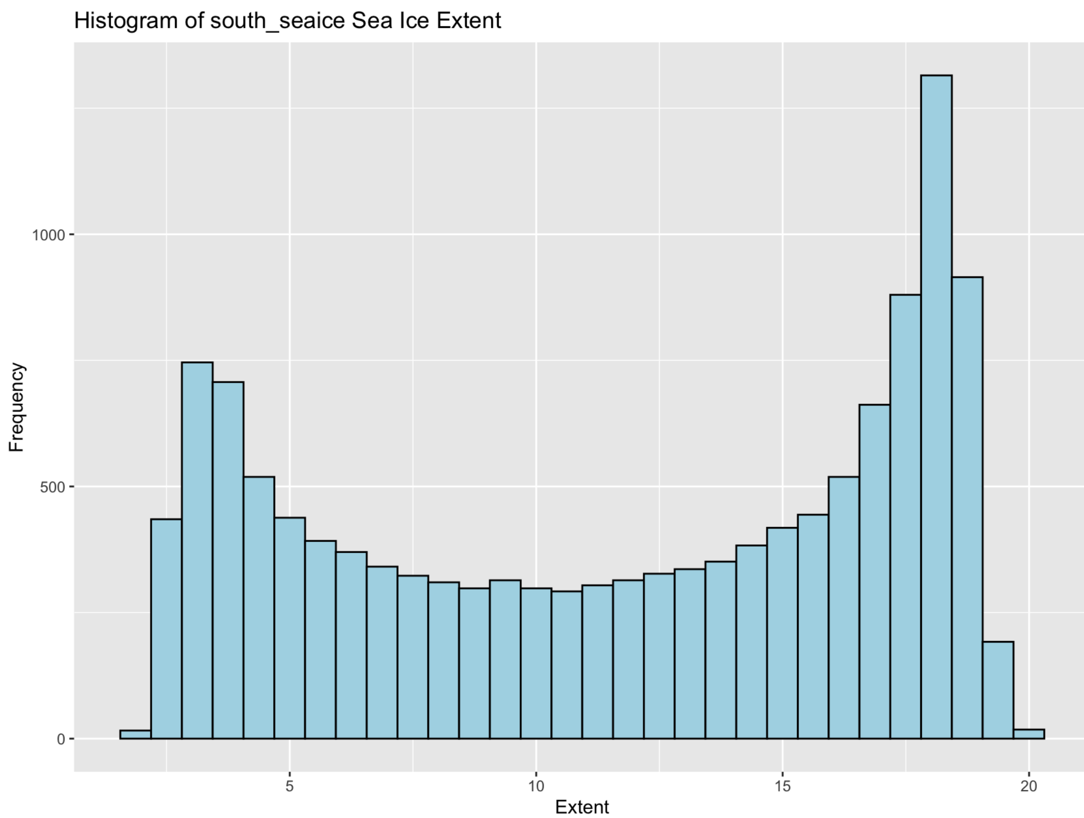
**Dickey-Fuller and an KPSS unit root test:**

```
> # 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
```

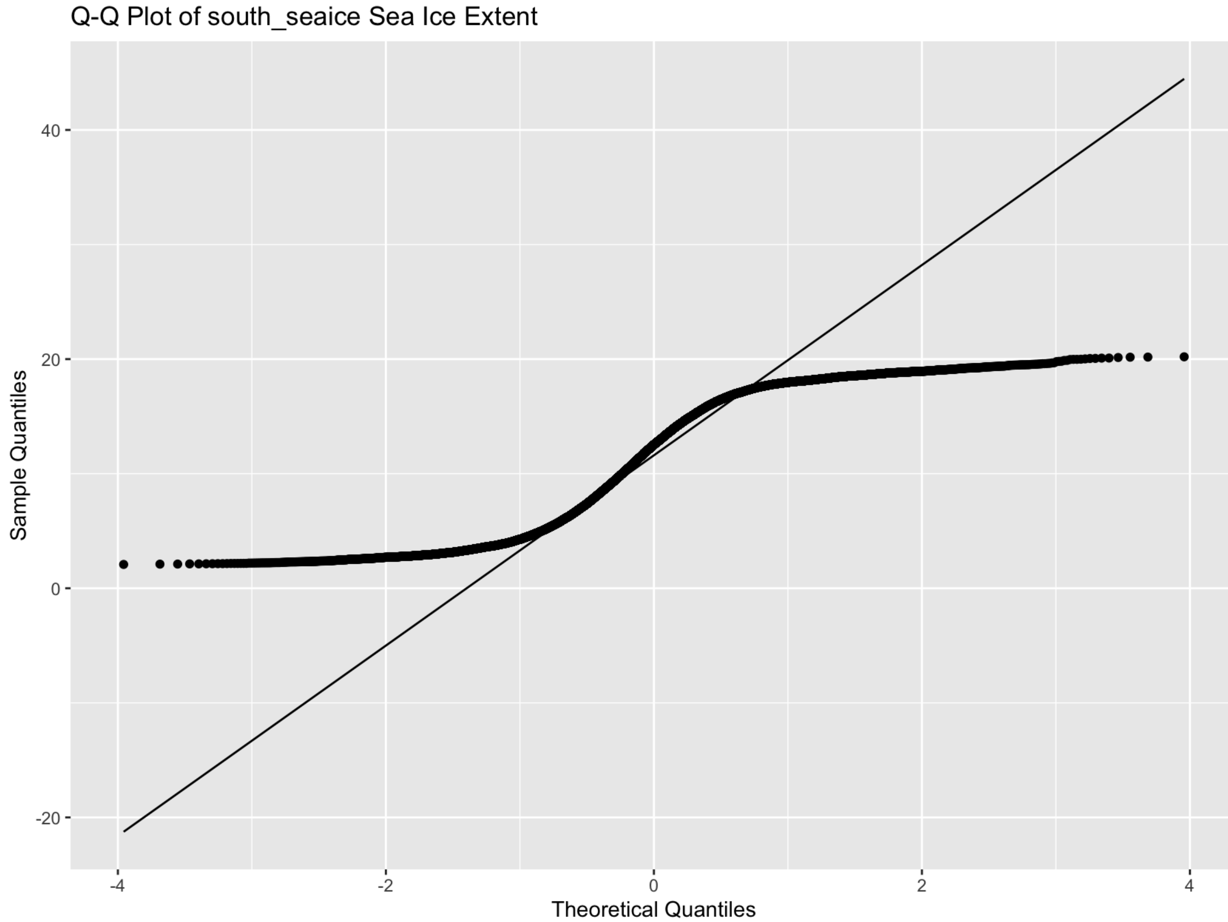
Based on the results of the Dickey-Fuller test, the p-values for all three types (no drift no trend, with drift no trend, and with drift and trend) are greater than 0.05. This suggests that we fail to reject the null hypothesis of the Dickey-Fuller test, indicating that the series is non-stationary. Similarly, the KPSS unit root test results show that the p-values for all three types (no drift no trend, with drift no trend, and with drift and trend) are greater than 0.01. This implies that we fail to reject the null hypothesis of the KPSS test, indicating non-stationarity in the series.
In summary, both the Dickey-Fuller and KPSS tests indicate that the series is non-stationary, meaning it does not exhibit a constant mean and variance over time.

**South Hemisphere:**

Histogram of south_seaice Sea Ice Extent:



**South Hemisphere QQ Plot:J**



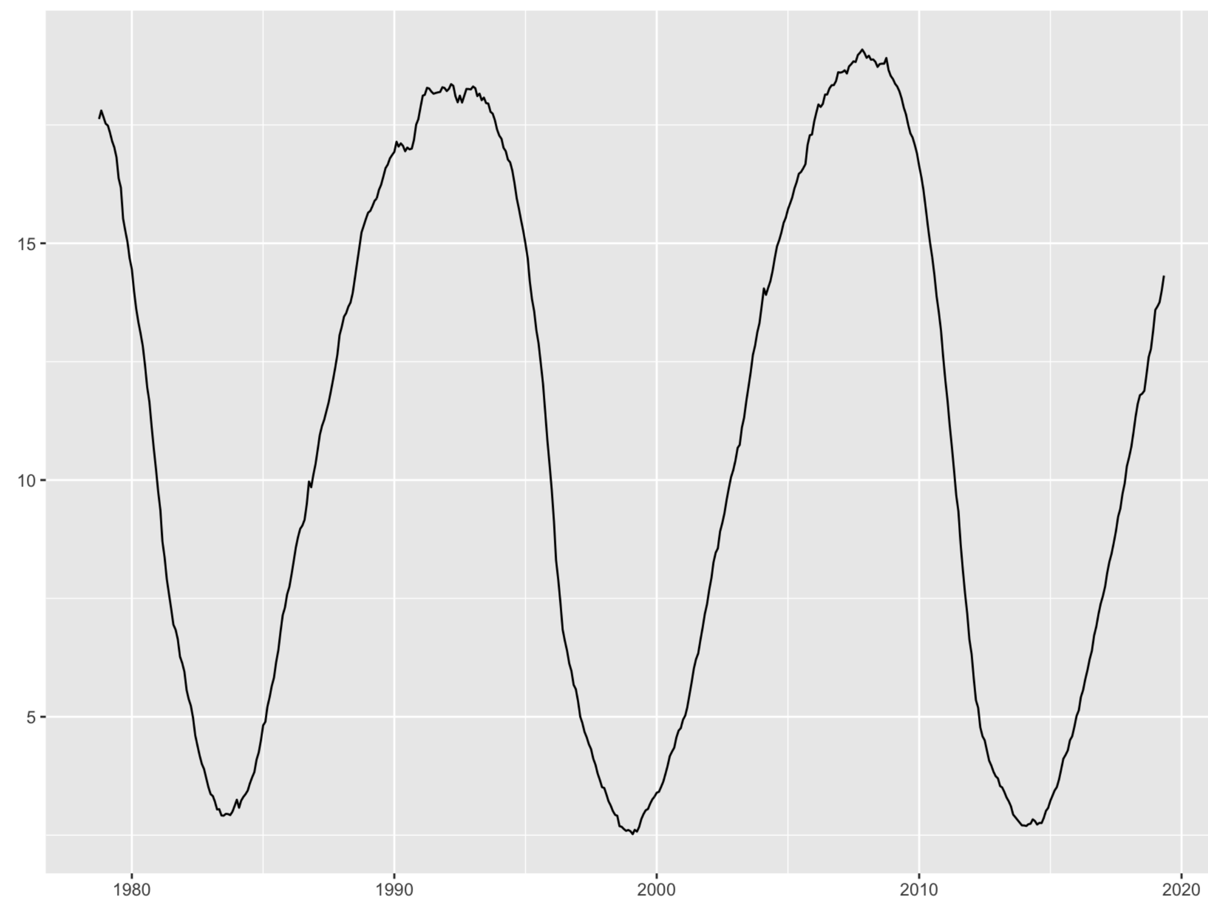

=> Based on the above plot we can interpret that south data is not normally distributed.

**JB Test:**

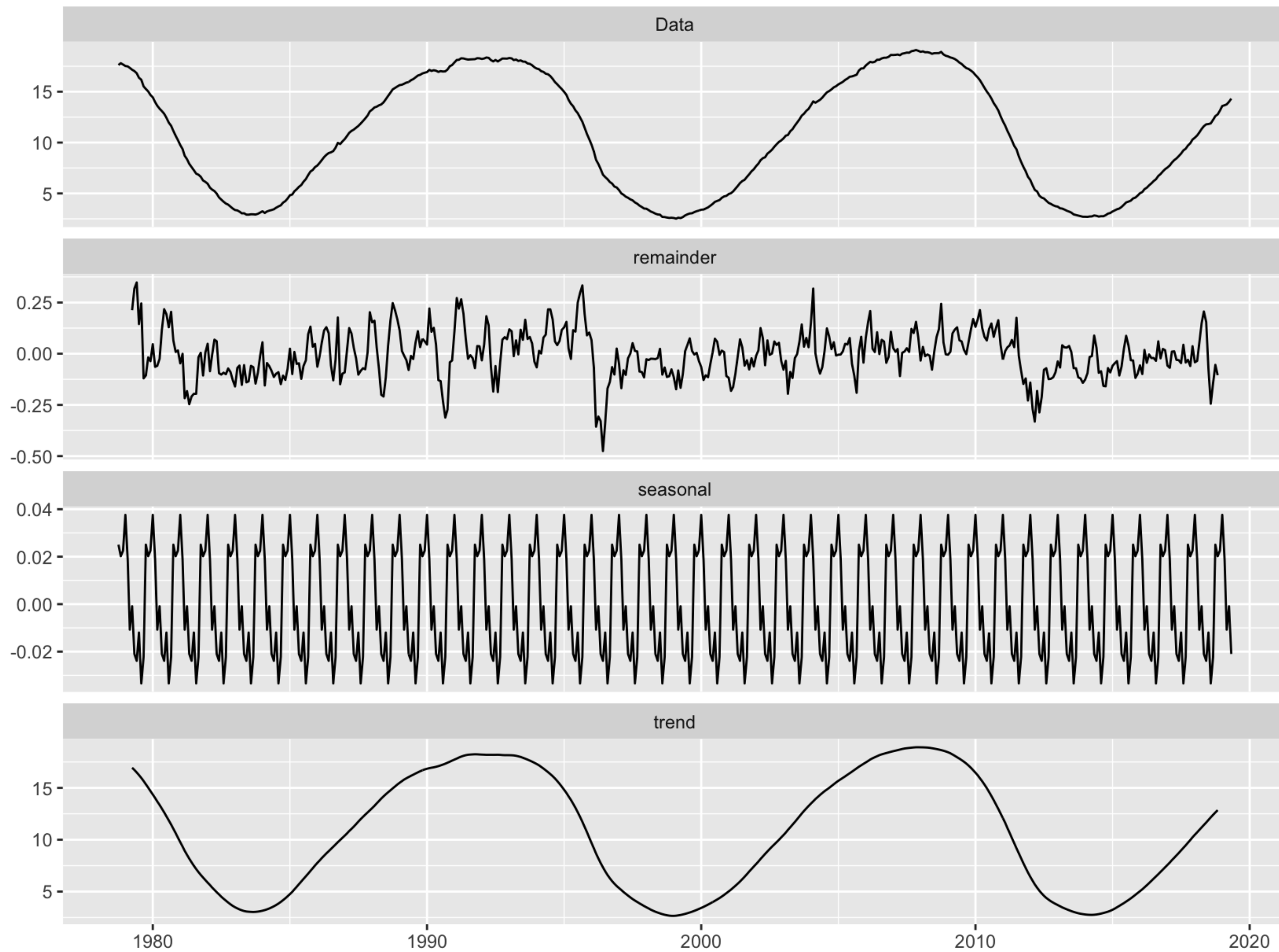
```
> jb_test <- jarque.bera.test(south_seaice$Extent)
> # Print the test results
> print(jb_test)

	Jarque Bera Test

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

>
```

=> Based on the above results of JB test, we see that degree of freedom is 2, X-squared value is 1285.9 and p-values is less than significant value(0.05). The low p-value tells us that we can reject null hypothesis. This suggests us that our data is not normally distributed.

**Creating Time series:**

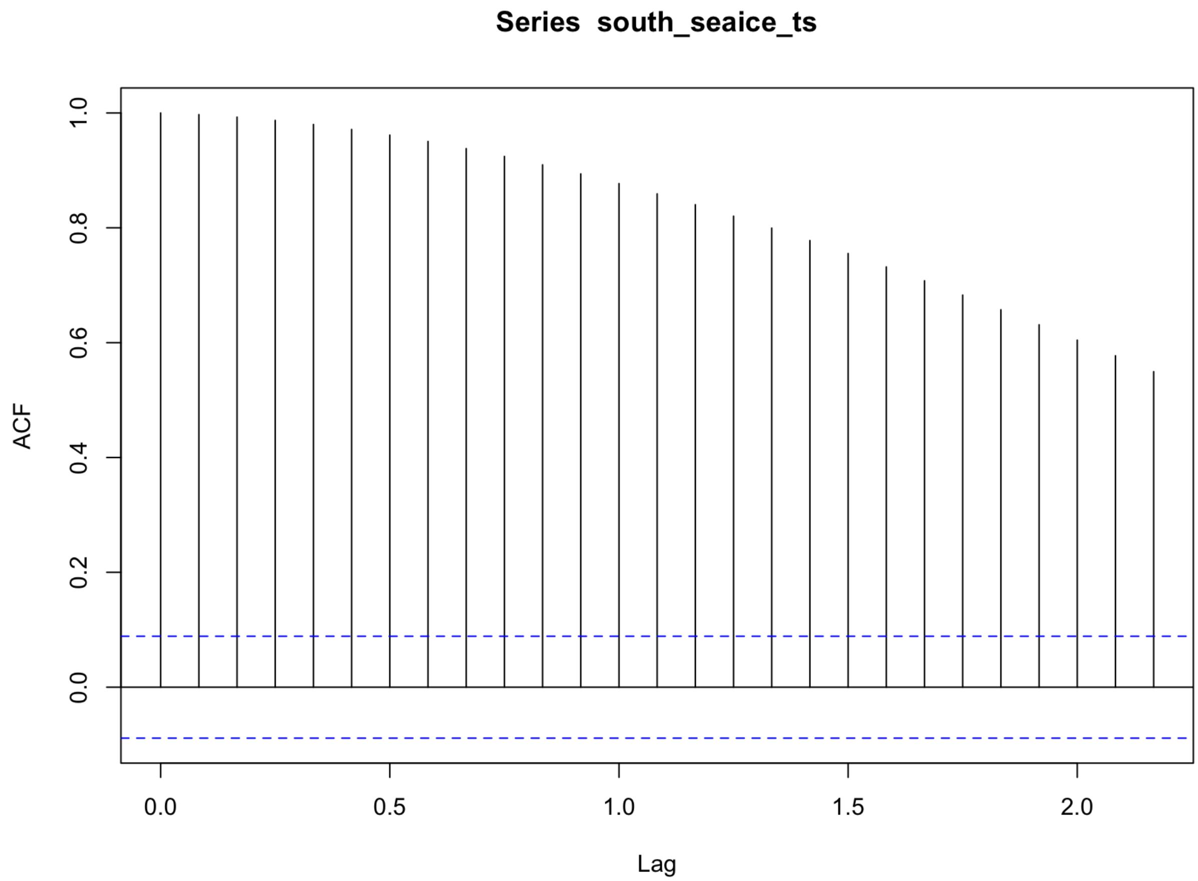


**Decompose**:



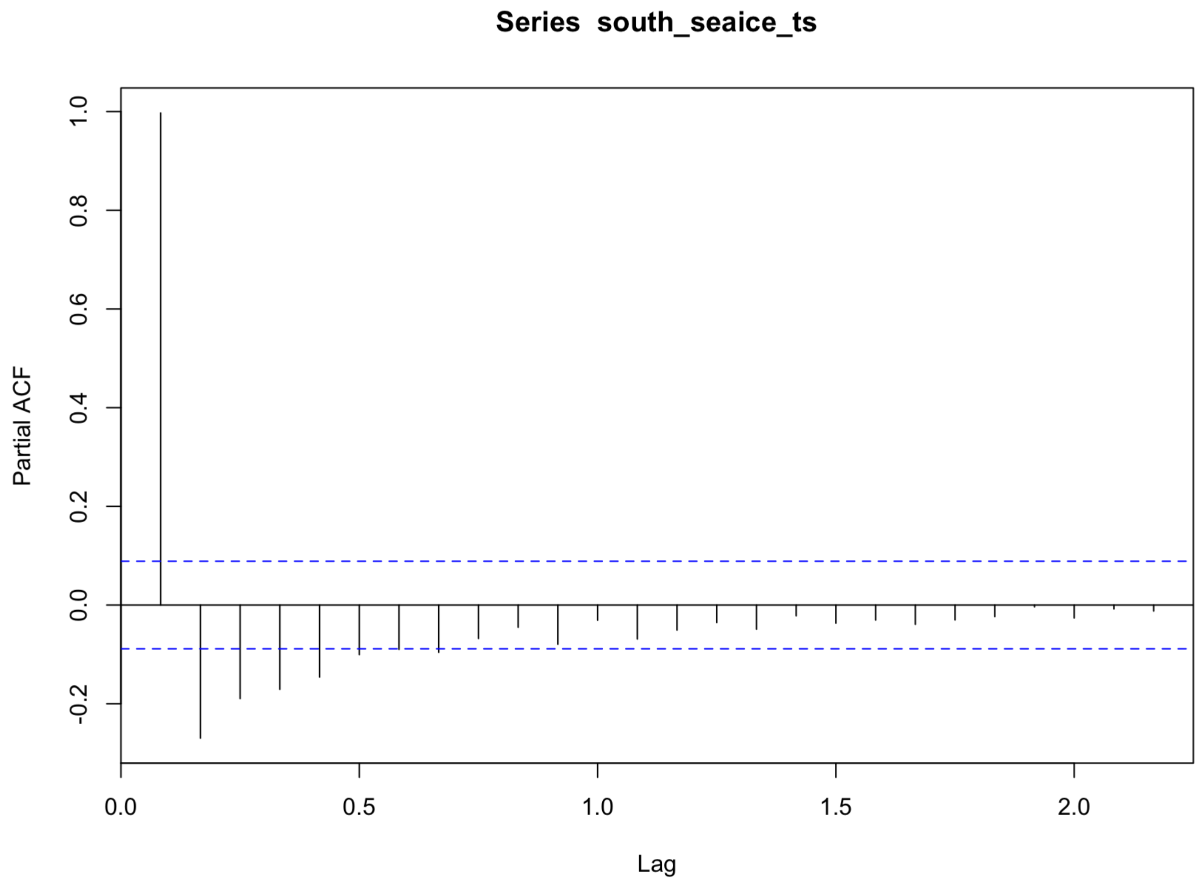

The above plot tell us about trend, seasonality and remainder. The trend component tells us that the some percentage of decline in sea ice extent. The seasonal component tells us that there is some cyclic pattern in sea ice extent indicating that there is some seasonality. The remainder plot tell us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

**ACF:**



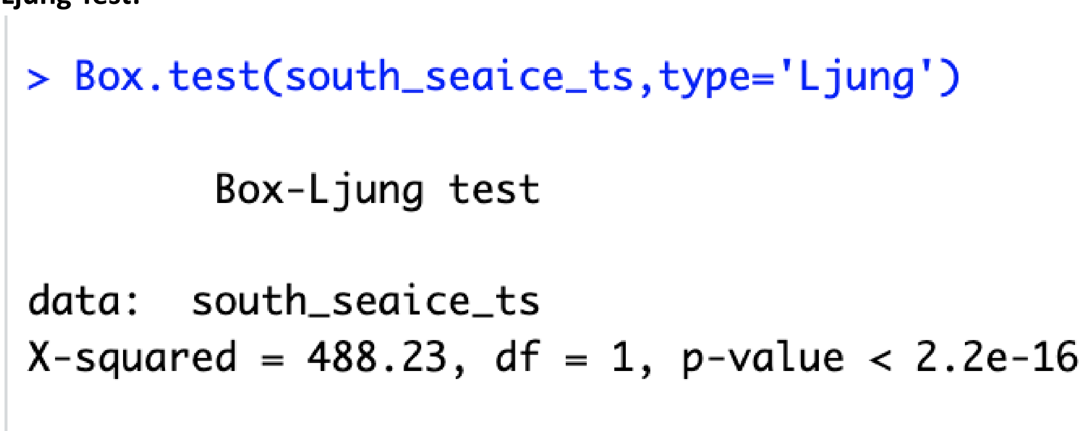

Based on the ACF we can see that there is slow decay and the series is non-stationary.

**PCAF**:




=> Based on the above PACF plot we can see that there is significant negative correlation at lag 2,3 and 4

**Ljung Test:**



```
> Box.test(south_seaice_ts,type='Ljung')

        Box-Ljung test

data:  south_seaice_ts
X-squared = 488.23, df = 1, p-value < 2.2e-16
```

Based on the above results of the Ljung test, we see that the degree of freedom is 1, X-squared value is 488.23 and p-values are less than significant value(0.05). The low p-value tells us that we can reject null hypothesis.

**Dickey-fuller test:**

```
> # 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
```

Based on the above results of the Dickey-Fuller test, we see that 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. We need further investigation.

**Kpss test:**

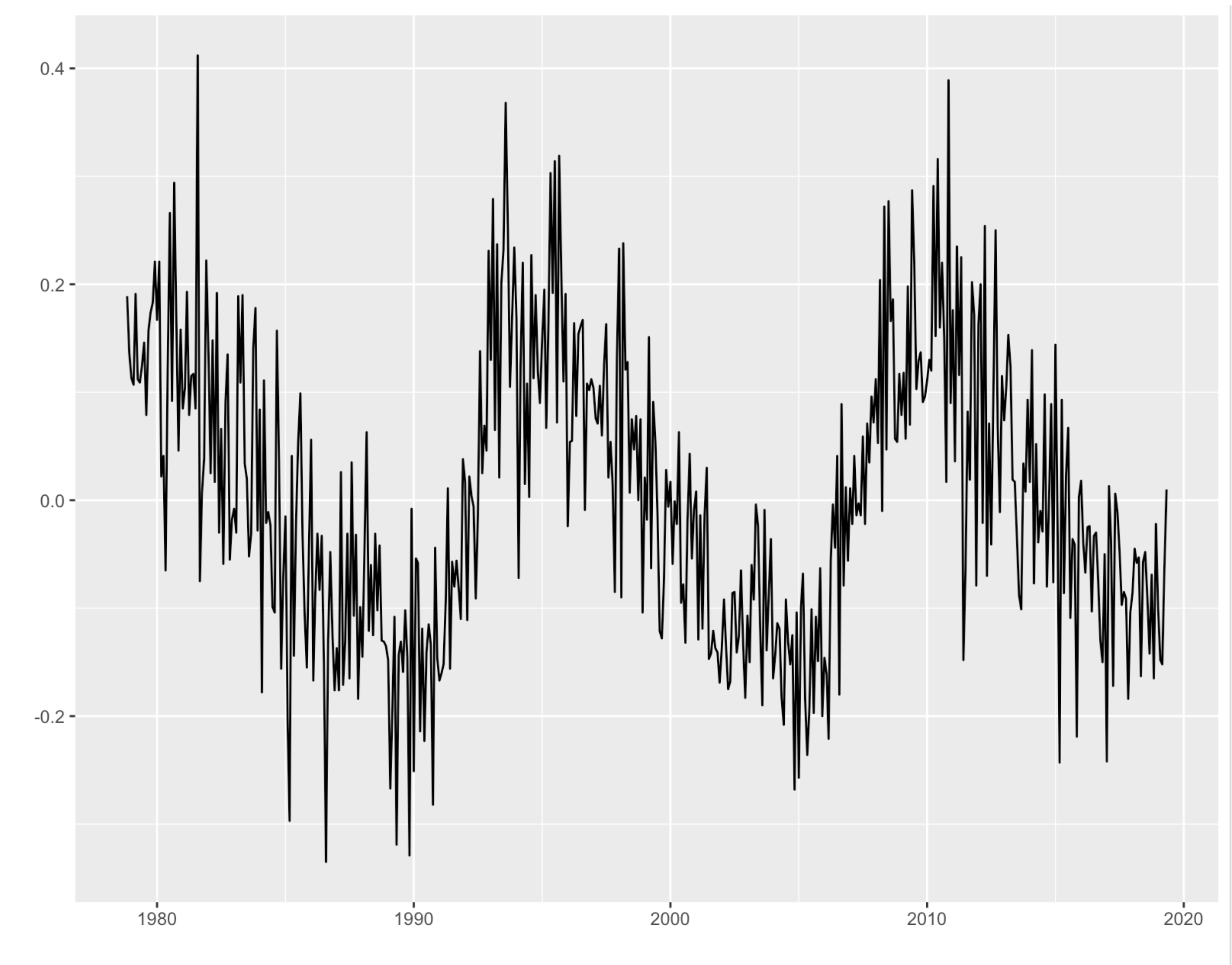
```
> kpss.test(south_seaice_ts)

        KPSS Test for Level Stationarity

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

> 
```

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. Therefore we fail to reject the null hypothesis of level stationary. This mean that the time series is stationary at the given level. But in contrast the p-value is greater than significant value which means that we need more investigation.

**Taking difference of model:**

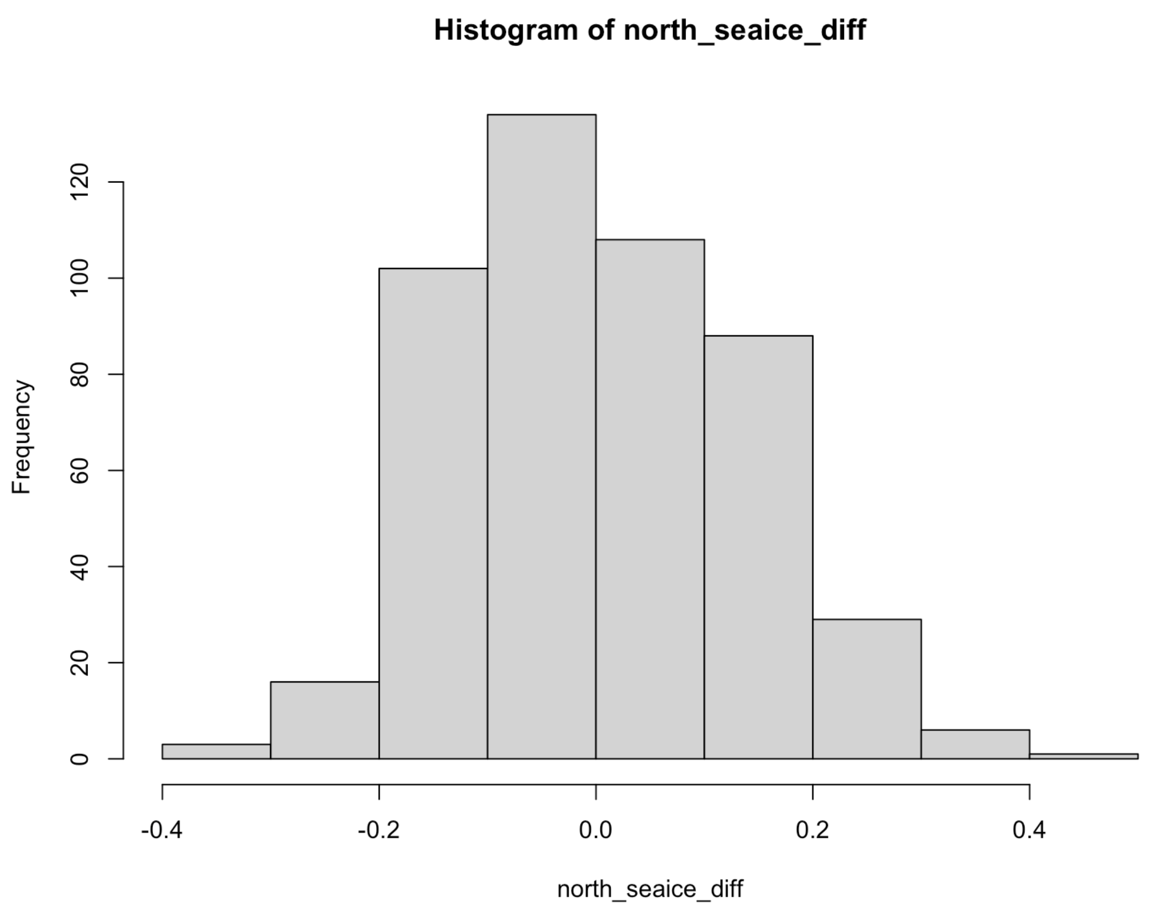
**North Model:**

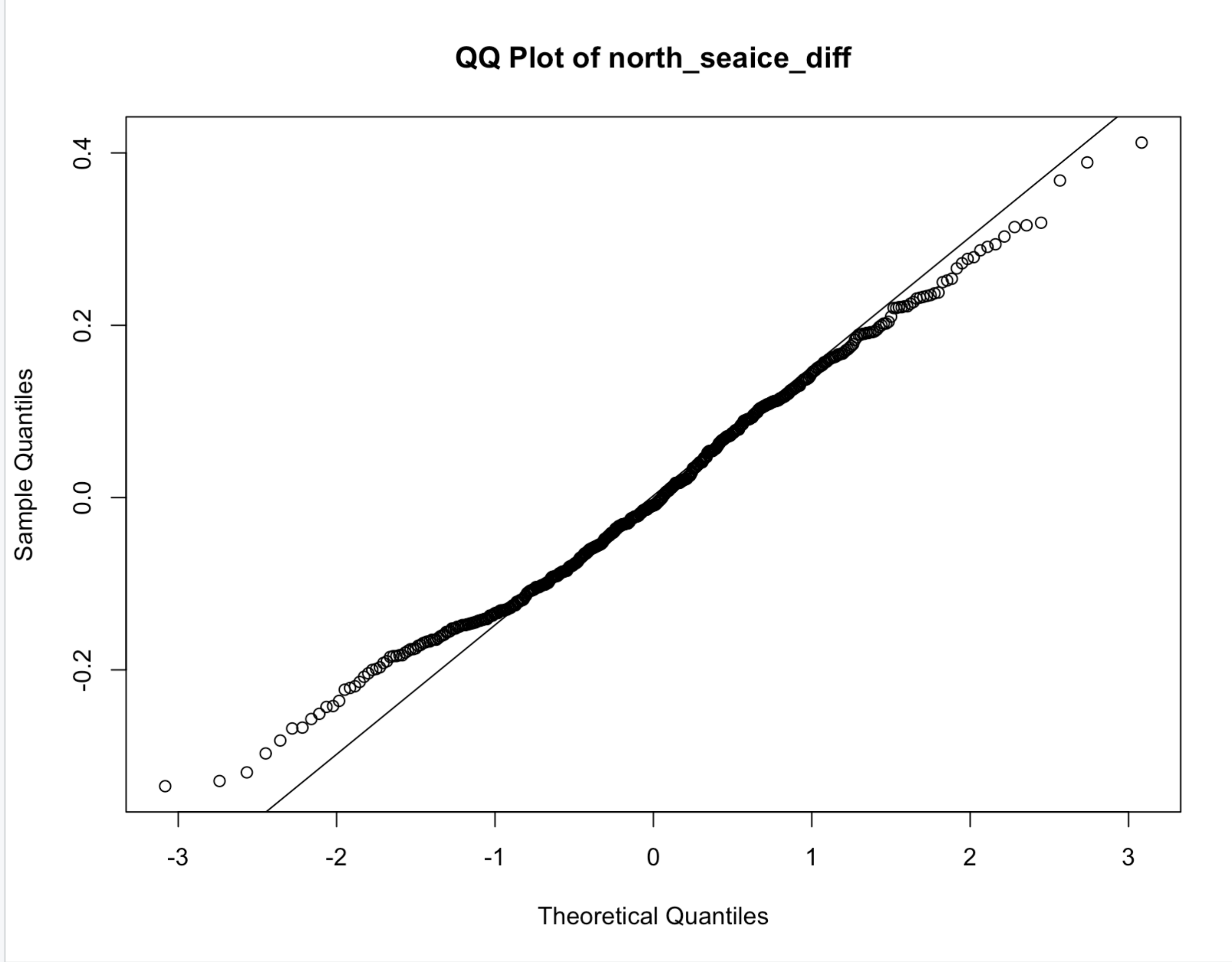
```
north_seaice_diff <- diff(north_seaice_ts)
autoplot(north_seaice_diff)
```




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 graphic also implies that the data has seasonality.Seasonality in the plot indicates that the time series data exhibits repeating trends.

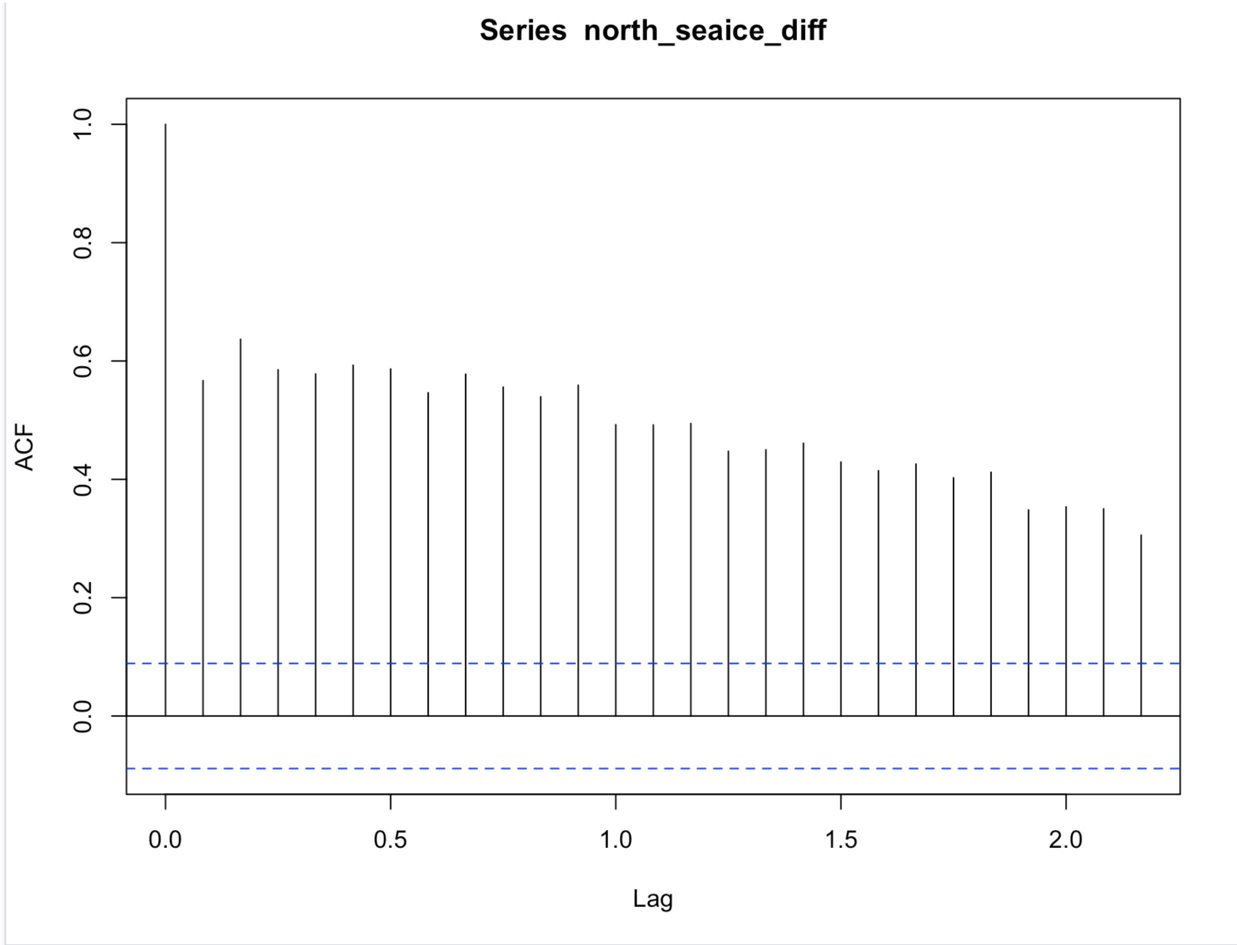
**Histogram of difference:**






The normal quantile plot shows variations from a completely straight line, indicating that the rate series is not fully normal.

**ACF:**




The acf suggests that the diff is still non-stationary, but because we have a time trend in our data, we are unsure.

**Checking for stationary using the adf and KPSS tests:**

```
> 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
>
```

The p-values for all lags are equal to 0.01, implying that the null hypothesis of non-stationarity is rejected in favor of stationarity.

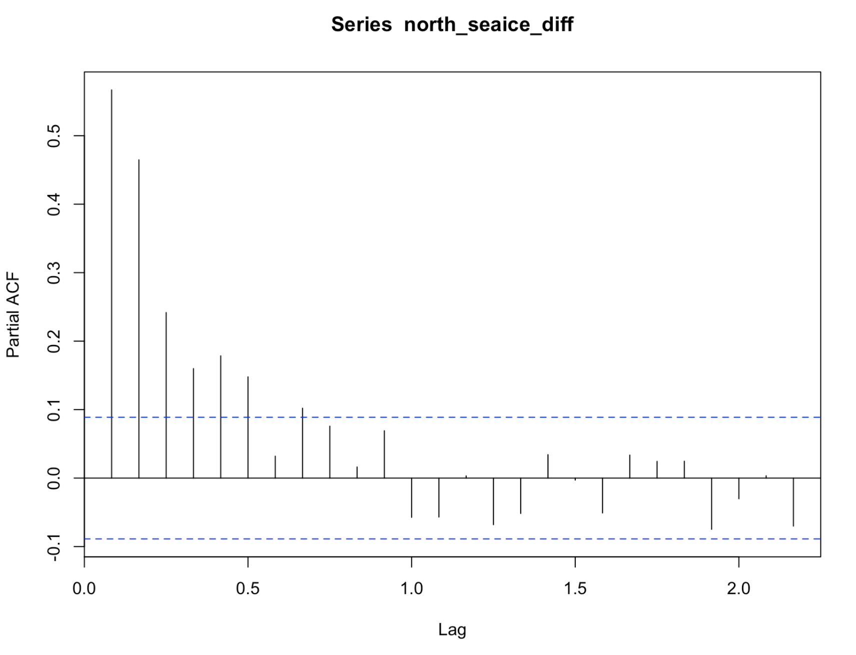
In conclusion, based on the ADF test results, the differenced series "north_seaice_diff" is stationary under all three scenarios, regardless of drift or trend.

**KPSS:**

```
> 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
>
```

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

**Pacf:**




From the pacf plot, it seems that there may be AR(2)or MA(1) .

**EACF:**

```
> 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
> 
```

From the eacf plot there might be ARMA(2,1) model or ARMA(2,0,2).

**Models:**

**Model1:**

```
> model1 = Arima(north_seaice_diff, order=c(2, 0, 2))
> summary(model1)
Series: north_seaice_diff
ARIMA(2,0,2) with non-zero mean

Coefficients:
         ar1     ar2      ma1      ma2    mean
      0.5810  0.3977  -0.5137  -0.1949  0.0083
s.e.  0.2427  0.2395   0.2533   0.2079  0.0495

sigma^2 = 0.008038:  log likelihood = 485.24
AIC=-958.48   AICc=-958.31   BIC=-933.35

Training set error measures:
                       ME       RMSE        MAE  MPE MAPE      MASE          ACF1
Training set -0.002098388 0.08919492 0.06959622 -Inf  Inf 0.6798602 -0.002088243
```

```
> coeftest(model1)

z test of coefficients:

            Estimate Std. Error z value Pr(>|z|)
ar1        0.5810346  0.2427132  2.3939  0.01667 *
ar2        0.3976770  0.2395209  1.6603  0.09685 .
ma1       -0.5136882  0.2532812 -2.0281  0.04255 *
ma2       -0.1948511  0.2079294 -0.9371  0.34871
intercept  0.0083103  0.0494658  0.1680  0.86658
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Box.test(model1$residuals, lag=10, type="Ljung")

        Box-Ljung test

data:  model1$residuals
X-squared = 7.292, df = 10, p-value = 0.6976
```

The ARIMA(2,0,2) model with a non-zero mean is fitted to the differenced series "north_seaice_diff". The estimated coefficients indicate the strength and direction of the relationships between current and lagged values. The model's sigma^2 represents the estimated residual variance. The AIC and BIC values are provided for model selection, with lower values indicating better fit. Training set error measures assess the model's performance, including ME, RMSE, MAE, MPE, MAPE, MASE, and ACF1. Coefficients are tested for significance using z-tests. The Box-Ljung test shows no remaining autocorrelation in the residuals, indicating a satisfactory fit. Overall, the ARIMA(2,0,2) model fits the data well.

**Model2:**

```
> 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 Akaike Information Criterion (AIC) is -959.77, the corrected AIC (AICc) is -959.65, and the Bayesian Information Criterion (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.

**AutoArima with aic:**

```
> 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 estimated coefficients for the ARIMA(2,0,1) model are as follows:

AR(1): 0.8158
AR(2): 0.1660
MA(1): -0.7532
The standard errors (s.e.) for the coefficients are also provided. The estimated residual variance (sigma^2) is 0.008017, and the log likelihood of the model is 484.87. The AIC value for the model is -961.74, the corrected AIC (AICc) value is -961.65, and the BIC value is -944.98.

In summary, 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 differenced series "north_seaice_diff" based on the AIC criterion.

```
> 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
```

In 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.

Interpreting the result, with a significance level of 0.05 (or any chosen threshold), 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.

In summary, based on the Box-Ljung test result, there is no strong indication of residual autocorrelation in the ARIMA model residuals selected using the AIC criterion.

**AutoArima with bic:**

```
> model_bic <- auto.arima(north_seaice_diff, ic = "bic")
> print(model_bic)
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
> coeftest(model_bic)

z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)
ar1  0.815783   0.055112  14.8022 < 2.2e-16 ***
ar2  0.165986   0.052512   3.1609  0.001573 **
ma1 -0.753163   0.036850 -20.4384 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

        Box-Ljung test

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

The auto.arima() function with the BIC criterion selected an ARIMA(2,0,1) model for the differenced series "north_seaice_diff". The model's coefficients are as follows:

AR(1): 0.8158
AR(2): 0.1660
MA(1): -0.7532
The standard errors (s.e.) of the coefficients are also provided. The estimated residual variance (sigma^2) is 0.008017, and the log likelihood of the model is 484.87. The AIC, AICc (AIC with small sample size correction), and BIC values are reported as well.

To further assess the significance of the coefficients, the coeftest() function is used. The z-test results show that all the coefficients (ar1, ar2, and ma1) are highly significant, with very low p-values ($p < 0.001$).

Additionally, the Box-Ljung test is performed on the residuals of the model using the Box.test() function. The test evaluates the independence of the residuals at different lags. In this case, the test statistic is 7.9673 with 10 degrees of freedom, and the corresponding p-value is 0.632. The relatively high p-value suggests that there is no evidence of significant autocorrelation remaining in the model's residuals at lags up to 10.

In summary, the ARIMA(2,0,1) model selected by the BIC criterion appears to be a good fit for the differenced series "north_seaice_diff". The coefficient estimates are significant, indicating that they contribute to explaining the behavior of the series, and the residuals do not exhibit significant autocorrelation.

**Backtesting:**

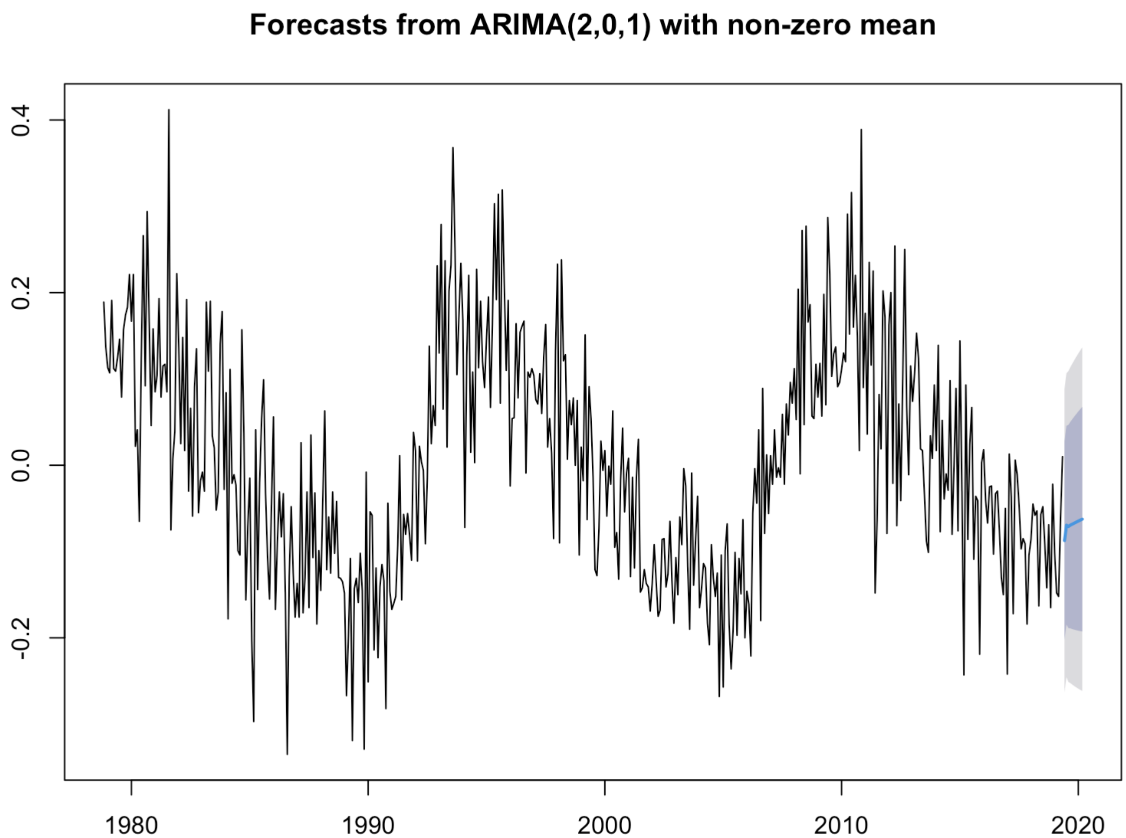
```
> b1 = backtest(model1,north_seaice_diff, h=1, orig=.8*n)
[1] "RMSE of out-of-sample forecasts"
[1] 0.09288803
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.07241832
[1] "Mean Absolute Percentage error"
[1] 1.680944
[1] "Symmetric Mean Absolute Percentage error"
[1] 1.103815
> b2 = backtest(model2,north_seaice_diff, 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
> b4=backtest(model_bic,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.

Model 2 appears to be the most straightforward, with an ARIMA(2,0,1) structure. It has statistically significant coefficients, indicating a good fit to the data. As a result, if simplicity and interpretability are important considerations, Model 2 may be the best option.

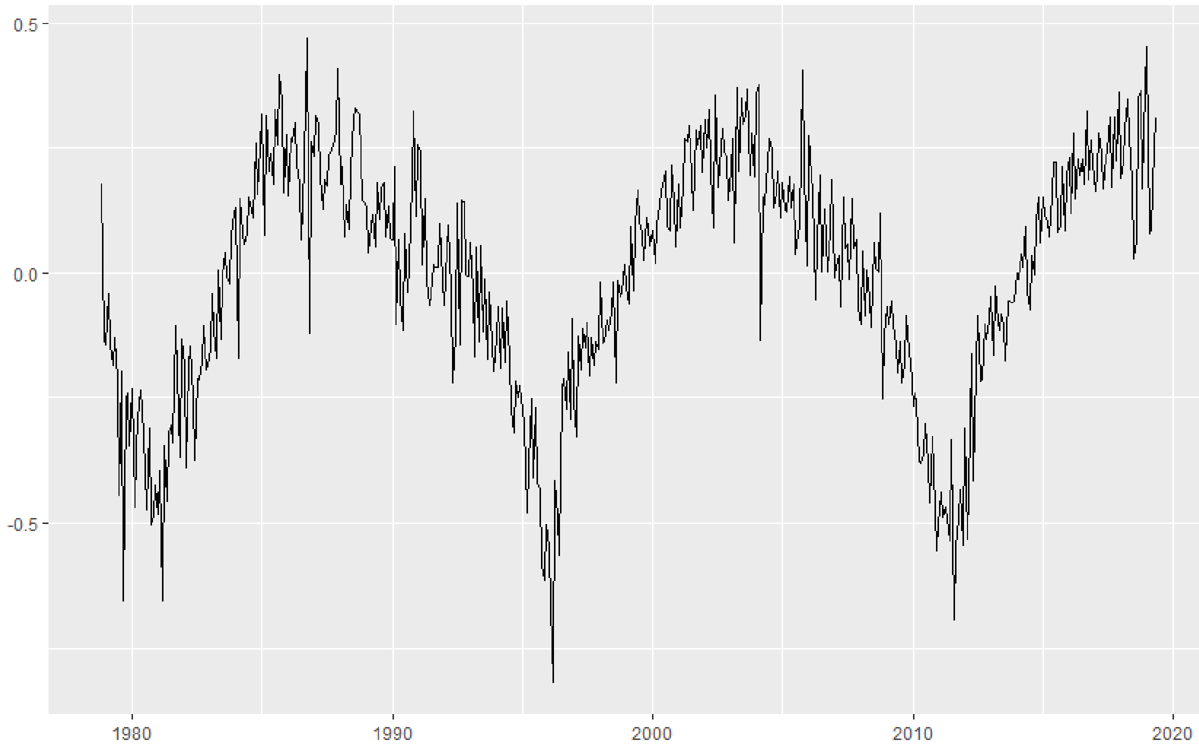
**Forecast:**

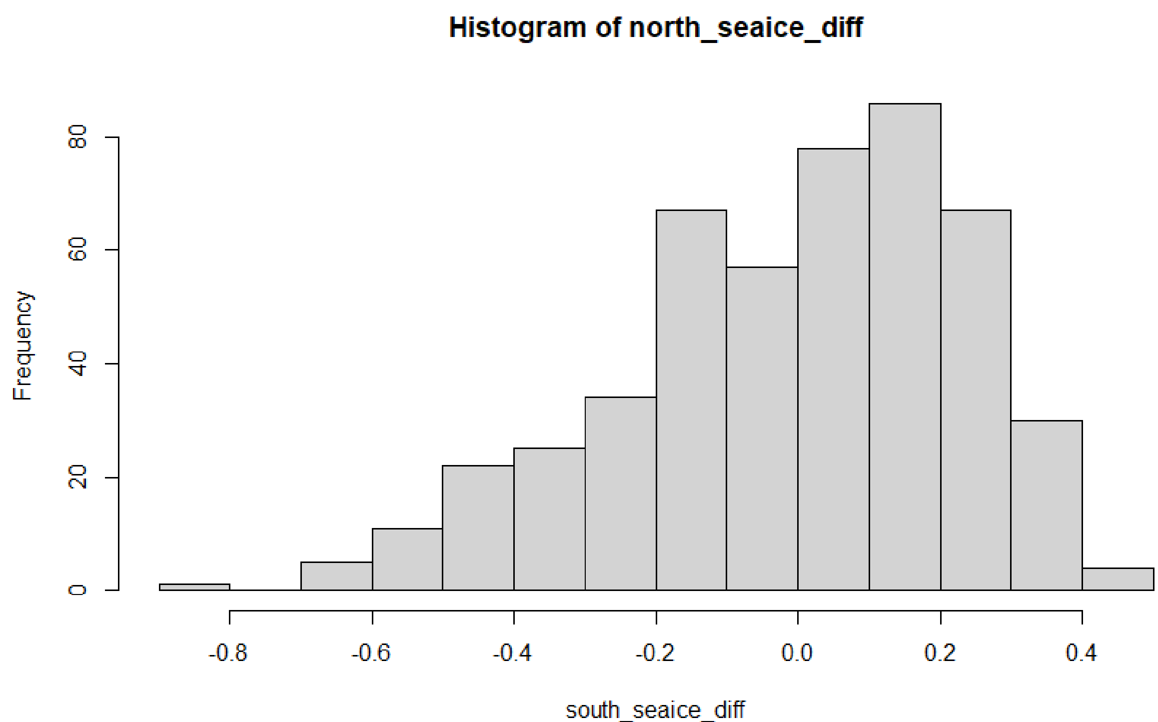
```
> forecast_model <- forecast(model2, h = 10)
> summary(forecast_model)

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
```



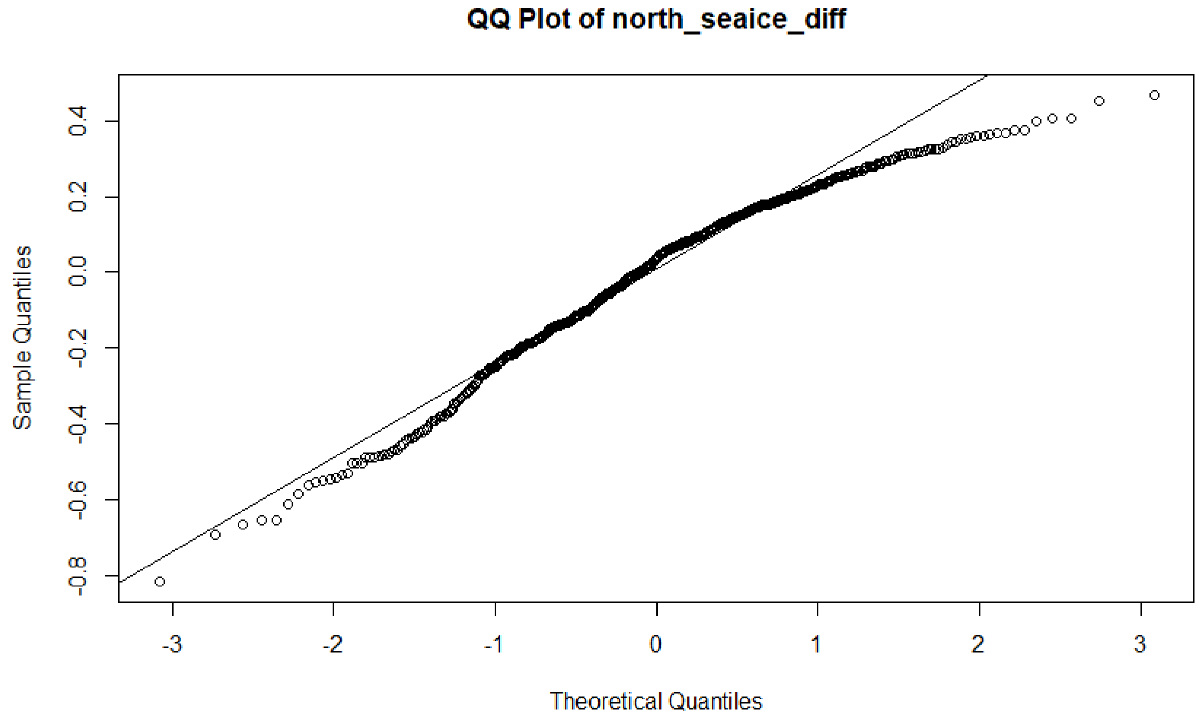

The model forecasts converge to a mean value of approximately 0.0076. This means that as the forecast horizon increases, the predicted values tend to approach this mean value.

South Hemisphere:

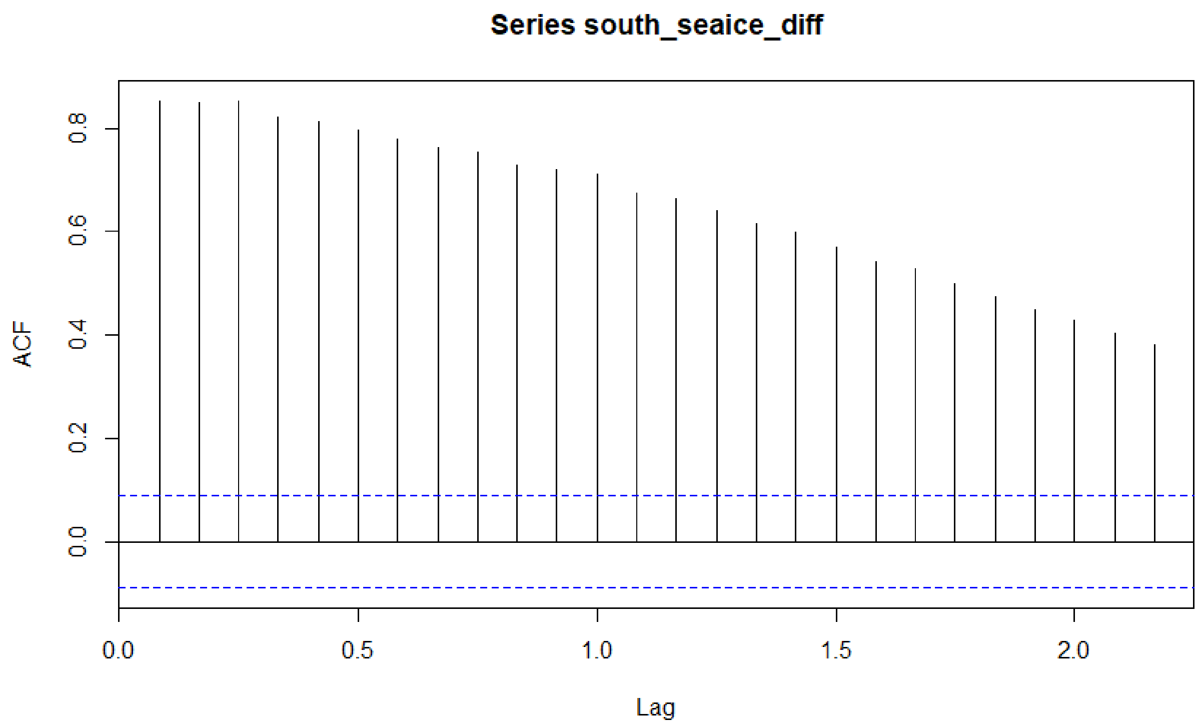


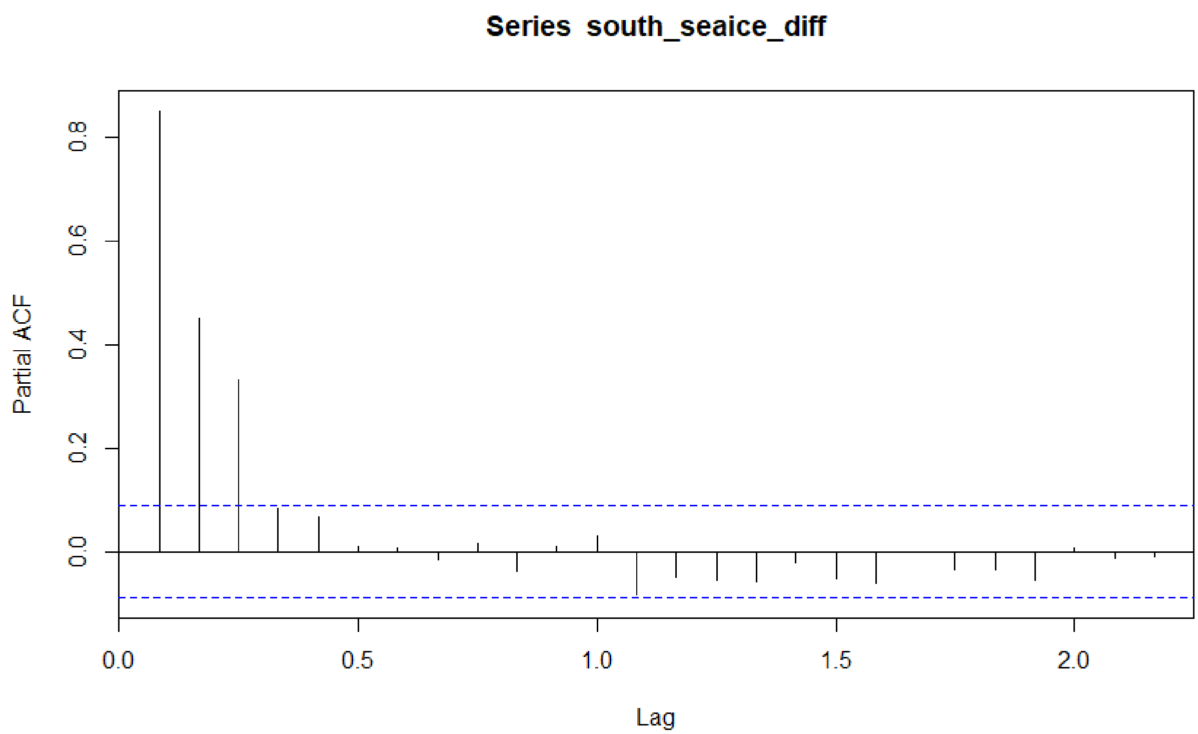
From the histogram we can see that the graph is slightly  skewed on the right side.




From the graph we can see that some points are not on the line they have more variance compared to others which indicates the it is not normally distributed.




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 not completely stationary.




Significant autocorrelation is a symptom of non-stationary behavior since it implies that the series remembers or is dependent on its previous values.

From the above graph we can see that

```
> 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
```

ADF:
Since 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. As a consequence, the'south_seaice_diff' series is probably non-stationary based on the findings of the ADF test.

KPSS:
Since the 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.

```
> model1 = Arima(south_seaice_diff, order=c(1, 0, 3), seasonal=list(order=c(2, 0, 0), seas
onal=12))
> summary(model1)
Series: south_seaice_diff
ARIMA(1,0,3)(2,0,0)[12] with non-zero mean

Coefficients:
         ar1      ma1     ma2     ma3    sar1    sar2    mean
      0.9764  -0.7048  0.0528  0.0703  0.1168  0.0137  0.0166
s.e.  0.0106   0.0479  0.0616  0.0480  0.0488  0.0485  0.0891

sigma^2 = 0.01092:  log likelihood = 411.22
AIC=-806.44   AICc=-806.14   BIC=-772.94

Training set error measures:
                        ME      RMSE        MAE       MPE     MAPE      MASE
Training set -0.0004841047 0.1037556 0.07985227 -1.094832 109.4944 0.5737989
                    ACF1
Training set 0.002175397
> coeftest(model1)

z test of coefficients:

           Estimate Std. Error  z value Pr(>|z|)
ar1        0.976448   0.010588  92.2243  < 2e-16 ***
ma1       -0.704838   0.047907 -14.7127  < 2e-16 ***
ma2        0.052755   0.061588   0.8566  0.39168
ma3        0.070316   0.047979   1.4656  0.14277
sar1       0.116803   0.048810   2.3930  0.01671 *
sar2       0.013739   0.048509   0.2832  0.77700
intercept  0.016585   0.089136   0.1861  0.85239
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model has an AIC of -806.44, an AICc of -806.14, and a log probability of 411.22. The MAPE (mean absolute percentage error) is 109.49, while the RMSE (root mean squared error) is 0.1038.

It is clear from the coefficients for ar1 and ma1 how crucial they are to the model. Ma2, Ma3, Sar1, Sar2, and Mean coefficients do not have statistical significance.

```
> Box.test(model1$residuals, lag=10, type="Ljung")

        Box-Ljung test

data:  model1$residuals
X-squared = 4.2904, df = 10, p-value = 0.9333
```

According to the test findings, X-squared has a test statistic of 4.2904 with 10 degrees of freedom. The corresponding p-value is 0.9333, above the usual significance threshold of 0.05. Therefore, model1 lacks compelling evidence of residual autocorrelation.

The non-significant p-value from the Box-Ljung test implies that model1 appropriately represents the autocorrelation structure in the data.

```
> 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 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.

Both models adequately account for the serial correlation in the data since their residuals show no substantial autocorrelation.In comparison to Model 2 (ARIMA(2,0,1)(1,0,0)[12]), Model 1 (ARIMA(1,0,3)(2,0,0)[12]) has a somewhat greater log likelihood, a lower AIC, and a lower AICc.However, Model 2's performance metrics (RMSE and MAPE) are marginally better, indicating marginally higher prediction accuracy.Model 2 (ARIMA(2,0,1)(1,0,0)[12]) may be favored above Model 1 when taking into account both statistical criteria (AIC, AICc) and performance measurements.

```
> 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 significance 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'south_seaice_diff' series well.

```
> model_bic <- auto.arima(south_seaice_diff, ic = "bic")
> # View the BIC-selected model
> print(model_bic)
Series: south_seaice_diff
ARIMA(1,0,1)(1,0,0)[12] with zero mean

Coefficients:
         ar1      ma1    sar1
      0.9825  -0.6483  0.1218
s.e.  0.0085   0.0327  0.0483

sigma^2 = 0.01101:  log likelihood = 407.37
AIC=-806.74   AICc=-806.66   BIC=-789.99
> coeftest(model_bic)

z test of coefficients:

       Estimate Std. Error  z value Pr(>|z|)
ar1   0.9825418  0.0084505 116.2709  < 2e-16 ***
ma1  -0.6483037  0.0327316 -19.8067  < 2e-16 ***
sar1  0.1218500  0.0482977   2.5229  0.01164 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

	Box-Ljung test

data:  model_bic$residuals
X-squared = 12.392, df = 10, p-value = 0.2597
```

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 three coefficients (ar1, ma1, and sar1) in this instance are all extremely significant (p 2e-16).

The test results show that the X-squared test statistic has 10 degrees of freedom and a value of 12.392. The corresponding p-value is 0.2597, which is above the usual

significance threshold of 0.05. Therefore, based on the Box-Ljung test, there is no meaningful evidence of residual autocorrelation in the chosen model.
The residual autocorrelation is not very strong and the BIC has a low BIC value. These results imply that the model is a fair match for the'south_seaice_diff' series, however it might not be as excellent as compared to AIC.

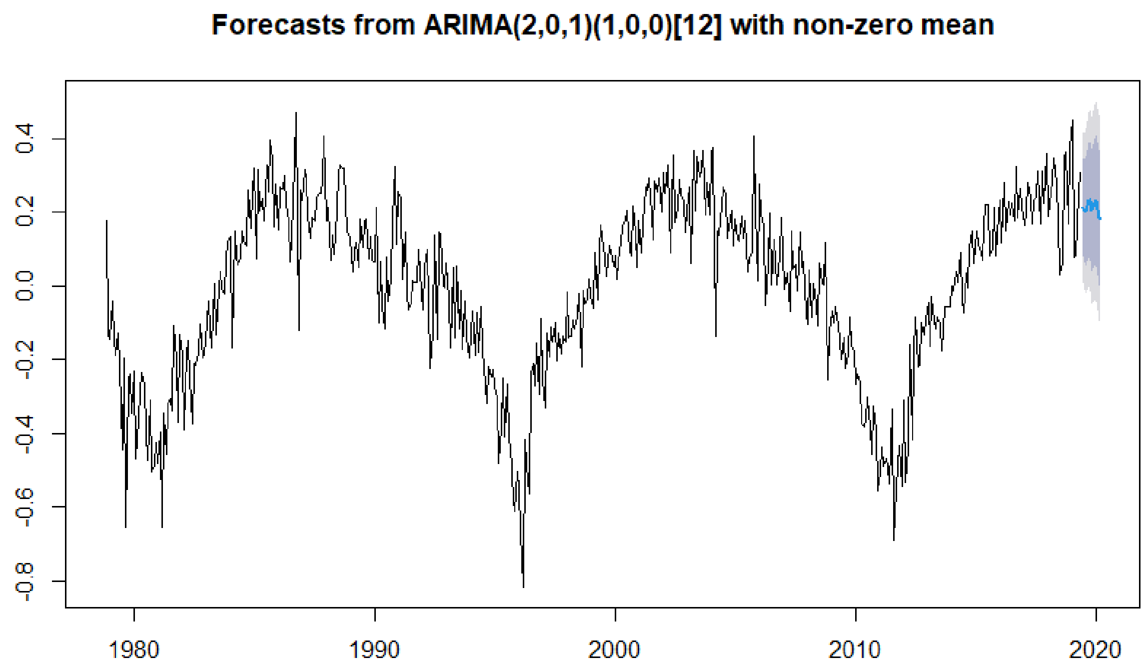
```
> b1 = backtest(model1,south_seaice_diff, h=1, orig=.8*n)
[1] "RMSE of out-of-sample forecasts"
[1] 0.0974513
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.0751902
[1] "Mean Absolute Percentage error"
[1] 0.8792194
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.54578
> b2 = backtest(model2,south_seaice_diff, 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
> b4=backtest(model_bic,south_seaice_diff, h=1, orig=.8*n)
[1] "RMSE of out-of-sample forecasts"
[1] 0.09739296
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.0756426
[1] "Mean Absolute Percentage error"
[1] 0.9537672
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.5577879
```

The comparison shows that the models' RMSE, MAE, and SMAPE performances are comparable. But when compared to the other models, the model that the BIC chose

(Model 4) had a little higher MAPE. It's important to note that the performance variations are minimal.

Because they have comparable performance metrics and a little lower MAPE than Model 2 and the BIC-selected model, Model 1 (ARIMA(1,0,3)(2,0,0)[12]) and the AIC-selected model (ARIMA(1,0,3)(2,0,0)[12]) can be thought of as good possibilities.

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

Forecast method: ARIMA(2,0,1)(1,0,0)[12] with non-zero mean

Model Information:
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

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

Forecasts:
         Point Forecast        Lo 80     Hi 80        Lo 95     Hi 95
Jun 2019      0.2144940  0.080356617 0.3486314  0.009348571 0.4196394
Jul 2019      0.2035980  0.064630607 0.3425654 -0.008934278 0.4161302
Aug 2019      0.2025915  0.055409960 0.3497730 -0.022503218 0.4276861
Sep 2019      0.2335479  0.079400106 0.3876957 -0.002200796 0.4692966
Oct 2019      0.2316114  0.070981062 0.3922417 -0.014051490 0.4772742
Nov 2019      0.2053362  0.038715746 0.3719567 -0.049487807 0.4601603
Dec 2019      0.2264431  0.054255380 0.3986308 -0.036895297 0.4897815
Jan 2020      0.2324305  0.055051915 0.4098092 -0.038846661 0.5037078
Feb 2020      0.1849873  0.002754318 0.3672203 -0.093713994 0.4636886
Mar 2020      0.1827917 -0.003992256 0.3695757 -0.102869727 0.4684532
> plot(forecast_model)
```




The predicted values from the model converge to a mean value of about 0.02. This indicates that the anticipated values tend to approach this mean value as the prediction horizon lengthens.