Time Series Analysis of Oil and Gold Prices

Contents

1 Introduction

Oil and natural gas are major industries in the energy market and play an influential role in the global economy as the world's primary fuel sources.Crude oil is used to make fuels needed for energy and transportation and to produce petrochemical products such as plastics. In 2020, approximately 4.2 billion metric tons of crude oil was produced worldwide. Oil prices can be very volatile due to fluxes in supply and demand and the fact that most oil is traded using futures contracts. These contracts fix prices for traders at a future date leading to either a deficit or profit based on the current market price ("Oil Industry Worldwide" 2021). The expectation that oil prices will invariably change leads to investors hedging their investments or at least diversifying their portfolios with more stable investments,like gold.
Gold is a precious metal used for making coinage, jewelry, various technologies for diverse industries like aerospace to dentistry, and for reserve investments by both private and public financial institutions. After World War II, under the Bretton Woods agreement, the most influential world currencies had their values attached to the value of gold- this was called the Gold Standard. Although this standard ended in 1971, most countries continue to invest in gold, and perceive it as an alternative to currency. As of 2020, it has been reported that around 53,000 metric tons of gold is in reserve worldwide. Investments in gold tend to increase during recessions. This persistence to consistently invest in gold all over the world has allowed gold prices to increase on average over time giving higher rates of return for most investors. Generally, gold is perceived as a more stable investment ("Gold Mining Worldwide" 2021).

In the modern economy, oil is just as important, if not more important, than gold. Everything depends on oil, it heats homes, runs cars and machinery, and produces electricity which in turn runs computers, the internet, entertainment, lights and HVAC. However, the price of oil can be extremely difficult to forecast accurately. Being able to effectively predict prices of such an essential, yet volatile, commodity would be beneficial to organizations and investors.
Also, another way to measure or predict inflation may be to compare one commodity with another. Being able to predict inflation would also be very helpful to investors. Theoretically, since both oil and gold are commodities, if it were truly democratic, inflation would

raise all prices equally. Thus the ratio of the price of gold to the price of oil would remain constant. So then, as the value of the dollar depreciated (due to an increased supply of dollars) the price of both gold and oil would increase in tandem.

2 Problem Statement

The purpose of this project is to examine and analyze time series data for gold and oil prices, and to compare various forecasting techniques.

3 Purpose Statement

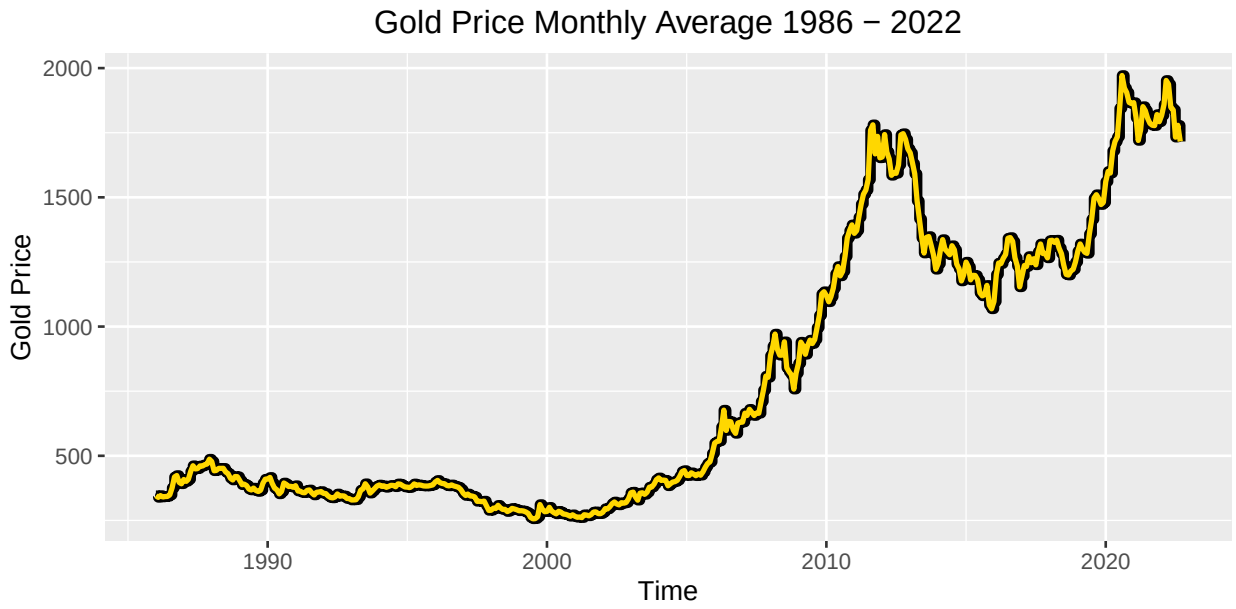
In this project, time series analysis will apply on the data sets of gold and oil prices separately. At the beginning, by Using acf and pacf functions, auto.arima and eacf, probable models will identify then model evaluation will apply to find the best fit for the data, at the end forecast techniques will use to forecast the prices of oil and gold.

4 Data Exploration

The gold price and oil price data were obtained from two different csv files. These files were procured from the website Macrotrends and Kaggle. Both data sets provide the daily closing value prices for each commodity starting on January 02, 1986 and ending on September 2022. The average monthly closing price for each dataset was calculated for each year and those averages were used for the time series analysis.

4.1 Gold

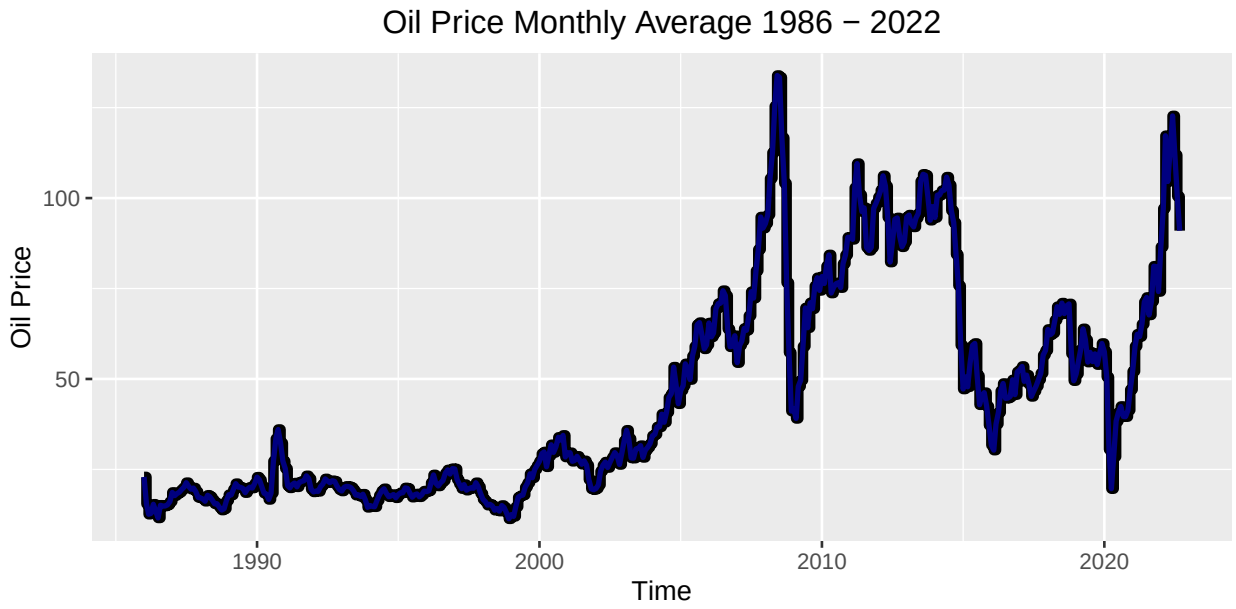
The Gold data set has 9284 records and the time series data starts from 1986-01-02 to 2022-09-02.It has 0 missing value. In order to do the analysis the average of monthly price is calculated and store in a new data frame called GoldPriceMonthly. The new data frame has 441 rows including the average of monthly price of gold since January 1986 to September 2022.
The first approach before starting to check the data in time series theories is plotting the data over the time to see the overall patterns and seasonality. In the plot below the monthly average of price of gold has shown over the time.



The time series plot above displays the average monthly gold closing prices from 1986 to 2022. Gold prices remained fairly steady, under $500, between 1985 until around 2005. A steep incline in gold prices started around 2005 until about 2012. Gold prices went from around $500 to $1,750 during this incline. Prices dropped around the year 2012, and then stayed fairly steady from 2013 to 2020. A steep rise occurred from 2020 to 2022 accompanied with some fluctuations where the highest gold price was close to $2,000.

4.2 Oil

The Oil data set has 9257 records and the time series data starts from 1986-01-02 to 2022-09-21.It has 0 missing value. In order to do the analysis, the average of monthly price is calculated and store in a new data frame called OilPriceMonthly. The new data frame has 441 rows including the average of monthly price of oil since January 1986 to September 2022.
Same as before, plotting the data will apply to see the overall patterns and seasonality. In the plot below the monthly average of price of oil has shown over the time.



The time series plot above displays the average monthly oil closing prices from 1986 to 2022. Oil prices remained fairly steady, except for a brief spike around 1991, from 1985 until around 1999. A steady incline in oil prices started around the year 2000, and then a major spike occurred around the year 2008 with an equally major decline around the year 2009. Oil prices again increased between the years 2009 to 2015. Another sharp decline occurred around 2015. Oil prices had smaller increases and decreases from 2015 to 2020. Then a large decrease in 2020 occurred followed by a large increase in the year 2021 and decrease after.

From 1985 until about 2000, gold and oil price fluctuations look fairly similar. Gold and oil prices both had significant increases in the early 2000s. Both commodities had major price declines from around 2015 to 2020. However, from the plots above, gold prices appear overall less volatile than oil prices.

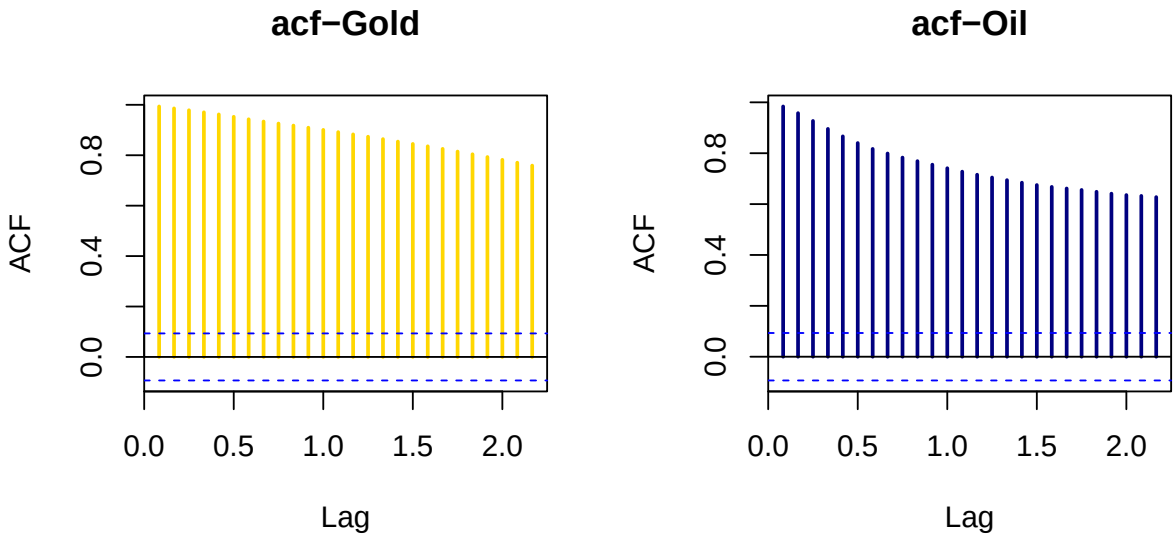
5 Model Identification

In this project, three ways will use to identify the time series models. The acf (Auto Correlation Function) and pacf(Partial Auto Correlation Function) will check at first to have a better understanding of the data, Then auto.arima and eacf functions will use in model identification.

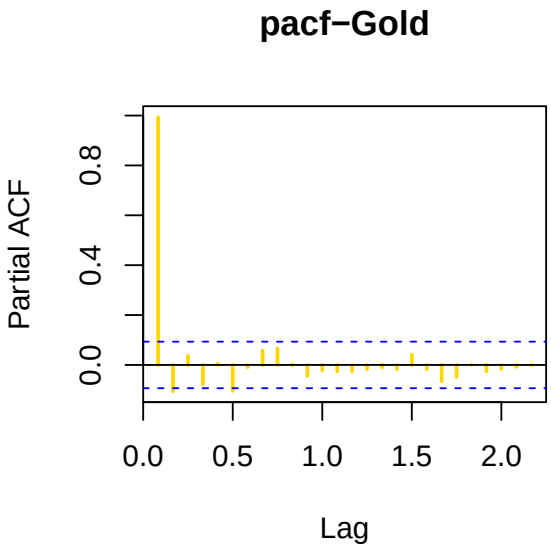
5.1 The acf and pacf Functions

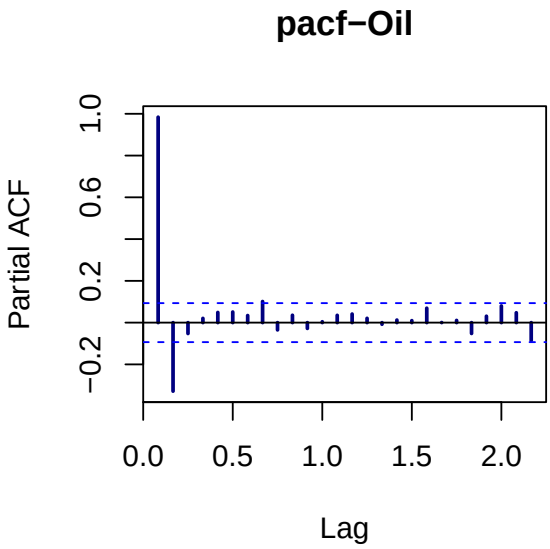
A common assumption in many time series techniques is that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure

do not change over time.
The autocorrelation function (ACF) reveals how the correlation between any two values of the signal changes as their separation changes. It is a time domain measure of the stochastic process memory, and does not reveal any information about the frequency content of the process(Mohamed Nounou, n.d.).

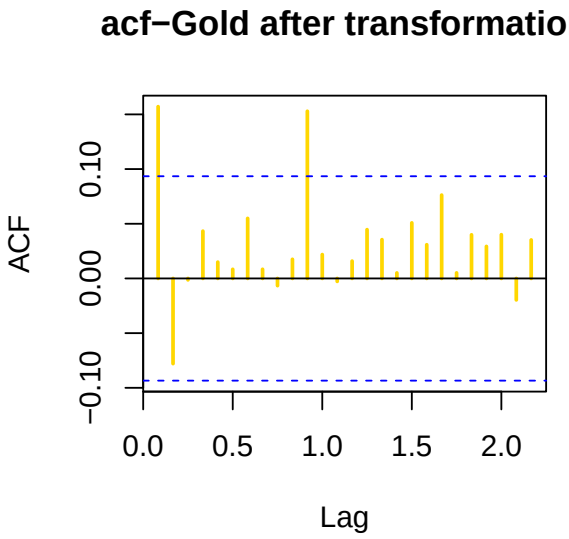


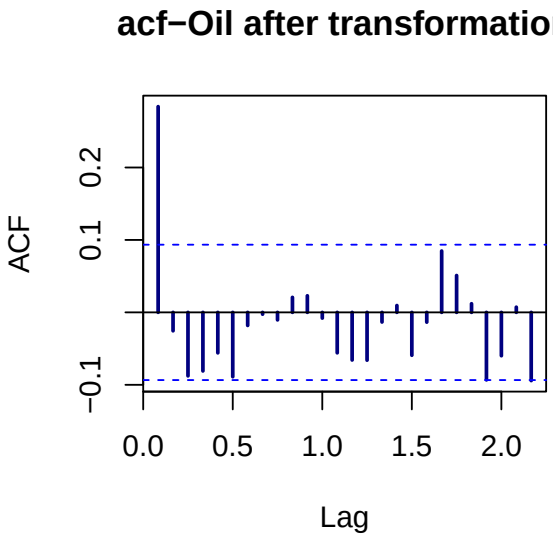
The acf plots above of both the oil and gold prices depict slowly declining or tapering off data. This indicates that both data sets are non-stationary. Both data sets need to be stationarized in order to continue the time series analysis.One way to make the data stationary is the log function, but before using log function, it is useful to check the pacf plot as well.
In time series analysis, the partial autocorrelation function gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags.This function plays an important role in data analysis aimed at identifying the extent of the lag in an autoregressive (AR) model. The use of this function was introduced as part of the Box–Jenkins approach to time series modelling, whereby plotting the partial autocorrelative functions one could determine the appropriate lags p in an $AR(p)$ model or in an extended $ARIMA(p, d, q)$ model(“Partial Autocorrelation Function,” n.d.).





The pacf plots for both gold and oil prices cuts off after lag one and two. The graphs above illustrate the possible **AR(1)** model for gold since the ACF plot tails off slowly and the PACF cuts off after the lag 1 while one of the possible models for oil is **AR(2)** since the ACF plot tails off slowly and the PACF cuts off after the lag 2. Another possible model is **ARIMA(1,1,0)** for gold and **ARIMA(2,1,0)** for oil because of applying differencing to make the data stationary.





In order to help stationarize the data, we transformed the average closing price data into a logarithmic scale and took the difference of the data. As a result the oil data looks stationary after log transformation since all of the spikes locate between two blue lines, However the acf plot for gold might need another log transformation because of another spike in between, and it might be the indication of seasonality in the data. Furthermore

auto.arima function will apply to see what is the recommended model by R. Hence these models has identified for further evaluation:

Gold	Oil
AR(1)	AR(2)
ARIMA(1,1,0)	ARIMA(2,1,0)

5.2 The auto.arima() Function

The $auto.arima()$ function in R uses a variation of the Hyndman-Khandakar algorithm, which combines unit root tests, minimisation of the AICc and MLE to obtain an ARIMA model. The arguments to auto.arima() provide for many variations on the algorithm (Khandakar, n.d.).
The result of auto.arima function on the gold data after applying log transformation is as follows:

```
Series: log(GoldPriceTs)
ARIMA(3,1,0)(1,0,0)[12] with drift

Coefficients:
         ar1      ar2     ar3     sar1   drift
      0.1789  -0.1119  0.0324  -0.0055  0.0036
s.e.  0.0485   0.0483  0.0480   0.0494  0.0018

sigma^2 = 0.001184:  log likelihood = 860.78
AIC=-1709.56   AICc=-1709.36   BIC=-1685.04
```

The suggested model for the gold data after differencing is **ARIMA(3,1,0)(1,0,0)[12]** which is illustrate a seasonality in the model and the auto.arima function identified P=1 for the seasonality. Thus the validity of the model will check in the model evalution section.
As well as gold data, the same function will apply for the oil data after log transformation, and the results are as follows:

```
Series: log(OilPriceTs)
ARIMA(2,1,1)(1,0,2)[12]

Coefficients:
         ar1      ar2      ma1     sar1    sma1     sma2
      1.1865  -0.3293  -0.8923  -0.0847  0.0861  -0.0496
s.e.  0.1207   0.0461   0.1230   0.7310  0.7326   0.0510

sigma^2 = 0.008155:  log likelihood = 436.54
AIC=-859.08   AICc=-858.82   BIC=-830.47
```

As it has shown above, the suggested model for the oil data after differencing is **SARIMA(2,1,1)(1,0,2)[12]** which is obviously is a seasonality in the model and the auto.arima function identified P=1 and Q=2 for the seasonality. Similarly the validity of the model will check in the model evalution section.

5.3 The eacf Function

The EACF stands for *Extended Auto Correlation Function* which allows for the identification of ARIMA models (differencing is not necessary). The quantlet generates a table of the extended (sample) autocorrelation function (EACF) for the time series y. You have to specify the maximal number of AR lags (p) and MA lags (q).
The eacf result for the gold data is as follow:

```
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  x
1 x o o o o o o o o o x  o  o  o
2 x o o o o o o o o o x  o  o  o
3 x o x o o o o o o o x  o  o  o
4 x x x o o o o o o o x  o  o  o
5 x x o x o o o o o o x  o  o  o
6 x o x o x o o o o o x  o  o  o
7 x x x x o o o o o o x  o  o  o
```

The result of eacf shows some sort of the seasonality in the data and perhaps the **SARIMA(1,1,0)(1,0,0)[10]** can be checked as the identify model in this section for further model evaluation. The eacf result for the oil data is as follow:

```
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  x
1 x o o o o o o o o o o  o  o  o
2 x o o o o o o o o o o  o  o  o
3 x o o 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 o o o o o o o o o  o  o  o
6 x x o o o o o o o o o  o  o  o
7 x o x o x o o o o o o  o  o  o
```

Also for the oil data, we might have the option of **AR(2)** as a possible model to check in model evaluation.

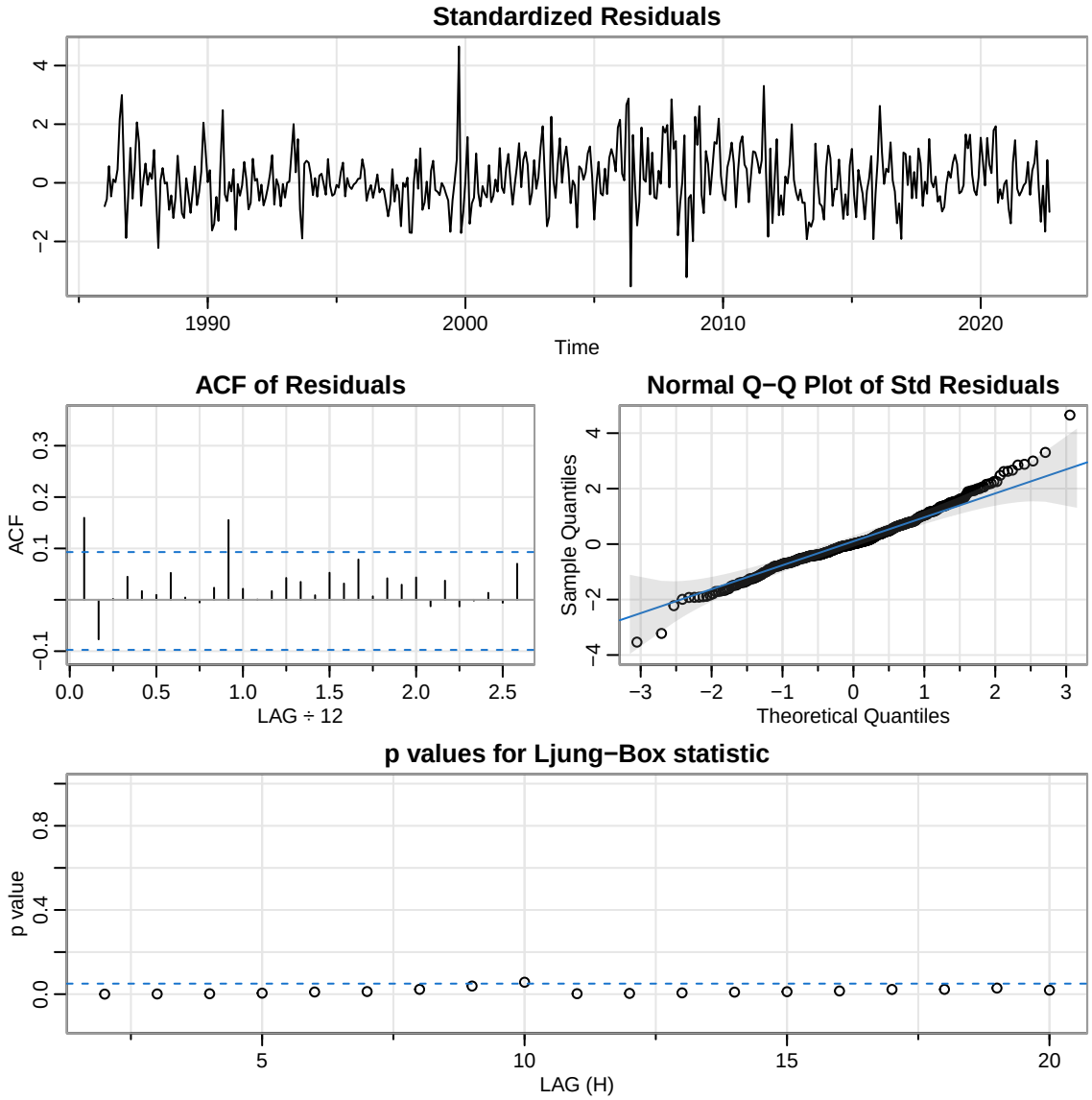
6 Model Evaluation

One of the R functions that helps in model evaluation is $sarima()$ function. By looking at the results of the sarima() function in the plots and comparing AIC , AICc and BIC, the best model will select and will use in forecasting.

6.1 Gold

During the model identification step, AR(1), ARIMA(1,1,0), SARIMA(3,1,0)(1,0,0)[12] and ARIMA(1,1,0)(1,0,0)[10] has identified as the possible models for gold data.In this section sarima() function will apply for the mentioned models and the results will compare.

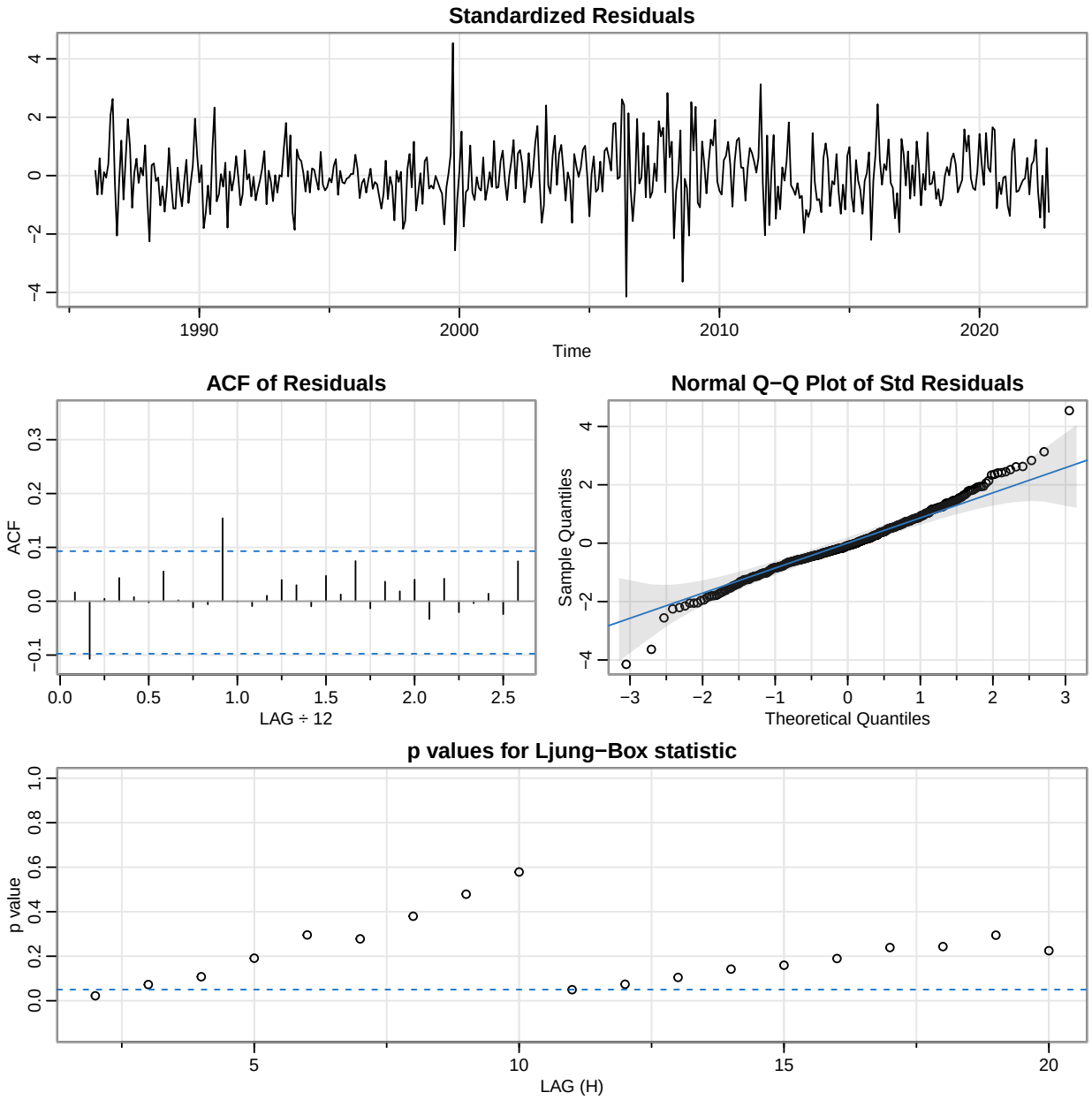
6.1.1 Model-1 : AR(1)



The standard residual plot has some spikes and it looks like a pattern so it does not like a white nose process that is stochastic. Similarly in the ACF of the residuals, some of the spikes passed the blue lines which is the indication of some sort of pattern in the residuals.

The Ljung–Box test is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the “overall” randomness based on a number of lags. In another word, If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit. As it has shown in the Ljung-Box plot, all of the p-values are equal or less than 0.05 which means the model exhibits lack of fit.

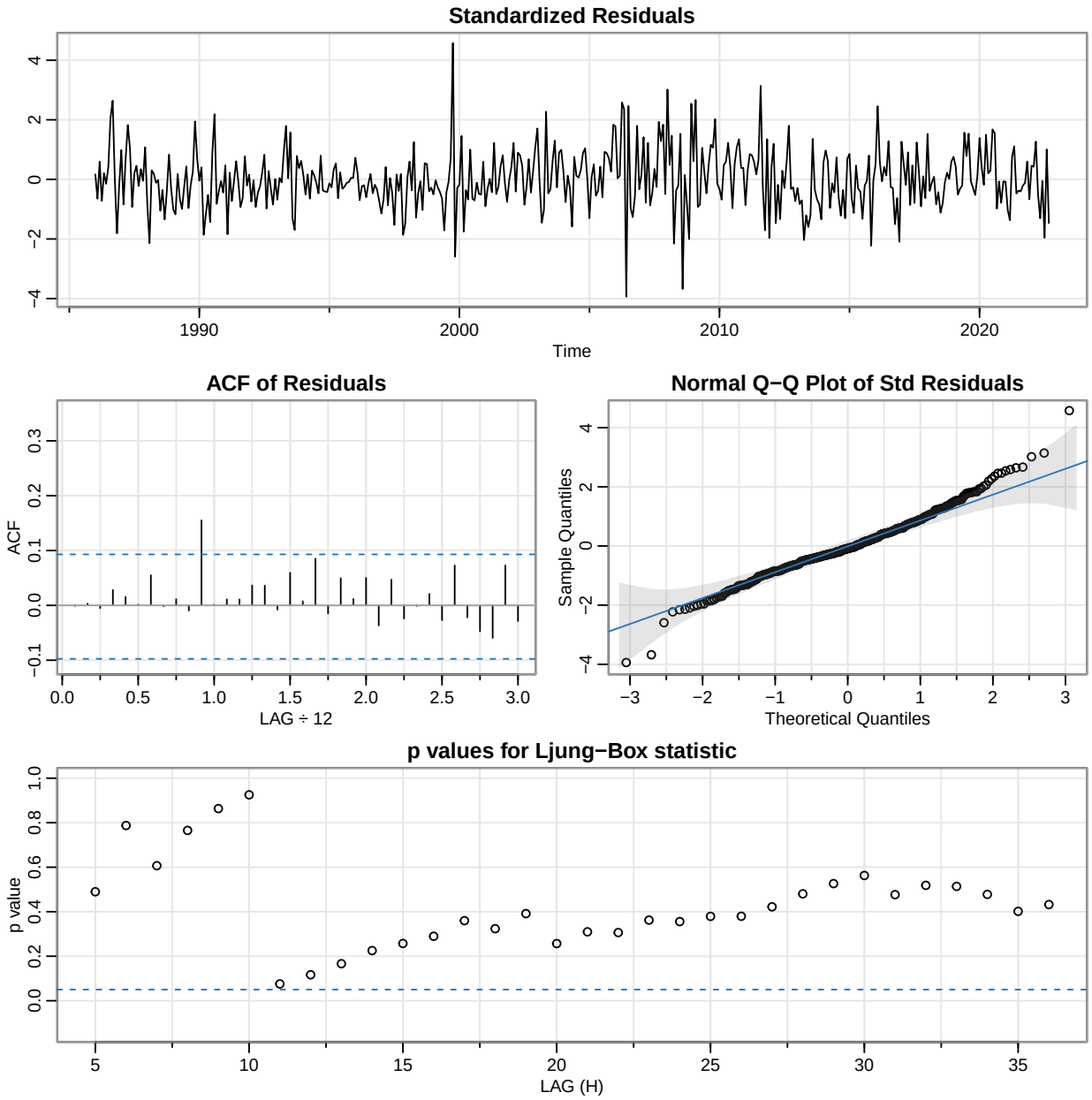
6.1.2 Model-2 : ARIMA(1,1,0)



Looking at the result of the ARIMA(1,1,0) shows improvement in the model since the residuals looks more to the white nose process and most of the p-values for Ljung-Box statistics plot are above the blue line which indicates the model does not exhibit lack of fit. Now other parameters need to compute and compare including AIC, AICc and BIC.

AIC	AICc	BIC
-3.89	-3.89	-3.86

6.1.3 Model-3 : SARIMA(3,1,0)(1,0,0)[12]

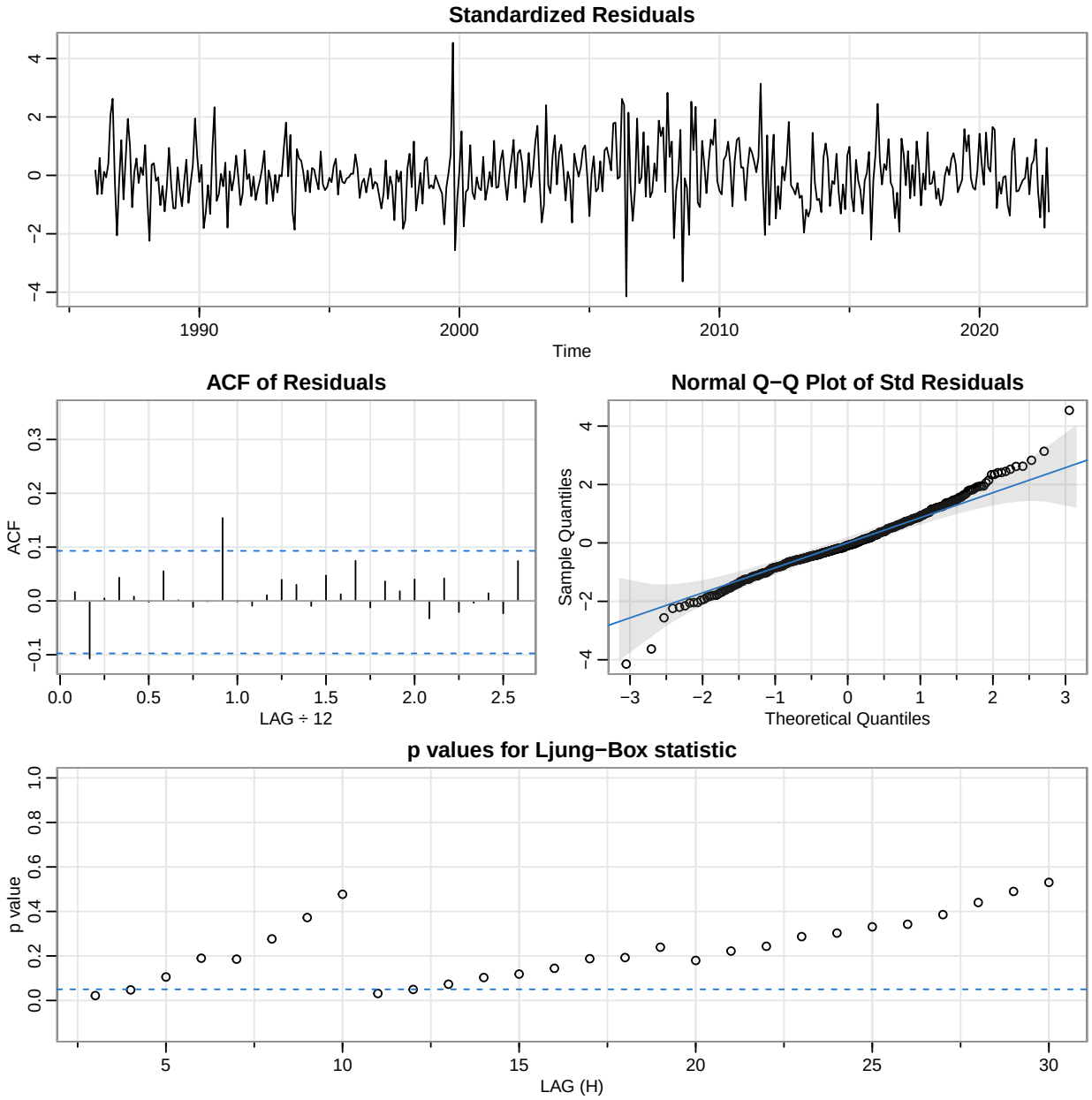


The third models shows more improvement in compare of the second model since all of the p-values of the Ljung-Box plot locate on top of the blue line which is the indication of not exhibition in the lack of fit in this model. Almost all of the ACF of residuals' spikes are between the two blue line and it is more similar to the stochastic process of white noise. Other criteria table for this model including AIC, AICc and BIC are as follows:

AIC	AICc	BIC
-3.89	-3.89	-3.83

The result of the AIC and AICc for the model-2 (ARIMA(1,1,0)) and model-3 (SARIMA(3,1,0)(1,0,0)[12]) are mostly similar but BIC for the model-3 is smaller which is the indication of better fit for that model.

6.1.4 Model-4 : SARIMA(1,1,0)(1,0,0)[10]

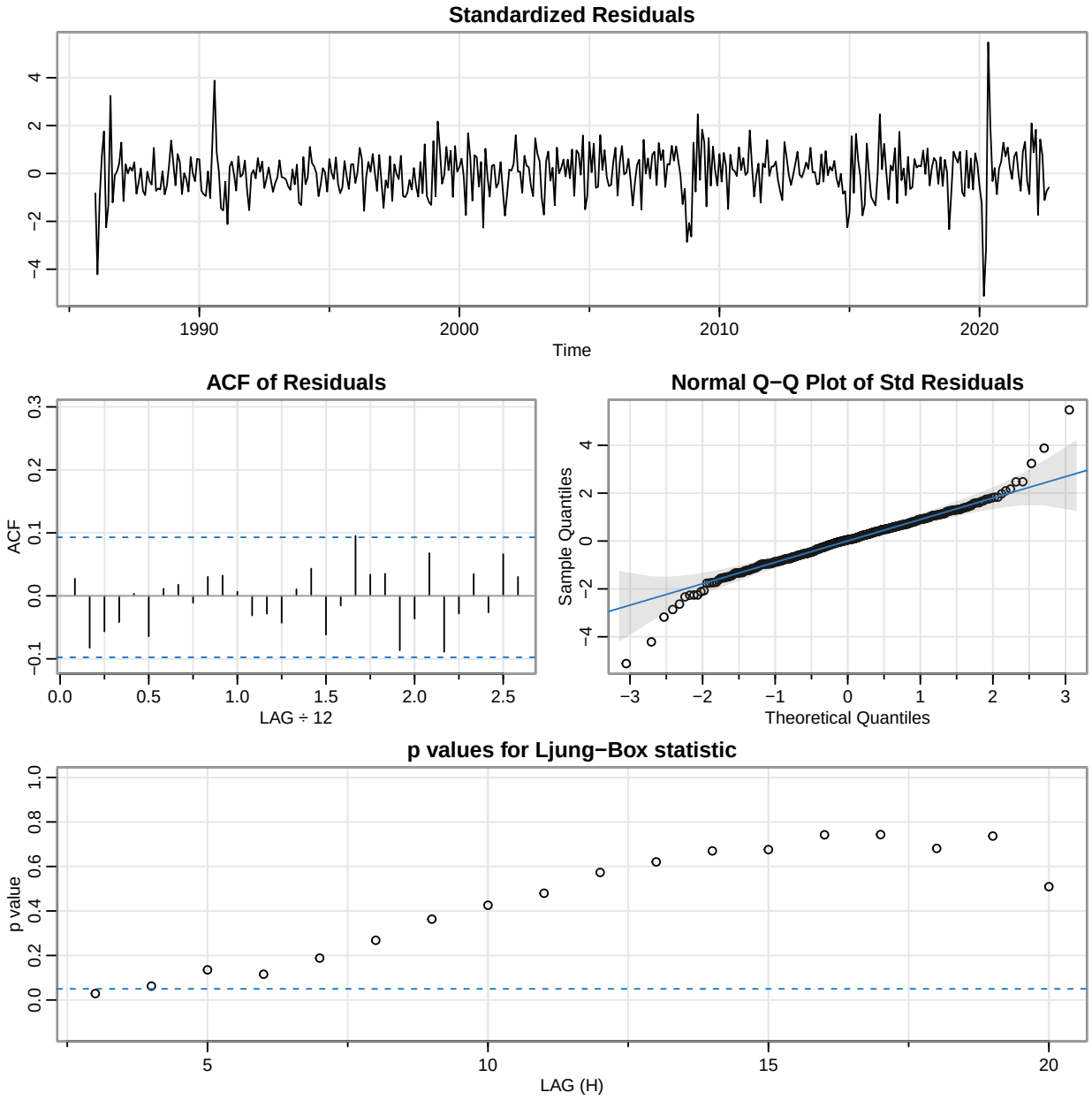


The results of the Ljung-Box plot for model-4 shows some of the p-values are less than blue line and in compare of model-3, that all of them were above the 0.05 level. Comparing the AIC and AICc between Model-3 and Model-4 shows some increase which is not in favor and some decrease for BIC. AIC is -3.88 , AICc is equal to -3.88 and BIC is -3.85.

6.2 Oil

During the model identification step, AR(2), ARIMA(2,1,0) and SARIMA(2,1,1)(1,0,2)[12], has identified as the possible models for oil data.In this section sarima() function will apply for the mentioned models and the results will compare.

6.2.1 Model-1 : AR(2)

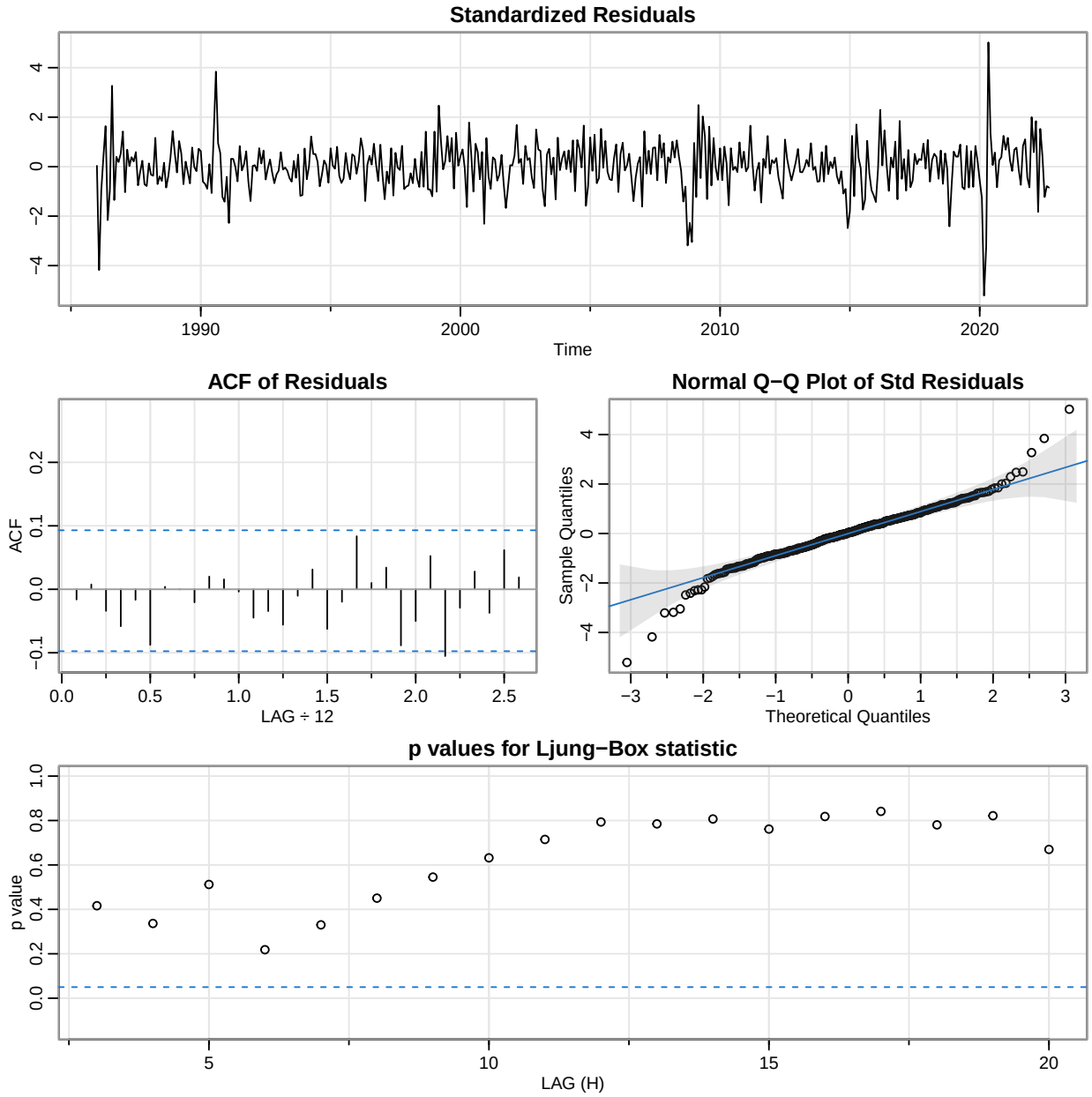


The result graphs of the sarima function that applied for evaluation of Model-1 are shown above. p-values were above the dotted line and residual values were mainly within adequate bounds. Very few outliers were noted on the Normal Q-Q plot. The important

criteria has shown in the table below:

AIC	AICc	BIC
-1.94	-1.94	-1.9

6.2.2 Model-2 : ARIMA(2,1,0)

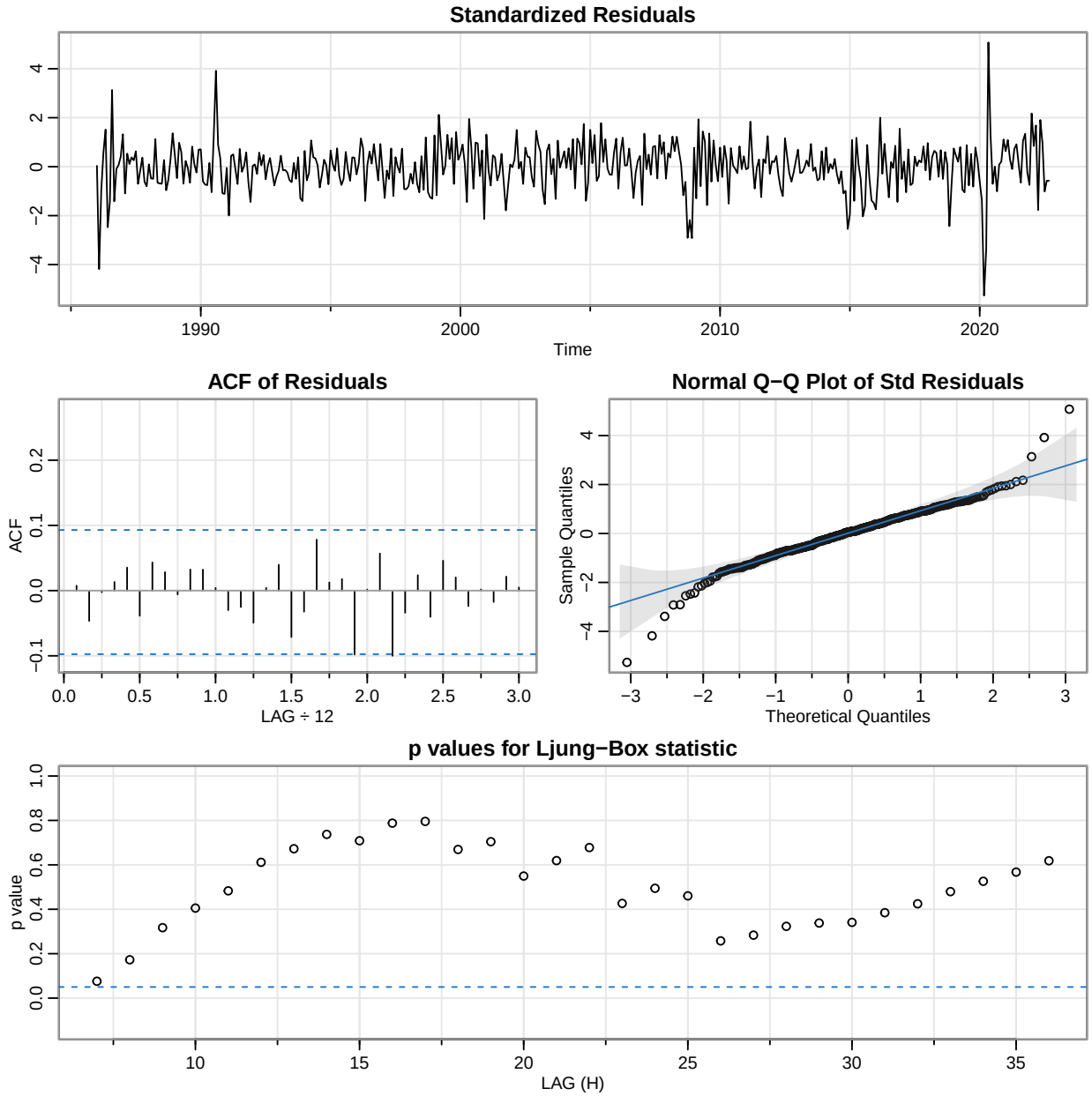


The second models shows more improvement in compare of the first model since all of the p-values of the Ljung-Box plot locate on top of the blue line which is the indication of not exhibition in the lack of fit in this model. All of the ACF of residuals' spikes are

between the two blue line and it is more similar to the stochastic process of white noise. Other important criteria table for this model including AIC, AICc and BIC are as follows:

AIC	AICc	BIC
-1.95	-1.95	-1.92

6.2.3 Model-3 : SARIMA(2,1,1)(1,0,2)[12]



The results of the Ljung-Box plot for model-3 shows all of the p-values are above the blue line, but the Normal Q-Q plot reveals some outliers. The important criteria values have shown in the table below:

AIC	AICc	BIC
-1.95	-1.95	-1.88

6.2.4 Comparing Models and Model Selection

6.2.4.1 Gold In comparison of the results of the AIC, AICc and BIC,it seems the Model-2 is a better fit for the gold data since the mentioned criteria is less than others and the p-values forLjung-Box plot are all above the blue line.Also with considering the principle of parsimony, the simpler model is better.

Gold	AIC	AICc	BIC
Model-1	-3.84	-3.84	-3.81
Model-2	-3.89	-3.89	-3.86
Model-3	-3.89	-3.89	-3.83
Model-4	-3.88	-3.88	-3.85

As a result, the model-2 (ARIMA(1,1,0)) is the better fit for the gold data and it will use to forecast the price of gold.

6.2.4.2 Oil Similarly the same comparison will apply between models that identified for the oil data. In the table below the AIC, AICc and BIC criteria for three models have shown. Since less value of those parameters are favorable, with considering the principle of parsimony, it can concluded the model-2 is the better fit for the oil data set.

Oil	AIC	AICc	BIC
Model-1	-1.94	-1.94	-1.9
Model-2	-1.95	-1.95	-1.92
Model-3	-1.95	-1.95	-1.88

So Model-2 which is ARIMA(2,1,0) is the winner model in the process of model selection and will use to forecast the oil price.

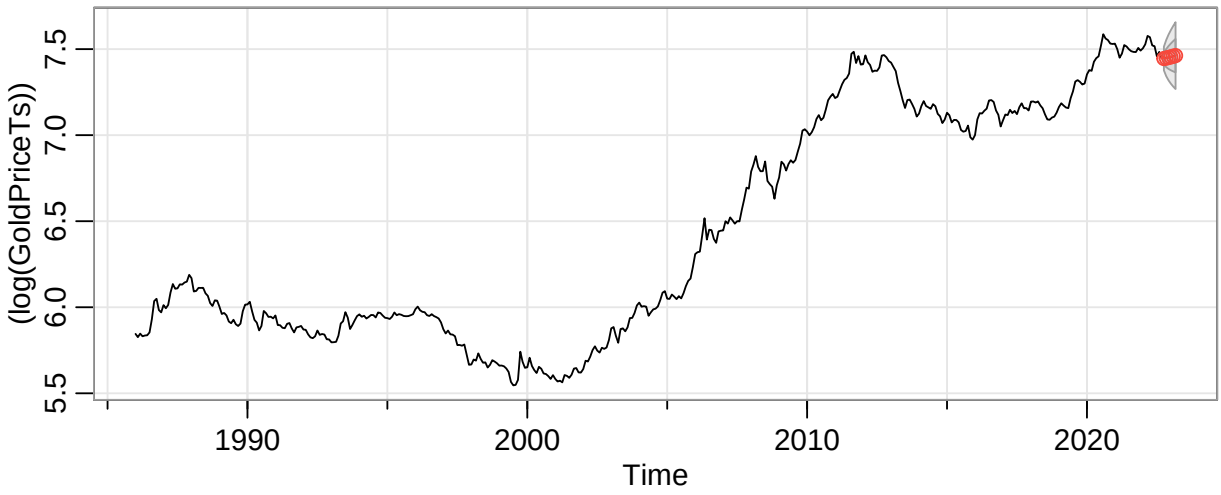
6.3 Forecasting

6.3.1 Using The Results of the SARIMA

In the previous section, model selection recognized the best fit for each data set. The best model for gold data is ARIMA(1,1,0) and for hte oil data set, ARIMA(2,1,0) is the best fit. Now the $sarima.for()$ function will use in forecasting the prices of gold and oil for the

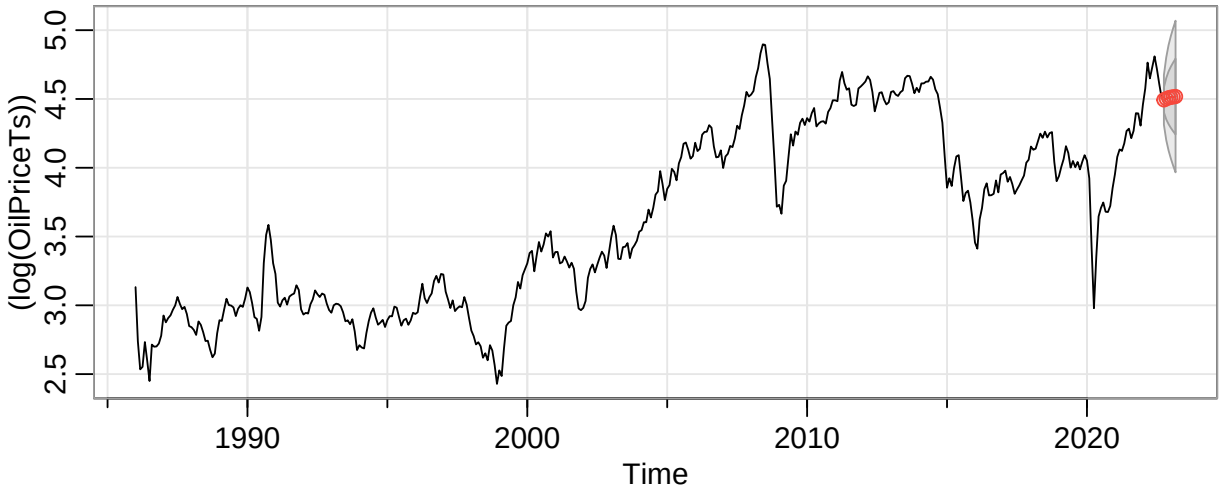
next 6 month, including the 95% confidence interval to estimate the upper and the lower prices.

6.3.1.1 Gold



Month	Lower	Predicted	Upper
Oct22	7.411	7.445	7.480
Nov22	7.395	7.448	7.501
Dec22	7.385	7.451	7.518
Jan23	7.377	7.455	7.533
Feb23	7.370	7.459	7.547
Mar23	7.365	7.462	7.559

6.3.1.2 Oil



Month	Lower	Predicted	Upper
Oct22	4.403	4.493	4.584
Nov22	4.352	4.502	4.652
Dec22	4.317	4.509	4.700
Jan23	4.289	4.512	4.736
Feb23	4.264	4.515	4.766
Mar23	4.242	4.518	4.793

6.3.2 Using ARFIMA

7 Conclusion

Oil and gold are both essential global commodities. Oil prices tend to fluctuate more greatly than gold prices over time. However, pricing predictions for both commodities varied greatly depending upon the type of forecasting technique utilized in the analysis. For gold prices, ARFIMA produced a more conservative prediction by including a larger range of values within the confidence intervals and by having a smaller mean value. SARIMA produced predictions with less variation, and the Linear Regression produced the highest mean predicted value. For oil prices, the Linear Regression model produced the more conservative prediction with the greatest range of values and the lowest mean value. SARIMA again provided the least amount of variation for the predicted values, but had the predicted highest mean value. ARFIMA had variation as well, but not as great as the Linear Regression model. Using different forecasting techniques provided greater insight into the commodity pricing data which allowed for better understanding of the data overall. This better understanding could lead to more effective decision making regarding future investing and commodity trading. https://www.macrotrends.net/

References

"Gold Mining Worldwide." 2021. https://www-statista-com.ezproxy.uhd.edu/study/12567/gold-statista-dossier.

Khandakar, Hyndman &. n.d. "Automatic Time Series Forecasting: The Forecast Package for r." https://www.jstatsoft.org/article/view/v027i03.

Mohamed Nounou, Bhavik Baksh. n.d. "Walvets in Chemistry." https://www.sciencedirect.com/topics/chemistry/autocorrelation-function.

"Oil Industry Worldwide." 2021. https://www-statista-com.ezproxy.uhd.edu/study/10750/global-oil-industry-and-market-statista-dossier.

"Partial Autocorrelation Function." n.d. https://en.wikipedia.org/wiki/Partial_autocorrelation_function.