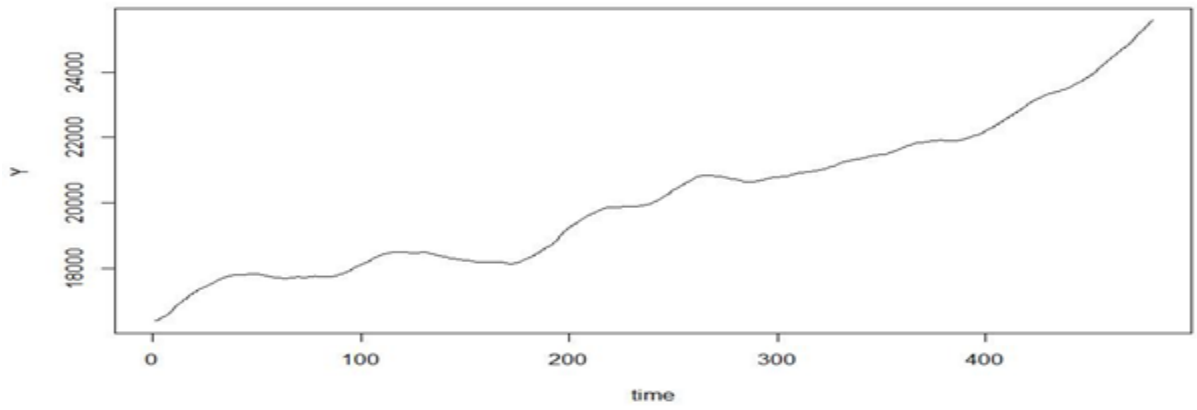
Time series analysis final project

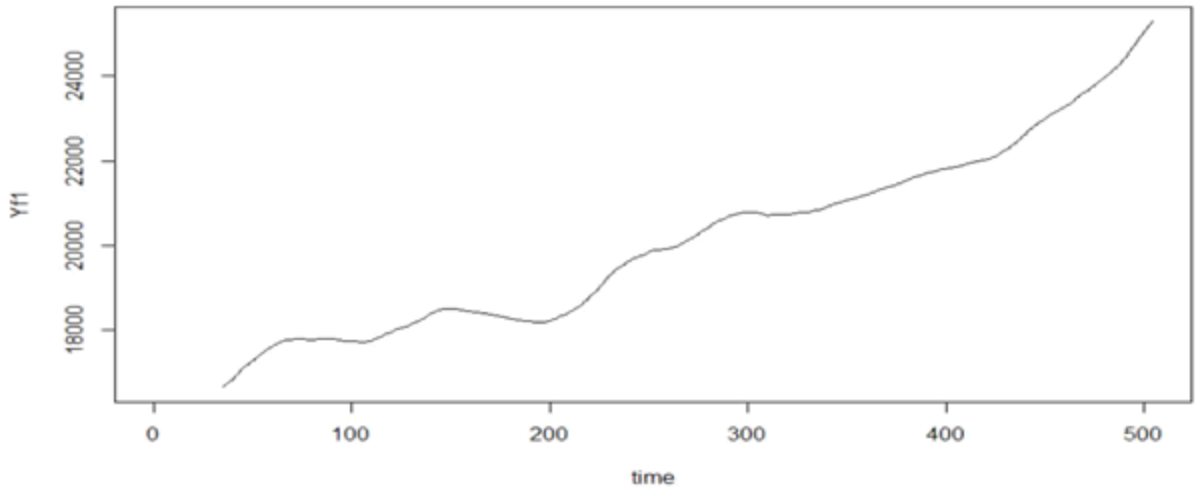
Introduction and hypothesis:

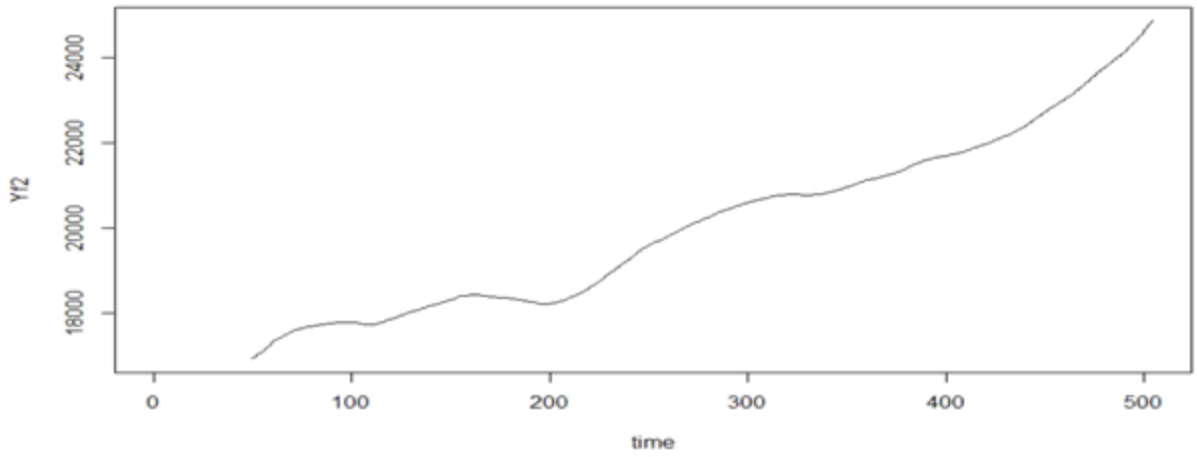
In the time series analysis project, the goal of the project is to determine the Dow jones industrial average (Dow 30) that is the major market index in (US market) i.e the output variable, on the Australian securities exchange(Australia) and the Shanghai stock exchange (China) as the input variables, impact on American stock exchange(US market) by Australian and Chinese stock market respectively. Forecasting the trends of stock market is of extreme importance and profitable to stock market traders and also to the researchers who are always trying to find an analogy to describe the behaviour of stock market. First of all what is stock market? In short A stock market, equity market or share market is the aggregation of buyers and sellers (a loose network of economic transactions, not a physical facility or discrete entity) of stocks (also called shares), which represent ownership claims on businesses; these may include securities listed on a public stock exchange as well as those only traded privately. American stock exchange is located at 11 Wall Street, Lower Manhattan, New York City, New York. It is by far the **world's largest stock exchange by market capitalization of its listed companies at US$21.3 trillion as of June 2017**.The average daily trading value was approximately US$169 billion in 2013. Regarding Australian stock exchange (ASX), he Australian Securities Exchange (ASX, sometimes referred to outside Australia as the Sydney Stock Exchange) is Australia's primary securities exchange. It is owned by the Australian Securities Exchange Ltd, or ASX Limited, an Australian public company (ASX). Prior to December 2006 it was known as the Australian Stock Exchange, which was formed on 1 April 1987. **ASX has an average daily turnover of A$4.685 billion and a market capitalization of around A$1.6 trillion**, making it one of the world's top 15 listed exchange groups. The major market index is the S&P/AX200, an index made up of the top 200 shares in the ASX. Then comes the shanghai stock exchange (China). The Shanghai Stock Exchange (SSE) is a stock exchange that is based in the city of Shanghai, China. It is one of the two stock exchanges operating independently in the People's Republic of China, the other is the Shenzhen Stock Exchange. Shanghai Stock Exchange is the world's 3rd largest stock market by **market capitalization at US$5.5 trillion.** In contrast to Dow 30 industrial average(Dow 30) and the Australian S&P/AX200 , the major index which is the skeleton of the Chinese stock exchange is the SSE, which is the average of the major 50 companies in Chinese market. The dataset consists of Dow 30 ,S&P/AX200 & SSE for the period of roughly 2 years(2016-2018) and the data was downloaded from the yahoo finance website. We will try to predict daily stock movements of US market(Dow 30) based on Australian and Chinese stock markets and the built various linear regression models based on the above data.

Plotting the time series data:

The time series of the Dow 30 industrial average is shown in the figure (1)below, the plot shows the value of Dow jones(y axis) is increasing with the increase in the time (x axis).The smoothed output with window with size of varying 30 to 50 is shown below in the figures 2 & 3 respectively(yf1 vs time)& (yf2 vs time), the smoothing effect is done to analyze & vizualize the trend in the clear pattern in the market of US stock market.

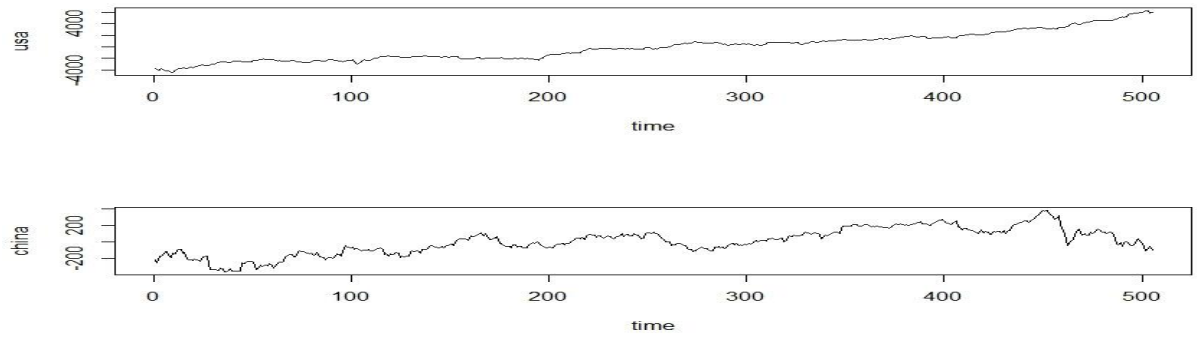






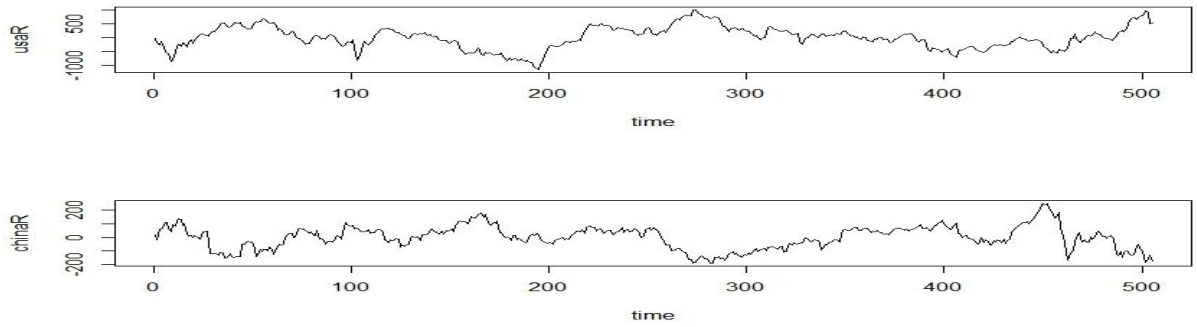
Detrending the time series variable using polynomial of order 2& estimating variance :

Time series variables residuals of USA and Chinese market where plotted and from the plot it could be seen that there was no visible trend observed in the Chinese market, but there was no visible trend observed in the plot,and thus tried to do the regression analysis in which polynomial of degree 3 was fitted ,

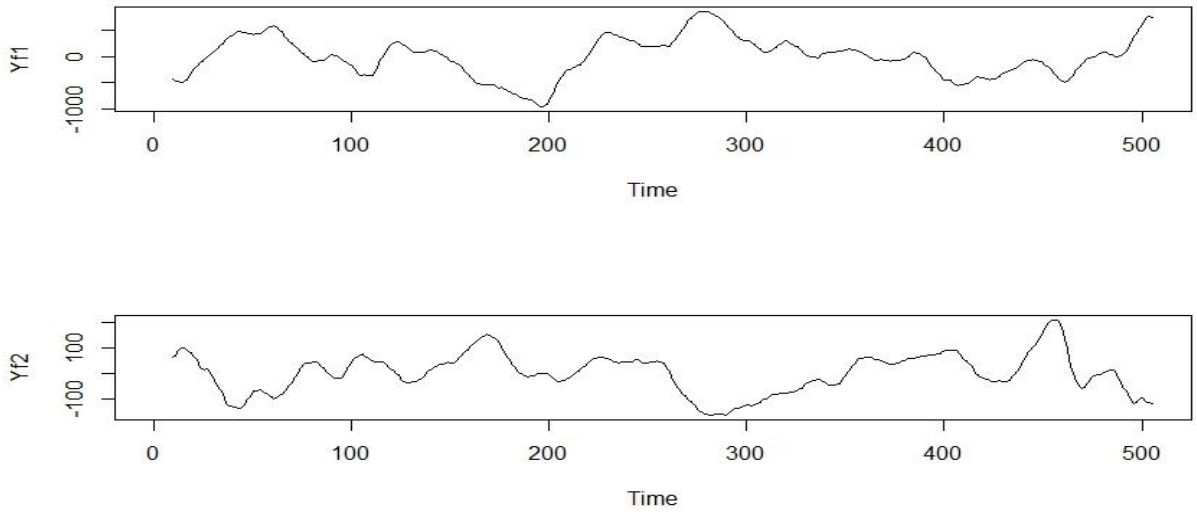



```
> summary(res_usa)

Call:
lm(formula = usa ~ time + T3 + T2)

Residuals:
     Min       1Q   Median       3Q      Max
-1176.08  -251.24    22.85   259.73  1013.72

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.775e+03  7.220e+01 -52.283  < 2e-16 ***
time         1.962e+01  1.235e+00  15.895  < 2e-16 ***
T3           9.017e-05  7.362e-06  12.248  < 2e-16 ***
T2          -4.814e-02  5.666e-03  -8.495 2.27e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 402.6 on 501 degrees of freedom
Multiple R-squared:  0.9724,   Adjusted R-squared:  0.9722
F-statistic:  5875 on 3 and 501 DF,  p-value: < 2.2e-16

> summary(res_china)

Call:
lm(formula = china ~ time + T2 + T3)

Residuals:
     Min       1Q   Median       3Q      Max
-189.910  -51.808    4.821   56.307  251.225

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.384e+02  1.504e+01 -15.849  < 2e-16 ***
time         3.896e-01  2.572e-01   1.515     0.13
T2           5.259e-03  1.181e-03   4.455 1.04e-05 ***
T3          -9.545e-06  1.534e-06  -6.224 1.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

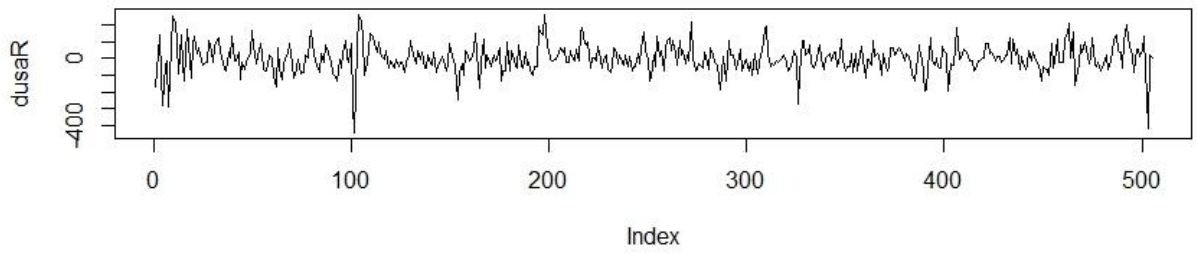
```
Residual standard error: 83.88 on 501 degrees of freedom
Multiple R-squared:  0.7138,   Adjusted R-squared:  0.7121
F-statistic: 416.6 on 3 and 501 DF,  p-value: < 2.2e-16
```

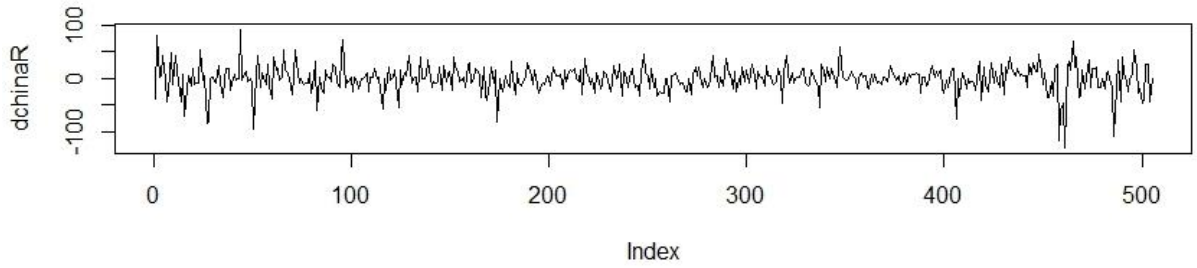


From the plot above ,the usaR and chinaR it can be observed that the residuals of china and usa do not show any pattern and are not similar to time series plot in the figures above. For the window size of size 10,to show the variance is increasing ,decreasing or stationary, the signal with window size of 10 was plotted and from the figure below it can be seen that the variance was stationary as there was no increasing or decreasing pattern observed in the plot.

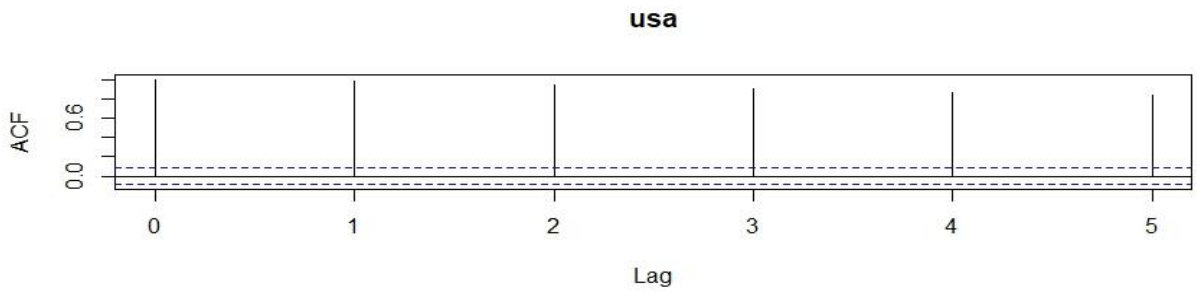


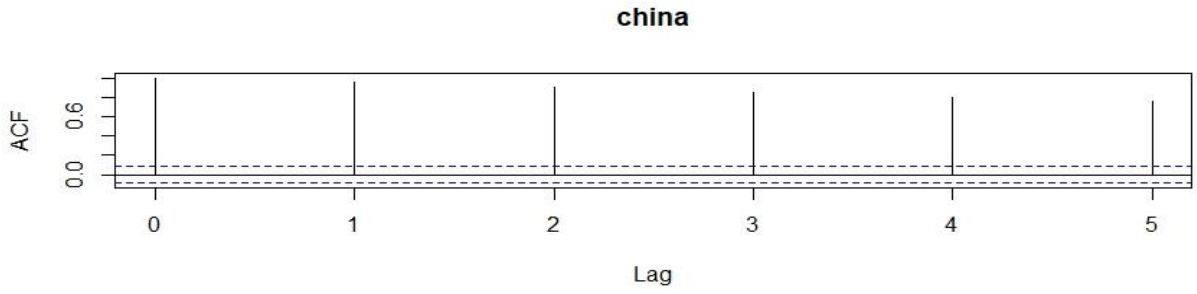
Differenciating the plot one way to make a time series stationary — compute the differences between consecutive observations. This is known as differencing. Differencing can help stabilize the mean of a time series by removing changes in the level of a time series, and so eliminating trend and seasonality. From the plot it can be seen that the variance is stationary as there was no pattern.

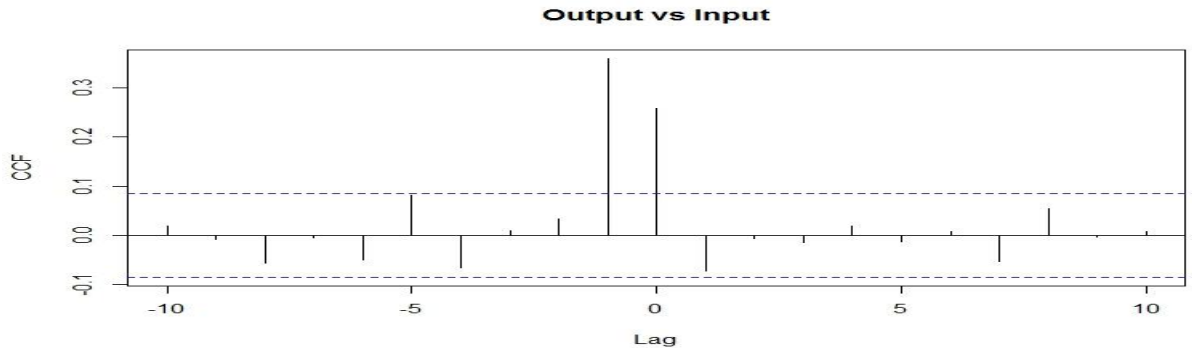




Auto-correlation and cross correlation: Error terms correlated over time are said to be autocorrelated or serially correlated, Cross correlation is a measure of similarity of two series as a function of the displacement of one relative to the other. This is also known as a sliding dot product or sliding inner-product.

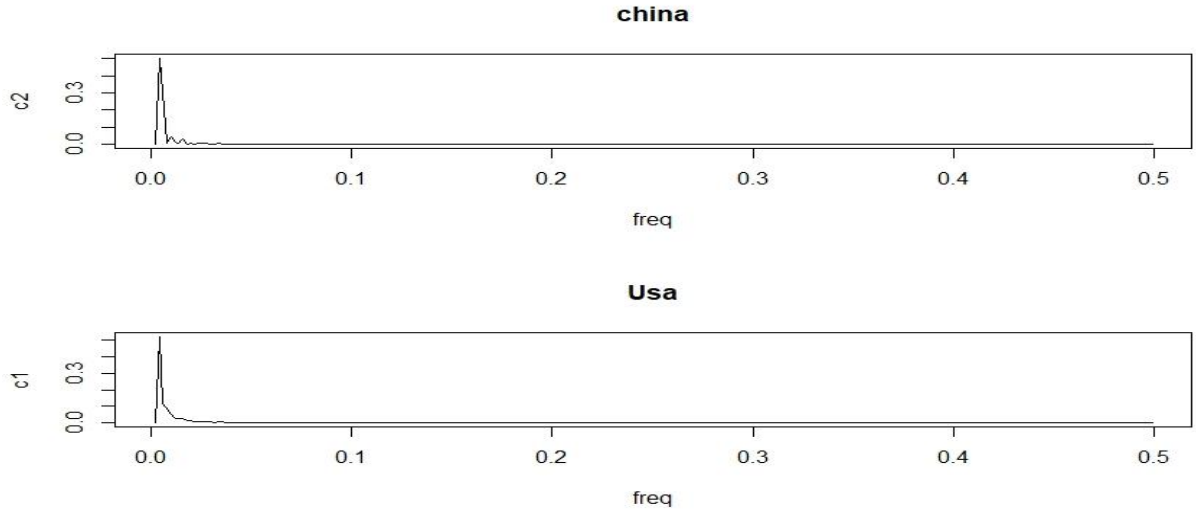




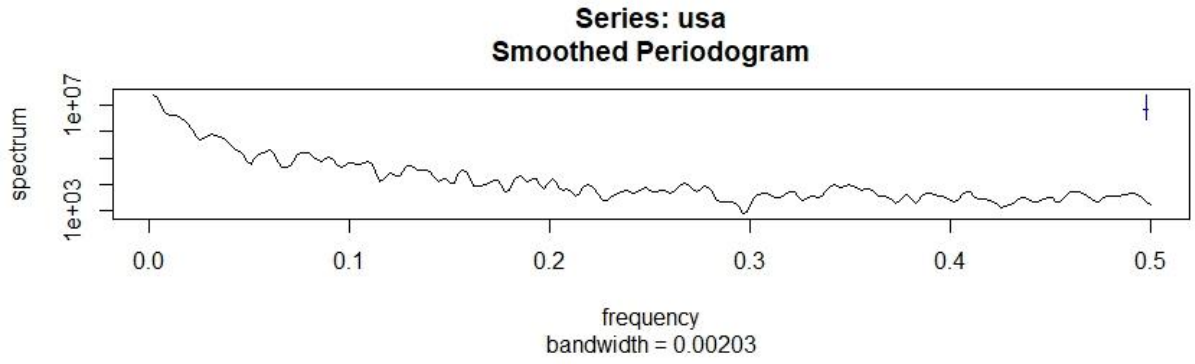


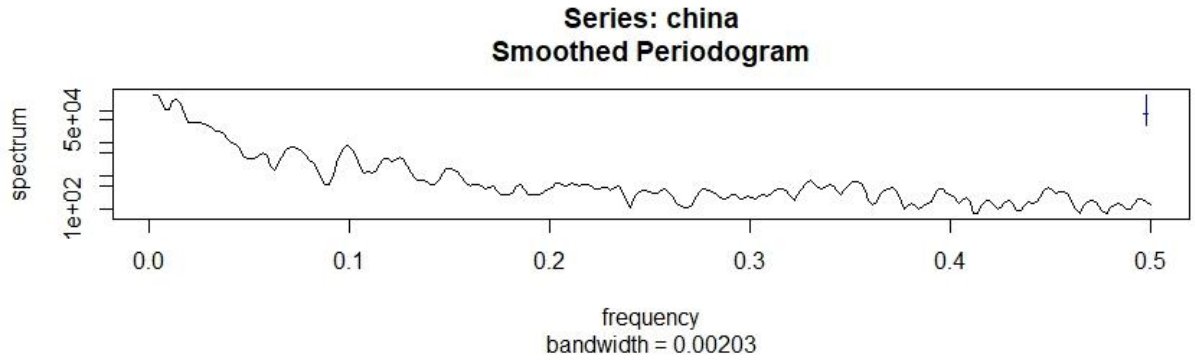
The cross correlation between the Usa and Chinese market is shown in the figure above,it can be seen that The dotted lines in the plot are the significance levels. Negative lags are the responses which are above the significance level. If the lags are negative then the CCF function is given by Cov[Y(t)X(t+k)]. Hence, the Y(t) predicts X(t). In our model, USA(Dow 30) predicts Chinese(SSE) market.

Plotting the periodograms:
Periodogram in time series analysis is used to identify the dominant periods (or frequencies) of a time series. This can be a helpful tool for identifying the dominant cyclical behavior in a series, particularly when the cycles are not related to the commonly encountered monthly or quarterly seasonality.From the plot it can be seen that there were no systematic cycles observed in the periodogram.

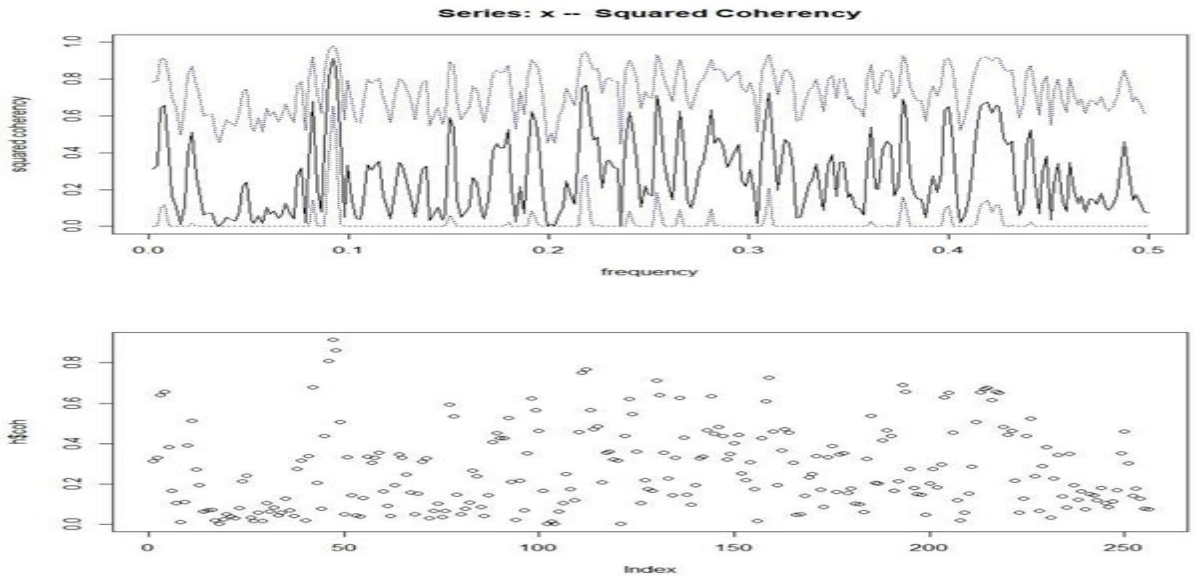


Smoothing of periodograms of window size 2 was done as the time spread is more and more a during the smoothing stock exchange rates across more than 2 weeks underfits the model and the plot shows that there is peak at starting point .Here the frequency bandwidth of 0.00203 is taken.



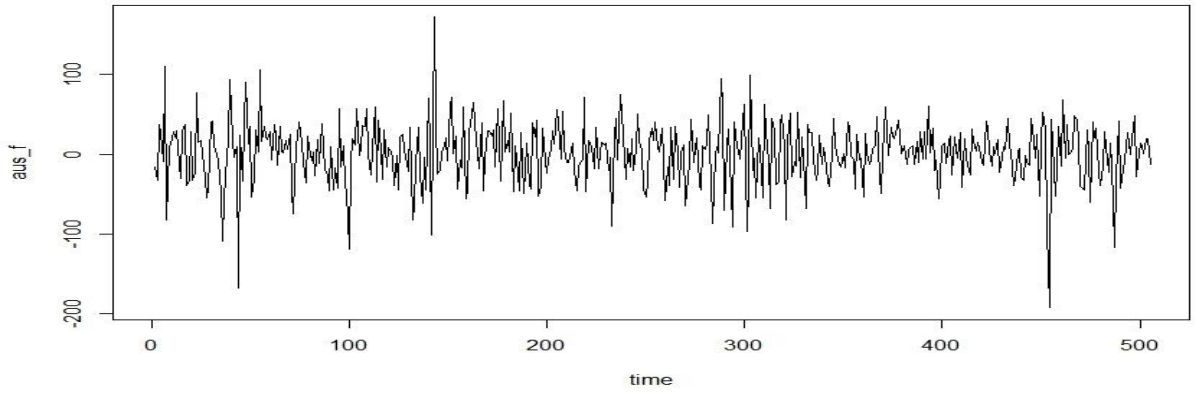


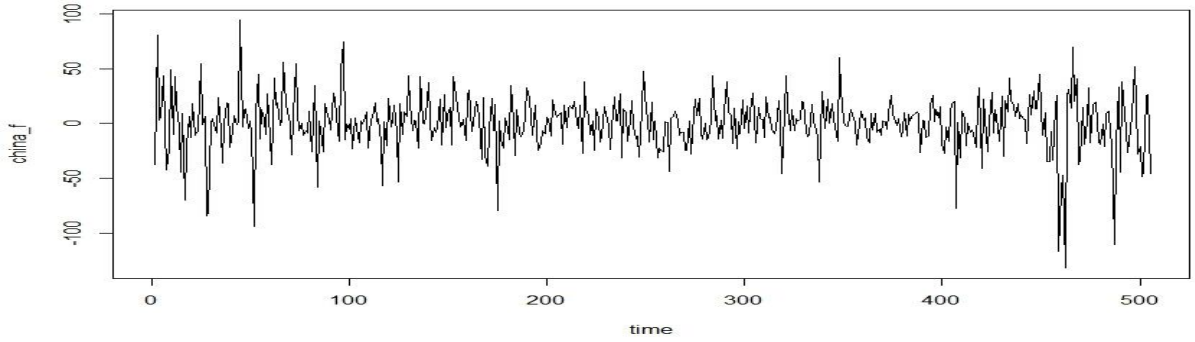
Plotting of coherence in time series: In time series analysis, and particularly in spectral analysis, it is used to describe the strength of association between two series where the possible dependence between the two series is not limited to simultaneous values but may include leading, lagged and smoothed relationships. From the plot, we can see that the signals are both positively and negatively correlated at different frequencies.

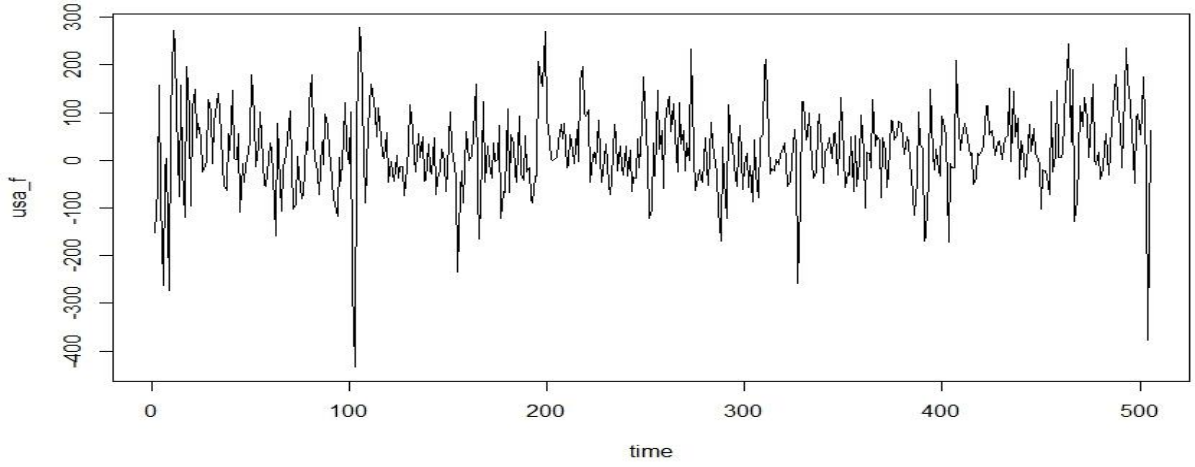


<u>Introducing the 3rd variable(Australian stock exchange)</u>:Until now the analysis was done with only USA and Chinese markets, the US market will be predicted with 2 variables (Chinese(SSE) and Australian(AX 200).

The 3 plots below show below the spectral analysis of Chinese, Australian and American stock exchanges ,there are few spikes between 150 and 450 day, while there are spikes between 0-100 & 400-500 time slots of Chinese markets and there are spikes between 100 & 500 in us stock market.







Comparing regression models and anova & BIC tests:

3 regression models where build a simple and a complex, with the simple being nested within the complex and the a third model that is not nested within the other two models.First being united states ,Australia & china, second being united states & third being united states , Australia and china stock market respectively.

```
> res1 = lm(usa~usa_1+usa_2+aus_1+china_1)
> res2 = lm(usa~usa_1+usa_2)
> res3 = lm(usa~usa_1+aus+china)
```

Anova test is Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. , ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups. ANOVA is useful for comparing (testing) three or more means (groups or variables) for statistical significance. The number of degrees of freedom DF can be partitioned in a similar way: one of these components (that for error) specifies a chi-squared distribution which describes the associated sum of squares, while the same is true for "treatments" if there is no treatment effect. A chi-squared test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

```
> anova(res1,res2,test = "Chisq")
Analysis of Variance Table

Model 1: usa ~ usa_1 + usa_2 + aus_1 + china_1
Model 2: usa ~ usa_1 + usa_2
  Res.Df     RSS Df Sum of Sq Pr(>Chi)
1    498 3281319
2    500 3295910 -2    -14591   0.3305
```

Anova test shows that there is no evidence against the null hypothesis favoring simple,the model 2 i.e simple model is better model 2:usa~usa_1+usa_2.

BIC Test: Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).The BIC test of the output shows that model 2 is better. From both tests it can be assumed that model that is res2 is the best among three.

```
> BIC(res1)
[1] 5882.708
> BIC(res2)
[1] 5872.499
> BIC(res3)
[1] 5919.304

> summary(res2)

Call:
lm(formula = usa ~ usa_1 + usa_2)

Residuals:
    Min      1Q  Median      3Q     Max
-396.65  -44.54   -5.15   44.59  247.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.51845   30.86068   -0.60    0.549
usa_1         1.30003    0.04254   30.56  < 2e-16 ***
usa_2        -0.29843    0.04269   -6.99 8.83e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 81.19 on 500 degrees of freedom
```

```
Multiple R-squared:  0.9989,   Adjusted R-squared:  0.9989
F-statistic: 2.203e+05 on 2 and 500 DF,  p-value: < 2.2e-16
```

Examining the residuals : A residual plot is a graph that shows the residuals on the vertical axis and the ind ependent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around th e horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate. From the examination of residuals in the plot given below shows that the dots are randomly dispersed and linear regression is more suitable for the table.

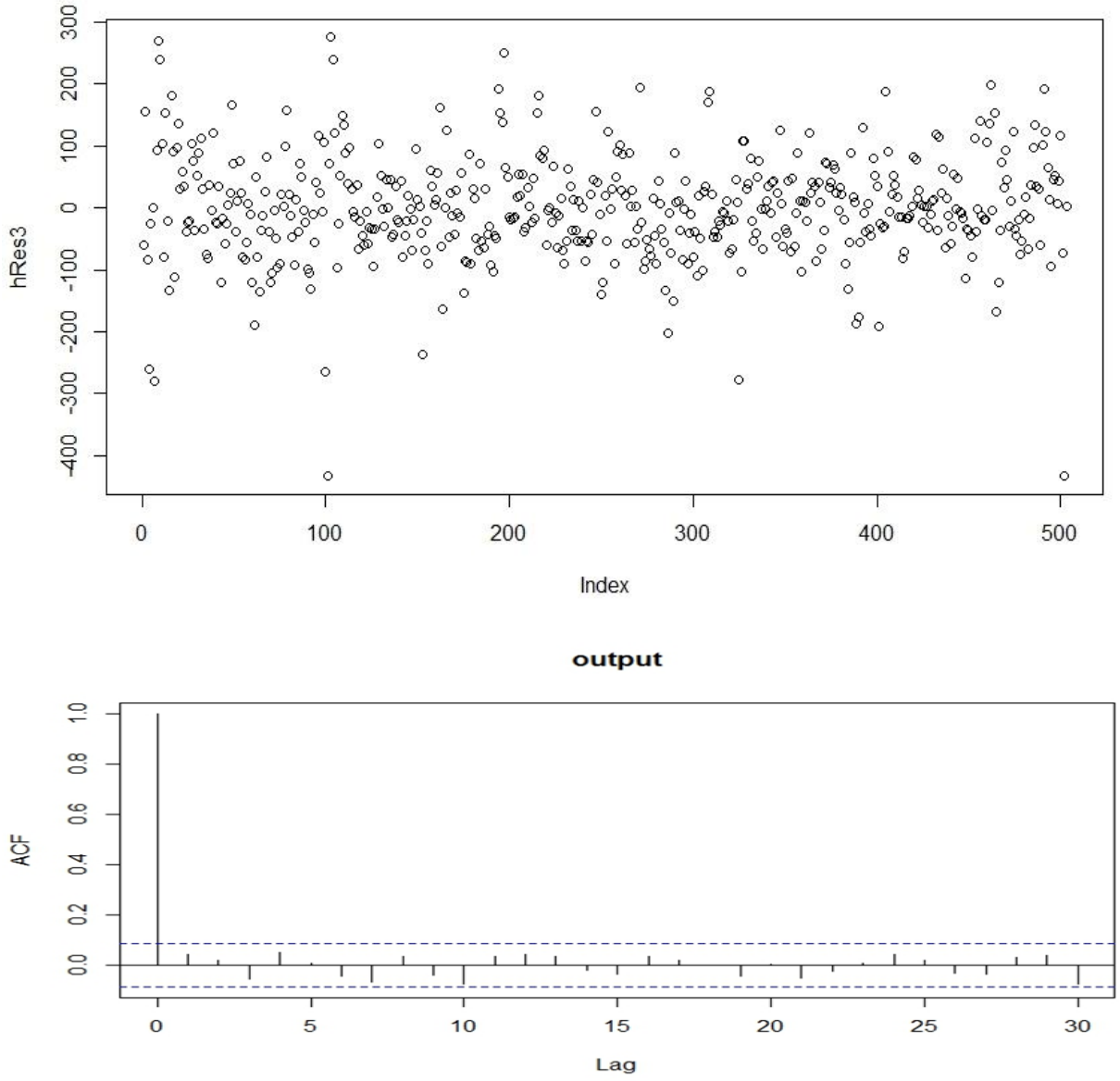
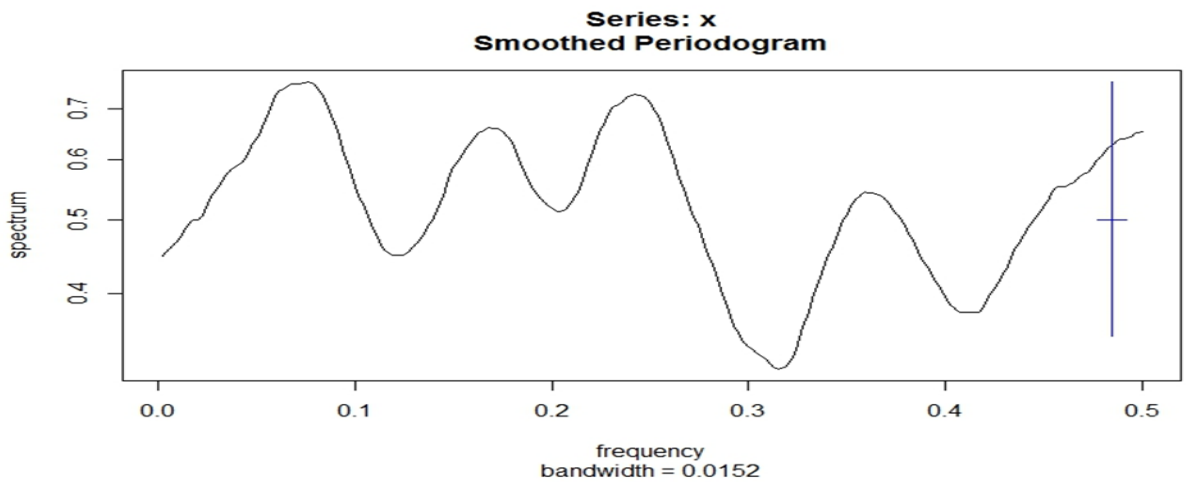


Fig.ACF vs Lag



From the plot above, of ACF vs Lag (figure)it can be seen that none of the lag components have a significa nt autocorrelation values, hence the noise in the model can be classified as white noise. From the plot of Spectrum vs frequency bandwidth i.e the periodogram the spectrum doesn't grow on average with the frequency and hence the noise present is white noise.The residuals are white noise.

Introduction of 3rd variable Australian stock exchange:

Weekly USA stock exchange data was picked with two input variables (Australian and Chinese markets) and an output variable USA stock exchange. Four models were formulated:

1. Lagged output variable, lagged output variable with lag = 2 (usa_1 and usa_2), with two lags of one of the input variables (china_1 and china_2). AR 1 and MA 1.
2. Same features with AR 2 and MA 2.
3. Third model with two lags of output and two lags of both the inputs. AR 2 and MA 2.
4. Two lags of outputs and two lags of the australian stock market inputs. AR 2 and MA 1

ARMAX models are useful when you have dominating disturbances that enter early in the process, such as at the input. For example, a wind gust affecting an aircraft is a dominating disturbance early in the proc ess. The ARMAX model has more flexibility than the ARX model in handling models that contain disturbances.The 4 models are,

```
> X1 = cbind(usa_1,usa_2,china_1,china_2)
> res1 = arima(usa, xreg=X1, order=c(1,0,1))
> res2 = arima(usa, xreg=X1, order=c(2,0,2))
> X3 = cbind(usa_1,usa_2,china_1, china_2, aus_1, aus_2)
> res3 = arima(usa, xreg=X3, order=c(2,0,2))
> X4 = cbind(usa_1, usa_2, aus_1, aus_2)
> res4 = arima(usa, xreg=X4, order=c(2,0,1))
```

AIC and BIC tests:

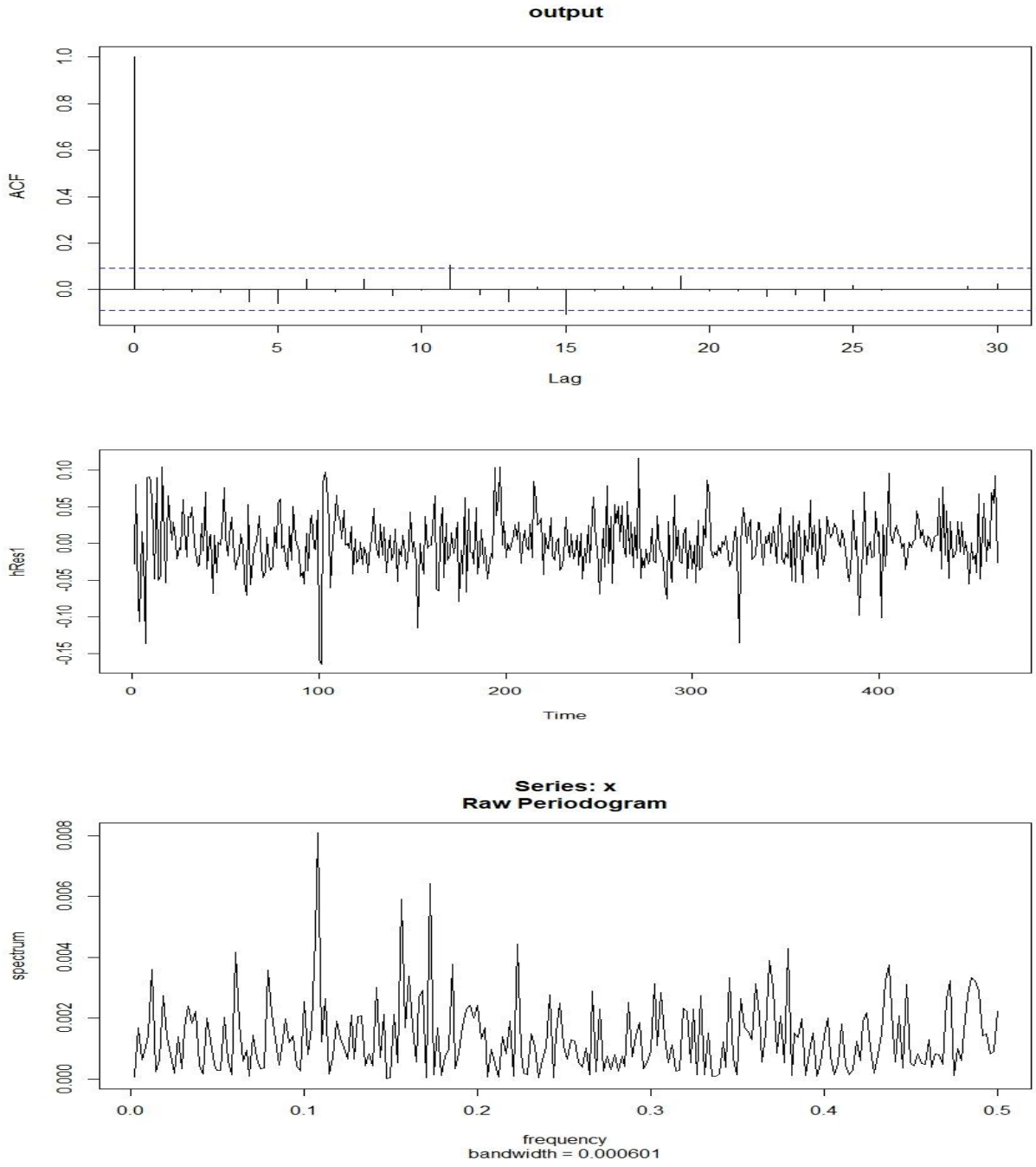
The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a give n set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative t o each of the other models. Thus, AIC provides a means for model selection. AIC is founded on informatio n theory: it offers an estimate of the relative information lost when a given model is used to represent th e process that generated the data. (In doing so, it deals with the trade-off between the goodness of fit of the model and the simplicity of the model.)

Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models; the mo del with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related t o the Akaike information criterion (AIC). BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered

to be more likely to be the true model. Both the criterias are based on various assumptions and asymptotic approximations.

```
> cbind(AIC(res1,res2,res3,res4), BIC(res1,res2,res3,res4))
     df       AIC df       BIC
res1  8 -1674.362  8 -1641.260
res2 10 -1674.762 10 -1633.385
res3 12 -1678.688 12 -1629.035
res4  9 -1678.008  9 -1640.768
```

From the results it can be seen, First model has the least BIC but there is a disagreement with AIC. Since we are focusing only on BIC, one should choose model 1.

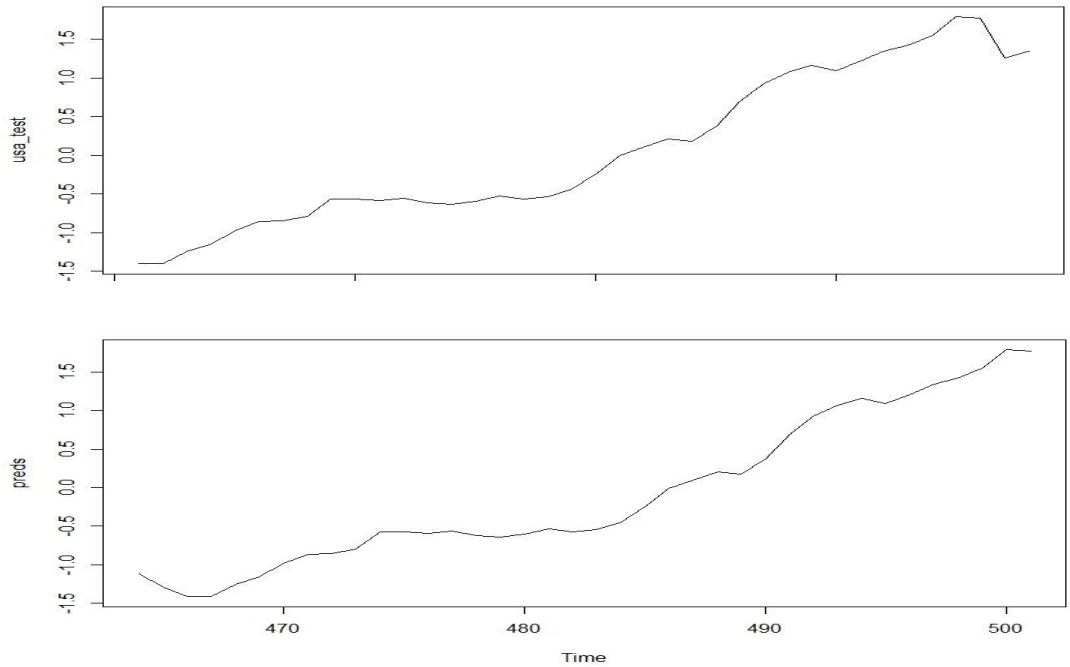


From the autocorrelation plot and the periodogram plot, it can be seen that the signal noise is white noise since no lag components have significant autocorrelation. The hre1 vs time graph is shown spikes near 100 & between 300 & 400 time period. From the above summary, it can be seen that the standard error for the ar and ma components is on the higher side, but majority of the contribution of the model is from these components (coefficients close to 1 and -1), the predictors have relatively low contributions, b ut the standard error is a small quantity. Out of the predictors, usa with a lag of 1 has the most contribution to the model.The output of R code is show below.

```
> res1
Call:
arima(x = usa, order = c(1, 0, 1), xreg = X1)

Coefficients:
          ar1     ma1  intercept   usa_1    usa_2  china_1  china_2
      -0.2678  0.5492     0.0078  1.0914  -0.0899   0.0008  -0.0020
s.e.   0.2529  0.1315     0.0025  0.1494   0.1501   0.0126   0.0128

sigma^2 estimated as 0.00152:  log likelihood = 845.18,  aic = -1674.36

> coef(res1)
     ar1           ma1     intercept         usa_1         usa_2        china
_1       china_2
-0.2677778166  0.5491822295  0.0077862921  1.0914087327 -0.0899155837
0.0008461482 -0.0020182033
```

Plotting the predictions:

Actual vs predicted:The plot below shows actual plot and predicted plot, the actual plot and predicted look quite similar with slopes and coeffients nearly same.



Future work:

More and more market exchanges from different developed counties can be added to the model(such United Kingdom,korea,canada), implementation of machine learning algorithms (Such as neural networks ,deep learning techniques)can also be applied which consists of keras and lstm libraries).Bayesian and Monte carlo simulation can also be applied which can help predict the behavior more deeply.