Time Series Analysis

Final Project Report

Submitted by Vinay Vernekar

Table of contents

Time Series Analysis – STAT-650

Final Project report - Analyzing S&P and Nikkei Index

<u>Data:</u>

The data set consists of data from two indexes. SNP Index associated with United States and Nikkei Index related to Japans economic activity. The data source is Yahoo Finance.

The attributes taken for the analysis are the volume of securities traded on both of these stock indexes. We have taken the daily closing value of securities traded on both these stock exchanges with reference to the same date. Here the Japan market will close before the US market and hence the hypothesis is that it has some effect on the volume traded on US markets. But since the U.S. economy is so big there is also a possibility that SNP index can influence Nikkei.

We have taken daily figures from Jan 2012 till Dec 2015. Since the stock market operates for 5 days we have close to 1001 observations. There are cases when one of the markets may be closed due to Public holidays, external factors etc. So for such mismatches we have taken the previous days values (keeping it constant). The Data looks something like below.

Date	Serial	SNP_volume	Nikkei_volume
1/9/2012	1	3371600000	101400
1/10/2012	2	4221960000	112400
1/11/2012	3	3968120000	106200
1/12/2012	4	4019890000	84800
1/13/2012	5	3692370000	109800
1/17/2012	6	4010490000	85400
1/18/2012	7	4096160000	136200
1/19/2012	8	4465890000	125000
1/20/2012	9	3912620000	177600
1/23/2012	10	3770910000	119800

Date: It's the trading or business day

Serial: Is unique identifier

SNP_volume: Is the closing figure of the volume of securities traded on SNP

Nikkei_volume: Is the closing figure of the volume of securities traded on Nikkei

SNP=Standard and Poor

NIK= Nikkei

AR= Autoregressive

MA= Moving Average

<u>Objective:</u>

The objective is to understand the seasonal properties of these two series and to create a model that can predict the Nikkei volume of securities traded based on the input of SNP Index volume of securities traded or vice versa based on The Granger causality test, the analysis may include the use of ARIMAX model.

<u>The Plotting of the Original time series is as follows:</u>

a) Values plotted for SNP index

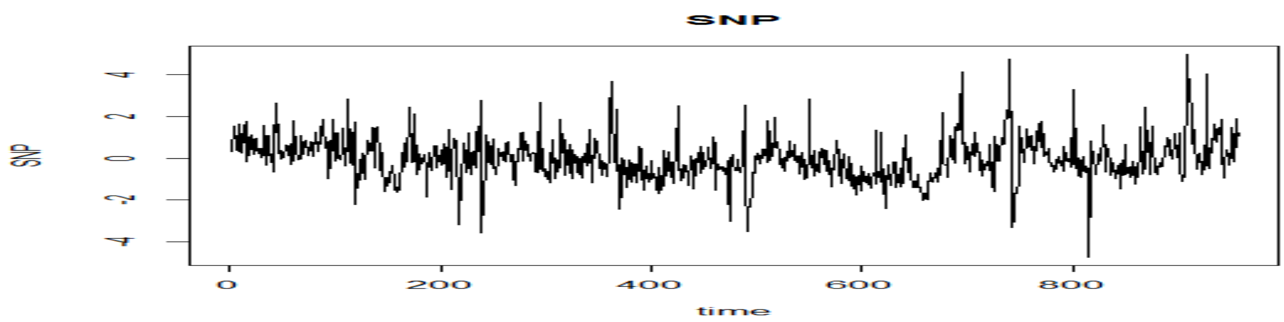

Figure 1

b) Values plotted for Nikkei index

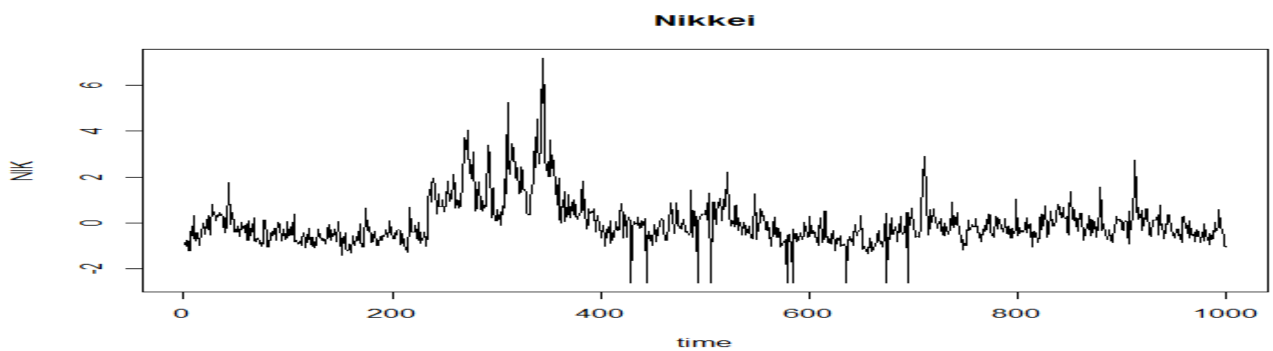

Figure 2

The SNP index and Nikkei index chosen, are composite index consisting for top 500 and 225 companies based on the market capitalization respectively. These index are relatively stable as they consists of companies from different backgrounds. These indexes have strong resistance (maximum) and support (minimum). That means that they will resist strong changes from the long standing average. Hence seeing spikes in ACF and Periodogram will help in better analysis of these series with reference to time.

Plotting the Periodogram for SNP index Volume

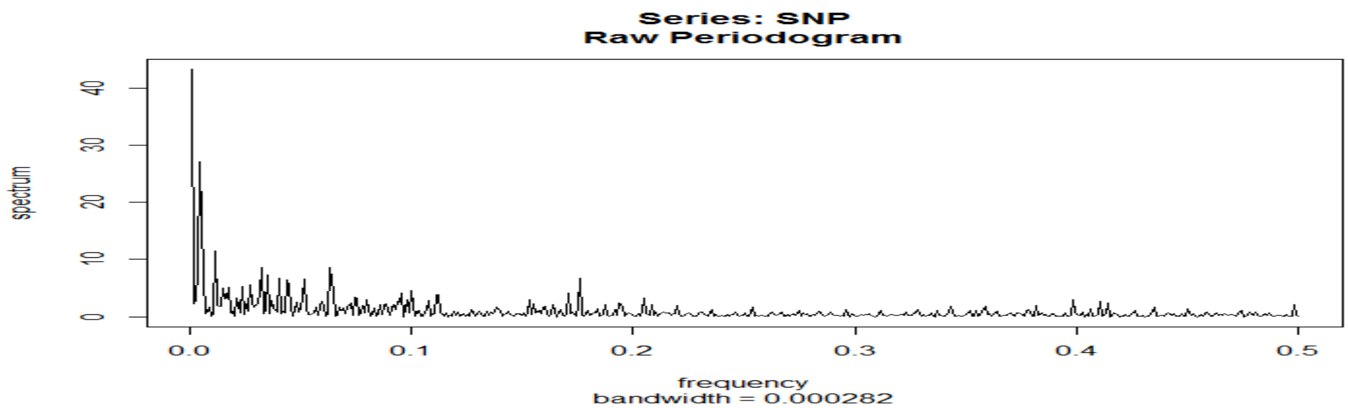

Figure 3

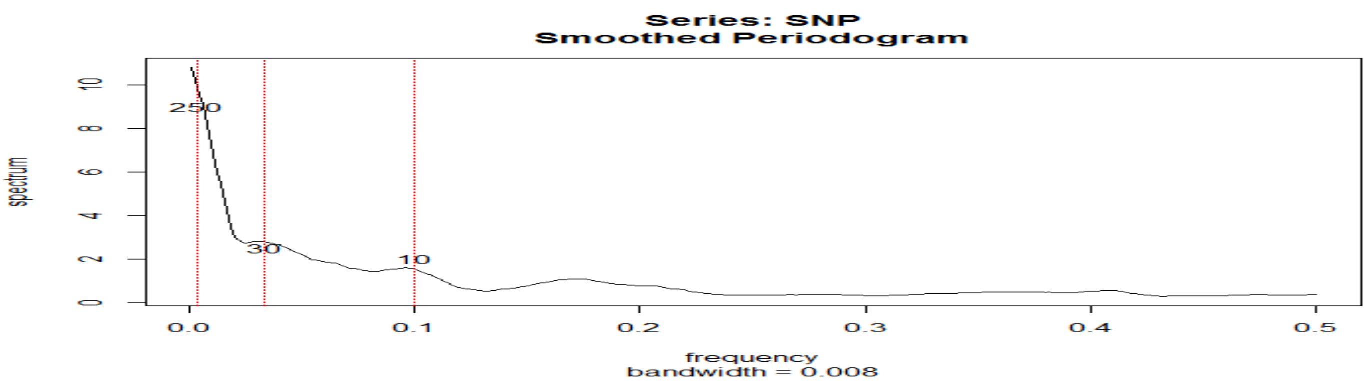

Figure 4

Observations: Looking at the Smoothed periodogram, we can try to estimate the frequency at which we see max fluctuations or spike. By adding lines to the periodogram for spikes (by Trial and error) and overlapping them, we can see at what points the spikes are taking place. The figure helps us see that the highest fluctuations is at the frequency ~ 0.004 corresponding ~ 250 days and this figure represents the annual cycle in a financial calendar. We can infer another peak in the periodogram that happens on the 30th day i.e. Frequency of ~0.0333. The third peak corresponds to a frequency of ~0.1 and this corresponds to the 10th day.

The observations at 30th day and 10th day mostly correspond to monthly and biweekly activities of the stock market. We have monthly review of performance and bi-weekly reporting or reviews for many of the investment firms.

Plotting the Periodogram for Nikkei index Volume:

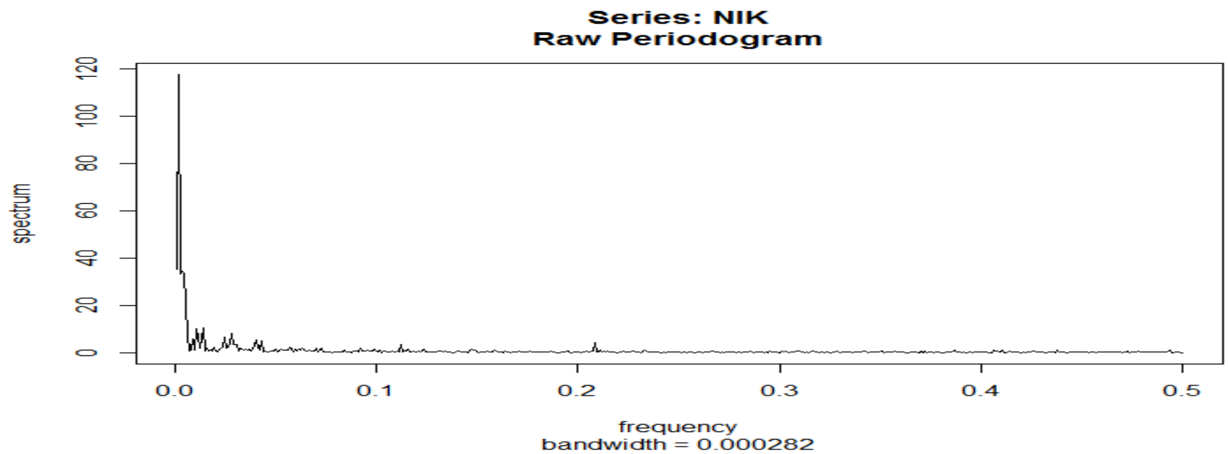

Figure 5

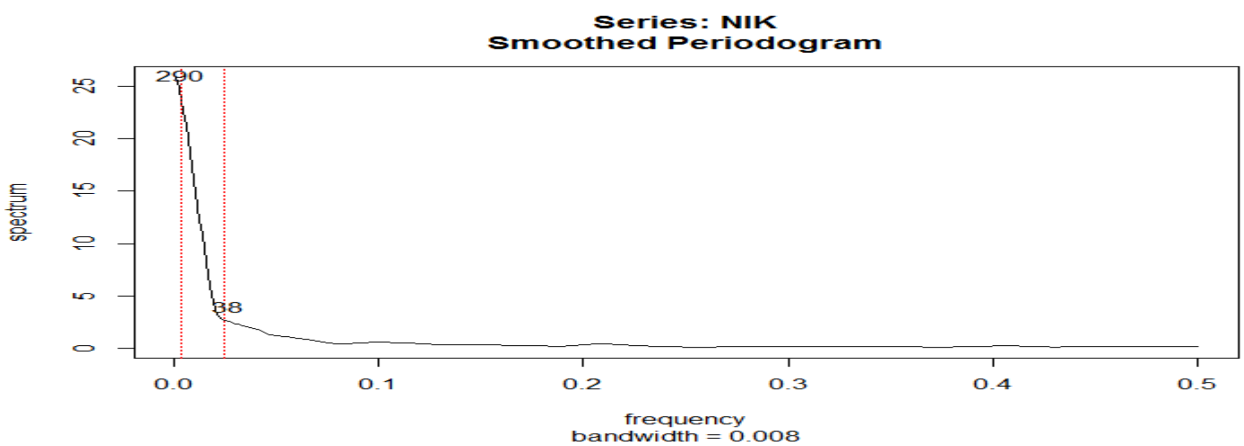

Figure 6

Observations:

The highest fluctuations is at the frequency ~ 0.0035 corresponding ~ 280-290 days and this figure represents the annual cycle and there is another not so significant spike at the frequency ~0.026 corresponding ~ 35-38 days.

Hence, we can conclude that SNP and Nikkei have similar seasonal properties. Like annual spike and close to 30 days spike.

<u>Stationarity</u>:

A stationary time series is one whose properties like mean, variance, autocorrelation, etc. are all same over time.

The predictions for the stationary series can be changed or mathematically engineered by reversing whatever mathematical transformations were previously used, to obtain predictions for the original series.

We know that we can generally make a series stationary by taking different lags of the series.

<u>Plotting the first lag for both SNP and Nikkei:</u>

a) Plot for SNP

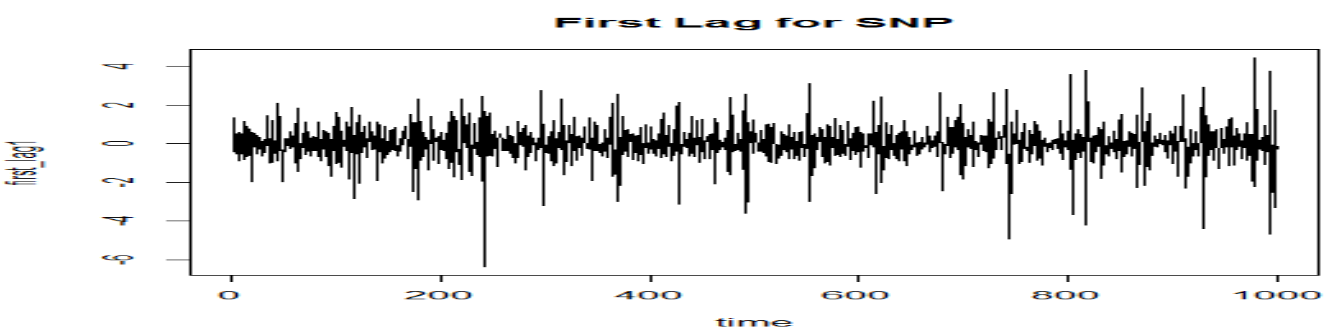

Figure 7

b) Plot for Nikkei

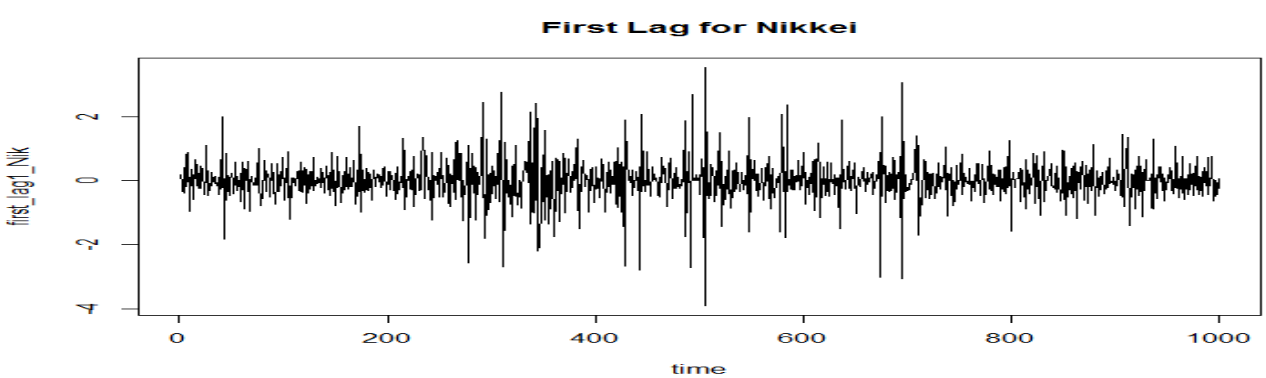

Figure 8

Observation:

The above plots appears weakly stationary.

After making the series stationary we need to look at the ACF and PACF in order to determine the order of the process like how many Autoregressive, Moving Average and Integration are required and also think about the ARIMA/ ARIMAX model.

Looking at the stationary time series ACF and PACF

a) For SNP:

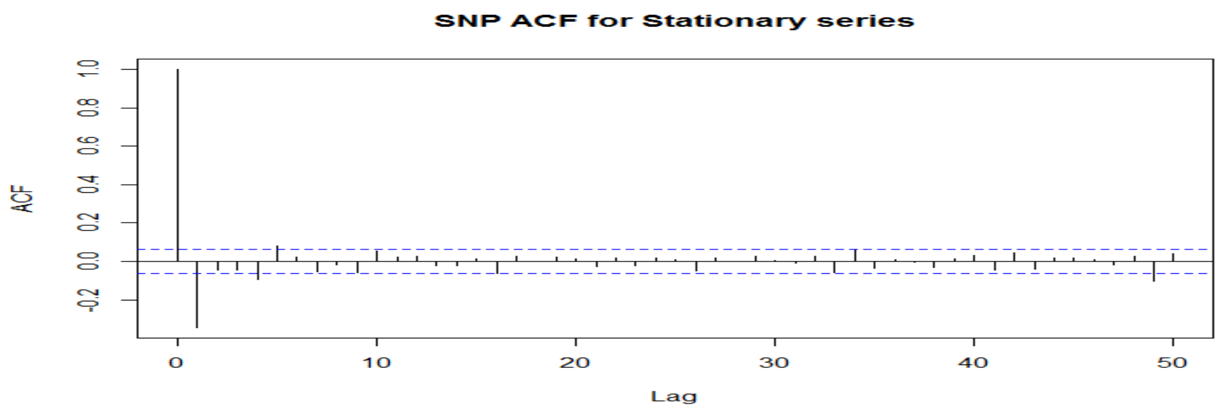

Figure 9

Looking at the Plot we see that the ACF goes negative at lag 1. The series then fluctuates a little but is largely within the threshold limits. The negative correlation at lag 1 points towards the presence of a MA process.

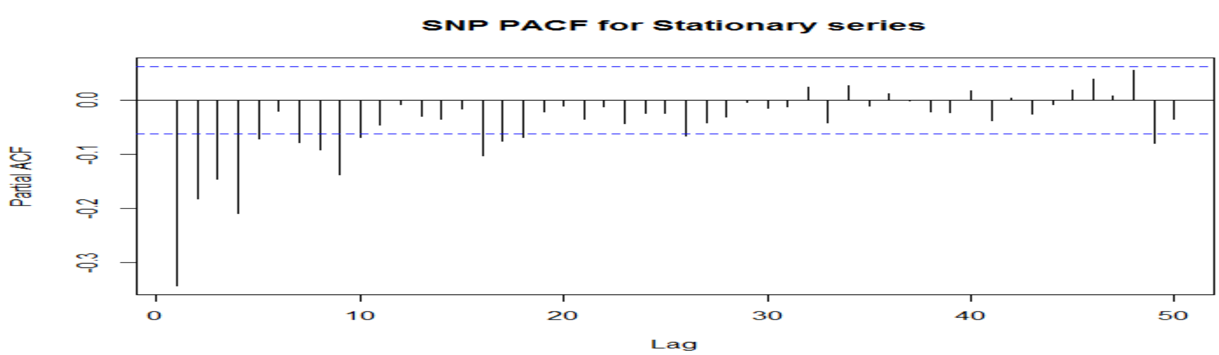

Figure 10

The PACF indicates negative correlations in the beginning and then makes a few transitions in the positive domain and back to the negative domain.

b) For Nikkei

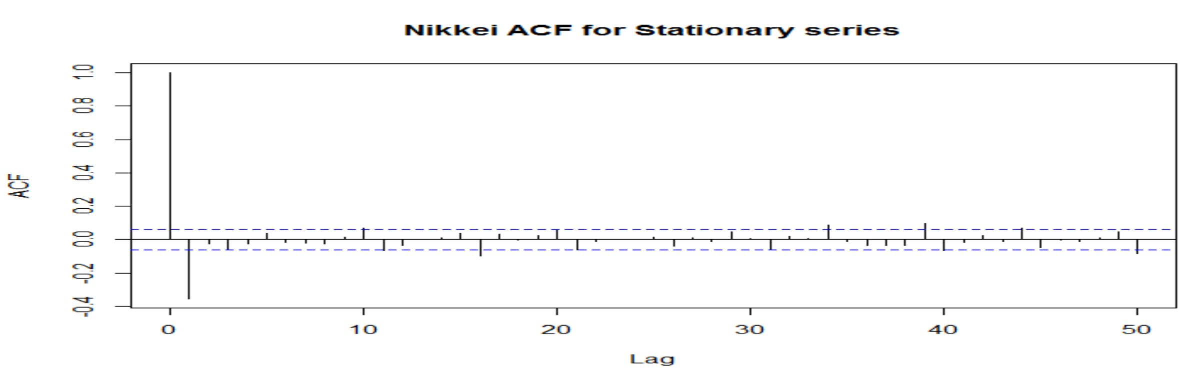

Figure 11

The First lag is negative and the consecutive lags slowly die down. But we do see correlation at a few future lags to be significant.

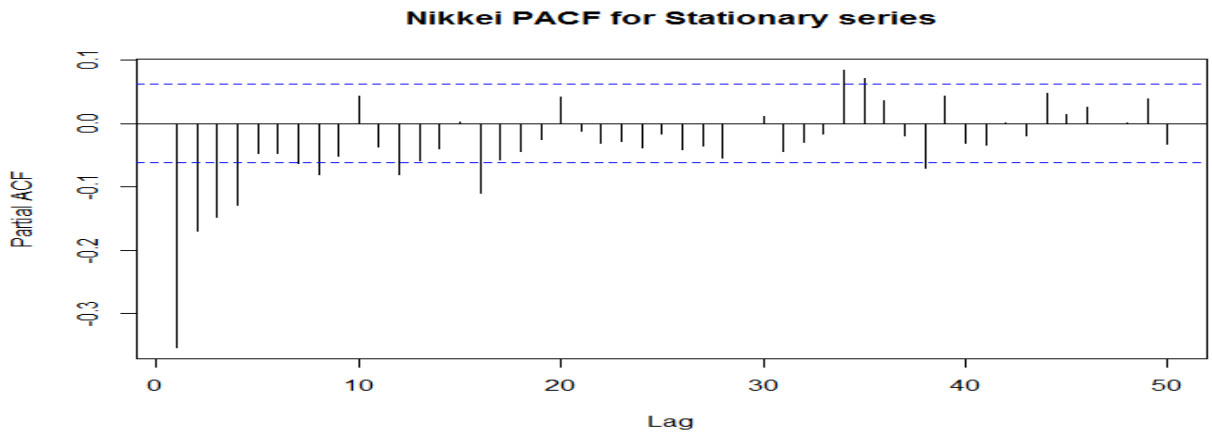

Figure 12

The PACF is significant for first few lags and then slowly dies down, however there are a few spikes later. This leads to the conclusion that we have an MA process (close to second order).

We also know that both the series have similar seasonality or high spike frequency.

Examining Cross Correlation between the two series

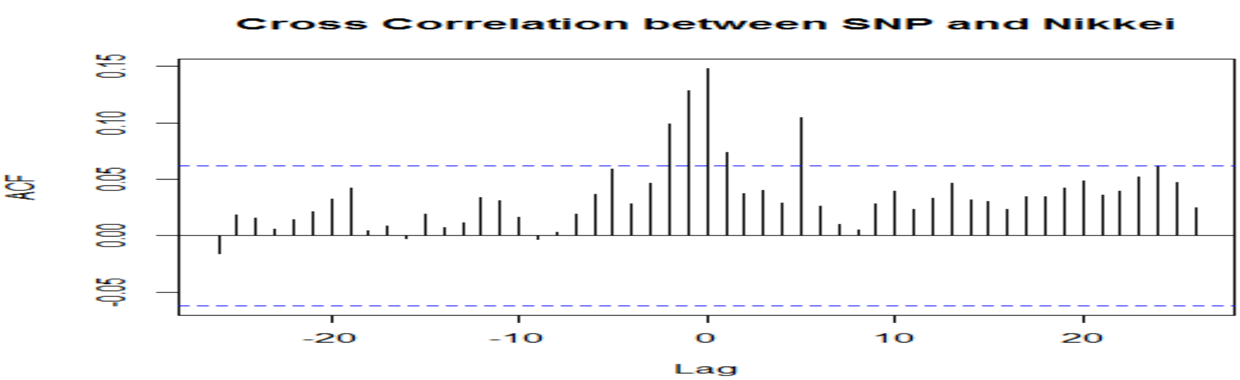

Figure 13

We can assay the above graph indicating that an above average value of SNP volume is likely to lead to an above average value of Nikkei (+ ve correlation). This is especially true for past lags. We do see that the future lag of 1 and 5 also have significant positive correlation.

A better way is to visualize this is through the scatter graph plotted below which gives the correlation between the variables for different time lags.

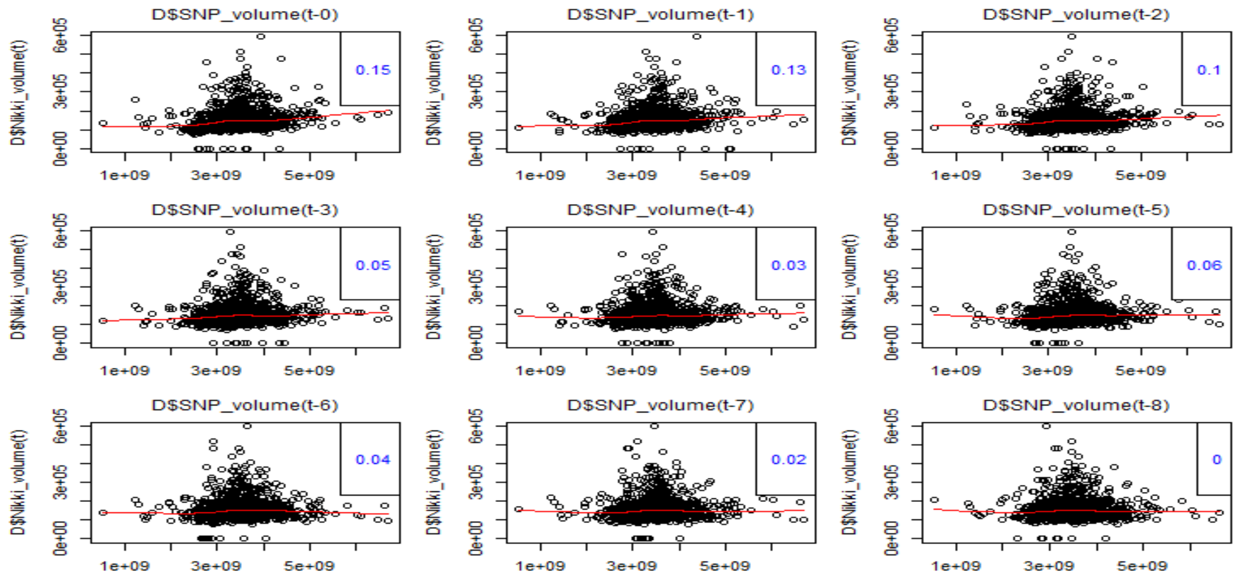

Figure 14

Observations:

All Lags have mostly positive correlations, indicating that both series have correlation and that Japans Nikkei's performance has impact on SNP or vice versa. Apart from the first (t-1) and the second (t-2) lag the fifth (t-5) has the third highest correlation value.

Granger causality test:

We need to check if SNP can be used to predict Nikkei or is the relation reverse. For this we used the Granger causality test and found that at Lag 3 we can say that SNP Granger-Cause Nikkei. This was tried for different lags for both the direction i.e. SNP causing Nikkei or Nikkei causing SNP, but the causality was produced at lag 3 for the assumption "SNP Granger-Causing Nikkei"

```
> grangertest(NIK ~ SNP, order = 3, data = mainData)
Granger causality test

Model 1: NIK ~ Lags(NIK, 1:3) + Lags(SNP, 1:3)
Model 2: NIK ~ Lags(NIK, 1:3)
  Res.Df Df     F    Pr(>F)
1    991
2    994 -3 5.49 0.0009645 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Figure 15

Model selection and Forecasting:

For forecasting we know that the process followed by time series is an AR process and more orders MA process, we see that there is a max seasonality of for ~ 250 days which corresponds to a year's cycle, as financial calendars has close to 250 working days and is present in both the series i.e. SNP and Nikkei.

We have also see in Cross Correlation plot that there is correlation between SNP and Nikkei and this correlation is high for the first Lag in the past and 5th lag in the future and we also have the inference from Granger causality test that SNP Granger-Cause Nikkei, hence maybe we can use an ARIMAX model to replicate or predict the behavior of SNP index volume on Nikkei.

After trying different orders and inputs for the model, the model which provided the best fit was selected. In this study, the ARIMAX model with order (1,1,2) gave the lowest AIC and BIC. Hence this model was selected.

Table with AIC and BIC for all Trial Models comparison

Order-Model	Model No.	Input	AIC	BIC	SigmaSq
1,1,2	Model 1	SNP_3	1712.089	1736.4	0.3471
1,1,2	Model 2	SNP_2	1712.84	1737.15	0.3473
1,1,2	Model 3	SNP_1	1706.353	1730.67	0.345
1,0,1	Model 4	SNP_2	1725.919	1750.24	0.3513
1,0,1	Model 5	SNP_1	1717.34	1741.66	0.3482
2,1,2	Model 6	SNP_1	1708.322	1737.5	0.345
2,1,2	Model 7	SNP_2	1714.269	1743.45	0.3471

Note: SNP_1= Lag 1(SNP), SNP_2= Lag 2(SNP), SNP_3=Lag 3(SNP) and Input= exogenous input

Observation: Model with order (1,1,2) highlighted in green i.e. 1 Autoregressive, 1 Integration and 2 Moving average terms with Lag 1 of SNP as exogenous input gives the lowest AIC and BIC values and the variance is also less. Hence this model is selected as the final model.

<u>The Coefficients for the Final model:</u>

Features	Co-eff	Std Error	P-Value
AR1	0.78734	0.055974	< 2.2e-16
MA1	-1.30487	0.071521	< 2.2e-16
MA2	0.34147	0.0602	1.41E-08
SNP_1	0.07308	0.023658	0.002007

Inference:

The Z values for each Coefficient is calculated by dividing co-efficient by its respective Standard Error. The P-value for each coefficient is significant when compared to a 95% level of confidence.

If we fit 1 AR coefficient and 2 MA coefficients with 1 Integration we can see that for the given data all the coefficients are significant. Hence this model does add value to the explanation of the movement in the series.

Diagnostics Plots for the final model:

Plot showing the diagnosis of the residuals of ARIMAX model of the order (1,1,2) and using lag 1 of SNP Volume as an exogenous input. Below we can see the model residual plot, QQ-norm plot and AFC for the residual of the final model.

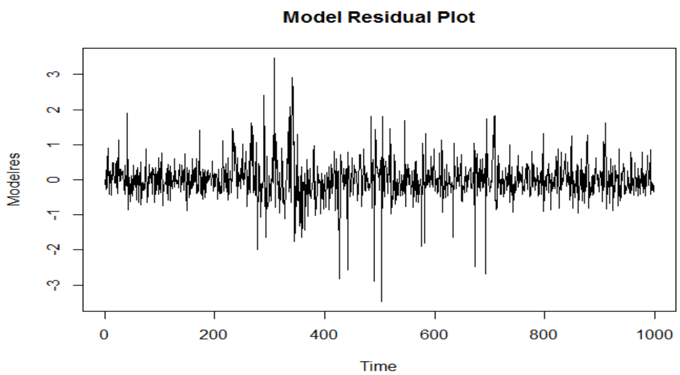

Figure 16

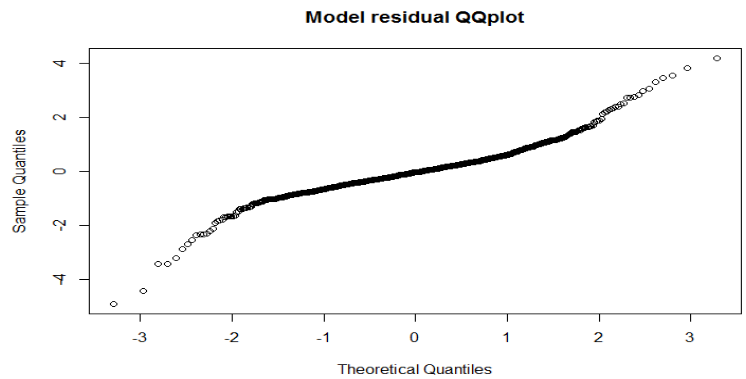

Figure 17

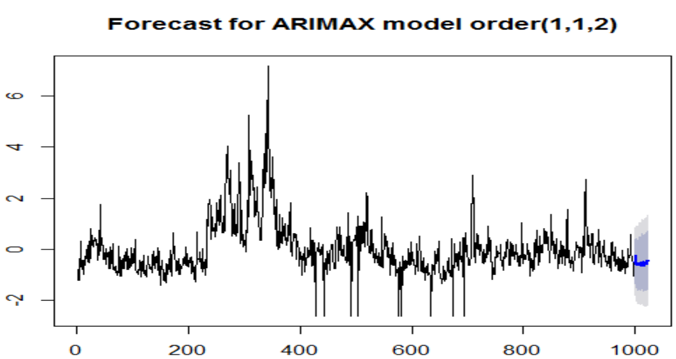

Figure 18

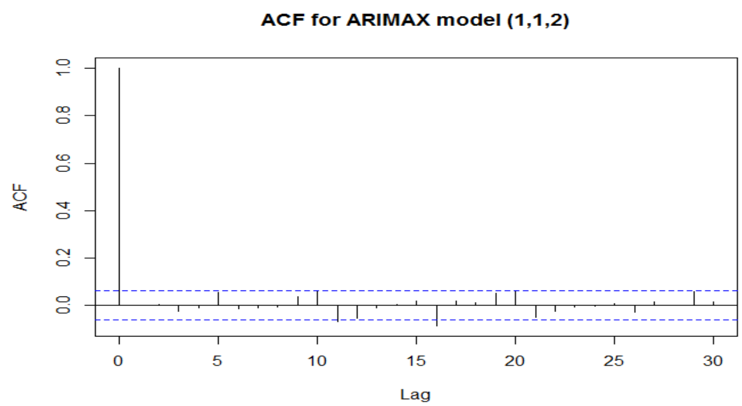

Figure 19

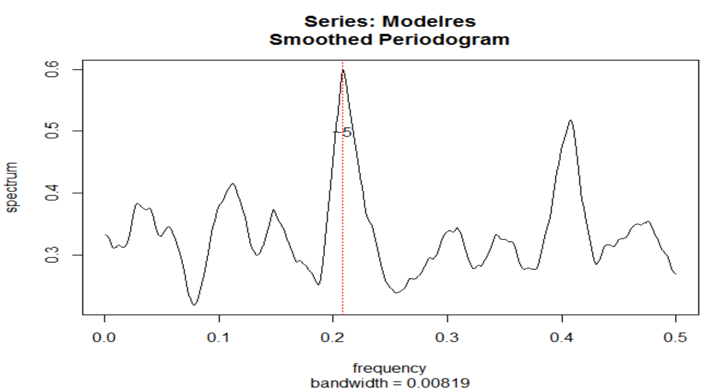

Figure 20

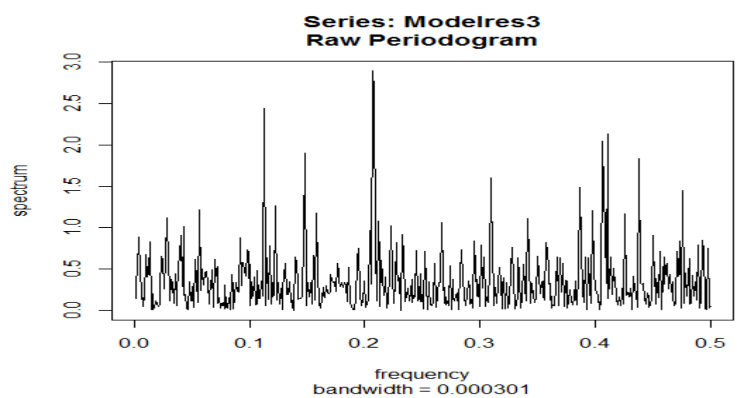

Figure 21

Observations:

The Residual plot Figure 16, looks somewhat stationary having a fluctuations close to the mean and the QQ norm plot of the residual appears almost straight diagonal line indicating Normality of distribution.

The figure 19, ACF value for the 15th Lag appears to be high however, the Box- Pierce test supports the Null Hypothesis stating that it's not significant i.e. p-value = 0.1055. The p-values for the Box-Pierce all are well above .05, indicating "non-significance."

The Forecasted value in figure 18, is well within the confidence level of 95%.

The periodogram of the residual is mostly wavy with one high peak, corresponding to frequency of 0.2040 and this translates to roughly 5 days. This 5 days can be interpreted as the weekly correction that takes place due to the long weekend (Discontinuous operations) as many things may have occurred after market hours in weekend and thus having an impact in the time series.

Conclusion:

Hence we can say that we can predict Nikkei Volume based on SNP volume.

Additional Notes:

The seasonal model was tried but since the annual cycle was of 250, the code did not give an output

The original code is provided below this page in the code section and also contains the code for plotting ACF and PACF for original series of SNP and Nikkei

Code:

Code for Final Project report: You can search for the code by searching Figure "Number". This will be easier to look for the code for a specific figure. J = Lag 1(SNP), L= Lag 2(SNP), M = Lag 3(SNP)

#Mandatory Code

```
library(tseries)

library(forecast)

library(lmtest)

library(astsa)

setwd("C:\\Second Sem\\Time Series Analysis\\Data")

D=read.csv("time_series_final.csv")

x=c(7,13,15)

mainData=D[,x]

head(mainData)

SNP = (mainData$SNP_volume - mean(mainData$SNP_volume))/sd(mainData$SNP_volume)

NIK = (mainData$Nikki_volume - mean(mainData$Nikki_volume))/sd(mainData$Nikki_volume)
```

 # Plotting ACF and PACF for original series

```
acf(SNP,50,main="SNP ACF Original data")

pacf(SNP,50,main="SNP PACF Original data ")

acf(NIK,50,main="Nikkei ACF Original data ")
```

```
pacf(NIK,50,main="Nikkei PACF Original data ")
```

 # Code For Figure 1 and Figure 2

```
###Plotting original- normalized data#####
# Code for Figure 1
plot(SNP,type="l",xlab="time", main="SNP")
# Code for Figure 2
plot(NIK,type="l",xlab="time", main="Nikkei")
```

 # Code For Figure 3, Figure 4, Figure 5 and Figure 6

```
# plotting periodogram
# Code for Figure 3
hrper = spec.pgram(SNP,taper=0,log="no")
# Code for Figure 4
spec.pgram(SNP,spans=c(20,20),taper=0,log="no")
abline(v=1/250, lty="dotted",col="red")
abline(v=1/30, lty = "dotted",col="red")
abline(v=1/10, lty = "dotted", col="red")
text(1/250,9, "250")
text(1/30,2.5, "30")
text(1/10,2, "10")
# Code for Figure 5
hrper = spec.pgram(NIK,taper=0,log="no")
# Code for Figure 6
spec.pgram(NIK,spans=c(20,20),taper=0,log="no")
abline(v=1/280, lty="dotted",col="red")
abline(v=1/40, lty = "dotted",col="red")
text(1/290,26, "290")
text(1/38,4, "38")
```

#Stationary Series Plot Figure 7 and Figure 8

```
# taking first lag and plotting them
# Code for Figure 7
```

```
first_lag1=diff(SNP,1)

plot(first_lag1,type="l",xlab="time", main="First Lag for SNP")
```

 # Code for Figure 8

```
first_lag1_Nik=diff(NIK,1)

plot (first_lag1_Nik,type="l",xlab="time", main="First Lag for Nikkei")
```

 # Code For Figure 9, Figure 10, Figure 11, Figure 12 ,Figure 13 and Figure 14

 # Code for Figure 9

```
acf(first_lag1,50,main="SNP ACF for Stationary series")
```

Code for Figure 10

```
pacf(first_lag1,50,main="SNP PACF for Stationary series")
```

 # Code for Figure 11

```
acf(first_lag1_Nik,50,main="Nikkei ACF for Stationary series")
```

 # Code for Figure 12

```
pacf(first_lag1_Nik,50,main="Nikkei PACF for Stationary series")
```

 # Code for Figure 13

```
ccf(SNP,NIK, main="Cross Correlation between SNP and Nikkei")
```

 # Code for Figure 14

```
lag2.plot(SNP,NIK,10)
```

#Granger causality test Figure 15

```
# NIK causing SNP

grangertest(SNP ~ NIK, order = 1, data = mainData)

grangertest(SNP ~ NIK, order = 2, data = mainData)

grangertest(SNP ~ NIK, order = 3, data = mainData)

# SNP causing NIK

grangertest(NIK ~ SNP, order = 1, data = mainData)

grangertest(NIK ~ SNP, order = 2, data = mainData)
```

 # Code for Figure 15

```
grangertest(NIK ~ SNP, order = 3, data = mainData)
```

 # ARIMAX Modelling Mandatory Code

```
B = c(0,1)

SNP_1 = filter(SNP,B,sides=1)

NIK_1 = filter(NIK,B,sides=1)

B = c(0,0,1)

SNP_2 = filter(SNP,B,sides=1)

NIK_2 = filter(NIK,B,sides=1)

B = c(0,0,0,1)

SNP_3 = filter(SNP,B,sides=1)

NIK_3= filter(NIK,B,sides=1)

N = 960

SNP = SNP[4:N]

SNP_1 = SNP_1[4:N]

NIK_1 = NIK_1[4:N]

SNP_2 = SNP_2[4:N]

SNP_3 = SNP_3[4:N]

NIK_2 = NIK_2[4:N]

J=cbind(SNP_1)

L=cbind(SNP_2)

M=cbind(SNP_3)
```

#Code for the main ARIMAX Model with order (1,1,2) and SNP lag 1 as input

```
Model3 = arima(NIK[4:N], xreg=J, order=c(1,1,2)) # Here xreg can be changed to "L" or "M" as well

Model3

coef(Model3)

AIC(Model3)

BIC(Model3)

Modelres3= residuals(Model3)

Modelres3
```

 # Code for Figure 16

```
plot(Modelres3, main="Model Residual Plot")
```

 # Code for Figure 17

```
qqnorm(Modelres3, main = "Model residual QQplot")
```

 # Code for Figure 18

```
pv<-forecast(Model3,h = 25,xreg=J[c(426:450),])

plot(pv, main="Forecast for ARIMAX model order(1,1,2)")
```

 # Code for Figure 19

```
acf(Modelres3, 30, main="ACF for ARIMAX model (1,1,2)")

Box.test (Modelres3, lag = 16)
```

 # Code to give significance of the coefficient model

```
coeftest(Model3)
```

 # Code for Figure 21

```
hrper= spec.pgram(Modelres3,taper=0,log="no")
```

 # Code for Figure 20

```
spec.pgram(Modelres3,spans=c(20,20),taper=0,log="no")

abline(v=1/4.8, lty="dotted",col="red")

text(1/4.8,0.5, "~5")
```

#Code for all other models used to achieve the best model

 # Model 1 with order=(1,1,2) and with regressor as M

```
Model1 = arima(NIK[4:N], xreg=M, order=c(1,1,2)) # Here xreg can be changed to "L" or "M" as well

Model1

coef(Model1) # This part givs the Coefficient of the best model

AIC(Model1)

BIC(Model1)

Modelres= residuals(Model1)

Modelres

plot(Modelres, main="Model Residual Plot") # Code for figure 16

qqnorm(Modelres, main = "Model residual QQplot") # Code for figure 17

coeftest(Model1)

hrper= spec.pgram(Modelres,taper=0,log="no")

spec.pgram(Modelres,spans=c(20,20),taper=0,log="no")

abline(v=1/4.8, lty="dotted",col="red")

text(1/4.8,0.5, "~5")
```

```
plot(Modelres,main = "Model residual Plot")
acf(Modelres, 30, main="ACF for ARIMAX model (1,1,2)")
Box.test (Modelres, lag = 16)
pv<-forecast(Model1,h = 25,xreg=M[c(426:450),])
plot(pv, main="Forecast for ARIMAX model order(1,1,2)")
```

 # Model 2 with order=(1,1,2) and with regressor as L

```
Model2 = arima(NIK[4:N], xreg=L, order=c(1,1,2)) # Here xreg can be changed to "L" or "M" as well
Model2
coef(Model2)
AIC(Model2)
BIC(Model2)
Modelres2= residuals(Model2)
Modelres2
plot(Modelres2, main="Model Residual Plot")
coeftest(Model2)
hrper= spec.pgram(Modelres2,taper=0,log="no")
spec.pgram(Modelres2,spans=c(20,20),taper=0,log="no")
abline(v=1/4.8, lty="dotted",col="red")
text(1/4.8,0.5, "~5")

plot(Modelres2,main = "Model residual Plot")
qqnorm(Modelres2, main = "Model residual QQplot")
acf(Modelres2, 30, main="ACF for ARIMAX model (1,1,2)")
Box.test (Modelres2, lag = 16)
pv<-forecast(Model2,h = 25,xreg=L[c(426:450),])
plot(pv, main="Forecast for ARIMAX model order(1,1,2)")
```

 # Model 4 with order=(1,0,1) and with regressor as L

```
Model4 = arima(NIK[4:N], xreg=L, order=c(1,0,1)) # Here xreg can be changed to J as well
```

```
Model4
coef(Model4)
AIC(Model4)
BIC(Model4)
Modelres4= residuals(Model4)
Modelres4
coeftest(Model4)
plot(Modelres4)
qqnorm(Modelres4)
acf(Modelres4, 30, main="output")
Box.test (Modelres4, lag = 16)
pv<-forecast(Model4,h = 25,xreg=L[c(426:450),])
plot(pv)
accuracy(Model4)
```

 # Model 5 with order=(1,0,1) and with regressor as J

```
Model5 = arima(NIK[4:N], xreg=J, order=c(1,0,1)) # Here xreg can be changed to L as well
Model5
coef(Model5)
AIC(Model5)
BIC(Model5)
Modelres5= residuals(Model5)
Modelres5
coeftest(Model5)
plot(Modelres5)
qqnorm(Modelres5)
acf(Modelres5, 30, main="output")
Box.test (Modelres5, lag = 16)
pv<-forecast(Model5,h = 25,xreg=J[c(426:450),])
plot(pv)
accuracy(Model5)
```

 # Model 6 with order (2,1,2) and with regressor as J

```
Model6 = arima(NIK[4:N], xreg=J, order=c(2,1,2))# Here xreg can be changed to J as well

Model6

coef(Model6)

AIC(Model6)

BIC(Model6)

Modelres6= residuals(Model6)

Modelres6

plot(Modelres6)

qqnorm(Modelres6)

acf(Modelres6, 30, main="output")

Box.test (Modelres6, lag = 16)

Fcast<-forecast(Model6,h = 25,xreg=J[c(426:450),]) # Forecasting the series

plot(Fcast)

accuracy(Model6)
```

 # Model 7 with order (2,1,2) and with regressor as L

```
Model7 = arima(NIK[4:N], xreg=L, order=c(2,1,2))# Here xreg can be changed to J as well

Model7

coef(Model7)

AIC(Model7)

BIC(Model7)

Modelres7= residuals(Model7)

plot(Modelres7)

qqnorm(Modelres7)

acf(Modelres7, 30, main="output")

Box.test (Modelres7, lag = 16)

Fcast<-forecast(Model7,h = 25,xreg=L[c(426:450),]) # Forecasting the series

plot(Fcast)

accuracy(Model7)
```