Forecasting of the asset value of S&P500

Haozhe Chen

Contents

Abstract

The financial markets contain a plethora of statistical patterns. The behavior of those patterns is similar to the behavior of the natural phenomena patterns. That means that both are affected by unknown and unstable variables. Which leads to high unpredictability and volatility. That makes hard to forecast future behavior.

Introduction

The purpose of this project is to analyze and compare different time series models for the financial forecasting of the asset value of the S&P 500. The data source is from Yahoo Finance, the path is where you can find the data:

'https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC'

Instead of forecasting the asset value directly, I have focused on the net flows and reconciled the projected flows to the asset by the capital gain rate. Therefore, the model performance evaluation here is not only based on the metrics, such as MAPE but also evaluated by the percentage difference on the average asset forecasting, which is from a practical point of view.

The data is from 2018-01-01 to 2020-3-31. In the modeling part, I use the data before 2020 as the train data and data after 2020 as testing data for model comparison. I used the ARIMA as the baseline model, then compared the SARIMA model and HoltWinter model. When the number of forecasted series is big, the ARIMA forecasting converges to the mean, which is not robust. SARIMA could capture the seasonality, but it may have overfitting problems due to the complexity and variability of the data. It turns out the HoltWinter is better than the ARIMA and SARIMA.

Date preparation

- Date description
  Here is how the original dataset looks like. The useful variables here are Date, Adj Close, Volume.
  Date — Date ranges from 2018-01-02 to 2020-3-31. The dataset does not include the weekends and holidays, so the dates are not consecutive.

Adj Close — The adjusted close price, it was adjusted after dividend and fund split. It represents the standardized daily net asset value.
Volume — The total number of shares that are traded (bought and sold).

Date	Open	High	Low	Close	Adj Close	Volume
2018-01-02	2683.73	2695.89	2682.36	2695.81	2695.81	3367250000
2018-01-03	2697.85	2714.37	2697.77	2713.06	2713.06	3538660000
2018-01-04	2719.31	2729.29	2719.07	2723.99	2723.99	3695260000
2018-01-05	2731.33	2743.45	2727.92	2743.15	2743.15	3236620000
2018-01-08	2742.67	2748.51	2737.60	2747.71	2747.71	3242650000
2018-01-09	2751.15	2759.14	2747.86	2751.29	2751.29	3453480000

- Date preparation
  This is one of the main parts of this project, the original data are not consecutive, and lack of useful information. For example, the flow data is not available. Hence, I have implemented certain computations to process the data, got the information that I want. I have written several functions and wrapped them into the file 'functions_EDA.R', feel free to play with it. Here are some details of how I have processed the data.

1. Use 'Adj Close' as Net Asset Value(NAV), multiplied by the number of volume to get the asset value, differencing the asset value to get the daily net flows. Scale the asset value and the daily net flows as millions.
2. Add time ticks, including year, months, weeks, and weekdays to the data frame.
3. Label outliers: there are many spikes in the flows. I used lowess to fit the flow and then use the residual to check for the outlier. I calculated the z score of the absolute value of residuals. picked 'z' greater than 3 as outliers, then replaced outliers by the smoothed value from lowess. The smoothing is using the closest 30 data points, approximately 1.5 months. This smoothing factor is achieved by empirical observations on the stock. It might not be optimal, but useful, as least it can avoid the impact of those spikes on the forecasted value.
4. Adding 5 points moving average flow and 21 moving average flow to the data, since on average, there are 5 days a week and 21 days a month in the dataset. The EDA is done by the time ticks, but to save space, I won't include the EDA here.
5. Eliminate useless columns and clean the data, get the following data frame.

Date	Asset	net_flow	log_return	de_outlier	ma_5	ma_21
2018-01-02	9077466	NA	NA	NA	NA	NA
2018-01-03	9600597	523130.69	0.0063784	523130.69	523130.687	523130.69
2018-01-04	10065851	465254.14	0.0040205	465254.14	465254.142	465254.14
2018-01-05	8878534	-1187317.41	0.0070091	-1187317.41	84811.741	-1187317.41
2018-01-08	8909862	31327.87	0.0016610	31327.87	45607.036	31327.87
2018-01-09	9501525	591663.42	0.0013021	591663.42	2344.109	591663.42

- Data visualization
  The plots here give you a visualily comparison of how the data was processed.
  The first plot is the adjusted close price (NAV), we could see that there are huge decreases around March 2020, which might due to the pandemic of COVID-19.
  Correspondingly, the flow around March, which observed from the second plot oscillates deeper. Note that the flows on the plot were scaled by Z scores, and centered at different positions, for making comparisons.We could see that the de-outlier flow has fewer spikes, but it was not smoothed, it is used to eliminate the impact of the spikes on forecasting.
  The third plot shows the comparison between de-outlier flows and smoothed flow. The two tails of smoothed data were not smoothed because the moving average at those points cannot be computed, so

I replaced those points by the original net flow.

Forecasting models

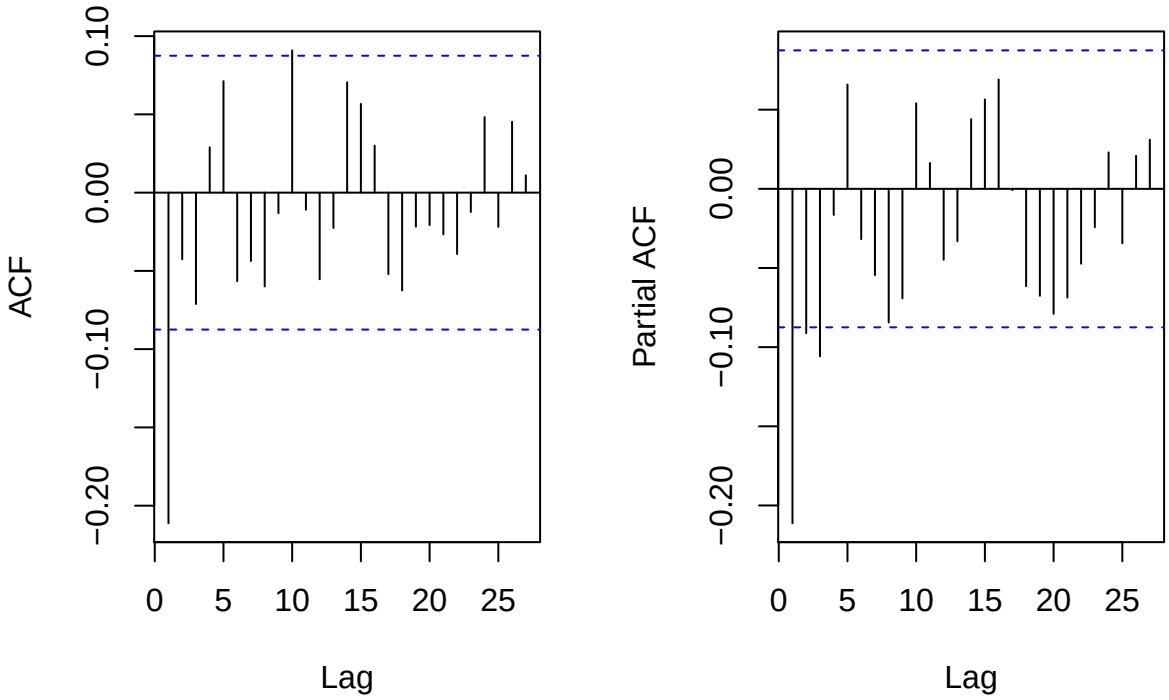
ARIMA

- Stationarity Test
  The variance of the flow series looks stable, so I won't use transformation. First test the stationary of the series.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  train
## Dickey-Fuller = -9.2795, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

```
## Number of differences required for a stationary series is 0
```

- Model Identification
  From the correlation plots, we could see that many ACF and PACF are marginally significant, and they both decay to zero, thus it may have a lower ARMA order, it could be p=1, q=1. The PACF cuts-off at certain lag, hence I compared ARIMA(1,0,1), ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(2,0,0) , and the model returned from auto.arima().

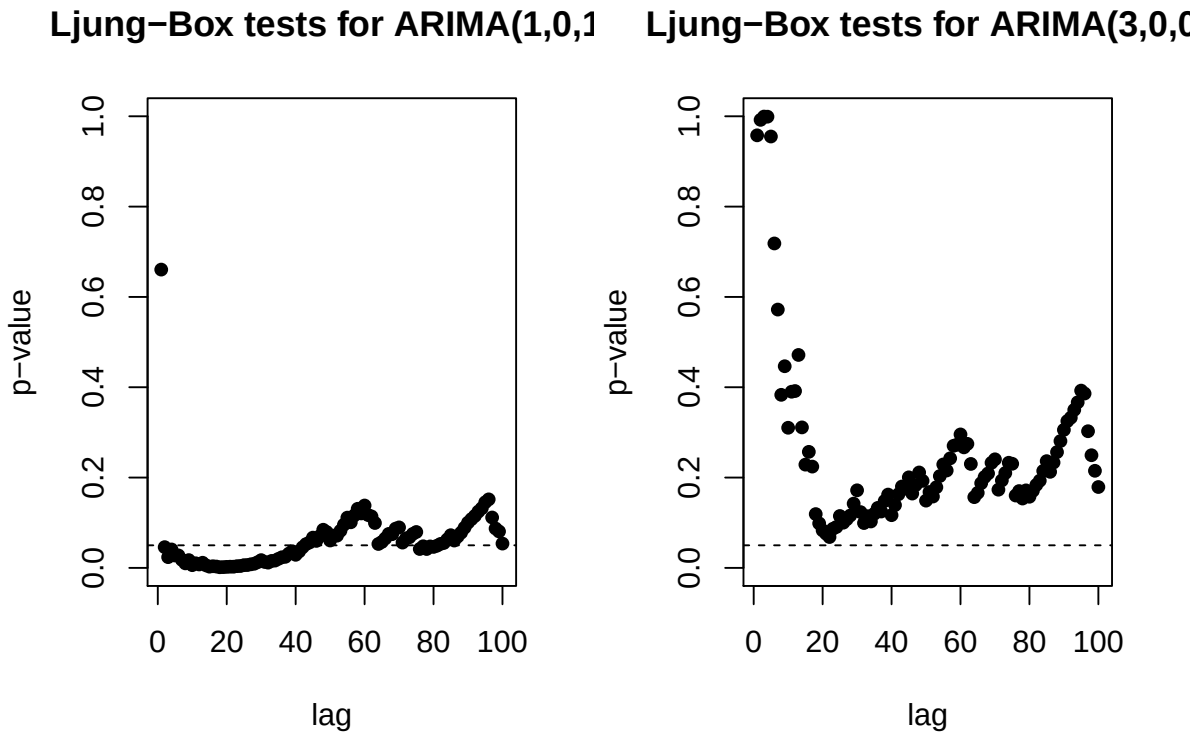


The AICc table ends up being:

Model	AICc
ARIMA(0,0,2)	15420.1
ARIMA(1,0,1)	15419.3
ARIMA(1,0,0)	15425.43
ARIMA(2,0,0)	15423.27
ARIMA(3,0,0)	15419.66

Model Evaluation

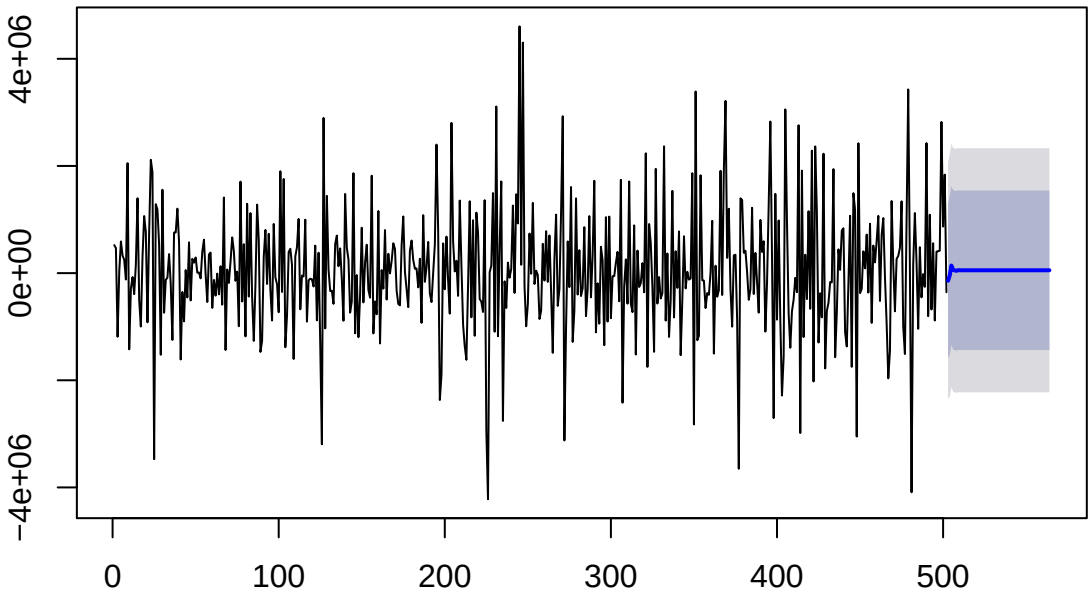
It turns out that ARIMA(1,0,1), and ARIMA(3,0,0) have almost the same lowest AICc. Instead of looking at the ACF, and making a conclusion by eyes, I implemented the Ljung-Box test. The plots show the p-values of the Ljung-Box test from lag 1 to lag 100, the dashed line is the 0.05. Since $H_o : \rho_1 = \rho_2 = ... = \rho_k = 0$, we can make the decision that the residuals of ARIMA(3,0,0) is a white noise, whereas the residuals of ARIMA(1,0,1) is not.



When looking at the coefficient estimations, the 95% confidence interval does not across 0, hence we are safe about using ARIMA(3,0,0) for forecasting.

```
## Series: train
## ARIMA(3,0,0) with non-zero mean
##
## Coefficients:
##           ar1      ar2      ar3      mean
##       -0.2397  -0.1159  -0.1060  53000.13
## s.e.   0.0444   0.0454   0.0445  34254.88
##
## sigma^2 estimated as 1.265e+12:  log likelihood=-7704.77
## AIC=15419.54   AICc=15419.66   BIC=15440.63
```

- Forecasting




```
## The testing errors of ARIMA forecasting are
```

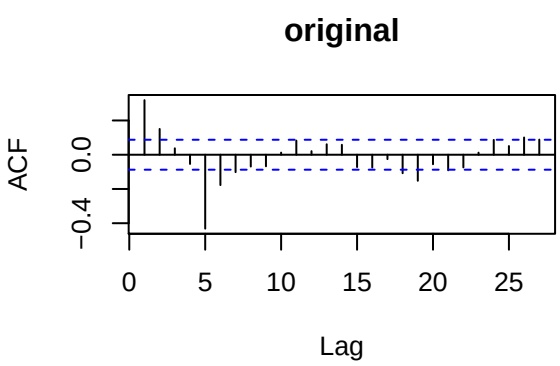
```
##       RMSE        MAE       MAPE
## 1536471.43 1206828.09      99.12
```

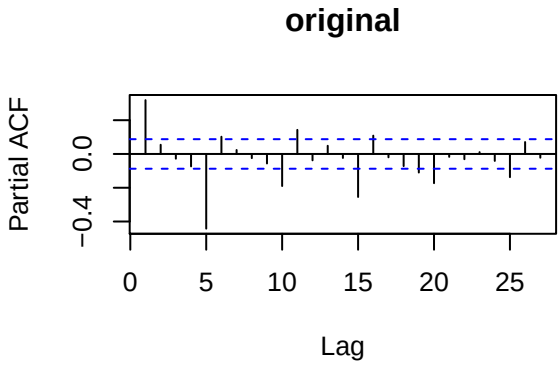
SARIMA

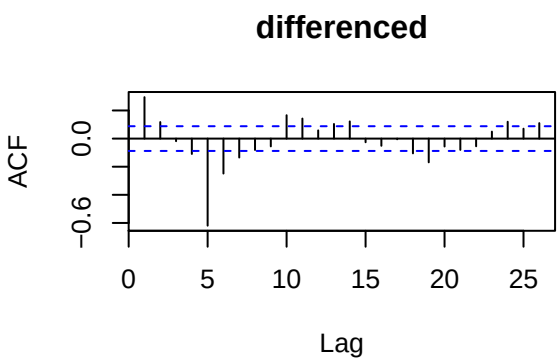
We have already seen how ACF and PACF of de-outlier flow look like, the PACF displays some seasonalities, but it is hard to analyze seasonal trends based on the unsmoothed flow data due to the noisiness of the data. Hence I used the 5 and 21 moving average data to fit the SARIMA model to see what happens.

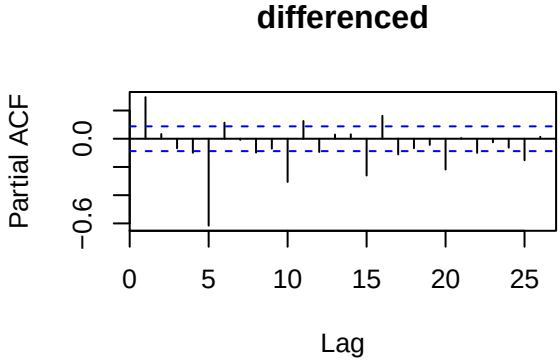
Since the moving average smooths the data by computing average in a certain range, it reduces the impact of spikes and makes easier to find seasonalities.

The data is stationary, I am not going to show the test.let's take a look at the ACF and PACF plots. I differenced the data by 1 lag and compared its ACF and PACF to the undifferenced series. It turns out the models could be $SARIMA(1,0,0)(0,0,1)_5$, $SARIMA(1,0,1)(0,0,1)_5$, $SARIMA(0,0,2)(0,0,1)_5$, $SARIMA(1,0,1)(1,1,1)_5$, $SARIMA(1,0,1)(0,1,2)_5$.





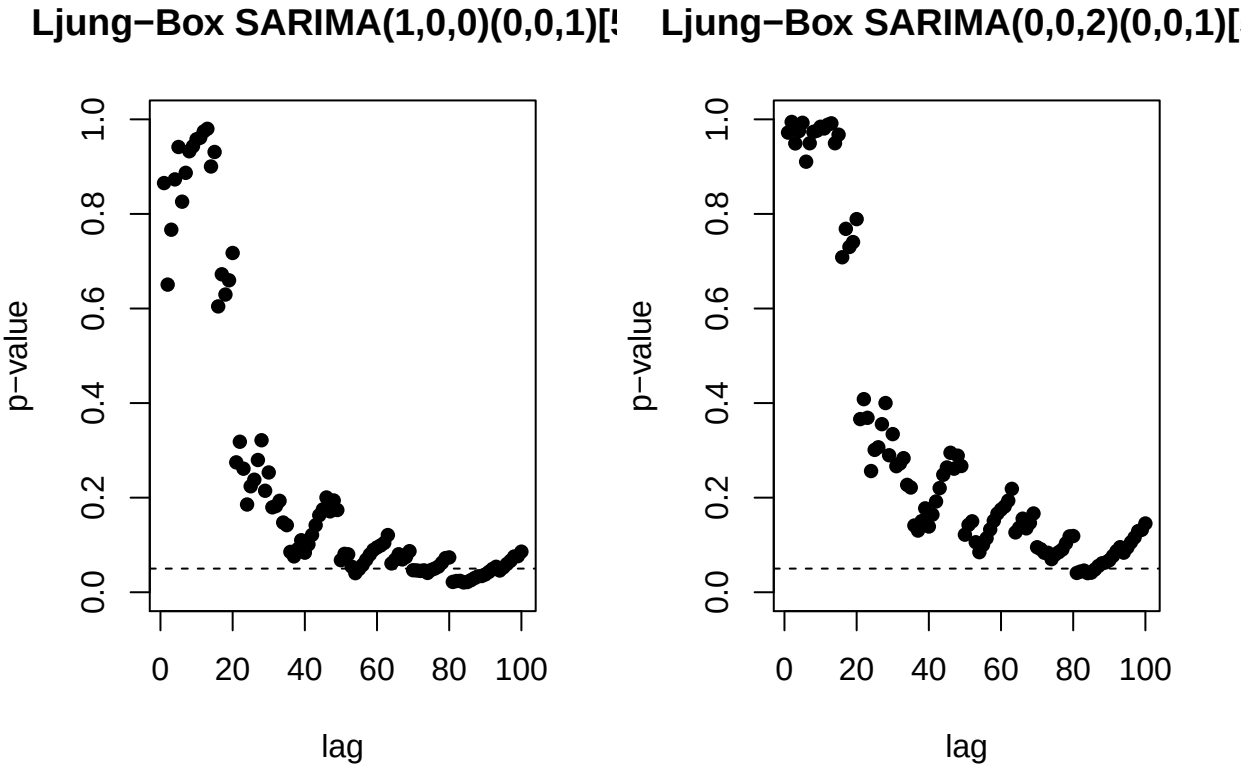




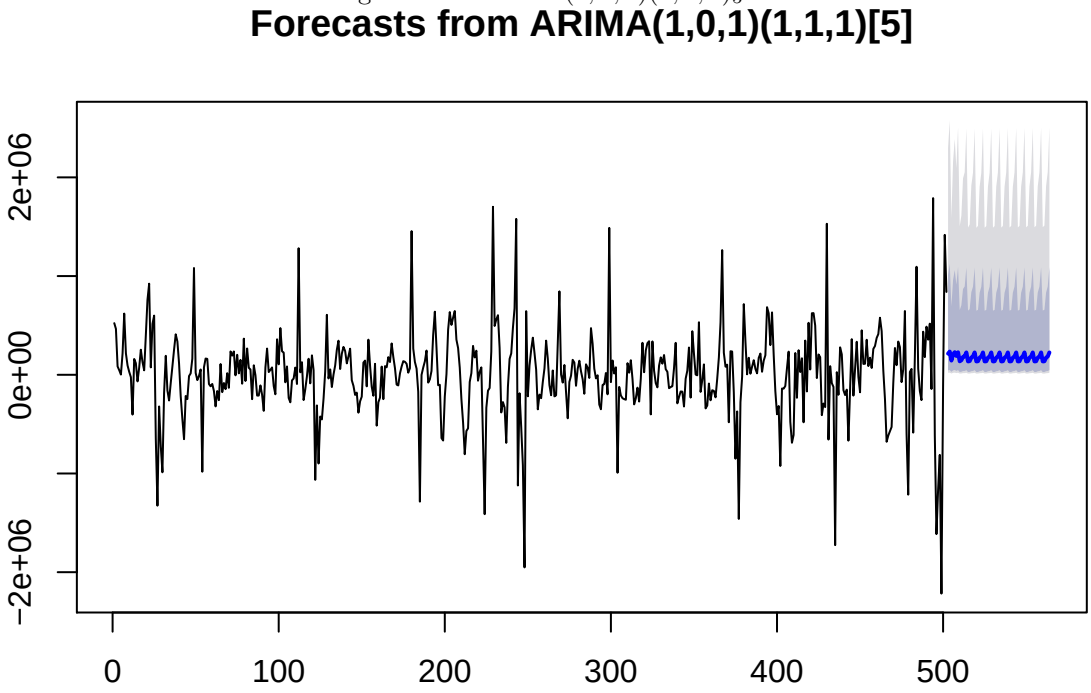
The AICc are shown below:

Model	AICc
SARIMA(1,0,0)(0,0,1)[5]	878.12
SARIMA(1,0,1)(0,0,1)[5]	879.72
SARIMA(0,0,2)(0,0,1)[5]	879.58
SARIMA(1,0,1)(1,1,1)[5]	881.61
SARIMA(1,0,1)(0,1,2)[5]	881.62

After making comparison, we check the $SARIMA(1,0,0)(0,0,1)_5$ and $SARIMA(0,0,2)(0,0,1)_5$ togther. I decided to use $SARIMA(0,0,2)(0,0,1)_5$ for forecasting.



- Forecasting
  Here is the result of forecasting from $SARIMA(0,0,2)(0,0,1)_5$.

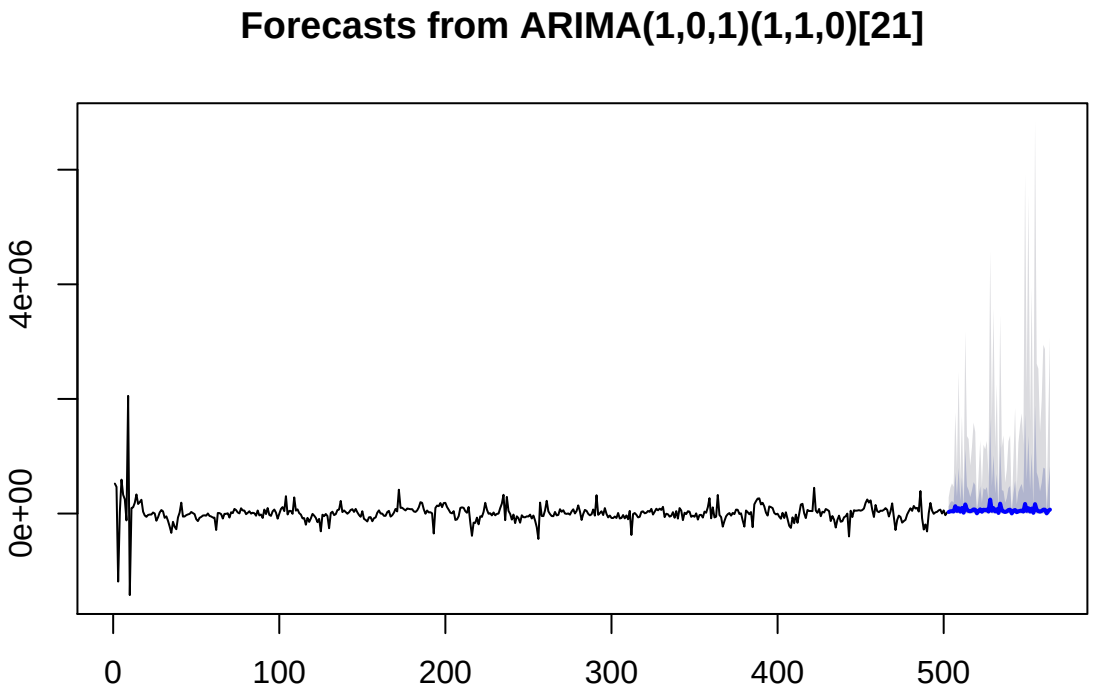



```
## The testing errors of SARIMA(0,0,2)(0,0,1)[5] forecasting are:

##       RMSE       MAE      MAPE
## 631885.22 466786.68    373.61
```

I have implemented the same process to the 21 moving average series. For this forecasting, I used

$SARIMA(1,0,1)(3,1,0)_{21}$ The forecasting results are shown below.




```
## The testing errors of SARIMA(1,0,1)(3,1,0)[21] forecasting are:

##      RMSE       MAE      MAPE
## 992374.84 466659.90    172.27
```

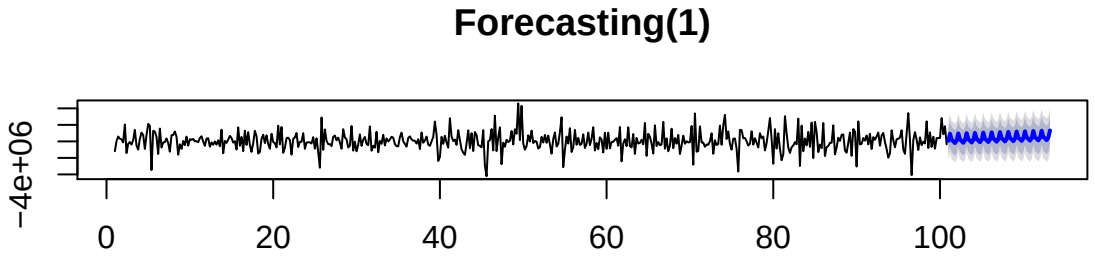
HoltWinters

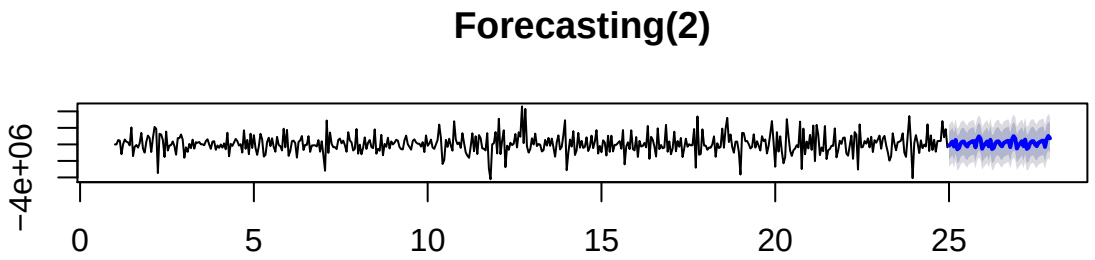
The Holt-Winters forecasting algorithm allows users to smooth a time series exponentially and use that data to forecast areas of interest. Exponential smoothing assigns exponentially decreasing weights and values against historical data to decrease the value of the weight for the older data. In other words, more recent historical data is assigned more weight in forecasting than the older results.

Since HoltWinters model assumes that the data has periods, I trimmed the train data into certain periods, I used 5 and 21 respectively as the frequency, and then fit the model. Here I make two forecastings with different trimmed training series, they are

- Forecasting(1): de-oulier flow with frequency 5
- Forecasting(2): de-oulier flow with frequency 21

The forecasting plots and results are shown below:






```
## The testing errors of Holt_Winters Forecasting(1) are:

##       RMSE        MAE       MAPE
## 1555312.79 1218672.12     143.76

## The testing errors of Holt_Winters Forecasting(2) are:

##       RMSE        MAE       MAPE
## 1441587.48 1111399.71     104.09
```

Metric Deisgn

Since the flows are always over billions, the metrics from flow forecasting models may not be interesting. Also, since the SARIMA model uses the smoothed series to forecast, simply comparing the modeling metrics among different models may not give us a good conclusion here. Instead, I wanted to compare the difference in average assets. The idea is reconciling the projected flows to the asset by the capital gain rate and comparing the difference with the reconciled flows and the actual de-outlier flows.

In order to illustrate my computations, I use the following denotations:

- R: reconciled asset
- r: capital gain rate
- d: de-outlier flow
- f: net flow
- $\hat{y_t}$: forecasted value
- y_t : actual value

  The formulas are $R_t = R_{t-1}e^{r_t} + d_i$, and $r_t = ln(f_t/f_{t-1})$. The percentage difference on the average asset forecasting is calculated by $\frac{1}{n}\sum_i^n(\hat{y_t} - y_t)/y_t$, and $t = 1, 2, ..., n$. The y_t here refers to the de-outlier flow, but we can also use the real flow for comparison.

  Feel free to play with the functions reconcile_asset() and eval_asset() in the 'functions_EDA.R' file.

Model Comparison

	ARIMA	SARIMA on 5 moving average	SARIMA on 21 moving average
RMSE	1536471.43	631885.22	992374.84
MAE	1206828.09	466786.68	466659.90
MAPE	99.12	373.61	172.27
Percentage difference on asseet	16.04	8.70	14.04

	Holt-Winters with freqency 5	Holt-Winters with freqency 21
RMSE	1555312.79	1441587.48
MAE	1218672.12	1111399.71
MAPE	143.76	104.09
Percentage difference on asseet	40.53	3.09

Conclusion

The simple ARIMA has the lowest MAPE, but higher percentage difference on asset value. The drawback of ARIMA is that the point forecasting converges to the mean when doing long-term forecasting, which means that ARIMA does not capture the variability of the data.

SARIMA models have the highest MAPE, but lower percentage difference on asset value compared to the ARIMA model, which means SARIMA captures some seasonalities, but on the other hand they may have overfitting problem.

HoltWinter performs awful on the data with frequency 5, but much better on the data with frequency 21. The weekly frequency can be dangerous than monthly frequency because the actual weekdays may not match with each other over time due to the cut-off of the dates.