Project III – Apple Stock Prices Analysis

Li Hantao, G2101725H, MSAI, hli038@e.ntu.edu.sg

The data are from the daily historical Apple stock prices(open, high, low, close and adjusted prices) from February 1, 2002, to January 31, 2017, extracted from the Yahoo Finance website. The data has logged the prices of the Apple stock every day and comprises of the open, close, low, high and the adjusted close prices of the stock for the span of 15 years.

The goal of the project is to discover an interesting trend in the apple stock prices over the past 15 years (3775 attributes) and to design and develop the best model for forecasting.

I. Original Data Analysis

This project asks us to fit an appropriate forecasting model for the stock price of Apple. First of all, we will look in the original dataset, trying to find some attributes about the time series data. We chose the Adjusted Closing Price for analysis in this project. The original plot of data is shown in Fig. 1. We can effortlessly know that the maximum and minimum of data are 33.25 and 0.23, with a mean value of 10.78.

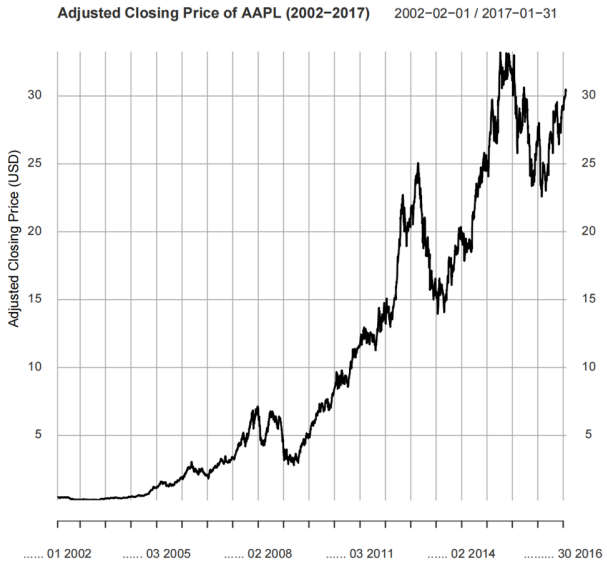

Fig. 1. Original Data Plot.

Observing the original data plot, we can find an apparent increasing trending component with a volatility pattern. This phenomenon is called 'Volatility Clustering,' usually found in a finical data curve. Also, the increasing trend shows a non-linear attribute that rapidly increases in the later period.

For the seasonality pattern it may contain, we can quickly know that the period of it is 12 months. Firstly, we will implement the seasonal decomposition on the original data to get the knowledge for more detailed information with the slt() function. Before that, we convert the data into the monthly format. The figure of them is shown in Fig. 2. Secondly, we can calculate the original data's ACF and PACF plots, shown in Fig. 3, which indicate that the data is non-stationary and non-linear.

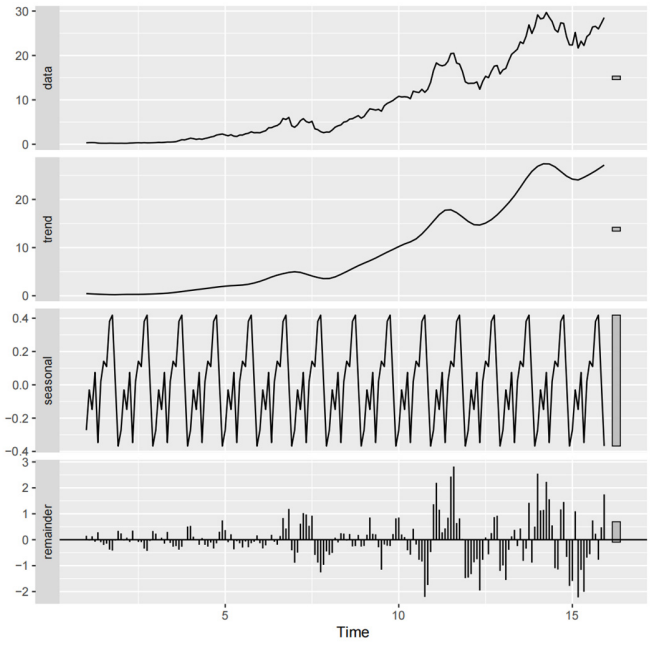

Fig. 2. Seasonal Decomposition on Original Data.

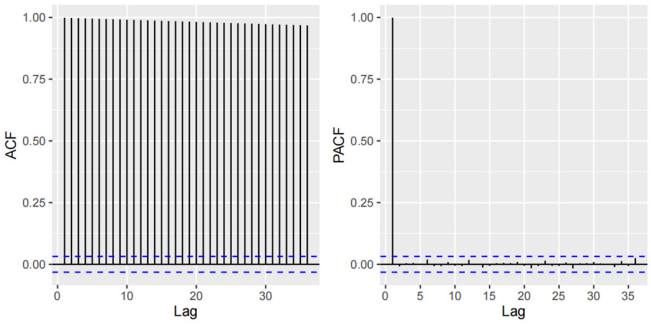

Fig. 3. ACF and PACF of Original Data.

We can observe a noticeable increase in the seasonality variance from the data plot and remainder plot in Fig. 2. Thus, we should try to implement the Box-Cox transformation. Moreover, the Augmented Dickey-Fuller Test gives the p-value of 0.4114, showing that the data is non-stationary.

II. Transformation

Box-Cox transformation can stabilize the data with the fluctuation of variance. There is a hyper-parameter in Box-Cox, the lambda. This project will implement the transformation with a zero lambda. On the other hand, we utilize the first-order differencing to remove the trending and make the data stable. The data after the Box-Cox transformation and first-order differencing is shown in Fig. 4, which is multiplied by 100 to be interpreted as percentage changes in the price. The ADF test shows that the p-value of 0.01, exhibiting the stationarity of the data.

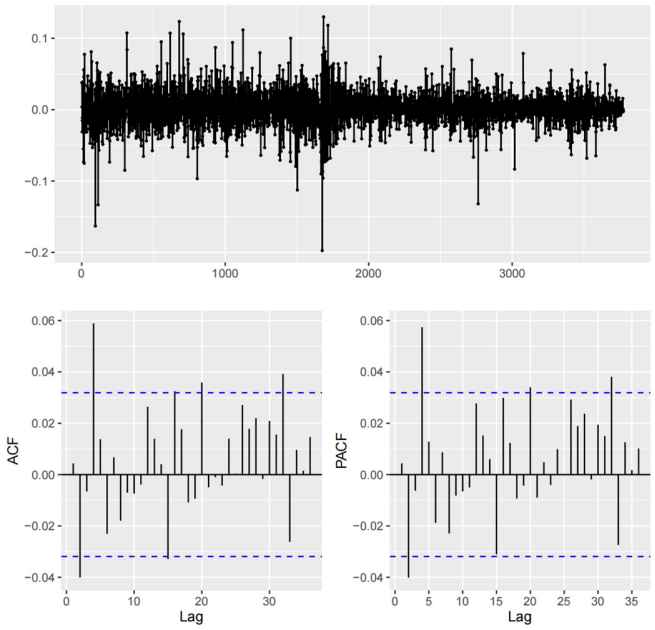

Fig. 4. Data after Transformers (Log-return Data).

For the reason that we have observed the volatility clustering in the log-return data, we should further analyze the ACF/PACF plots of the absolute and squared log-return data, which are shown in Fig. 5 and Fig. 6. Observed from the ACF/PACF of those two return data, we can ensure that the returns are not independently and identically distributed.

Furthermore, we implement the QQ-plot checking to explore the data distributional shape of the AAPL returns, which is shown in Fig. 7. The QQ-plot suggests that the distribution of returns has a thicker tail while somewhat being skewed to the left, referred to as a heavy-tailed distribution. We also utilize kurtosis and skewness testing to test the returns data, where their answers are 5.4356 and -0.1900, respectively. These two results verify our observation above. The positive kurtosis score indicates the heavy-tailed distribution, while the negative skewness score led to a left-skewed pattern.

In summary, the log-returns data are serially uncorrelated, having a heavy-tailed and left-skewed distribution. The ARCH/GARCH model will be a suitable framework for modeling and analyzing time series with such patterns.

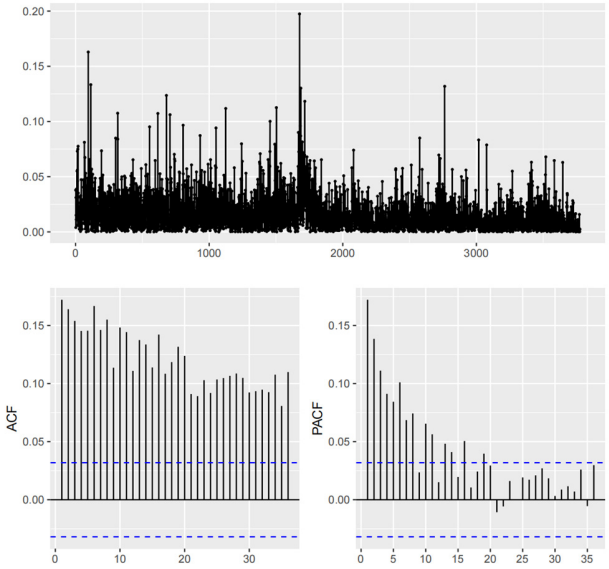

Fig. 5. Absolute Log-return Data.

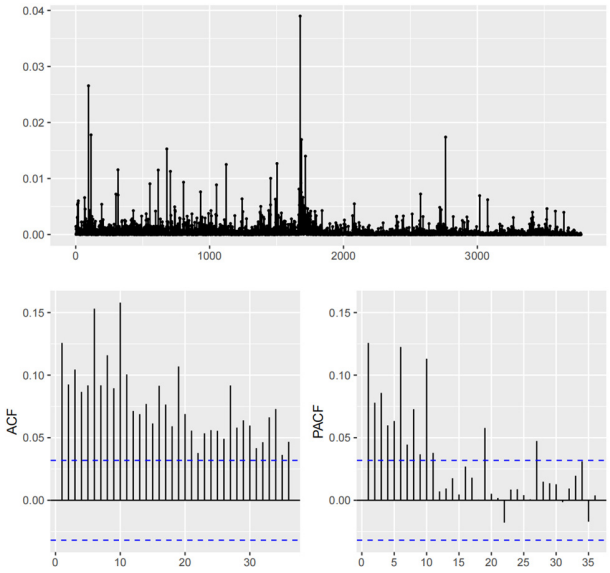

Fig. 6. Squared Log-return Data.

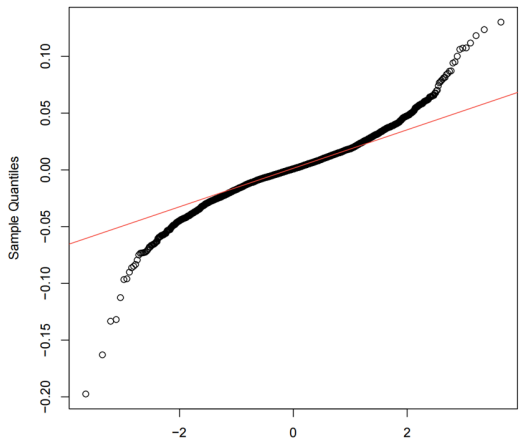

Fig. 7. QQ-Plot of Log-return Data.

III. Split Dataset

Before we begin the model selection, we first split the data into the training and validation sets. Due to the volatility clustering in the time series data, it is unnecessary to forecast the stock price for longer than one month; thus, we set the validation length to 30 days instead of setting a specific ratio for the training and validation set.

After the splitting, there are 3745 data points in the training set, while 30 data points in the validation set. We will not touch the validation data except for the prediction results.

IV. GARCH Model Fitting

Firstly, we utilize the extended autocorrelation function (EACF) to obtain the optimal parameter setting of the ARMA model. Fig. 8 presents the EACF results of log-returns, absolute log-returns, and squared log-returns, respectively.

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 o x o x o o o o o o  o  o  o
1 x x o x o o o o o o  o  o  o
2 x x o x o x o o o o  o  o  o
3 x x o o o x o o o o  o  o  o
4 x x x x o x o o o o  o  o  o
5 x x x x x x o o o o  o  o  o
6 x x o x x x o o o o  o  o  o
7 x x x x x x x o o o  o  o  o
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x  x  x  x
1 x o o o o o o x x o o  x  o  o
2 x x o o o o o o x o o  x  o  o
3 x o x o o o o o o o o  o  o  o
4 x x o o o o o o o o o  o  o  o
5 x x x x x o o o o o o  o  o  o
6 x x x x x x o o o o o  o  o  o
7 x x x x x x x o o o o  o  o  o
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x  x  x  x
1 x o o o o x x o o x o  o  o  o
2 x x o o o x o o o x o  o  o  o
3 x x o o o x o o o x o  o  o  o
4 x x x x o x o o o x o  o  x  o
5 x x x x x o o o o x o  x  o  o
6 x x x x x x o o o x o  o  x  o
7 x x x x x x x o o x o  o  x  o
```

Fig. 8. EACF Results.

The EACF of daily returns suggests the parameter of (4,0); the EACF of absolute returns suggests the parameter of (1,1), (2,2), or (3,3); the EACF of squared returns suggests the parameter of (1,1). Considering the above three suggestions together, we chose the (1,1) as the optimal parameter setting of the data, which indicates the GARCH(1,1) model.

Then, we can implement the diagnostic check of the GARCH(1,1) model to examine the attribute of this model. The result of standardized residuals checking is shown in Fig. 9. The QQ-plot of it is shown in Fig. 10 as well. Fig. 11 shows the result of the Generalized Portmanteau Test of the squared residuals of the GARCH(1,1). Fig. 12 shows the ACF/PACF plots of the squared residuals.

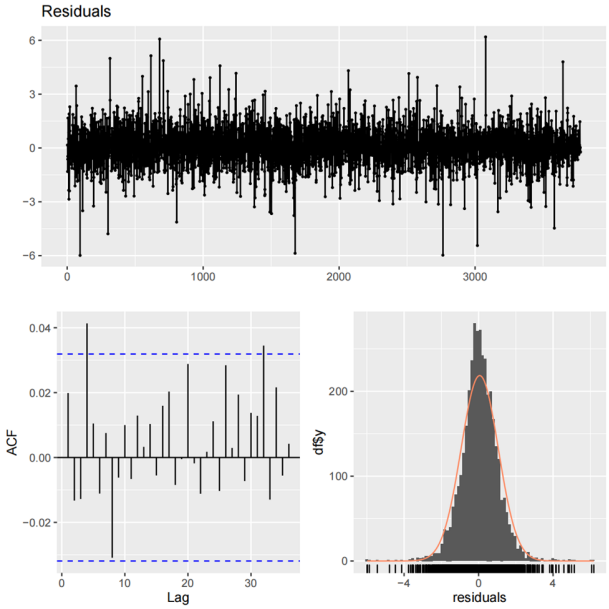

Fig. 9. Diagnostic Check of Residuals of GARCH(1,1).

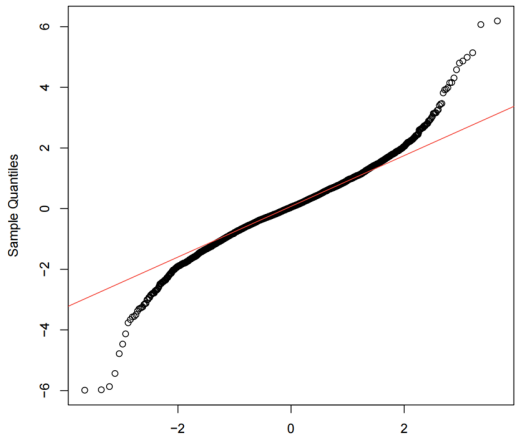

Fig. 10. QQ-Plot of GARCH(1,1).

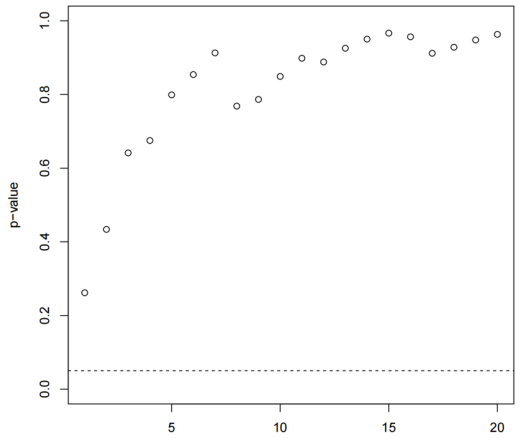

Fig. 11. Generalized Portmanteau Test.

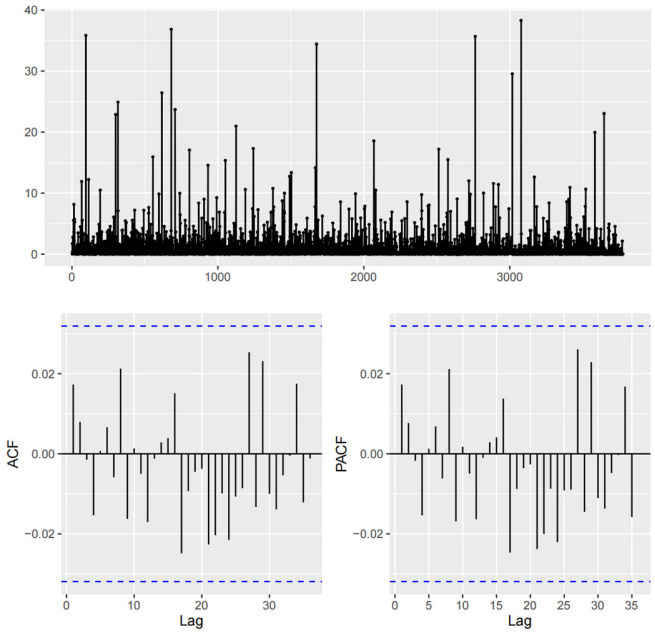

Fig. 12. ACF/PACF plots of the squared residuals.

The p-values in Fig. 11 are all higher than 0.05, suggesting that the squared residuals are uncorrelated over time, and hence the standardized residuals may be independent.

Secondly, we utilize the rugarch library in R to find out the optimal GARCH(1,1) model for both data distributions and sub-types of GARCH models.

To determine the most appropriate distribution setting, we implement a series of sGARCH(1,1) models with different data distribution presets to examine their likelihood (the higher, the better) and information criterion (Akaike, the lower, the better). The results are shown in Table I.

TABLE I
LIKELIHOOD AND AKAIKE VALUE OF DIFFERENT DISTRIBUTIONS

Distribution	Likelihood	Akaike
Normal Distribution	9296.009	-4.9229
Skew Normal Distribution	9296.242	-4.9225
T-Distribution	**9472.314**	**-5.0158**
Skew T-Distribution	**9472.777**	**-5.0155**
Generalized Error Distribution	9449.263	-5.0036
Skew Generalized Error Distribution	9450.581	-5.0038
Normal Inverse Gaussian Distribution	9468.302	-5.0131
Generalized Hyperbolic Distribution	9472.504	-5.0148
Johnson's S_U Distribution	9471.499	-5.0148

From the results shown in the table above, we can find out the (Skew-) T-Distribution performs better than others; thus, we select the T-Distribution as the distribution setting.

To determine the most appropriate sub-model under the fGARCH model (or other GARCH models), similar to the distribution selection, we implement a series of testing to find out the optimal sub-model. The results are shown in Table II.

TABLE II
LIKELIHOOD AND AKAIKE VALUE OF DIFFERENT SUB-MODELS

Sub-Model / Model	Likelihood	Akaike
fGARCH	9472.314	-5.0158
fGARCH - TGARCH	9494.400	-5.0270
fGARCH - AVGARCH	9494.357	-5.0264
fGARCH - NGARCH	9482.028	-5.0204
fGARCH - NAGARCH	9487.081	-5.0231
fGARCH - APARCH	9494.421	-5.0264
fGARCH - GJRGARCH	9482.289	-5.0206
fGARCH - ALLGARCH	9495.035	-5.0262
eGARCH	**9496.097**	**-5.0279**
gjrGARCH	9482.289	-5.0206
apARCH	9494.421	-5.0264
iGARCH	9471.993	-5.0162
csGARCH	9486.041	-5.0220

The results shown in the table above indicate that the eGARCH(1,1) is the optimal model to fit the time series data returns. Consequently, we chose eGARCH(1,1) as our final model for implementing forecasting of the AAPL. In addition, all of the above models have passed the diagnostic testing and Ljung-Box testing.

V. FORECASTING

From now, we will focus on the training set data instead of the whole dataset. We re-implement the model fitting to test the likelihood and Akaike score of the training data. The results of them are 9385.934 and -5.009311, respectively.

We can easily generate the forecasting graphs with the N-Roll configuration, as shown in Fig. 14. The yellow shaded area indicates the 95% upper and lower bounds. The left parts are 30-days forecast series with unconditional 1-Sigma and forecast unconditional sigma plots.

The rugarch package also allows the simulation of GARCH models. Fig. 13 shows five 8-years simulation plots of the eGARCH(1,1) model using the rugarch package.

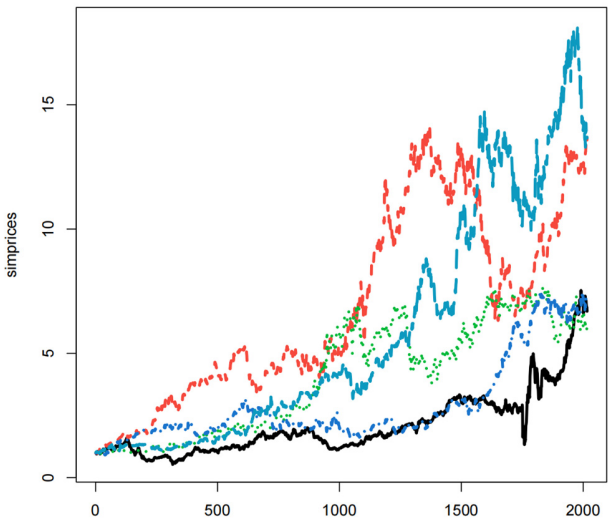

Fig. 13. 8-years simulations of eGARCH(1,1).

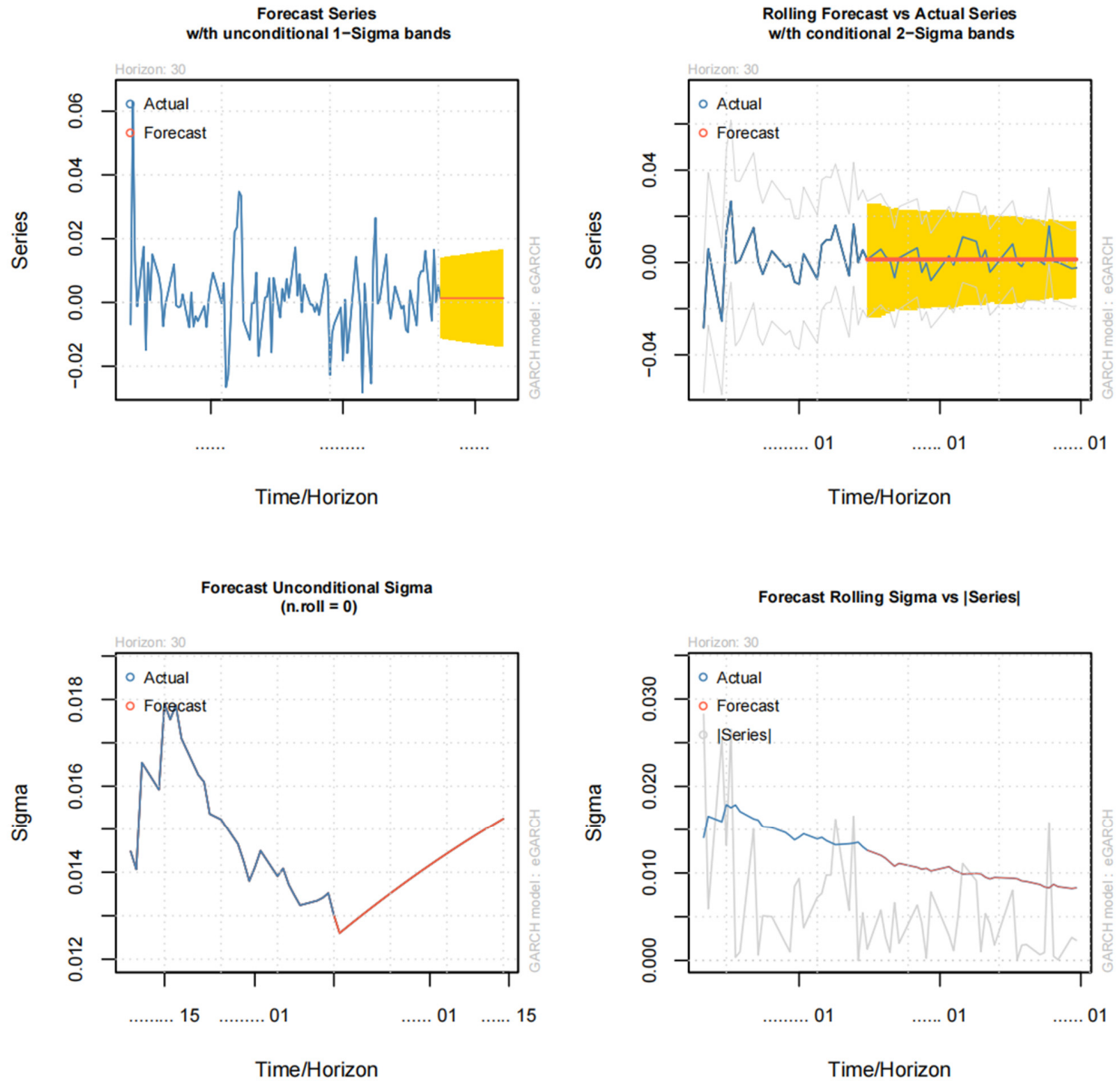

Fig. 14. forecasting graphs.