EC4304 Economic and Financial Forecasting

Group Project Report

AY20/21 Semester 1

Group Members

Chan You Zhong (A0182994L)

Lam Si Jie Joey (A0189610B)

Lim Yi Xuan (A0187612B)

Tay Jia Yi (A0187638L)

Wong Chun Kitt (A0154205R)

1. Introduction

The aim of this paper is to investigate whether adding oil price to a model containing more "conventional" leading indicators like term spread, high-yield bond spread, housing starts, inventories to sales ratio and total retail trade can improve forecasts of real Gross Domestic Product (GDP) growth in the United States (US). Oil price has been shown to be a leading indicator of the business cycle in the US but we are interested in whether this remains true when oil price is added into a model containing other leading indicators. Another aim is to construct forecasting models with oil price to forecast US GDP growth, selecting both on Akaike information criterion (AIC) and Bayesian information criterion (BIC). Since BIC penalises additional parameters more heavily than AIC does, we are interested in whether selection by BIC results in a better model when it comes to forecasting data with a smaller sample size, although AIC is more widely used as the selection criteria for forecasting models.

We first chose a benchmark Auto-Regressive (AR) model followed by creating combined models containing "conventional" leading indicators. We chose a few best combined models by AIC and a few best combined models by BIC. Subsequently, we added lags of oil price to each of the models. We then chose the best AIC and BIC model with oil price and compared it to the best AIC and BIC model without oil price respectively.

The Diebold-Mariano test indicated that the best model with oil price did not have better forecast accuracy as compared to the best model without oil price for both AIC and BIC chosen models. However, adding lags of oil price to models containing "conventional" indicators can result in lower AIC and BIC for the combined model.

Since combined forecasts tend to perform better than individual forecasts, we combined forecasts from the 6 best AIC models using Weighted AIC (WAIC) and the 6 best BIC models using Bayesian Model Averaging (BMA) to implement true out-of-sample forecasting of US GDP growth for the next 4 quarters from 2020 Quarter 3 to 2021 Quarter 2.

The remainder of the paper is structured as follows. We first provide a brief literature review on the usage of oil price in forecasting GDP growth in section 2. In section 3, we discuss the variables used for this forecasting exercise and why they are leading variables. This is followed by section 4 where we discuss methodology and model selection. In section 5, we evaluate the models and this is followed by a discussion of the results in section 6.

2. Literature Review

Many papers have used oil price as a leading indicator for forecasting GDP growth (Kilian & Vigfusson, 2013, Hamilton, 2003 and Nonejad 2020). Hamilton (2003) cited papers showing a positive correlation between the percentage change in oil price and GDP growth rate. Theoretically, he argued that a sharp increase in oil price would contribute to a reduction or postponement in the purchase of consumer and capital goods which require the use of crude oil and its refined products, such as cars and machineries. The resultant fall in consumption and investment would negatively affect GDP growth.

However, Hamilton casts doubt on whether the relationship between oil price and GDP growth is symmetrical as some papers suggest, as he believes that a fall in oil price would not have the same positive effect on the economy as the negative effect on the economy with an increase in oil price. He thus adopts a non-linear, non-symmetrical approach in forecasting GDP growth with oil price. However, Kilian and Vigfusson (2013) found that the non-linearity of the Hamilton model was responsible for the success of the model, and not the asymmetrical approach that he took.

Looking at the plot of crude oil prices (West Texas Intermediate) coupled with recession periods in the US, we are even more convinced that oil price is a leading indicator of recessions in the US. (Figure 1).

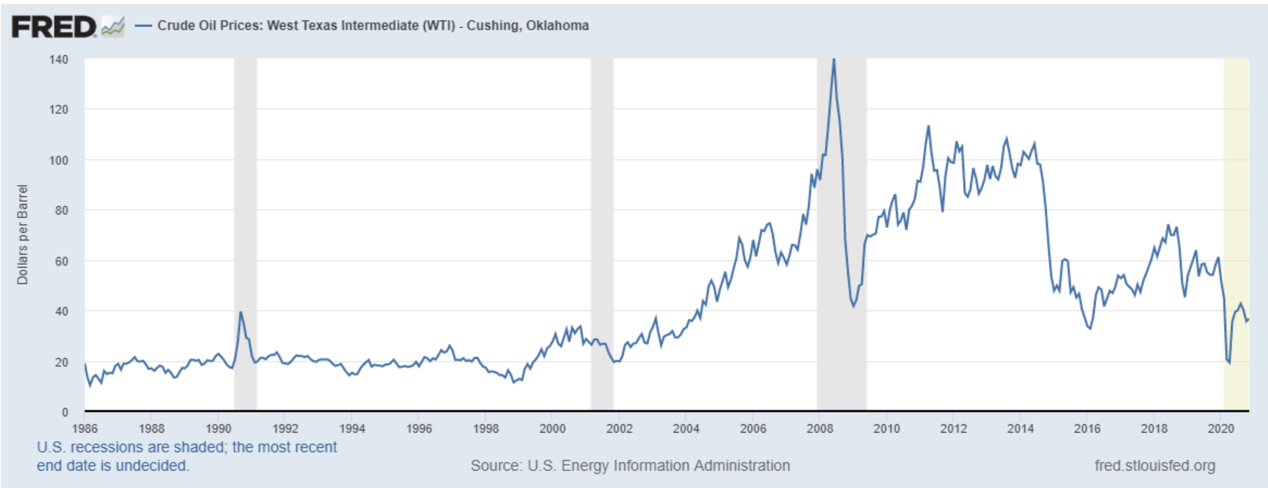

Figure 1 Crude Oil Prices: West Texas Intermediate (WTI) - Cushing Oklahoma from 1986 to 2020

Other variables like the 10-year term spread (10 year-3 month), high-yield bond spread (Baa-Aaa), housing starts and total retail trade are more “conventional” leading indicators of GDP growth (Banerjee & Marcellino, 2006 and Kuosmanen & Vataja, 2014). The papers on oil price used oil price as the sole variable in forecasting GDP growth. We are therefore interested in investigating whether oil price can still result in better forecast when added as a leading indicator to a model containing “conventional” leading indicators as compared to the same model without oil price.

In addition, both Kilian & Vigfusson (2013) and Nonejad (2020) used percentage change in crude oil prices to forecast GDP growth. We are therefore interested in whether crude oil remains a valid leading indicator when we use the absolute price of crude oil as a predictor instead of percentage change.

3. Data

We chose to forecast real GDP growth in the US using the following variables: Housing Starts, Inventory-to-Sales Ratio, 10-Year Term Spread, High-Yield Spread, Total Retail Trade and Oil Prices. The selection of the US as our country of study is driven by both the availability of a long quarterly GDP growth rate series and the fact that the US economy is the world's largest economy.

All the data are quarterly data and have been seasonally adjusted with the exception of oil prices. We did not adjust oil prices for seasonality because there appears to be no significant seasonal pattern in oil prices after we adjusted oil prices to reflect quarterly data by taking the average of oil prices in a quarter of a year. As seen from Figure 3.1, there is no recurring seasonal pattern in each year when we plot oil prices against time.

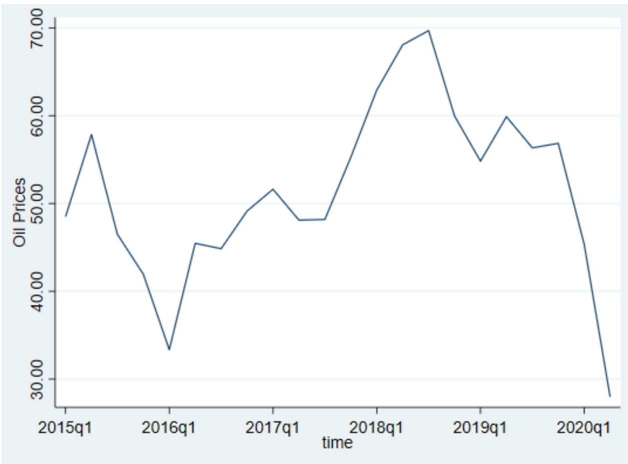

Figure 3.1 Crude oil (WTI) prices from 2015 to 2020

The data used in this project is sourced from the Federal Reserve Bank of St. Louis, which organizes data from multiple sources to build the Federal Reserve Economic Data (FRED).

The GDP growth rates from 1947 to 2020 were taken from the U.S. Bureau of Economic Analysis, while the database for Housing Starts from 1959 to 2020 is provided by the U.S. Census Bureau, where one housing start is indicated by the start of the excavation for the foundation of a building. The database for the Inventory-to-Sales Ratio from 1947 to 2020 is provided by the Bureau of Economic Analysis and measures the ratio of private inventories to final sales of domestic business.

The variable 10-Year Term Spread is created by finding the difference between 10-Year Treasury Constant Maturity Rate and 10-Year 3-Month Treasury Bill: Secondary Market Rate from 1953 to 2020, where both datasets are obtained from the Board of Governors of the Federal Reserve System.

Similarly, the variable High-Yield Spread is created by finding the difference between Moody's Seasoned Aaa Corporate Bond Yield and Moody's Seasoned Baa Corporate Bond Yield from 1919 to 2020, where both datasets are obtained from Moody's.

The database for Total Retail Trade in the US, from 1960 to 2020, is provided by the Organization for Economic Co-operation and Development (OECD) while the database for Oil Prices from 1986 to 2020 is taken from the U.S. Energy Information Administration and measures the price of the crude oil, West Texas Intermediate, in Cushing, Oklahoma. The crude oil, West Texas Intermediate, serves as one of the main global oil benchmarks for oil prices.

When we plot each variable against time, we can see that they are closely linked to GDP growth rate. For the case of Housing Starts, the number of constructions started reached its lowest in the years 1981, 1990 and 2008 (Figure 3.2). This corresponds to the 1979 Energy Crisis which led to a recession in the early 1980s; the recession in the early 1990s caused by a combination of consumer pessimism, high interest rates and an increase in oil prices; and the 2008 Global Financial Crisis. While it is less obvious in the case for Inventory-to-Sales ratio, we can see that there is a peak in Inventory-to-Sales Ratio in 1980 and 2008, indicating an increase in inventories level in firms because of lower levels of consumption of goods and services (Figure 3.3).

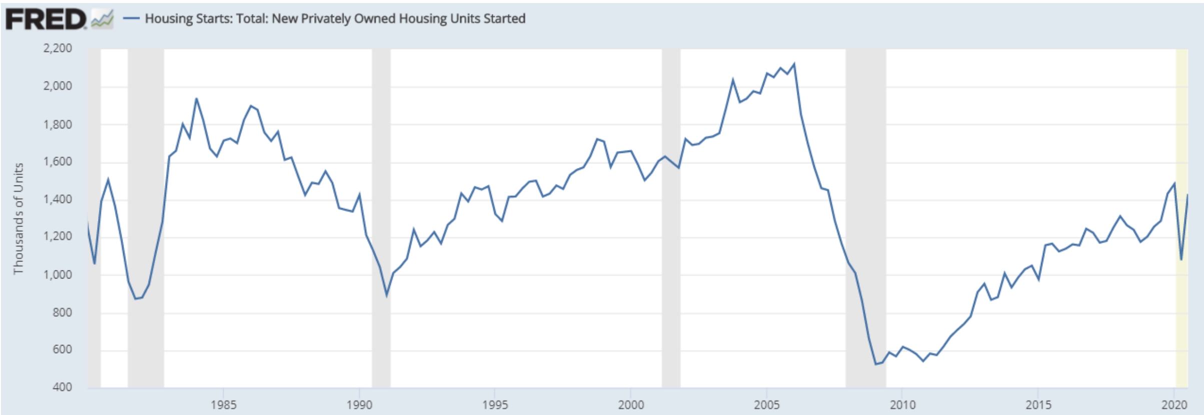

Figure 3.2 Housing Starts in the US from 1980 to 2020

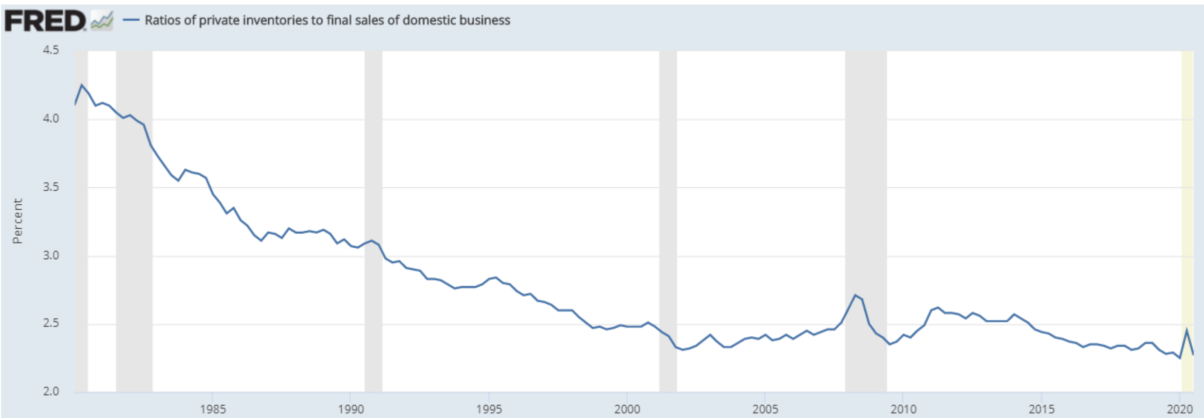

Figure 3.3 Ratio of private Inventories to final Sales of domestic business in the US from 1980 to 2020

Similarly, for interest rate spreads, there are obvious peaks in both the 10-Year Term Spread and High-Yield Spread during the early 1980s recession and the 2008 Global Financial Crisis, indicating higher market risk in the financial market. However, the peaks are less obvious for the other economic recessions during the early 1990s and 2000s respectively (Figure 3.4).

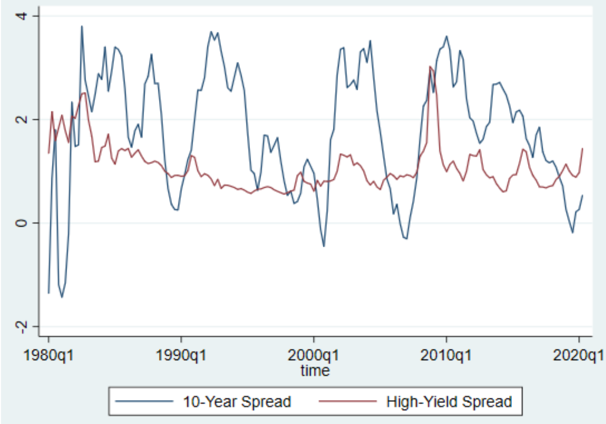

Figure 3.4 10-Year Term Spread and High-Yield Spread in the US from 1980 to 2020

In the case of Total Retail Trade, we can see that there is a dip in retail trade around 1982, 1992 and 2008 which corresponds to periods of financial recessions (Figure 3.5) Similarly, for the case of oil prices, there are small spikes in oil price in the early 1990s and early 2000s and a large spike in oil prices in 2008. This suggests that oil prices is possibly a good indicator of GDP growth, since the magnitude of spikes correspond to the severity of the recessions occurring during that particular period: the Global Recession in 2008 had the largest impact - up to 8.4% of negative GDP growth while the other two recessions were of a smaller scale, leading to 3.6% and 1.7% decline in GDP growth respectively (Figure 1).

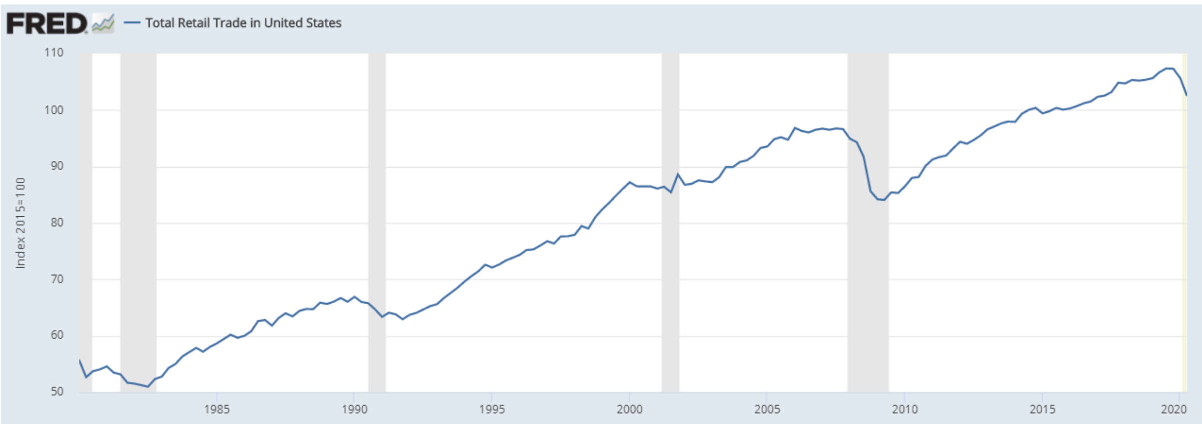

Figure 3.5 Total Retail Trade in the US from 1980 to 2020

4. Methodology

This section describes how we select the model used for forecasting real GDP growth. Throughout our analysis, we chose to use both AIC and BIC when selecting the optimal model. Both information criteria evaluate how well a model fits the data it is generated and penalises complexity (number of parameters used) to prevent overfitting. The model with the smallest AIC or BIC value is considered superior. AIC is widely preferred as the method of selecting forecasting models, but it works better in larger samples. On the other hand, BIC is generally used when the aim is to select the true model, but may work better than AIC in selecting forecast models in smaller samples as it penalises additional parameters more heavily than AIC does. As our aim is to find the best out-of-sample forecasts and our sample size is relatively small (137 observations if we include all variables), we chose to use both AIC and BIC to select the optimal model. We then compared the forecasting model selected by each information criterion with the benchmark model.

First, we estimated the AR model for GDP growth by choosing the optimal number of lags for GDP growth using both AIC and BIC, considering up to 6 lags of GDP growth.

After we chose the optimal number of lags *p* for GDP growth, this Auto-Regressive model, AR(*p*) was set as the benchmark model which we used to compare with our final forecast models. AIC selected AR(3) while BIC selected AR(1).

Table 4.1 AIC and BIC of AR(1) to AR(6)

Model (247 observations)	**AIC**	**BIC**
AR(1)	1362.444	**<u>1369.463</u>**
AR(2)	1360.934	1371.462
AR(3)	**<u>1359.304</u>**	1373.342
AR(4)	1359.964	1377.511
AR(5)	1360.301	1381.357
AR(6)	1361.52	1387.086

Next, we selected the optimal number of lags for each leading indicator by running an Auto-Regressive Distributed Lag Model (ADL) with the optimal lags of GDP growth selected above, and up to 6 lags of a single leading indicator. We repeated this process for all five of the common leading indicators used for forecasting GDP: Housing Starts, Inventory-to-Sales Ratio, 10-Year Term Spread, High-Yield Spread and Total Retail Trade.

The optimal lags selected by AIC for each leading indicator are as follows: 3 lags of Housing Starts, 0 lags of Inventory-to-Sales Ratio, 2 lags of 10-Year Term Spread, 5 lags of High-Yield Spread and 3 lags of Total Retail Trade. We then compared the AIC of the ADL model with 3 lags of GDP growth and different combinations of the optimal lags of these variables:

Table 4.2 AIC of models with different combinations of the optimal lags of leading indicators selected by AIC.
Note: *houst*: Housing Starts, *spr10*: 10-Year Term Spread, *hyspr*: High-Yield Spread, *rtrade*: Total Retail Trade; the numbers before each variable represents the number of lags of that variable

No.	**Model (196 observations)**	**AIC**
1	**AR(3)**	1024.992
2	**AR(3) + 3houst**	991.3635
3	**AR(3) + 2spr10**	1019.78
4	**AR(3) + 5hyspr**	1012.904
5	**AR(3) + 3rtrade**	1001.284
6	**AR(3) + 3houst + 2spr10**	994.5581
7	**AR(3) + 3houst + 5hyspr**	**982.7434**
8	**AR(3) + 3houst + 3rtrade**	**982.4078**
9	**AR(3) + 2spr10 + 5hyspr**	1009.692
10	**AR(3) + 2spr10 + 3rtrade**	995.7321
11	**AR(3) + 5hyspr + 3rtrade**	998.6357
12	**AR(3) + 3houst + 2spr10 + 5hyspr**	986.1672
13	**AR(3) + 3houst + 2spr10 + 3rtrade**	984.191
14	**AR(3) + 2spr10 + 5hyspr + 3rtrade**	993.375
15	**AR(3) + 3houst + 5hyspr + 3rtrade**	**979.4676**
16	**AR(3) + 3houst + 2spr10 + 5hyspr + 3rtrade**	**981.0973**

We selected the models with AIC that are the lowest and closest to each other, so that we can use them later to construct our combined forecast model. The selected models based on AIC were, in order, Model 15, Model 16, Model 8 and Model 7.

We then repeated the same steps to find the optimal lags selected for each variable by BIC which are: 3 lags of Housing Starts, 0 lags of Inventory-To-Sales Ratio, 1 lag of 10-Year Term Spread, 2 lags of High-Yield Spread and 3 lags of Total Retail Trade. Again, we compared the BIC of the ADL model with 1 lag of GDP growth and different combinations of the optimal lags of these variables:

Table 4.3 BIC of models with different combinations of the optimal lags of leading indicators selected by BIC.
Note: *houst*: Housing Starts, *spr10*: 10-Year Term Spread, *hyspr*: High-Yield Spread, *rtrade*: Total Retail Trade; the numbers before each variable represents the number of lags of that variable

No.	Model (196 observations)	BIC
17	**AR(1)**	1035.963
18	**AR(1) + 3houst**	**1011.806**
19	**AR(1) + 1spr10**	1035.383
20	**AR(1) + 2hyspr**	1029.73
21	**AR(1) + 3rtrade**	1016.06
22	**AR(1) + 3houst + 1spr10**	1016.755
23	**AR(1) + 3houst + 2hyspr**	**1007.179**
24	**AR(1) + 3houst + 3rtrade**	**1007.999**
25	**AR(1) + 1spr10 + 2hyspr**	1027.885
26	**AR(1) + 1spr10 + 3rtrade**	1014.317
27	**AR(1) + 2hyspr + 3rtrade**	1020.343
28	**AR(1) + 3houst + 1spr10 + 2hyspr**	**1011.386**
29	**AR(1) + 3houst + 1spr10 + 3rtrade**	**1011.777**
30	**AR(1) + 1spr10 + 2hyspr + 3rtrade**	1017.861
31	**AR(1) + 3houst + 2hyspr + 3rtrade**	**1010.704**
32	**AR(1) + 3houst + 1spr10 + 2hyspr + 3rtrade**	1013.172

We then selected the best models based on the lowest BIC. The selected models based on BIC were, in order, Model 23, Model 24, Model 31, Model 28, Model 29 and Model 18.

As the aim of this project is to find out whether oil prices will improve the forecast model that includes only "conventional" leading indicators for GDP growth, we subsequently introduced up to 6 lags of oil prices and compared the AIC and BIC of these models with the models that we found in the previous steps. We found that including up to 1 or 2 lags of oil prices reduces both the AIC and BIC of the selected models. The final models selected to construct the combined forecast models were as follows:

Table 4.4 Final models selected to construct the combined forecast models.
Note: *houst*: Housing Starts, *spr10*: 10-Year Term Spread, *hyspr*: High-Yield Spread, *rtrade*: Total Retail Trade; the numbers before each variable represents the number of lags of that variable

Model (selected by AIC)	**Model (selected by BIC)**
AR(3) + 3houst + 3rtrade + 1oilp	AR(1) + 3houst + 2hyspr + 1oilp
AR(3) + 3houst + 5hyspr + 1oilp	AR(1) + 3houst + 2hyspr + 2oilp
AR(3) + 3houst + 5hyspr + 2oilp	AR(1) + 3houst + 1oilp
AR(3) + 3houst + 5hyspr + 5oilp	AR(1) + 3houst + 3rtrade + 1oilp
AR(3) + 3houst + 3rtrade + 2oilp	AR(1) + 3houst + 3rtrade
AR(3) + 3houst + 5hyspr + 3rtrade + 1oilp	AR(1) + 3houst + 1spr10 + 2hyspr + 1oilp

We proceeded to combine the models selected by AIC using the weighted AIC (WAIC) and combine the models selected by BIC using the Bayesian Model Averaging (BMA). To combine the models using weighted AIC, we accorded the models with weights proportional to their AIC value. The models with the lowest AIC will be accorded the most weight while the models with the highest AIC will be accorded the least weight when constructing the combined forecast model. The same logic applies to the Bayesian Model Averaging, but we accorded weights according to the BIC values instead.

Therefore, we ended up with 2 optimal combined forecast models, one selected based on AIC and constructed using WAIC; the other selected based on BIC and constructed using BMA.

We then withheld 40 observations from our sample size so that we could compare the pseudo-out-of-sample forecasts of these two combined forecast models and the two benchmark models – AR(3) and AR(1), with the actual GDP growth rates using a rolling window. For our project, we chose to use a rolling window because rolling window has traditionally been used to forecast output growth. Stock and Watson (2003) found that when the same models are used to forecast GDP growth using inflation with different windows, using a rolling window tends to yield lower mean square forecast error than using a recursive window.

In addition, the choice to use 40 observations for our pseudo out-of-sample forecasts stems from the fact that the research division of the Federal Reserve Bank of St. Louis, where our datasets are obtained from, also choose to use 40 observations to conduct pseudo out-of-sample forecasts when forecasting GDP growth using data from 1953 Quarter 4 (Clark and McCraken, 2008).

5. Results

5. 1 Pseudo-Out-Of-Sample Forecasts

Our first findings come from the pseudo-out-of-sample (POOS) forecasts using the two combined forecast models - WAIC and BMA, along with the two benchmark models - AR(3) and AR(1). The results are shown in Fig 5.1.1.

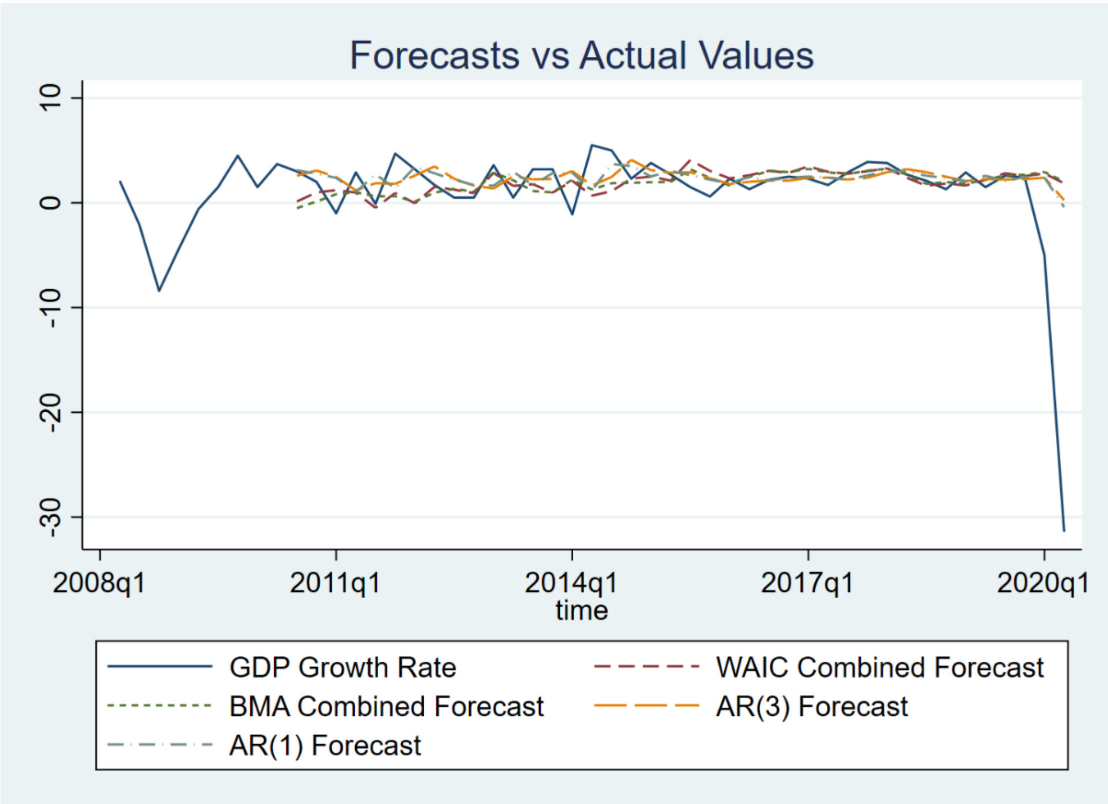

Figure 5.1.1

From Fig 5.1.1, it can be seen that the two benchmark models performed relatively better as it generated forecasts that were closer to the actual real GDP growth as compared to the two combined models. It is worth noting that towards the end of the forecast horizon, the benchmark models reflected the sharp dip in growth rate more accurately than the combined models. This could suggest that the benchmark models are more suited at forecasting sharp changes in growth rate while the combined models are more apt for minor fluctuations.

5.2 True Out-Of-Sample Forecasts

We then computed the true out-of-sample forecasts for each model. The figures below show the results for each model.

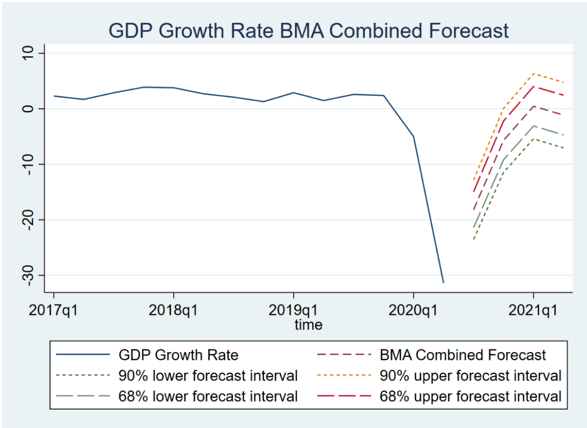

Figure 5.2.1 BMA Combined Forecast

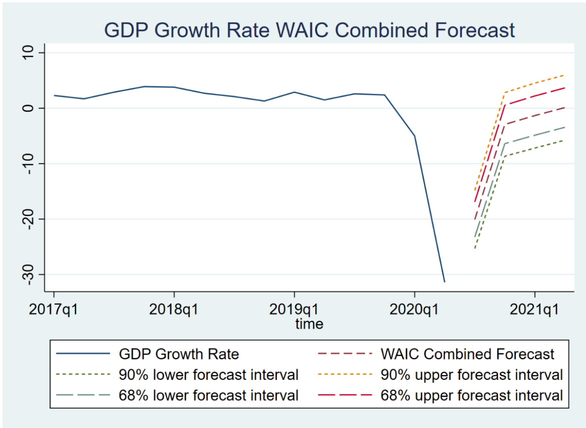

Figure 5.2.2 WAIC Combined Forecast

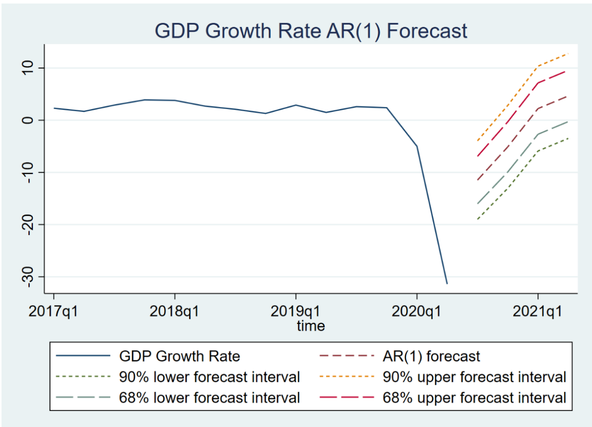

Figure 5.2.3 AR(1) Forecast

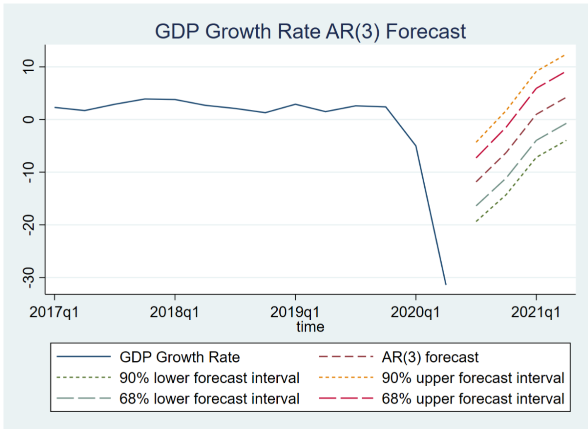

Figure 5.2.4 AR(3) Forecast

5.3 Forecast Evaluation

5.3.1 Test 1-Step Ahead for White Noise

To ensure that the forecasts are optimal, 1-step forecasts were constructed for each model and tested for serial correlation in the forecast errors using the Ljung-Box statistical test and the ACF (Auto-Correlation Function) plots. The figures below show the results for each model.

. corrgram er_combined_bma

LAG	AC	PAC	Q	Prob>Q
1	0.2149	1.5354	1.9903	0.1583
2	0.0115	0.0965	1.9962	0.3686
3	0.0337	0.4428	2.0477	0.5626
4	0.0066	-0.3695	2.0497	0.7266
5	-0.0182	0.2196	2.0655	0.8400
6	0.0338	0.2877	2.1221	0.9081
7	0.0003	0.0271	2.1221	0.9528
8	-0.0298	-0.5738	2.1686	0.9754
9	-0.0198	0.6266	2.1899	0.9881
10	-0.0036	0.4277	2.1907	0.9947
11	-0.0085	-0.2018	2.1948	0.9977
12	0.0628	0.2822	2.4316	0.9984
13	0.0349	-0.9361	2.5074	0.9992
14	0.0021	1.1793	2.5077	0.9997
15	0.0363	0.7677	2.5964	0.9998
16	0.0301	0.0483	2.6597	0.9999
17	-0.0103	0.1641	2.6675	1.0000
18	0.0500	0.2400	2.8588	1.0000

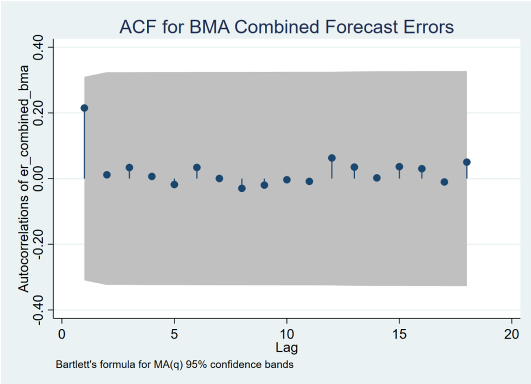

Figure 5.3.1 Ljung-Box Statistics and ACF plot for BMA combined model

Note for Ljung-Box statistics: number of observations, $t = 40$, lag $=\sqrt{t} \approx 6$. At a p-value of 0.9081, there is insufficient evidence to reject the null hypothesis that jointly, the autocorrelations are zero up to lag 6. Likewise, the ACF plot suggests that there is no serial correlation between the forecast errors for the BMA combined model as all the points fall within Bartlett's band.

. corrgram er_combined_waic

LAG	AC	PAC	Q	Prob>Q
1	0.2184	1.5154	2.0552	0.1517
2	0.0086	-0.0803	2.0585	0.3573
3	0.0410	0.4676	2.1348	0.5449
4	0.0027	-0.5068	2.1351	0.7109
5	-0.0340	0.2157	2.1906	0.8222
6	0.0247	0.2912	2.2206	0.8983
7	-0.0129	-0.0395	2.2291	0.9461
8	-0.0406	-0.4534	2.3158	0.9698
9	-0.0187	0.5657	2.3348	0.9850
10	-0.0085	0.1482	2.3388	0.9930
11	-0.0092	-0.1160	2.3437	0.9969
12	0.0603	0.4450	2.5615	0.9979
13	0.0380	-0.5921	2.6514	0.9989
14	-0.0055	0.8418	2.6533	0.9995
15	0.0319	0.5453	2.7217	0.9998
16	0.0332	0.2792	2.7989	0.9999
17	-0.0005	0.2598	2.799	1.0000
18	0.0705	0.4329	3.1787	1.0000

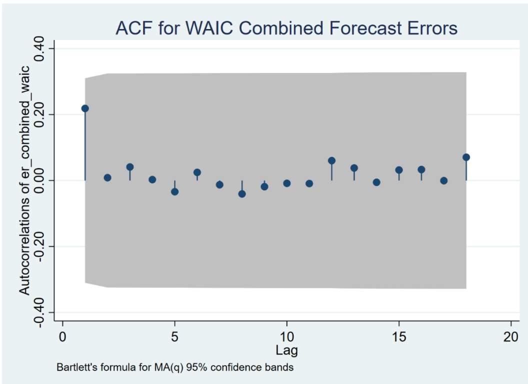

Figure 5.3.2 Ljung-Box Statistics and ACF plot for WAIC combined model

Note for Ljung-Box statistics: number of observations, $t = 40$, lag $=\sqrt{t} \approx 6$. At a p-value of 0.8983, there is insufficient evidence to reject the null hypothesis that jointly, the autocorrelations are zero up to lag 6. Likewise, the ACF plot suggests that there is no serial correlation between the forecast errors for the WAIC combined model as all the points fall within Bartlett's band.

. corrgram er_ar1

LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	0.1610	1.1229	1.1161	0.2908		
2	-0.0073	0.4584	1.1185	0.5716		
3	-0.0007	0.2643	1.1185	0.7726		
4	-0.0079	-0.2891	1.1214	0.8909		
5	-0.0161	0.1798	1.1338	0.9511		
6	0.0209	0.3076	1.1555	0.9790		
7	0.0188	0.6078	1.1734	0.9915		
8	-0.0315	-0.7416	1.2254	0.9964		
9	-0.0465	0.0413	1.3425	0.9982		
10	-0.0296	0.2467	1.3915	0.9992		
11	-0.0542	-0.3462	1.5615	0.9995		
12	0.0288	0.0365	1.6113	0.9998		
13	-0.0074	-0.6715	1.6147	0.9999		
14	-0.0287	0.0945	1.6678	1.0000		
15	-0.0087	-0.8559	1.673	1.0000		
16	0.0049	0.3515	1.6747	1.0000		
17	-0.0250	0.5844	1.7203	1.0000		
18	0.0283	0.3747	1.7816	1.0000		

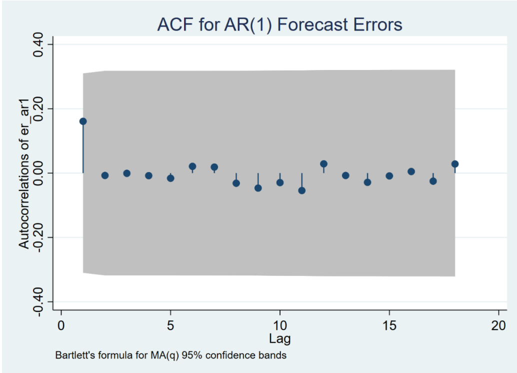

Figure 5.3.3 Ljung-Box Statistics and ACF plot for AR(1) model

Note for Ljung-Box statistics: number of observations, t = 40, lag $=\sqrt{t} \approx 6$. At a p-value of 0.9790, there is insufficient evidence to reject the null hypothesis that jointly, the autocorrelations are zero up to lag 6. Likewise, the ACF plot suggests that there is no serial correlation between the forecast errors for the AR(1) model as all the points fall within Bartlett's band.

. corrgram er_ar3

LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	0.1713	1.2420	1.2642	0.2609		
2	-0.0276	0.0892	1.2977	0.5226		
3	0.0038	0.3542	1.2984	0.7295		
4	-0.0161	-0.3329	1.3105	0.8596		
5	-0.0184	0.3272	1.3267	0.9322		
6	0.0382	0.4320	1.3988	0.9659		
7	0.0279	0.6573	1.4384	0.9844		
8	-0.0286	-0.7311	1.4814	0.9930		
9	-0.0490	0.4409	1.6118	0.9962		
10	-0.0330	0.0530	1.6727	0.9983		
11	-0.0480	-0.1358	1.8061	0.9991		
12	0.0246	0.1074	1.8425	0.9996		
13	-0.0047	-0.7584	1.8438	0.9999		
14	-0.0401	0.1562	1.9478	0.9999		
15	-0.0121	-1.1855	1.9578	1.0000		
16	-0.0059	0.4245	1.9602	1.0000		
17	-0.0312	0.2306	2.0314	1.0000		
18	0.0375	0.4962	2.1387	1.0000		

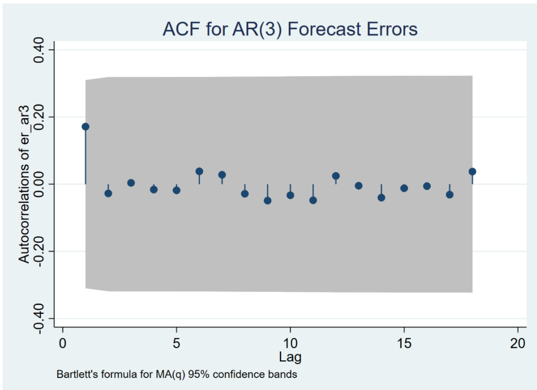

Figure 5.3.4 Ljung-Box Statistics and ACF plot for AR(3) model

Note for Ljung-Box statistics: number of observations, t = 40, lag $=\sqrt{t} \approx 6$. At a p-value of 0.9659, there is insufficient evidence to reject the null hypothesis that jointly, the autocorrelations are zero up to lag 6. Likewise, the ACF plot suggests that there is no serial correlation between the forecast errors for the AR(3) model as all the points fall within Bartlett's band.

Both tests provided strong evidence that the forecast errors from all the models are white noise and thus, unforecastable, which should be what we expect for optimal forecast errors.

5.3.2 Comparison of Forecast Risk - Diebold-Mariano Test

A simple comparison of the bias between models using estimates such as the RMSE (Root Mean Squared Errors) might not be sufficiently rigorous. To account for forecast risk, the Diebold-Mariano (DM) test is used to test for equal predictive ability between two forecasts. Since our holdout sample is small, we did a finite sample correction to the t-statistics as proposed by Harvey, Leybourne and Newbold (1998).

First, we conduct a DM test between the AR(3) forecast and the WAIC forecast, and between the AR(1) forecast and BMA forecast. Since the p-values are high for both, we cannot reject the null hypothesis of equal predictive ability at the 10% significance level. Hence, we are unable to conclude that the AR(3) forecast is better than the WAIC combined forecast, or that the AR(1) forecast is better than the BMA combined forecast in the pseudo out-of-sample period. As such, we cannot conclude whether or not combined forecasts performed better than individual forecasts in the pseudo out-of-sample period (Figure 5.3.9 and Figure 5.3.10).

Figure 5.3.9 AR(3) and WAIC

```
Diebold-Mariano forecast comparison test for actual : gdpgr
Competing forecasts:  combined_waic versus ar3forecast
Criterion: MSE over 40 observations
Maxlag = 0   Kernel : bartlett

Series                 MSE

combined_waic          32.29
ar3forecast            28.99
Difference             3.298

By this criterion, ar3forecast is the better forecast
H0: Forecast accuracy is equal.
S(1) =     1.253  p-value = 0.2101
```

$t_{DM} = 1.253$; $P = 40$; $h = 1$

$t_{HLN} = 1.2372384$

2-sided p-value = t_{39} = 0.22339657

Figure 5.3.10 AR(1) and BMA

```
Diebold-Mariano forecast comparison test for actual : gdpgr
Competing forecasts:  combined_bma versus ar1forecast
Criterion: MSE over 40 observations
Maxlag = 0   Kernel : bartlett

Series                 MSE

combined_bma           32.53
ar1forecast            27.98
Difference             4.551

By this criterion, ar1forecast is the better forecast
H0: Forecast accuracy is equal.
S(1) =     1.161  p-value = 0.2457
```

$t_{DM} = 1.161$; $P = 40$; $h = 1$

$t_{HLN} = 1.1463956$

2-sided p-value = t_{39} = 0.25861672

We also conducted a DM test between the WAIC combined forecast and the BMA combined forecast, and between the AR(1) forecast and the AR(3) forecast. Since the p-values are high for both, we cannot reject the null hypothesis of equal predictive ability at the 10% significance level. Thus, we cannot conclude that the WAIC combined forecast performs better than the BMA combined forecast. As such, we are unable to conclude which information criteria, AIC or BIC is better in selecting the optimal forecast model for forecasts in the pseudo out-of-sample period. Similarly, we also cannot conclude that the AR(1) forecast does better than the AR(3) forecast. (Figure 5.3.11 and Figure 5.3.12)

Figure 5.3.11 WAIC and BMA

```
Diebold-Mariano forecast comparison test for actual : gdpgr
Competing forecasts:  combined_bma versus combined_waic
Criterion: MSE over 40 observations
Maxlag = 0   Kernel : bartlett

Series                    MSE
______________________________________
combined_bma              32.53
combined_waic             32.29
Difference                .2357

By this criterion, combined_waic is the better forecast
H0: Forecast accuracy is equal.
S(1) =      .5808  p-value = 0.5614
```

$t_{DM} = 0.5808$; $P = 40$; $h = 1$

$t_{HLN} = 0.5734940$

2-sided p-value = $t_{39} = 0.56960276$

Figure 5.3.12 AR(1) and AR(3)

```
Diebold-Mariano forecast comparison test for actual : gdpgr
Competing forecasts:  ar3forecast versus ar1forecast
Criterion: MSE over 40 observations
Maxlag = 0   Kernel : bartlett

Series                    MSE
______________________________________
ar3forecast               28.99
ar1forecast               27.98
Difference                1.017

By this criterion, ar1forecast is the better forecast
H0: Forecast accuracy is equal.
S(1) =      .9734  p-value = 0.3304
```

$t_{DM} = 0.9734$; $P = 40$; $h = 1$

$t_{HLN} = 0.96115549$

2-sided p-value = $t_{39} = 0.34239558$

Therefore, we are unable to conclude that our combined forecasts perform better than the benchmark model AR(1) and AR(3) respectively.

In addition, to test whether oil prices improve our combined forecast, we also conduct the DM test between the model that includes oil prices and the model that did not include oil prices. We compare the best model selected by AIC that includes oil prices: ADL model with 3 lags of GDP growth, 3 lags of housing starts, 3 lags of total retail trade and 1 lag of oil prices, with the same ADL model that does not include oil price. Similarly, we also compare the best model selected by BIC that includes oil prices: ADL model with 1 lag of GDP growth, 3 lags of housing starts, 2 lags of high-yield spread and 1 lag of oil prices, with the same ADL model that does not include oil prices.

Figure 5.3.13 AIC Selected Oil and Non-Oil Price

```
Diebold-Mariano forecast comparison test for actual : gdpgr
Competing forecasts:  aicnonoilforecast versus am8oilpL1forecast
Criterion: MSE over 40 observations
Maxlag = 0   Kernel : bartlett

Series                    MSE
______________________________________
aicnonoilforecast         30.98
am8oilpL1forecast         30.35
Difference                .6344

By this criterion, am8oilpL1forecast is the better forecast
H0: Forecast accuracy is equal.
S(1) =       1.06  p-value = 0.2889
```

$t_{DM} = 1.06$; $P = 40$; $h = 1$

$t_{HLN} = 1.0466661$

2-sided p-value = $t_{39} = 0.30169742$

Figure 5.3.14 BIC Selected Oil and Non-Oil Price

```
Diebold-Mariano forecast comparison test for actual : gdpgr
Competing forecasts:  bm7oilpL1forecast versus bicnonoilforecast
Criterion: MSE over 40 observations
Maxlag = 0   Kernel : bartlett

Series                    MSE
______________________________________
bm7oilpL1forecast          33.8
bicnonoilforecast         32.95
Difference                 .854

By this criterion, bicnonoilforecast is the better forecast
H0: Forecast accuracy is equal.
S(1) =      1.196  p-value = 0.2318
```

$t_{DM} = 1.196$; $P = 40$; $h = 1$

$t_{HLN} = 1.1809554$

2-sided p-value = $t_{39} = 0.24477257$

Since the p-values are high for both tests, we cannot reject the null hypothesis of equal predictive ability at the 10% significance level. Thus, we cannot conclude that the AIC selected forecast with oil price performs better than the AIC selected forecast without oil price. Similarly, we are also unable to conclude that the BIC selected combined forecast with oil price performs better than the BIC selected combined forecast without oil price. Therefore, we are unable to conclude that oil prices improve our forecast from the DM test (Figure 5.3.13 and 5.3.14).

6. Discussion

Our results show that adding oil prices improves the AIC and BIC for the respective combined models. However, results from the DM test concluded that the models with and without oil prices for both AIC and BIC do not differ in their predictive ability. This outcome is rather unexpected as models with oil prices included are expected to perform better as per AIC/BIC. One possible explanation for this occurrence could be due to the linear modelling of oil price that we used. This might not reflect the true relationship between GDP growth rate and oil prices. A non-linear transformation of oil prices might have performed better and possible approaches have been brought up in studies by Hamilton (2003) and Kilian and Vigfusson (2013). Hence, a natural extension of the paper is to re-enact the models using a non-linear model in oil prices instead of a linear model in oil prices and rerun the DM test.

Additionally, the DM tests have shown that models chosen based on AIC and BIC both produce forecasts of equal predictive ability in the pseudo out-of-sample period, and no single information criterion beats the other even when forecasting with a small sample size. Hence, even though BIC penalises additional parameters more than AIC, it does not choose a model that produces better forecasts in the pseudo out-of-sample period. This could simply be an anomalous case in the pseudo out-of-sample period, or it could be that while the sample size is small in this case, it is not extremely small and BIC could potentially perform better with smaller sample sizes. While BIC chosen models reduce estimation uncertainty, they also provide much less information for forecasting. Thus, it could be that the effect of the reduced estimation uncertainty in the BIC chosen models and the effect of better information for forecasting in the AIC chosen models cancel out in this case, such that no model produces better forecasts for the pseudo out-of-sample period.

The true out-of-sample forecasts for the BMA combined forecast model, WAIC combined forecast model, AR(1) model and AR(3) model derived in Section 5 all predict an increase in real GDP growth. This is in line with an article we found from the Chicago FED where they forecasted US GDP growth in the next few quarters using a mixed-frequency Bayesian vector autoregressive model called ALEX (Butters, et.al., 2020). ALEX forecasted GDP growth in 2021 quarter 1 at 0.6%, lower than the 4.23% and 4.64% for AR(1) and AR(3) but higher than the 0.115% and -1.12% for BMA. It is interesting that the forecast lies between the simple AR model forecasts and the WAIC and BMA combined model forecasts, with the AR model providing a very optimistic forecast. The forecast from ALEX was derived after factoring in fresh data from 2020 quarter 3, thus explaining the divergence.

While the AR(1), AR(3), and WAIC combined forecasts predict that GDP growth will see a relatively sharper increase before slowing down, the BMA combined forecast predicts not only a slowdown, but also a decrease in 2021 between quarters 1 and 2. On the other hand, while the WAIC combined forecast predicts a similar trend to the benchmark models, it also predicts the steepest increase in GDP growth rate for the last 2 quarters of 2020, as compared to all the other forecast models.

7. Conclusion

This paper examined whether adding oil prices improves the forecast accuracy of the ADL model which includes "conventional" leading indicators used for forecasting real GDP growth. Although the ADL models that included oil prices result in lower AIC and BIC, we are unable to conclude that including oil prices improves forecast accuracy in the pseudo out-of-sample period from the DM test. However, we noticed that even when we conduct a DM test between the best ADL model selected by AIC that did not include oil prices and the benchmark model, AR(3), we could not reject the null hypothesis that both models had equal predictive ability. Hence, this suggests that perhaps the leading indicators or the lags we selected for the leading indicators based on AIC and BIC may not be optimal.

We would also like to highlight some of the limitations in our study. As we were unable to find data on oil prices in the US before 1986, we are limited by the number of observations we have. As a result, we were unable to select the best model based on Predictive Least Squares (PLS) criterion which selects models with the lowest estimated out-of-sample mean square forecast errors from pseudo out-of-sample forecasts in a separate "hold-out" sample. Selection based on AIC and BIC assumes conditional homoskedasticity in our residuals but this condition is unlikely to hold in GDP growth. In addition, it may be better to take into account not just oil prices in our study, but also prices for future contracts for crude oil since the price of these contracts reflect market expectations of future oil prices. Since investment projects tend to take time, whether a firm chooses to invest now may be based on expectations of oil prices in the future and not oil prices now. However, we are unable to find comprehensive data on future contracts for crude oil and thus, this variable was not considered.

Recent research seems to favor using a Vector Auto-Regressive (VAR) model to predict real GDP growth, using indicators such as unemployment, inflation and asset prices that have a two-way correlation with GDP growth. This is perhaps an extension of our research that can be expanded on in the future.

8. References

Bannerjee, A. & Marcellino, M. (2006). Are there any reliable leading indicators for US inflation and GDP growth?. *International Journal of Forecasting,* 22(1), 137-151.

Brave, S., Butters, A. & Forgarty, M. (2020, October). Looking Down the Road with ALEX: Forecasting U.S. GDP. *Chicago Fed Letter, No. 447.* Retrieved from https://www.chicagofed.org/publications/chicago-fed-letter/2020/447#:~:text=Yet%2C%20notably%2C%20this%20predicted%20growth,to%20its%20pre%2Dpandemic%20levels.

Clark, T., & McCracken, M. (2009). Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts. *International Economic Review, 50*(2), 363-395. Retrieved November 9, 2020, from http://www.jstor.org/stable/20486866

Federal Reserve Bank of St. Louis and Board of Governors of the Federal Reserve System (US), 3-Month Treasury Bill: Secondary Market Rate (TB3MS), retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/TB3MS, November 1, 2020.

Federal Reserve Bank of St. Louis and Board of Governors of the Federal Reserve System (US), 10-Year Treasury Constant Maturity Rate (GS10), retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GS10, November 1, 2020.

Federal Reserve Bank of St. Louis and Moody's, Moody's Seasoned Baa Corporate Bond Yield-Moody's Seasoned Aaa Corporate Bond Yield, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/graph/?g=f09, November 1, 2020.

Federal Reserve Bank of St. Louis and Organization for Economic Co-operation and Development, Total Retail Trade in United States (USASARTMISMEI), retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/USASARTMISMEI, November 1, 2020.

Federal Reserve Bank of St. Louis and U.S. Bureau of Economic Analysis, Ratios of private inventories to final sales of domestic business (A811RC2Q027SBEA), retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/A811RC2Q027SBEA, November 1, 2020.

Federal Reserve Bank of St. Louis and U.S. Bureau of Economic Analysis, Real Gross Domestic Product (A191RL1Q225SBEA), retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/A191RL1Q225SBEA, November 1, 2020.

Federal Reserve Bank of St. Louis and U.S. Census Bureau, Housing Starts: Total: New Privately Owned Housing Units Started (HOUST), retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/HOUST, November 1, 2020.

Federal Reserve Bank of St. Louis and U.S. Energy Information Administration, Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma (DCOILWTICO), retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DCOILWTICO, November 1, 2020.

Harvey, D., Leybourne, S., & Newbold, P. (1998). Tests for Forecast Encompassing. Journal of Business & Economic Statistics, 16(2), 254-259. doi:10.2307/1392581

Hamilton, J. D. (2003). What is an oil shock?. *Journal of Econometrics,* 113(1), 363-398.

Kilian, L. & Vigfusson, R.J. (2013). Do oil prices help forecast U.S. real GDP? The role of nonlinearities and asymmetries. *Journal of Business & Economic Statistics,* 31(1), 78-93.

Kuosmanen, P. & Vataja, J. (2014). Forecasting GDP growth with financial market data in Finland: Revisiting stylized facts in a small open economy during the financial crisis. *Review of Financial Economics,* 23(1), 90-97.

Nonejad, N. (2020). Crude oil price changes and the United Kingdom real gross domestic product growth rate: An out-of-sample investigation. *The Journal of Economic Asymmetries,* 21(1), 1-12.

Scott A. Brave, R. A. (October, 2020). Looking down the road with ALEX: Forecasting U.S. GDP. *Chicago Fed Letter*. The Federal Reserve Bank of Chicago. Retrieved from Federal Reserve Bank of Chicago.

Stock, James, H., and Mark Watson. 2003. "Forecasting Output and Inflation: The Role of Asset Prices." *Journal of Economic Literature*, 41 (3): 788-829. DOI: 10.1257/002205103322436197