Extent of Sea Ice Case Study

: By Annu Chauhan

Table of Contents

Contents

Executive Summary

The main objective of this project is to perform an analytical study on the monthly extent of ice flow over a time. Through this project, we are trying to see the impact of season and year over year extent of ice flow, to establish or identify the presence of trends and seasonality by performing the time series analysis. This analysis can be use forecasting the ice flow as well.

For carrying out this analytical and predictive study, our team has used following major R packages.

- dplyr
- readr
- kendall
- tseries

We had wrote a python script for acquiring the dataset from the website:
ftp://sidads.colorado.edu/DATASETS/NOAA/G02135/

This provided the monthly datasets/ csv files from 1978 until September 2022.

The objective of extracting this solid evidence from the above-mentioned packages are derived because each of the package mentioned above has so much to offer in its capacity.

Let us look at the description of each package in brief.

- **dplyr**: This package is used for doing the data manipulation.
- **readr**: This package is used for reading and combining the multiple csv files into single dataset. Basically, it read and combine all the csv files present in the folder and create a dataframe.
- **Kendall**: This package is used for evaluating the timeseries for the presence of any trend and seasonality.
- **tseries:** This R package is used for performing the time series analysis and computation.

Since we are clear now about the role of each package in this project, let us explore other aspects of this projects. In the further sections, we will see what are all the operations we performed to over the constraints in this project.

Methodology

In this section, we will see the methodology we used for executing this project. This section is divided into following categories.

- Data Description
- Data Preprocessing
- Data Exploration
- Model Building and Evaluation

Let us dive into each subsection one by one.

Data Description

As already mentioned, the dataset for this project has been taken from NOAA website.

As the suggestion provided in the project, we have downloaded the separate CSV files for north and south regions. However, the dataset variables are the same for both region

We will look at the data in each file as follows.

This file contains a total of 6 columns, which are described as follows.

Column Name	Description
Year	Year of the ice flow data
Mo	Month of ice flow data
Data type	NASA data Center name
Region	Region of Hemisphere – North of South
Extent	the Extent of the sea ice, which is measured in millions of square kilometers
Area	Area of extent

Data Preprocessing

This section deals with all the operations we performed as a part of Data Preprocessing. This includes the following parts.

- Missing Values
- Data Transformation

Let us look at each of these sections individually. Here our target variable is extent.

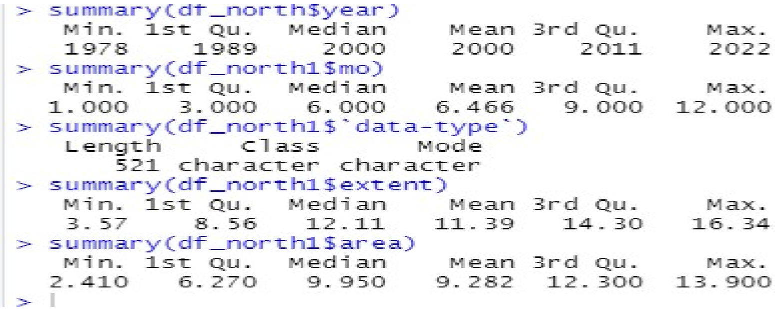
Missing Values

Upon performing the check of missing values, we don't find the missing data as such, but we did find a value which is set to default -9999, possibly indicting the missing data. Since this value doesn't hold any value to us. Moreover, it will affect our ability to properly create time series model we have decided to remove it

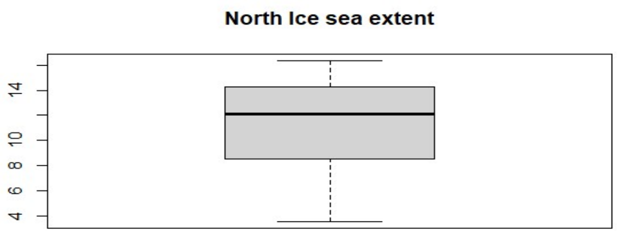
Data Exploration

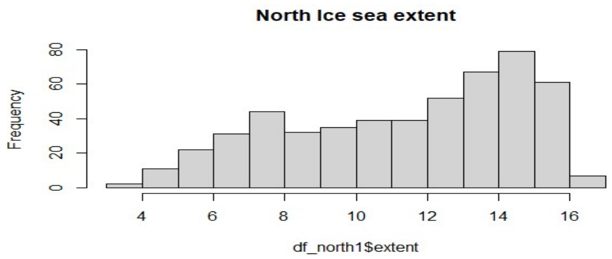
We performed data exploration / descriptive statistics on the data Separately for both regions.

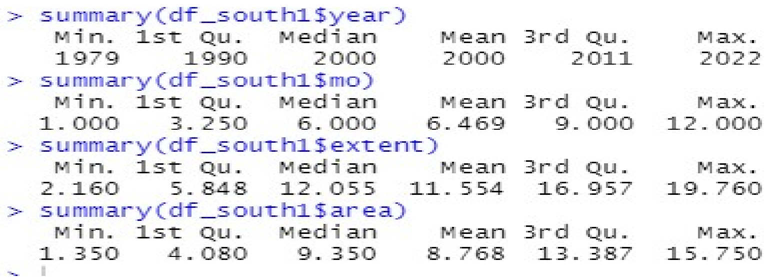
As we can see here, data ranges from 1978 until 2022. However, we are interested in data since 1979 (mentioned in project requirement). Ice extent ranges from 3.57 to 16.34 million sq km in case of north region and 2.16 to 19.76 million sq km in the south region.



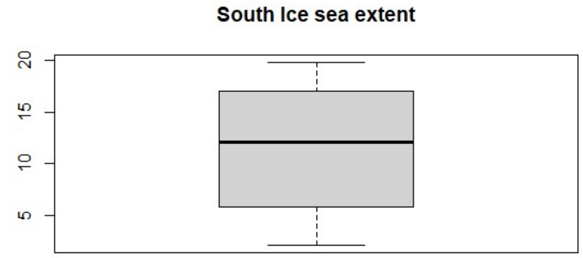
```
> summary(df_north$year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1978    1989    2000    2000    2011    2022 
> summary(df_north1$mo)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.000   6.000   6.466   9.000  12.000 
> summary(df_north1$`data-type`)
   Length     Class      Mode 
      521 character character 
> summary(df_north1$extent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3.57    8.56   12.11   11.39   14.30   16.34 
> summary(df_north1$area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.410   6.270   9.950   9.282  12.300  13.900 
> 
```

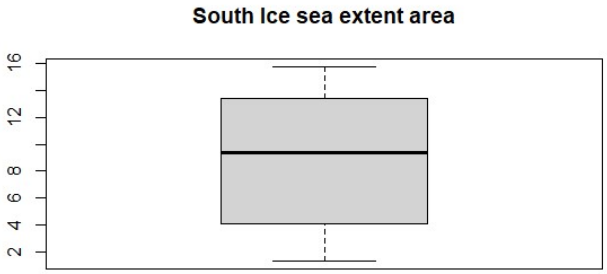

North Ice sea extent

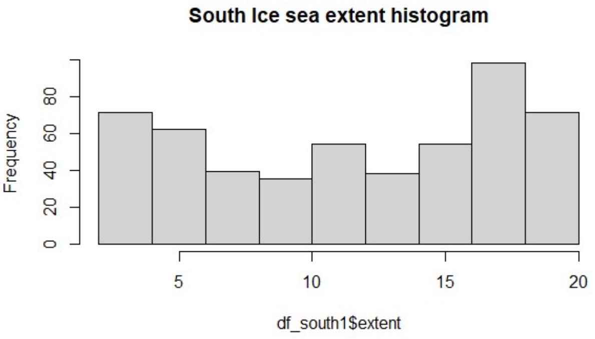

North Ice sea extent



```
> summary(df_south1$year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1979    1990    2000    2000    2011    2022 
> summary(df_south1$mo)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.250   6.000   6.469   9.000  12.000 
> summary(df_south1$extent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.160   5.848  12.055  11.554  16.957  19.760 
> summary(df_south1$area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.350   4.080   9.350   8.768  13.387  15.750 
>
```


South Ice sea extent


South Ice sea extent area



We have created boxplot for extent in both regions to check if there is any outliers. We can see that the histogram of south region ice extent is more symmetrical than the north region.

Data Transformation

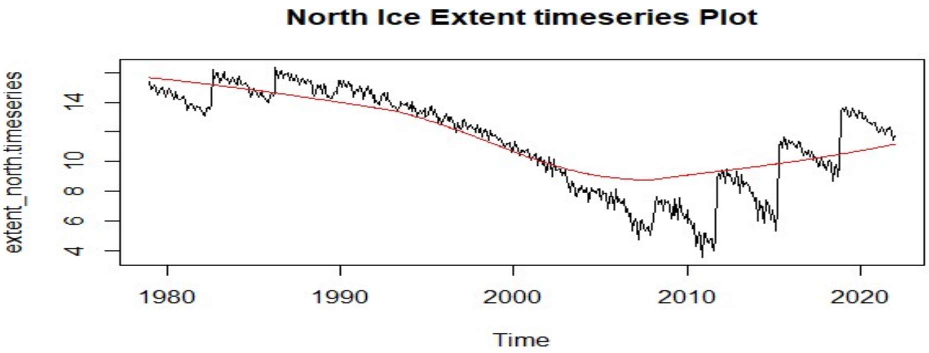
The time series analysis contains the major portion of the data transformation part.

Since extent is our dependent variable and year and month is our independent variable. We have decided the extracted extent from the data alone and transform into time series. Taking data from 1979 and dividing the year into months i.e., frequency of 12.

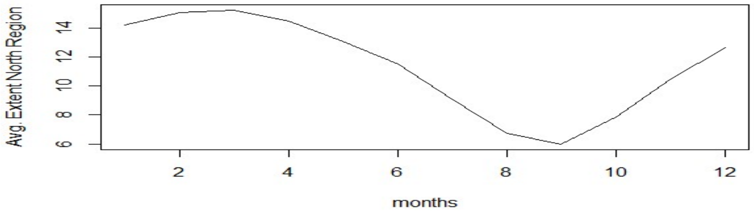
Extent_north = df_north1$extent

extent.timeseries = ts(data = extent_north,start = 1979 ,end = 2022 ,frequency = 12)

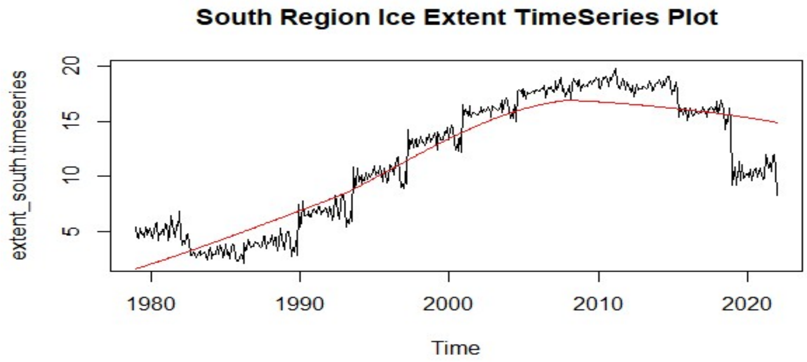
So, once we have transformed the data into timeseries, we have plotted against the time. So that we identify any trend and seasonality. By looking into these above graphs, we can say that we do have trend and seasonality.



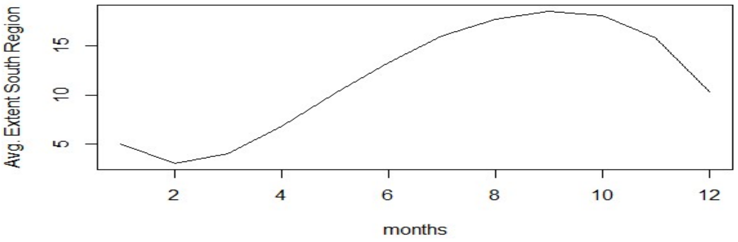
Based on the above north region chart, we can say that the since 1979 indeed rapid changes have been occurring in the north region ice until 2012, where the ice coverage has been declining at a substantial rate. This might be driven global warming explosion and industrialization After 2012 there is slow but continues improvement in the ice extent. However, it has cyclical pattern associated with it.



We also took monthly average of the data and plotted against the months to see the seasonality associated with the north region ice extent. Based on the above chart it seems there is seasonality associated with the ice formation. Ice extent decreases sharply during the summer months, and increases thereafter.



In contrast to the north region, in the south region the sea ice coverage has been increasing (1979 to 2014) although at a lesser rate than the decreases in the north region. However, after 2014 we can observe continues and sharp decline in ice extent of south region showing the excessive impact of global warming. We also see some stabilization in ice extent during 2020 which indicates some positive impact of COVID spread on south region ice extent. Seasonality associated with it indicate the ice extent changes over a period of the year based on the season.



We also took monthly average of the data and plotted against the months to see the seasonality associated with the north region ice extent. Based on the above chart it seems there is seasonality associated with the ice formation. Ice extent in south region start increasing march onwards, and peak around September.

Trend and Seasonality Analysis

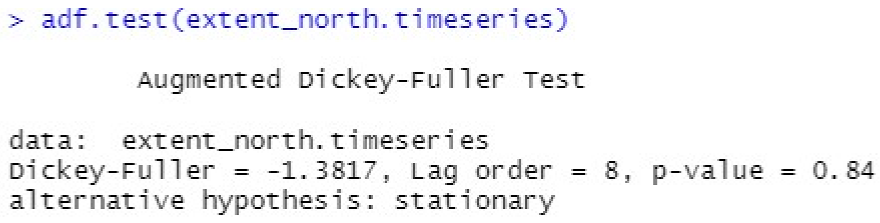
We have also decided to employ statistical test like Kendall and ADF (Augmented Dickey-Fuller test) to check whether timeseries is stationary or not. A time series is said to be “stationary” if it has no trend, exhibits constant variance over time, and has a constant autocorrelation structure over time.

ADF Test:

H0: The time series is non-stationary. In other words, it has some time-dependent structure and does not have constant variance over time.

HA: The time series is stationary.

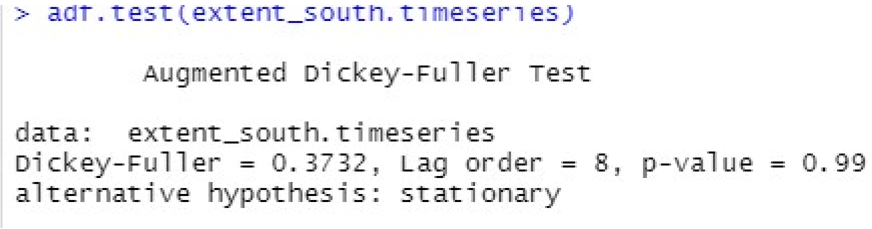
If the p-value from the test is less than some significance level (e.g. α = .05), then we can reject the null hypothesis and conclude that the time series is stationary.



```
> adf.test(extent_north.timeseries)

        Augmented Dickey-Fuller Test

data:  extent_north.timeseries
Dickey-Fuller = -1.3817, Lag order = 8, p-value = 0.84
alternative hypothesis: stationary
```



```
> adf.test(extent_south.timeseries)

        Augmented Dickey-Fuller Test

data:  extent_south.timeseries
Dickey-Fuller = 0.3732, Lag order = 8, p-value = 0.99
alternative hypothesis: stationary
```

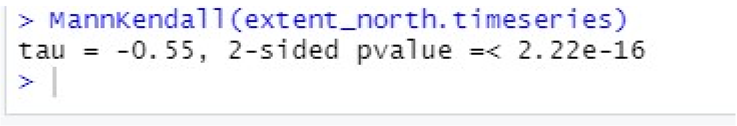
Since the p value is greater than the significance level of 0.05 , we fail to reject the null hypothesis. Thus the time series is not stationary and does have trend.

Kendall Test:

H0: The time series is stationary.

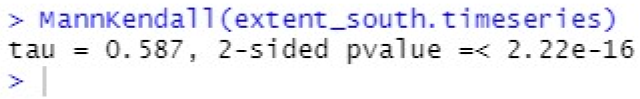
HA: The time series is non-stationary. That is trend exists

We have performed Mann Kendell test for both Trend and seasonality.



```
> MannKendall(extent_north.timeseries)
tau = -0.55, 2-sided pvalue =< 2.22e-16
> |
```

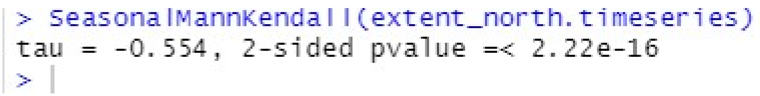
The test statistic is -0.55, and the two-sided p-value is less than 0.05. We reject the null hypothesis of the test and conclude that a trend exists in the data because the p-value is much less than 0.05.



```
> MannKendall(extent_south.timeseries)
tau = 0.587, 2-sided pvalue =< 2.22e-16
>
```

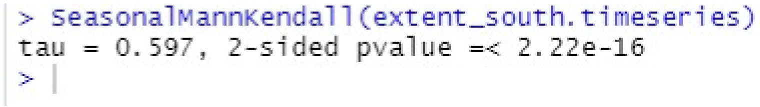
The test statistic is 0.587, and the two-sided p-value is less than 0.05. We reject the null hypothesis of the test and conclude that a trend exists in the data because the p-value is much less than 0.05.

We use SeasonalMannKendall Test to account for any seasonality in the data



```
> SeasonalMannKendall(extent_north.timeseries)
tau = -0.554, 2-sided pvalue =< 2.22e-16
>
```

The test statistic is -0.554, and the two-sided p-value is less than 0.05. We reject the null hypothesis of the test and conclude that a seasonality exists in the data because the p-value is much less than 0.05.



```
> SeasonalMannKendall(extent_south.timeseries)
tau = 0.597, 2-sided pvalue =< 2.22e-16
>
```

The test statistic is 0.597, and the two-sided p-value is less than 0.05. We reject the null hypothesis of the test and conclude that a seasonality exists in the data because the p-value is much less than 0.05.

Decomposing

Immediately we can “decompose” the time series — which in this case means separating out the 3 main components that make up the time series:

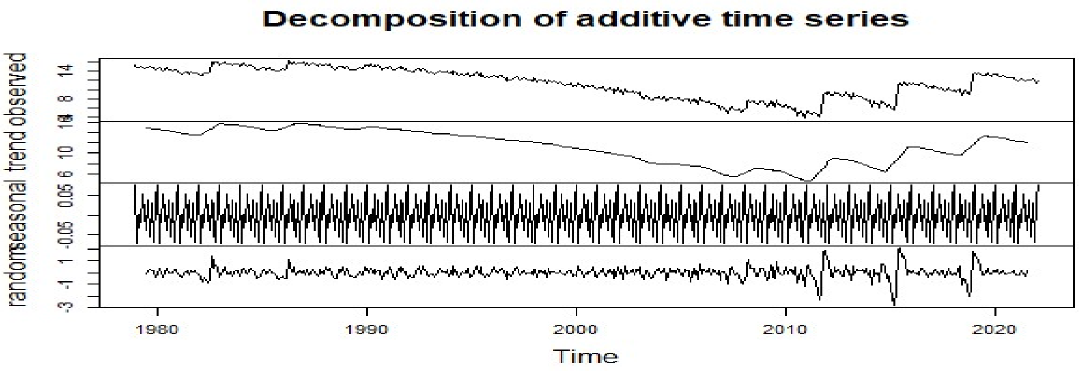
1.trend: the long-term trends in the data

2.seasonal: the repeated seasonal signal adder

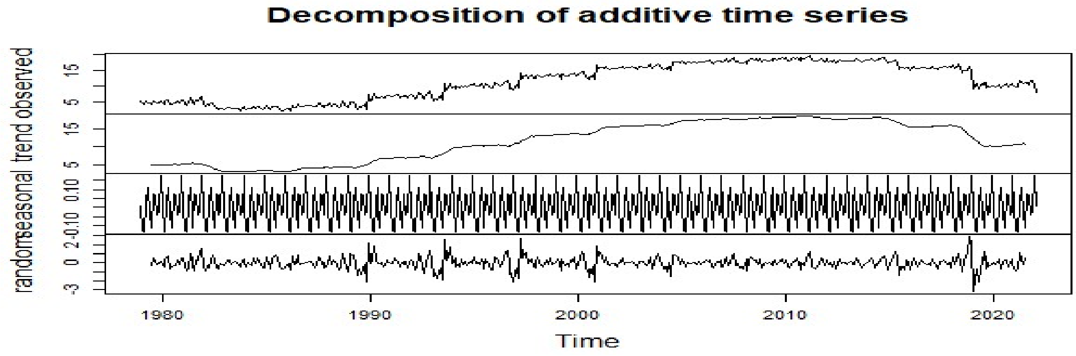
3.random: the “left-over” components that aren’t expected from the seasonality or trend components.

We can easily extract these components and plot them with:

North Region:



South Region:



As we look at the decomposition components, we can visually see how they can add up to our “observed” value (our real values).

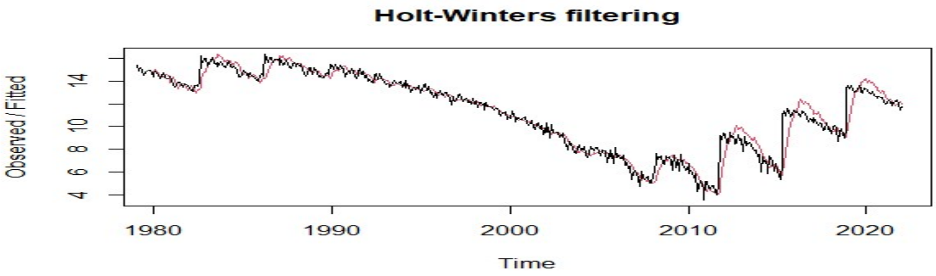
Modelling and Evaluation

Since both north and South region ice flow timeseries does have trend as well as the seasonality we have decided to use the Holt-winters algorithm for model building.

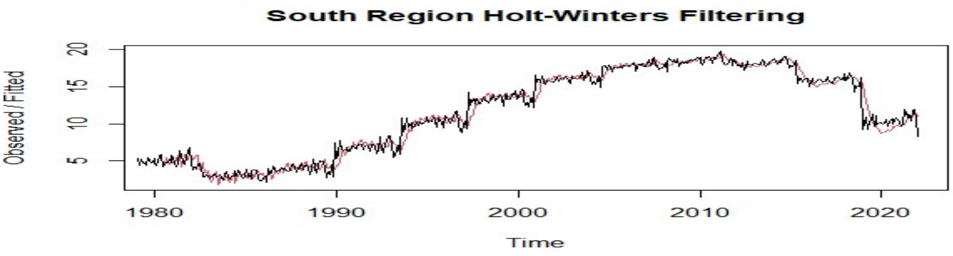
HW1 <- HoltWinters(extent_north.timeseries, alpha=0.2, beta=0.1, gamma=0.1)

HW1_south <- HoltWinters(extent_south.timeseries, alpha=0.2, beta=0.1, gamma=0.1)

Once we build the timeseries model we perfom the fitting of the timeseries to see how close we are to the observed values.

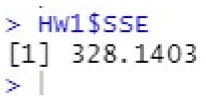


we can see in the above charts we are quite closely fitted the timeseries data of north region.

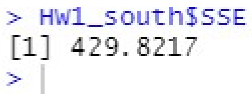


Similarly, we can see in the above charts we are quite closely fitted the timeseries data of south region.

Evaluation: We have also evaluated the models – North and South Region on SSE metric, which shows we are very close the observed values of the data, and overall good model fit.

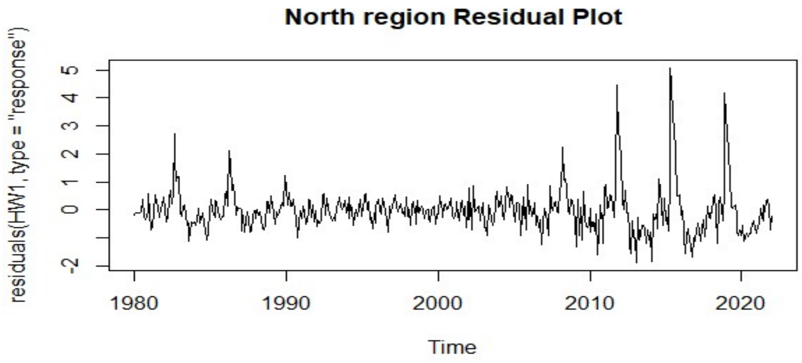


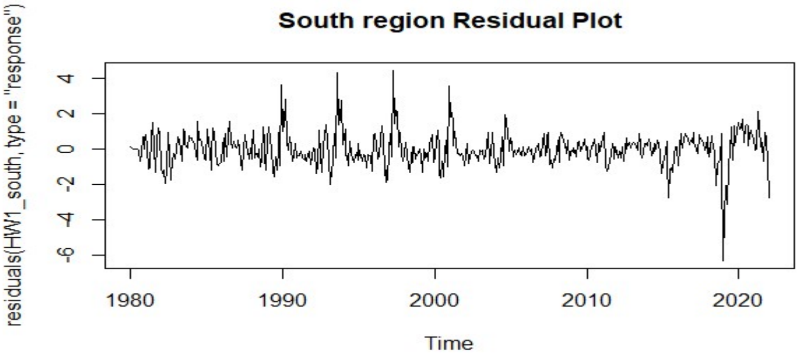
```
> HW1$SSE
[1] 328.1403
>
```



```
> HW1_south$SSE
[1] 429.8217
>
```

Finally, we produce the residual plot for both north and south timeseries models





Conclusion

In conclusion, we would like to mention the following things.

- This is evident there is significant shrinkage of the extent of the north and south ice extent over the time. As we saw in the above time series graphs in the north region ice cover reduce substantially from 1979 until 2012, however there some improvement in recent years. In contrast to north region south region sea ice increase from 1979 to 2014, however it reduces drastically since then. It indicates the ice formation between these two regions work differently and affected the geography apart from other factors.
- Since the Ice extent has decreased significantly compared to 1979, it confirms the effects of global warming.

- Further, based on finding (month avg. charts) we can infer the presence of seasonality in the ice formation which differ region wise.
- We can see stabilization in south region ice extent during 2020 after sharp decline which indicates reduction in greenhouse gases release during this time which make sense considering COVID spread during this time period which effected the globe (industries closure, less use of automobiles etc.)

Appendix

None

References

- https://www.statisticshowto.com/mann-kendall-trend-test/
- https://analyticsindiamag.com/a-guide-to-different-evaluation-metrics-for-time-series-forecasting-models/
- https://rdrr.io/cran/forecast/man/residuals.forecast.html