Master's Thesis Financial Economics

Multi-Horizon Forecast Comparison of Linear and Non-linear Methods for GDP Growth and Inflation

Erasmus School of Economics

Erasmus University Rotterdam

Xinyu Xing 411780

Supervisor: Dr. Rogier Quaedvlieg

Abstract

Forecasting macroeconomic variables is an important issue for decision makers. In this paper, we analyze 13 univariate models which include linear and non-linear methods to forecast GDP growth and inflation to determine a benchmark. We not only examine the models in single horizon by the Diebold-Mariano test, but also apply the multi-horizon analysis from Quaedvlieg (2018). By conducting the research on 4 European countries with different fundamentals, we find that *Autoregressive method* with fixed lags performs the best in our sample, which is the benchmark model that we propose. The multi-horizon superior predictive ability(SPA) test can be an useful tool in comparing large set of forecasts which delivers reliable results.

JEL classifications: C22, C52, C53, C58

Keywords: Forecasting; Multi-horizon testing; Superior Predictive Ability; Growth: Inflation

May 6, 2020

Contents

1 Introduction

Predicting the future evolution of macroeconomic variables has always been a central concern in economics. Among which GDP growth and inflation are the most vital indicators for investors, government and economists to make certain decisions. Numerous research has been done to analyze the models for forecasting GDP growth and inflation. Some of these models are economic theory-based while the others are time series models. Good time series models can not only provide outstanding forecasts but also indicate to what extent economic theory adds value in forecasting. As for long time panel, simple linear models may not be flexible enough to explain the patterns in macroeconomics time series due to the continuous changes in politics, economics and culture. In this regard, time-varying and neural network models, which are more flexible, could have a comparative advantage in forecasting. Marcellino (2007) compares the performance of 55 univariate models to evaluate whether the more sophisticated times series methods outperform standard linear models when forecasting US GDP growth and inflation. By comparing individual horizon analysis, Marcellino (2007) concludes that the general linear time series models cannot be easily beaten if they are carefully specified. However, the forecast performance should not only be compared at the individual horizon level. The full forecast path is also important in many cases while making fiscal and monetary decisions. Moreover, when comparing the performance of a large number of models, the individual horizon analysis may lead to incoherent results. As some models provide better predictability in the short run, others can work well in the long run, or give impressive results in some specific horizons. Hence, the traditional approach which compares the performance at different horizons independently is less intuitive and also less efficient when comparing multiple models.

Recently, Quaedvlieg (2018) presents a test method which compares the forecasting performance at a multi-horizon level instead of individual horizon level. This has a superior advantage to the original way of determining which model outperform the other. The author has proposed two types of test statistics regarding the concept of multi-horizon superior predictive ability(SPA). The statistic of uniform SPA (uSPA) indicates a model has a lower loss at each individual horizon while the statistic of average SPA(aSPA) considers the average loss level of a forecast. Quaedvlieg (2018) applies the multi-horizon test on the research results by Marcellino et al. (2006), which evaluates the performance difference in iterated and direct forecasting method. The test results imply that the multi-horizon analysis can provide a more comprehensive and more consistent picture in comparing forecast path accuracy.

In this research, we compare the forecasting performance of 13 univariate models in 8 horizons for GDP and inflation growth. The selected models cover 3 categories: (1)Linear models, (2)Time-varying models, and (3)Artificial neural network (ANN) models[1]. We not only investigate the performance of each model in the individual horizon but also jointly take all horizons into account by using aSPA and uSPA test. The method we used to compare the performance in individual horizon is Diebold-Mariano test. The input of all tests is the loss function defined by mean squared prediction error (MSE) . The analysis of each model can be applied both in-sample and out-of-sample. As our emphasis is on the complete forecasting path of GDP and inflation change, we only assess the out-of-sample performance.

The data set that used in this study is the quarterly GDP growth and consumer price index(CPI) of France, Greece, The Netherlands and Portugal. The country selection is based on the GDP correlation between European countries. The time span ranges from 1969q1 to 2019q4[2] which covers the evolution of GDP and inflation for 50 years. This is a rather long time span which covers a few business cycles and different behaviors of GDP growth and inflation. We are forecasting with a rolling window of 124 quarters which is from 1969q1 to 1999q4 and the out-of-sample forecasting start from 2000q1.

The methodology framework is extracted from Stock and Watson (1998) and Marcellino (2007), which both analyze a large number of univariate models on forecasting US macroeconomic time series. They compare the predictability of each type of model without prespecifying them. The forecasting methods that used in this study lie in the same categories of these literature, but for non-linear methods, we choose different models. The aim of this study is to further research on the forecasting performance of univariate models from both

[1] <The detailed model specification is shown in table 3 >

[2] <For GDP growth ranges from 1969q1 to 2019q3 >

individual horizon level and multi-horizon level. In addition, this analysis also investigates the performance of these models when forecasting various times series.

The structure of this paper is reported as follow. Section 2 discusses the existing literature of macroeconomic forecasting and the performance of different univariate models. Section 3 illustrates the data sets that we used, followed by section 4 provides the model specification and introduction of the tests. In section 5, we investigate the empirical results of the chosen models. Section 6 discusses the application of the finding and limitations of this study. Last but not least, section 7 concludes.

2 Literature Review

A large variety of models are proposed to analyse and forecast macroeconomic variables, and the styles have evolved more and more sophisticated along time. In general, there are more researches about factor models than univariate models in this area. Rünstler et al. (2009) has conducted a real-time forecast exercise for GDP growth with large datasets where they included uni-variate models as a benchmark to test the predictability of other popular factor models. In this section, we first have an insight into former literature which compares the performance of univariate models. Then discuss the previous studies on evaluating model's forecasting abilities.

2.1 Empirical research on univariate models

Several popular research regarding forecast macroeconomics variables with univariate models are conducted by Massimiliano Marcellino, James H. Stock and Mark W. Watson since 1990s. Stock and Watson (1998) analyze 49 univariate models to see if the non-linear methods outperform the simple linear models in practice and further test the benefit of using non-parametric approaches. They also combine different methods to tested the predictability of the pooling procedure. With the evidence of modeling and predicting 215 U.S. monthly macroeconomic time series, they find the best performance is the pre-specified AR model, and this result can be further improved by combining with other methods. Marcellino et al. (2003) and Marcellino (2004) make research in this line for European area. The latter literature evaluates 58 univariate models and 18 pooling procedures of about 500 macroeconomics variables for 11 countries in European Monetary Union (EMU). They again find the pooling methods usually give good results, but for some series, simple linear models can outperform any other methods. Last but not least, non-linear models forecast best for unstable series which lead to outstanding results for some specific variables of some countries.

Later, Marcellino et al. (2006) conduct research to compare the forecasting performance of direct and iterated methods with univariate and bivariate models. By simulating the out-of-sample forecast to 170 U.S. macroeconomics time series, they conclude that iterated method provides a better forecast in most of the time and the difference is more distinct when the model is specified with long-lag selection. One year later, Stock and Watson (2007) analyze the structure terms of U.S. inflation and reveal the challenge of forecasting it. In this study, they find the univariate models outperform multi-variate models in describing the patterns of inflation and the time-varying method with moving average process gives the best results. Same year Marcellino (2007) propose further research base on Marcellino (2004) regarding forecasting GDP growth and inflation in the U.S who compares 55 univariate models in order to find a linear benchmark. In general, they find the linear methods still outperform more sophisticated time series models. However, in some cases, non-linear specifications can yield substantial gains.

In both Stock and Watson (1998) and Marcellino (2007), artificial neural network(ANN) models are analyzed as comparison to linear methods. However, the performance of such models were not as stable and accurate as the linear methods in their sample. Since the computing power are better now than ever, researches never stop in finding more accurate machines learning(ML) models. Inoue and Kilian (2008) show that integrating bagging method in predicting macroeconomic series can reduce mean squared prediction errors(MSE) significantly and outperform the univariate benchmark in forecasting U.S. inflation. A newly propose research Medeiros et al. (2019) also further study on the base of Stock and Watson (1998) regarding ML methods. They analyze a vast of ML models in forecasting U.S. inflation and find some of them can give better results than the AR benchmark for several horizons. More advanced models are tested by Cook and Hall (2017) who aim to improve the accuracy of predicting macroeconomic variables with deep neural network methods. Their results show that most of the neural network methods outperform the benchmark in the early horizons while the model based on encoder-decoder architectures give the most superior results up to the fourth quarter in forecasting U.S. unemployment rate. In our analysis, we only adopt single layer artificial neural network model as comparison with different hidden units for the simplicity.

2.2 Forecast evaluation methods

There are alternative ways to evaluate the forecasting ability of models. One typical method is the forecast encompassing test which proposed by Chong and Hendry (1986) and Fair and Shiller (1988). This method analyzes two sets of forecast and test whether one provides additional information than the other. The test approach of Diebold and Mariano (1995) emphasizes the accuracy difference of two competing forecasts which can be applied in many cases regardless of multi-types of forecast error. Later on, the Diebold-Mariano(DM) test becomes popular to test for the significant difference between two set of forecasts on the individual horizon level. Base on the framework of Diebold and Mariano (1995), White (2000) develop a direct procedure to test if the best model among a bunch of methods has benefit from data reusing which is practically unavoidable in forecasting time series. Also to tackle the problem of data snooping, Hansen (2005) updates the SPA test of White (2000) by implementing the reality check(RC). The test from Giacomini and White (2006) is more focus on the out-of-sample predictive ability and take into account the conditional evaluation. Their method help greatly in finding the most performing model under different economic conditions. All above tests compare the accuracy of forecasting performance on the individual horizon level. Based on them, Quaedvlieg (2018) introduced a multi-horizon SPA test which jointly compares the forecasting results from each horizon. He has segmented the test into i) *uniform* multi-horizon SPA(uSPA) test which compares the loss at individual horizons; ii) *average* multi-horizon SPA(aSPA) which allows the superior results at some horizons to cover the loss at poor performance horizons. By applying the uSPA and aSPA pairwise test, one can easily compare the forecasting ability of two models for the whole forecast paths. From Diebold and Mariano (1995) to Quaedvlieg (2018), their goal is to test for multi-horizon SPA within a finite sample, which compares the accuracy of forecasts generated by the estimated value of parameters. A similar concept is to test for the SPA in a population level which compares the accuracy of forecasts implied by population value of parameters. More literature in this line can be consulted in Clark and McCracken (2013).

The superior predictive ability(SPA) tests use loss functions as input, which are derived from respective forecasting errors. Some typical loss functions are the common mean absolute error(MAE) and the mean square predict error (MSE). The loss level itself can also be analyzed to see the forecasting performance. However, a lower loss level does not imply the difference between two set of forecasts is significant. Hence, further tests are needed.

In this study, we are interested in finding a benchmark model in forecasting GDP and inflation growth and have two extensions comparing to previous studies: I) The benchmark model is not only selected by individual horizon comparison, but also through multi-horizon comparison. To do so, we first analyze the MSE of all models, then apply the DM, uSPA, aSPA tests to access the forecasting performances. II) We analyze 3 types of univariate model which include simple AR models, time-varying models and more sophisticated NN models to 4 European countries with different fundamentals. We aim to have a more comprehensive understanding of these models while forecasting diverse time series. Therefore, we know the reliability of a benchmark model to predict various time series.

3 Data

In this section, we give a more detailed description of data selection and corresponding descriptive statistics. The data used in this research is obtained from the OECD Database. First, we collect the quarterly GDP growth and quarterly CPI for 27 European Union (EU) countries from 1969q1 to 2019q4 [3]. Then according to the data completeness, 12 countries are maintained for the further research. For the purpose of this paper, we desire variety in time series. Hence a correlation matrix is generated to find the correlation of GDP between countries. Later we rank the level of correlations for each country, and the 1st, 4th, 8th and 12th rank are chosen for the further research. The correlation matrix for CPI has also been generated; however, the correlation between CPI of all countries are considerably high as you can find both correlation matrix in table 14.a, 14.b in Appendix. Therefore, we only choose the countries by the correlation level of GDP. The selected countries are France, Greece, The Netherlands and Portugal where Greece is the least correlated country with the other 11 countries regarding GDP, France is the most correlated country with the rest countries, and The Netherlands and Portugal are the 4th and 8th correlated countries respectively.

3.1 GDP Growth

The data collected for GDP is measured by the percentage change of the previous period, and the data frequencies are quarterly. Hence it can be interpreted as the quarterly GDP growth. The length of the time series covers 50 years until the third quarter of 2019. The total observation is 203 for each country, and there is no missing value. The descriptive statistics of four countries are shown in table 1 and the plot of GDP growth revolution can be seen in figure1. The unit-root tests have been applied to 4 time series and the augmented Dickey–Fuller (ADF) statistics are listed in the table. All ADF statistics are significant at 1% confidence level, which indicates all time series is stationary. Moreover, the Breusch-Pagan test is also applied to all time series to detect for heteroskedasticity. Observe from the χ^2 values, Greece and the Netherlands do not have constant variance and hence indicate

[3] <Sample length is from 1969q1 to 2019q3 for GDP due to most data availability. >

heterskedaticity in the time series.

	Abbv.	Mean	Med.	St.Dev	Min	Max	ADF	χ^2
France	FRA	0.45	0.50	0.56	-1.66	2.45	-5.22***	0.46
Greece	GRC	0.25	0.45	2.62	-6.85	10.43	-5.03***	12.21***
The Netherlands	NLD	0.52	0.62	1.13	-4.91	5.83	-4.76***	7.33***
Portugal	PRT	0.47	0.57	1.20	-2.52	4.90	-5.83***	1.80

Table 1: Descriptive statistic for GDP growth rate with ADF, χ^2 test statistics. *** implies the respective p-value is lower than 0.01.

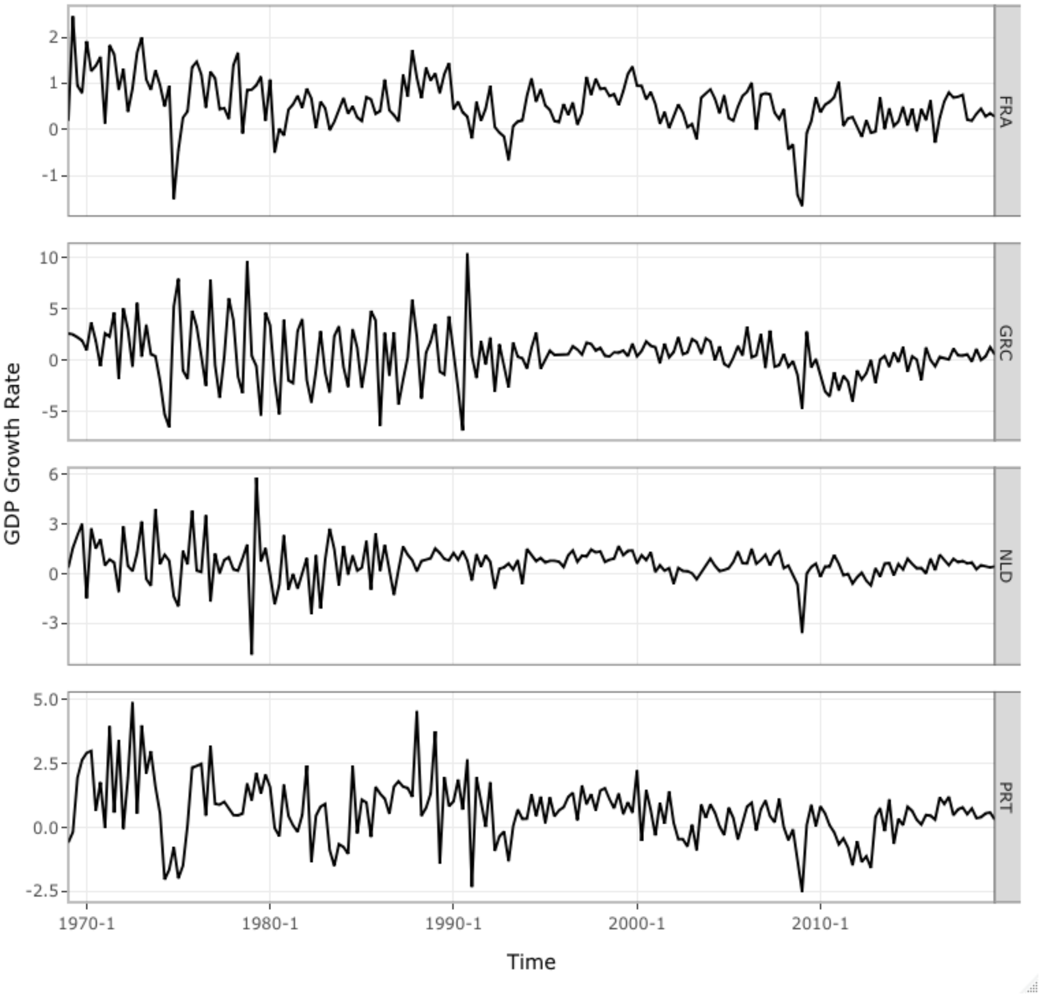

Figure 1: GDP growth evolution of 4 countries

In general, the Netherlands has the highest average growth rate in GDP from 1969 to 2019, while Greece had the lowest average GDP growth. In the meantime, France had the smallest volatility over the four countries, while Greece had the most volatile GDP growth. From

Figure1 one can observe that before 1995, all countries showed comparatively larger volatility then beyond and around 2008, there is a significant dip in GDP growth for all countries. For each country, GDP growth fluctuates around a corresponding fix level.

3.2 Inflation Growth

Inflation is a quantitative measurement of the increase in the average price level of a basket of goods and services purchased by people or household in an economy over a time period. It can also indicate the purchasing power of the nation's currency. If the inflation of a country increases, the corresponding currency depreciates. In this research, we also interested in forecasting inflation growth. The data is obtained from the OECD database, which covers 1969q1 to 2019q4. The total observation is 204, and there is no missing value. The original data series is the Consumer Price Index(CPI) with the base year 2015, and we calculate the inflation growth in year t by the following method:

$$\Delta Inflation_t = \frac{CPI_t}{CPI_{t-1}} - 1$$

The descriptive statistics of CPI growth are shown in table 2 and the evolution charts are demonstrated in figure 2. In general, we can see that CPI growth is less volatile than GDP growth. Among the 4 countries, Greece and Portugal have relatively higher inflation growth which are around 2%. Meanwhile, The Netherlands and France have lower inflation growth which are both around 1%. Different from GDP, inflation growth does not fluctuate around a fixed number but sometimes show a trend. For instance, CPI growth of France from 1970 to 1990 first shows an upward trend then a downward trend. The volatility of CPI growth before 1990 is larger than after for every country and the level of CPI growth is also higher. After 1995, most countries have relatively steady change in CPI growth. The stationary test is also applied to each time series. The inflation growth of Portugal and Greece are stationary at 1% and 5% significant level respectively. However, France and The Netherlands do not have stationary CPI growth rate. As we are forecasting with roiling window and more focus on the out-of-sample performance, further test of stationary is conducted to the out-of-sample time series. Then for both France and The Netherlands, CPI growth rate is stationary at 1% significant level from 2000q1 to 2019 q4. As for heteroskedasticity, only the CPI growth for Netherlands has a constant variance and the rest are all heteroskedastic.

	Abbv.	Mean	Med.	St.Dev	Min	Max	ADF	χ^2
France	FRA	0.01	0.01	0.01	-0.01	0.04	-2.57	4.56**
Greece	GRC	0.02	0.02	0.03	-0.03	0.12	-3.72**	32.30***
The Netherlands	NLD	0.01	0.01	0.01	-0.01	0.05	-2.891	2.62
Portugal	PRT	0.02	0.01	0.03	-0.05	0.14	-4.08***	67.62***

Table 2: Descriptive statistic for Inflation growth with ADF, χ^2 statistics. * implies the respective p-value is lower than 0.1; ** implies the respective p-value is lower than 0.05; *** implies the respective p-value is lower than 0.01.

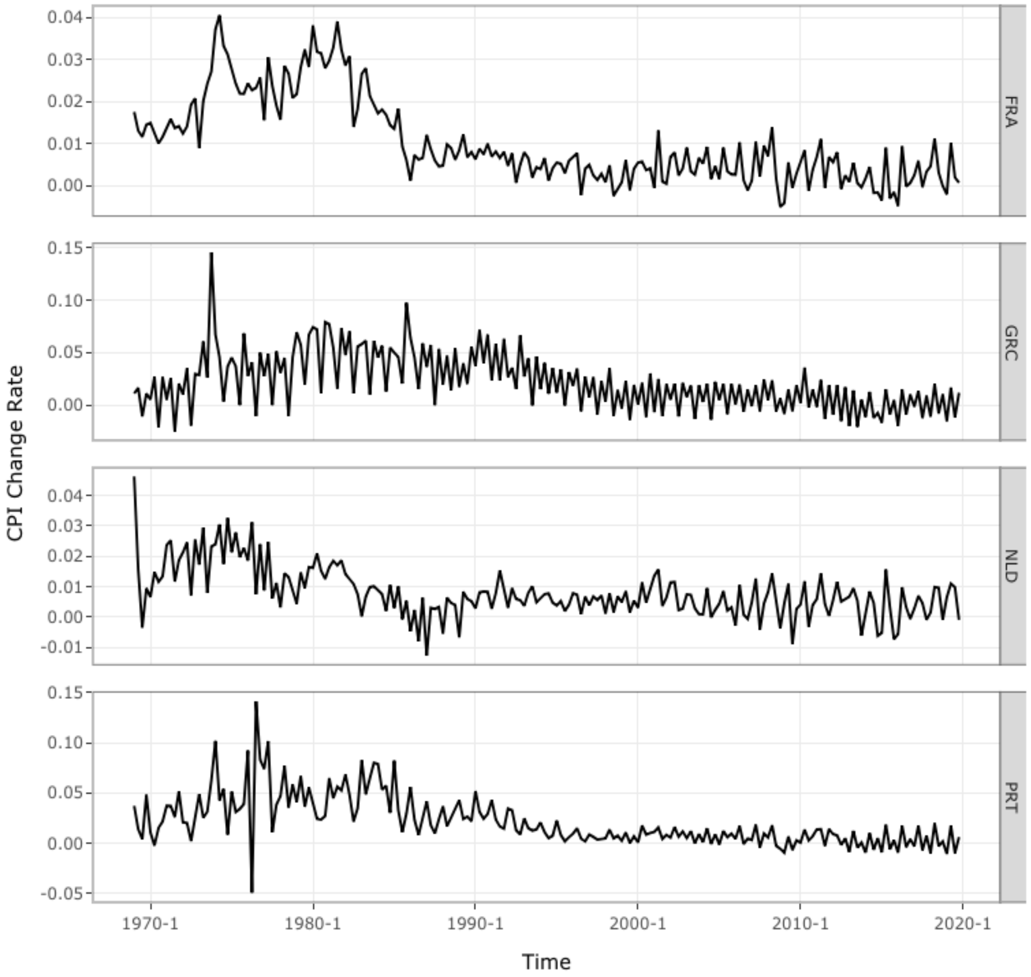

Figure 2: CPI growth evolution of 4 countries

4 Methodology

This section demonstrates the models that are used to forecast GDP and CPI growth and the tests to compare the forecasting performance. We consider four representative univariate model categories, which range from simple autoregressive models to more sophisticated artificial neural network setups. First, we introduce the general forecasting format, which can be written as:

$$y_{t+h} = f(X_t; \delta_t^h) + \epsilon_{t+h}, \tag{1}$$

where y_t represents the GDP and CPI growth of each country in time t, h indicates the forecast horizon, X_t is a vector of explanatory variables, δ_t is a vector of parameters and ϵ_t is the error term. Forecasting models are determined by both function f and regressors X_t. The h-step ahead forecast is given by:

$$\hat{y}_{t+h} = f(X_t; \hat{\delta}_t^h), \tag{2}$$

and the respective forecast error is the difference between (1) and (2):

$$\varepsilon_{t+h} = y_{t+h} - \hat{y}_{t+h}. \tag{3}$$

We not only concern the 1-step ahead forecast but also the full forecast path of each model. Therefore, we obtain the forecast value at h=1,2,3...8, which enable us to compare the forecast performance from 1 quarter to 2 years ahead. In subsection 4.1, we give the details of model specified, and in subsection 4.2, 2 measurements of forecasting performance are introduced.

4.1 Forecasting Models

4.1.1 Linear methods

Autoregressive(AR) Models. An autoregressive model is structured as the current value of a variable, y, depends on only the values it took in previous periods plus an error term. Box and Jenkins (1970) has introduced AR models comprehensively in predicting time series of economics. In practice, AR models were proved to have superior results by many researchers. They can be hardly beaten if the components are correctly specified. Marcellino (2007) The AR model is model(1) with a linear f function, and the detailed format can be written as:

$$y_{t+h} = \sum_{j=0}^{L-1} \delta^h y_{t-j} + \mu^h + \epsilon_{t+h}, \tag{4}$$

where L is the number of chosen lags and μ is either a constant or constant with a linear trend. The number of lags is chosen by three methods: either a fixed number of 4 or a moving number chosen by AIC and BIC with a maximum of 6. With different combinations, we have 6 models in this category.

Random Walk(RW) Model. Random walk model, which also known as the No change model, can be written as $y_{t+h} = y_t$. This method indicates the value of the next horizon is the same value as the current observation. Random walk model is popular not only for its simplicity but also the hard beaten results in some cases. Artis and Marcellino (2001) has used a random walk model to forecast government deficit and found a better performance than the IMF forecast for most of the countries.

4.1.2 Non-linear methods

Time-varying autoregressive (TVAR) models. A TVAR model allows the coefficients of the explanatory valuables to vary over time. Unlike Marcellino (2007), Nyblom (1989) and Stock and Watson (1996) who assume the coefficients follow a random walk model, this research we adopt a non-parametric time-varying coefficient model. In this case, the model is defined as same as model(4) while $\delta_t^h = \left(\delta_{t,0}^h, \delta_{t,1}^h, \ldots, \delta_{t,j}^h\right)^\top$ is a vector of functions of time t. Follow the research of Robinson (1989),Cai (2007) and Chen et al. (2018), we assume that

$$\delta_{t,d}^h = \delta_d^h\left(\tau_t\right), \tau_t = \frac{t}{n}, \quad t = 1, \ldots, n, \quad d = 0, \ldots, j \tag{5}$$

where n is the number of observations. Meanwhile, the function of δ_t^h is estimated with a consistent non-parametric kernel regression. [4] This assumption implies an increasing intensity of data points in the interval of [0,1]. Note that this method allows for heteroscedasticity in error terms which may yield better forecasts when dealing with heterogeneous time series. We have 3 TVAR models which regress on constant with 3 lags and number of lags chosen by AIC, BIC.

Artificial neural network(NN) models. Artificial neural network method becomes popular since 1990s for the various use in modelling and forecasting. A NN model can approximate almost any non-linear function very close. Hence, when dealing with a truly non-linear dynamic time series, the NN model will detect the term-structure and display a superior fit comparing to simple linear or time-varying models. However, a good in-sample fit does not imply an outstanding out-of-sample forecast. As NN model can be very flexible to have a nearly perfect in-sample fit, they sometimes overfit. Considering the complexity and overfitting of the model, we adopt a single layer neural network model; more details see Franses et al. (2000). The single-layer neural network model with q hidden units and activation function g can be represented as :

$$y_{t+h} = \delta_0^h + \sum_{j=1}^{q} \delta_j g\left(\gamma_{0j} + \sum_{i=1}^{m} \gamma_{ij} y_{t-(i-1)d}\right) + \epsilon_{t+h} \tag{6}$$

where m is the number of embedding dimension and d is the time delay. As different researchers have chosen distinct value of the embedding dimensions, in this research, we define $m = 3$ and $d = 1$. The model is estimated via Conditional Least Square (CLS) and we report the results when q=1,2,3. In order to get consistent results, we use a fixed seed of 5.

Overall, we consider 13 models in the forecast comparison: 7 linear models, 3 time-varying models and 3 NN models. Specific model description can be seen in table3.

[4] More of the kernel regression see Hart (1991).

	Lables	Model type	Constant(C) or trend(T)	Lag chosen
1	ARC_4	Autoregressive	C	4
2	ART_4	Autoregressive	T	4
3	ARC_{AIC}	Autoregressive	C	AIC
4	ART_{AIC}	Autoregressive	T	AIC
5	ARC_{BIC}	Autoregressive	C	BIC
6	ART_{BIC}	Autoregressive	T	BIC
7	RW	Random walk	/	/
8	TV_3	Time varying AR	C	3
9	TV_{AIC}	Time varying AR	C	AIC
10	TV_{BIC}	Time varying AR	C	BIC

	Lables	Model type	Number of hidden units	Embedding dimension
11	NN_1	Artifical Neural Network	1	3
12	NN_2	Artifical Neural Network	2	3
13	NN_3	Artifical Neural Network	3	3

Table 3: Model Specification

4.2 Measurement of forecast ability

The first comparison of forecasting performances is conducted by the mean square predict error(MSE) which is also used as the input of DM, uSPA and uSPA tests. The MSE of h-step ahead forecast at time t for model i is calculated as:

$$MSE_{i,t+h} = \frac{1}{T}\sum_{t=1}^{T} \varepsilon^2_{i,t+h}. \tag{7}$$

where $\varepsilon_{i,t+h}$ is the forecasting error for model i at time t for the h horizon.

The Diebold-Mariano(DM) test and multi-horizonal SPA tests all work pair-wisely. The DM test which proposed by Diebold and Mariano (1995) is commonly used to verify the significant difference between two sets of forecast at single horizon level while the uSPA and aSPA tests from Quaedvlieg (2018) are used to test for multi-horizon predictive power. To apply all tests, we need the loss differential. For each model i, we have a loss matrix $L_{i,t}$ whose elements $L^h_{i,t}$= $l_{i,t+h}$. Then for two sets of forecast i and j, the loss differential is calculated as:

$$d_{ij,t} = L_{i,t} - L_{j,t}, \tag{8}$$

which is a matrix of $t \times h$. The null hypothesis of the DM test is there is no difference between the accuracy of two sets of forecasts or the population mean of the loss differential series is zero Diebold and Mariano (1995). The null hypothesis against the alternative can be written as :

$$H_0^{DM} : E(d_{ij,t}) = 0 \;\; \forall\, t; \;\; H_1^{DM} : E(d_{ij,t}) \neq 0 \tag{9}$$

where $E(d_{ij,t})$ is the population mean of the loss differential. The sample mean of loss differential is calculated as $\bar{d_{ij}} = \frac{1}{T}\sum_{t=1}^{T} d_{ij,t}$ and the test statistic of DM is derived as:

$$DM = \frac{\bar{d_{ij}}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}} \tag{10}$$

where $\hat{f}_d(0)$ is a consistent estimator of $f_d(0)$. $f_d(0)$ is the spectral density at frequency 0 which is calculated with the autocovariance of the loss differential between 2 competeing forecasts(More to see Diebold and Mariano (1995)). The DM test is a two-sided z-test and the null hypothesis of equal accuracy is rejected when $|DM| > z_{\alpha/2}$. According to Harvey et al. (1997),the DM test statistic follows an asymptotically N(0,1) distribution under the null hypothesis. From the simulation study of Diebold and Mariano (1995), the DM test tend to reject the null too often when the sample size is too small due to the serial correlation between the prediction errors.

The second measurement is based on two test statistics which both compare the performance of models at multiple horizons jointly within a forecast path. The first statistic analyze the uniform superior predictive ability(uSPA), which indicates a model outperform the other at each individual horizon. The second statistic tests the average superior predictive ability(aSPA) , which implies a model outperform the other on average among all horizons. The latter statistic has a much looser standard as it allows poor performance at some horizons to be compensated by superior performance at other horizons. To apply the tests, we follow the procedure of Quaedvlieg (2018). Again, we have the loss differential matrix $d_{ij,t}$ from equation 8. The test statistic of SPA is derived from the expected value of loss differentials: $u_{ij}^h = E(d_{ij,t}^h)$. For uSPA, we aim to test if a model provide superior forecast results for every horizon. Hence the test statistic for uSPA is defined as:

$$u_{ij}^{(Unif)} = \min_h u_{ij}^h. \tag{11}$$

Under the circumstance that $u_{ij}^{(Unif)} > 0$, model j outperforms model i for all horizons. The corresponding null and alternative hypothesis for uSPA test are:

$$H_{0,uSPA} : u_{ij}^{(Unif)} \leq 0; \; H_{1,uSPA} : u_{ij}^{(Unif)} > 0. \tag{12}$$

When the p-value of the uSPA test result is less than 0.05, the null hypothesis is rejected on 5% significant level. Then we conclude model j has an uniform multi-horizon superior predictive ability(uSPA) than model i at 5% significant level.

The test for aSPA considers the average performance of all horizons which requires a weight factor in the calculation. The test statistic of aSPA is shown as(Quaedvlieg (2018)):

$$u_{ij}^{(Avg)} = w' u_{ij} = \sum_{h=1}^{H} w_h u_{ij}^h \tag{13}$$

where $w = (w_1, ...w_H)$ is the weight factor for different horizons. In this research, we assume the weights for all horizons are equal which implies $w_1 = ... = w_H = \frac{1}{8}$. Similarly, the null and alternative hypothesis for aSPA test are:

$$H_{0,aSPA} : u_{ij}^{(Avg)} \leq 0; \; H_{1,aSPA} : u_{ij}^{(Avg)} > 0 \tag{14}$$

and we conclude model j has an average multi-horizon superior predictive ability(aSPA) than model i at 5% significant level when p-value is less than 0.05.

5 Empirical Results

In this section, we report the forecasting results of all tested models. The in-sample period is from 1969q1 to 1999q4, which is also the length of the rolling window. The forecasting period is from 2000q1 to 2019q3 for GDP growth and for CPI growth is one quarter longer. There are three subsections that discuss the models' performance for GDP and inflation growth that compared by MSE, DM test, uSPA and aSPA tests respectively.

5.1 Forecast evaluation by MSE

The MSE results of forecasting GDP and inflation growth of 4 countries are shown in table 4 and 5 respectively. In order to compare them more easily, the MSE has been normalized to the results of ARC_4 (Autoregressive model which regress with 4 lags and a constant) which we use as the benchmark in this section. The original level of MSE can be seen from figure 3 to 10 in the appendix, which also indicate the evolution of the forecasting performance among each horizon.

5.1.1 GDP growth

Consulting figure 3 to 6 and table 4, we can compare the forecasting performance of each model from their loss level. First, we discuss the results of linear methods which are AR models and random walk. It can be observed that AR models usually give more consistent results for all horizons comparing to other methods for all countries. ARC_4 is the best AR model in most of the cases while sometimes ART_4, ARC_{AIC} and ARC_{BIC} perform slightly better in the shorter horizons (max h=2). Moreover, we see that AR models with a linear trend specification do not improve the forecasting performance of GDP growth. As for the random walk, it shows a more volatile performance than other AR models and its MSE increases in longer horizons. Time-varying AR models have similar MSE as to AR models and TV_3 performs slightly better than other time-varying methods in most scenarios. The artificial neural network methods are the least performing methods while giving inconsistent results for different horizons. In particular, NN_3, which contains more hidden units in the network always present a spike in MSE for all countries' GDP. When the number of hidden units increases in the NN model, the results become more unstable and worse in our sample.

Overall, AR and TV methods provide more stable and satisfying results among all methods. In particular, models with fix lags tend to perform better than lags chosen by AIC and BIC. Colored cell in table 14.a indicates the best performance for each horizon. For France and The Netherlands, the best forecasting models are highly possible to be ARC_4 and TV_3 respectively. However, for Greece and Portugal, we cannot conclude which model performs the best as there is no single model excel on all horizons. If we focus on the actual level of MSE of each country, it can be found that most models have a value between 0 to 1 for France, The Netherlands and Portugal. As for Greece, all models have MSE exceed 1.4, which indicates that Greece GDP is hard to forecast only by univariate models.

5.1.2 Inflation growth

Figure 7 to 10 and table 5 demonstrate the evolution of MSE of all horizons and the normalized MSE of forecasting CPI for each country. Comparing to the forecasting results of GDP where most models show a consistent performance for all horizons(except for NN methods), the forecast of CPI demonstrate more fluctuated outcomes. AR models still present more consistent forecast and lower MSE while ARC_4 and ART_4 are the most stable and superior among all models. Other AR models with the lags chosen by AIC and BIC perform significantly worse than these two. Meantime, we find adding a linear trend to the AR model yield to small gains for some countries. Regarding Random walk, which offers a fluctuated forecast for different horizons, performs significantly better when h=4 and 8. This may imply a seasonal pattern in inflation growth. TV models cannot compete with ARC_4 and ART_4 in forecasting CPI. They show parallel forecast paths as AR models but with higher MSE. NN methods demonstrate similar results to TV methods while NN_3 is the least consistent one which gives abnormal forecast occasionally. Different from GDP growth, where the forecast paths tangle together, the results of CPI are more distinct. From the table and figures, we can conclude ARC_4 gives a superior forecast for France and The Netherlands while ART_4

performance better in most horizons for Greece and Portugal.

Table 4: Normalized MSE of forecasting GDP growth for each country

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
ARC_4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	0.993	1.006	1.028	1.054	1.064	1.088	1.096	1.099
ARC_{AIC}	1.009	1.030	1.068	1.096	1.142	1.179	1.188	1.208
ARC_{BIC}	1.009	1.036	1.067	1.105	1.135	1.188	1.192	1.211
ART_{AIC}	1.006	1.030	1.074	1.108	1.158	1.196	1.206	1.227
ART_{BIC}	1.001	1.034	1.066	1.114	1.147	1.202	1.206	1.222
RW	1.093	1.134	1.234	1.319	1.333	1.419	1.440	1.484
TV_3	1.171	1.069	1.033	1.053	1.063	1.052	1.043	1.044
TV_{AIC}	1.256	1.120	1.027	1.068	1.074	1.047	1.052	1.044
TV_{BIC}	1.237	1.108	1.065	1.076	1.071	1.045	1.046	1.041
NN_1	1.357	2.647	3.362	3.754	2.984	3.796	3.393	3.508
NN_2	4.325	2.115	1.867	1.515	2.260	2.287	2.997	1.256
NN_3	1.124	1.229	1.745	1.772	6.943	1.682	2.092	1.186

Table 4.a France

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
ARC_4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	1.079	1.061	1.054	1.061	1.096	1.108	1.113	1.125
ARC_{AIC}	0.977	0.996	1.118	1.091	1.078	1.153	1.164	1.189
ARC_{BIC}	0.978	0.998	1.120	1.095	1.081	1.160	1.172	1.193
ART_{AIC}	1.091	1.131	1.281	1.268	1.258	1.353	1.397	1.420
ART_{BIC}	1.091	1.133	1.282	1.271	1.260	1.359	1.403	1.424
RW	1.035	0.945	0.990	1.197	0.915	1.060	1.271	1.136
TV_3	1.164	1.173	1.032	0.914	0.896	1.029	1.057	1.011
TV_{AIC}	0.890	0.884	1.017	1.092	1.018	1.051	1.053	1.051
TV_{BIC}	0.891	0.884	1.016	1.093	1.018	1.051	1.052	1.051
NN_1	1.064	1.156	1.095	1.101	1.026	1.062	1.083	1.176
NN_2	1.260	1.119	1.196	1.133	1.145	1.101	1.261	1.157
NN_3	1.229	1.183	1.881	1.382	1.166	1.353	1.338	1.368

Table 4.b Greece

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
ARC_4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	1.060	1.035	1.045	1.046	1.040	1.050	1.051	1.049
ARC_{AIC}	0.984	1.086	1.075	1.075	1.078	1.104	1.081	1.056
ARC_{BIC}	0.993	1.087	1.076	1.070	1.076	1.097	1.073	1.039
ART_{AIC}	1.022	1.140	1.137	1.143	1.148	1.172	1.160	1.135
ART_{BIC}	1.023	1.130	1.129	1.129	1.140	1.164	1.145	1.117
RW	0.998	1.182	1.249	1.244	1.329	1.337	1.410	1.464
TV_3	0.932	1.042	1.027	0.984	0.993	0.996	0.997	0.997
TV_{AIC}	0.899	1.024	1.025	1.037	1.043	1.035	1.014	1.007
TV_{BIC}	0.928	1.015	1.016	1.022	1.036	1.021	1.009	1.001
NN_1	0.941	1.048	1.005	1.019	1.003	0.996	1.014	2.416
NN_2	1.753	1.073	1.283	1.048	1.049	1.118	1.013	1.051
NN_3	1.597	1.169	1.110	1.124	1.525	1.146	2.683	1.125

Table 4.c The Netherlands

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
ARC_4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	1.005	1.008	1.016	1.026	1.034	1.041	1.044	1.049
ARC_{AIC}	1.024	1.135	1.090	1.069	1.043	1.024	1.048	1.016
ARC_{BIC}	1.013	1.110	1.074	1.056	1.030	1.021	1.047	0.994
ART_{AIC}	1.037	1.156	1.116	1.100	1.079	1.059	1.080	1.048
ART_{BIC}	1.018	1.118	1.092	1.084	1.068	1.056	1.080	1.026
RW	1.148	1.165	1.157	1.213	1.254	1.285	1.163	1.274
TV_3	0.955	1.025	1.026	1.017	1.021	1.023	1.023	1.019
TV_{AIC}	1.140	1.067	1.009	0.995	0.997	1.015	1.006	0.997
TV_{BIC}	1.083	1.029	1.001	1.002	1.009	1.016	1.015	1.007
NN_1	1.006	1.055	1.059	0.980	0.992	0.998	1.011	1.001
NN_2	1.562	1.061	1.146	1.114	1.108	1.111	1.118	1.175
NN_3	0.988	1.072	1.192	2.149	1.153	1.195	1.237	1.205

Table 4.d Portugal

Note : The colored cell indicates the best performance in each horizon.

Table 5: Normalized MSE of forecasting CPI growth for each country

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
ARC_4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	1.022	1.036	1.043	1.048	1.048	1.068	1.077	1.093
ARC_{AIC}	1.068	1.268	1.217	1.177	1.030	1.195	1.194	1.165
ARC_{BIC}	1.150	1.287	1.245	1.159	1.092	1.211	1.166	1.095
ART_{AIC}	1.076	1.281	1.224	1.179	1.020	1.194	1.186	1.177
ART_{BIC}	1.141	1.280	1.229	1.176	1.095	1.198	1.163	1.122
RW	1.470	1.475	1.534	1.046	1.385	1.403	1.445	1.062
TV_3	1.193	1.164	1.202	1.144	1.080	1.073	1.049	1.067
TV_{AIC}	1.206	1.195	1.142	1.121	1.096	1.116	1.156	1.150
TV_{BIC}	1.202	1.209	1.173	1.141	1.107	1.139	1.162	1.165
NN_1	1.192	1.208	1.212	1.131	1.123	1.138	1.133	1.775
NN_2	1.201	1.167	1.178	1.091	1.110	1.084	1.073	1.124
NN_3	1.202	1.162	1.176	1.113	1.124	1.089	1.070	1.051

Table 5.a France

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
ARC_4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	0.978	0.950	0.922	0.933	0.875	0.849	0.814	0.822
ARC_{AIC}	1.039	3.586	1.270	3.466	0.916	2.859	1.010	2.695
ARC_{BIC}	1.038	3.576	1.271	3.459	0.919	2.853	1.006	2.703
ART_{AIC}	0.995	3.564	1.221	3.442	0.839	2.838	0.944	2.668
ART_{BIC}	0.987	3.556	1.218	3.432	0.848	2.830	0.946	2.677
RW	3.740	1.351	3.571	0.962	3.028	1.123	2.758	0.824
TV_3	1.828	1.774	2.227	2.252	2.273	2.222	2.243	2.290
TV_{AIC}	2.693	1.495	2.058	1.295	1.956	1.267	1.910	1.416
TV_{BIC}	3.061	1.486	2.288	1.301	2.106	1.262	2.021	1.435
NN_1	1.791	1.861	2.303	2.320	2.310	2.265	2.311	2.315
NN_2	1.798	1.826	2.262	2.345	2.334	2.499	2.350	2.293
NN_3	1.811	1.733	2.022	2.254	2.490	3.544	17.272	12.863

Table 5.b Greece

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
ARC_4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	1.041	1.045	1.043	1.047	1.056	1.062	1.068	1.070
ARC_{AIC}	1.011	1.512	1.696	1.555	0.965	1.309	1.477	1.387
ARC_{BIC}	0.997	1.513	1.696	1.556	0.961	1.306	1.482	1.376
ART_{AIC}	1.050	1.537	1.717	1.581	1.005	1.340	1.504	1.417
ART_{BIC}	1.035	1.538	1.717	1.584	1.001	1.337	1.510	1.407
RW	1.767	2.087	1.828	1.034	1.503	1.761	1.647	1.028
TV_3	1.387	1.402	1.250	1.250	1.147	1.136	1.124	1.123
TV_{AIC}	1.479	1.514	1.326	1.163	1.156	1.148	1.155	1.094
TV_{BIC}	1.445	1.505	1.330	1.157	1.149	1.148	1.158	1.096
NN_1	1.400	1.384	1.273	1.260	1.144	1.132	1.141	1.124
NN_2	1.417	1.410	1.250	1.230	1.142	1.141	1.125	1.118
NN_3	1.420	1.410	1.248	1.231	1.142	1.139	1.125	1.117

Table 5.c Netherland

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
ARC_4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	1.080	1.062	0.977	0.921	0.864	0.823	0.781	0.733
ARC_{AIC}	1.086	1.807	1.178	1.562	0.872	1.278	0.917	1.142
ARC_{BIC}	1.139	1.850	1.228	1.593	0.912	1.300	0.945	1.147
ART_{AIC}	1.027	1.772	1.096	1.507	0.786	1.222	0.836	1.077
ART_{BIC}	1.035	1.766	1.106	1.502	0.789	1.213	0.834	1.059
RW	2.083	1.242	1.836	0.745	1.470	0.898	1.326	0.615
TV_3	1.510	1.423	1.541	1.441	1.374	1.354	1.379	1.353
TV_{AIC}	1.710	1.458	1.582	1.438	1.416	1.343	1.384	1.297
TV_{BIC}	1.797	1.513	1.717	1.549	1.515	1.414	1.463	1.359
NN_1	1.515	1.444	1.604	1.403	1.327	1.310	1.341	1.328
NN_2	1.490	1.363	1.476	1.356	1.310	1.293	1.318	1.287
NN_3	1.486	1.351	1.449	1.334	1.288	1.275	1.313	1.288

Table 5.d Portugal

Note : The colored cell indicates the best performance in each horizon.

5.2 Single-horizon forecast comparison

From the previous analysis, it is hard to rank the models with only MSE since the forecasting performance for different models fluctuate in different ways. Moreover, there are circumstances that one model performs well in the earlier horizons, but the other exceeds it in the longer horizons (see ARC_4 and ART_4 in figure 8) and we do not know if these difference is significant. The Diebold-Mariano(DM) test is widely used to test for the significant equality in forecasting accuracy and it works pire-wisely. We conduct the test between each two sets of forecasts at each horizons for all data sets. Hence we have 8 (time series) * 8 (horizons) * 13(models) *12 (the other models) test results. The original test results are too tedious to be included in this paper, hence we summarize the significant output in table 6 and 7. We calculate how many times the forecasts of a model outperform the others at each specific horizon at 5% significant level.

5.2.1 GDP growth

Table 6 illustrates the number of significant DM test results for GDP growth. From this table, we see ARC_4 and TV models give the most significant results. For the GDP growth of France, the times that ARC_4 perform significantly better than other models are leading for all horizons. ART_4 performs slightly worse than ARC_4 which is inline with the results of MSE analysis. TV models begin to perform well in the later horizons. When forecasting the GDP growth of Greece, TV_{AIC} and TV_{BIC} give better results than ARC_4 at first three horizons. In the later horizons, TV_3 and ARC_4 performs better. Note that a higher grade in the table does not imply one model outperform the other. As we do not have a number of 12 in table 6 which indicates there is no model outperform all other models at a single horizon. ARC_4 and TV_3 continue to perform well in forecasting The Netherlands and Portugal's GDP growth. In the fourth horizon of the Netherlands, TV_3 outperforms 11 other models significantly. TV_3 and ARC_4 begin to perform well after the third horizon for Portugal. To conclude, by using the DM test, we can easily know which model outperform the others at a specific horizon and by our summery table, ARC_4 and TV methods outperform the others in most of the cases.

Table 6: Diebold-Mariano test results summary of GDP growth

FRA	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
$h=1$	8	6	2	3	2	2	0	0	0	0	1	0	1
$h=2$	6	6	1	1	2	1	0	1	0	0	0	1	2
$h=3$	5	5	1	1	1	1	0	2	1	0	0	0	0
$h=4$	4	4	3	2	3	2	0	2	2	2	0	0	0
$h=5$	4	4	2	2	2	3	0	2	2	2	0	0	0
$h=6$	7	4	1	1	2	1	0	5	6	6	0	0	0
$h=7$	6	5	1	1	1	1	0	3	5	5	0	0	0
$h=8$	6	5	1	1	1	1	0	7	6	6	0	1	5
Ave.	5.8	4.9	1.5	1.5	1.8	1.5	0.0	2.8	2.8	2.6	0.1	0.3	1.0

GRC	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
$h=1$	5	1	7	7	2	2	1	0	11	11	1	0	0
$h=2$	6	1	5	4	1	0	0	0	10	10	0	0	0
$h=3$	8	3	3	3	0	0	3	5	7	7	1	0	0
$h=4$	7	1	4	4	0	0	1	10	3	3	1	1	0
$h=5$	7	2	4	4	0	0	2	9	7	7	4	0	0
$h=6$	6	1	3	3	0	0	0	6	4	4	1	1	0
$h=7$	9	2	3	2	0	0	0	8	8	8	4	0	0
$h=8$	8	1	3	3	0	0	0	8	8	8	1	1	0
Ave.	7.0	1.5	4.0	3.8	0.4	0.3	0.9	5.8	7.3	7.3	1.6	0.4	0.0

NLD	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
$h=1$	3	1	4	2	1	1	1	8	4	2	3	0	0
$h=2$	7	4	1	1	0	0	0	5	1	1	1	0	0
$h=3$	6	3	0	0	0	0	0	3	3	4	4	0	0
$h=4$	6	1	0	1	0	0	0	11	2	5	1	0	0
$h=5$	7	3	0	1	0	0	0	7	4	5	3	0	0
$h=6$	8	5	0	0	0	0	0	10	6	7	6	0	0
$h=7$	4	1	1	1	0	0	0	6	5	4	4	1	0
$h=8$	3	1	1	1	1	1	0	6	6	6	0	1	1
Ave.	5.5	2.4	0.9	0.9	0.3	0.3	0.1	7.0	3.9	4.3	2.8	0.3	0.1

PRT	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
$h=1$	2	0	1	2	0	1	0	5	0	0	1	0	0
$h=2$	5	2	0	1	0	1	0	5	2	2	0	0	0
$h=3$	6	3	1	1	0	0	0	3	6	6	1	0	0
$h=4$	5	2	0	0	0	0	0	5	5	5	5	0	0
$h=5$	4	1	1	1	0	0	0	4	6	6	4	0	0
$h=6$	3	1	2	2	0	0	0	3	5	3	2	1	0
$h=7$	3	1	2	3	1	1	0	3	3	3	3	1	0
$h=8$	3	2	3	3	2	2	0	3	3	3	2	0	0
Ave.	3.9	1.5	1.3	1.6	0.4	0.6	0.0	3.9	3.8	3.5	2.3	0.3	0.0

*Note: The original results of Diebold-Mariano test are too tedious to present as we need 4 (countries)*8 (horizons) tables to show the pairwise test results of 13 models against each other. This table displays how many times the forecast of a model is significantly better that the others at each individual horizon (at 5% significant level). The best two forecasts of each horizon are colored.*

Table 7: Diebold-Mariano test results summary of inflation growth

FRA	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
$h=1$	10	8	4	1	2	1	0	1	1	1	1	1	1
$h=2$	11	7	1	1	1	1	0	1	1	2	1	1	1
$h=3$	10	5	1	3	1	3	0	1	7	3	1	2	3
$h=4$	10	1	0	0	0	0	0	0	0	1	0	2	2
$h=5$	9	1	4	2	2	1	0	1	1	1	1	1	1
$h=6$	10	2	1	1	0	0	0	5	1	1	1	3	3
$h=7$	7	3	1	1	1	1	0	8	3	1	1	5	6
$h=8$	6	0	0	2	0	1	0	2	0	0	0	0	2
Ave.	9.1	3.4	1.5	1.4	0.9	1.0	0.0	2.4	1.8	1.3	1.2	1.9	2.4

GRC	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
$h=1$	7	7	7	7	7	7	0	3	2	1	3	3	3
$h=2$	11	11	0	0	2	2	8	6	7	7	4	4	5
$h=3$	11	11	7	7	7	7	0	1	2	1	1	1	4
$h=4$	10	10	0	0	2	2	10	5	8	8	4	4	4
$h=5$	7	7	7	7	7	7	0	2	2	2	1	0	0
$h=6$	8	9	0	0	0	0	6	2	6	6	0	0	0
$h=7$	6	10	6	6	6	6	0	2	4	3	0	0	0
$h=8$	7	7	0	0	0	0	7	0	7	7	0	0	0
Ave.	8.4	9.0	3.4	3.4	3.9	3.9	3.9	2.6	4.8	4.4	1.6	1.5	2.0

NLD	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
$h=1$	7	7	7	8	7	7	0	1	1	2	1	1	1
$h=2$	11	11	1	1	1	1	0	5	1	1	7	5	5
$h=3$	7	7	2	2	0	0	0	8	5	5	7	7	9
$h=4$	10	4	0	0	0	0	4	4	6	7	4	5	6
$h=5$	8	1	9	11	4	4	0	1	1	1	2	1	2
$h=6$	12	5	3	3	1	2	0	3	3	3	4	3	3
$h=7$	12	7	2	2	0	0	0	6	5	5	5	5	5
$h=8$	11	4	1	1	0	0	4	5	4	4	4	5	5
Ave.	9.8	5.8	3.1	3.5	1.6	1.8	1.0	4.1	3.3	3.5	4.3	4.0	4.5

PRT	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
$h=1$	8	7	8	7	7	7	0	3	2	1	3	3	3
$h=2$	11	10	0	0	0	0	6	4	4	4	0	6	6
$h=3$	9	10	8	7	7	7	0	2	2	0	0	4	4
$h=4$	10	10	0	0	0	0	11	0	0	0	0	3	4
$h=5$	7	7	8	8	8	8	0	0	0	0	0	0	1
$h=6$	5	9	0	0	0	0	6	0	0	0	0	0	2
$h=7$	6	7	7	7	6	6	0	0	0	0	0	0	0
$h=8$	4	6	0	0	0	0	7	0	0	0	0	0	0
Ave.	7.5	8.3	3.9	3.6	3.5	3.5	3.8	1.1	1.0	0.6	0.4	2.0	2.5

*Note: The original results of Diebold-Mariano test are too tedious to present as we need 4 (countries)*8 (horizons) tables to show the pairwise test results of 13 models against each other. This table displays how many times the forecast of a model is significantly better that the others at each individual horizon (at 5% significant level). The best two forecasts of each horizon are colored.*

5.2.2 Inflation growth

Table 7 illustrates the DM test results for forecasting inflation growth. Comparing to GDP, we observe more significant results in forecasting inflation and ARC_4, ART_4 are clearly the most performing ones. ARC_4 is still the most consistent model which performs better than most of the models in all horizons. However, when forecasting Greece and Portugal, ART_4 shows slightly more significant results than ARC_4. RW is likely to perform well in the fourth and eighth horizon for these two countries. The rest of the models also give good results in some specific horizons which can be observed from the colored cells.

By conducting the DM test, we can easily compare the single horizons performance between models. Nonetheless, it is still not possible to decide which model is more superior than the others for all horizon and the test results are tedious when multiple data sets with more horizons are tested. From individual horizon evaluation, we find ARC_4 and TV models provide better results among all methods in forecasting GDP growth while ARC_4 and ART_4 performs well in forecasting inflation.

5.3 Multi-horizon forecast comparison

The measurements that we adopt for multi-horizon superior predictive ability (SPA) are from Quaedvlieg (2018)[5]. In this section, we introduce the pairwise test results of both uniform multi-horizon SPA (uSPA) and average multi-horizon SPA(aSPA). Both tests are applied for every two models in all data sets. From table 8,9,11 and 12, we show the p-value results of uSPA and aSPA test for GDP and Inflation growth. The first column and first row indicate model i and model j. When the p-value is smaller than 0.05, the cell is colored, which implies model j has a uSPA/aSPA regarding model i. The number of colored cells in column j also shows how many times model j has beaten other models on uSPA/aSPA level. Hence in the last row, we give a sum of the number of colored cells in one column. The number in the last row can also be seen as the score that model j have in the uSPA and aSPA test. In table 10 and 13, we give a summary of the scores for each methods under uSPA and aSPA test

[5] The Matlab code is available at https://www.tandfonline.com/doi/full/10.1080/07350015.2019.1620074?.

respectively.

5.3.1 GDP growth

First, we discuss the uSPA test results in forecasting GDP growth and the corresponding p-values are demonstrated in table 8. In general, there is not a model which has uSPA to all other forecasts. As for France, the best performing models are ARC_4 and ART_4 who both have uSPA to 8 other models. However, they do not have uSPA to each other, nor to TV_{AIC}. For Greece, the leading performance can be seen from ARC_4, TV_{AIC} and TV_{BIC} which have uSPA over 6 other models. TV methods continue performing well in The Netherlands which have uSPA to all other models except for ARC_4 and ART_4. As for Portugal, only ARC_4 shows a leading result of defeating 6 models on uSPA level. AR models with lags chosen by AIC and BIC and NN methods show poor results as they seldom have uSPA to other methods. Although in most of the cases, ARC_4 has uSPA to more models than ART_4, it does not have uSPA over ART_4 except for Greece. By consulting table 10.a, the top 3 performing methods in forecasting GDP growth are ARC_4, TV_{BIC} and TV_{AIC} among all models. Although ART_4 and TV_3 do not lead in scores of uSPA test, they are not easily beaten by other models, which implies they give outstanding forecasting results for at least one single horizon.

Table 8.a P-values of uSPA test on GDP growth of France

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.955	0.874	0.907	0.889	0.903	0.922	0.942	0.950	0.971	0.829	0.824	0.953
ART_4	0.319	/	0.702	0.684	0.818	0.710	0.886	0.723	0.707	0.760	0.820	0.801	0.952
ARC_{AIC}	0.004	0.006	/	0.625	0.725	0.669	0.960	0.526	0.531	0.559	0.807	0.809	0.953
ARC_{BIC}	0.001	0.007	0.282	/	0.600	0.614	0.953	0.447	0.544	0.545	0.799	0.803	0.926
ART_{AIC}	0.063	0.010	0.300	0.321	/	0.280	0.963	0.517	0.486	0.526	0.814	0.769	0.935
ART_{BIC}	0.138	0.027	0.336	0.469	0.401	/	0.961	0.487	0.528	0.549	0.841	0.808	0.929
RW	0.006	0.002	0.001	0.001	0.000	0.000	/	0.280	0.345	0.334	0.815	0.660	0.549
TV_3	0.017	0.540	0.686	0.727	0.724	0.709	0.803	/	0.458	0.515	0.830	0.851	0.872
TV_{AIC}	0.063	0.647	0.707	0.744	0.675	0.693	0.803	0.138	/	0.468	0.811	0.813	0.633
TV_{BIC}	0.016	0.611	0.651	0.692	0.650	0.659	0.757	0.200	0.231	/	0.813	0.848	0.628
NN_1	0.006	0.005	0.003	0.002	0.005	0.001	0.016	0.019	0.042	0.040	/	0.327	0.275
NN_2	0.000	0.001	0.009	0.006	0.010	0.009	0.407	0.002	0.002	0.000	0.533	/	0.280
NN_3	0.009	0.024	0.191	0.245	0.290	0.242	0.434	0.260	0.318	0.318	0.477	0.610	/
Sum	8	8	3	3	3	3	1	2	2	2	0	0	0

Table 8.b P-values of uSPA test on GDP growth of Greece

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.983	0.927	0.928	0.987	0.986	0.852	0.716	0.732	0.729	0.993	0.997	0.999
ART_4	0.018	/	0.446	0.494	0.957	0.972	0.710	0.499	0.219	0.220	0.615	0.908	0.968
ARC_{AIC}	0.328	0.590	/	0.959	0.983	0.980	0.440	0.652	0.056	0.055	0.764	0.871	0.928
ARC_{BIC}	0.320	0.612	0.103	/	0.977	0.983	0.433	0.687	0.038	0.035	0.735	0.838	0.929
ART_{AIC}	0.015	0.084	0.004	0.005	/	0.925	0.153	0.359	0.000	0.000	0.290	0.676	0.534
ART_{BIC}	0.017	0.089	0.006	0.006	0.144	/	0.135	0.355	0.000	0.000	0.273	0.693	0.496
RW	0.318	0.653	0.689	0.713	0.939	0.953	/	0.619	0.513	0.511	0.739	0.894	0.938
TV_3	0.713	0.805	0.789	0.768	0.821	0.819	0.820	/	0.728	0.785	0.870	0.990	0.993
TV_{AIC}	0.743	0.886	0.915	0.938	0.988	0.988	0.711	0.670	/	0.274	0.846	0.988	0.998
TV_{BIC}	0.752	0.901	0.926	0.918	0.987	0.983	0.707	0.666	0.332	/	0.815	0.988	0.994
NN_1	0.000	0.481	0.409	0.430	0.921	0.925	0.606	0.486	0.019	0.017	/	0.987	0.970
NN_2	0.000	0.032	0.091	0.110	0.829	0.815	0.297	0.072	0.000	0.000	0.059	/	0.864
NN_3	0.000	0.001	0.000	0.000	0.251	0.269	0.009	0.002	0.000	0.000	0.001	0.016	/
Sum	6	2	3	3	0	0	1	1	6	6	1	1	0

Table 8.c P-values of uSPA test on GDP growth of The Netherlands

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.917	0.917	0.907	0.949	0.926	0.949	0.505	0.682	0.509	0.459	0.975	0.999
ART_4	0.063	/	0.745	0.781	0.885	0.907	0.907	0.353	0.285	0.218	0.460	0.603	0.888
ARC_{AIC}	0.432	0.685	/	0.463	0.919	0.880	0.932	0.016	0.012	0.026	0.447	0.631	0.749
ARC_{BIC}	0.309	0.671	0.583	/	0.928	0.923	0.945	0.003	0.006	0.009	0.478	0.630	0.883
ART_{AIC}	0.084	0.610	0.057	0.106	/	0.202	0.835	0.020	0.032	0.031	0.587	0.652	0.630
ART_{BIC}	0.063	0.563	0.036	0.096	0.726	/	0.865	0.011	0.021	0.033	0.566	0.646	0.651
RW	0.126	0.292	0.066	0.074	0.173	0.159	/	0.024	0.006	0.025	0.281	0.412	0.400
TV_3	0.958	0.880	0.962	0.948	0.963	0.951	0.982	/	0.873	0.702	0.660	0.977	0.995
TV_{AIC}	0.613	0.830	0.880	0.849	0.941	0.940	0.983	0.393	/	0.571	0.463	0.772	0.992
TV_{BIC}	0.516	0.851	0.914	0.891	0.940	0.941	0.990	0.383	0.516	/	0.434	0.876	0.994
NN_1	0.384	0.553	0.645	0.631	0.674	0.649	0.640	0.211	0.470	0.295	/	0.613	0.634
NN_2	0.001	0.164	0.248	0.244	0.398	0.402	0.871	0.001	0.031	0.008	0.305	/	0.336
NN_3	0.000	0.000	0.002	0.001	0.139	0.137	0.408	0.000	0.000	0.000	0.169	0.123	/
Sum	2	1	2	1	0	0	0	7	7	7	0	0	0

Table 8.d P-values of uSPA test on GDP growth of Portugal

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.801	0.980	0.966	0.973	0.943	0.995	0.802	0.900	0.621	0.602	0.954	0.996
ART_4	0.343	/	0.832	0.745	0.976	0.954	0.987	0.410	0.800	0.642	0.452	0.917	0.987
ARC_{AIC}	0.006	0.401	/	0.007	0.704	0.594	0.976	0.062	0.592	0.271	0.017	0.613	0.966
ARC_{BIC}	0.030	0.674	0.923	/	0.810	0.760	0.992	0.154	0.689	0.348	0.100	0.730	0.967
ART_{AIC}	0.060	0.065	0.253	0.125	/	0.072	0.945	0.108	0.638	0.419	0.064	0.613	0.939
ART_{BIC}	0.103	0.151	0.450	0.310	0.966	/	0.980	0.196	0.726	0.531	0.098	0.686	0.943
RW	0.001	0.000	0.024	0.010	0.049	0.014	/	0.007	0.063	0.030	0.017	0.354	0.361
TV_3	0.588	0.713	0.933	0.835	0.863	0.750	0.989	/	0.853	0.695	0.544	0.764	0.991
TV_{AIC}	0.117	0.734	0.930	0.943	0.934	0.879	0.999	0.472	/	0.275	0.386	0.973	0.994
TV_{BIC}	0.045	0.548	0.898	0.817	0.873	0.772	0.997	0.469	0.660	/	0.501	0.881	0.988
NN_1	0.274	0.809	0.790	0.806	0.864	0.829	0.996	0.459	0.852	0.447	/	0.949	0.996
NN_2	0.000	0.005	0.360	0.239	0.433	0.321	0.536	0.014	0.090	0.010	0.029	/	0.571
NN_3	0.047	0.094	0.341	0.189	0.396	0.251	0.359	0.001	0.422	0.238	0.058	0.310	/
Sum	6	2	1	2	1	1	0	3	0	2	3	0	0

Table 8

Note: p-values that less than 0.05 are colored, which indicates model j has uSPA regarding model i.

The results of aSPA test are shown in table 9. Recall that aSPA test is less strict than uSPA test, which considers the average loss difference between models. In most of the cases, if model j has uSPA to model i, it also has aSPA, but not vice versa. We observe the same pattern in the aSPA results. The models which perform well under uSPA test still give good results in aSPA test. ARC_4 is the most performing model among all methods in the uSPA test for all countries. In particular, when forecasting GDP growth of France, ARC_4 has aSPA to all other methods. Time-varying methods deliver the second best performances, especially in forecasting GDP growth in Greece and The Netherlands. According to pre-test results, these two data sets do not have constant variances. Hence TV methods may have an advantage comparing to other methods as they are non-parametric estimation which allows for heteroscedasticity in error terms. Not surprisingly, we found NN methods and random walk are the least performing models for all countries. AR models with lags chosen by AIC and BIC perform slightly better than the worst models. From table 10.b which shows the summary of the scores, we see that ARC_4 has the highest score of aSPA test and the scores of other models also in line with the patterns that we discussed.

Table 9.a P-values of aSPA test on GDP growth of France

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.979	0.980	0.985	0.968	0.976	0.997	0.974	0.971	0.989	0.965	0.965	0.916
ART_4	0.012	/	0.893	0.917	0.905	0.919	0.996	0.575	0.798	0.849	0.959	0.973	0.936
ARC_{AIC}	0.016	0.085	/	0.786	0.796	0.793	1.000	0.026	0.101	0.121	0.955	0.969	0.926
ARC_{BIC}	0.014	0.082	0.203	/	0.737	0.715	1.000	0.018	0.082	0.087	0.951	0.958	0.944
ART_{AIC}	0.030	0.077	0.174	0.285	/	0.315	0.999	0.039	0.098	0.098	0.954	0.955	0.937
ART_{BIC}	0.020	0.083	0.179	0.300	0.668	/	1.000	0.030	0.106	0.106	0.951	0.963	0.935
RW	0.004	0.006	0.001	0.000	0.000	0.001	/	0.002	0.001	0.004	0.947	0.953	0.925
TV_3	0.022	0.426	0.969	0.979	0.965	0.964	0.999	/	0.841	0.875	0.960	0.977	0.928
TV_{AIC}	0.033	0.212	0.907	0.922	0.907	0.905	0.999	0.193	/	0.506	0.952	0.976	0.940
TV_{BIC}	0.011	0.150	0.900	0.932	0.905	0.903	1.000	0.142	0.489	/	0.959	0.970	0.937
NN_1	0.042	0.034	0.036	0.044	0.054	0.050	0.044	0.060	0.046	0.048	/	0.078	0.221
NN_2	0.044	0.035	0.039	0.023	0.038	0.029	0.039	0.035	0.043	0.037	0.909	/	0.782
NN_3	0.061	0.068	0.057	0.055	0.061	0.079	0.070	0.065	0.066	0.069	0.787	0.198	/
Sum	11	3	3	3	2	2	2	6	3	3	0	0	0

Table 9.b P-values of aSPA test on GDP growth of Greece

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.989	0.975	0.979	0.997	0.997	0.735	0.935	0.573	0.573	1.000	1.000	1.000
ART_4	0.011	/	0.545	0.598	0.989	0.986	0.373	0.214	0.014	0.008	0.567	0.976	1.000
ARC_{AIC}	0.028	0.432	/	0.942	0.992	0.987	0.373	0.047	0.007	0.005	0.493	0.896	0.999
ARC_{BIC}	0.020	0.429	0.053	/	0.986	0.986	0.350	0.041	0.005	0.000	0.453	0.917	0.997
ART_{AIC}	0.008	0.015	0.011	0.012	/	0.910	0.073	0.012	0.002	0.006	0.037	0.155	0.723
ART_{BIC}	0.007	0.005	0.012	0.008	0.109	/	0.052	0.006	0.003	0.001	0.039	0.108	0.734
RW	0.272	0.638	0.619	0.649	0.933	0.930	/	0.456	0.326	0.341	0.651	0.904	0.997
TV_3	0.076	0.791	0.968	0.957	0.996	0.985	0.563	/	0.154	0.130	0.960	1.000	0.999
TV_{AIC}	0.406	0.996	0.995	0.995	0.999	0.996	0.668	0.858	/	0.596	0.931	1.000	1.000
TV_{BIC}	0.428	0.996	0.995	0.996	0.997	0.999	0.689	0.851	0.394	/	0.919	1.000	1.000
NN_1	0.002	0.454	0.498	0.556	0.962	0.966	0.335	0.044	0.077	0.066	/	1.000	0.999
NN_2	0.000	0.021	0.083	0.099	0.872	0.887	0.082	0.000	0.001	0.000	0.000	/	0.998
NN_3	0.000	0.000	0.002	0.002	0.259	0.257	0.000	0.000	0.000	0.000	0.000	0.002	/
Sum	8	4	3	3	0	0	1	7	7	7	4	1	0

Table 9.c P-values of aSPA test on GDP growth of The Netherlands

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.915	0.984	0.986	0.973	0.979	0.997	0.242	0.777	0.730	0.941	0.994	0.990
ART_4	0.064	/	0.843	0.738	0.984	0.995	0.987	0.071	0.148	0.099	0.891	0.985	0.979
ARC_{AIC}	0.021	0.174	/	0.354	0.943	0.931	0.993	0.014	0.033	0.023	0.869	0.975	0.982
ARC_{BIC}	0.022	0.265	0.661	/	0.920	0.935	0.991	0.009	0.030	0.014	0.883	0.982	0.980
ART_{AIC}	0.020	0.007	0.043	0.073	/	0.169	0.931	0.017	0.035	0.018	0.753	0.839	0.949
ART_{BIC}	0.031	0.008	0.049	0.083	0.830	/	0.951	0.018	0.028	0.017	0.769	0.871	0.970
RW	0.006	0.006	0.014	0.010	0.050	0.042	/	0.007	0.012	0.006	0.543	0.299	0.854
TV_3	0.760	0.929	0.988	0.993	0.981	0.977	0.992	/	0.873	0.861	0.927	0.996	0.993
TV_{AIC}	0.237	0.849	0.972	0.977	0.967	0.962	0.987	0.107	/	0.260	0.931	0.990	0.991
TV_{BIC}	0.277	0.897	0.979	0.985	0.979	0.976	0.996	0.160	0.735	/	0.931	0.996	0.990
NN_1	0.064	0.105	0.138	0.107	0.228	0.250	0.474	0.063	0.057	0.064	/	0.344	0.751
NN_2	0.005	0.012	0.036	0.018	0.189	0.153	0.715	0.003	0.004	0.005	0.647	/	0.970
NN_3	0.016	0.016	0.017	0.019	0.043	0.025	0.150	0.012	0.011	0.018	0.249	0.034	/
Sum	7	5	5	3	1	2	0	7	7	7	0	1	0

Table 9.d P-values of aSPA test on GDP growth of Portugal

i \ j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.770	0.998	0.999	0.970	0.933	0.986	0.860	0.972	0.894	0.692	0.999	0.994
ART_4	0.247	/	0.705	0.600	0.985	0.978	0.994	0.415	0.465	0.401	0.309	0.987	0.995
ARC_{AIC}	0.001	0.280	/	0.021	0.736	0.616	0.976	0.033	0.039	0.026	0.097	0.975	0.980
ARC_{BIC}	0.002	0.412	0.978	/	0.870	0.771	0.984	0.119	0.122	0.108	0.133	0.984	0.981
ART_{AIC}	0.042	0.018	0.253	0.131	/	0.029	0.945	0.129	0.104	0.094	0.053	0.906	0.966
ART_{BIC}	0.066	0.025	0.356	0.242	0.955	/	0.967	0.188	0.163	0.162	0.115	0.936	0.978
RW	0.012	0.020	0.024	0.012	0.049	0.030	/	0.023	0.018	0.015	0.012	0.366	0.861
TV_3	0.139	0.578	0.973	0.878	0.893	0.830	0.981	/	0.726	0.604	0.455	0.988	0.990
TV_{AIC}	0.033	0.532	0.957	0.863	0.894	0.820	0.987	0.310	/	0.224	0.280	0.994	0.991
TV_{BIC}	0.091	0.581	0.965	0.886	0.898	0.837	0.989	0.413	0.781	/	0.383	0.994	0.988
NN_1	0.325	0.673	0.916	0.871	0.943	0.915	0.989	0.509	0.687	0.616	/	0.996	0.989
NN_2	0.002	0.013	0.021	0.019	0.111	0.056	0.609	0.013	0.008	0.004	0.003	/	0.833
NN_3	0.009	0.009	0.010	0.008	0.036	0.024	0.169	0.013	0.011	0.011	0.012	0.155	/
Sum	7	5	3	4	2	3	0	4	4	4	3	0	0

Table 9

Note: p-values that less than 0.05 are colored, which indicates model j has aSPA regarding model i.

Table 10.a Score of uSPA test of GDP growth

	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
FRA	8	8	3	3	3	3	1	2	2	2	0	0	0
GRE	6	2	3	3	0	0	1	1	6	6	1	1	0
NLD	2	1	2	1	0	0	0	7	7	7	0	0	0
PRT	6	2	1	2	1	1	0	3	0	2	3	0	0
Sum	22	13	9	9	4	4	2	13	15	17	4	1	0

Table 10.b Score of aSPA test of GDP growth

	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
FRA	11	3	3	3	2	2	2	6	3	3	0	0	0
GRE	8	4	3	3	0	0	1	7	7	7	4	1	0
NLD	7	5	5	3	1	2	0	7	7	7	0	1	0
PRT	7	5	3	4	2	3	0	4	4	4	3	0	0
Sum	33	17	14	13	5	7	3	24	21	21	7	2	0

Table 10

Overall, the results of the multi-horizon test are in line with what we see form MSE and DM test. However, by using aSPA and uSPA test, one can easily compare the forecast performance of each two methods. Although in most of the cases, there is no single model that has aSPA or uSPA to all other methods, hence cannot say one model outperform all other models over the whole forecasting horizons. Nonetheless, the model performance varies for different data sets.

5.3.2 Inflation growth

The uSPA test results of inflation growth are demonstrated in table 11. We observe a very different pattern of CPI forecast than GDP. TV methods do not perform well in forecasting CPI which did not beat any models on uSPA level. The best performing model, in general, is still ARC_4 while ART_4 forecasts better in Greece. Adding a linear trend in the AR model does improve the forecasts for inflation in Greece. NN_2 and NN_3 have uSPA to TV_3, TV_{BIC} and NN_1 in forecasting Portugal inflation. The rest of the methods do not show significant results. From uSPA analysis, we see that ARC_4 and ART_4 have over average results, but we cannot rank the rest of the methods as they do not have uSPA over each other. This is also in line with the MSE plots that the forecast paths of these models fluctuate over different horizons.

Last but not least, we analyze the results of aSPA test of inflation growth. The most significant change of the pattern of aSPA results comparing to the previous test is the score of ART_4 is catching up. For The Netherlands and Portugal, ART_4 has aSPA to all other methods except for ARC_4. A lot more methods show significant results in aSPA test. In forecasting inflation in France and The Netherlands ARC_4 has aSPA to all other methods, and for the other two countries, it also performs well. TV methods show some significant results in Greece and The Netherlands but perform badly in Portugal and France. Consulting table 11, confirm again with our results. ARC_4 and ART_4 are the best two models in forecasting inflation by multi-horizon test. ARC_4 has a stable performance for most data sets while ART_4 yield better forecast in some cases (i.e. PRT).

Table 11.a P-value of uSPA test of inflation growth in France

i \ j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.9332	0.999	1.000	1.000	1.000	0.998	0.996	0.999	0.996	0.999	0.999	0.999
ART_4	0.014	/	0.996	0.997	1.000	1.000	0.999	0.940	0.980	0.969	0.974	0.984	0.988
ARC_{AIC}	0.000	0.151	/	0.858	0.379	0.804	0.992	0.536	0.923	0.920	0.838	0.798	0.773
ARC_{BIC}	0.000	0.056	0.566	/	0.590	0.430	0.998	0.180	0.341	0.760	0.391	0.403	0.395
ART_{AIC}	0.008	0.337	0.299	0.887	/	0.845	0.995	0.636	0.874	0.893	0.839	0.847	0.854
ART_{BIC}	0.000	0.000	0.389	0.403	0.532	/	0.997	0.437	0.521	0.724	0.472	0.639	0.659
RW	0.000	0.142	0.735	0.734	0.656	0.764	/	0.495	0.503	0.561	0.438	0.219	0.316
TV_3	0.000	0.462	0.972	0.978	0.931	0.849	1.000	/	0.949	0.944	0.725	0.697	0.867
TV_{AIC}	0.000	0.013	0.879	0.972	0.917	0.926	0.996	0.708	/	0.572	0.613	0.507	0.564
TV_{BIC}	0.000	0.009	0.657	0.852	0.713	0.784	0.997	0.543	0.152	/	0.436	0.197	0.349
NN_1	0.000	0.001	0.345	0.389	0.412	0.425	0.638	0.087	0.099	0.179	/	0.019	0.011
NN_2	0.000	0.281	0.922	0.938	0.884	0.867	0.998	0.756	0.920	0.923	0.697	/	0.293
NN_3	0.000	0.492	0.934	0.965	0.907	0.874	0.999	0.909	0.935	0.934	0.671	0.363	/
Sum	12	4	0	0	0	0	0	0	0	0	0	1	1

Table 11.b P-value of uSPA test of inflation growth in Greece

i \ j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.241	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
ART_4	0.937	/	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996
ARC_{AIC}	0.645	0.054	/	0.398	0.068	0.032	1.000	0.999	0.999	1.000	1.000	0.998	0.995
ARC_{BIC}	0.605	0.042	0.403	/	0.105	0.070	1.000	0.999	0.999	1.000	0.999	0.998	0.993
ART_{AIC}	0.867	0.359	0.899	0.887	/	0.391	1.000	0.999	1.000	1.000	0.995	1.000	0.996
ART_{BIC}	0.804	0.219	0.828	0.872	0.182	/	1.000	0.998	1.000	1.000	1.000	1.000	0.996
RW	0.760	0.028	1.000	1.000	1.000	1.000	/	0.998	0.944	0.933	0.999	1.000	0.999
TV_3	0.000	0.001	0.999	1.000	1.000	0.999	0.999	/	0.929	0.965	0.947	0.964	0.554
TV_{AIC}	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.997	/	0.816	0.998	0.996	0.996
TV_{BIC}	0.000	0.000	1.000	1.000	1.000	1.000	0.998	0.998	0.095	/	0.999	0.998	0.997
NN_1	0.000	0.000	1.000	1.000	0.998	0.999	0.998	0.691	0.921	0.966	/	0.641	0.598
NN_2	0.000	0.000	1.000	1.000	1.000	0.999	0.999	0.181	0.902	0.963	0.577	/	0.515
NN_3	0.000	0.000	0.573	0.598	0.605	0.607	0.604	0.517	0.487	0.511	0.506	0.488	/
Sum	6	8	0	0	0	1	0	0	0	0	0	0	0

Table 11.c P-value of uSPA test of inflation growth in The Netherlands

i \ j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.991	1.000	0.644	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.999
ART_4	0.015	/	0.999	0.983	1.000	0.644	1.000	0.995	0.994	0.996	0.988	0.991	0.994
ARC_{AIC}	0.496	0.849	/	0.395	0.970	0.978	0.991	0.987	0.959	0.967	0.988	0.992	0.990
ARC_{BIC}	0.639	0.885	0.645	/	0.978	0.979	0.994	0.985	0.966	0.959	0.984	0.988	0.993
ART_{AIC}	0.051	0.653	0.020	0.016	/	0.457	0.991	0.983	0.964	0.963	0.978	0.977	0.975
ART_{BIC}	0.082	0.752	0.045	0.025	0.466	/	0.995	0.966	0.957	0.955	0.978	0.987	0.978
RW	0.010	0.228	0.974	0.972	0.982	0.972	/	0.737	0.607	0.592	0.785	0.681	0.672
TV_3	0.000	0.030	0.998	0.994	0.995	0.999	0.998	/	0.834	0.898	0.620	0.916	0.838
TV_{AIC}	0.000	0.054	0.993	0.996	0.994	0.995	1.000	0.762	/	0.416	0.806	0.620	0.664
TV_{BIC}	0.000	0.047	0.999	0.995	0.999	0.998	0.999	0.845	0.461	/	0.822	0.704	0.691
NN_1	0.000	0.025	0.997	0.991	0.992	0.998	1.000	0.322	0.653	0.656	/	0.521	0.417
NN_2	0.000	0.055	0.997	0.996	0.998	0.998	0.997	0.911	0.805	0.856	0.803	/	0.161
NN_3	0.000	0.030	0.999	0.996	0.998	0.999	0.998	0.932	0.821	0.843	0.784	0.396	/
Sum	8	4	2	2	0	0	0	0	0	0	0	0	0

Table 11.d P-value of uSPA test of inflation growth in Portugal

j i	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.698	0.989	0.998	0.978	0.969	0.969	0.997	0.998	0.998	0.997	0.989	0.993
ART_4	0.934	/	0.982	0.993	0.931	0.955	0.991	0.988	0.995	0.996	0.990	0.982	0.984
ARC_{AIC}	0.966	0.167	/	0.981	0.186	0.207	0.962	0.988	0.982	0.995	0.992	0.981	0.983
ARC_{BIC}	0.937	0.070	0.154	/	0.072	0.101	0.974	0.993	0.976	0.985	0.997	0.980	0.986
ART_{AIC}	0.883	0.950	0.863	0.905	/	0.196	0.970	0.991	0.990	0.992	0.993	0.984	0.986
ART_{BIC}	0.905	0.960	0.936	0.916	0.485	/	0.971	0.992	0.988	0.999	0.993	0.983	0.980
RW	0.957	0.839	0.979	0.985	0.974	0.976	/	0.994	0.991	0.991	0.986	0.983	0.984
TV_3	0.000	0.000	0.914	0.969	0.882	0.878	0.965	/	0.815	0.933	0.482	0.003	0.009
TV_{AIC}	0.000	0.000	0.774	0.818	0.698	0.715	0.928	0.486	/	0.982	0.326	0.143	0.142
TV_{BIC}	0.000	0.000	0.665	0.775	0.628	0.581	0.934	0.158	0.018	/	0.072	0.034	0.023
NN_1	0.000	0.000	0.855	0.911	0.803	0.793	0.969	0.754	0.783	0.931	/	0.000	0.000
NN_2	0.000	0.001	0.958	0.983	0.908	0.931	0.982	0.994	0.876	0.971	0.883	/	0.001
NN_3	0.000	0.000	0.969	0.977	0.929	0.923	0.975	0.988	0.863	0.966	0.960	0.488	/
Sum	6	6	0	0	0	0	0	0	1	0	0	3	4

Table 11

Note: p-values that less than 0.05 are colored, which indicates model j has uSPA regarding model i.

To summarize, ARC_4 gives most stable and accurate forecasts for both GDP and inflation growth which is in line with many former literature (Stock and Watson (1998), Marcellino (2004), Marcellino (2007) etc.). Time-varying methods show good performance in forecasting GDP growth. ART_4 provides satisfying results in predict inflation growth and can outperform ARC_4 in particular data set. AR methods with lags chosen by AIC and BIC yield moderate results as they provide more inconsistent forecasts with low accuracy. NN methods are the most unstable in forecasting. In forecasting GDP, we find poorer performance as the number of units increases in the hidden layer. As for inflation, more units in NN models yield to better results. It is hard to give a convincing explanation as we only have a limited number of NN models. However, one possible reason can be the high fluctuation in inflation growth which can be seen from figure 1 and 2. From the period 2000q1 and onward, inflation growth fluctuates more than GDP growth. NN models are smart with in-sample fitting, and they are perfectly flexible. Hence they deliver better performance when dealing with more fluctuate inflation growth but tend to overfit in forecasting GDP.

Table 12.a P-value of aSPA test of inflation growth in France

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.974	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.999	1.000	1.000
ART_4	0.021	/	0.994	0.994	1.000	0.999	1.000	0.876	0.982	0.987	0.997	0.941	0.923
ARC_{AIC}	0.000	0.009	/	0.772	0.587	0.721	1.000	0.130	0.275	0.485	0.900	0.136	0.121
ARC_{BIC}	0.000	0.001	0.219	/	0.355	0.464	1.000	0.032	0.195	0.304	0.870	0.031	0.022
ART_{AIC}	0.000	0.000	0.456	0.663	/	0.819	1.000	0.184	0.303	0.445	0.869	0.208	0.178
ART_{BIC}	0.000	0.000	0.287	0.512	0.172	/	1.000	0.146	0.239	0.391	0.824	0.175	0.147
RW	0.000	0.000	0.000	0.000	0.000	0.001	/	0.001	0.000	0.007	0.259	0.001	0.000
TV_3	0.001	0.130	0.893	0.972	0.821	0.848	1.000	/	0.784	0.848	0.969	0.705	0.634
TV_{AIC}	0.001	0.014	0.715	0.825	0.704	0.760	0.999	0.213	/	0.762	0.948	0.227	0.198
TV_{BIC}	0.001	0.017	0.497	0.693	0.504	0.628	0.992	0.162	0.249	/	0.904	0.129	0.120
NN_1	0.001	0.004	0.122	0.127	0.149	0.184	0.740	0.045	0.058	0.082	/	0.044	0.036
NN_2	0.000	0.053	0.844	0.960	0.803	0.827	0.999	0.291	0.764	0.859	0.976	/	0.270
NN_3	0.000	0.083	0.861	0.977	0.818	0.867	0.998	0.341	0.768	0.886	0.951	0.707	/
Sum	12	8	1	1	1	1	0	3	1	1	0	3	3

Table 12.b P-value of aSPA test of inflation growth in Greece

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.057	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	0.944
ART_4	0.937	/	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.938
ARC_{AIC}	0.000	0.000	/	0.359	0.082	0.063	0.968	0.185	0.000	0.000	0.303	0.377	0.934
ARC_{BIC}	0.000	0.000	0.710	/	0.085	0.077	0.995	0.177	0.000	0.000	0.309	0.397	0.936
ART_{AIC}	0.000	0.000	0.929	0.896	/	0.323	0.992	0.237	0.000	0.000	0.386	0.473	0.928
ART_{BIC}	0.000	0.000	0.939	0.926	0.649	/	0.995	0.289	0.000	0.000	0.398	0.461	0.927
RW	0.000	0.000	0.021	0.007	0.011	0.003	/	0.111	0.000	0.000	0.182	0.231	0.927
TV_3	0.000	0.001	0.788	0.794	0.751	0.733	0.885	/	0.030	0.065	0.992	0.995	0.940
TV_{AIC}	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.969	/	0.998	0.981	0.992	0.945
TV_{BIC}	0.000	0.000	1.000	1.000	1.000	1.000	1.000	0.944	0.004	/	0.955	0.968	0.944
NN_1	0.000	0.000	0.674	0.698	0.628	0.580	0.816	0.002	0.015	0.043	/	0.875	0.937
NN_2	0.000	0.000	0.582	0.565	0.512	0.529	0.758	0.004	0.008	0.032	0.120	/	0.932
NN_3	0.069	0.051	0.074	0.072	0.062	0.064	0.065	0.062	0.062	0.064	0.074	0.060	/
Sum	10	10	1	1	1	1	0	2	9	7	0	0	0

Table 12.c P-value of aSPA test of inflation growth in The Netherlands

i j	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ART_4	0.003	/	0.997	0.999	1.000	0.999	1.000	0.999	1.000	0.999	0.996	0.994	0.997
ARC_{AIC}	0.000	0.000	/	0.096	0.988	0.965	1.000	0.001	0.000	0.001	0.001	0.001	0.001
ARC_{BIC}	0.000	0.000	0.903	/	0.993	0.985	1.000	0.000	0.000	0.001	0.002	0.004	0.000
ART_{AIC}	0.000	0.000	0.012	0.011	/	0.160	0.997	0.002	0.000	0.000	0.002	0.003	0.003
ART_{BIC}	0.000	0.000	0.032	0.017	0.852	/	0.999	0.001	0.000	0.000	0.002	0.003	0.001
RW	0.000	0.002	0.000	0.000	0.000	0.001	/	0.001	0.000	0.000	0.000	0.000	0.000
TV_3	0.000	0.002	0.997	0.997	0.997	0.999	1.000	/	0.872	0.849	0.722	0.849	0.831
TV_{AIC}	0.000	0.000	0.998	1.000	1.000	1.000	1.000	0.116	/	0.044	0.186	0.140	0.131
TV_{BIC}	0.000	0.001	1.000	0.997	1.000	0.999	1.000	0.179	0.950	/	0.236	0.214	0.192
NN_1	0.000	0.005	0.999	1.000	0.996	1.000	1.000	0.258	0.818	0.767	/	0.379	0.335
NN_2	0.000	0.002	0.999	0.999	1.000	0.999	1.000	0.147	0.858	0.803	0.640	/	0.348
NN_3	0.000	0.006	0.999	0.998	0.999	0.999	1.000	0.168	0.862	0.801	0.659	0.667	/
Sum	12	11	3	3	1	1	0	5	5	6	5	5	5

Table 12.d P-value of aSPA test of inflation growth in Portugal

j i	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
ARC_4	/	0.061	0.987	0.993	0.856	0.856	0.977	0.995	0.999	1.000	0.998	0.998	0.999
ART_4	0.932	/	0.991	0.991	0.997	0.999	1.000	0.999	0.997	1.000	0.996	0.996	0.996
ARC_{AIC}	0.014	0.007	/	0.984	0.146	0.128	0.846	0.971	0.980	0.986	0.972	0.958	0.939
ARC_{BIC}	0.005	0.011	0.010	/	0.072	0.059	0.724	0.965	0.979	0.994	0.953	0.932	0.901
ART_{AIC}	0.168	0.001	0.853	0.918	/	0.079	0.974	0.973	0.969	0.983	0.964	0.954	0.949
ART_{BIC}	0.154	0.003	0.875	0.943	0.927	/	0.978	0.981	0.972	0.987	0.973	0.955	0.947
RW	0.027	0.007	0.138	0.299	0.026	0.017	/	0.878	0.858	0.940	0.845	0.799	0.721
TV_3	0.003	0.000	0.026	0.031	0.023	0.018	0.133	/	0.637	0.920	0.119	0.011	0.005
TV_{AIC}	0.000	0.000	0.016	0.023	0.035	0.029	0.174	0.380	/	0.991	0.270	0.154	0.107
TV_{BIC}	0.002	0.003	0.011	0.015	0.018	0.015	0.066	0.078	0.004	/	0.058	0.037	0.028
NN_1	0.003	0.003	0.017	0.039	0.026	0.026	0.141	0.855	0.708	0.937	/	0.003	0.000
NN_2	0.000	0.009	0.059	0.077	0.049	0.035	0.246	0.993	0.873	0.968	0.998	/	0.015
NN_3	0.000	0.008	0.059	0.089	0.052	0.056	0.273	0.993	0.880	0.973	0.999	0.987	/
Sum	9	11	5	4	6	6	0	0	1	0	0	3	4

Table 12

Note: p-values that less than 0.05 are colored, which indicates model j has aSPA regarding model i.

Table 13.a Score of uSPA test of inflation growth

p-val uSPA	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
FRA	12	4	0	0	0	0	0	0	0	0	0	1	1
GRE	6	8	0	0	0	1	0	0	0	0	0	0	0
NLD	8	4	2	2	0	0	0	0	0	0	0	0	0
PRT	6	6	0	0	0	0	0	0	1	0	0	3	4
Sum	32	22	2	2	0	1	0	0	1	0	0	4	5

Table 13.b Score of aSPA test of inflation growth

p-val aSPA	ARC_4	ART_4	ARC_{AIC}	ARC_{BIC}	ART_{AIC}	ART_{BIC}	RW	TV_3	TV_{AIC}	TV_{BIC}	NN_1	NN_2	NN_3
FRA	12	8	1	1	1	1	0	3	1	1	0	3	3
GRE	10	10	1	1	1	1	0	2	9	7	0	0	0
NLD	12	11	3	3	1	1	0	5	5	6	5	5	5
PRT	9	11	5	4	6	6	0	0	1	0	0	3	4
Sum	43	40	10	9	9	9	0	10	16	14	5	11	12

Table 13

6 Discussion

In this study, we analyze 13 linear and non-linear models in forecasting GDP and inflation growth of 4 countries and try to find a benchmark. The traditional way to access forecast performance is to compare the outcomes of each horizon individual. However, only by loss evaluation or single horizon predictive ability test is not efficient when comparing a lat set of models. In our research, the pairwise tests from Quaedvlieg (2018) are applied to 13 sets of forecasts and the superior predictive ability of each two methods can be accessed easily. Loss function evaluation is necessary as it gives intuitive results of predict error. When the whole forecast path is considered, the multi-horizon test has advantages in at least two aspects: first, it disentangles the forecasts according to multi-horizon performances; second, it distinguishes the SPA from spurious results that caused by analyzing multiple horizons individually.

The pairwise test is proposed to find the existence of SPA between two sets of forecasts. In our study, we design a scoring system to calculate how many times does each model have SPA to the others. We try to rank the models in this way and find a global benchmark. However, the model which won the "Game" does not imply it has uSPA/aSPA to all other methods in all data sets. If one wants to know the more superior model of the two, he should still consult the p-value tables of the test. When there are no significant results showing one model has uSPA/aSPA to another, they do not outperform the other on the multi-horizon level.

A more convenient way to compare a large number of models, can be the Model Confidence Set (MCS) which was introduced by Hansen et al. (2003) and Hansen et al. (2011). Later on, Quaedvlieg (2018) has further developed a multi-horizon version. As we want to have a comparison of Diebold-Mariano test with its multi-horizon version(uSPA and aSPA), we did not include MCS in this study. Further researches can use MCS from Quaedvlieg (2018) as the multi-horizon measurement when comparing a large set of models.

Our research also has several limitations regarding the data set and methodology. First, we use quarterly data instead of monthly data. Data frequency can have a significant impact on forecasting performance. Second, 4 European countries are selected for this study which cannot fully present the situation of all European countries. From the plots of MSE, we observe that the same model can perform differently in different data sets. Third, we only estimate the simplest artificial neural network models (with one hidden layer) which cannot fully present this method. Due to the lack of technical knowledge, we did not choose the most appropriate NN methods. Further research can be done from these perspectives.

7 Conclusion

This paper evaluates 13 linear and non-linear models in forecasting GDP and inflation growth and aims to find a benchmark model from both individual and multi-horizon perspective. We follow the same structure of Marcellino et al. (2006) in the first part of this research which compares the chosen models by MSE. Then in the second part, we compare the DM test with uniform and average superior predictive ability test from Quaedvlieg (2018). Combining the results from all measurements, the autoregressive model with 4 lags gives the most consistent and accurate forecast among all methods for every data sets. In addition, time-varying autoregressive models perform well in forecasting GDP growth while the autoregressive model with 4 lags and a linear trend gives outstanding forecast in inflation growth. Overall, we propose the autoregressive model with 4 lags as the multi-horizon benchmark in forecasting GDP growth and inflation. Moreover, the multi-horizon superior predictive ability tests from Quaedvlieg (2018) are examined to be efficient and reliable in comparing multiple models' forecasting performance of the whole forecast path.

Finding a consistent and reliable benchmark is essential for macroeconomic researchers as it extracts the information from the past behaviour of the variable and displays the real value added by economic theory. In this sense, forecast at multiple horizons should be even valuable since the economic theories are often working for an extended period. Hence, further research is welcome in this area.

References

[1] Artis, M. and Marcellino, M. (2001), Fiscal forecasting: The track record of the imf, oecd and ec, *The Econometrics Journal* 4(1), 20–36.

[2] Box, G. E. and Jenkins, G. M. (1970), Time series analysis: Forecasting and control holden-day, *San Francisco* p. 498.

[3] Cai, Z. (2007), Trending time-varying coefficient time series models with serially correlated errors, *Journal of Econometrics* 136(1), 163–188.

[4] Chen, X. B., Gao, J., Li, D. and Silvapulle, P. (2018), Nonparametric estimation and forecasting for time-varying coefficient realized volatility models, *Journal of Business & Economic Statistics* 36(1), 88–100.

[5] Chong, Y. Y. and Hendry, D. F. (1986), Econometric evaluation of linear macro-economic models, *The Review of Economic Studies* 53(4), 671–690.

[6] Clark, T. and McCracken, M. (2013), Advances in forecast evaluation, *in* 'Handbook of economic forecasting', Vol. 2, Elsevier, pp. 1107–1201.

[7] Cook, T. and Hall, A. S. (2017), Macroeconomic indicator forecasting with deep neural networks, *Federal Reserve Bank of Kansas City, Research Working Paper* (17-11).

[8] Diebold, F. and Mariano, R. (1995), Comparing predictive accuracy. journal of business and economics statistics, v. 13.

[9] Fair, R. C. and Shiller, R. J. (1988), 'The informational content of ex ante forecasts'.

[10] Franses, P. H., Van Dijk, D. et al. (2000), *Non-linear time series models in empirical finance*, Cambridge University Press.

[11] Giacomini, R. and White, H. (2006), Tests of conditional predictive ability, *Econometrica* 74(6), 1545–1578.

[12] Hansen, P. R. (2005), A test for superior predictive ability, *Journal of Business & Economic Statistics* 23(4), 365–380.

[13] Hansen, P. R., Lunde, A. and Nason, J. M. (2003), Choosing the best volatility models: the model confidence set approach, *Oxford Bulletin of Economics and Statistics* 65, 839–861.

[14] Hansen, P. R., Lunde, A. and Nason, J. M. (2011), The model confidence set, *Econometrica* 79(2), 453–497.

[15] Hart, J. D. (1991), Kernel regression estimation with time series errors, *Journal of the Royal Statistical Society: Series B (Methodological)* 53(1), 173–187.

[16] Harvey, D., Leybourne, S. and Newbold, P. (1997), Testing the equality of prediction mean squared errors, *International Journal of forecasting* 13(2), 281–291.

[17] Inoue, A. and Kilian, L. (2008), How useful is bagging in forecasting economic time series? a case study of us consumer price inflation, *Journal of the American Statistical Association* 103(482), 511–522.

[18] Marcellino, M. (2004), Forecast pooling for european macroeconomic variables, *Oxford Bulletin of Economics and Statistics* 66(1), 91–112.

[19] Marcellino, M. (2007), A comparison of time series models for forecasting gdp growth and inflation, *Bocconi University, Italia* .

[20] Marcellino, M., Stock, J. H. and Watson, M. W. (2003), Macroeconomic forecasting in the euro area: Country specific versus area-wide information, *European Economic Review* 47(1), 1–18.

[21] Marcellino, M., Stock, J. H. and Watson, M. W. (2006), A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series, *Journal of econometrics* 135(1-2), 499–526.

[22] Medeiros, M. C., Vasconcelos, G. F., Veiga, Á. and Zilberman, E. (2019), Forecasting inflation in a data-rich environment: the benefits of machine learning methods, *Journal of Business & Economic Statistics* pp. 1–22.

[23] Nyblom, J. (1989), Testing for the constancy of parameters over time, *Journal of the American Statistical Association* 84(405), 223–230.

[24] Quaedvlieg, R. (2018), Multi-horizon forecast comparison.

[25] Robinson, P. M. (1989), Nonparametric estimation of time-varying parameters, *in* 'Statistical analysis and forecasting of economic structural change', Springer, pp. 253–264.

[26] Rünstler, G., Barhoumi, K., Benk, S., Cristadoro, R., Den Reijer, A., Jakaitiene, A., Jelonek, P., Rua, A., Ruth, K. and Van Nieuwenhuyze, C. (2009), Short-term forecasting of gdp using large datasets: a pseudo real-time forecast evaluation exercise, *Journal of forecasting* 28(7), 595–611.

[27] Stock, J. H. and Watson, M. W. (1996), Evidence on structural instability in macroeconomic time series relations, *Journal of Business & Economic Statistics* 14(1), 11–30.

[28] Stock, J. H. and Watson, M. W. (1998), A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, Technical report, National Bureau of Economic Research.

[29] Stock, J. H. and Watson, M. W. (2007), Why has us inflation become harder to forecast?, *Journal of Money, Credit and banking* 39, 3–33.

[30] White, H. (2000), A reality check for data snooping, *Econometrica* 68(5), 1097–1126.

8 Appendix

Table 14.a Correlation matrix of GDP growth for 12 countries

	AUT	BEL	DEU	DNK	ESP	FIN	**FRA**	**GRC**	ITA	**NLD**	**PRT**	SWE
AUT	1.00											
BEL	0.44	1.00										
DEU	0.45	0.39	1.00									
DNK	0.27	0.34	0.28	1.00								
ESP	0.32	0.49	0.30	0.16	1.00							
FIN	0.28	0.34	0.36	0.12	0.33	1.00						
FRA	0.47	0.60	0.47	0.32	0.47	0.43	1.00					
GRC	0.15	0.22	0.26	0.10	0.32	0.13	0.22	1.00				
ITA	0.39	0.48	0.32	0.20	0.35	0.28	0.59	0.16	1.00			
NLD	0.29	0.34	0.42	0.22	0.32	0.25	0.36	0.08	0.18	1.00		
PRT	0.26	0.43	0.33	0.28	0.53	0.28	0.52	0.27	0.36	0.29	1.00	
SWE	0.28	0.35	0.28	0.21	0.20	0.38	0.39	0.10	0.22	0.11	0.12	1.00

Table 14.b Correlation matrix of CPI growth for 12 countries

	AUT	BEL	DEU	DNK	ESP	FIN	**FRA**	**GRC**	ITA	**NLD**	**PRT**	SWE
AUT	1.00											
BEL	0.91	1.00										
DEU	0.89	0.78	1.00									
DNK	0.82	0.83	0.73	1.00								
ESP	0.76	0.79	0.66	0.90	1.00							
FIN	0.89	0.91	0.76	0.88	0.87	1.00						
FRA	0.84	0.88	0.76	0.94	0.89	0.89	1.00					
GRC	0.57	0.57	0.55	0.66	0.69	0.67	0.70	1.00				
ITA	0.81	0.86	0.72	0.92	0.92	0.89	0.95	0.76	1.00			
NLD	0.88	0.82	0.82	0.74	0.67	0.79	0.77	0.34	0.69	1.00		
PRT	0.69	0.73	0.58	0.80	0.87	0.79	0.84	0.76	0.85	0.56	1.00	
SWE	0.76	0.75	0.70	0.82	0.86	0.84	0.85	0.76	0.87	0.63	0.82	1.00

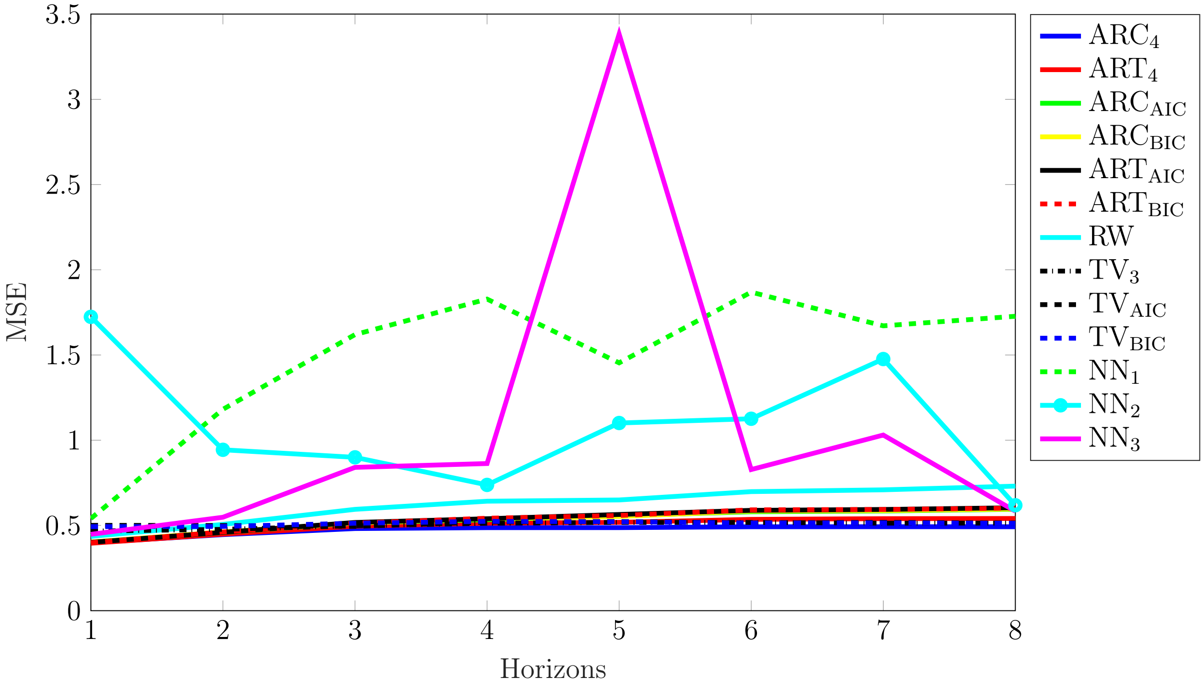

Figure 3: MSE of forecasting GDP growth in France

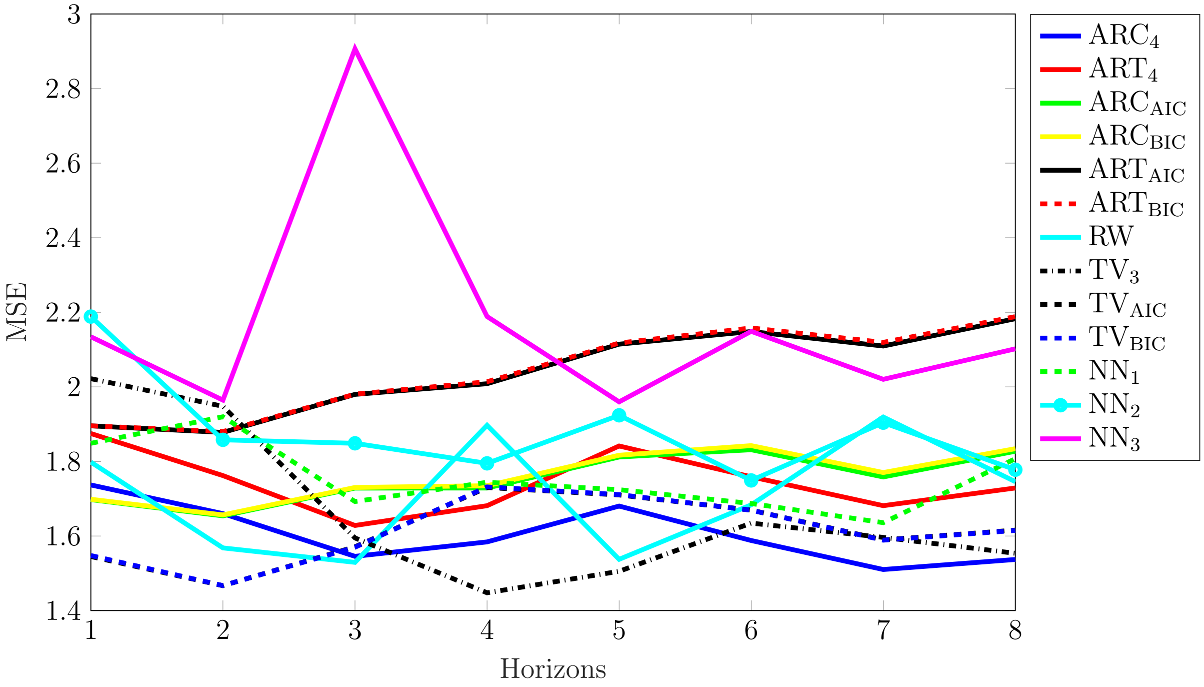

Figure 4: MSE of forecasting GDP growth in Greece

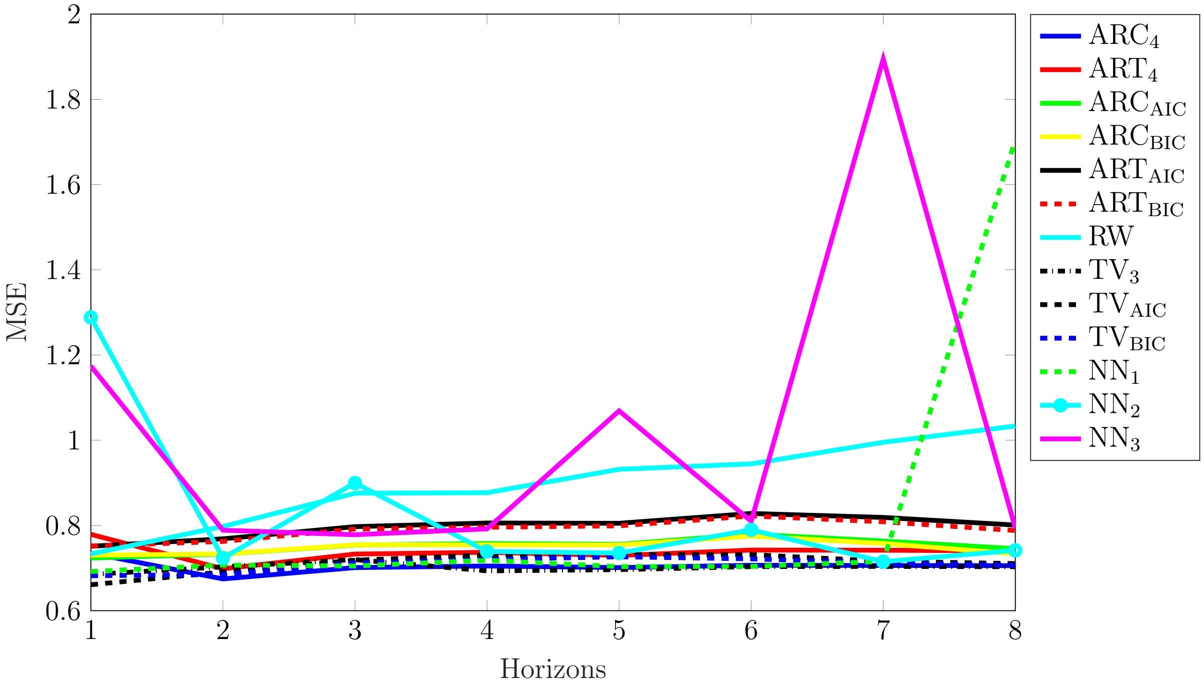

Figure 5: MSE of forecasting GDP growth in The Netherlands

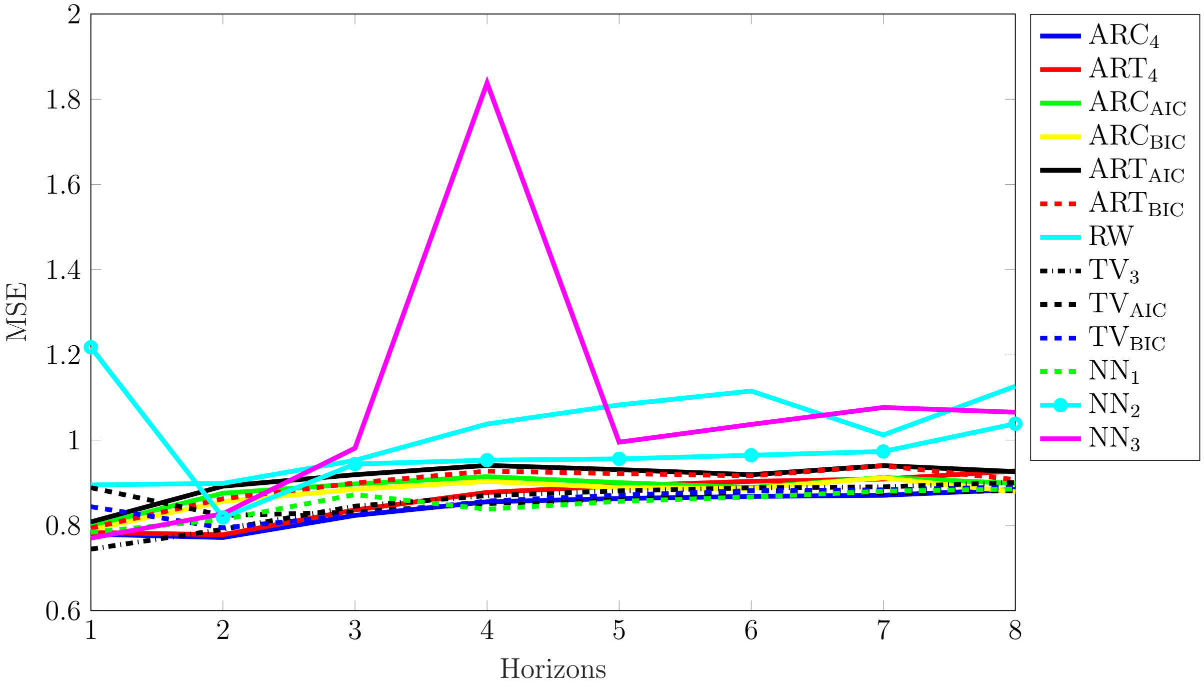

Figure 6: MSE of forecasting GDP growth in Portugal

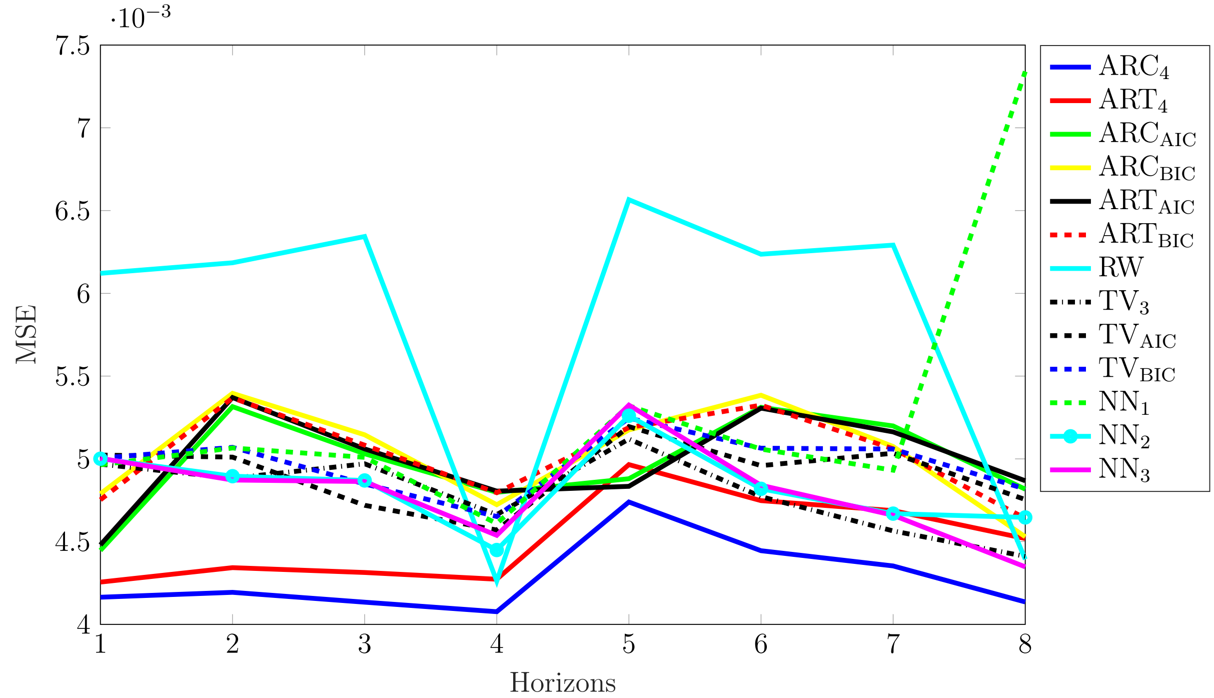

Figure 7: MSE of forecasting CPI growth in France

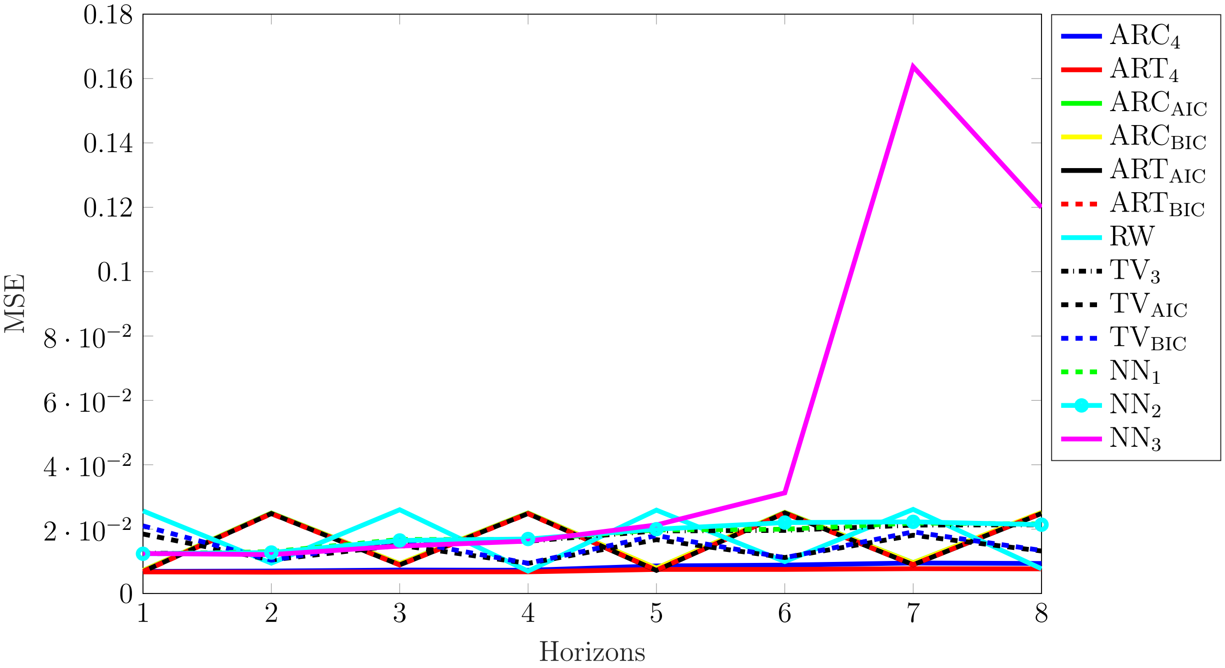

Figure 8: MSE of forecasting CPI growth in Greece

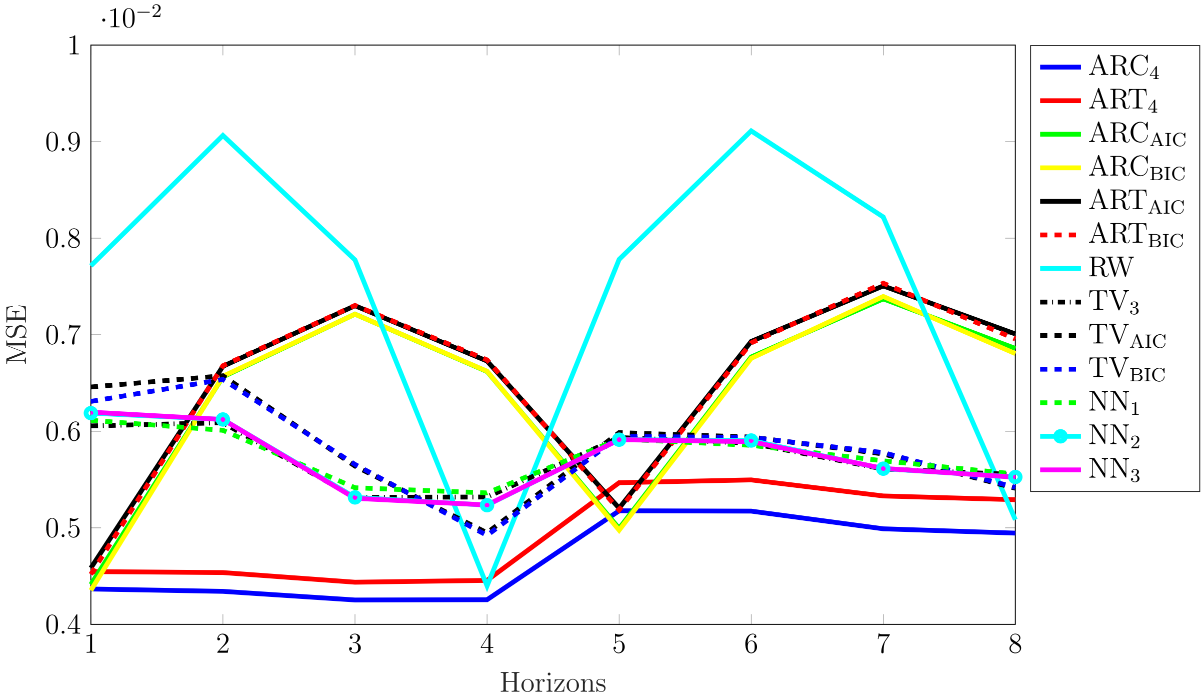

Figure 9: MSE of forecasting CPI growth in The Netherlands

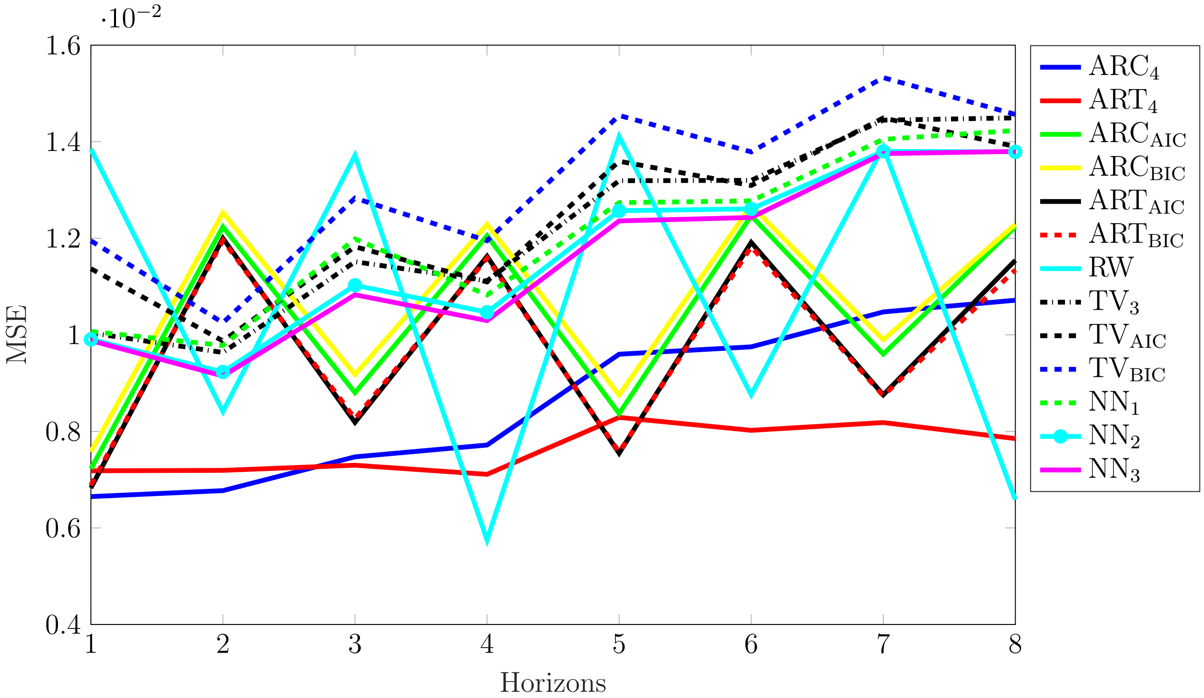

Figure 10: MSE of forecasting CPI growth in Portugal