Stock Market Data Analysis

Tianyi Li

Abstract:

This paper analyzes the trend term and seasonality of time series, studies the method of time series smoothing, establishes the SARIMA model of time series, and forecasts it. The stock closing price is used as an example for data validation.

I. Introduction

Time series analysis is very important in life. This paper analyzes the characteristics of time series, such as trend term and seasonality, and studies ACF and PACF charts of time series. The stationarity of time series is analyzed, and the logarithmic transformation and difference method are used for stationarization. Finally, the SARIMA model of time series is established and predicted. The stock closing price is used as an example for data validation.

II. Analysis Section

A. feature analysis

Taking Bharat Petroleum Corporation Ltd(BPCL) as an example, Fig. 1 shows the daily close prices of BPCL from 2000 to 2020, and Fig. 2 shows the daily close prices in 2000. It can be seen that the daily close price of BPCL has a linear trend.

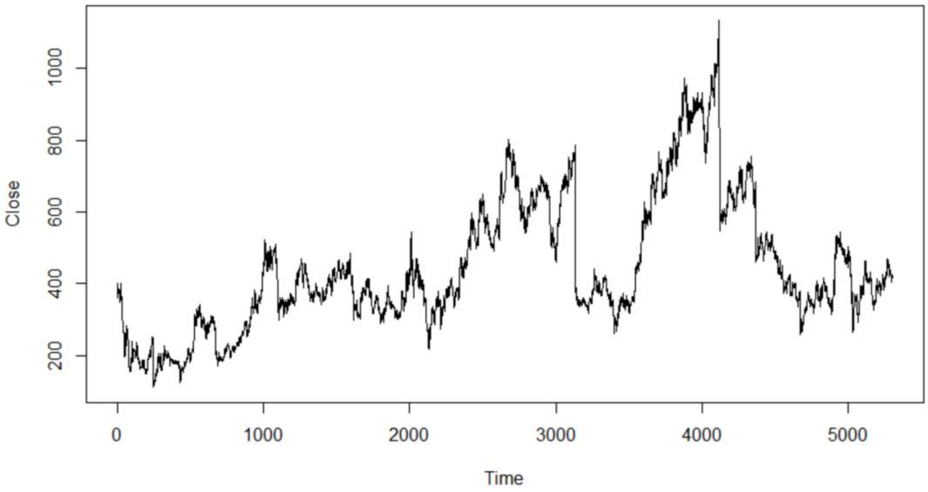

Fig. 1 Close price of day for BPCL over time

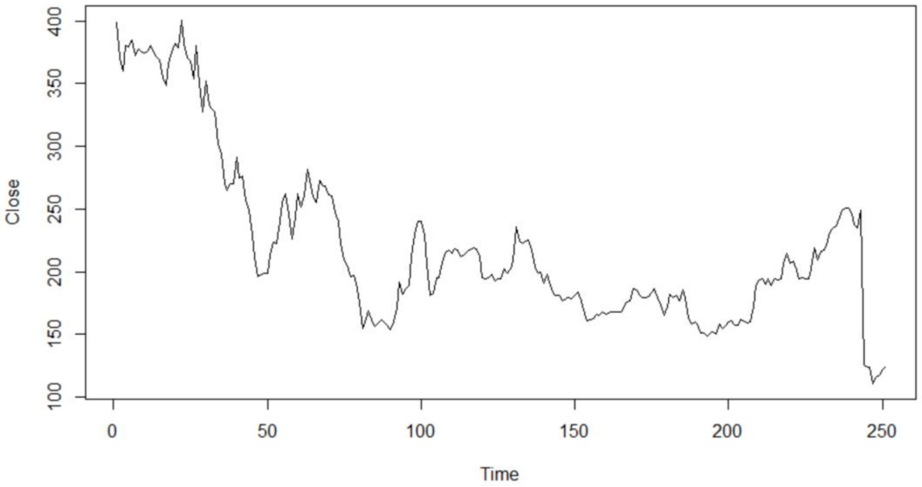

Fig. 2 Close price of day for BPCL in 2020

Fig. 3 is the periodogram of Close price of day for BPCL in over time, and frequency of BPCL is small.

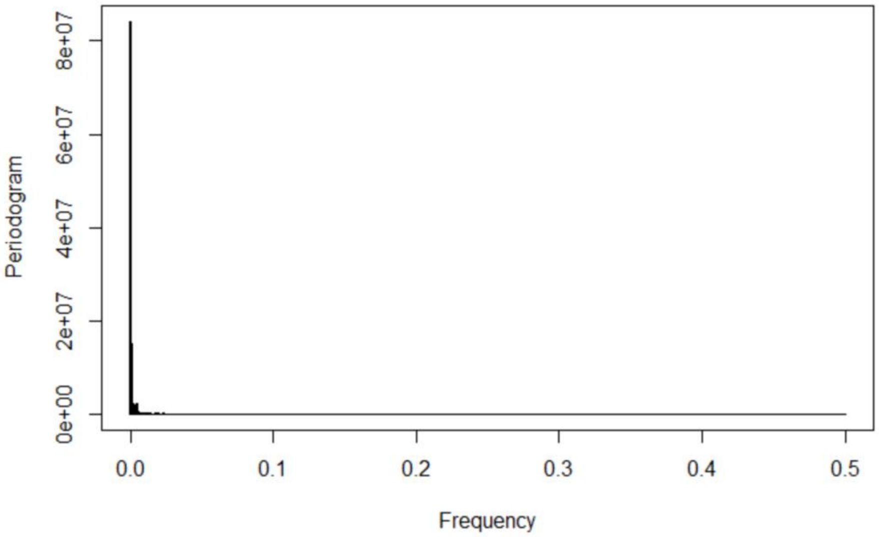

Fig. 3 Periodogram of close price of day for BPCL

Fig. 4 is a first-order difference chart of the close price. The mean and variance after the difference are 0.00425 and 17.025, respectively. It can be seen from the chart that the time series has changed dramatically.

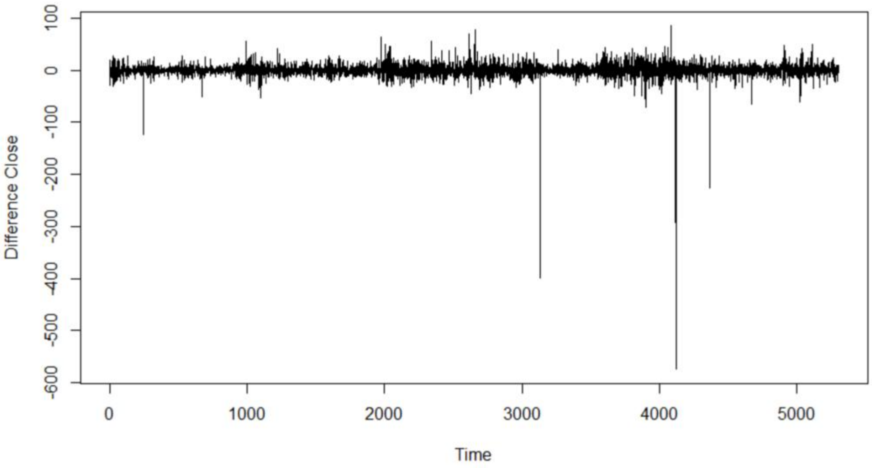

Fig. 4 Difference Close

B. transformations

This part tests the normality of time series. The histogram of the closing price is shown in Fig. 5. It can be seen that histogram is sketched.

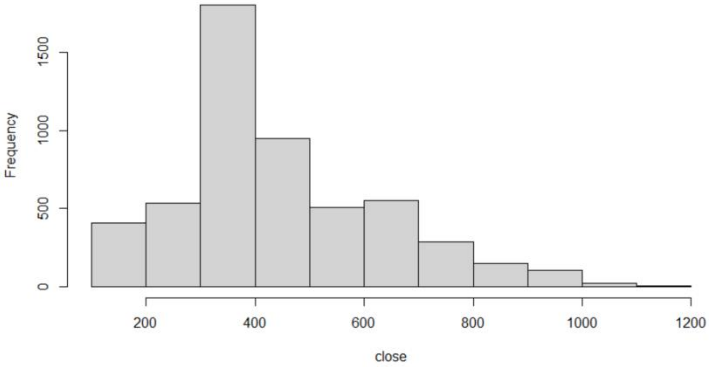

Fig. 5 Histogram of close

Then use Box-Cox transformation for normalization. Fig. 6 shows the Box-Cox figure, corresponding to the highest point λ=- 0.0202。 λ= 0(log) is also in the confidence interval and very close to λ= 0.02020202. Therefore, choose log transform.

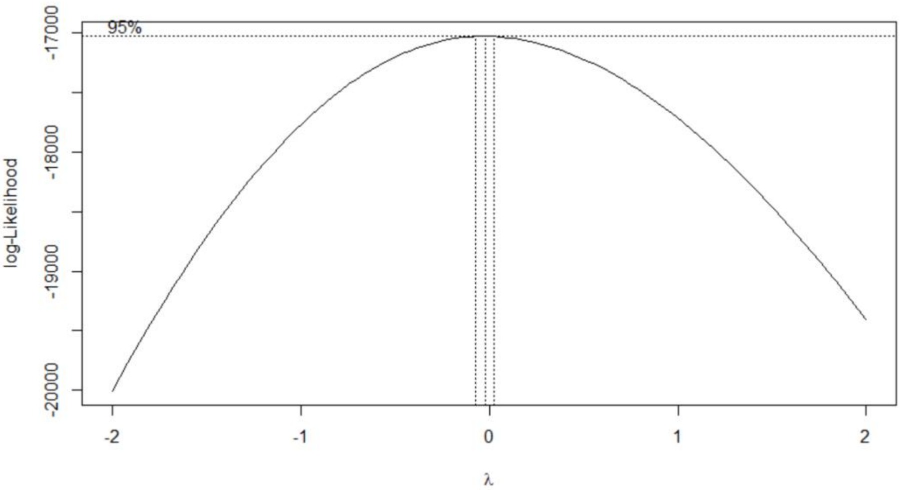

Fig. 6 Box-Cox

Fig. 7 is the histogram after log transformation. It can be seen that it basically conforms to the normal distribution, and the variance is 0.1752, which is very small compared with the variance before the transformation.

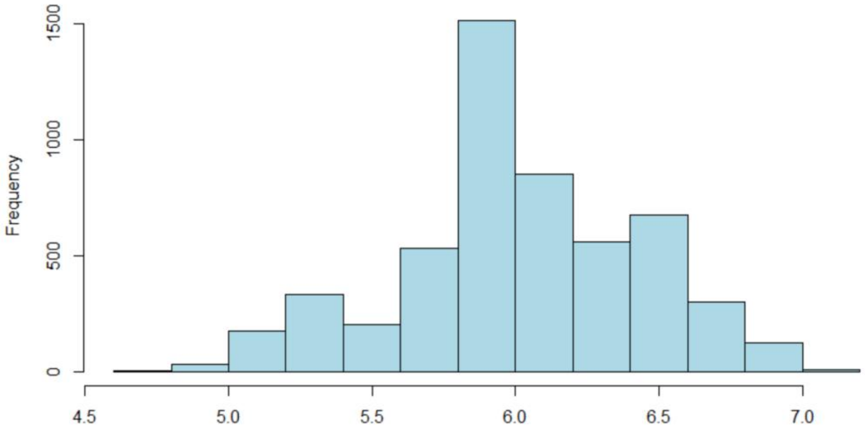

Fig. 7 Histogram of log close

Fig. 8 shows the decomposition of ln (U), showing the trend, seasonality and randomness of the time series. It can be seen that the time series has obvious seasonality.

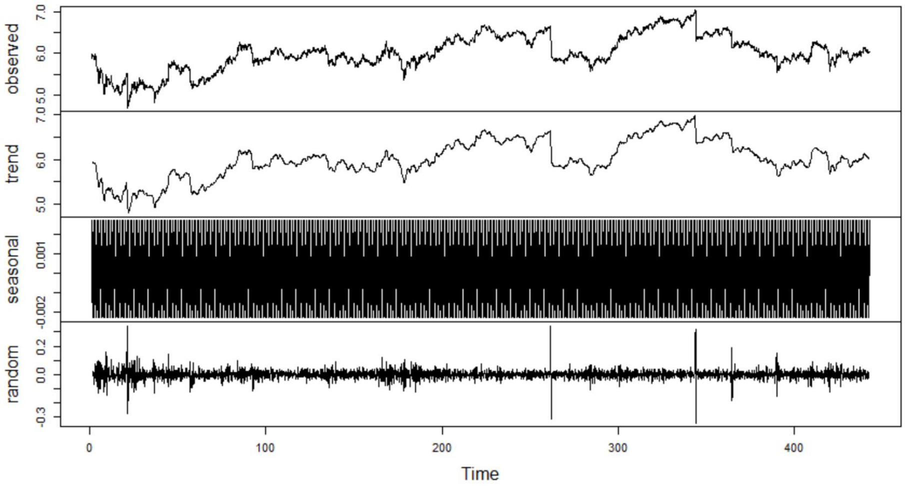

Fig. 8 Decomposition of additive time series

Then the difference method is used to eliminate the tendency, seasonality and randomness, so that the time series becomes a stationary signal. It can be seen from Figure 8 that the seasonality is in months, so the lag of the difference is set to 12. Figure 9 is the difference diagram when lag=12. It can be seen from Figure 9 that seasonality does not appear and the trend item does not exist. In addition, the variance is 0.0121, which is also very small.

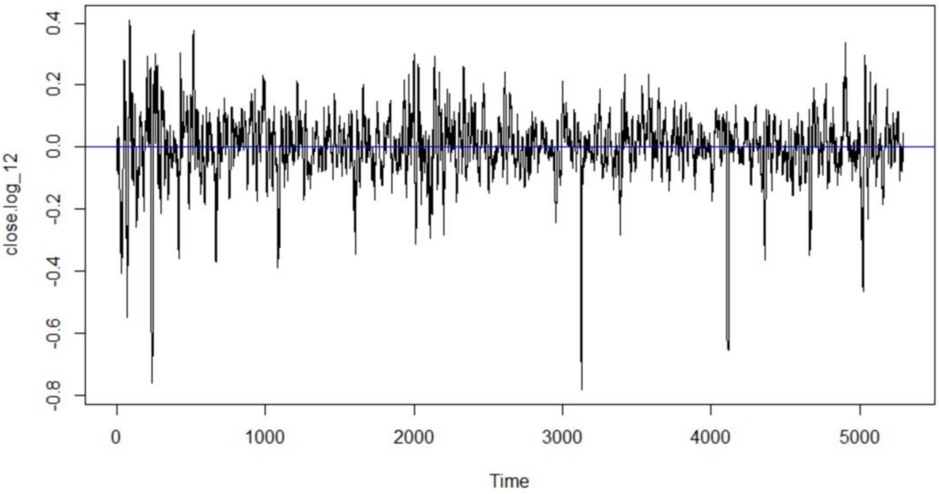

Fig. 9 Difference at lag 12

The results in Figure 9 are not enough to show that the time series after difference is stable, and ACF needs to be further tested. Figure 10 shows the slow attenuation of ACF, indicating that the time series is still unstable at this time, and further difference is needed. Figure 11 shows the ACF of the time series after two differences, and Figure 12 shows the ACF of the original closing price. Compared with Figure 12, we can see that ACF deal responses to a stationary process.

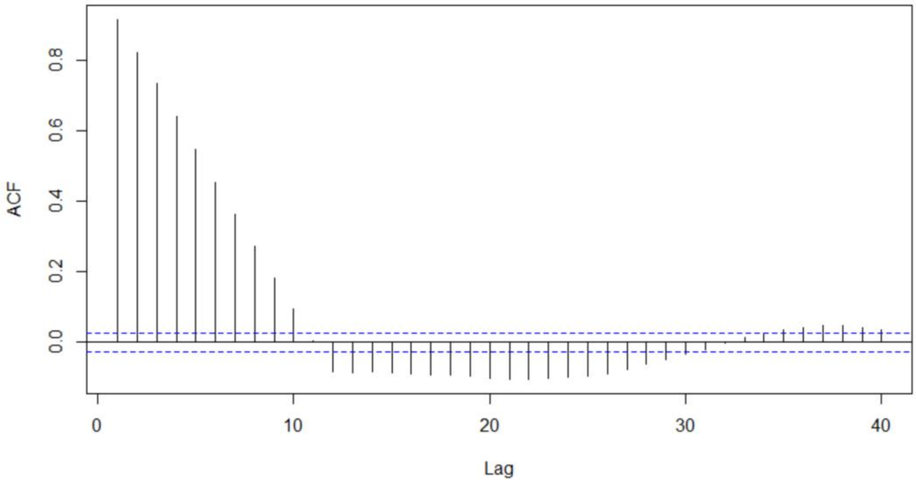

Fig. 10 ACF of the log(U), differenced at lag 12

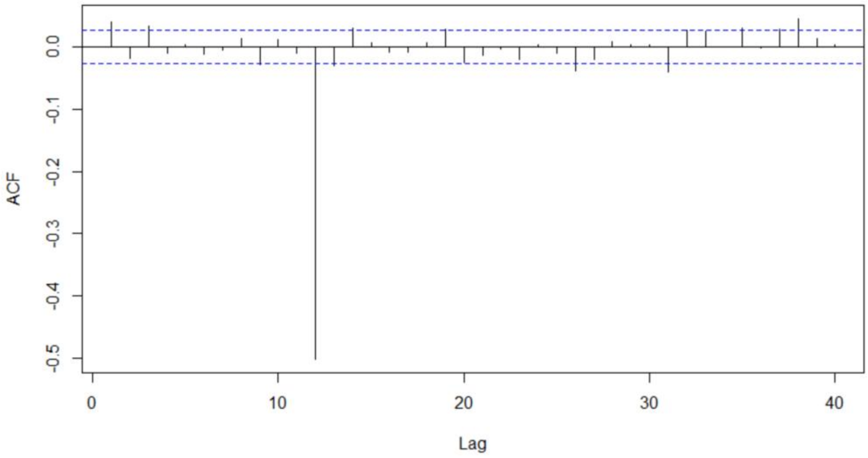

Fig. 11 ACF of the log(U), differenced at lag 12 and 1

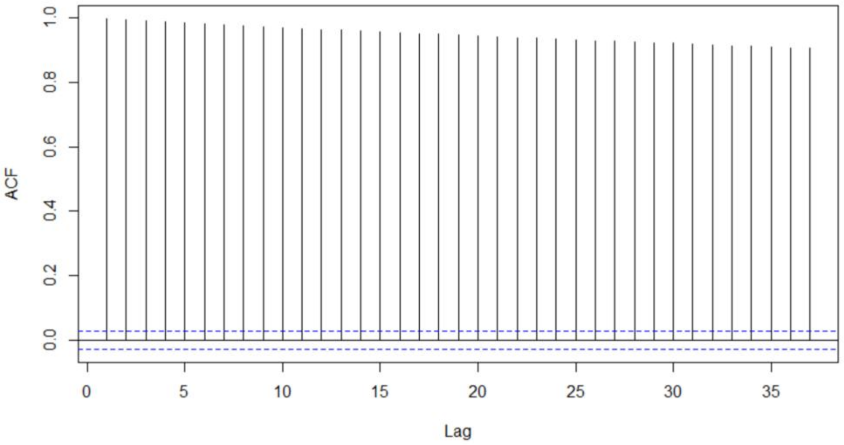

Fig. 12 ACF of original Close

C. models fitting

Figure 13 shows the PACF after twice difference. PACF outside confidence intervals: Lags 1, may be 3,9,12,15. From fig.12, ACF outside confidence intervals: Lags 1, may be 3,12. Figure 8 shows the seasonality of the time series, so the SARIMA model is selected in this section.

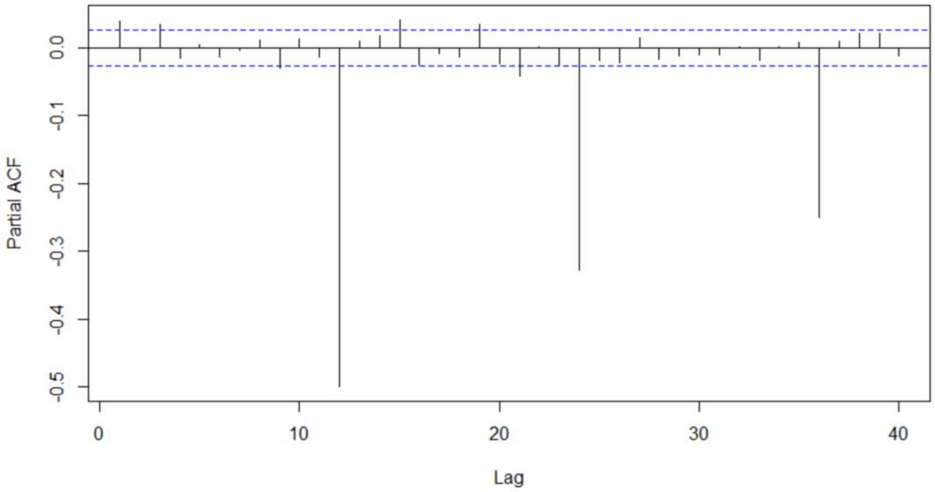

Fig. 13 PACF of the log(U), differenced at lag 12 and 1

The available parameters are shown in Table 1, and then the AIC criteria are used to select the best and second-best models. Table 2 shows the AIC and parameter combination results of each model.

As shown in Table 2, the minimum AIC is -21382.05, and the corresponding model is,

Model A：

$$\nabla_1\nabla_{12}\ln(U_t) = (1+0.0217_{(0.0138)}B-0.0068_{(0.0141)}B^2\text{-}\ 0.0335_{(0.0141)}B^3)(1-0.9974_{(0.0040)}B^{12})Z_t$$

$$\hat{\sigma} = 0.001017$$

The second smallest is -21380.16, and the corresponding model is,

Model B：

$$\nabla_1\nabla_{12}\ln(U_t) = (1+0.0199_{(0.0139)}B)(1-0.9970_{(0.003\,7)}B^{12})Z_t$$

$$\hat{\sigma} = 0.001018$$

Table 1 Parameters to be selected

Param.	Value 1	Value 2
p	1	3
d	1	\
q	1	3
P	1	\
D	1	\
Q	1	2
s	12	\

TABLE 2 Parameter combination and corresponding model AIC

AIC.	p	d	q	P	D	Q	s
-21380.16	1	1	1	1	1	1	12
-21379.61	3	1	1	1	1	1	12
-21382.05	1	1	3	1	1	1	12
-21378.95	3	1	3	1	1	1	12
-21375.77	1	1	1	1	1	2	12
-21376.97	3	1	3	1	1	2	12

Next, check Model A and B

From Fig.14，Fig.15，and Fig.16，Model A:No trend, no visible change of variance, no seasonality. ACF of residuals is within confidence intervals and can be counted as zeros. Histogram and Q-Q plot look OK.

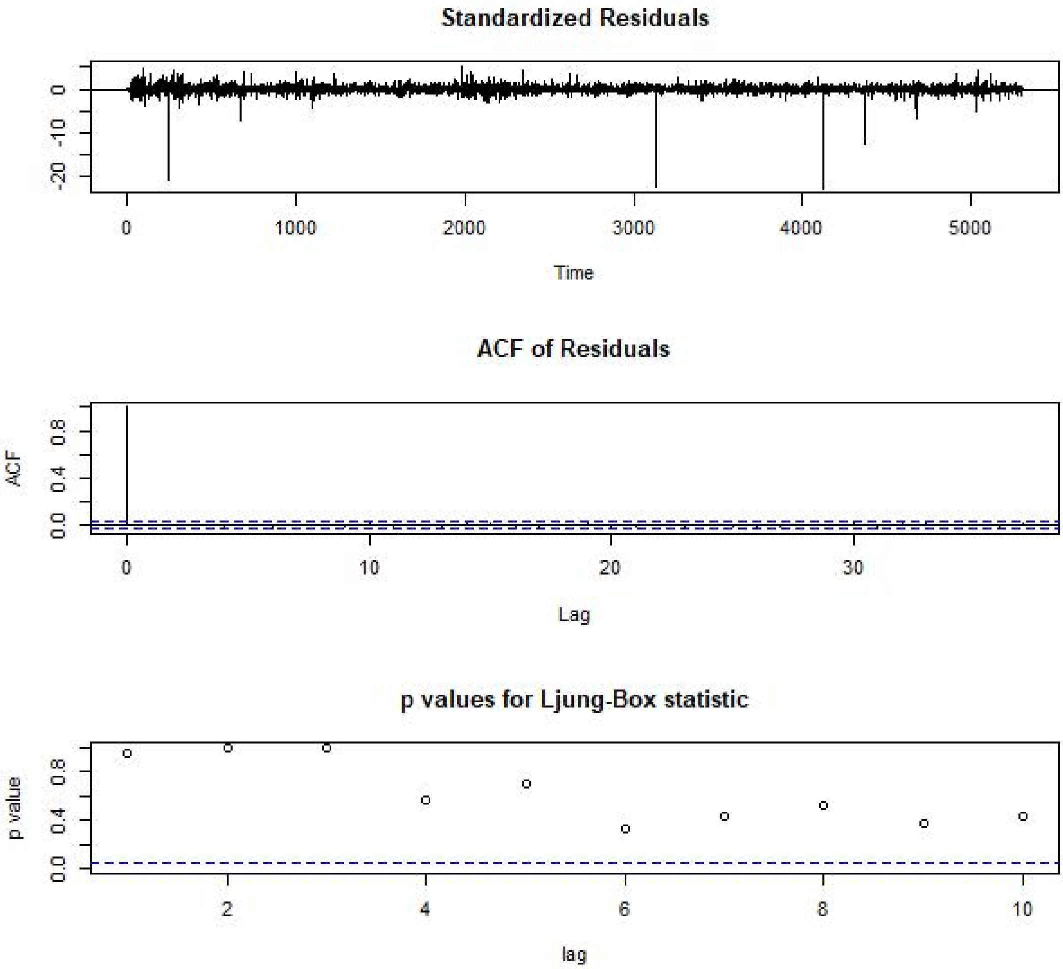

Fig. 14 Residual correlation property diagram(Model A)

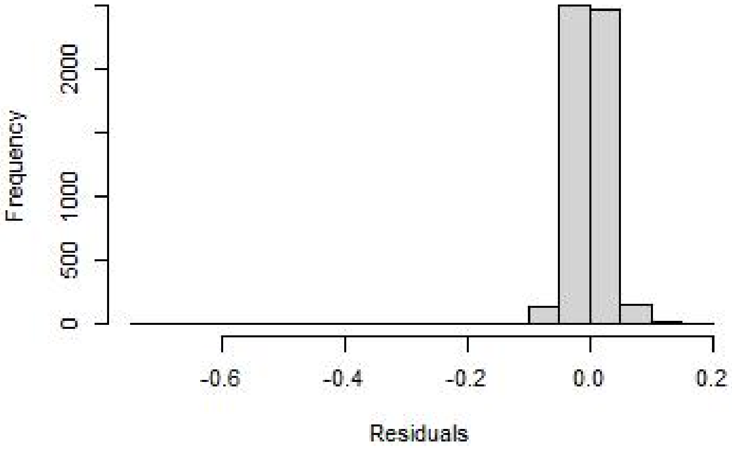

Fig. 15 Histogram of residuals(Model A)

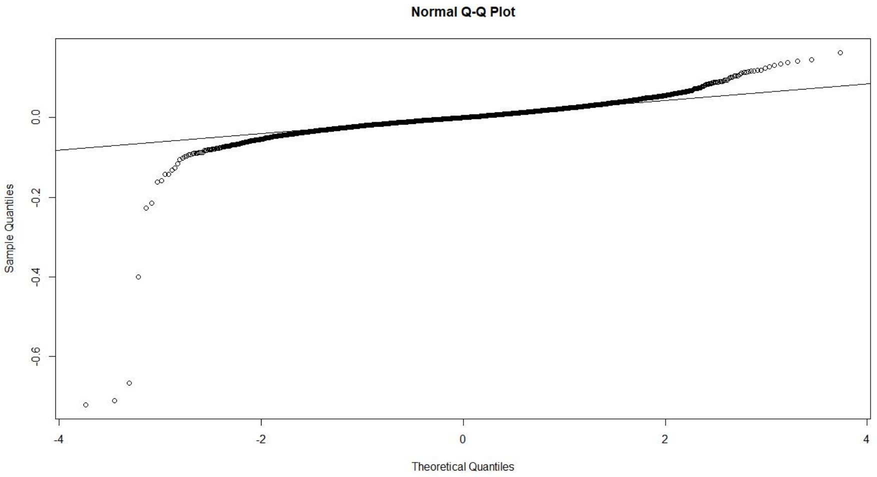

Fig. 16 Normal Q-Q(Model A)

The residual error of Model A is tested, and the test results are shown in the table. Except that the p-value of shapiro is less than 0.05, the rest are greater than 0.05. Fitted residuals to AR(0), i.e.WN, is 0.001014. So, model A passes diagnostic checking and can be used for forecasting.

Table 3 Test result

Test Approach	Parm.	p-value
shapiro.test	\	2.2e-16
Box.test	Box-Pierce/ fitdf=3	0.2803
Box.test	Ljung-Box/ fitdf=3	0.279
Box.test	Ljung-Box/ fitdf=0	0.5331

From Fig.17,Fig.18,and Fig.19， Model B:No trend, no visible change of variance, no seasonality. ACF of residuals is within confidence intervals and can be counted as zeros. Histogram and Q-Q plot look OK.

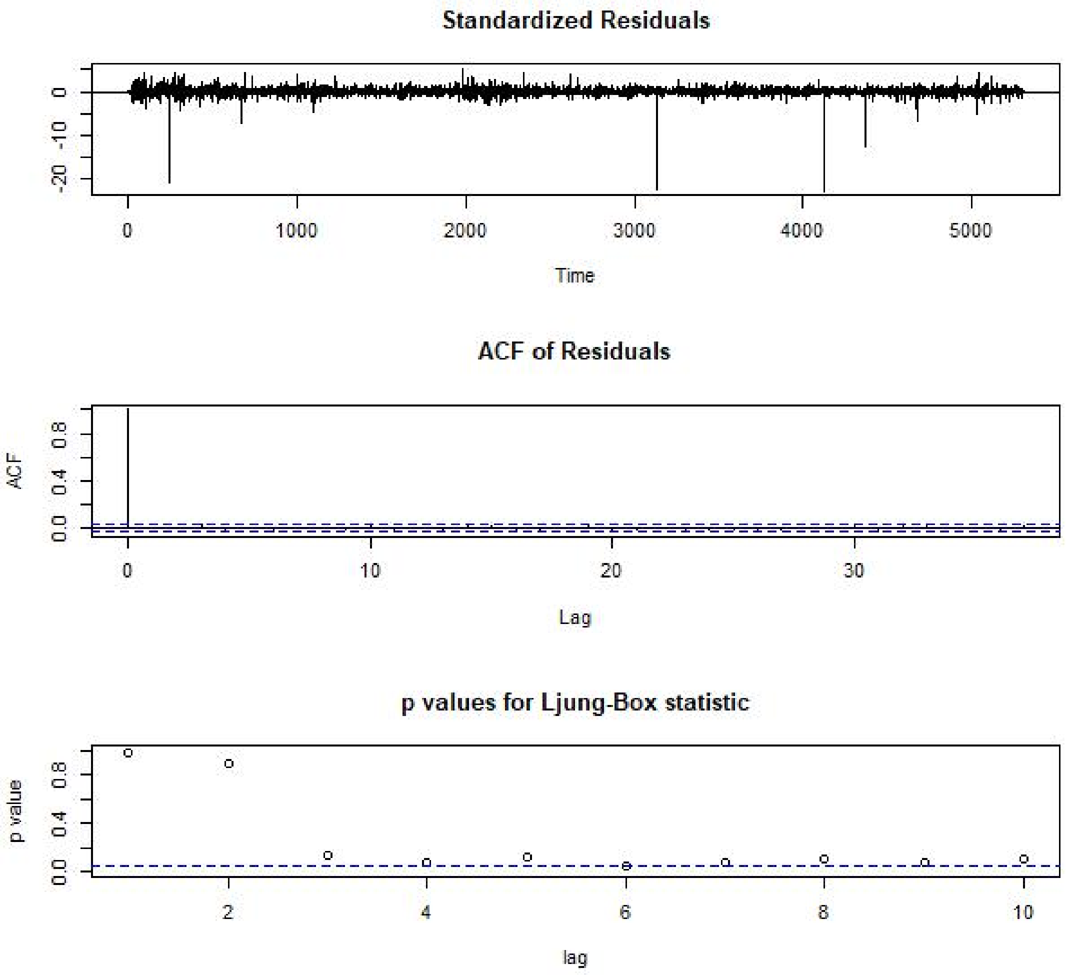

Fig. 17 Residual correlation property diagram(Model B)

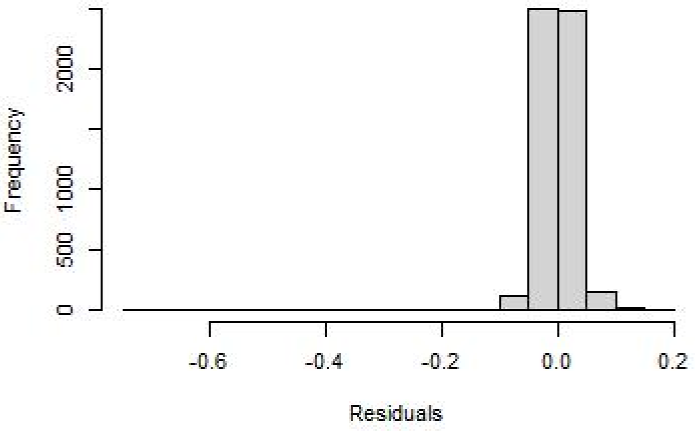

Fig. 18 Histogram of residuals(Model B)

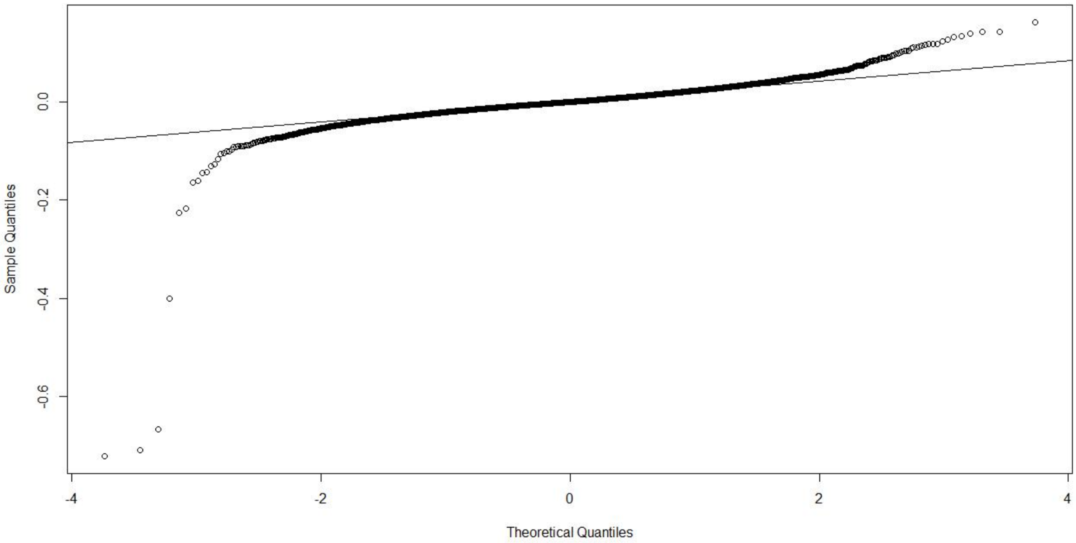

Fig. 19 Normal Q-Q(Model B)

The residual error of Model B is tested, and the test results are shown in Table 4. All p-values are not much greater than 0.05. Fitted residuals to AR(0), i.e.WN, is 0.001014.

Table 4 Test result

Test Approach	Parm.	p-value
shapiro.test	\	2.2e-16
Box.test	Box-Pierce/ fitdf=3	0.05265
Box.test	Ljung-Box/ fitdf=3	0.05225
Box.test	Ljung-Box/ fitdf=0	0.158

D. forecasting

Through the analysis in C, model A is the final model, which can be used for prediction. The results of 12 time prediction using model A are shown in Fig. 20 and Fig. 21. Fig. 22 shows the comparison between the real value and the predicted value. The black circle represents the predicted value, the red line represents the real value, and the blue dotted line represents confidence intervals.

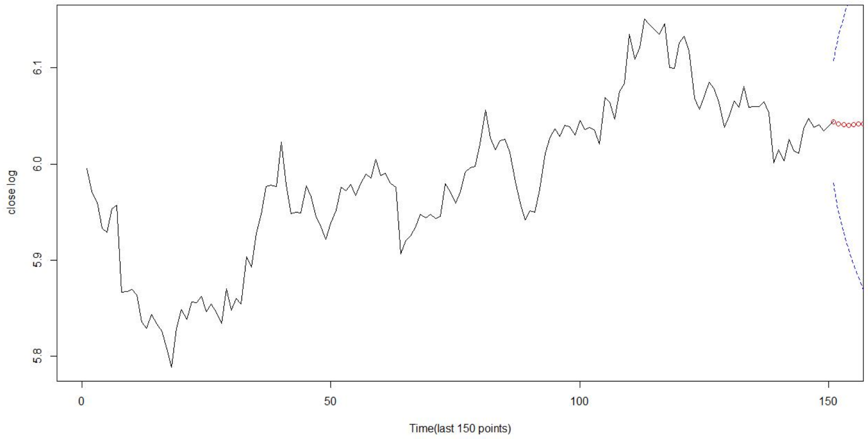

Fig. 20 Log forecast of transformed data using model A

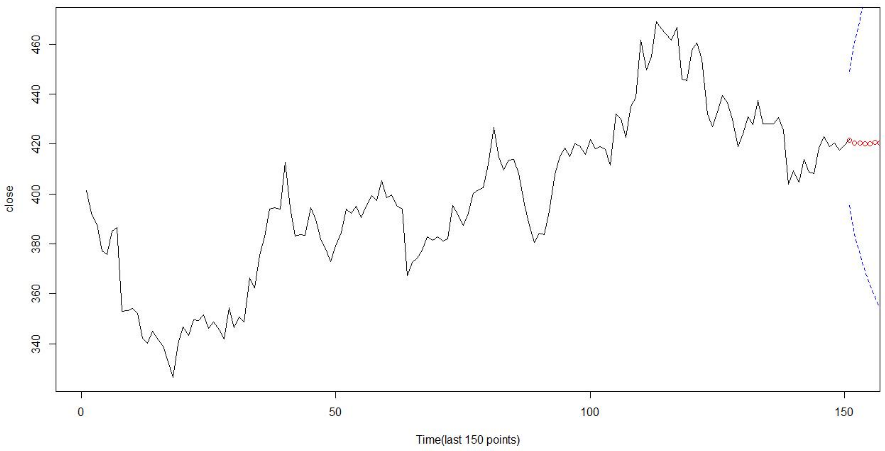

Fig. 21 Forecast of transformed data using model A

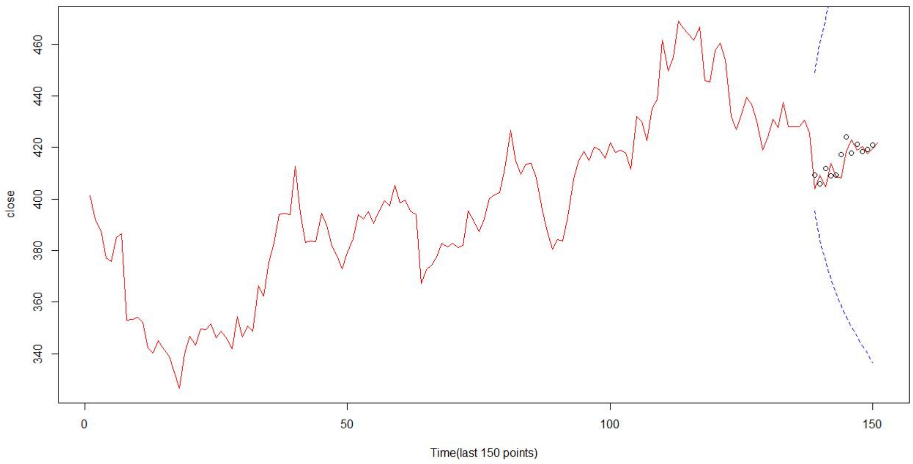

Fig. 22 Original vs. Forecast

III. Conclusion

This paper analyzes time characteristics such as trend term and seasonality, and analyzes the stationarity of time series. The SARIMA model of time series is established using stock data, and the prediction is made.

IV. Reference

Data source: https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data
Software used: RStudio, RMarkdown, Word
Partly self-learned from: Github Datacamp
Individuals helped: Chris, Aoao, Karin (my PSTAT 174 classmates)

V. Appendix

```
library(feasts)
library(tidyverse)
library(lubridate)
library(TSA)
library(ggplot2)
library(ggfortify)
library(MASS)
library(ggplot2)
library(ggfortify)
library(MASS)
library(forecast)

 data <- read.csv('StockArchive/BPCL.csv',encoding="UTF-8")
 data$Date <- ymd(data$Date)
 data_tsi <- as_tsibble(data, index=Date)
 plot(data$Close,type='l',xlab='Time',ylab='Close')
 periodogram(data$Close)

 #data_tsi %>%
 #gg_season(data_tsi$Close)

 diff = data$Close[2:length(data$Close)] - data$Close[1:length(data$Close)-1]
 diff_mean = mean(diff)
 diff_var = var(diff)
 qq <- quantile(diff)
 sig <- 1.5 * (qq[4]-qq[2])

 plot(diff,type='l',xlab='Time',ylab='Difference Close')
 par(new = TRUE)
 plot(diff(which(diff>=sig)),col = "red",ylab='')
 par(new = TRUE)
 plot(diff(which(diff<=-sig)),col = "red",ylab='')

data <- read.csv('StockArchive/BPCL.csv',encoding="UTF-8")
#data$Date <- ymd(data$Date)
#data_tsi <- as_tsibble(data, index=Date)
close = data$Close

plot(close,type='l',xlab='Time',ylab='Close')

hist(close)
# acf(log(close))
bcTransform <- boxcox(close~ as.numeric(1:length(close)))
bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]

close.bc = (1/lambda)*(close^lambda-1)
close.log <- log(close)
```

```
plot.ts(close.bc)
plot.ts(close.log)

hist(close.log, col="light blue", xlab="", main="histogram; ln(U_t)")
hist(close.bc, col="light blue", xlab="", main="histogram; bc(U_t)")

y <- ts(as.ts(close.log), frequency = 12)
decomp <- decompose(y)
plot(decomp)

log_var = var(close.log)
close.log_12 <- diff(close.log, lag=12)
plot.ts(close.log_12, main="Ln(U_t) differenced at lag 12")
log12_var = var(close.log_12)
fit <- lm(close.log_12 ~ as.numeric(1:length(close.log_12)));
abline(fit, col="red")
log12_mean = mean(close.log_12)
abline(h=mean(close.stat), col="blue")

log12_mean = mean(close.log_12)
close.stat <- diff(close.log_12, lag=1)
plot.ts(close.stat, main="Ln(U_t) differenced at lag 12 and lag 1")
# plot.ts(close.stat, main="Ln(U_t) differenced at lag 12 & lag 1")
fit <- lm(close.stat ~ as.numeric(1:length(close.stat)));
abline(fit, col="red")
stat_mean = mean(close.stat)
abline(h=mean(close.stat), col="blue")
stat_var = var(close.stat)

acf(close.log, lag.max=40, main="ACF of the log(U_t)")
acf(close.log_12, lag.max=40, main="ACF of the log(U_t), differenced at lag 12")
acf(close.stat, lag.max=40, main="ACF of the log(U_t), differenced at lags 12 and 1")
hist(close.stat, col="light blue", xlab="", main="histogram; ln(U_t) differenced at lags 12 & 1")

hist(close.stat, density=20,breaks=20, col="blue", xlab="", prob=TRUE)
m <- mean(close.stat)
std <- sqrt(var(close.stat))
curve( dnorm(x,m,std), add=TRUE )

pacf(close.stat, lag.max=40, main="PACF of the ln(U_t), differenced at lags 12 and 1")
```

```
p = 0
d = 1
q = 3
P = 1
D = 1
Q = 1

# modeA
modelA = arima(close.log, order=c(p,d,q), seasonal = list(order = c(P,D,Q),
period = 12), method="ML")
summary(modelA)

win.graph(width=6.5,height=6);
tsdiag(modelA)
win.graph(width=4,height=3,pointsize=8)
hist(residuals(modelA),xlab='Residuals')
win.graph(width=6.5,height=6)
qqnorm(residuals(modelA))
qqline(residuals(modelA))

res_A = residuals(modelA)
shapiro.test(res_A[1:5000])
Box.test(res_A,lag=12,type=c("Box-Pierce"),fitdf=3)
Box.test(res_A,lag=12,type=c("Ljung-Box"),fitdf=3)
Box.test(res_A,lag=12,type=c("Ljung-Box"),fitdf=0)

ar(res_A,aic = TRUE,order.max=NULL,method=c("yule-walker"))

# modelB

modelB = arima(close.log, order=c(0,1,1), seasonal = list(order = c(1,1,1), period
= 12), method="ML")
summary(modelB)

win.graph(width=6.5,height=6);
tsdiag(modelB)
win.graph(width=4,height=3,pointsize=8)
hist(residuals(modelB),xlab='Residuals')
win.graph(width=6.5,height=6)
qqnorm(residuals(modelB))
qqline(residuals(modelB))

res_B = residuals(modelB)
```

```
    shapiro.test(res_B[1:5000])
    Box.test(res_B,lag=12,type=c("Box-Pierce"),fitdf=3)
    Box.test(res_B,lag=12,type=c("Ljung-Box"),fitdf=3)
    Box.test(res_B,lag=12,type=c("Ljung-Box"),fitdf=0)

    ar(res_B,aic = TRUE,order.max=NULL,method=c("yule-walker"))


data <- read.csv('StockArchive/BPCL.csv',encoding="UTF-8")
#data$Date <- ymd(data$Date)
#data_tsi <- as_tsibble(data, index=Date)
close = data$Close

plot(close,type='l',xlab='Time',ylab='Close')

hist(close)
bcTransform <- boxcox(close~ as.numeric(1:length(close)))
bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]

close.bc = (1/lambda)*(close^lambda-1)
close.log <- log(close)
plot.ts(close.bc)
plot.ts(close.log)

fit.A <- arima(close.log[1:(length(close.log)-12)], order=c(0,1,1), seasonal = list(order
= c(1,1,1), period = 12), method="ML")
forecast(fit.A)
# To produce graph with 12 forecasts on transformed data:
pred.tr <- predict(fit.A, n.ahead = 12)
U.tr= pred.tr$pred + 2*pred.tr$se # upper bound of prediction interval
L.tr= pred.tr$pred - 2*pred.tr$se # lower bound

win.graph(width=6.5,height=6);
ts.plot(close.log[(length(close.log)-150):(length(close.log))],
        gpars=list(xlab="Time(last 150 points)",ylab='close log'))
a = 150+1
b = 150+12
lines(a:b,U.tr, col="blue", lty="dashed")
lines(a:b,L.tr, col="blue", lty="dashed")
points(a:b,pred.tr$pred, col="red")

# To produce graph with forecasts on original data:
win.graph(width=6.5,height=10);
```

```
pred.orig <- exp(pred.tr$pred)
U= exp(U.tr)
L= exp(L.tr)
ts.plot(close[(length(close.log)-150):(length(close.log))],
            gpars=list(xlab="Time(last 150 points)",ylab='close'),lty=c(1:3))
a = 150+1
b = 150+12
lines(a:b, U, col="blue", lty="dashed")
lines(a:b, L, col="blue", lty="dashed")
points(a:b, pred.orig, col="red")

# To zoom the graph, starting from entry 100:
win.graph(width=6.5,height=10);
ts.plot(close[(length(close.log)-150):(length(close.log))],
            gpars=list(xlab="Time(last 150 points)",ylab='close'),lty=c(1:3))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
a = 150+1
b = 150+12
points(a:b, pred.orig, col="red")
```