Maastricht University

School of Business and Economics

EBC4097 Quantitative Techniques for Financial Economics

GROUP ASSIGNMENT

WHICH TIME SERIES MODEL OFFERS THE BEST PERFORMANCE IN FORECASTING OIL PRICES RETURNS?

School of Business and Economics, Maastricht University

Course number: EBC4097

TABLE OF CONTENTS

1. INTRODUCTION

Oil and more in general fossil fuels have significant consequences for the global economy. With a proliferating population worldwide, oil has always represented a vital energy source to meet the needs of world economies: it has been the single greatest source of energy since the mid-1950s. The derivates of oil still underpin modern society.

Being oil a vital commodity for the global market, one can think that its prices can be rather volatile. Moreover, geopolitical events like the ones that have been experienced lately can have the potential to disrupt the flow of oil to the market, and subsequently, leading to uncertainty about future supply and demand.

Keeping the aforementioned statements, this paper aims to present the reader a time series analysis concerning oil price returns from 1986 until 2020. Specifically, this paper examines the returns related to WTI oil prices. The West Texas Intermediate (WTI) is one of the three primary oil benchmarks (along with Brent and Dubai crude) and refers to a specific grade of crude oil sourced primarily in Texas.

The ultimate goal of our research is to assess the forecasting performance of different time series models, providing a panoramic that could help in choosing the appropriate model in terms of accuracy when trying to predict oil prices returns. In particular, the present study is focused on investigating whether a single time series model can be considered as best choice when forecasting oil price returns in the proposed setting.

First, the paper investigates the auto regressive moving average (ARMA) models. Second, we try acquainting the reader with a vector autoregressive (VAR) model by including explanatory variables into the model, namely supply of crude oil and petroleum products and inflation.

Third, an inspection about the effects of the application of rolling window methodology on the previously explored time series models is proposed.

Finally, analysis and comparisons between the forecasting performance of the implemented methodologies are provided in order to reach a conclusion with respect to the objective of the study.

2. DATA DESCRIPTION

We use data about WTI oil prices with monthly frequency of the period between January 1986 and January 2020 in our analysis. The dataset is retrieved from the U.S. Energy Information Administration and contains 409 non-seasonally adjusted price observations measured in dollars per barrel[1].

The series considers spot prices, which can be defined as the current delivery price of a commodity traded in the spot market, in which goods are sold for cash and delivered immediately[2].

[1] *U.S. Energy Information Administration, Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma [MCOILWTICO], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/MCOILWTICO*

[2] *Nasdaq Stock Market (NASDAQ); https://www.nasdaq.com/glossary/s/spot-price*

Figure 1 shows a huge spike in prices followed by a significant decrease during the 2008 global financial crisis. From a preliminary visual analysis, we can suspect our series to be non-stationary.

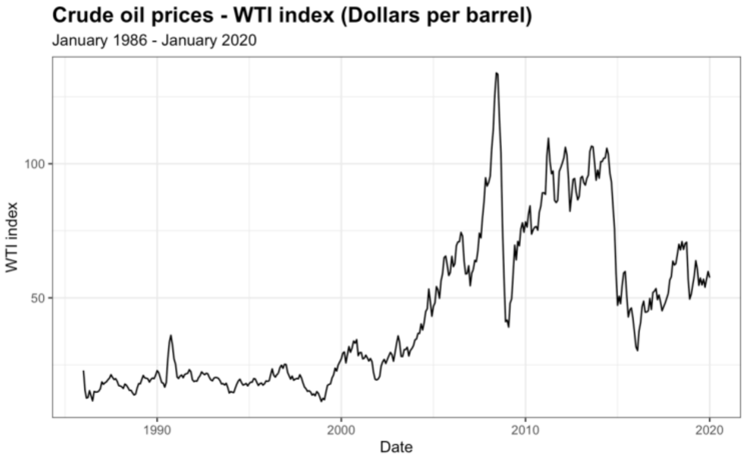


Figure 1: Monthly crude oil prices (WTI index) between Jan 1986 and Jan 2020 – not seasonally adjusted.

We proceed by looking for data about additional explanatory variables used in the VAR model, such as the supply of crude oil and petroleum products in the USA and the Consumer Price Index (CPI) as a measure of inflation. We were able to retrieve the needed information regarding supply from a collection of data published by the U.S. Energy Information Administration[3]. With petroleum products we refer to what is obtained from the processing of crude oil, natural gas, and other hydrocarbon compounds.

The dataset refers to the monthly amount of barrels produced and supplied in the USA from January 1986 to January 2020 (409 observations). This variable is measured in thousands of barrels (a barrel is a unit of volume equal to 42 U.S. gallons). Note that data are not seasonally adjusted.

When looking at the visual representation of our data (Figure 2), we can see that the series exhibits some kind of seasonality pattern other than being non-stationary.

[3] *U.S. Energy Information Administration, U.S. Product Supplied of Crude Oil and Petroleum Products [U.S. Product Supplied of Crude Oil and Petroleum Products Thousand Barrels; https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=MTTUPUS1&f=M*

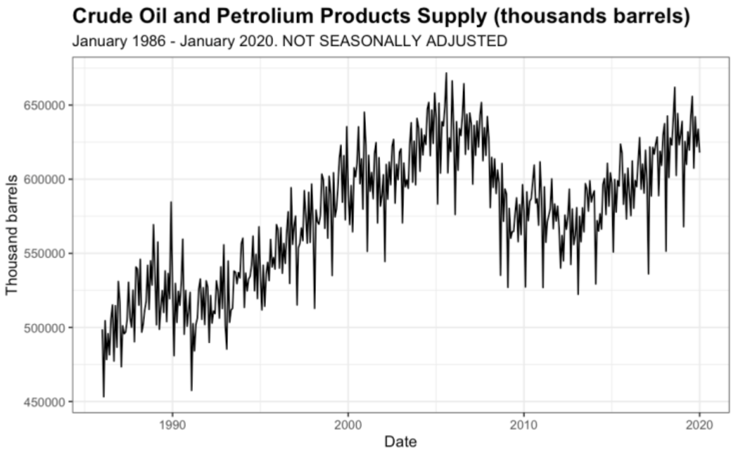


Figure 2: Supply of crude oil and petroleum products (January 1986 – January 2020) – not seasonally adjusted.

For the inflation measure, we gather information about the monthly Consumer price index (CPI) development for all urban consumers. The indicator is a seasonally adjusted price index of a basket of goods and services paid by urban consumers. All the details and information about the CPI index were published by the U.S. Bureau of Labor Statistics.

This particular CPI index includes roughly 88 percent of the total population, accounting for wage earners, clerical workers, technical workers, self-employed, short-term workers, unemployed, retirees, and those not in the labor force[4].

The U.S. Bureau of Labor Statistics states that percent changes in the price index measure the inflation rate between any two time periods.

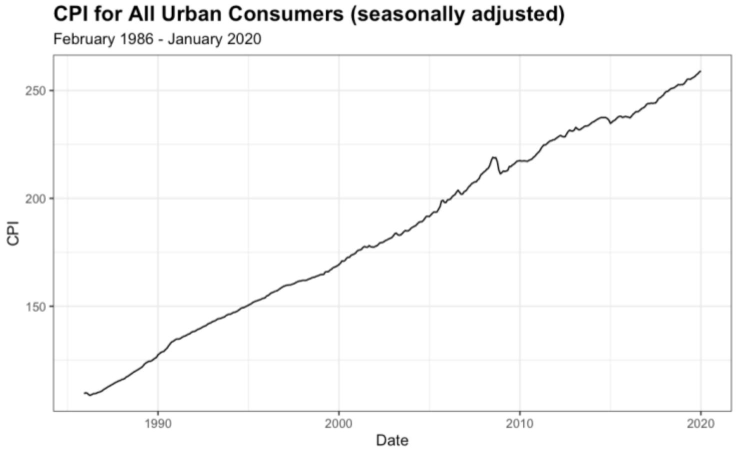


Figure 3: CPI for All Urban Consumers (January 1986 – January 2020) - seasonally adjusted.

[4] *U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis;* [*https://fred.stlouisfed.org/series/CPIAUCSL*](https://fred.stlouisfed.org/series/CPIAUCSL)

The dataset that will be used during our analysis refers to 409 observations that attain to monthly measurements about the considered Consumer Price Index.

From the visual representation provided in Figure 3, we can observe that the time series is characterized by an upward trend and we can suspect it to be non-stationary.

3. METHODOLOGY

In this section we will go through all the procedures we employ in order to conduct our research. The main scheme followed in each step consists in testing for stationarity, differencing (to make series stationary) and explaining the choice of the model that could fit better our purposes. We will try to corroborate our decisions through statistical evidence.

3.1 ARMA AND AR MODEL

As demonstrated in the previous section (see Figure 1 presented in the "DATA DESCRIPTION" section), the oil prices series clearly is non-stationary from a visual point of view. The first step will be testing for the presence of a unit root and making it stationary, if necessary.

Since the gathered data about oil prices are not seasonally adjusted, a decomposition of the time series is performed. As Figure 4 shows, no clear seasonal pattern is exhibited. Therefore, we decide to proceed without adjusting for seasonality.

In order to corroborate our analysis with the visual aspect of the series being stationary, we run the Augmented Dickey Fuller (ADF test). The ADF test is a commonly used test which tests the presence of a unit root. The value of the ADF t-statistic obtained is -3.1562, as shown in Figure 1.A in Appendix. Therefore, we fail to reject the null hypothesis since the critical value for the 5 percent significance level is -3.42. This means that the series has a unit root, being the test a left tailed test.

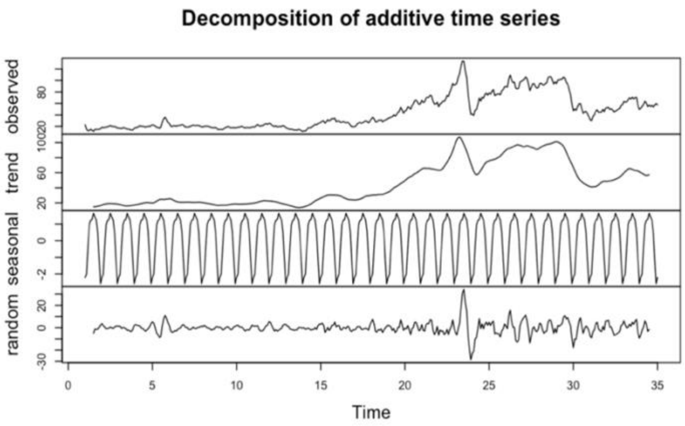


Figure 4: Decomposition of oil price time series. The plot above shows the original time series (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated irregular component (bottom)

We proceed by, first, taking the log transformation of the oil prices time series with the aim of stabilizing the variance and, second, taking the first difference in order to make it stationary. The series resulting from this procedure is depicted in Figure 5.

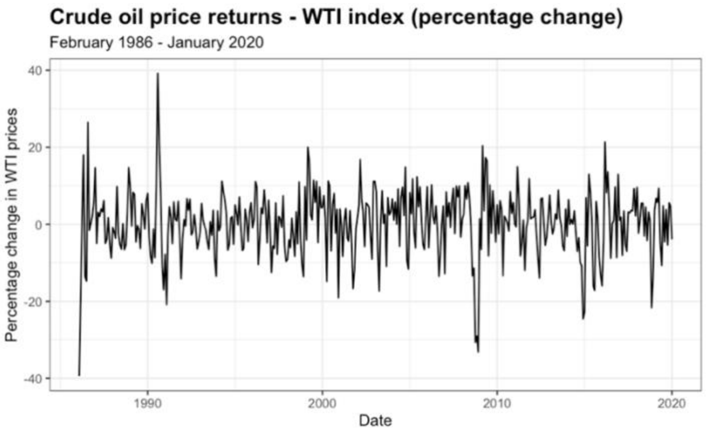


Figure 5: Monthly crude oil price returns (WTI index) between February 1986 and January 2020

The dataset looks stationary. In order to confirm this visual hypothesis, we must once again run the ADF test. Indeed, the value of the t statistic is -13.0216 (see Figure 2.A in Appendix), which lies to the left of the 5% critical value that delimits the rejection region (-3.42). Therefore, we can reject the null hypothesis that the series has a unit root: the series is now stationary and captures oil price returns.

By looking at the autocorrelation function (ACF) plot and partial autocorrelation function (PACF) plot in Figure 6, we can see that our series is now stationary. However, the two plots do both show a sinusoidal pattern, suggesting that an autoregressive moving average model (ARMA) could be appropriate for our case.

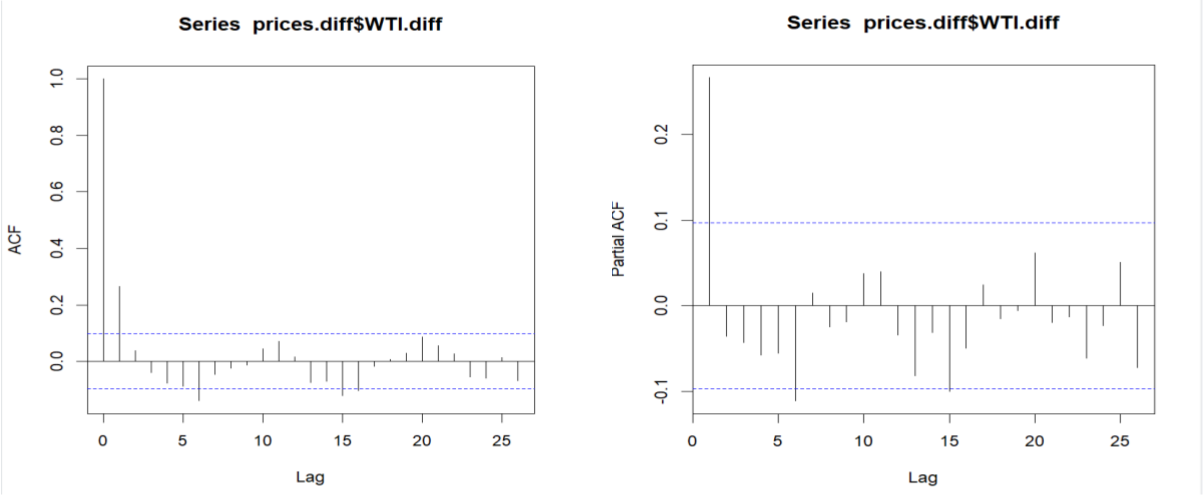


Figure 6: ACF plot (left) and PACF plot (right) for oil price returns series

Each of the two correlograms shows one significant spike at lag one. This can be interpreted as a sign of the fact that an ARMA(1, 0, 1) model can be taken into consideration.

In an autoregressive model (AR), we forecast the variable of interest using a linear combination of past values of the variable. Rather than using past values of the forecast variable in a regression, a moving average model (MA) uses past forecast errors in a regression-like model. An ARMA model combines the features of an AR and a MA model (Hyndman, R.J., Athanasopoulos, G., 2018). For our analysis, we consider oil price returns as a dependent variable. Conceptually: in our case the AR part could model the lagged past oil price returns, and the MA part could capture the shocks not explained in the AR part of the model.

We start by implementing the auto.arima function in R, which suggests the best possible ARMA model after taking as input the maximum order of the desired model parameters (namely p for the AR part and q for the MA part; d in our case is not considered since our series is already stationary and doesn't need differencing). The auto.arima function determines that the ARMA(1, 0, 0) model is the best when setting the maximum order for both p and q parameters equal to 2. Note that we can refer to the ARMA(1, 0, 0) model as an AR(1) model. To be certain that the AR(1) is the best model for forecasting, we estimate it along with a few other models, in order to have an overview about which ones could fit better our forecasting purpose.

The ARMA(1, 0, 0) returns the $ar1$ lag to be significant. The ARMA(1, 0, 1) model returns that both the $ar1$ and $ma1$ lags are statistically insignificant. We then proceed to testing ARMA(2, 0, 1) where all lags, namely $ar1$, $ar2$ and $ma1$, are significant. Next, we test the ARMA(2, 0, 2) where only the $ar1$, and $ma1$ lags are significant. Finally, we test the ARMA(1, 0, 2) for which it is observed that no coefficient is significant. A summary of the findings is provided in Figure 3.A contained in Appendix.

The use of information criteria is another way that can guide us through the selection of which model could fit better our data. Two types of information criteria utilised in our research are the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). Both criteria suggest implementing those models which return a lower value. The best models are the ARMA(2, 0, 1), when considering AIC, and the AR(1) when using the BIC. Since these two models turn out to be the best in terms of the two indicators mentioned, we use them for further analyses. A table summarising the values for the two information criteria is provided here below.

MODEL	*AIC*	*BIC*
ARMA(1,1)	2893.808	2909.853
ARMA(1,2)	2895.593	2915.649
ARMA(2,1)	2888.295	2908.351
ARMA(2,2)	2890.081	2914.149
AR(1)	2892.139	2904.172

Table 1: BIC and AIC values summary for all the ARMA models tested

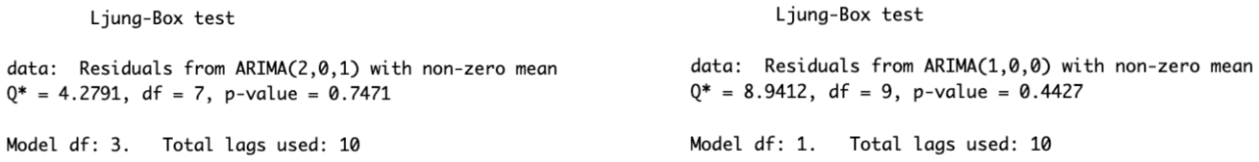
Before proceeding, the equations for the two chosen model must be specified, where r labels our dependent variable, which are the oil price returns:

AR(1) model: $r_t = \beta_0 + \beta_1 r_{t-1} + \varepsilon_t$

ARMA(2, 0, 1) model: $r_t = \beta_0 + \beta_1 r_{t-1} + \beta_2 r_{t-2} + \emptyset_1 \varepsilon_{t-1} + \varepsilon_t$

We proceed by testing for autocorrelation in the residuals of both models. The chosen test is the Ljung Box Text, in which the null hypothesis states that there is no autocorrelation in the residuals. A p-value higher than the 5 percent significance level leads to the null hypothesis not being rejected, meaning that there is no autocorrelation in residuals and therefore they're white noise.

If we look at Figure 7, the p-values attained indicate that we fail to reject the null hypothesis in both cases and we can conclude that there is no autocorrelation in the residuals for the AR(1) model and for the ARMA(2, 0, 1) model. Visually, the residuals seem to not exhibit autocorrelation and the correlogram for residuals appear to represent the case in which all spikes are within the threshold, suggesting errors are not autocorrelated (see Figure 4.A in Appendix).



```
	Ljung-Box test

data:  Residuals from ARIMA(2,0,1) with non-zero mean
Q* = 4.2791, df = 7, p-value = 0.7471

Model df: 3.   Total lags used: 10
```

```
	Ljung-Box test

data:  Residuals from ARIMA(1,0,0) with non-zero mean
Q* = 8.9412, df = 9, p-value = 0.4427

Model df: 1.   Total lags used: 10
```

Figure 7: Output from Ljung-Box test for ARMA(2, 0, 1) (left) and AR(1) (right)

3.2 VAR MODEL

For the next step of our analysis, we focused on inspecting whether the forecasting performance could be improved by exploring a multivariate analysis. The model chosen to test this hypothesis is the vector autoregressive (VAR) model, which is specifically used to model inter-dependencies between economic time series. Let's point out that "a criticism that VARs face is that they are atheoretical; that is, they are not built on some economic theory that imposes a theoretical structure on the equations" (Hyndman, R.J., Athanasopoulos, G., 2018). However, in this research we will not go deep into interdependencies, but we will pursue our main goal of evaluating the predictive performance of the model in forecasting oil price returns.

The two additional explanatory variables, added in order to model them jointly with oil price returns, were percentage change in crude oil supply and inflation. The further analysis is dedicated on determining whether using these two variables improves forecasting performance, compared to the AR and ARMA models previously examined.

Oil supply

We choose supply of crude oil and petroleum products in the US, from here on simply labelled as “oil supply”, as an additional explanatory variable because it is part of the supply side of oil. According to economic theory, prices are negatively related to supply. Therefore, assuming demand to stay constant, a decrease in oil extraction should be reflected in a rise in oil prices.

Since VAR models are multivariate extensions of standard autoregressive (AR) models, we need stationary series.

A visual test of the raw data shows non-stationarity (see Figure 2 in “DATA DESCRIPTION” section). Furthermore, since our data were not seasonally adjusted, a decomposition analysis indicates a seasonality pattern (Figure 8), which we adjust for.

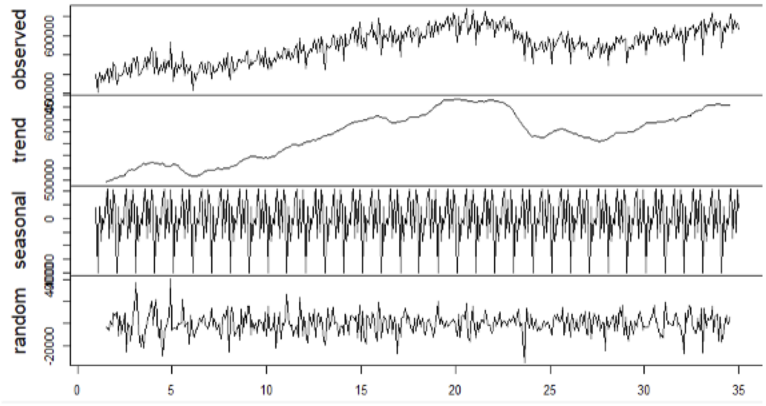


Figure 8: Decomposition of supply of crude oil and petroleum products time series

At this point, the Augmented-Dickey-Fuller test is applied to oil supply data to test for the presence of a unit root. Since the statistical value of -3.0963 is above the critical value for 5 percent significance value (-3.42), the null hypothesis of the presence of a unit root cannot be rejected (see Figure 5.A in Appendix). To make the series stationary, we proceed by taking the log difference. The resulting differenced series allows us to reject the null hypothesis set by the Augmented Dickey Fuller test for the presence of a unit root: the value of statistical test (-25.0353) lies in the rejection region delimited by the 5 percent significance critical value (-3.42; see Figure 6.A in Appendix). The variable $supply$ refers to seasonally adjusted month-over-month percentage changes in supply.

A visual analysis of the time series can help us corroborate the stationarity (see Figure 9).

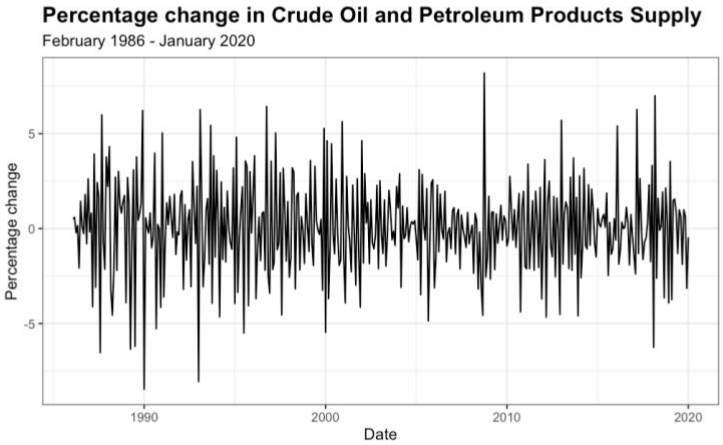


Figure 9: Percentage change in crude oil and petroleum products supply

Inflation

We choose inflation as a third variable for the VAR model. The reason for this is, because a spike in commodity prices reflected in inflation, could also be related to the oil price returns. The dataset used as an indicator for inflation is the monthly consumer price index (CPI) for all urban consumers in the USA. The CPI shows a long upward trend in the past 40 years (see Figure 3 in "DATA DESCRIPTION" section). To align to commonly used inflation calculation, we calculate monthly (non-annualised) inflation as follows:

$$\pi_m = \frac{CPI_m - CPI_{m-1}}{CPI_{m-1}} * 100$$

The monthly percentage change in CPI, which from now on we will refer to as "inflation" (π), passes the visual stationarity test (see Figure 10). When running the Augmented Dickey Fuller test on the data, the null hypothesis gets rejected (see Figure 7.A in Appendix) since the test-statistic value (-13.3901) lies in the rejection area delimited by the critical value of -3.42 corresponding to 5 percent significance level. The series is therefore already stationary and needs no further processing.

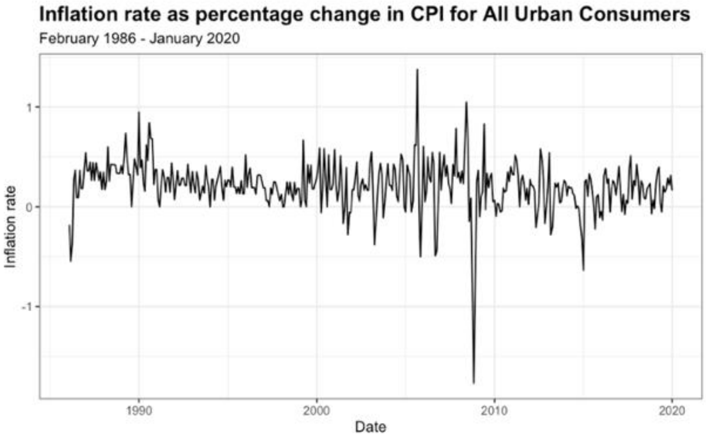


Figure 10: Inflation rate as percentage of CPI

To summarize, we implement the VAR model considering oil returns r (log differences of oil prices), $supply$ (log difference of monthly supply data) and inflation π (percentage change of monthly CPI) as endogenous variables.

To select the optimal number of lags to be used in the VAR model we use the VAR.select function from vars R package, which suggests time lags according to different information criteria, namely AIC, HQ, BIC and FPE. Each information criterion suggests implementing a model with two time lags (see Figure 8.A. in Appendix) and, therefore, we proceed by constructing a VAR(2) model. Our model equations with two time lags and three explanatory variables can be expressed as follows:

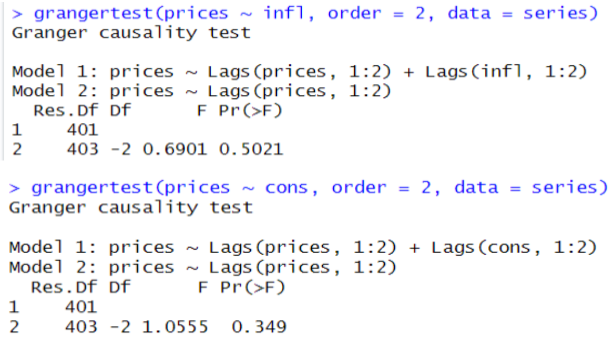
$$r_t = \beta_0 + \beta_{11,1} r_{t-1} + \beta_{11,2} r_{t-2} + \beta_{12,1} supply_{t-1} + \beta_{12,2} supply_{t-2} + \beta_{13,1} \pi_{t-1} + \beta_{13,2} \pi_{t-2}$$

$$supply_t = \beta_0 + \beta_{21,1} r_{t-1} + \beta_{21,2} r_{t-2} + \beta_{22,1} supply_{t-1} + \beta_{22,2} supply_{t-2} + \beta_{23,1} \pi_{t-1} + \beta_{23,2} \pi_{t-2}$$

$$\pi_t = \beta_0 + \beta_{31,1} r_{t-1} + \beta_{31,2} r_{t-2} + \beta_{32,1} supply_{t-1} + \beta_{32,2} supply_{t-2} + \beta_{33,1} \pi_{t-1} + \beta_{33,2} \pi_{t-2}$$

When looking at the significant coefficients, the only variable that is statistically significant when modelling returns r_t is the oil return itself at time lag 1 (r_{t-1}) (see Figure 9.A in Appendix).

To test whether variables are useful in predicting the others, a Granger causality test is performed. The Granger test leads us to accept the null hypothesis of no Granger causality in both cases for inflation and supply, given that the p-value is greater than 5 percent significance level. This confirms that neither of them seems to help predicting oil price returns (see Figure 11).

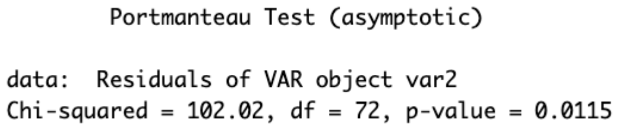


```
> grangertest(prices ~ infl, order = 2, data = series)
Granger causality test

Model 1: prices ~ Lags(prices, 1:2) + Lags(infl, 1:2)
Model 2: prices ~ Lags(prices, 1:2)
  Res.Df Df      F Pr(>F)
1    401
2    403 -2 0.6901 0.5021

> grangertest(prices ~ cons, order = 2, data = series)
Granger causality test

Model 1: prices ~ Lags(prices, 1:2) + Lags(cons, 1:2)
Model 2: prices ~ Lags(prices, 1:2)
  Res.Df Df      F Pr(>F)
1    401
2    403 -2 1.0555  0.349
```

Figure 11: Output of the Granger causality test

To illustrate the impact of a unit shock in both inflation and supply variables on oil price returns, we create an impulse response function. The impulse response function indicates that a shock in inflation or supply has only a limited effect on oil returns. The red lines depict the high statistical uncertainty in the graph (see Figure 10.A in Appendix).

Furthermore, to test whether the model is able to capture all relevant information we apply a Portmanteau test, whose null hypothesis of no serial correlation in residuals is rejected (see Figure 12): the p-values for the test is lower than 5 percent significance level. This implies that there exists serial correlation in residuals. For modelling purposes, to further reduce correlation in residuals, including more variables or time lags, has to be considered. However, our main focus is to forecast oil returns and not modelling interdependencies: since the VAR.select function proposed to use two lags, we decided not to extend the time lags. Moreover, adding additional lags could lead to overfitting.



```
                Portmanteau Test (asymptotic)

data:  Residuals of VAR object var2
Chi-squared = 102.02, df = 72, p-value = 0.0115
```

Figure 12: Portmanteau test on VAR(2) model residuals. The low p-value leads us to reject the null hypothesis of no serial correlation when applied to residuals

To summarize, the constructed VAR model, which includes inflation and supply as additional variables, turns out to exhibit issues in terms of significance of coefficients and in terms of Granger causality. As mentioned previously in the introduction, the primary motivation about implementing a VAR model was trying to improve the forecasting performance related to returns, so we proceed by checking whether adding supply and inflation can still (even though insignificant) help in this sense. It is important to mention, that adding other variables could lead to a construction of a better model. However, due to time constraints, we proceed in our analysis with this setting: incorporating other variables to be used in the VAR model is left for future research.

3.3 ROLLING WINDOW ANALYSIS

A rolling analysis of a time series model is often used to assess the model's stability over time (Zivot E., Wang, J., 2003). Usually when a model is implemented, the assumption that its parameters remain constant is made. However, data can capture different scenarios and changes in the environment, making this assumption not ideal: rolling window analysis can help considering these changes. In our research we use this methodology to pursue the main goal of the research question: evaluating forecasting performance of the chosen models.

The standard procedure that is carried out when forecasting, involves splitting the whole data available into two subsets: an estimating sample and a predicting sample. The model is fitted on the first subsample, then h-period ahead predictions are generated. Given that the predicting sample contains observed values of the considered dependent variable values, a prediction error measure can be generated.

The rolling window analysis however implies a different approach. Such an analysis is commonly used to backtest a statistical model on historical data to evaluate stability and predictive accuracy (Zivot E., Wang, J., 2003). Backtesting generally works in the following way: a restricted number of consecutive observations (to which we refer as window) starting from the oldest one is considered at a time. The unique feature of the rolling analysis consists in making the window "rolling" ahead by a given increment (which from now on will be labeled as "step"). The window is rolled by dropping the oldest observations contained in it for a number corresponding to the chosen increment, then the same amount of more recent consecutive observations is added in order to create a new window with the same size as the previous one. The described procedure is repeated until no more predictions can be computed, meaning that all the available data have been considered. Figure 13 provides a visual representation of the process in order to have a better understanding of the implications.

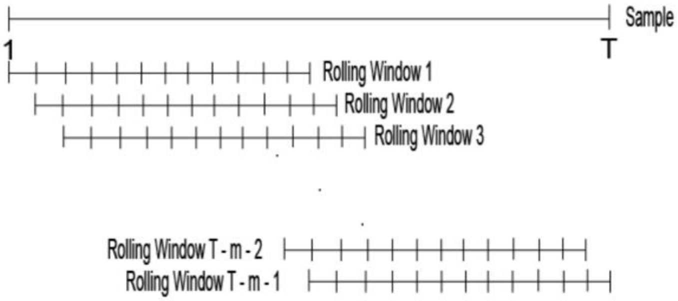


Figure 13: Visual representation of how windows are "rolled" ahead in rolling window analysis with an increment equal to 1. Source: MathWorks.com

Each window is split into two subsets as usual when performing forecasting: an estimation and a predicting sample. The model is fitted on the estimating subsample, then predictions are generated. The predicting sample contains h observations, so h predictions are generated and consequently an error measure is computed. Since multiple windows are generated throughout the process, the chosen model is fitted multiple times on different estimating samples and therefore its parameters will change across the analysis.

Supposing to consider the mean squared forecasting error (MSFE) as the chosen error measure, it will be computed according to the following formula for each single window:

$$MSFE = \frac{1}{h} \cdot \sum_{i=m-h+1}^{m} (\hat{y}_i - y_i)^2$$

where h represents the number of periods for which predictions are computed (corresponding to the number of observations in the predicting sample), m represents the window size, $\hat{y}_i$ the generated prediction for the i-th observation and y_i the observed value for the same observation.

Given that each window allows to calculate a prediction error, the overall error for the rolling analysis will be given by the mean of all the ones generated through the whole procedure.

Let the number of generated windows be K, the MSFE for the rolling analysis applied to the time series model will be:

$$MSFE = \frac{1}{K} \cdot \sum_{i=1}^{K} MSFE_i$$

where $MSFE_i$ is the mean forecasting error for the i-th window generated during the procedure.

Since there's no objective rule or analytical formula that can be followed to determine neither the appropriate window size nor appropriate step size, we decide to perform a recursive implementation of the rolling analysis to find out which combination of the two parameters allows to return the lowest MSFE. First, a range of values is chosen for both the two parameters, then a grid containing all the possible combinations between step size values and window size values is generated. The rolling analysis is then run multiple times using a different combination for each iteration. The term iteration refers to the single repetition of the computational procedure described before, in which a couple of parameters is used.

Performing the analysis in this way allows to return a MSFE for each single combination of parameters. The obtained error measures are then compared and the combination corresponding to the lowest one is considered the optimal combination of parameters to be considered.

The motivation lying behind running such a procedure is that different combinations of parameters can affect the forecasting performance of the model:

- a small step could generate many overlapping windows leading to high computational cost, while a huge step could preclude the possibility of capturing certain changes in the data;
- a longer window could return smoother estimates compared to a shorter one since a larger window size allows to build a lower number of windows throughout the analysis.

The analysis displayed in this section is applied to all the models that have been analyzed in the present study: AR(1), ARMA(2, 0, 1) and VAR(2).

4. DISCUSSION OF RESULTS

Finally, we generate forecasts by implementing the models and the procedures presented until now.

The standard forecasting procedure is straightforward. The considered dataset is split into two subsets: an estimating sample and a predicting/test. The former is used to fit the model, the latter instead is used to assess the predicting performance.

Forecasting accuracy is then evaluated with different measures. For our purposes we use the Mean Squared Forecasting Error metric (MSFE).

MSFE is computed as follows:

$$MSFE \ = \frac{1}{H} \sum_{i=T-H+1}^{T} (\hat{y}_i - y_i)^2$$

where H is the chosen forecasting horizon (corresponding to the number of observations in the predicting sample), T is the number of total observations in the time series, $\hat{y}_i$ the generated prediction for the i-th observation and y_i the observed value for the i-th observation contained in the predicting sample. MSFE is measured as the mean of squared residuals.

For the present analysis two different forecasting horizons are chosen, with the aim of evaluating the performance of the explored models and methodologies with different settings. The considered horizons are 50 periods and 20 periods.

Table 2 summarizes the MSFE errors obtained by performing forecasting on both the selected forecasting periods.

MODEL	*50-PERIOD HORIZON*	*20-PERIOD HORIZON*
AR(1)	63.04929	60.84797
ARMA(2, 0, 1)	**62.90668**	**58.86873**
VAR(2)	63.14641	61.48843

Table 2: MSFE values referring to AR(1), ARMA(2, 0, 1) and VAR(2) models for a 20-period forecasting horizon and a 50-period forecasting horizon

Looking at the results in the table, one thing should be noted: MSFEs are generally lower for the 20-period forecasting horizon compared to the ones obtained when trying to run forecasts on 50 periods. This finding seems to suggest that the explored models tend to perform better when a shorter forecasting horizon is considered. Moreover, the best performing model according to the error measures is ARMA(2, 0, 1) with reference to both the chosen horizons.

The VAR(2) model turns out to be the one that provides the worst forecasting accuracy between the models here considered, even though only slightly worse than AR(1) model. The VAR model constructed with supply and inflation as additional explanatory variable therefore does not offer an improved performance compared to the examined univariate models.

We proceed by examining the results coming from the rolling analysis. The first remark that has to be made regards the optimal combination of parameters (namely window size and step size) that has to be used when performing the analysis. For our purposes, we arbitrarily decided to run the iterative procedure displayed in section “3.3 ROLLING WINDOW ANALYSIS” with the following set of values:

- from 100 to 300 (every 10 numbers) for the window size;
- from 1 to 12 for the step size.

MODEL	*50-PERIOD HORIZON*	*20-PERIOD HORIZON*
AR(1) ROLLING WINDOW	**29.68278**	31.30085
ARMA(2, 0, 1) ROLLING WINDOW	29.89157	31.32429
VAR(2) ROLLING WINDOW	29.96265	**31.05293**

Table 3: MSFE values referring to rolling window analysis applied to AR(1), ARMA(2, 0, 1) and VAR(2) models for a 20-period forecasting horizon and a 50-period forecasting horizon

The iterative procedure returns 110 for window size and 12 for step size as the optimal combination for every considered model, as it allows to return the lower MSFE. The combination is defined ideal referring to both chosen forecasting horizon.

Table 3 displays a summary of the MSFEs obtained by running the rolling analysis with the optimal set of parameters on the three inspected models. The table shows a reverted outcome with respect to Table 2: the error measures tend to be lower when the analysis is performed to generate predictions for a 50-period forecasting horizon.

In this second setting there's no unique model that performs better than the others when generating predictions for both horizons: AR(1) performs better when a 50-period horizon is considered, while the VAR(2) model performs better when a 20-period horizon is considered.

5. CONCLUSIONS

The present study provides some insights about comparison between predicting performances for multiple time series models when forecasting oil prices.

Looking at the results, the study suggests that the explored models offer a better performance when generating prediction for a smaller period: as seen in the previous section, it is clear how all the models perform more effectively in terms of forecasting accuracy with a 20-period horizon compared to the 50-period horizon.

In both cases, ARMA(2, 0, 1) offers the best forecasting performance according to the considered error measure (MSFE).

Rolling window analysis opens a different scenario, in which the considered models offer a better performance when generating prediction for a longer period. Furthermore, AR(1) model shows a better forecasting accuracy for a 50-period horizon, while VAR(2) seems to be the best-performing model when a 20-period horizon is considered.

Given these findings, the present research leads to the conclusion that there is no univocal model that can be declared as the best performing model when forecasting oil prices returns. As it can be seen from the presented results, the model choice depends on a plurality of factors, namely the considered forecasting horizon and the methodology with which the predicting procedure is carried out.

One limitation of the approach adopted here is connected to the implemented VAR model. It has been presented how the additional variables chosen to implement the model do not seem to help predicting oil price returns. To enrich the presented study, different additional explanatory variables could be considered in order to improve the model and test its forecasting performance. Another limitation concerns the choice in parameters for the rolling window analysis: it was stated how the range of values for window size and step size considered to determine their optimal combination were arbitrarily chosen. A different range of values could lead to different results, so this has to be taken into consideration if trying to improve the research.

To conclude, the approach proposed considered only two forecasting horizons: additional forecasting periods could be examined in order to evaluate model performances and accuracy. Performing forecasting on short-term horizons or on different horizon in general can represent an upgrade for our analysis.

REFERENCES

Hyndman, R.J., Athanasopoulos, G. (2018). *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia.

https://otexts.com/fpp2/ . Accessed on October 12th 2023.

James, G., Witten, D., Hastie T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, New York, NY.

https://doi.org/10.1007/978-1-0716-1418-1

MathWorks.com

https://www.mathworks.com/help/econ/rolling-window-estimation-of-state-space-models.html#buhn26v . Accessed on October 12th 2023.

Nasdaq Stock Market (NASDAQ).

https://www.nasdaq.com/glossary/s/spot-price . Accessed on October 12th 2023.

Zivot, E., Wang, J. (2003). *Rolling Analysis of Time Series. In: Modeling Financial Time Series with S-Plus®*. Springer, New York, NY.

https://doi.org/10.1007/978-0-387-21763-5_9

DATA SOURCES

U.S. Energy Information Administration, Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma [MCOILWTICO], retrieved from FRED, Federal Reserve Bank of St. Louis.

https://fred.stlouisfed.org/series/MCOILWTICO

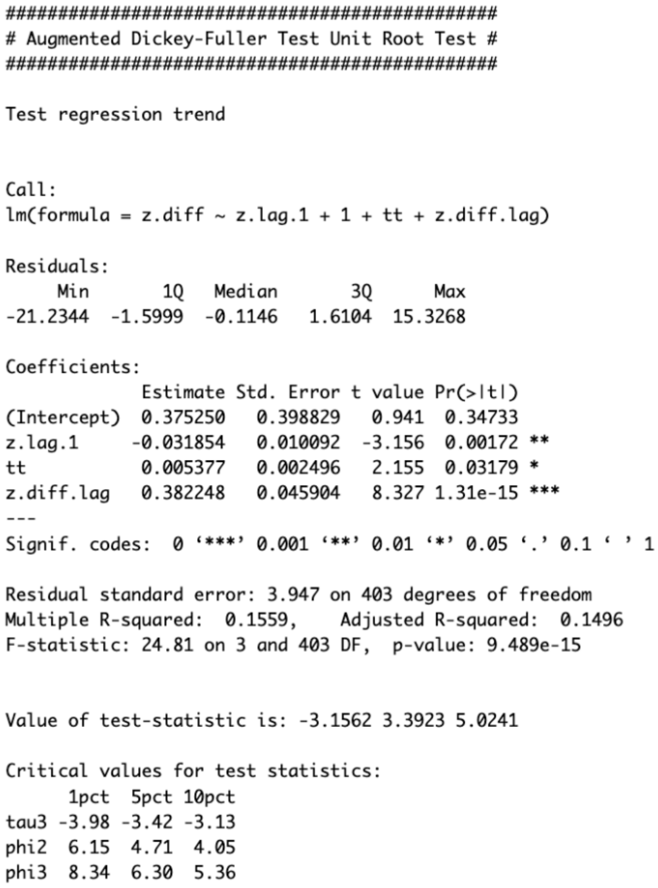
U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis.

https://fred.stlouisfed.org/series/CPIAUCSL

U.S. Energy Information Administration, U.S. Product Supplied of Crude Oil and Petroleum Products [U.S. Product Supplied of Crude Oil and Petroleum Products Thousand Barrels].

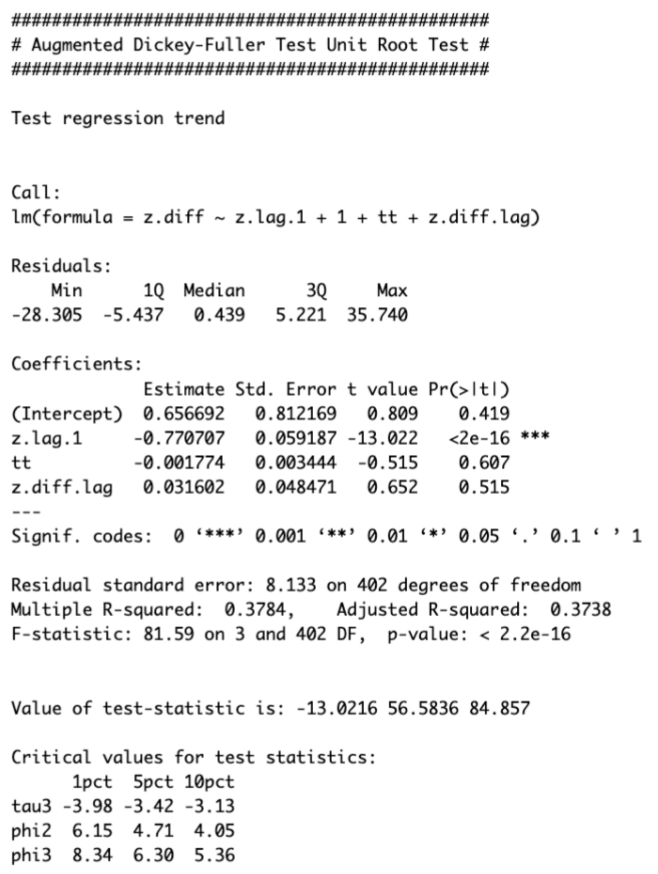
https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=MTTUPUS1&f=M

APPENDIX



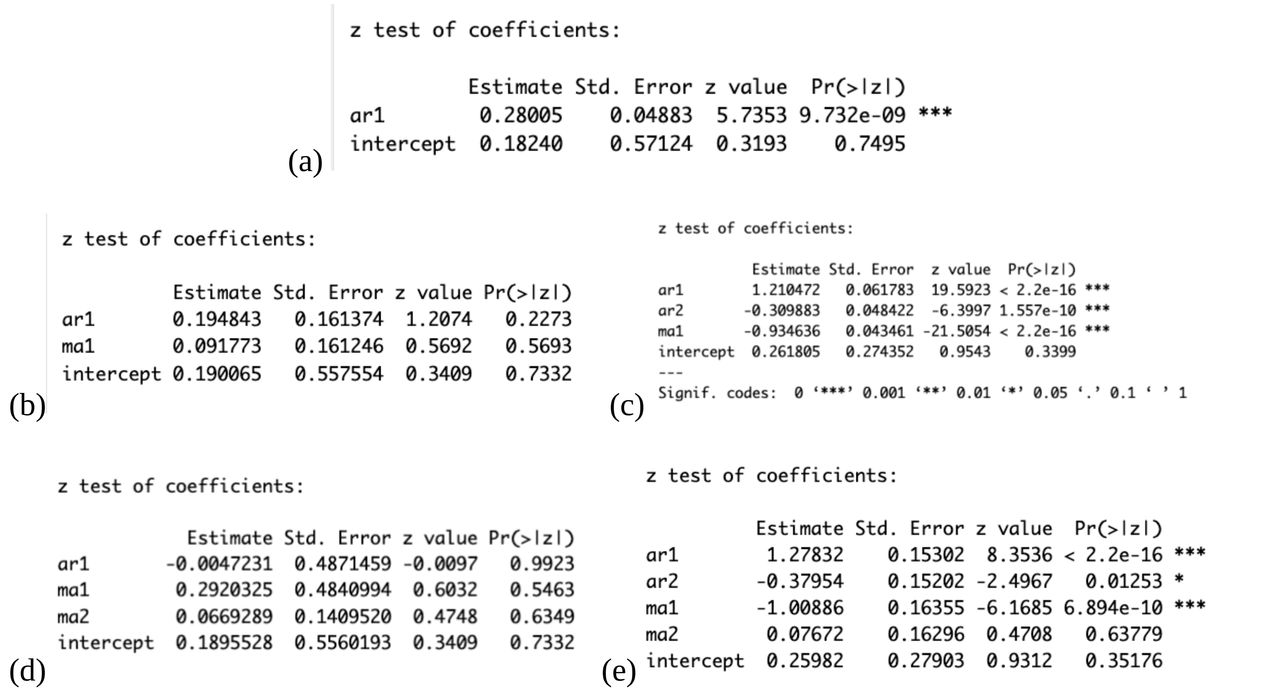
```
################################################
# Augmented Dickey-Fuller Test Unit Root Test #
################################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-21.2344  -1.5999  -0.1146   1.6104  15.3268

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.375250   0.398829   0.941  0.34733
z.lag.1     -0.031854   0.010092  -3.156  0.00172 **
tt           0.005377   0.002496   2.155  0.03179 *
z.diff.lag   0.382248   0.045904   8.327 1.31e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.947 on 403 degrees of freedom
Multiple R-squared:  0.1559,    Adjusted R-squared:  0.1496
F-statistic: 24.81 on 3 and 403 DF,  p-value: 9.489e-15


Value of test-statistic is: -3.1562 3.3923 5.0241

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Figure 1.A: Augmented Dickey-Fuller test for oil prices series



```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-28.305  -5.437   0.439   5.221  35.740

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.656692   0.812169   0.809    0.419
z.lag.1     -0.770707   0.059187 -13.022   <2e-16 ***
tt          -0.001774   0.003444  -0.515    0.607
z.diff.lag   0.031602   0.048471   0.652    0.515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.133 on 402 degrees of freedom
Multiple R-squared:  0.3784,	Adjusted R-squared:  0.3738
F-statistic: 81.59 on 3 and 402 DF,  p-value: < 2.2e-16


Value of test-statistic is: -13.0216 56.5836 84.857

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Figure 2.A: Augmented Dickey-Fuller test for oil price returns series



```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1         0.28005    0.04883  5.7353 9.732e-09 ***
intercept   0.18240    0.57124  0.3193    0.7495
```

(a)

```
z test of coefficients:

          Estimate Std. Error z value Pr(>|z|)
ar1       0.194843   0.161374  1.2074   0.2273
ma1       0.091773   0.161246  0.5692   0.5693
intercept 0.190065   0.557554  0.3409   0.7332
```

(b)

```
z test of coefficients:

           Estimate Std. Error  z value  Pr(>|z|)
ar1        1.210472   0.061783  19.5923 < 2.2e-16 ***
ar2       -0.309883   0.048422  -6.3997 1.557e-10 ***
ma1       -0.934636   0.043461 -21.5054 < 2.2e-16 ***
intercept  0.261805   0.274352   0.9543    0.3399
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(c)

```
z test of coefficients:

            Estimate Std. Error z value Pr(>|z|)
ar1       -0.0047231  0.4871459 -0.0097   0.9923
ma1        0.2920325  0.4840994  0.6032   0.5463
ma2        0.0669289  0.1409520  0.4748   0.6349
intercept  0.1895528  0.5560193  0.3409   0.7332
```

(d)

```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1         1.27832    0.15302  8.3536 < 2.2e-16 ***
ar2        -0.37954    0.15202 -2.4967   0.01253 *
ma1        -1.00886    0.16355 -6.1685 6.894e-10 ***
ma2         0.07672    0.16296  0.4708   0.63779
intercept   0.25982    0.27903  0.9312   0.35176
```

(e)

Figure 3.A: Coefficients’ estimates and significance for (a) AR(1) model, (b) ARMA(1, 0, 1) model, (c) ARMA(2, 0, 1), (d) ARMA(1, 0, 2), (e) ARMA(2, 0, 2)

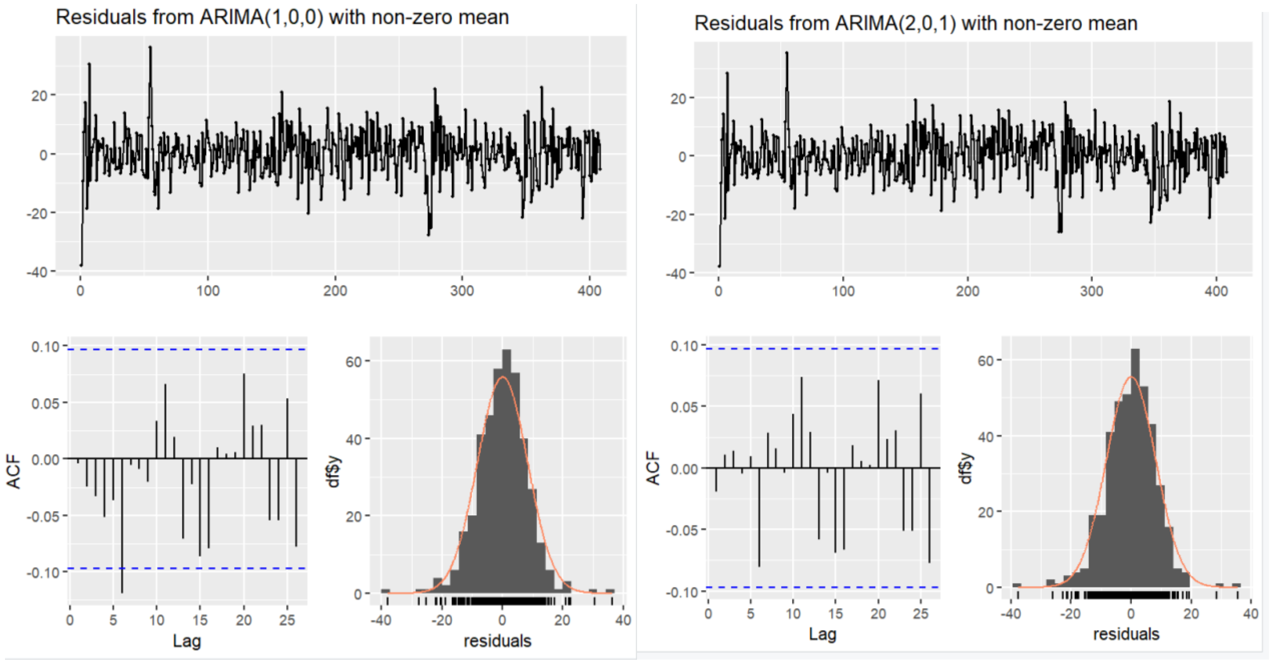
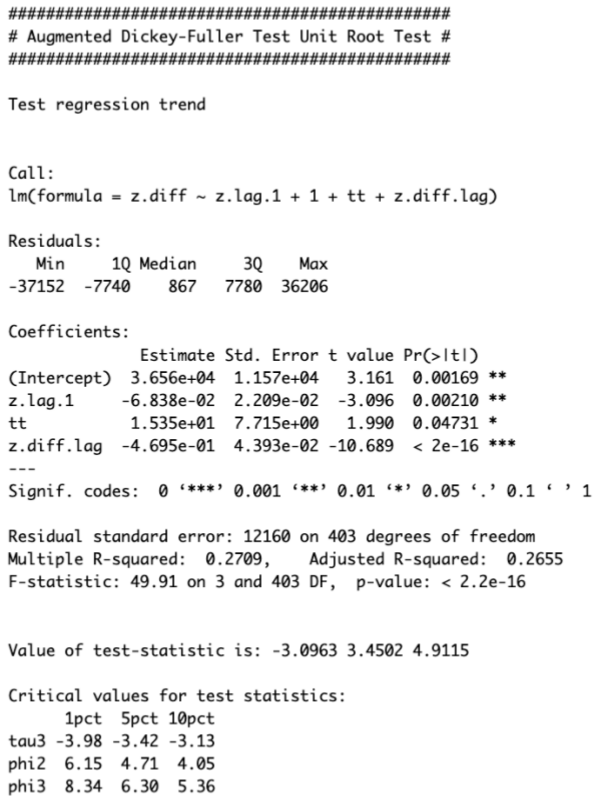
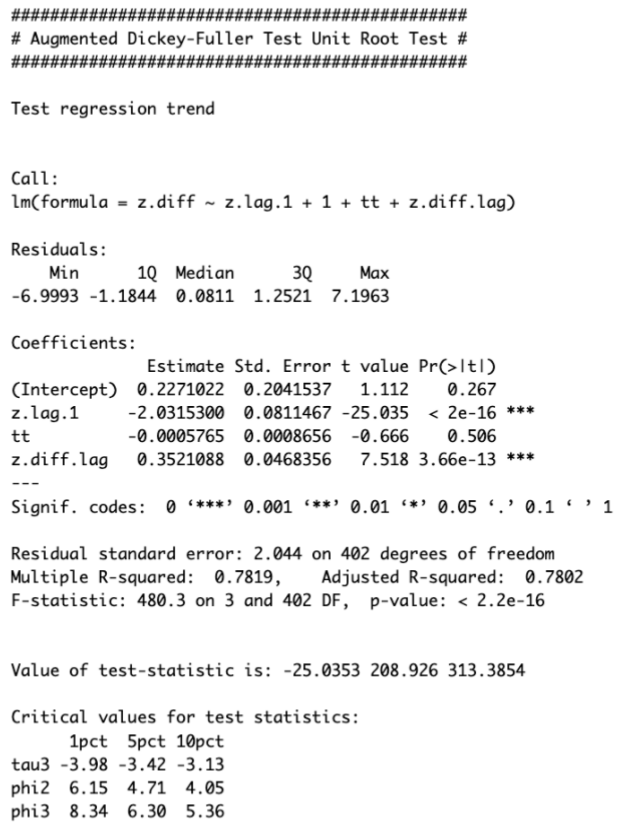


Figure 4.A: Visual outputs from Ljung-Box test for AR(1) (on the left) and for ARMA(2, 0, 1) (on the right)



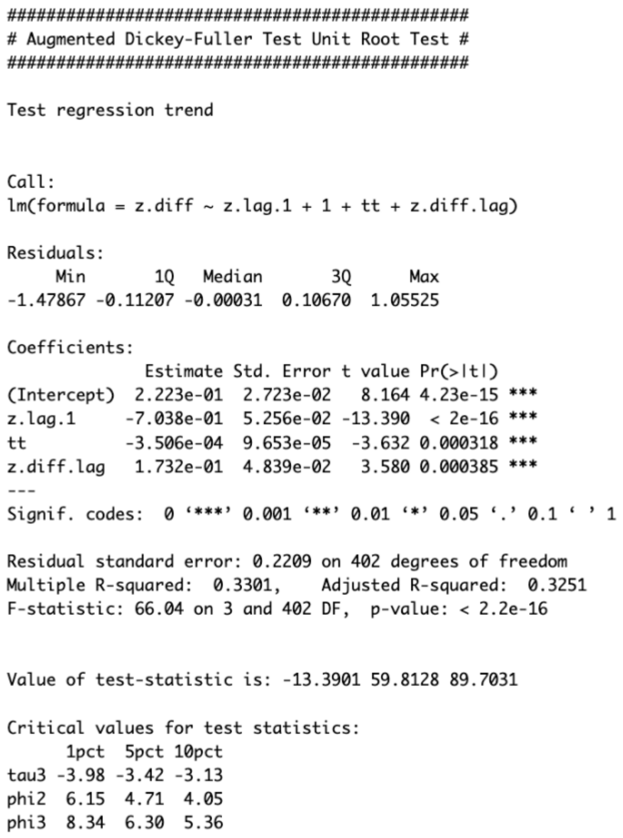
```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
   Min     1Q Median     3Q    Max
-37152  -7740    867   7780  36206

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.656e+04  1.157e+04   3.161  0.00169 **
z.lag.1     -6.838e-02  2.209e-02  -3.096  0.00210 **
tt           1.535e+01  7.715e+00   1.990  0.04731 *
z.diff.lag  -4.695e-01  4.393e-02 -10.689  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12160 on 403 degrees of freedom
Multiple R-squared:  0.2709,    Adjusted R-squared:  0.2655
F-statistic: 49.91 on 3 and 403 DF,  p-value: < 2.2e-16


Value of test-statistic is: -3.0963 3.4502 4.9115

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Figure 5.A: Augmented Dickey-Fuller test for seasonally adjusted supply of crude oil and petroleum products series



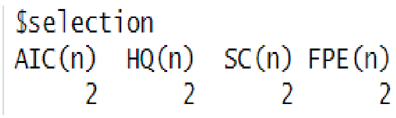
```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9993 -1.1844  0.0811  1.2521  7.1963

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2271022  0.2041537   1.112    0.267
z.lag.1     -2.0315300  0.0811467 -25.035  < 2e-16 ***
tt          -0.0005765  0.0008656  -0.666    0.506
z.diff.lag   0.3521088  0.0468356   7.518 3.66e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.044 on 402 degrees of freedom
Multiple R-squared:  0.7819,    Adjusted R-squared:  0.7802
F-statistic: 480.3 on 3 and 402 DF,  p-value: < 2.2e-16


Value of test-statistic is: -25.0353 208.926 313.3854

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Figure 6.A: Augmented Dickey-Fuller test for change in oil supply series



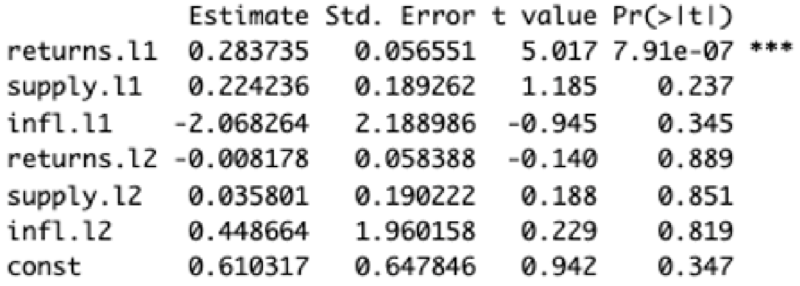
```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47867 -0.11207 -0.00031  0.10670  1.05525

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.223e-01  2.723e-02   8.164 4.23e-15 ***
z.lag.1     -7.038e-01  5.256e-02 -13.390  < 2e-16 ***
tt          -3.506e-04  9.653e-05  -3.632 0.000318 ***
z.diff.lag   1.732e-01  4.839e-02   3.580 0.000385 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2209 on 402 degrees of freedom
Multiple R-squared:  0.3301,    Adjusted R-squared:  0.3251
F-statistic: 66.04 on 3 and 402 DF,  p-value: < 2.2e-16


Value of test-statistic is: -13.3901 59.8128 89.7031

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.98 -3.42 -3.13
phi2  6.15  4.71  4.05
phi3  8.34  6.30  5.36
```

Figure 7.A: Augmented Dickey-Fuller test for inflation series



```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     2      2      2      2
```

Figure 8.A: Output of the Var.Select function showing lags suggested from the following information criteria: Akaike information criterion (AIC), Hannan-Quinn information criterion (HQ), Schwarz criterion (SC), Akaike's Final Prediction Error criterion (FPE).



```
             Estimate Std. Error t value Pr(>|t|)    
returns.l1  0.283735   0.056551   5.017 7.91e-07 ***
supply.l1   0.224236   0.189262   1.185    0.237    
infl.l1    -2.068264   2.188986  -0.945    0.345    
returns.l2 -0.008178   0.058388  -0.140    0.889    
supply.l2   0.035801   0.190222   0.188    0.851    
infl.l2     0.448664   1.960158   0.229    0.819    
const       0.610317   0.647846   0.942    0.347    
```

Figure 9.A: Test results of the VAR(2) model with returns, supply, and inflation as explanatory variables.

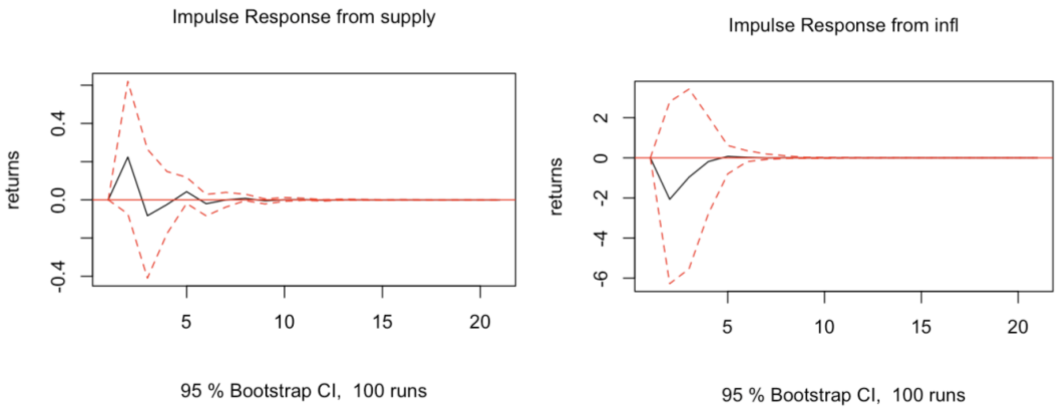


Figure 10.A: Impulse response functions for supply (left) and inflation (right) as impulses.

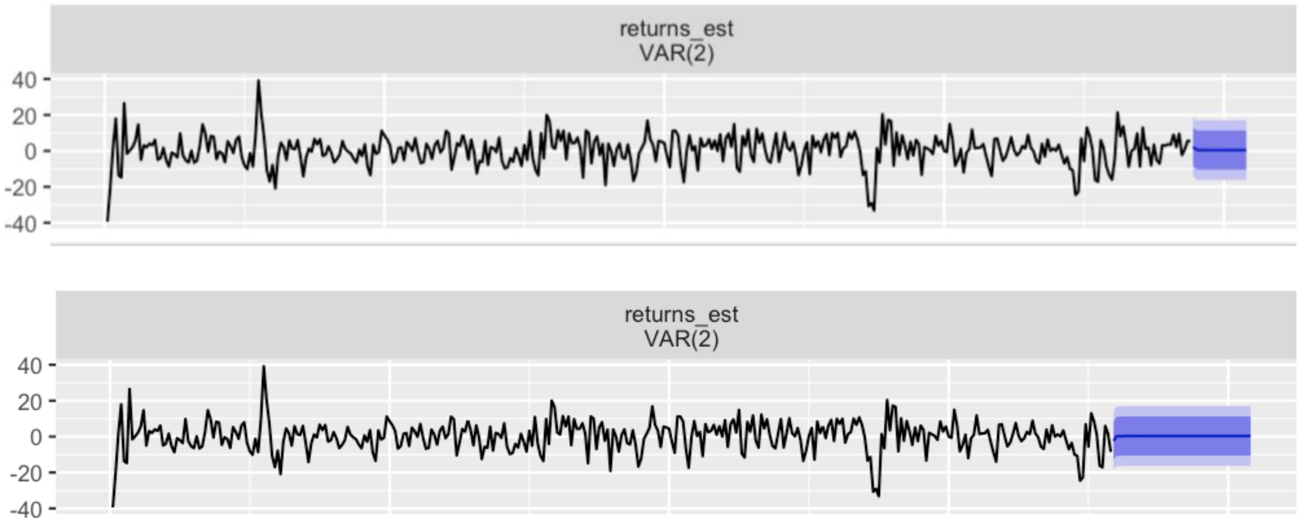


Figure 11.A: VAR(2) forecasts for a 20-period forecasting horizon (top) and a 50-period forecasting horizon (bottom).

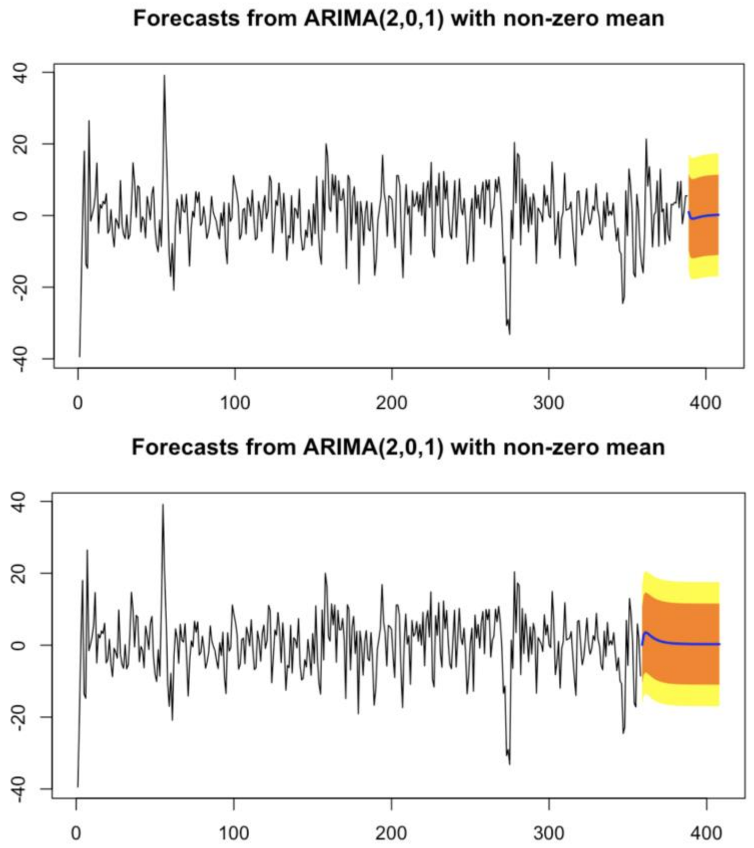


Figure 12.A: ARMA(2, 0, 1) forecasts for a 20-period forecasting horizon (top) and a 50-period forecasting horizon (bottom).

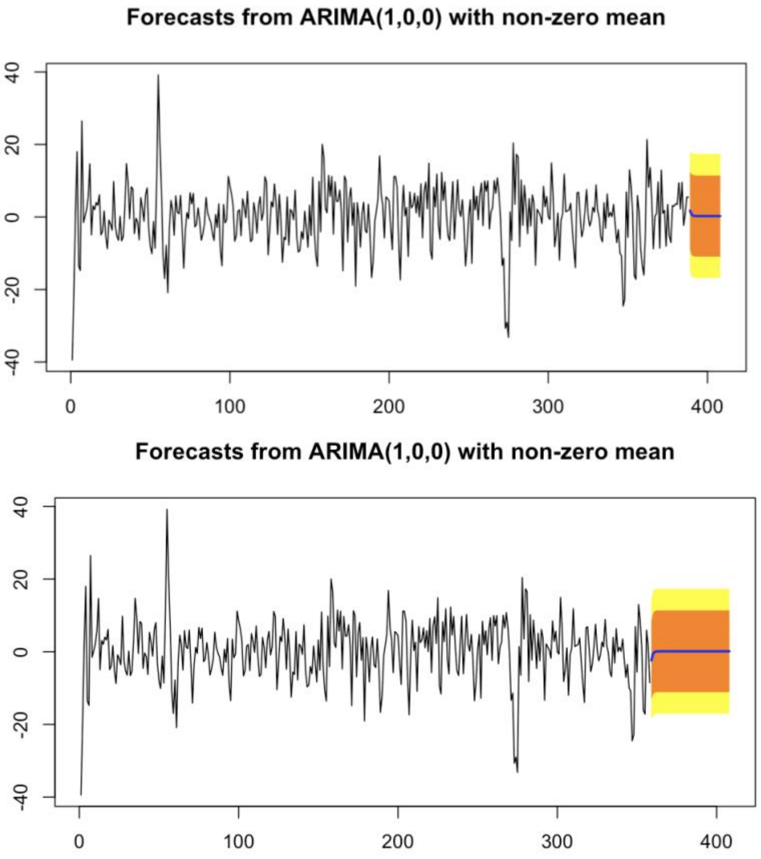


Figure 13.A: AR(1) forecasts for a 20-period forecasting horizon (top) and a 50-period forecasting horizon (bottom).