Crude Oil Prediction

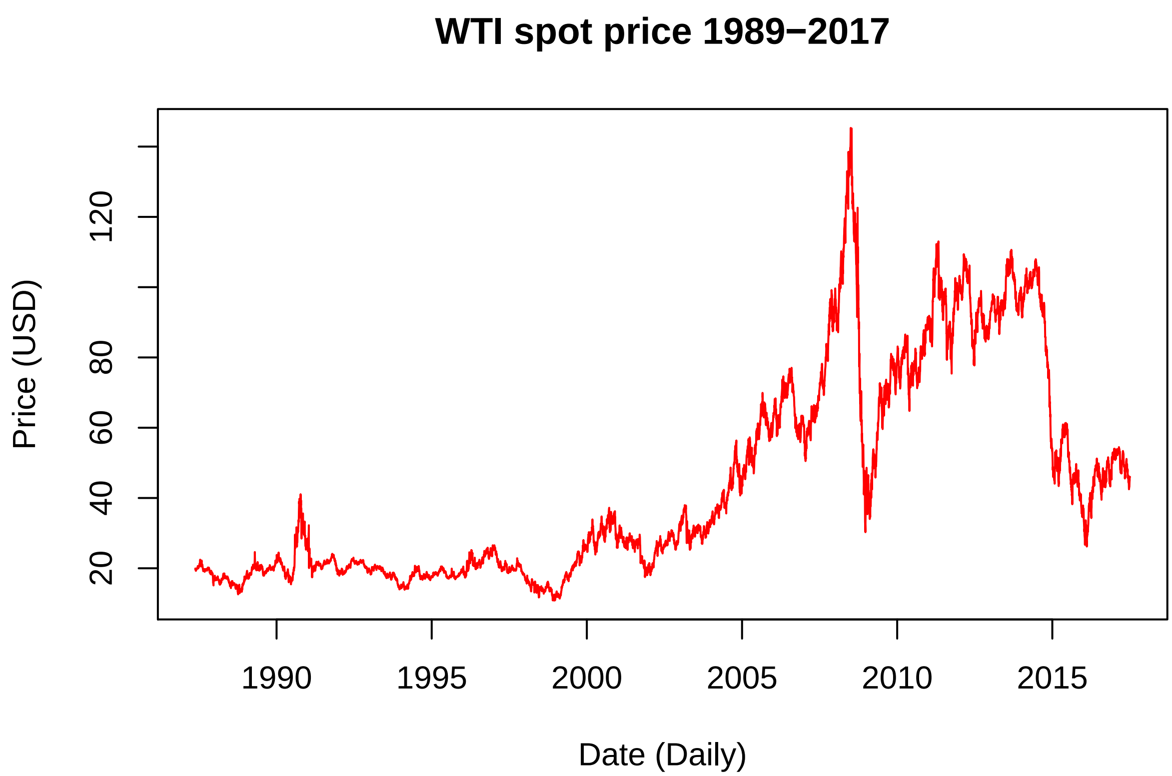
Guillaume Ostrom - ECP 2017

July 8th 2017

Introduction

In this paper we will try to modeling and forecasting an important commodity for economics agents : the **oil price**. We will analyse the **West Texas Intermediate (WTI)** in this document.

Here is the daily WTI (Red) crude oil price from the U.S. Energy Information Administration.



Source:
https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=D

We saw a 78% sell-off in 2008 because of the global financial crisis bubble explosion. For the first 2017 semester the oil price is around 50$.

Here is a little summary:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.82   19.86   30.05   44.29   65.38  145.30
```

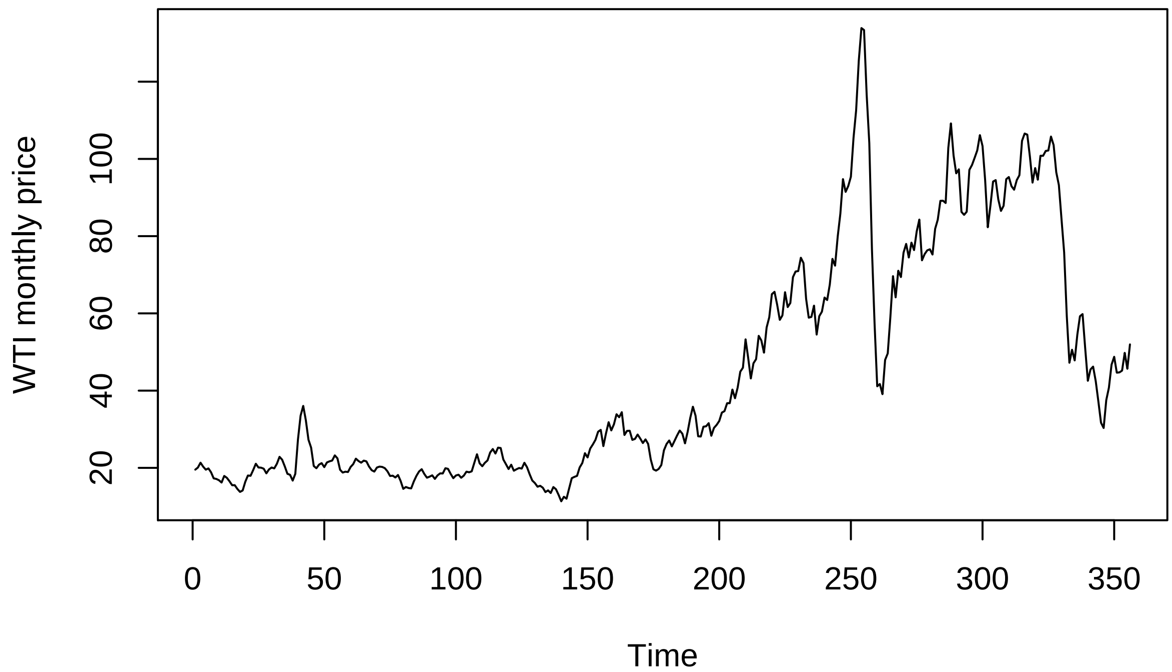
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    9.10   18.50   28.55   45.06   65.22  144.00
```

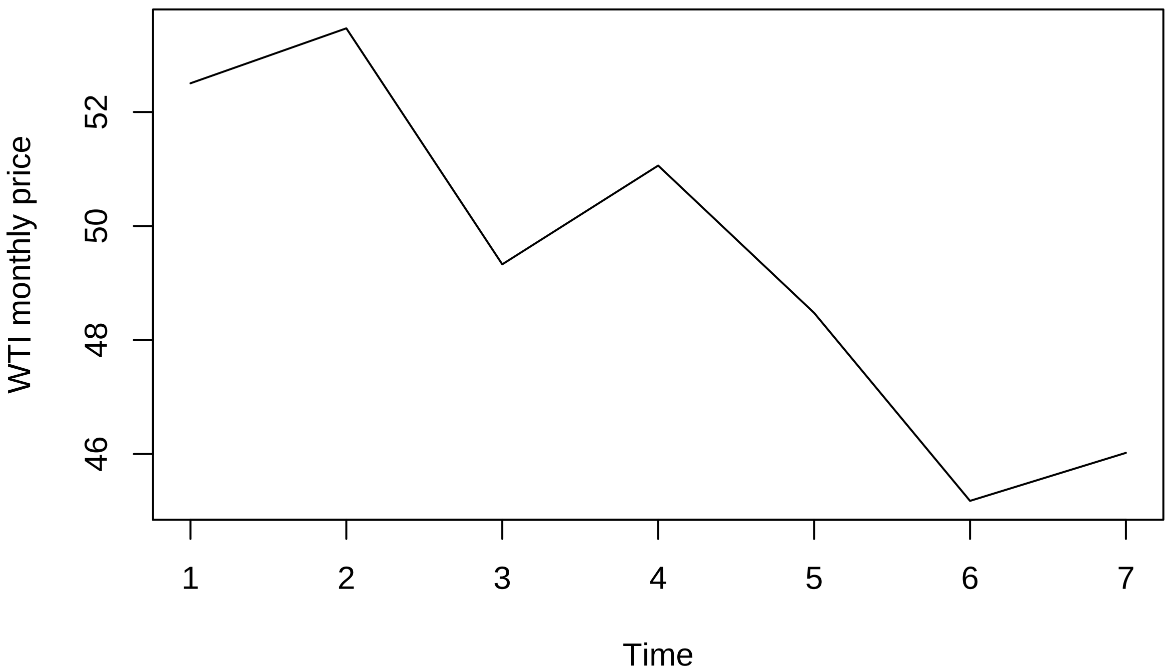
Part I : Prepare the data

We will analyse the monthly price of the WTI. The data before december 2016 will be our training set for modeling. The data after december 2017 will be our testing set for prediction.

```
plot(wti$price,type='l',xlab='Time',ylab='WTI monthly price')
```



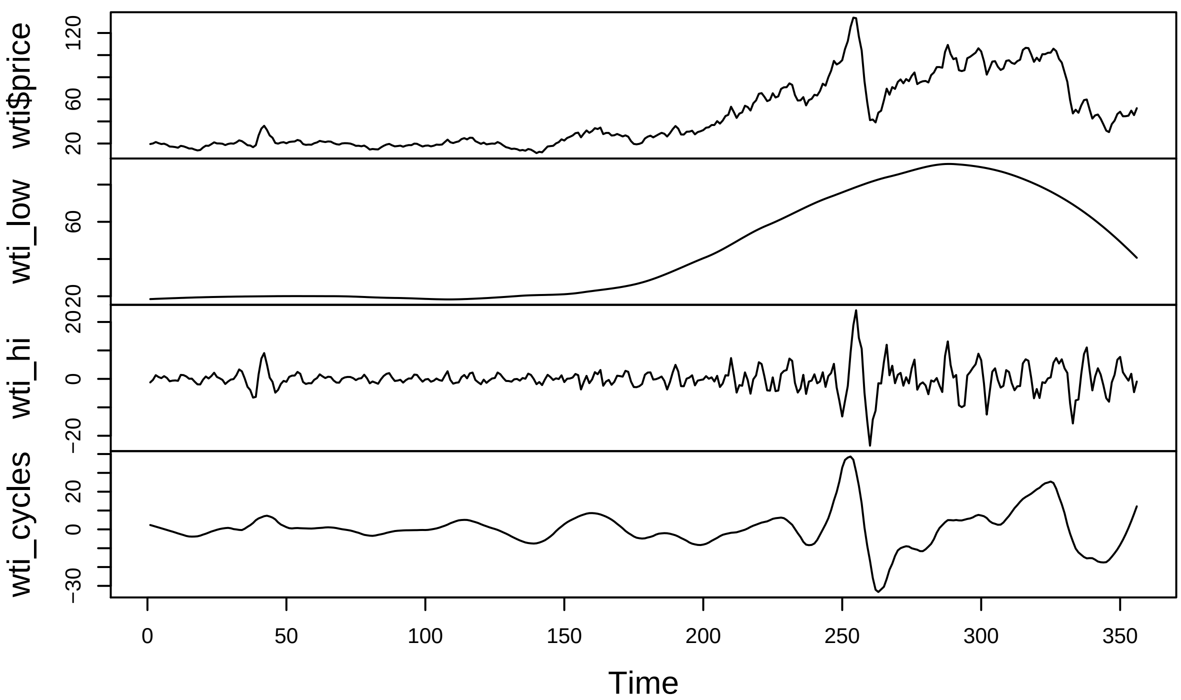

```
plot(wti_test$price,type='l',xlab='Time',ylab='WTI monthly price')
```



To use a ARMA(p,q) model we need to detrend the monthly data first. We use a filter (exponential smoothing) to extract a cycle. We consider high frequency variation as a "noise", the low frequency variation as the trend and the middle frequencies as the cycle. A band of mid-range frequencies might be considered to correspond to the cycle.

```
plot(ts.union(wti$price, wti_low,wti_hi,wti_cycles),
     main="WTI monthly = trend + noise + cycles")
```

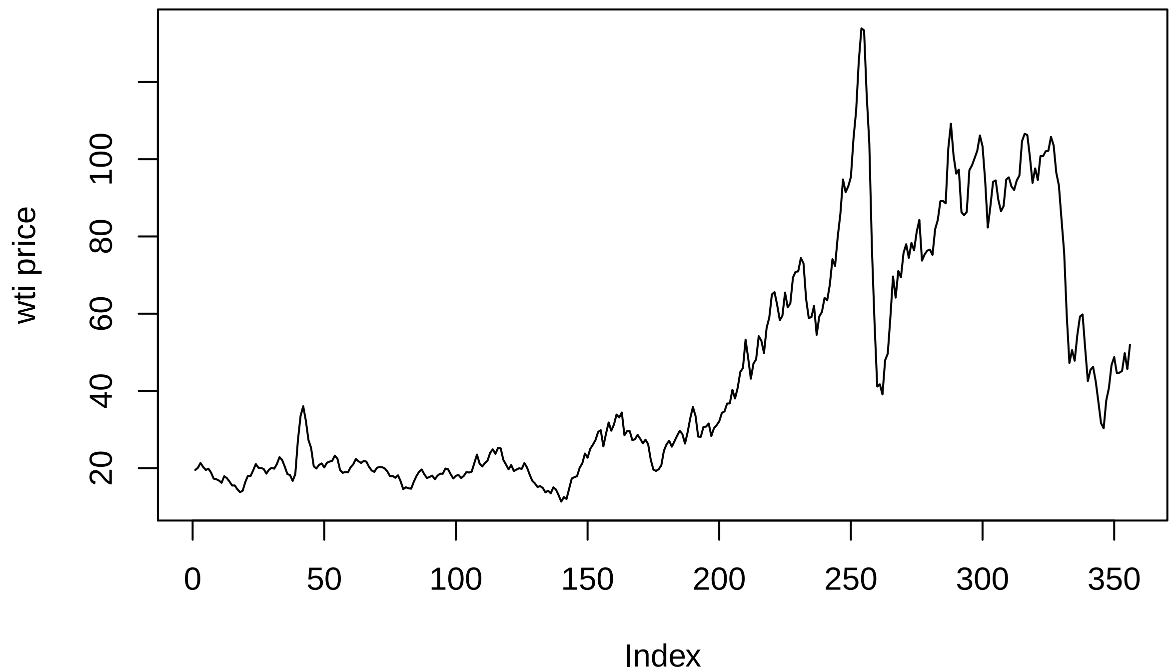
WTI monthly = trend + noise + cycles



This transformations is not enough because of the high frequencies.

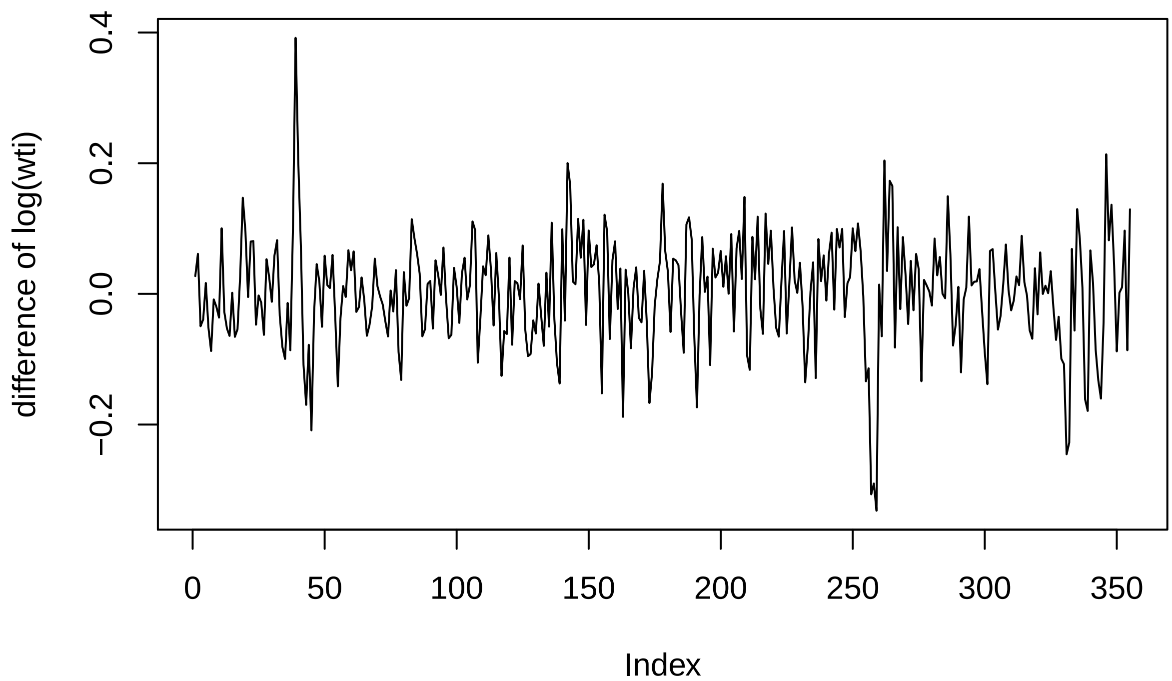
We tried 5 box-Cox transformations, the log difference gave the best results.

```
plot(wti$price,type='l',ylab='wti price')
```

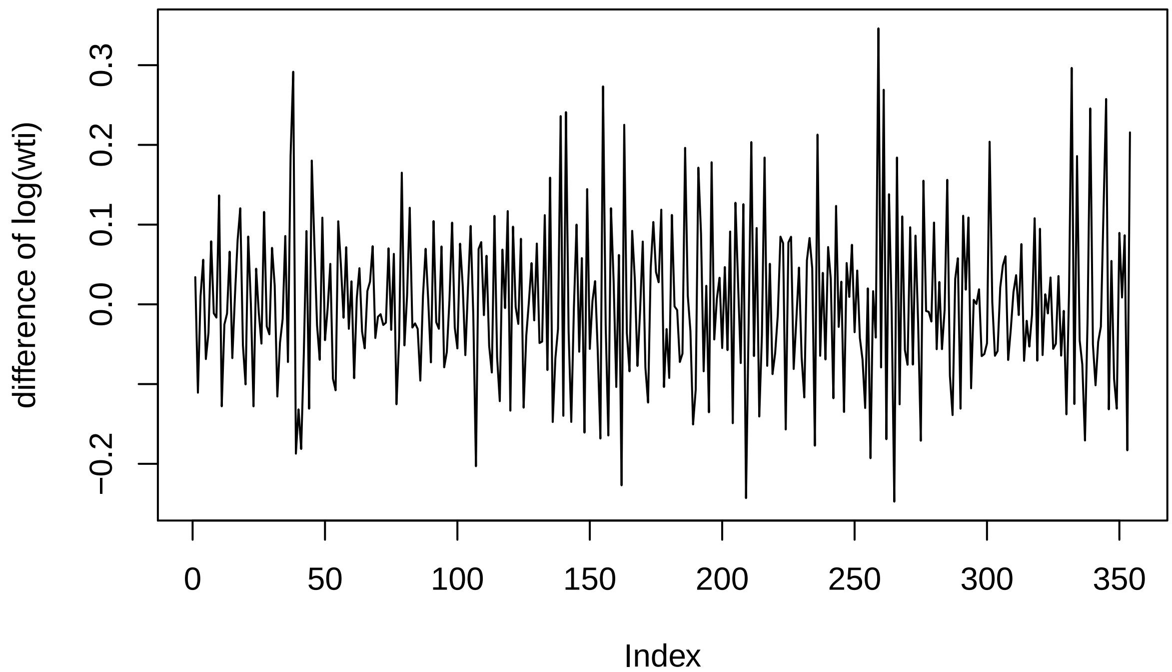



```r
plot(log(wti$price),type='l',ylab='log(wti)')
```


```r
plot(diff(log(wti$price),differences = 1),type='l',ylab='difference of log(wti)')
```




```
plot(diff(log(wti$price),differences = 2),type='l',ylab='difference of log(wti)')
```

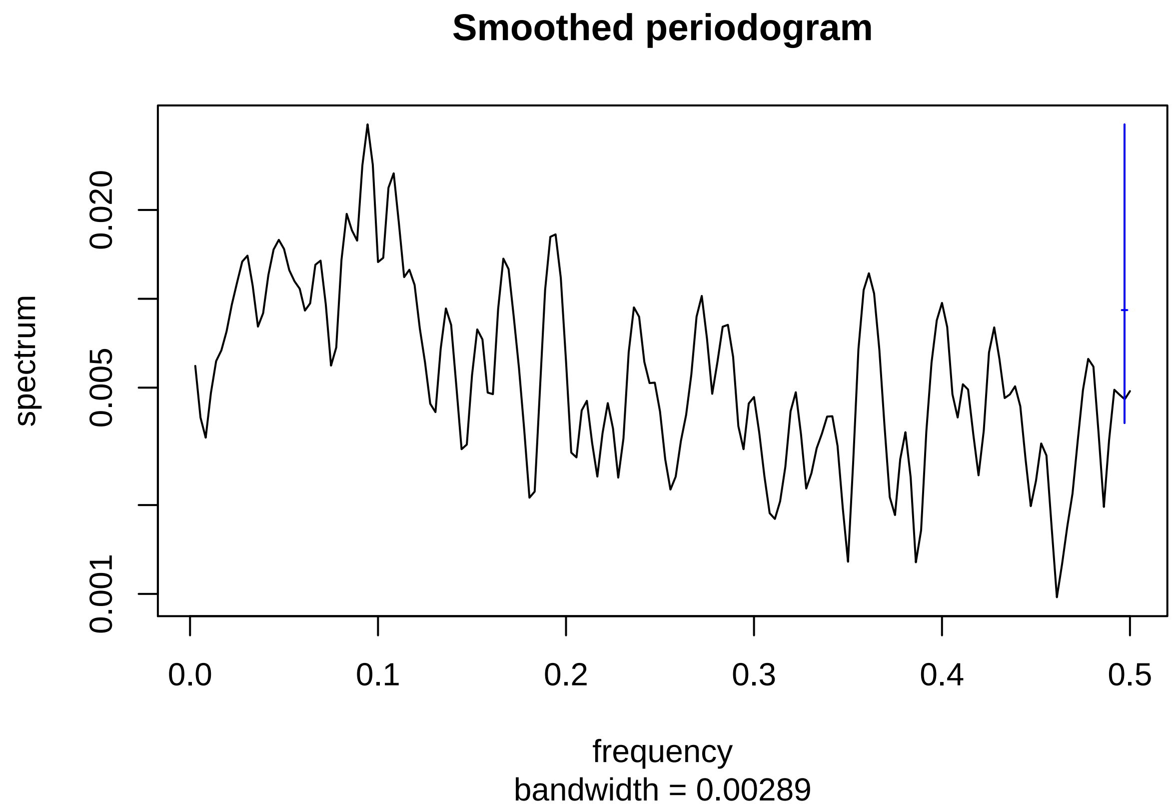


Now the time serie looks stationnary. We can try an ARMA(p,q) model.

Part II : Modeling

We check if there is some seasonality. A smooth the log difference.

```
diff_logprice=diff(log(wti$price),differences = 1)
f=spectrum(diff_logprice,spans=c(2,2), main="Smoothed periodogram")
```




```r
f$freq[which.max(f$spec)]
```

```
## [1] 0.09444444
```

```r
1/f$freq[which.max(f$spec)]
```

```
## [1] 10.58824
```

The max frequency is 0.094, the max period is 10.6. So we will try SARIMA$(p, 1, q) \times (1, 0, 1)_{12}$ under the null hypothesis that the time series are stationary. Where :

$$\phi(B)\Phi(B^{12})\big((1-B)X_n - \mu\big) = \psi(B)\Psi(B^{12})\epsilon_n$$

where ϵ_n is a Gaussian white noise process, the intercept μ is the mean of the differenced process $(1-B)X_n - \mu$, and we have

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

$$\Phi(B^{12}) = 1 - \phi_1 B^{12}$$

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_q B^q$$

$$\Psi(B^{12}) = 1 + \psi_1 B^{12}$$

To chose the p (AR) qnd the q (MA) we use the low AIC method.

	MA0	MA1	MA2	MA3	MA4
AR0	-748.53	-777.59	-779.22	-777.43	-777.19
AR1	-780.72	-778.72	-777.30	-776.22	-780.15
AR2	-778.73	-776.76	-781.61	-774.61	-778.16
AR3	-778.18	-781.68	-779.68	-777.76	-776.43
AR4	-779.01	-779.68	-778.09	-783.86	-781.98

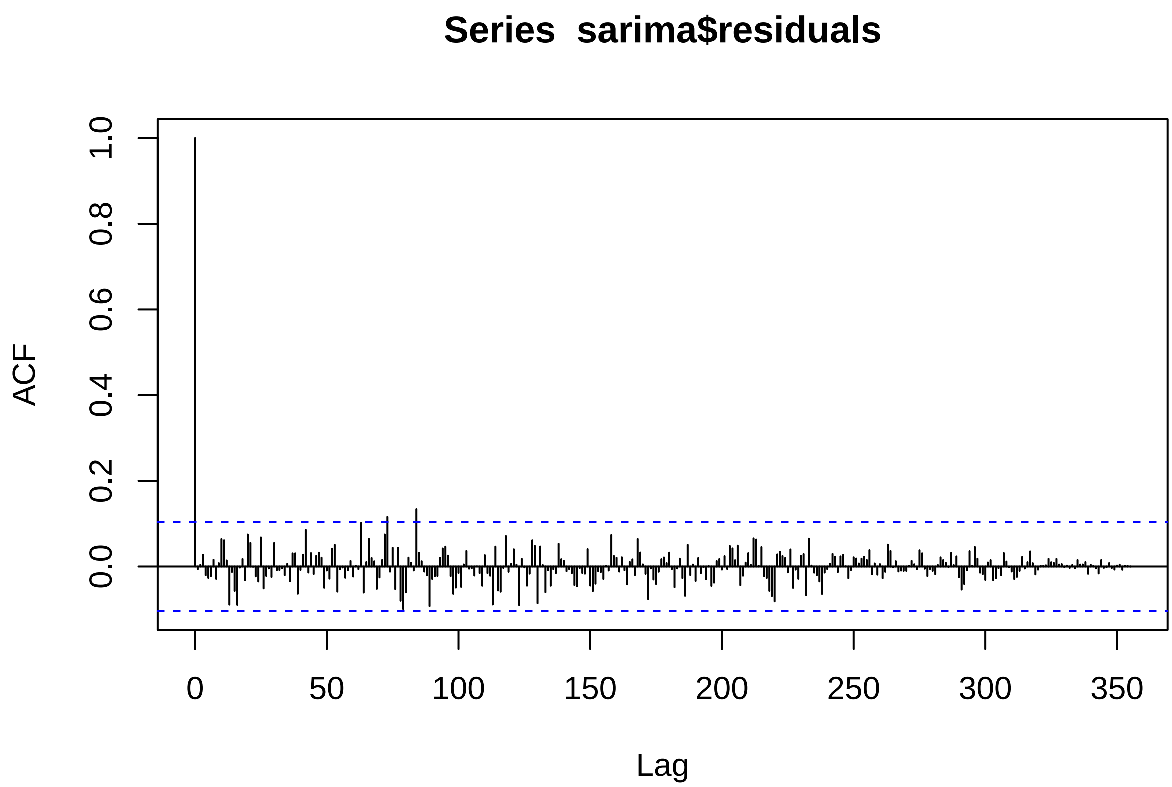
We select $\text{SARIMA}(4,1,3) \times (1,0,1)_{10}$

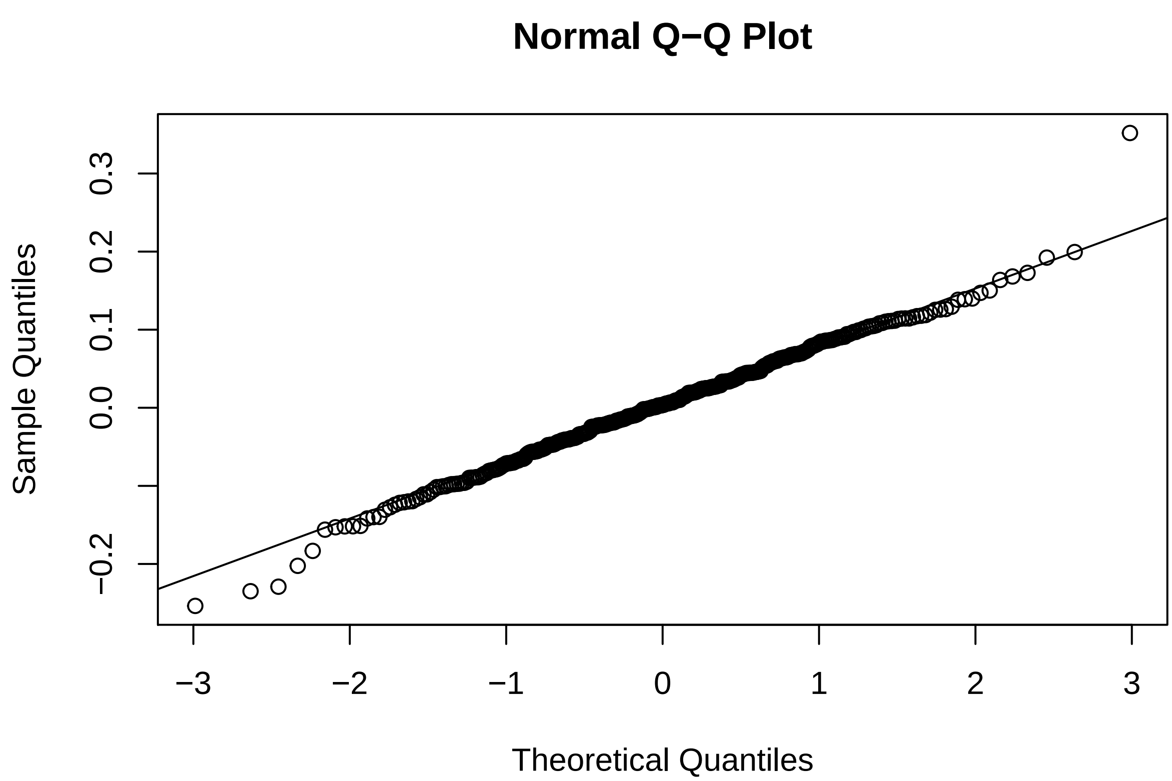
```r
sarima=arima(log(wti$price),order=c(4,1,3),seasonal=list(order=c(1,0,1),period=12))
sarima
```

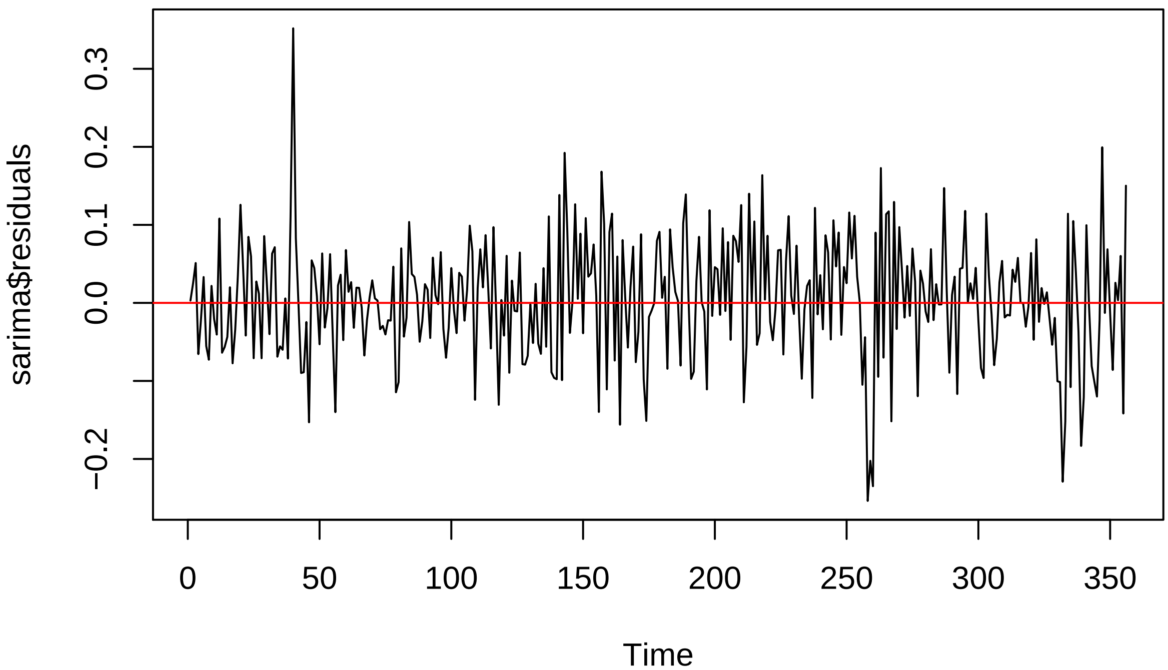
```
##
## Call:
## arima(x = log(wti$price), order = c(4, 1, 3), seasonal = list(order = c(1, 0,
##     1), period = 12))
##
## Coefficients:
##          ar1     ar2     ar3      ar4     ma1      ma2      ma3     sar1
##       0.0476  0.1835  0.7840  -0.3328  0.2677  -0.1001  -0.9235  -0.8209
## s.e.  0.0680  0.0602  0.0506   0.0508  0.0557   0.0626   0.0582   0.1557
##         sma1
##       0.9109
## s.e.  0.1283
##
## sigma^2 estimated as 0.005983:  log likelihood = 401.93,  aic = -783.86
```

Part III : Forecasting and back-testing

To check our model we need to check our residuals with the: -ACF plot under the 5% level acceptation under a null hypothesis.
-QQplot to see if the residuals are normally distributed.
-check if there is not extremly high residual point.

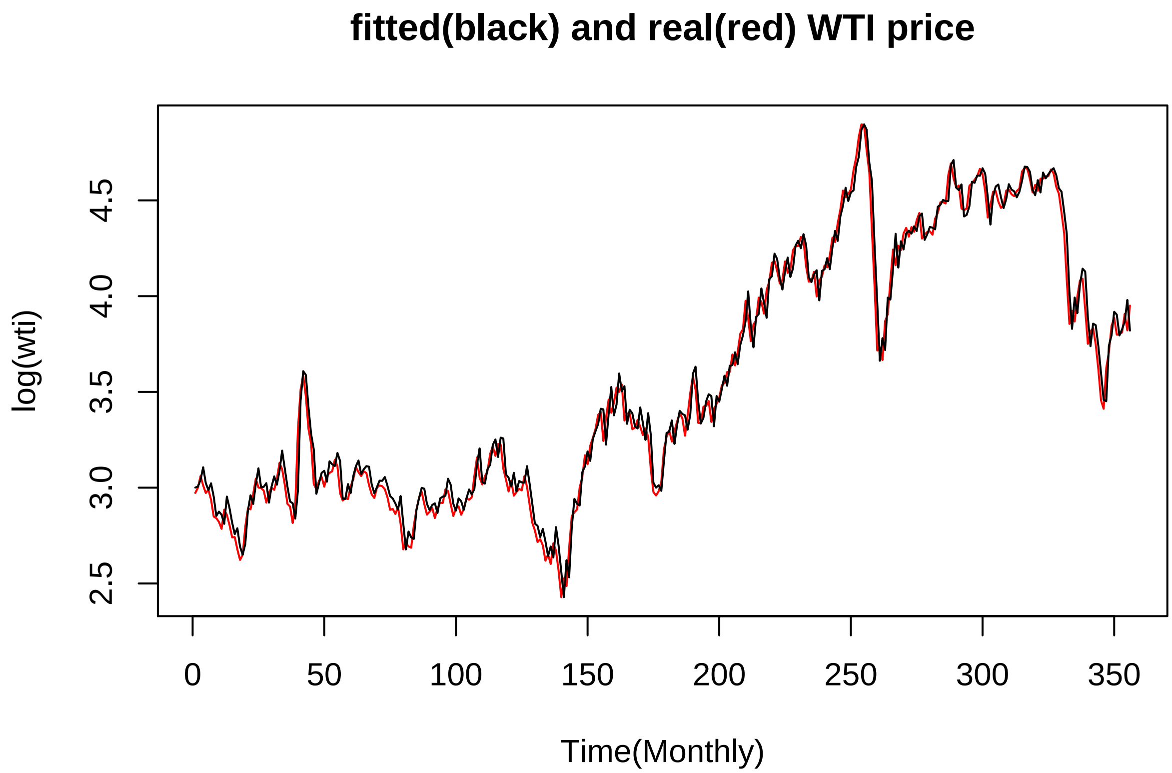

Series sarima$residuals
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Lag


Normal Q−Q Plot
Sample Quantiles
−0.2
0.0
0.1
0.2
0.3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

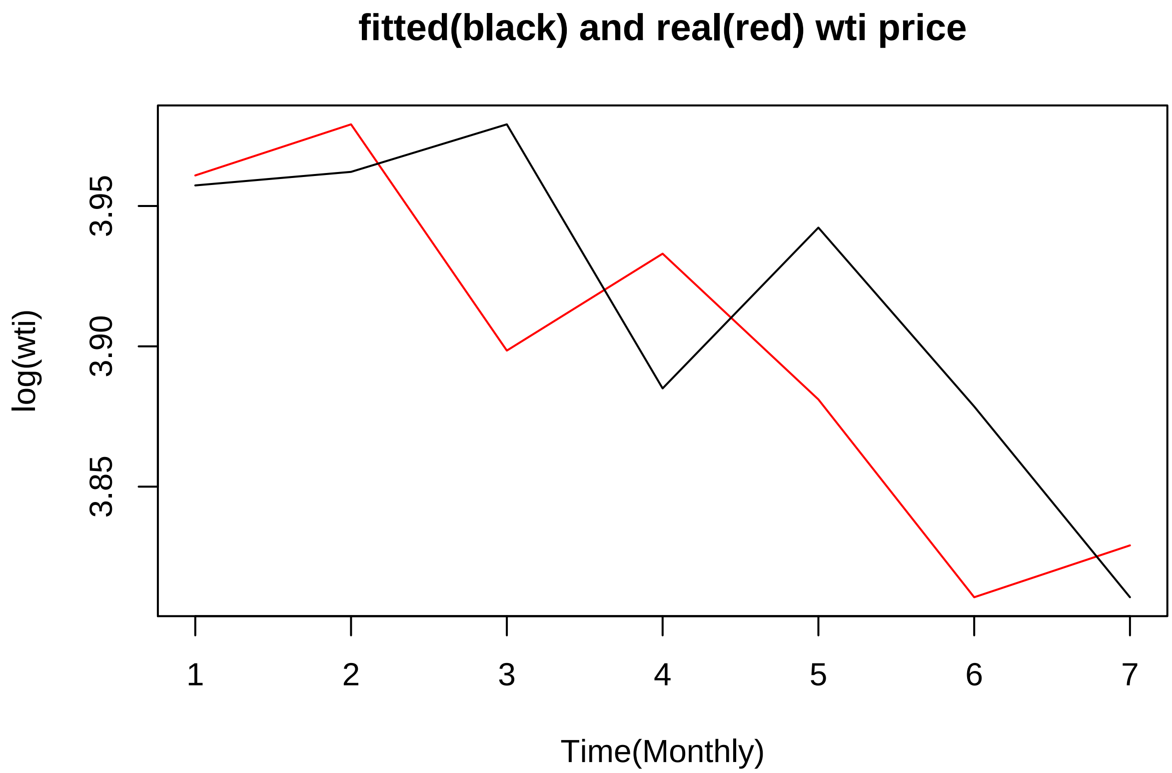


In conclusion we cannot reject the null hypothesis of Gaussiqn noise. The the model is good to test.

```
## Series: log(wti$price)
## ARIMA(4,1,3)(1,0,1)[12]
##
## Coefficients:
##          ar1     ar2    ar3      ar4     ma1      ma2      ma3     sar1
##       0.0476  0.1835  0.784  -0.3328  0.2677  -0.1001  -0.9235  -0.8209
## s.e.  0.0000  0.0000  0.000   0.0000  0.0000   0.0000   0.0000   0.0000
##         sma1
##       0.9109
## s.e.  0.0000
##
## sigma^2 estimated as 0.005983:  log likelihood=401.93
## AIC=-801.86   AICc=-801.85   BIC=-797.98
```



Now we can try to predict the 2017 WTI price and compare.



Conclusion

We choose a good SARIMA$(4, 1, 3) \times (1, 0, 1)_{12}$ model with a low AIC criterion. So SARIMA is good for prediction. However the model have a 1-month lag, we need to go deeper to understand why and t