TIME SERIES ANALYSIS PROJECT

‘Oil.price Dataset’

Dataset description

The ‘oil.price’ dataset belongs to the ‘TSA’ R package, and contains monthly spot prices for crude oil, Cushing, OK (in U.S. dollars per barrel), for a total of 241 observations, starting from January 1986 up to January 2006.

The dataset’s format is the following: Time series [1:241] with 12 columns representing each month containing the respective spot price, divided per year:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1986	22.93	15.45	12.61	12.84	15.38	13.43	11.58	15.10	14.87	14.90	15.22	16.11
1987	18.65	17.75	18.30	18.68	19.44	20.07	21.34	20.31	19.53	19.86	18.85	17.27
1988	17.13	16.80	16.20	17.86	17.42	16.53	15.50	15.52	14.54	13.77	14.14	16.38
1989	18.02	17.94	19.48	21.07	20.12	20.05	19.78	18.58	19.59	20.10	19.86	21.10
1990	22.86	22.11	20.39	18.43	18.20	16.70	18.45	27.31	33.51	36.04	32.33	27.28
1991	25.23	20.48	19.90	20.83	21.23	20.19	21.40	21.69	21.89	23.23	22.46	19.50
1992	18.79	19.01	18.92	20.23	20.98	22.38	21.77	21.34	21.88	21.68	20.34	19.41
1993	19.03	20.09	20.32	20.25	19.95	19.09	17.89	18.01	17.50	18.15	16.61	14.51
1994	15.03	14.78	14.68	16.42	17.89	19.06	19.65	18.38	17.45	17.72	18.07	17.16
1995	18.04	18.57	18.54	19.90	19.74	18.45	17.32	18.02	18.23	17.43	17.99	19.03
1996	18.85	19.09	21.33	23.50	21.16	20.42	21.30	21.90	23.97	24.88	23.70	25.23
1997	25.13	22.18	20.97	19.70	20.82	19.26	19.66	19.95	19.80	21.32	20.19	18.33
1998	16.72	16.06	15.12	15.35	14.91	13.72	14.17	13.47	15.03	14.46	13.00	11.35
1999	12.51	12.01	14.68	17.31	17.72	17.92	20.10	21.28	23.80	22.69	25.00	26.10
2000	27.26	29.37	29.84	25.72	28.79	31.82	29.70	31.26	33.88	33.11	34.42	28.44
2001	29.59	29.61	27.24	27.49	28.63	27.60	26.42	27.37	26.20	22.17	19.64	19.39
2002	19.71	20.72	24.53	26.18	27.04	25.52	26.97	28.39	29.66	28.84	26.35	29.46
2003	32.95	35.83	33.51	28.17	28.11	30.66	30.75	31.57	28.31	30.34	31.11	32.13
2004	34.31	34.68	36.74	36.75	40.27	38.02	40.78	44.90	45.94	53.28	48.47	43.15
2005	46.84	48.15	54.19	52.98	49.83	56.35	58.99	64.98	65.59	62.26	58.32	59.41
2006	65.48											

Note that, the value assigned for each month is an average of crude oil prices belonging to that period, and these data were obtained from the US Energy Information Administration.

Where from a preliminary analysis it is able to see that just by looking at the values displayed above, an increasing trend characterize these recorded data.

```
> summary(oil.price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  11.35   18.01   20.25   24.09   27.31   65.59 
```

Here, I’ve applied the summary function to the dataset since it provides a quick but useful summary returning the main information of it as a minimum value of 11.35 and a maximum of 65.59, with an overall mean of 24.09.

It returns also the first and third quartiles for each variable in the dataset, respectively of 18.01 and 27.31; and the median value of 20.25.

This function is useful in terms of overviewing the main tendency of the data, and even though it was clear also by directly looking at the data, there are no null values in the series that normally could affect the analysis if not preliminary identified.

After this first approach, another important step is about checking that the time series is stationary or not:

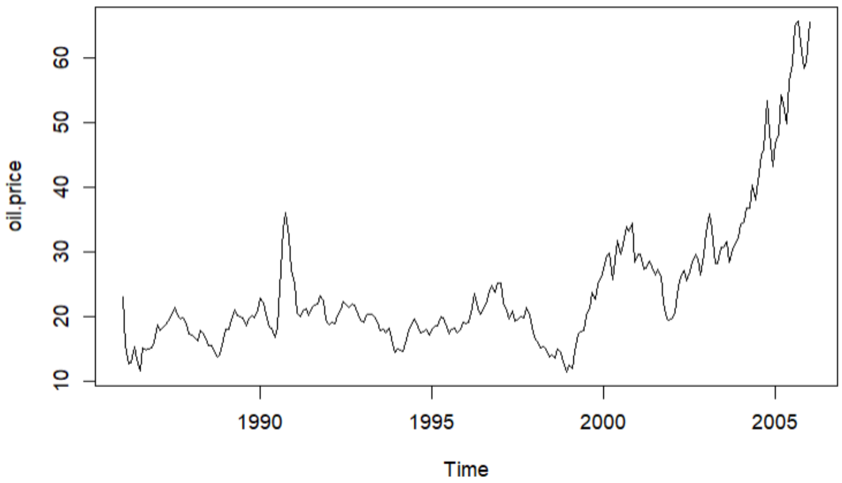
in the case of stationarity, the time series should require three conditions:

1. a constant mean across all t

2. a constant variance across all t

3. the autocovariance between the observations is only dependent on the distance between the observations (lag h)

So, from a first plot we obtain the following situation:



The plot clearly shows a considerable variation that is even more evident in the last period, with an upward trend from 2001 arriving to 2006.

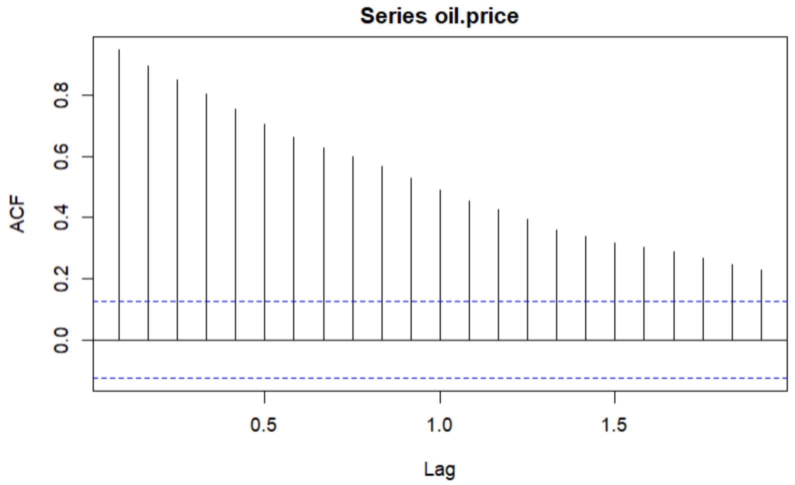
This already represents strong evidence that a stationary model will not be reasonable for this series.

But it’s important to go deeper in this analysis using ACF, PACF and the Augmented Dickey-Fuller Test:

So, I’ve checked the time series' Autocorrelation function through its plot: in general, an autocorrelation function (ACF) plot is used to examine the correlation between a time series and its lags. The ACF plot shows the correlation between a time series and its lags up to a specified number of lags.
In a Time series analysis, analysing an ACF plot is useful to determine if the residuals of a model are independent, which is an important assumption in many time series models.

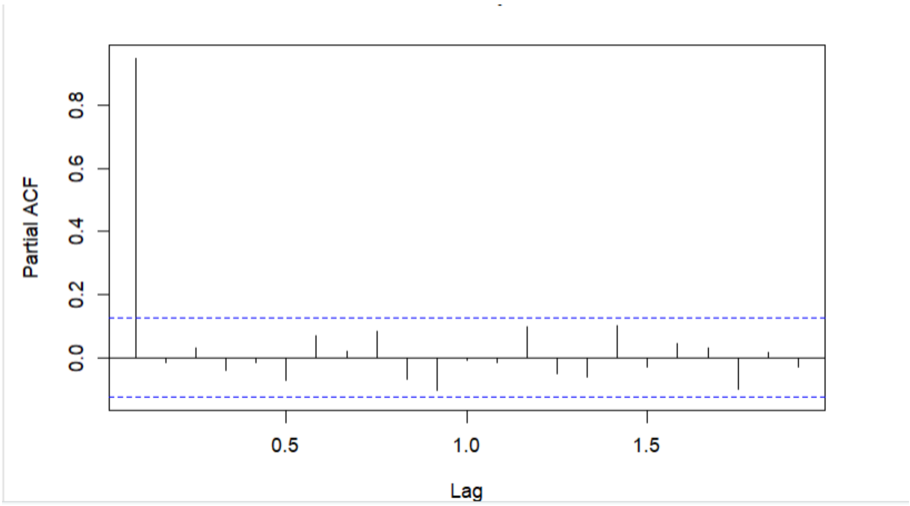
In the case of stationarity, the ACF, and as we'll see later the PACF plots, will show a quick drop-off in correlation after a small amount of lag between points.

On the opposite, the sample ACF computed for non-stationary series will also usually indicate the non-stationarity.



In this case, all values are significantly far from 0 and there is evidence of a pattern of a linear decrease with increasing lag, that could be explained by the common tendency for nonstationary series to drift slowly down with an apparent trend.

Based on this, the sample PACF is indeterminate as well:



An additional tool for confirming these results is given by the Augmented Dickey-Fuller Test:

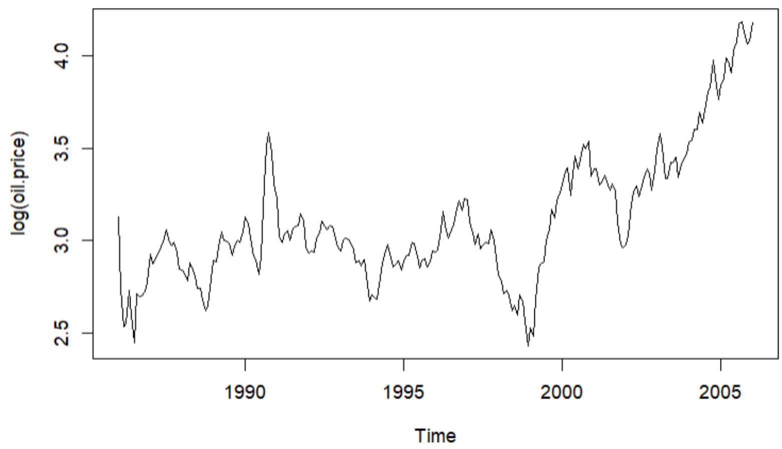
```
Warning in adf.test(oil.price) : p-value greater than printed p-value

        Augmented Dickey-Fuller Test

data:  oil.price
Dickey-Fuller = 0.7045, Lag order = 6, p-value = 0.99
alternative hypothesis: stationary
```

That confirms as well that the dataset is nonstationary, return a high p-value so that I can't reject the first hypothesis that affirms stationarity; a lag order of 6 and a Dickey-Fuller value of 0.7045.

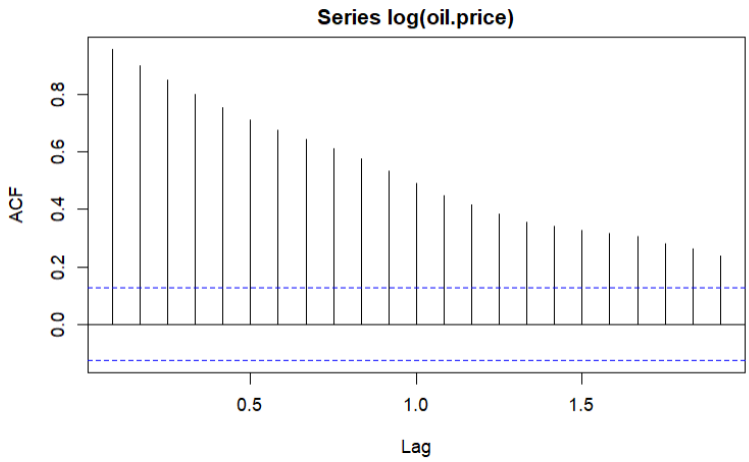
Thanks to this preliminary analysis I can conclude that a proper model for this kind of series could be represented by the nonstationary IMA(1,1) model; however, a better way to approach this dataset could be of transforming the whole series.

So here, let's see what happens by applying the logarithm to the series and then, the differencing technique to the transformed series.

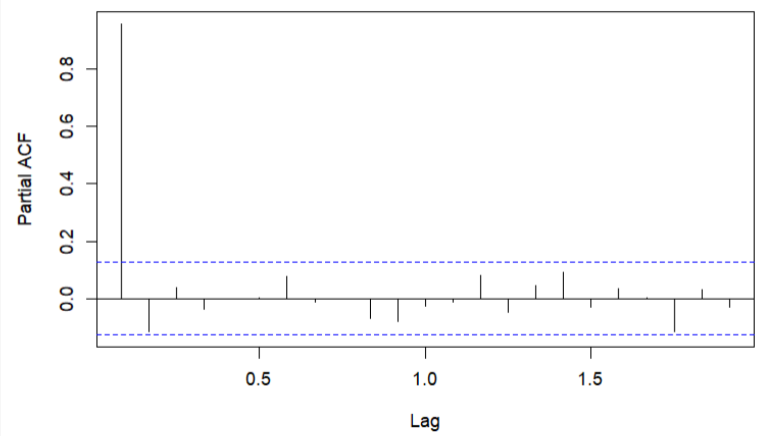
Let's check the first plot having applied the logarithm to the series:



The ACF of the modified series:

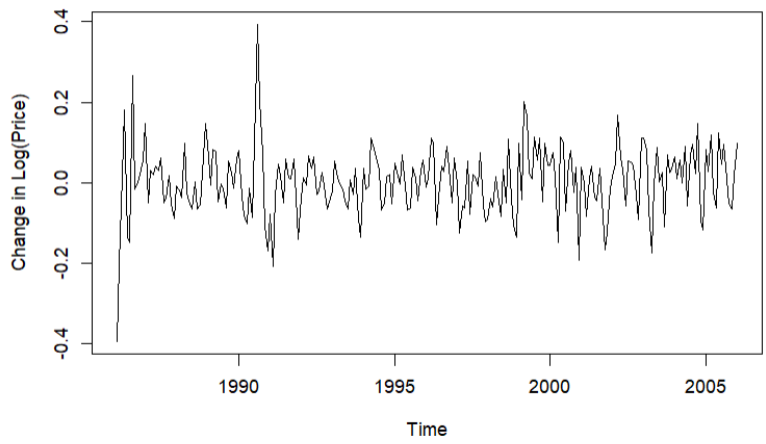


And the PACF:

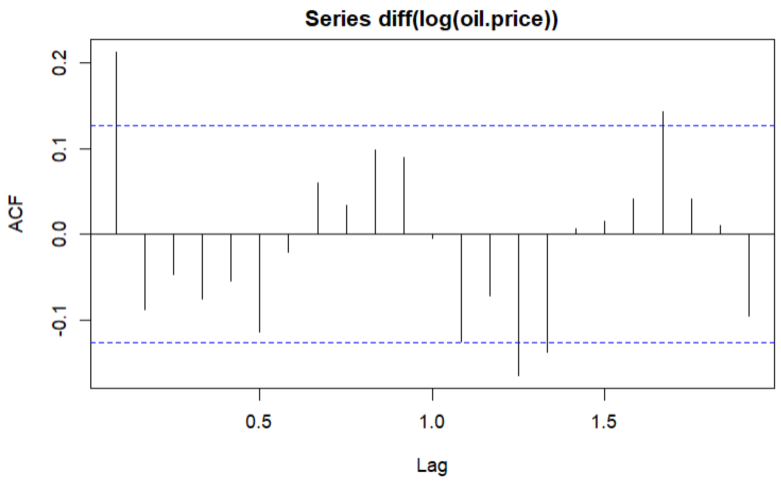


In none of the plots it's possible to see strong differences from the original series, this is why I've decided to apply the 'Differencing Technique' to this transformed series:

So, in the following plot it's possible to see that the series, thanks to the use of the 'diff' function applied on the logarithm of the series, it could be considered stationary; but before confirming it, let's check also the acf and pacf of the same as well.

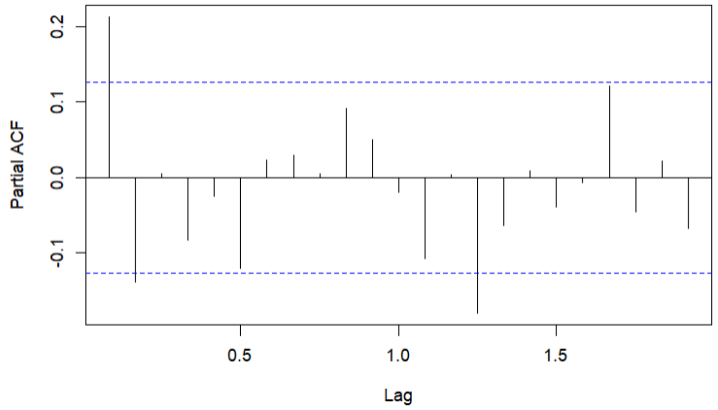


Here, the acf of the differenced log of the oil price series:



Here the lag 1 is significant and this suggests to apply a MA(1) model to apply.

And here the pacf:



Here instead, the partial autoregression function suggests to consider an AR(2) model instead.

```
> library(tseries)
> adf.test(log(oil.price))

        Augmented Dickey-Fuller Test

data:  log(oil.price)
Dickey-Fuller = -1.1119, Lag order = 6, p-value = 0.9189
alternative hypothesis: stationary
```

Here again, let's confirm this result through the Augmented Dickey-Fuller Test:

The test returns a statistic of -1.119 and a p-value of 0.9189, with 'stationarity' as alternative hypothesis.

So, this is a strong proof of non-stationarity and of appropriateness of taking a difference of the logs.

Now, let's discuss which kind of model is properly accurate for this specific case.

With this aim, I'll display below a table of 'Extended ACF' for the transformed series:

```
> eacf(diff(log(oil.price)))
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  o  o  o
1 x x o o o o o o o o x  o  o  o
2 o x o o o o o o o o o  o  o  o
3 o x o o o o o o o o o  o  o  o
4 o x x o o o o o o o o  o  o  o
5 o x o x o o o o o o o  o  o  o
6 o x o x o o o o o o o  o  o  o
7 x x o x o o o o o o o  o  o  o
```

This table suggests an ARMA Model with parameters defined as p=0 and q=1.

According to this first impression, and after many models comparisons, the best model in terms of fitting these data results to be an ARIMA(0, 1, 1) where parameter ‘d’ is equal to 1 representing the degree of differencing used in the transformation:

```
> arima(log(oil.price), order = c(0, 1, 1), method = 'ML')

Call:
arima(x = log(oil.price), order = c(0, 1, 1), method = "ML")

Coefficients:
         ma1
      0.2956
s.e.  0.0693

sigma^2 estimated as 0.006689:  log likelihood = 260.29,  aic = -518.58
```

With a parameter estimate of 0.2956 and a standard error of 0.0693.

In terms of output for checking the goodness of the model fitting we need the observe three measures: the sigma squared estimate, standing for the variance of the model, here returning a very small value of 0.006689 that implies a good result in terms of fitting; then we need to check the log likelihood value that higher it is, better the model fits the data; in this case the output returned a value of 260.29, the higher I’ve obtained in my trials; and finally the last value displayed above, the ‘AIC’ value.

$$AIC = 2k - 2\ln(\hat{L})$$

By definition, the Akaike Information Criterion (AIC) is a measure of the goodness of fit of a statistical model, where lower AIC values indicate a better fit. The AIC is defined as

where ‘k’ represents the number of parameters in the model and ‘L^’ represents the maximum likelihood of the model.

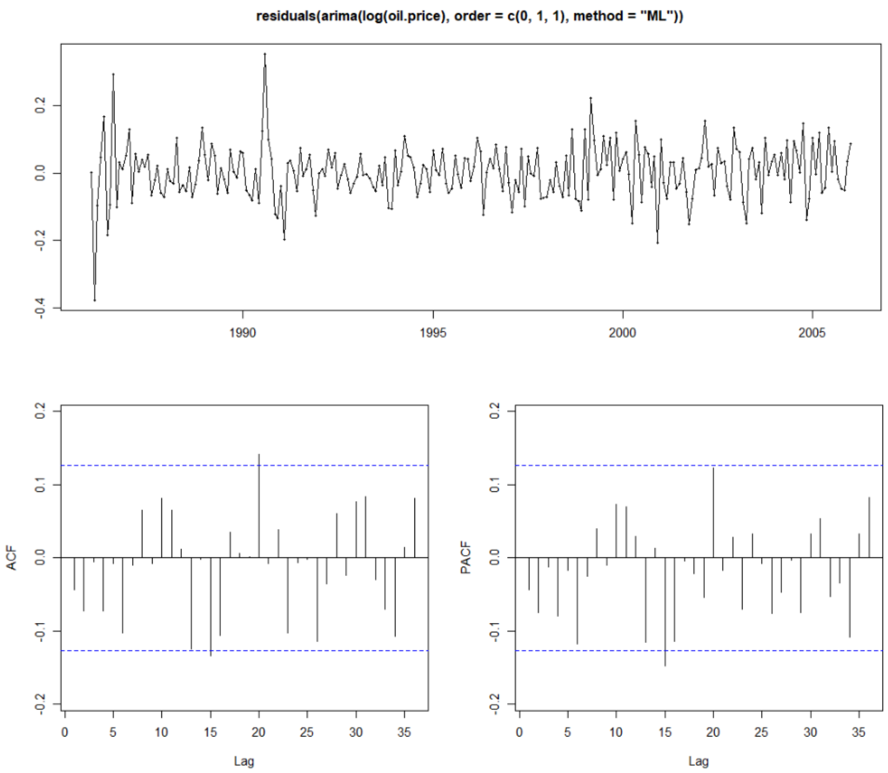
In this model’s results, an interesting indicator is a negative AIC value of -518.58.

This kind of result is a possible outcome, although it is not common since a negative AIC value indicates that the model is a very good fit to the data, but it does not have a straightforward interpretation in terms of model complexity or prediction accuracy, problem that I’ve faced later in the analysis by predicting some ‘future’ values.
So here, it is important to keep in mind that the AIC is only one of many possible measures of model fit, and that a low AIC value does not guarantee that a model will perform well in forecasting or other applications.

It is also important to consider other aspects of the model, such as the residuals, the autocorrelation function, and the distribution of the residuals, to ensure that the model is adequate for the data and the problem at hand.

So finally, I need to proceed with the model diagnostic phase in order to confirm that mine assumptions and analysis on the data and the model that I’ve looked for to be the most adapt are correct.

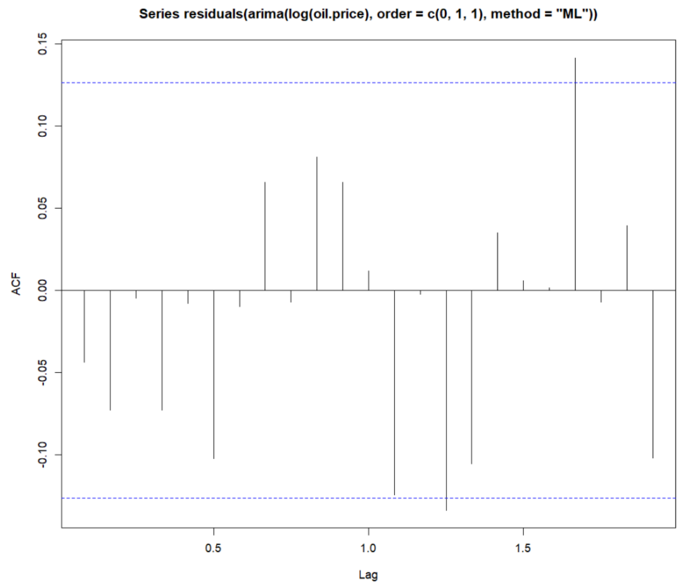
In this phase, I need to analyse the residuals of my model, and firstly I'm going to plot the residuals to visualize their distribution, identify any potential issues with the model and check if they are white noise or not.



In the first plot residuals seems to follow a quite constant trend that can be said to be stationary.

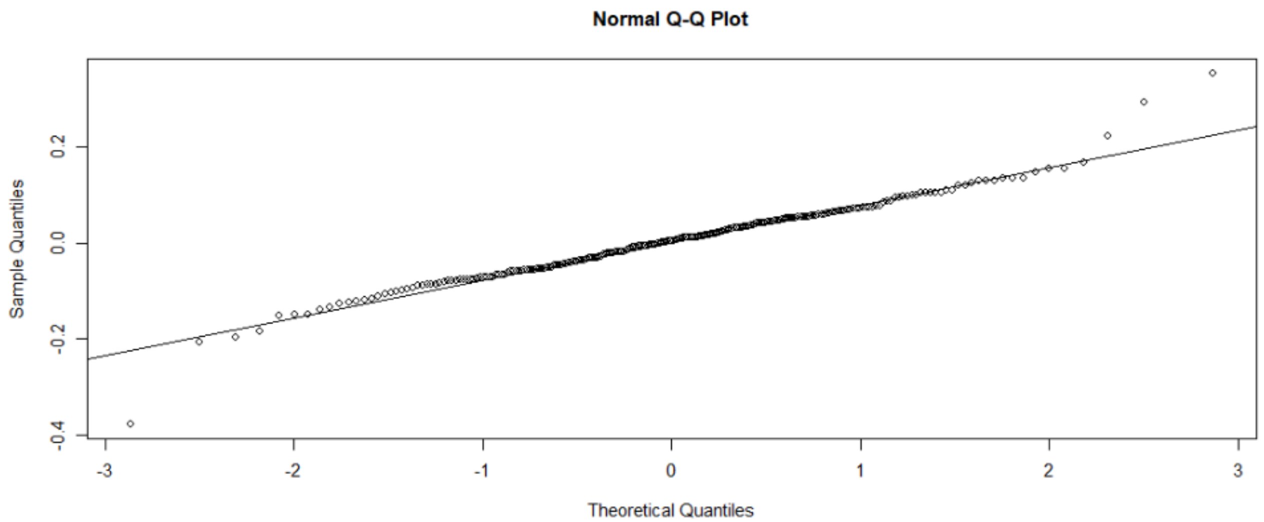
In the ACF and PACF plots, it's possible to analyse the autocorrelation and the partial autocorrelation functions of the residuals, both returning mostly all the values to be close to the 0, even though with come exception.

Looking for the independence assumption of the residuals using the autocorrelation function (ACF), let's control better lag 20:



Here, even though not perfectly, the mean of the residuals is quite close to zero, the only statistically significant correlation is at lag 20, and a smaller one at lag 15; but in general, considering them as exceptions, it's possible to say that the model has captured the dependence levels of the series and that there is no significant correlation.

Below a QQ – plot that compare the theoretical quantiles to the corresponding sample ones to check for normality of the residuals.



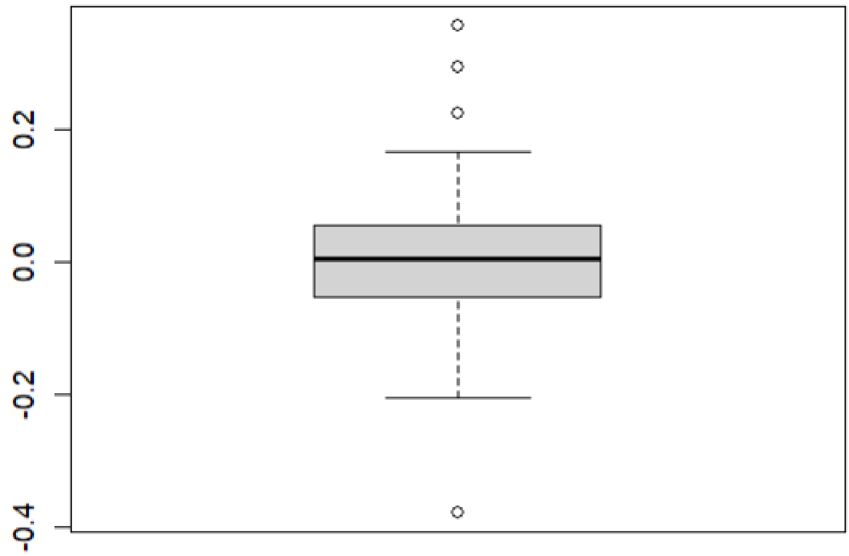
The residuals are very close to the straight line of reference, and this represents a good sign of residuals' normal distribution; however, it's possible to see that there are some outliers that must be considered in case of additional issues faced during the following analysis.

In general, if the residuals are normally distributed, independent, and have constant variance, it's possible to say that the ARIMA model is a good fit for the data.

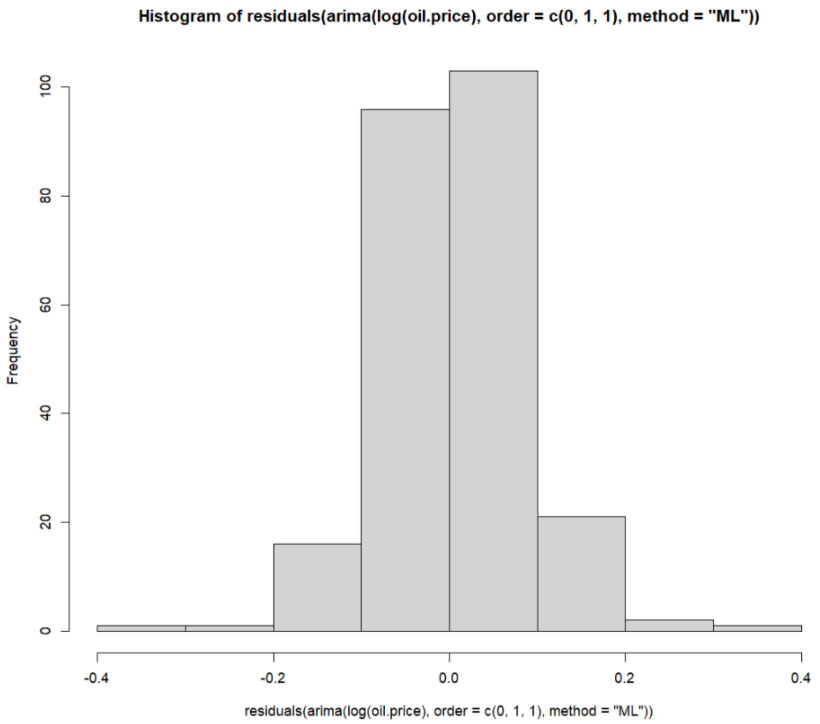
In this example, an ARIMA model is fitted to the 'oil.price' dataset, and the residuals are plotted using the acf function. The autocorrelation coefficients are plotted against the number of lags, and the dotted lines represent the upper and lower bounds of the 95% confidence interval for a white noise process. If the autocorrelation coefficients are inside the confidence interval, it indicates that the residuals are likely to be independent. If the autocorrelation coefficients are outside the confidence interval, it indicates that the residuals are not independent, and that the model may need to be modified to better capture the structure of the time series data.

In this case, the residuals of the ARIMA model appear to be independent, as the autocorrelation coefficients are inside the confidence interval.

A more accurate analysis of residuals can be represented by the following scatterplot, that shows the existence of four outliers as check before, that may be acceptable due to kind of the data I own since these outliers may belong to some economic or financial phenomenon that could have affected the oil prices even for a small period.



In addition, also the displaying of a histogram could be useful in terms of distribution, returning another confirm of the residuals' normal distribution.



Finally, having found a proper model for these modified data, I could proceed with the Forecasting phase.

Forecasting represents one of the primary objectives of building a model for a time series and it refers to the process of predicting future values of a time series based on its past values.

The goal of forecasting is to create a model that accurately captures the underlying patterns in the data and use that model to make predictions about future values and so, in these terms, it is a key tool for making predictions about future values of a time series, and it can help to inform decisions and planes based on the expected future behaviour of the time series.

Obviously, in this specific case it may not be useful to forecast future values since the dataset time period ends in 2006, so a very far period from now; but what could be interesting to analyse is to predict the 'future' values up to 2016 and to check if what the model predicts corresponds to what has effectively been happen during these years.

In addition, as I've already explained above, the negative values belonging to AIC and BIC, even if very good in terms of model fitting, may not be a good sign in terms of prediction accuracy and these problems effectively come out below.

By predicting the future values with respect to the series and having the possibility of comparing the predicted values with the actual values, I've had a confirm of this accuracy problem.

Here the main results in terms of forecasted values:

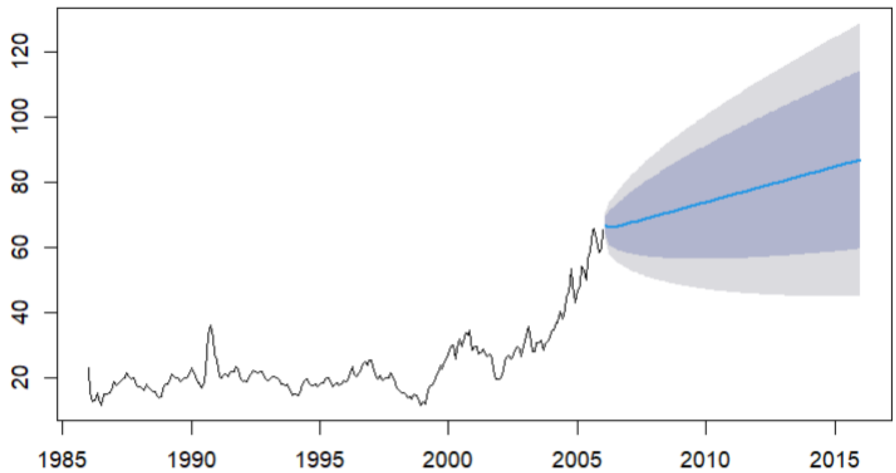
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Feb 2006	66.76524	64.08592	69.44457	62.66757	70.86292
Mar 2006	66.16306	61.94185	70.38426	59.70728	72.61883
Apr 2006	65.98593	60.89740	71.07445	58.20370	73.76816
May 2006	66.00322	60.26632	71.74013	57.22939	74.77706
Jun 2006	66.10945	59.82671	72.39219	56.50083	75.71807
Jul 2006	66.25635	59.48663	73.02608	55.90295	76.60975
Aug 2006	66.42186	59.20430	73.63942	55.38356	77.46016
Sep 2006	66.59588	58.95945	74.23230	54.91697	78.27478
Oct 2006	66.77379	58.74149	74.80608	54.48945	79.05812
Nov 2006	66.95348	58.54445	75.36250	54.09299	79.81397
Dec 2006	67.13398	58.36462	75.90333	53.72241	80.54555
Jan 2007	67.31486	58.19950	76.43021	53.37412	81.25559
Feb 2007	67.49590	58.04726	76.94455	53.04544	81.94636
Mar 2007	67.67703	57.90647	77.44759	52.73425	82.61981
Apr 2007	67.85819	57.77600	77.94038	52.43881	83.27757
May 2007	68.03937	57.65490	78.42383	52.15769	83.92104
Jun 2007	68.22055	57.54236	78.89874	51.88967	84.55143
Jul 2007	68.40174	57.43769	79.36578	51.63368	85.16980
Aug 2007	68.58293	57.34029	79.82556	51.38880	85.77706
Sep 2007	68.76412	57.24963	80.27861	51.15423	86.37401
Oct 2007	68.94531	57.16524	80.72538	50.92925	86.96137
Nov 2007	69.12650	57.08671	81.16629	50.71323	87.53977
Dec 2007	69.30769	57.01366	81.60172	50.50560	88.10978

Jan 2014	82.53462	57.98341	107.08582	44.98679	120.08244
Feb 2014	82.71581	58.03893	107.39268	44.97579	120.45583
Mar 2014	82.89700	58.09509	107.69890	44.96575	120.82824
Apr 2014	83.07819	58.15187	108.00450	44.95668	121.19970
May 2014	83.25938	58.20927	108.30948	44.94855	121.57021
Jun 2014	83.44057	58.26728	108.61385	44.94135	121.93979
Jul 2014	83.62176	58.32589	108.91763	44.93507	122.30845
Aug 2014	83.80295	58.38510	109.22081	44.92970	122.67621
Sep 2014	83.98414	58.44488	109.52340	44.92521	123.04307
Oct 2014	84.16533	58.50524	109.82543	44.92160	123.40906
Nov 2014	84.34652	58.56616	110.12688	44.91886	123.77418
Dec 2014	84.52771	58.62764	110.42778	44.91698	124.13845
Jan 2015	84.70890	58.68968	110.72813	44.91593	124.50188
Feb 2015	84.89010	58.75226	111.02794	44.91572	124.86447
Mar 2015	85.07129	58.81537	111.32720	44.91632	125.22625
Apr 2015	85.25248	58.87901	111.62594	44.91774	125.58722
May 2015	85.43367	58.94317	111.92416	44.91995	125.94738
Jun 2015	85.61486	59.00785	112.22187	44.92295	126.30677
Jul 2015	85.79605	59.07304	112.51906	44.92673	126.66537
Aug 2015	85.97724	59.13872	112.81576	44.93127	127.02321
Sep 2015	86.15843	59.20491	113.11196	44.93657	127.38029
Oct 2015	86.33962	59.27158	113.40767	44.94262	127.73663
Nov 2015	86.52081	59.33873	113.70290	44.94940	128.09222
Dec 2015	86.70200	59.40636	113.99765	44.95692	128.44709
Jan 2016	86.88319	59.47446	114.29193	44.96515	128.80124

As it's possible to see by comparing these results with the actual data displayed below, these forecast values aren't properly correct in terms of accuracy, due to the reasons just explained.

The predicted values follow an increasing trend, year by year, while the real values oscillate going up and down during this period.

Let's plot what it has been predicted:

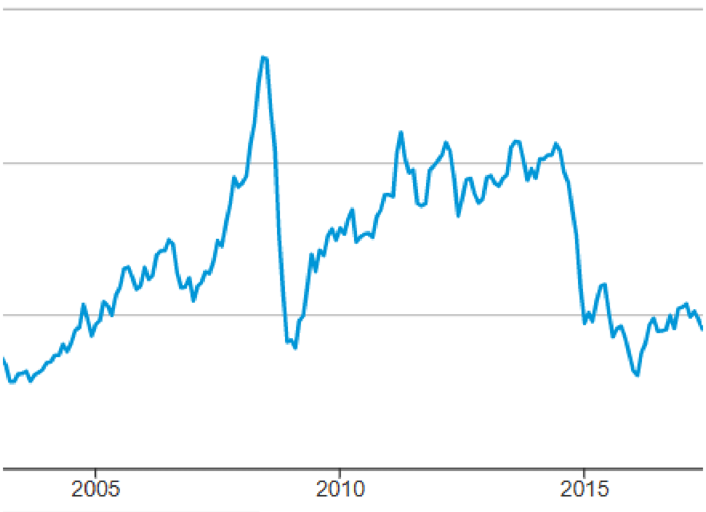


Nowadays, we are able to have information about these prices, and so I can confirm that this uptrend expectation in the years that follow the 2006 doesn't properly correspond to the actual data.

2005	46.84	48.15	54.19	52.98	49.83	56.35	59.00	64.99	65.59	62.26	58.32	59.41
2006	65.49	61.63	62.69	69.44	70.84	70.95	74.41	73.04	63.80	58.89	59.08	61.96
2007	54.51	59.28	60.44	63.98	63.46	67.49	74.12	72.36	79.92	85.80	94.77	91.69
2008	92.97	95.39	105.45	112.58	125.40	133.88	133.37	116.67	104.11	76.61	57.31	41.12
2009	41.71	39.09	47.94	49.65	59.03	69.64	64.15	71.05	69.41	75.72	77.99	74.47
2010	78.33	76.39	81.20	84.29	73.74	75.34	76.32	76.60	75.24	81.89	84.25	89.15
2011	89.17	88.58	102.86	109.53	100.90	96.26	97.30	86.33	85.52	86.32	97.16	98.56
2012	100.27	102.20	106.16	103.32	94.66	82.30	87.90	94.13	94.51	89.49	86.53	87.86
2013	94.76	95.31	92.94	92.02	94.51	95.77	104.67	106.57	106.29	100.54	93.86	97.63
2014	94.62	100.82	100.80	102.07	102.18	105.79	103.59	96.54	93.21	84.40	75.79	59.29
2015	47.22	50.58	47.82	54.45	59.27	59.82	50.90	42.87	45.48	46.22	42.44	37.19
2016	31.68	30.32	37.55	40.75	46.71	48.76	44.65	44.72	45.18	49.78	45.66	51.97

Cushing, OK WTI Spot Price FOB (Dollars per Barrel)

Graphically:



Where the blue line represents the Cushing, OK WTI Spot Price FOB (expression in dollars per barrel).

*These data have been found on [Cushing, OK WTI Spot Price FOB (Dollars per Barrel) (eia.gov)] from the Federal Reserve Economic Data (FRED).

Conclusions

Due to the non-stationarity condition of the series the best way to approach the problem seems to be the transformation of the data through the use of logarithm, applying then a first difference that has been sufficient for transforming the series in a stationary one.

The model that turned to be the best one in terms of goodness of fitting is the ARIMA(0, 1, 1) Model with the best measures of fitting on the data, and the best results in terms of residuals.

Another good model in terms of fitting was the ARIMA(1, 1, 0) Model, omitting the lag 4 term, that I tried to apply on the logarithm of the data, returning the following results:

```
Call:
arima(x = log(oil.price), order = c(1, 1, 0), method = "ML")

Coefficients:
         ar1
      0.2364
s.e.  0.0660

sigma^2 estimated as 0.006787:  log likelihood = 258.55,  aic = -515.11
```

This kind of model return good results in terms of goodness of fitting, but each measure returned to be a little bit worse than the model chosen for my analysis, with lower values of log likelihood, of AIC and a higher variance.

In terms of model's diagnostic, the residuals seem to respect all the conditions required for confirming that the model chosen is good.

The forecasting part is a critical point for this dataset for the reasons already explained, but the increasing trend predicted can be reasonable since in the period assumed for reference for the dataset a constant growth of the monthly price can be see year by year.

Generally, forecasting a non-stationary time series is difficult since it violates the assumption of stationarity commonly required by many models, but additional problem in these cases is to forecast time series for models that involve a transformation of the original series.

When a model involves a first difference to achieve a stationarity condition, as it was the case, we can consider two methods:

1. Forecasting the original non-stationary series
2. Forecasting the stationary differenced series and then undoing the difference by summing to obtain the forecast in original terms.

However, trying to apply this second strategy to my model, by undoing the differences to obtain the forecast in original terms, the results got even worse, predicting an even higher increase of the prices during the period assumed for reference.

Source of Dataset: 'Cryer JD an Chan K-S(2010) - Time Series Analysis: with applications in R, 2nd Edition' - page 473.