Sea Ice Extent

TIME SERIES ANALYSIS

Yiyang Yang | CSC 425 | 2018/03/07

Non-technical Summary

This report is talking about sea ice level time series analysis, in present, greenhouse effect is increasingly serious, as scientists said due to the sea level increasing, Japan will sink in future years, this is appalling. So, I decide to do some exploration about sea ice extent. Another reason for choosing this dataset is I think natural is related with every one of us, and as a human could do and find some change rules of nature and find a way to solve the natural problems.

Sea ice extent is another way to show how the sea level changes. This dataset is recorded the sea ice extent data about North and South pole from 1980 to 2017. In this 37 years period, sea ice extent has a very huge change, the whole sea ice extent from both pole decrease about 3 million square kilometers, this area is about one third of area of China. It is no doubt that one third of China ice melt into water, how the sea level increase.

Technical Summary

Situation

The reason why I want to analyze the sea ice level, because the greenhouse effect is increasingly serious nowadays, sea level increasing is one of the outstanding show the influence of greenhouse effect and from the sea ice extent, it will be easy to know how the sea level changes, and from the data of sea ice extent of past, the value in the future could be forecasted by using time series analysis.

Source Data

The dataset I use is from Kaggle.com, it is stored the sea ice extent from October 1978 to June 2017. There are 7 variables in this dataset, Year, Month, Day, Extent (10^6 sq km), Missing (10^6 sq km), Source.Data, hemisphere. There are 24908 rows data in this dataset, half north and half south.

	Year	Month	Day	Extent	Missing	Source.Data	hemisphere
1	1978	10	26	10.231	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
2	1978	10	28	10.420	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
3	1978	10	30	10.557	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
4	1978	11	1	10.670	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
5	1978	11	3	10.777	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
6	1978	11	5	10.968	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
7	1978	11	7	11.080	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
8	1978	11	9	11.189	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
9	1978	11	11	11.314	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north
10	1978	11	13	11.460	0	['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsf...	north

(A snapshot of the dataset)

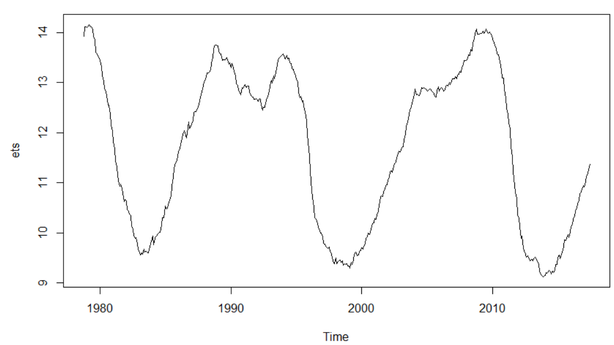
Data Clean and Preparation

For the data clean, I drop the 'Source.Data' column, since it doesn't have any relationship with the analysis. I replace 'North' and 'South' in hemisphere column with 'N' and 'S', and combine Year, Month, Day column into one column called 'Date" with 'YYYY-MM-DD' format. Since I want to find a whole trend of both pole sea ice extent, and each extent data of both pole has the same relational date, then I add the extent of both pole and get the mean into a new column called 'Extent'.

	Date	Extent
1	1978-10-26	13.9275
2	1978-10-28	14.1115
3	1978-10-30	14.1135
4	1978-11-01	14.0985
5	1978-11-03	14.1315
6	1978-11-05	14.1555
7	1978-11-07	14.1185
8	1978-11-09	14.1085
9	1978-11-11	14.0645
10	1978-11-13	13.9155

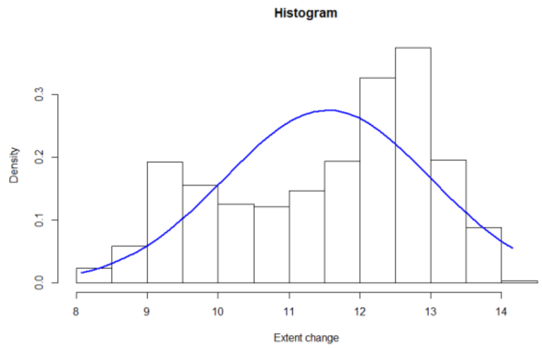
(A snapshot of the new dataset)

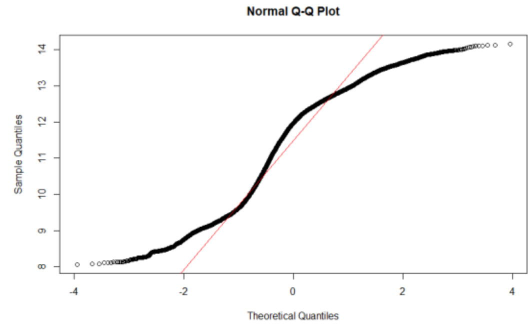
Time plot and basic stats



(Time plot for the dataset)

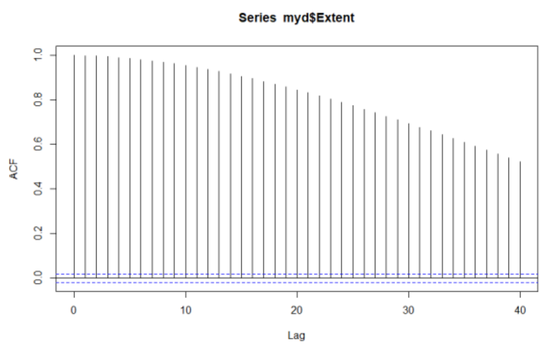
This is the time plot of the dataset, from the plot I can know that, the sea ice extent is changing periodic, but I notice the valley value of the periodic change is decreasing. The first valley value in about 1983, is about 9.5 million sq km; the second valley value in 1999, is about 9.3 million sq km; the third valley value in about 2015, is close to 9 million sq km. So, from the time plot, I can tentative predict that the sea ice extent is decreasing.

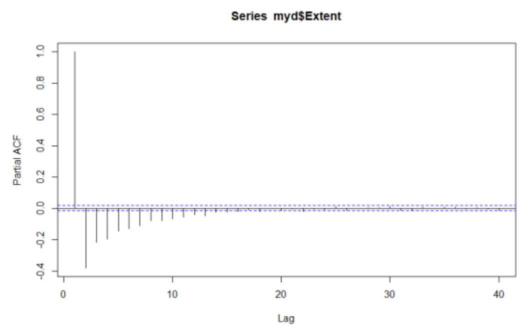




(Histogram) (Q-Q plot)

These two plots are the histogram plot and Normal Q-Q plot about the dataset, from the histogram, this dataset is not normal distributed, and from Q-Q plot this graph is not close to a straight line.



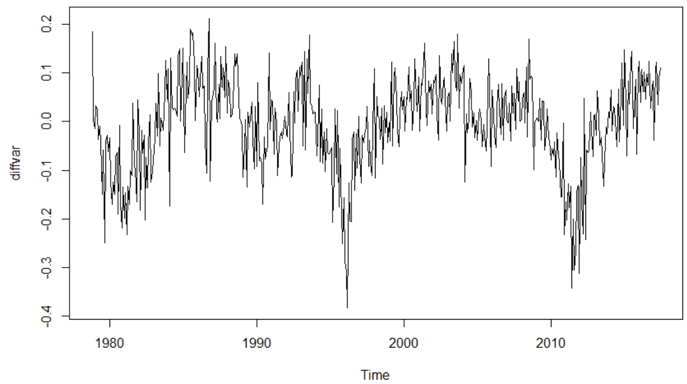


(ACF plot of extent) (PACF plot of extent)

These two are ACF and PACF of the dataset, from the ACF plot, the value of the ACF decays very slowly which means this dataset is a non-stationary time series.

Model Selection

Since it is non-stationary time series, then I deicide to do difference on the dataset, after first difference I get the following time plot, and then I apply the Dickey-Fuller test on the difference data, the p-value at lags-3 and lags-5 are smaller than 0.05, so the null-hypothesis of non-stationary is rejected.



(Time plot after first difference)

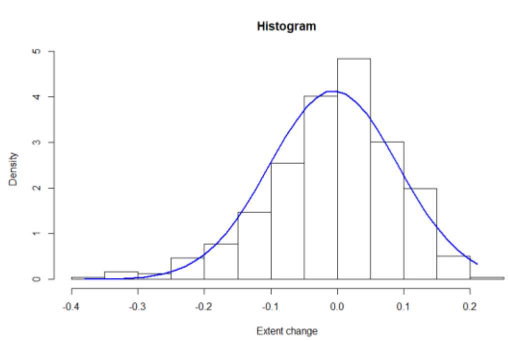
```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 3
  STATISTIC:
    Dickey-Fuller: -4.1979
  P VALUE:
    0.01

Description:
 Tue Mar 06 16:15:20 2018 by user: Yiyang Yang
```

(DF test Lags-3)

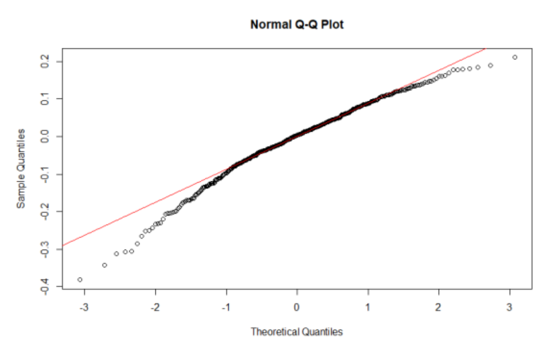
```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 5
  STATISTIC:
    Dickey-Fuller: -3.3144
  P VALUE:
    0.01619

Description:
 Tue Mar 06 16:16:04 2018 by user: Yiyang Yang
```

(DF test Lags-5)

Then I create the histogram and Q-Q plot of the difference part.



(Histogram after 1st difference)



(Q-Q plot after 1st difference)

The new histogram shows that the dataset is close to a normal distribution and the Q-Q plot is also close to a straight line, this is good and ready to do the time series analysis.

So, I decide to use ARIMA model on this dataset to begin my analysis.

ARIMA Model and Diagnostics

I choose to use auto.arima() to get a good model for the data. After applying with BIC criteria, I get the best model for the data is ARIMA (2, 1, 1) with the smallest BIC value.

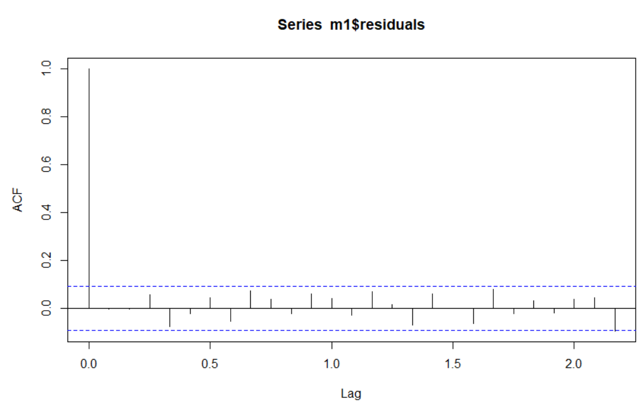
```
Fitting models using approximations to speed things up...

ARIMA(2,1,2)(1,0,1)[12] with drift         : -1093.112
ARIMA(0,1,0)            with drift         : -836.3266
ARIMA(1,1,0)(1,0,0)[12] with drift         : -983.4493
ARIMA(0,1,1)(0,0,1)[12] with drift         : -932.3853
ARIMA(0,1,0)                               : -840.9563
ARIMA(2,1,2)(0,0,1)[12] with drift         : -1094.527
ARIMA(2,1,2)            with drift         : -1100.053
ARIMA(1,1,2)            with drift         : -1095.327
ARIMA(3,1,2)            with drift         : -1094.45
ARIMA(2,1,1)            with drift         : -1105.744
ARIMA(1,1,0)            with drift         : -978.432
ARIMA(2,1,1)                               : -1111.868
ARIMA(2,1,1)(1,0,0)[12]                    : -1108.753
ARIMA(2,1,1)(0,0,1)[12]                    : -1106.328
ARIMA(2,1,1)(1,0,1)[12]                    : -1104.891
ARIMA(1,1,1)                               : -1095.895
ARIMA(3,1,1)                               : -1105.625
ARIMA(2,1,0)                               : -1070.512
ARIMA(2,1,2)                               : -1106.185
ARIMA(1,1,0)                               : -983.9818
ARIMA(3,1,2)                               : -1100.574

Now re-fitting the best model(s) without approximations...

ARIMA(2,1,1)                               : -1111.054
```

(ARIMA model selection)

Then I apply the model to the data get the following coefficient for the model. Then I get the model function for predicting: $(1 - 0.744B - 0.200B^2)X_t = (1 - 0.607B)a_t$.

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1  0.744108   0.071215 10.4488 < 2.2e-16 ***
ar2  0.199610   0.061176  3.2628  0.001103 **
ma1 -0.607148   0.062340 -9.7393 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To make sure if the model is good enough I do the diagnostics by using ACF for residual and Box-Ljung test.



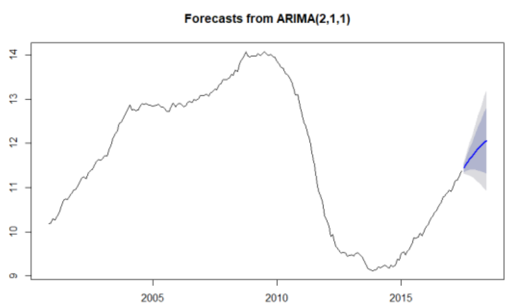
(ACF of residuals)

```
        Box-Ljung test

data:  m1$residuals
X-squared = 1.523, df = 2, p-value = 0.467
```

```
        Box-Ljung test

data:  m1$residuals
X-squared = 5.4534, df = 5, p-value = 0.3631
```

The ACF value are decaying quickly to close to zero and the p-value of the Box-Ljung is larger than 0.05, indicating non-significant. The ACF plot of residual and the Box-Ljung test both show a good result, which means the model is good for the forecast.

By using this model I predict the sea ice extent in one year period after the last month of the orginal dataset and get a trend graph.



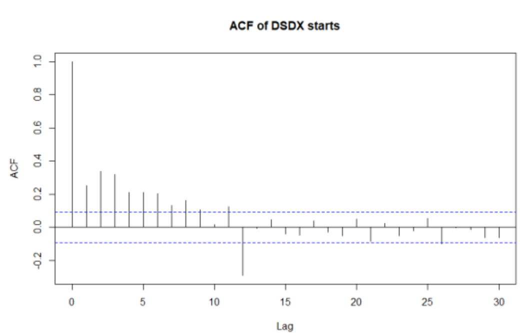
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jul 2017	11.43957	11.34817	11.53098	11.29978	11.57937
Aug 2017	11.51229	11.37388	11.65069	11.30061	11.72396
Sep 2017	11.57998	11.38908	11.77089	11.28802	11.87194
Oct 2017	11.64487	11.39928	11.89045	11.26927	12.02046
Nov 2017	11.70666	11.40387	12.00945	11.24358	12.16974
Dec 2017	11.76559	11.40351	12.12767	11.21184	12.31935
Jan 2018	11.82178	11.39862	12.24494	11.17462	12.46895
Feb 2018	11.87535	11.38961	12.36110	11.13248	12.61823
Mar 2018	11.92643	11.37685	12.47602	11.08592	12.76695
Apr 2018	11.97514	11.36067	12.58960	11.03539	12.91488
May 2018	12.02157	11.34137	12.70177	10.98130	13.06185
Jun 2018	12.06585	11.31923	12.81247	10.92399	13.20770

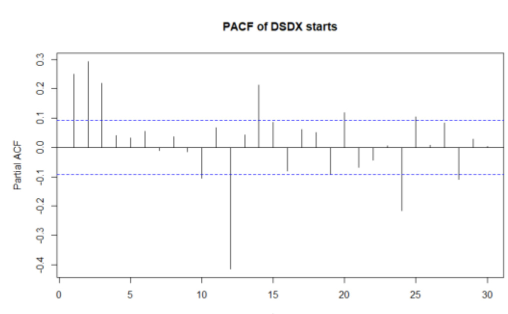
(Forecast plot) (Forecast values)

Then I think ARIMA model maybe not good enough, since the sea ice extent change is also influence by the climate in different seasons. So, I also do the SARIMA model for the data.

SARIMA Model and Diagnostics

At first, I make graph about ACF of Seasonal difference of the extent, and PACF of Seasonal difference of the extent.





(ACF of seasonal difference) (PACF of seasonal difference)

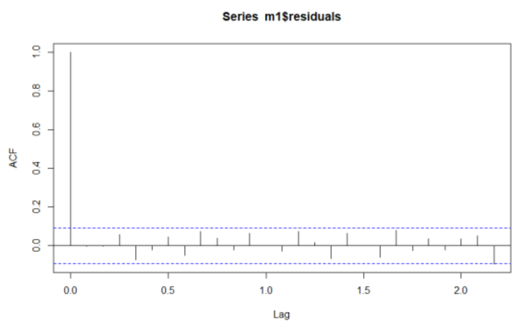
From the plots, I tentative get a model for the data which is AR(2) + MA(1) + SAR(1). Then I fit the model to the data to get the coefficients for the function.

```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1   0.731006   0.067788  10.7837 < 2.2e-16 ***
ar2   0.217419   0.060245   3.6089 0.0003075 ***
ma1  -0.590646   0.058230 -10.1433 < 2.2e-16 ***
sar1 -0.086803   0.048489  -1.7901 0.0734303 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then I get a function for the ARIMA(2, 1, 1)(1, 0, 0)$_{12}$ model:

$$(1 - 0.742B - 0.199B^2)(1 - 0.041B^{12})(1-B)X_t = (1 + 0.606B)a_t$$

Also to make sure if this model is good enough for forecast, I do the disgnostics.



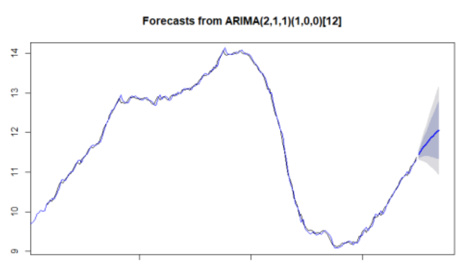
(ACF of residuals)

```
> acf(m1$residuals)
> Box.test(m1$residuals, 4, "Ljung-Box", fitdf = 2)

        Box-Ljung test

data:  m1$residuals
X-squared = 4.1035, df = 2, p-value = 0.1285

> Box.test(m1$residuals, 12, "Ljung-Box", fitdf = 2)

        Box-Ljung test

data:  m1$residuals
X-squared = 11.978, df = 10, p-value = 0.2865
```

(Snapshot of Box-Ljung test)

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 result of ACF plot and Box-Ljung test are good, which means this model is good for forecasting.

By using the SARIMA model, I also predict one year period after the last month of the original dataset and get a trend graph.



(Forecast plot)

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jul 2017	11.44031	11.34888	11.53174	11.30048	11.58014
Aug 2017	11.51309	11.37472	11.65145	11.30147	11.72470
Sep 2017	11.58283	11.39212	11.77354	11.29116	11.87450
Oct 2017	11.64581	11.40065	11.89097	11.27087	12.02074
Nov 2017	11.70690	11.40486	12.00893	11.24497	12.16882
Dec 2017	11.76625	11.40534	12.12716	11.21429	12.31821
Jan 2018	11.81780	11.39632	12.23928	11.17320	12.46239
Feb 2018	11.87278	11.38932	12.35624	11.13339	12.61217
Mar 2018	11.92589	11.37928	12.47251	11.08992	12.76187
Apr 2018	11.97305	11.36233	12.58377	11.03903	12.90707
May 2018	12.02021	11.34462	12.69580	10.98698	13.05343
Jun 2018	12.06617	11.32511	12.80723	10.93282	13.19952

(Forecast values)

Backtest

From both model, I notice they all get very closed results, but which one is the better model in a scientific way. I apply Backtest for both model to decide, which model is better.

```
[1] "RMSE of out-of-sample forecasts"
[1] 0.06847923
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.05426884
[1] "Mean Absolute Percentage error"
[1] 0.00494128
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.004942303
```

(Backtest of ARIMA)

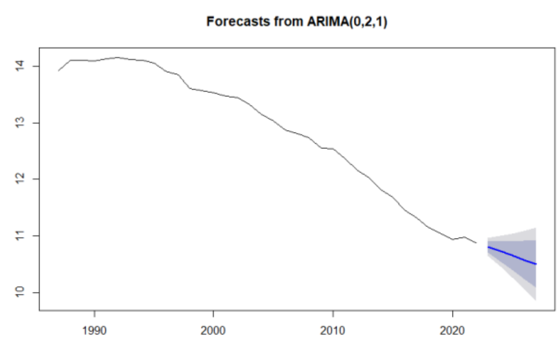
```
[1] "RMSE of out-of-sample forecasts"
[1] 0.08274096
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.06514572
[1] "Mean Absolute Percentage error"
[1] 0.005913751
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.005919107
```

(Backtest of SARIMA)

RMSE for ARIMA model is 0.0685, MAPE for ARIMA model is 0.0049; RMSE for SARIMA model is 0.0827, MAPE for SARIMA model is 0.0059. So according to backtest, the result shows that ARIMA model is the better model for the dataset due to the smaller MAPE value.

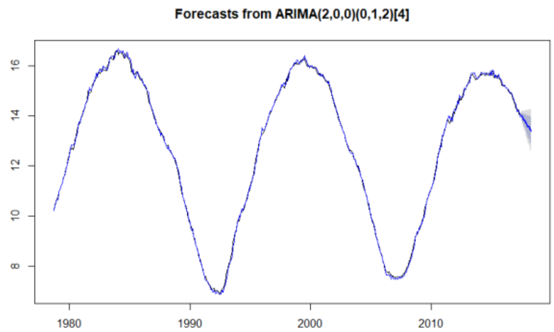
More Plots and Facts

I also use R to create more forecast plots, based on different situations. The first plot I create, is the annual sea ice extent of both pole changes trend, it shows that the global change trend of the sea ice extent in both pole is decrease from about 14 million sq km in 1989 decrease to about 11 million sq km in 2017, and it keep decrease in future years.



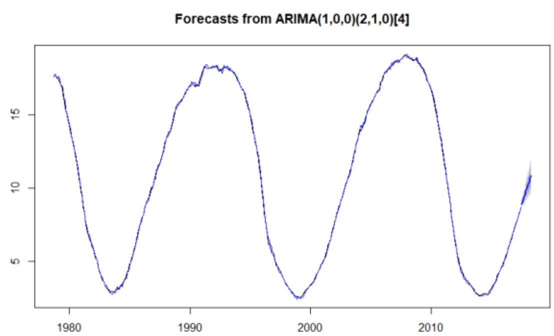
(Forecast of annual ice extent change)

The second plot shows the change trend in North pole, the sea ice extent is changing periodic in the north ploe, but I notice that, the range between peak and valley value is smaller, from another side to show the sea ice extent is decreasing.



(Forecast of north pole ice extent change)

The third plot shows the change trend in South pole, the sea ice extent is also changing periodic. But from this plot, I can't find some evidences that show the decreasing trend of sea ice extent.



(Forecast of south pole ice extent change)

Conclusion

After the analysis by using ARIMA and SARIMA models to the dataset, they both show very good results and forecast about the sea ice extent changes, the extent shows a decreasing trend same as the common fact. And the change range of the past data shocks me very much, from 14 million sq km to 11 million sq km in about 40 years, nearly one third of China area is melted into water. As an individual, I can't do more thing to this result, only have a good wish that greenhouse effect doesn't be more serious any more. By comparing ARIMA and SARIMA models with backtest, ARIMA is the better model to predict the furture, because of the smaller MAPE value.

From this project I learnt a lot of knowledge and skills by using R to do time series prediction with variety models. I think they are very helpful with my future works.

Appendix

R code

```
library(ggplot2)
library(forecast)
library(tseries)
library(fBasics)
library(lmtest)
library(zoo)
library(fUnitRoots)

#Read data
ds <- read.table("D:/CSC 425/seaice.csv", header = T, sep = ',')
df <- subset(ds, select = -c(6))

df$hemisphere <- gsub('north', 'N', df$hemisphere)
df$hemisphere <- gsub('south', 'S', df$hemisphere)

df$Date <- as.Date(with(df, paste(Year, Month, Day, sep = '-')), "%Y-%m-%d")

#Split into two dataframe based on hemisphere N or S
mydf <- subset(df, select = -c(1:3))
mydf <- mydf[, c(4, 1, 2, 3)]
mydfN <- split(mydf, mydf$hemisphere)[['N']]
mydfS <- split(mydf, mydf$hemisphere)[['S']]

newDate <- mydfN$Date
newExtent <- (mydfN$Extent + mydfS$Extent)/2

myd <- data.frame('Date' = newDate, 'Extent' = newExtent)

extts = ts(myd[, 2], start = c(1978, 10), end  = c(2017, 6), freq = 12)
extAts = ts(myd[, 2], start = c(1978, 10), end  = c(2017, 6), freq = 1)
basicStats(myd$Extent)

#Histogram and Q-Q plot
hist(myd$Extent, xlab = "Extent change", prob = TRUE, main = "Histogram")
xfit <- seq(min(myd$Extent), max(myd$Extent), length = 40)
yfit <- dnorm(xfit, mean = mean(myd$Extent), sd = sd(myd$Extent))
lines(xfit, yfit, col="blue", lwd = 2)

qqnorm(myd$Extent)
qqline(myd$Extent, col = 2)

plot(extts, ylab = 'Extent change')
```

```
acf(myd$Extent)
pacf(myd$Extent)

#First difference
adfTest(myd$Extent, lags = 5, type = c("ct"))
diffvar = diff(extts)
plot(diffvar)
adfTest(coredata(diffvar), lags = 5, type=c("c"))


hist(diffvar, xlab = "Extent change", prob = TRUE, main = "Histogram")
xfit <- seq(min(diffvar), max(diffvar), length = 40)
yfit <- dnorm(xfit, mean = mean(diffvar), sd = sd(diffvar))
lines(xfit, yfit, col="blue", lwd = 2)

qqnorm(diffvar)
qqline(diffvar, col = 2)

#ARMIA Model
auto.arima(extts, ic =c("bic"), trace = TRUE, allowdrift = TRUE)
m1 = Arima(extts, order = c(2,1,1), method = 'ML')
coeftest(m1)

acf(m1$residuals)

Box.test(m1$residuals,lag = 3,type = 'Ljung-Box', fitdf = 1)
Box.test(m1$residuals,lag = 6,type = 'Ljung-Box', fitdf = 1)

f = forecast(m1, h = 12)
f
plot(f, include = 200)

#SARIMA
x <- myd$Extent
ets = ts(x, frequency = 12, start = c(1978, 10), end  = c(2017, 6))
plot(ets, type = 'l')

hist(x, xlab = "Extent", freq = F)
xfit <- seq(min(x), max(x), length = 40)
yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
lines(xfit, yfit, col = "black", lwd = 2)

par(mfcol = c(1, 1))
acf(as.vector(ets), lag.max = 30, main = "ACF")

dx = diff(ets)
```

```
acf(as.vector(dx), lag.max = 26, main = "ACF of DX starts")

sdx = diff(dx, 12)
acf(as.vector(sdx), lag.max = 30, main = "ACF of DSDX starts")
pacf(as.vector(sdx), lag.max = 30, main = "PACF of DSDX starts")

m2 = Arima(ets, order = c(2, 1, 1), seasonal = list(order = c(1, 1, 0), period =
12), method = "ML")
m2
coeftest(m1)
acf(m1$residuals)

Box.test(m1$residuals, 4, "Ljung-Box", fitdf = 2)
Box.test(m1$residuals, 12, "Ljung-Box", fitdf = 2)

f1 = forecast(m2, h = 12)
f1
plot(f1, include = 200)
lines(ts(c(f1$fitted, f1$mean), frequency = 12, start = c(1978, 10), end = c(2017,
6)), col = "blue")

#Backtest
source('D:/CSC 425/backtest.R')
pm1 = backtest(m1, extts, 200, 1)
pm2 = backtest(m2, extts, 200, 1)

#Annual Change (ARIMA)
auto.arima(extAts, ic =c("aic"), trace = TRUE, allowdrift = TRUE)
mA1 = Arima(extAts, order = c(0,2,1), method = 'ML')
coeftest(mA1)

Box.test(m1$residuals,lag = 3,type = 'Ljung-Box', fitdf = 1)
Box.test(m1$residuals,lag = 6,type = 'Ljung-Box', fitdf = 1)

fA = forecast(mA1, h = 5)
plot(fA, include = 200)

#Tentative SARIMA Model (North pole)
xN <- mydfN$Extent
nts = ts(xN, frequency = 12, start = c(1978, 10), end  = c(2017, 6))
plot(nts, type = 'l')

hist(xN, xlab = "Extent", freq = F)
xfit <- seq(min(xN), max(xN), length = 40)
yfit <- dnorm(xfit, mean = mean(xN), sd = sd(xN))
lines(xfit, yfit, col = "black", lwd = 2)
```

```
par(mfcol = c(1, 1))
acf(as.vector(nts), lag.max = 30, main = "ACF")

ndx = diff(nts)
acf(as.vector(ndx), lag.max = 26, main = "ACF of DX extents")

nsdx = diff(ndx, 12)
acf(as.vector(nsdx), lag.max = 30, main = "ACF of DSDX extents")

mn = Arima(nts, order = c(2, 0, 0), seasonal = list(order = c(0, 1, 2), period =
4), method = "ML")
coeftest(mn)
acf(mn$residuals)

Box.test(mn$residuals, 4, "Ljung-Box", fitdf = 2)
Box.test(mn$residuals, 12, "Ljung-Box", fitdf = 2)

fn = forecast(mn, h = 10)
plot(fn, include = 1000)
lines(ts(c(fn$fitted, fn$mean), frequency = 12, start = c(1978, 10), end = c(2017,
6)), col = "blue")

#Tentative SARIMA Model (South pole)
xS <- mydfS$Extent
sts = ts(xS, frequency = 12, start = c(1978, 10), end = c(2017, 6))
plot(sts, type = 'l')

par(mfcol = c(1, 1))
acf(as.vector(sts), lag.max = 30, main = "ACF")

sdx = diff(sts)
acf(as.vector(sdx), lag.max = 26, main = "ACF of DX extents")

ssdx = diff(sdx, 12)
acf(as.vector(ssdx), lag.max = 30, main = "ACF of DSDX extents")

ms = Arima(sts, order = c(1, 0, 0), seasonal = list(order = c(2, 1, 0), period =
4), method = "ML")
coeftest(ms)
acf(ms$resid)

Box.test(ms$residuals, 4, "Ljung-Box", fitdf = 3)
Box.test(ms$residuals, 12, "Ljung-Box", fitdf = 3)

fs = forecast(ms, h = 10)
```

```
plot(fs, include = 1000)
lines(ts(c(fs$fitted, fs$mean), frequency = 12, start = c(1978, 10), end = c(2017,
6)), col = "blue")
```