Predictive Modeling of Monthly Oil Price Using Multiplicative Seasonal ARIMA Models

Introduction to Time Series Analysis

Final project

By

Soodabeh Ramezani

Instructor

Dr. Dexter Cahoy

Introduction and problem statement

The data under study, available in reference [1], consists of monthly WTI (west Texas intermediate) crude oil price from year 1986 to 2014 as listed below and plotted in Figure 1. The data shows several short- and long-term variations with potential weak seasonality as oil price relatively varies from summer to winter. The objective of this project is to explore and analyze the data and to implement several methods to adequately model the data. While we understand that oil price is a strong function of several economical, technological, geo-political, and environmental variables, it is particularly important to investigate possibility of a predictive model to fit the data and to perform data forecasting. In this project, six models will be analyzed and best fitted models will be used to forecast WTI crude oil price for a duration of 24 months.

22.93	15.46	12.61	12.84	15.38	13.43	11.59	15.1	14.87	14.9	15.22	16.11	18.65
17.75	18.3	18.68	19.44	20.07	21.34	20.31	19.53	19.86	18.85	17.28	17.13	16.8
16.2	17.86	17.42	16.53	15.5	15.52	14.54	13.77	14.14	16.38	18.02	17.94	19.48
21.07	20.12	20.05	19.78	18.58	19.59	20.1	19.86	21.1	22.86	22.11	20.39	18.43
18.2	16.7	18.45	27.31	33.51	36.04	32.33	27.28	25.23	20.48	19.9	20.83	21.23
20.19	21.4	21.69	21.89	23.23	22.46	19.5	18.79	19.01	18.92	20.23	20.98	22.39
21.78	21.34	21.88	21.69	20.34	19.41	19.03	20.09	20.32	20.25	19.95	19.09	17.89
18.01	17.5	18.15	16.61	14.52	15.03	14.78	14.68	16.42	17.89	19.06	19.66	18.38
17.45	17.72	18.07	17.16	18.04	18.57	18.54	19.9	19.74	18.45	17.33	18.02	18.23
17.43	17.99	19.03	18.86	19.09	21.33	23.5	21.17	20.42	21.3	21.9	23.97	24.88
23.71	25.23	25.13	22.18	20.97	19.7	20.82	19.26	19.66	19.95	19.8	21.33	20.19
18.33	16.72	16.06	15.12	15.35	14.91	13.72	14.17	13.47	15.03	14.46	13	11.35
12.52	12.01	14.68	17.31	17.72	17.92	20.1	21.28	23.8	22.69	25	26.1	27.26
29.37	29.84	25.72	28.79	31.82	29.7	31.26	33.88	33.11	34.42	28.44	29.59	29.61
27.25	27.49	28.63	27.6	26.43	27.37	26.2	22.17	19.64	19.39	19.72	20.72	24.53
26.18	27.04	25.52	26.97	28.39	29.66	28.84	26.35	29.46	32.95	35.83	33.51	28.17
28.11	30.66	30.76	31.57	28.31	30.34	31.11	32.13	34.31	34.69	36.74	36.75	40.28
38.03	40.78	44.9	45.94	53.28	48.47	43.15	46.84	48.15	54.19	52.98	49.83	56.35
59	64.99	65.59	62.26	58.32	59.41	65.49	61.63	62.69	69.44	70.84	70.95	74.41
73.04	63.8	58.89	59.08	61.96	54.51	59.28	60.44	63.98	63.46	67.49	74.12	72.36
79.92	85.8	94.77	91.69	92.97	95.39	105.45	112.58	125.4	133.88	133.37	116.67	104.11
76.61	57.31	41.12	41.71	39.09	47.94	49.65	59.03	69.64	64.15	71.05	69.41	75.72
77.99	74.47	78.33	76.39	81.2	84.29	73.74	75.34	76.32	76.6	75.24	81.89	84.25
89.15	89.17	88.58	102.86	109.53	100.9	96.26	97.3	86.33	85.52	86.32	97.16	98.56
100.27	102.2	106.16	103.32	94.66	82.3	87.9	94.13	94.51	89.49	86.53	87.86	94.76
95.31	92.94	92.02	94.51	95.77	104.67	106.57	106.29	100.54	93.86	97.63	94.62	

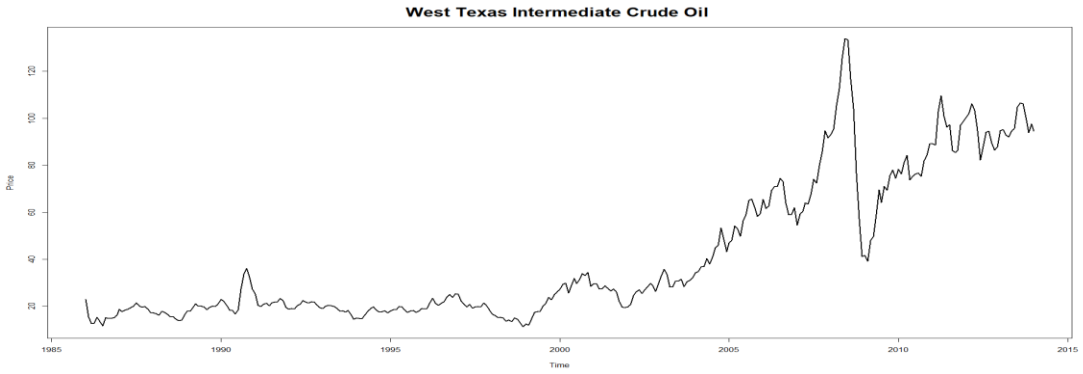


Figure 1: Monthly WTI crude oil price from 1986 to 2014

Data transformation

A close look at the ACF of the time series (Figure 2) shows that autocorrelation is significantly larger than zero for several lag times thus implying non-stattionaroty of the data. In addition, the data shows a relativeely wide range spanning from11.59 to 133.88. Therefore, data was transfomed to logarithmic space to narrow its range. Subsequently, a second order differecing (between the data and its first lag) was applied to make the data stationary. Figure 3 shows the time series after applying frist- and second-order differencing. ACF and PACF of the stationary signal is shown in Figure 4 confirming stationarity as the ACF cuts off after lag 1 or 2 and PACF tapers down [2]. In addition, ACF and PACF in Figure 4 show a

relatively weak seasonality with a period of 12 months which corresponds to slight variation of oil price from warm to cold seanons. Our anaylses showed that an AR(1) plue MA(1) model is sufficient to properly replicate seasonality. Seasonal variaiton of WTI oil price is also shown in Figure 5.

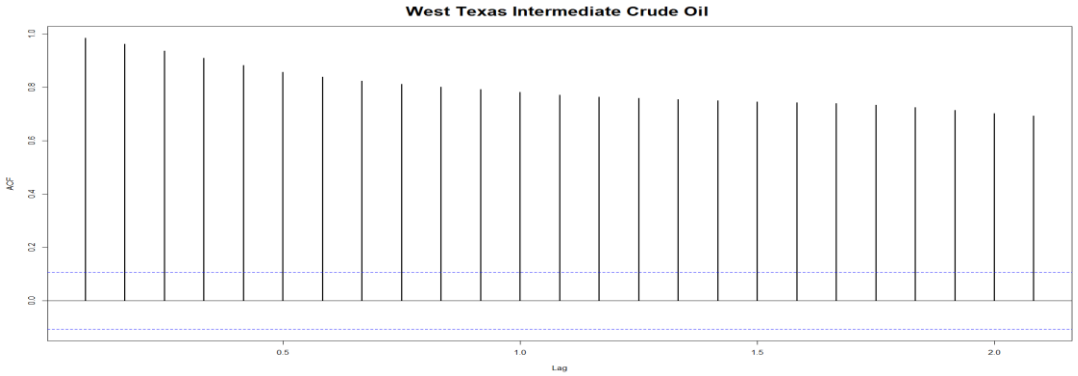


Figure 2: Sample ACF of the monthly WTI crude oil price implying non-stationarity

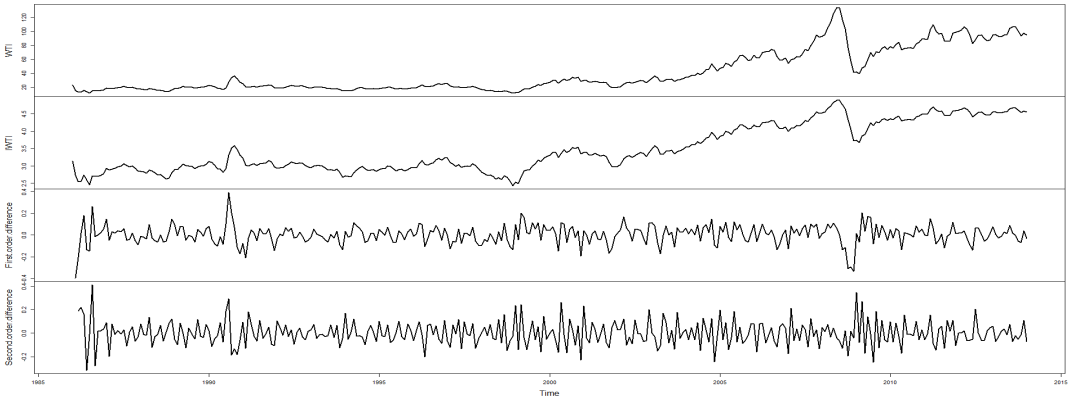


Figure 3: Data decomposition including first- and second-order-differenced data

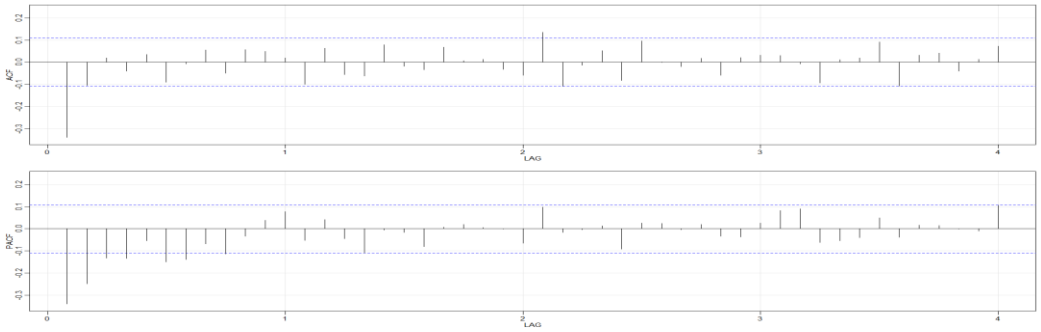


Figure 4: Sample ACF and PACF of the stationary data. ACF is cutting off while PACF is tailing off

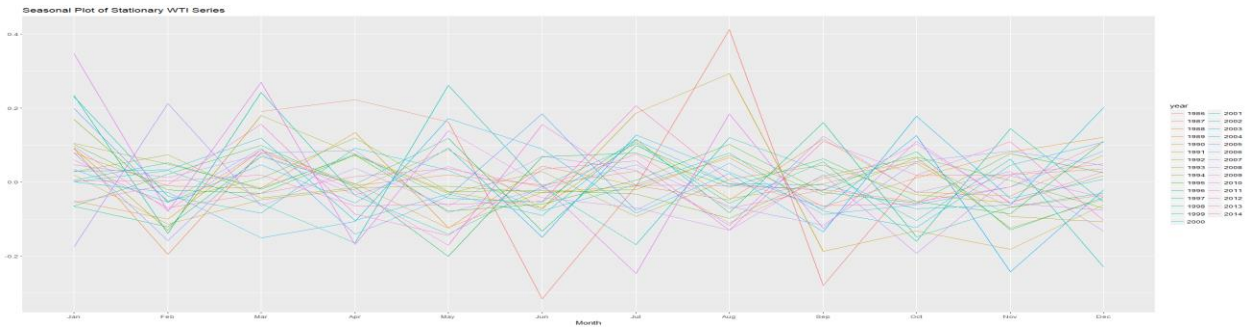


Figure 5: Seasonal behavior of WTI crude oil price

Estimation methods for ARMA models and their orders

Analysis of extended ACF matrix shown in Figure 6 suggests that ARMA(0,1), ARMA(1,0), and ARMA(1,1) are three appropriate models to honor auto-regressive moving-average characteristic of the time series under study. ARMA(0,1) model is also consistent with the cutting-off behavior of ACF after lag 1 as shown in Figure 4. In the following sections, we will show the results of these three models. Subsequently, based on statistical analyses of the results (e.g., AIC, BIC, Q-Q statistics) ARMA(1,1) model will be chosen as the best fit to data and will be used for data forecasting. To show the possible range of forecasted data, three additional estimation methods will be invoked to simulate and then forecast the data. In the first method, we fit the data using a polynomial of order 4 under the assumption of an auto-regressive-moving-average behavior of order (1,1) for the residuals. The second method uses MLE (maximum likelihood estimator) approach with no seasonality in ARIMA code. The third model applies Yule-Walker estimation method to fit the data and subsequently forecast the data.

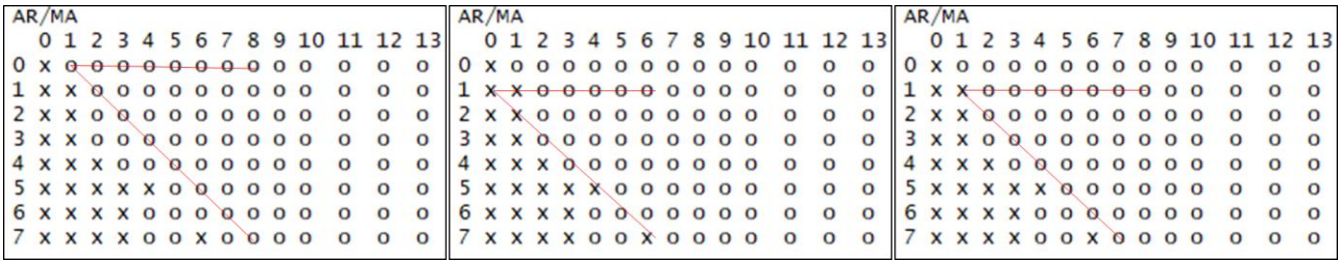
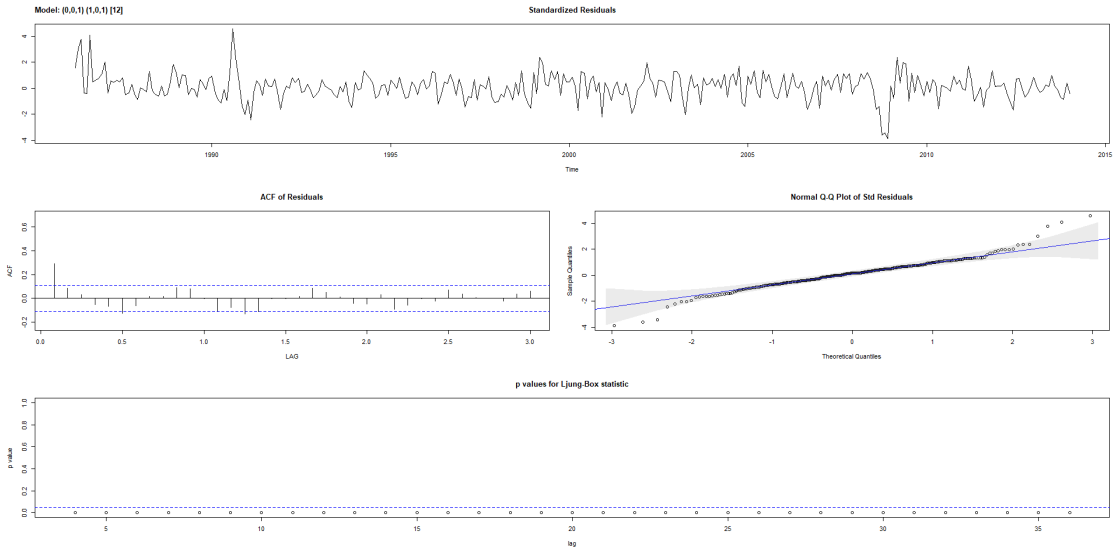


Figure 6: Extended ACF matrix and the estimated ARMA orders: left panel implies MA(1), middle panel implies AR(1), and right panel implies ARMA(1,1)

SARIMA models

As discussed previously, three models (i.e., ARMA(0,1), ARMA(1,0),and ARMA(1,1)) were implemented to simulate and fit the data. Model estimation was performed using SARIMA code in R. WTI crude price data in logarithmic space is used as data input. Subsequently, non-seasonal and seasonal differencing orders are assumed as 2 and 0, respectively, to ensure stationarity as stated before. Figures 7 through 9 demonstrate the quality of the fit for the three models. A comparison of AIC, BIC, and p-values for Ljung Box statistic test suggests that Model 3 achieves the best match as AIC and BIC values are the largest in absolute value and p-values are all higher than the significant level. Therefore, we use Model three for subsequent forecasting.

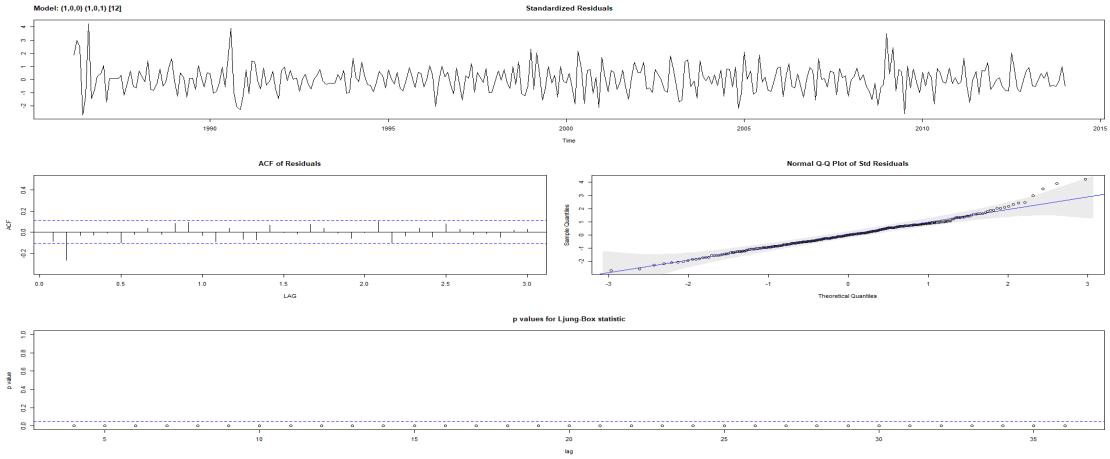



```
sarima(log(WTI), 0, 2, 1, 1, 0, 1, 12, no.constant = TRUE)
```

Figure 7: Residual analysis for Model 1 (i.e. ARIMA(0,2,1)×(1,0,1)$_{12}$) fit to WTI oil price data

The output of the SARIMA code for Model 1 and the estimated coefficients are shown below:

```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = xmean, include.mean = FALSE, optim.control = list(trace = trc,
##     REPORT = 1, reltol = tol))
##
## Coefficients:
## Warning in sqrt(diag(x$var.coef)): NaNs produced
##          ma1     sar1    sma1
##       -1.0000  -0.0044  0.0018
## s.e.   0.0121      NaN     NaN
##
## sigma^2 estimated as 0.007388:  log likelihood = 343.82,  aic = -679.64
##
## $degrees_of_freedom
## [1] 332
##
## $ttable
##      Estimate     SE  t.value p.value
## ma1   -1.0000 0.0121 -82.5527       0
## sar1  -0.0044    NaN      NaN     NaN
## sma1   0.0018    NaN      NaN     NaN
##
## $AIC
## [1] -3.889988
##
## $AICc
## [1] -3.883656
##
## $BIC
## [1] -4.855832
```

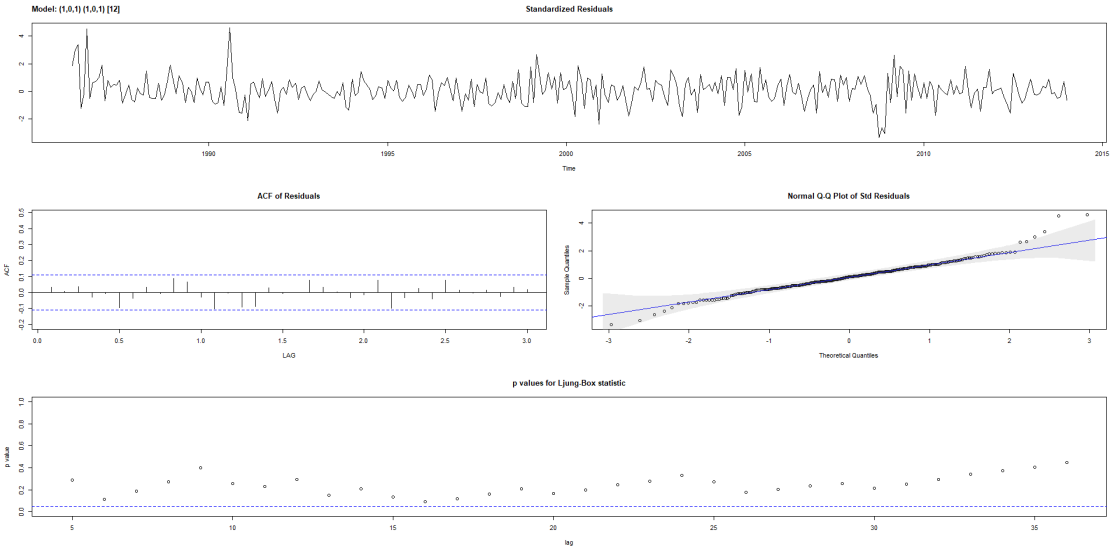



```
sarima(log(WTI), 1, 2, 0, 1, 0, 1, 12, no.constant = TRUE)
```

Figure 8: Residual analysis for Model 2 (i.e. ARIMA(1,2,0)×(1,0,1)$_{12}$) fit to WTI oil price data

The output of the SARIMA code for Model 2 and the estimated coefficients are shown below:

```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = xmean, include.mean = FALSE, optim.control = list(trace = trc,
##     REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     sar1    sma1
##       -0.3405  -0.8754  0.9193
## s.e.   0.0516   0.1167  0.1018
##
## sigma^2 estimated as 0.009183:  log likelihood = 309.92,  aic = -611.84
##
## $degrees_of_freedom
## [1] 332
##
## $ttable
##      Estimate     SE t.value p.value
## ar1   -0.3405 0.0516 -6.5927       0
## sar1  -0.8754 0.1167 -7.5017       0
## sma1   0.9193 0.1018  9.0327       0
##
## $AIC
## [1] -3.672437
##
## $AICc
## [1] -3.666105
##
## $BIC
## [1] -4.63828
```




```
sarima(log(WTI), 1, 2, 1, 1, 0, 1, 12, no.constant = TRUE)
```

Figure 9: Residual analysis for Model 3 (i.e. ARIMA(1,2,1)×(1,0,1)$_{12}$) fit to WTI oil price data

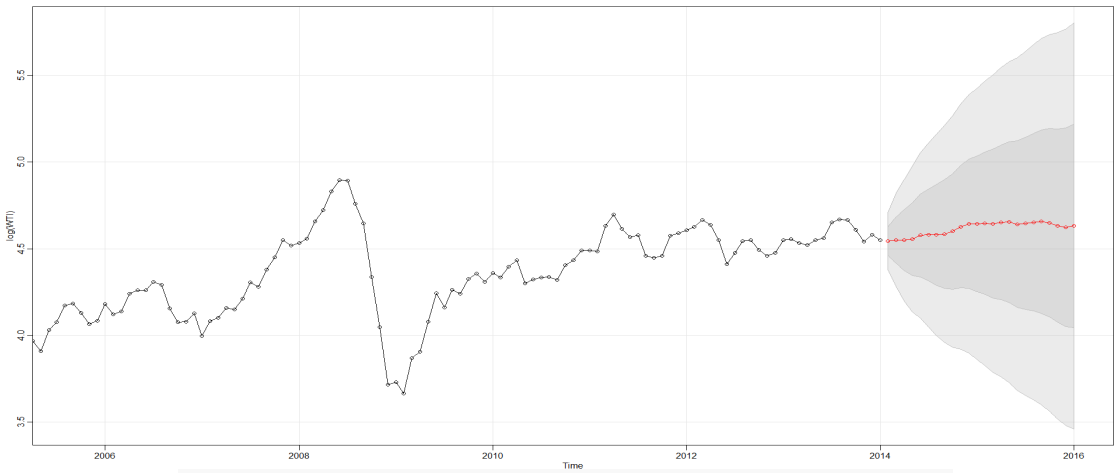
The output of the SARIMA code for Model 3 and the estimated coefficients are shown below:

```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = xmean, include.mean = FALSE, optim.control = list(trace = trc,
##     REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1      ma1     sar1    sma1
##       0.2799  -1.0000  -0.9006  0.9541
## s.e.  0.0544   0.0099   0.0960  0.0879
##
## sigma^2 estimated as 0.006767:  log likelihood = 357.96,  aic = -705.91
##
## $degrees_of_freedom
## [1] 331
##
## $ttable
##      Estimate     SE   t.value p.value
## ar1    0.2799 0.0544    5.1424       0
## ma1   -1.0000 0.0099 -101.0527       0
## sar1  -0.9006 0.0960   -9.3822       0
## sma1   0.9541 0.0879   10.8512       0
##
## $AIC
## [1] -3.971755
##
## $AICc
## [1] -3.96524
##
## $BIC
## [1] -4.926213
```

We have also shown the outputs of “auto.arima” code to automatically estimate a SARIMA model. The results show a SARIMA(2,1,2,1,0,0) with drift. This step is considered optional and the estimated model won’t be used in any forecasting.

```
auto.arima(lWTI)
```

```
## Series: lWTI
## SARIMA(2,1,2)(1,0,0)[12] with drift
##
## Coefficients:
##          ar1      ar2      ma1     ma2    sar1   drift
##       1.3374  -0.4386  -1.0875  0.1372  0.0211  0.0050
## s.e.  0.1868   0.1782   0.2064  0.2018  0.0604  0.0023
##
## sigma^2 estimated as 0.006741:  log likelihood=365.97
## AIC=-717.95   AICc=-717.6   BIC=-691.23
```




```
sarima.for(log(WTI), 24, 1, 2, 1, 1, 0, 1, 12, no.constant = TRUE)
```

Figure 10: Data forecasting for a 24-month duration using Model 3 including uncertainty ranges

The output of the SARIMA code for forecasting using Model 3 is shown below.

```
## $pred
##             Jan      Feb      Mar      Apr      May      Jun      Jul
## 2014          4.543929 4.550686 4.549536 4.555821 4.577306 4.580565
## 2015 4.644242 4.646719 4.645038 4.652515 4.653866 4.641687 4.645967
## 2016 4.632008
##             Aug      Sep      Oct      Nov      Dec
## 2014 4.580349 4.584970 4.600443 4.627486 4.644919
## 2015 4.653391 4.656460 4.649757 4.632633 4.624165
## 2016
##
## $se
##              Jan         Feb         Mar         Apr         May         Jun
## 2014          0.08242683 0.13407018 0.17502173 0.20922745 0.23898413
## 2015 0.39110251 0.40988419 0.42828949 0.44612552 0.46339280 0.48013498
## 2016 0.58610109
##              Jul         Aug         Sep         Oct         Nov         Dec
## 2014 0.26563406 0.28999458 0.31259088 0.33377865 0.35381021 0.37287138
## 2015 0.49640038 0.51223312 0.52767180 0.54275000 0.55749710 0.57193927
## 2016
```

The expression of ARIMA(1,2,1)×(1,0,1)$_{12}$ model is as follows:

$$x_t = \nabla\nabla(log(WTI))$$

$$x_t = 0.28x_{t-1} - 0.9x_{t-12} - 0.25x_{t-13} + w_t - w_{t-1} + 0.95w_{t-12} - .95w_{t-13}$$

Modeling using regression with auto-correlated errors

In this section, we explore the capability of trendline fitting under the assumption of auto-correlated errors. Once an adequate model is fitted to the WTI oil price data, the auto-correlation function of residuals is analyzed to establish an auto-regressive-moving-average model for the residuals. The final fitted model will be the summation of trendline and residuals. For this study, we tried multiple trendlines and concluded that a polynomial of order 4 (in the form of $a + bt + ct^2 + dt^4$ where t denotes to time) gives rise to a reasonable match as shown in Figure 11. It should be noted that coefficient d seems to be very small; however, the omission of that term resulted in a non-satisfactory fit. Therefore, we decided to keep that term for the regression calculation. The ACF and PACF of the residuals shown in Figure 12 imply that both ACF and PACF functions tailing off hence an ARMA(1,1) model can be established for the residuals. The simulated time series using the trendline (of fourth order polynomial) and errors (of ARMA(1,1) order) are shown in Figure 13. The outputs from "sarima" confirms whiteness of the final residuals. In addition, the p-values in Ljung Box statistic test show significantly high values. This fitted model is subsequently used for 24-month forecasting as plotted in Figure 14. The results show that the assumed trendline (i.e., polynomial of order four) highly influences the future prediction of oil price. In other words, the prediction will be dominated by the assumed trendline with an uncertainty range defined by the auto-correlated error. The R code and the associated outputs including quality of trendline fit are shown below:

```r
library(astsa)
par(mfrow=c(2,1))
ts.plot(lWTI,ylab="log(WTI)",lwd=2,cex.lab=2,cex.axis=2)
tm=time(lWTI)-mean(time(lWTI))
tm2=tm^2
tm4=tm^4
summary(fit<- lm(lWTI~tm1+tm2+tm4, na.action = NULL))
lines(fitted(fit), col="red", lwd=3)
acf2(resid(fit),52, lwd=3, cex.lab=2,cex.axis=2, main = "residuals")[0]

##
## Call:
## lm(formula = lWTI ~ tm + tm2 + tm4, na.action = NULL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69785 -0.11557 -0.01449  0.10838  0.62917
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.108e+00  2.136e-02  145.50   <2e-16 ***
## tm           7.263e-02  1.405e-03   51.69   <2e-16 ***
## tm2           1.095e-02  6.783e-04   16.14   <2e-16 ***
## tm4          -4.514e-05  3.846e-06  -11.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2091 on 333 degrees of freedom
## Multiple R-squared:  0.9031, Adjusted R-squared:  0.9022
## F-statistic:  1035 on 3 and 333 DF,  p-value: < 2.2e-16
```

The outputs of “sarima” code including coefficients of the estimated model are shown below:

```r
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = xreg, optim.control = list(trace = trc, REPORT = 1,
##     reltol = tol))
##
## Coefficients:
##          ar1     ma1  intercept      tm      tm2  tm4
##       0.8826  0.3104     3.1379  0.0693  0.0090   0
## s.e.  0.0265  0.0528     0.0848  0.0057  0.0015   0
##
## sigma^2 estimated as 0.006473:  log likelihood = 370.03,  aic = -726.06
##
## $degrees_of_freedom
## [1] 331
##
## $ttable
##           Estimate     SE t.value p.value
## ar1         0.8826 0.0265 33.2806   0e+00
## ma1         0.3104 0.0528  5.8755   0e+00
## intercept   3.1379 0.0848 36.9940   0e+00
## tm          0.0693 0.0057 12.1921   0e+00
## tm2          0.0090 0.0015  5.9600   0e+00
## tm4          0.0000 0.0000 -3.5677   4e-04
##
## $AIC
## [1] -4.004515
##
## $AICc
## [1] -3.99757
##
## $BIC
## [1] -4.936502
```

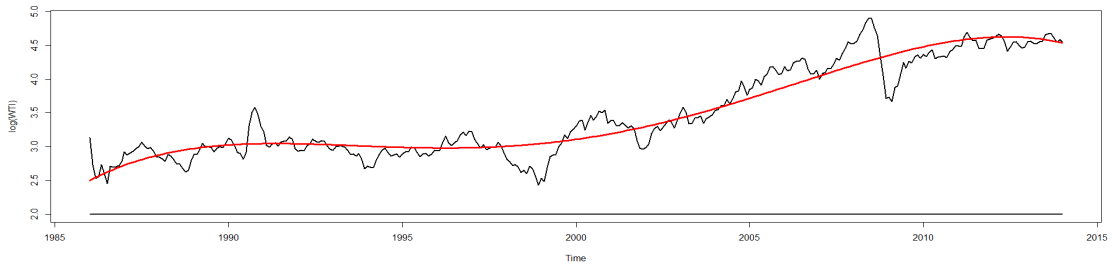


Figure 11: WTI oil price data shown with the fitted trendline using a polynomial of order 4

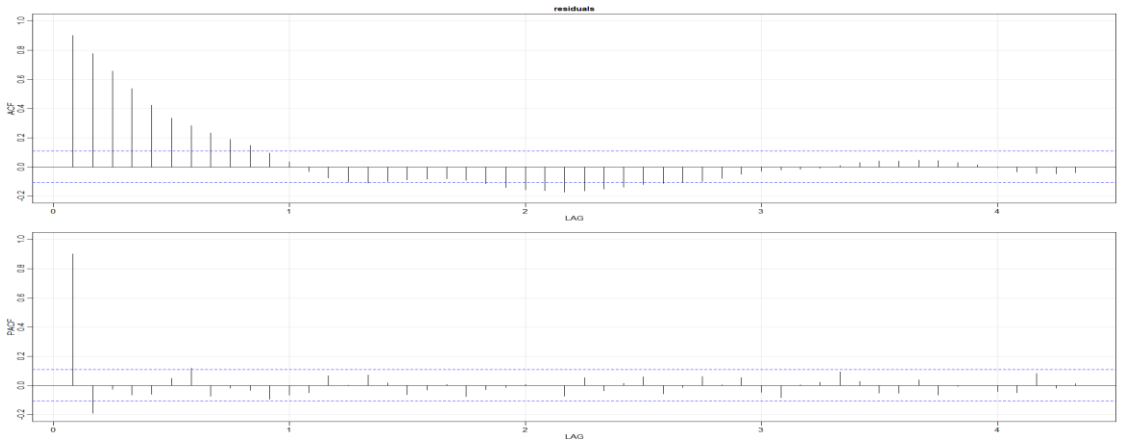
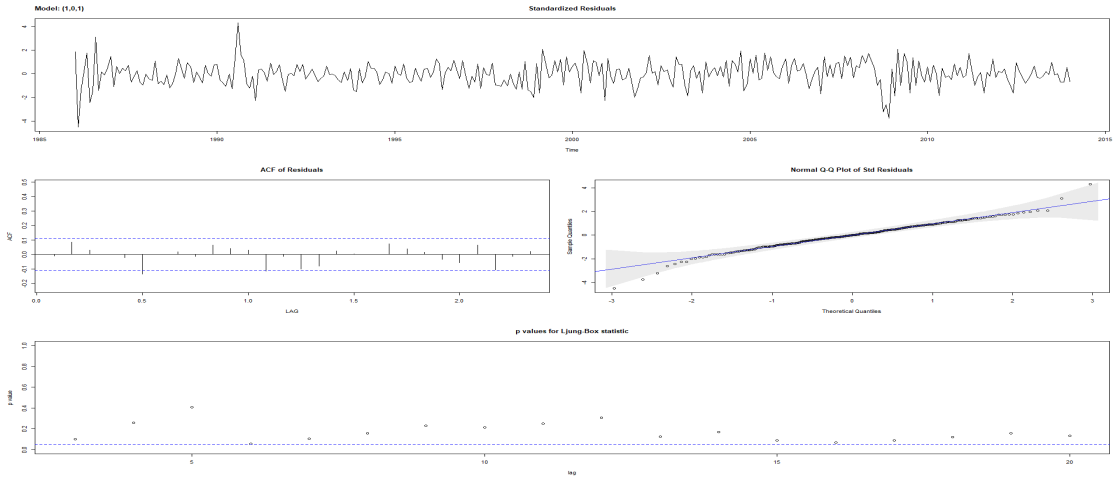


Figure 12: Sample ACF and PACF of oil price data residuals showing and ARMA(1,1) process




```
sarima(log(WTI), 1,0,1, xreg = cbind(x1,x2,x4), no.constant=FALSE)
```

Figure 13: Residual analysis for the model with regression and auto-correlated errors

The outputs of "sarima.for" code are shown below:

```
## $pred
##          Jan      Feb      Mar      Apr      May      Jun      Jul
## 2014          4.528495 4.517802 4.507236 4.496795 4.486478 4.476282
## 2015 4.417590 4.408209 4.398939 4.389778 4.380725 4.371780 4.362940
## 2016 4.312055
##          Aug      Sep      Oct      Nov      Dec
## 2014 4.466207 4.456251 4.446413 4.436691 4.427084
## 2015 4.354205 4.345573 4.337044 4.328615 4.320285
## 2016
##
## $se
##           Jan       Feb       Mar       Apr       May       Jun       Jul
## 2014           0.0825362 0.1332055 0.1685772 0.1970835 0.2214058 0.2428166
## 2015 0.3374578 0.3497036 0.3612613 0.3722011 0.3825821 0.3924541 0.4018601
## 2016 0.4503030
##           Aug       Sep       Oct       Nov       Dec
## 2014 0.2620413 0.2795412 0.2956317 0.3105405 0.3244384
## 2015 0.4108373 0.4194181 0.4276310 0.4355015 0.4430521
## 201
```

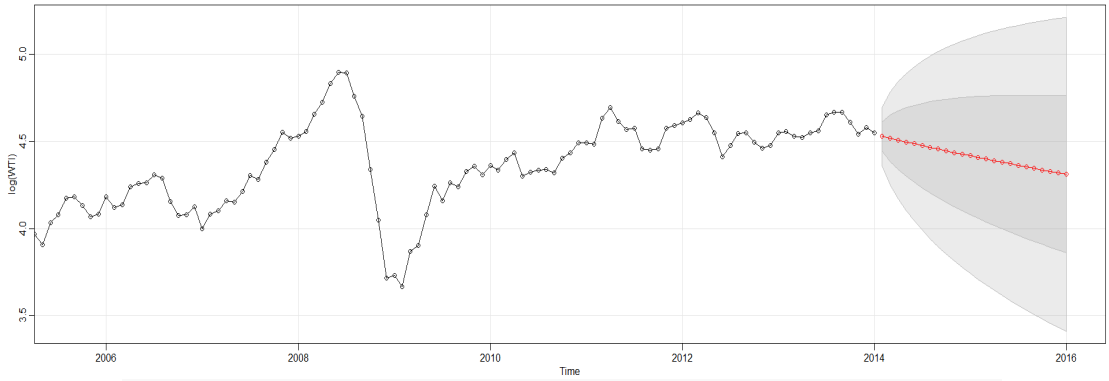
The model expression is written as:

$log(WTI) = y_t = 3.1 + 0.069tm + 0.009tm^2 - 0.00004tm^4 + e_t$,

where

$tm = time(log(WTI)) - mean\big(time(log(WTI))\big)$ and

$e_t = 0.88e_{t-1} + w_t + 0.31w_{t-1}$.




```
sarima.for(log(WTI),24,1,0,1, newxreg = cbind(x1,x2,x4), no.constant=FALSE)
```

Figure 14: Data forecasting for a 24-month duration using regression and auto-correlated errors

Model estimation using MLE method

As another method, we use maximum likelihood approach to estimate ARIMA model parameters. The R code to implement this method is listed below. The 24-month forecasted data is also shown in Figure 15.

```
MLE.WTI=arima(WTI, order = c(1, 2,1),seasonal = list(order = c(1,0,1),12),method ="ML")
fore=predict(MLE.WTI, n.ahead=24)
ts.plot(WTI, fore$pred, col=1:2, xlim=c(1985,2016), ylab="WTI", lwd=2)
U=fore$pred+fore$se ; L=fore$pred-fore$se
xx=c(time(U), rev(time(U))); yy=c(L, rev(U))
polygon(xx,yy, border=8, col=gray(.6,alpha = .2))
lines(fore$pred, type="p", col=2)
```

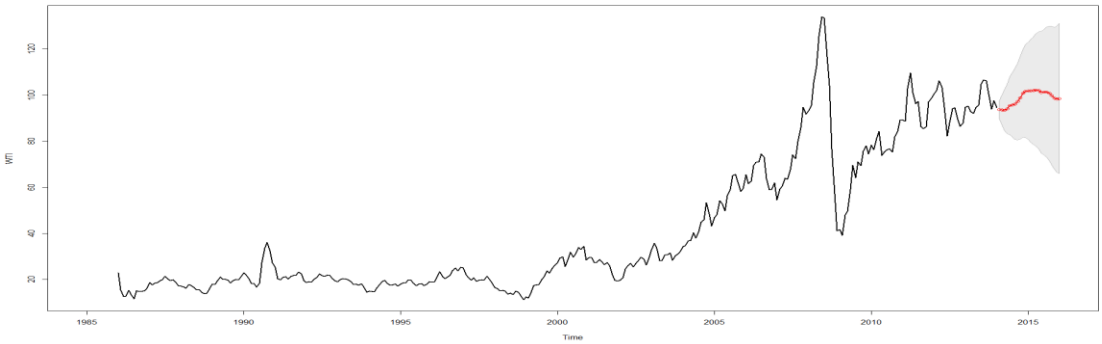


Figure 15: Data forecasting for a 24-month duration using MLE approach

With model expression as:

$x_t = 0.4x_{t-1} - 0.82x_{t-12} - 0.33x_{t-13} + w_t - w_{t-1} + 0.9w_{t-12} - .9w_{t-13}$ where $x_t = \nabla\nabla(log(WTI))$.

Model estimation using Yule-Walker method

Yule-Walker is another model estimation method which uses method of moments to equate the unknown population moments to the known sample moments [3]. This method subsequently back-calculate the model parameters using sample moments as the estimates of population moments. As Yule-Walker method bests works for AR process, we assume and AR process of order 1 for the stationary data (i.e., second-order-differenced WTI oil price in logarithmic scale). The R code is listed below with the forecasted data shown in Figure 16.

```
WTI.yw=ar.yw(Differenced_lWTI, order=1)
#Forcast
fore=predict(WTI.yw, n.ahead=24)
undiff=diffinv(fore$pre,differences = 2,xi=c(log(97.63),log(94.62)))
unlog=exp(undiff)
```

```
ts.plot(WTI,unlog, col=1:2, xlim=c(1985,2016), ylab="WTI", lwd=2)
lines(unlog, type="p", col=2)
```

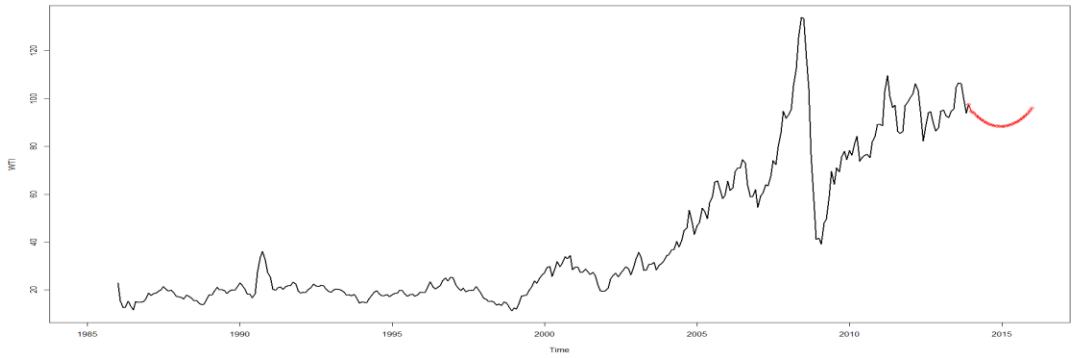


Figure 16: Data forecasting for a 24-month duration using Yule-Walker estimation approach

The model expression for Yule-Walker method is written as:

$x_t = -0.34x_{t-1}$ where $x_t = \nabla\nabla(log(WTI))$.

Summary, conclusion, and future studies

In this study, several methods were analyzed to match the monthly price of WTI crude oil for years between 1986 to 2014. Each method resulted in a predictive model based on multiplicative seasonal ARIMA approach. The data was first transferred to the logarithmic space and a second order differencing was applied to ensure stationarity. In one model, an ARIMA(1,2,1)×$(1,0,1)_{12}$ model was fitted to the data and used for 24-month forecasting. The model gave rise to a reasonable forecast with relatively large uncertainty which is to some extent in accordance with what is observed in reality. Another method applied a fourth-order polynomial relationship to extract data trendline and modeled the residuals as an ARMA(1,1) process. This approach led to a different forecast compared to that from ARIMA(1,2,1)×$(1,0,1)_{12}$ model with a lower range of uncertainty. However, we showed that the future predictions are largely affected by the assumed trendline meaning that the average forecasted values mainly follow the trendline with an auto-correlated error defined by ARMA(1,1) process. This effect can be considered as an advantage of the method if there is a high confidence on the future data trend. It could however cause erroneous results if the trendline significantly deviates from actual data in future. In conclusion, our analyses showed that several methods lead to a uniquely different forecasted oil price values with different uncertainty range. It is important to run various models to encompass the entire possible uncertainty ranges. In our study, time was the only variable used for regression. For a future work, it is recommended to bring additional independent variables to perform a more comprehensive lagged regression. In particular, parameters such as balance of oil supply and demand can be considered to obtain a more accurate predictive model.

References

[1] https://www.macrotrends.net/1369/crude-oil-price-history-chart

[2] Dettling, M. (2016). *Applied Time Series Analysis*. Winterthur, Switzerland: Zurich University of Applied Sciences.

[3] Shumway, R.H. and Stoffer, D.S. (2016). *Time Series Analysis and Its Applications with R Examples (4th Edition)*. New York, USA: Springer Texts in Statistics.