Universidad del Rosario

Facultad de Economia

A factor model to forecast Colombian inflation using a One Covariate at a Time - Multiple Testing approach (OCMT)

David Leonardo Gonzalez Leiva

Tesis de grado: Maestria en Economia

Asesor: Jesus Otero

Abril 2022

A factor model to forecast Colombian inflation using a One Covariate at a Time - Multiple Testing approach (OCMT)*

David L. Gonzalez†
Facultad de Economía
Universidad del Rosario
Bogotá, Colombia

May, 2022

Abstract

This paper studies the construction of an inflation model and its forecast using a large set of predictors. The decision about the optimal number of common factors is made using Bai and Ng (2002) information criteria; and, a One Covariate at a Time - Multiple Testing approach (OCMT) (Chudik et al., 2018) is implemented to choose the optimal predictors and lags of the dependent variable; and, afterwards, 1 to 12 month ahead forecasts are constructed to predict the inflation in Colombia using 60 macroeconomic variables from 2006 to 2021. During this period, a rolling-window approach is used. The OCMT model consistently shows significant better performance than a Phillips curve based Vector Autoregressive (VAR) model and an Autoregressive Integrated Moving Average (ARIMA) model when using the entire dataset and, also, performs closely to these models when estimated with a pre-COVID dataset.

Keywords: Common factors, OCMT, econometric model, forecast, inflation, Colombia.

*This paper would not have been possible without the exceptional support from my supervisor, Jesús G. Otero Cardona. I am thankful for his guidance and suggestions. I also thank Gustavo A. Hernandez Diaz, at Departamento Nacional de Planeacion, for his essential help providing the data used and insights that made this paper possible. Finally, to Luis F. Melo Velandia and Andrés Garcia, for their insights on the development and construction of this paper.

†E-mail: davidl.gonzalez@urosario.edu.co (David Gonzalez; corresponding author)

1 Introduction

Forecasting macroeconomic indicators like the Gross Domestic Product (GDP) or the Consumer Price Index (CPI) of a country has been a common practice in literature and economic research. Multiple approaches have been studied to find a reliable way to predict the future behavior of these important indicators, and theory is still being developed to try it on the field and improve our forecasts over time with these advances. Predicting these indicators in a reliable way helps countries and policy making agencies to develop more accurate and impactful policies and take better decisions for the overall welfare.

The exact procedures used by central banks and statistical agencies across the world to make their estimations are often private, but we can see some of these approaches looking at works like Castro (2003), where multivariate models are being used and where most of the discussion is based on how to use all the information available (that is quite big) to make a reliable prediction. This is where this paper comes in handy. The approach used here, starting with 60 macroeconomic variables, is inspired by the recent One Covariate at a time Multiple testing procedure, developed by Chudik et al. (2018), where econometric theory is developed as an alternative to penalized regressions to select the most appropiate variables for a model in a high-dimensional dataset context. The theory and economic motivation behind this paper supports the idea that macroeconomic variables, like inflation, are not explained always by the same variables or lags of themselves; yet, the predictors for some of these variables might change over time, specially under a high-volatility context, where multivariate models and factor models might be an step forward towards a better macroeconomic forecasting.

In this paper a factor model approach is also implemented to determine the ideal number of factors, following Bai and Ng (2002); consistent results are found on this paper with the three different criteria proposed by Bay &

Ng in their paper. With this result, a OCMT procedure is applied to the data, and the selected variables, alongside the factors from the previos stage (Bai and Ng, 2002), are used to model and predict the inflation in Colombia for the next twelve periods (out-of-sample). This procedure is done 70 times, using a rolling forecast approach with each window being 120 periods long (10 years). Some very interesting dynamics with the models and their reliability are found, and for most of these windows, the OCMT model produces a lower Root Mean Forecast Error (RMFE) than the VAR and ARIMA alternative, which are the benchmark models for this paper.

In the specific Colombian case, models in a large variable dataset context have been already studied. Castro (2003) uses a multivariate model system to forecast the Colombian GDP. In his paper Castro (2003) uses a dataset with more than 50 variables. Apart from this forecasting exercise, Castro (2003) presents an exceptional literature revision over different forecasting and prediction methodologies, where he concludes that there is no universally efficient and accepted methodology, nor an specific set of variables to use as indicators to predict macroeconomic variables precisely and consistently. This last idea motivates this paper to find the dynamic in variable selection for the OCMT procedure over time.

For their part, Espinoza et al. (2012), study the implementation of financial variables and indicators in GDP and inflation projection in Europe and in the United States (US). The authors use a VAR model and test different combinations of the available variables in their dataset, including financial variables, to find the best model possible. They find that, depending on the time spawn used to build and estimate the model, different variables become significant and part of the model. This last dynamic is seen in this paper, where the COVID-19 pandemic includes more variables to the model through the OCMT procedure[1].

[1]To see the dynamic for the OCMT procedure, see the online complementary material.

Apart from econometric and computational models used to make macroeconomic variable predictions, some countries and governments implement surveys as a valid instrument to get a bigger picture of how some important sectors and analysts think these variables are going to behave. An example of this can be found in Carlos and Gabriel (2010), where the authors analyze the results of the Encuesta sobre Expectativas de los Especialistas en Economia del Sector Privado del Banco de Mexico[2], and how the respondents to this survey made by Mexico's Central Bank respond to some signals. The period analyzed is 1995-2009, where the authors found that some experts were late to adjust their predictions when a fixed event (signal) was given on the market. They explain this behavior saying that the ego of some experts prevented them to recognize sooner that they needed to adjust their predictions. On the other hand, Carlos and Gabriel (2010) recognize that a comparative advantage of surveys like this, as a forecasting method, is the experience and knowledge of different sectors inside the economy, and pointed out the importance of this in high-volatility periods, where some econometric and statistic forecasting methods fail Bvuchete et al. (2021).

As shown before, there are many different methods available for variable modeling and forecasting and, in a context where data is not particularly scarce, how it's will be the key factor to make a good estimation and forecast. Stock and Watson (2002) implement a Principal Component Analysis (PCA) approach to reduce the number of dimensions in their forecasting scenario, where they had more than 200 variables to forecast GDP and CPI for the US in a 1970-1998 time-spawn, and reduced them into a few factors (diffusion indexes - DIs). Moreover, they conclude that it is possible to capture the variations of a large dimension dataset into a few factors to model and forecast the behavior of real economic activity variables. They achieve this by conducing a two-step process: first, they perform a PCA to estimate the

[2]Spanish for Expectatives survey by Economy and Private sector Specialists from the Mexico's Central Bank.

factors (DIs); and, second, they make a forecast of their variable of interest with the estimated DIs. Finally, they test the errors of their model against a simple VAR model, following the Phillips Curve approach, Gordon (1982) and some univariate autoregressive models, where they find their DI forecasts to have a better results than the benchmark models.

With the advances made in the macroeconomic forecasting literature made by Stock & Watson (Stock and Watson (2002), Stock and Watson (1989) and Stock and Watson (1999)), the discussion turned to how to properly reduce the number of dimensions of a large dataset. In this paper we propose a factor model for Colombian inflation under a large dimension dataset scenario. In this context, we have to determine the ideal number of factors for our model, so different tests have been created to make this decision (Onatski (2010) and Bai and Ng (2007)), yet Bai and Ng (2002) went deeper into the discussion and proposed a way to determine how many factors the large dimension dataset should be reduced to[3]; where the authors propose the criteria for factor selection in a large-dimensional panel context, Bai and Ng (2002) developed econometric theory to be able to determine the optimal number of factors from the data itself and not to assume it. They did this by estimating the factors for a model establishing different criteria based on the dimensions of the dataset and, according to asymptotic econometric theory, the authors found that the criteria proposed in their paper let them estimate the optimal number of factors consistently in approximate factor models.

Studies more focused on developing different procedures to improve forecasting precision have also been an important part of inspiration for this paper. Chudik et al. (2018) propose in their paper the One Covariate at a time Multiple-Testing, henceforth OCMT, procedure for large dimensional datasets. This is a two-step procedure where: i) they test the significance of all potential predictors when explaining the variable of interest one at a

[3]This same method is performed by Chudik et al. (2018) to reduce the dimensions of this dataset before performing the OCMT procedure for variable selection.

time, after this, they select the variables that surpassed a given threshold; and ii) only if there are hidden signals[4], a second stage is needed to include these covariates in the true model. The authors compare their OCMT model against a Lasso and Adaptive Lasso variation, where they found the OCMT model to have a good performance and successfully eliminate noise variables compared to the aforementioned variations.

When comparing forecasting models, Diebold and Mariano (1995) propose a test trying to find which, betweeen two models performs better[5]. Arruda et al. (2011) incorporates the Diebold Mariano test for linear and non-linear Phillips curve based VAR models constructed to forecast brazilian inflation, finding that the performance of the non-linear VAR model was significantly better than the linear VAR model. On the other hand, in this paper, the difference between Mean Forecast Error between the OCMT model and both benchmark models result significative according to this test.

This paper is divided in the following sections: Section 2 explains the OCMT methodology, the common factors theory and how it's applied to this paper; and, the benchmark models to which the OCMT model is compared against. Section 3 presents data, forecasting methodology and some robustness considerations. Section 4 shows and explains the results of the model construction process and the forecasting exercise. And Section 5 concludes. The supplemental material, organized in four appendixes and an online folder, provides an entire list of variables and how we organized and named them and all the rolling forecast results on this paper, comlpementary results for the Diebold and Mariano (1995) Tests and an Excel file with the dynamic for the OCMT variable selection procedure.

[4]This is when a variable is part of the true model but did not pass the threshold in the first step of the procedure.

[5]This test is based in the Mean Square Error of the models.

2 Econometric theory and methodology

The methodology implemented in this paper follows the idea that modelling the bahavior of a macroeconomic variable should not be done using the same variables and lags of the dependent variable anytime a forecast is needed. This is specially true when high-volatility periods occur, when the number of optimal factors, alongside the number of variables increases. The idea behind is that the forecaster can change and manipulate the number of variables to forecast with, alongside the variables itself, with time, depending on the number of periods to forecast ahead, the data available at the time or even the time of the year. On the other hand, This paper also bases on the idea that macroeconomic variability can be modeled by a few indicators, or *Diffusion Indexes* as named in Stock and Watson (2002). In the following subsections Bai and Ng (2002) criteria is explained and used, alongside the OCMT (Chudik et al., 2018) procedure which is an useful econometric tool to test our theory against two benchmark models that are explained in the following section.

2.1 Common factors

As exposed in the introduction, Bai and Ng (2002) propose some criteria to determine the optimal number of factors in a factor model, and Chudik et al. (2018) also implement these criteria into theirs. To understand how to determine the number of factors, it's necessary to understand the dimensions of the dataset used on this paper; at the end, it's the number of observations and variables employed that determine which of the three criteria proposed by Bai and Ng (2002) to use. After developing some econometric theory[6], Bai and Ng (2002) propose three different criteria:

[6]Profoundly explained in corollary 1 of Bai and Ng (2002).

$$IC_{p1}(k) = ln\big(V(f, \widehat{F^k})\big) + k\left(\frac{N+T}{NT}\right) ln\left(\frac{NT}{N+T}\right),$$

$$IC_{p2}(k) = ln\big(V(f, \widehat{F^k})\big) + k\left(\frac{N+T}{NT}\right) lnC_{NT}^2,$$

$$IC_{p3}(k) = ln\big(V(f, \widehat{F^k})\big) + k\left(\frac{lnC_{NT}^2}{C_{NT}^2}\right).$$

The process to determine the number of factors selected is the following:

$$MIN_r \; IC_{pi}(r) \; for \; i \in 1, 2, 3, \tag{1}$$

where N is the number of covariates or variables in the dataset, and T is the number of observations available. All three criteria are asymptotically equivalent, yet they give different results under a finite sample context. Based on Bai and Ng (2002), $IC_{p1}(k)$ is the standard penalty factor in most time series applications; $IC_{p2}(k)$ should only be used when $N << T$; and $IC_{p3}(k)$ performs well when N is considerably big relative to T. Once the number of factors is determined, they are included into the model and estimated using Principal Component Analysis (PCA) (Serge et al., 2019).

Taking into consideration that our dataset contains $N << T$, ($N = 60$ & $T = 190$), then, following Bai and Ng (2002) the $IC_{p2}(k)$ is the criteria used in this paper to estimate the number of factors for the model. Finally, when this criteria is applied to our dataset in the results section, as shown in Table 5 in Appendix B, an average of 4.6 factors were selected for each window.

2.2 OCMT

Once the ideal number of factors is determined from the previous stage, we can implement the OCMT procedure, which consists of the following two stages:

1. First stage: t-tests are performed for a regression for each covariate in the dataset against the interest variable. If the value of this t-test is greater than a pre-determined threshold or critic value, then this variable will be selected to enter the model. If don't, only under an exemption explained in the second stage that covariate will enter the model, otherwise it will not.

2. Second stage: if needed, this stage tests the so called *pseudo-signals*, which are variables that were not selected in the first stage, but belong to the true model. In order to do this, a determined increase in the critical value is made and the tests are performed again.

Chudik et al. (2018) start from a data generating process with the following form:

$$Y_t = \alpha' Z_t + \sum_{i=1}^{k} \beta_i x_{it} + \mu_i, \tag{2}$$

where Y_t is the variable of interest, and Z_t is a preselected set of variables and X_{it} is the vector representing the unknown signal variables. There are k signals, $k*$ pseudo-signals and $n - k - k*$ noise variables that we seek to identify[7].

The OCMT procedure starts by calculating the statistical significance of each variable when explaining the variations in Y_t, by performing an Ordinary Least Squares (OLS) regression as shown in equation 2. Then, a t-test is calculated for each of the regressions, and all the variables surpassing a given threshold are included in the model. The threshold is given by the following critical function Chudik et al. (2018):

$$c_p(n, \delta) = \Phi^{-1}\left(1 - \frac{p}{2f(n, \delta)}\right), \tag{3}$$

[7]The factors resulting from the Bai and Ng (2002) criteria implementation, are included in the model as the vector Z_t.

where Φ^{-1} represents the inverse function of a normal distribution. Also, δ denotes the critical value exponent and $0 < p < 1$, which represents the individual value of the t-ratio for each variable.

The second stage of the procedure follows the same statistical process with all the variables that were not selected in stage 1, and variables also have to surpass a given critical value in order to be selected. The critical function in stage 2 has the following form:

$$c_p(n, \delta*) = \Phi^{-1}\left(1 - \frac{p}{2f(n, \delta*)}\right), \tag{4}$$

where equation 3 and 4 are fairly similar, the only difference is $\delta*$, where $\delta < \delta*$.

The dynamic of this procedure for our dataset is shown in Table 5 in Appendix B, where an average of 2.8 variables were selected for each window of the rolling forecast and the model estimation[8].

2.3 Benchmark models

The OCMT model resulting from the last subsection's procedures is compared with two different models; a Phillips Curve-based VAR model, including inflation and unemployment rate; and an ARIMA model[9]. The VAR model used has the following form:

$$\begin{bmatrix} infl_t \\ zx3_t \end{bmatrix} = \begin{bmatrix} c_{infl} \\ c_{zx3} \end{bmatrix} + \begin{bmatrix} \alpha_{infl,1} & \alpha_{zx3,2} \\ \alpha_{infl,1} & \alpha_{zx3,2} \end{bmatrix} \begin{bmatrix} Infl_{t-1} \\ zx3_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{infl} \\ \epsilon_{zx3} \end{bmatrix} \tag{5}$$

The ARIMA model is estimated during each iteration of the rolling window forecast following the *arimaauto* command for the Stata software; yet,

[8]Without taking into account the selected factors following Bai and Ng (2002).

[9]These two models can be consistently found in literature as benchmark models when forecasting a macroeconomic time series with a different model. For further information, see Chudik et al. (2018) and Stock and Watson (2002).

for most of the rolling windows, the model recommended by the software results in an ARIMA(1,0,0). In consequence, the model used for simplicity has the following form:

$$infl_t - \alpha_1 infl_{t-1} - ... - \alpha_p infl_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q},$$

where $p = 1$ and $q = 0$. Ending up with the following model:

$$infl_t = c_1 + \alpha_1 infl_{t-1} + \epsilon_t \tag{6}$$

The benchmark models, to which the OCMT is compared agains to, are represented by equations 5 and 6.

3 Data and forecasting

In this section, I describe how the data used in this paper were manipulated in order to make the variables stationary, alongside the processes used to select them. Finally, the resulting model from the OCMT procedure was described, as well as the benchmark models used to compare its performance.

The dataset used in this paper contains 60 macroeconomic time series and indicators of the Colombian economy. All variables have a monthly periodicity and were obtained from different public entities and statistical departments. The complete dataset spans 2006:2 to 2021:11. Appendix A successfully lists all the variables used, their given name in the dataset, source, description and the code summarizing the transformations made to each variable following the nomenclature used by Stock and Watson (2002).

To use the time series, it was necessary that all of these variables were stationary and "well behaved"; to achieve this, each time series was submitted to unit root tests and, depending on the Auto correlation Function (ACF) and Partial Autocorelation Function (PACF), following Box (2008), differences were taken (only first, second and annual). Also, following Stock and Watson

(2002), logarithms were taken for all the non-negative variables that were not already in rates or percentage units. All variables were standardized. The following table summarizes this codification procedure for our dataset:

1. No transformation.
2. First differrence.
3. Annual difference.
4. Logarithm.
5. First difference of logarithms.
6. Annual difference of logarithms.
7. Annual difference of first difference.
8. Annual difference of first difference of logarithms.

A full description of each variable, its source and Stock & Watson's classification code is available in Table 5 in Appendix A.

3.1 Data considerations

An important consideration for the forecast is that, in order to make a prediction for H periods ahead (out-of-sample forecasting), observed values for the predictors are needed, this, for the OCMT model to be able to perform a dynamic forecast/prediction. Therefore, values for the predictors for these H periods ahead enter the database in three different styles, as a naive forecast, as a mean forecast and as an Autoregressive forecast. In the naive variation, the missing observations are replaced with the value of the last observation available for each predictor in the dataset and the difference is calculated between the dynamic prediction for the OCMT model, and later squared up to obtain the forecast error (FE). On the other hand, the mean variation

calculates the mean for the variable in the entire window and then compares the dynamic prediction against it to obtain the forecast error for the OCMT model. Following this procedure to calculate the FE, in the AR(p) forecast, the missing observations are assumed to follow an autoregressive behavior following each window.

Note that this is needed only in the last 11 observations, where 12 months ahead forecasts escape from our dataset. This point is reached in the 59th window of the rolling forecast and estimation, where one missing value is created, ending up with 11 in-dataset and 1 out-of-dataset observations to which the results of the dynamic prediction should be compared. In this order of ideas, in the 60th window, there will be 10 in-dataset and 2 out-of-dataset observations (missings), and so on for the remaining 10 windows, where, in the last window, there will be 12 out-of-dataset observations with missings that need to be generated (either via naive, mean or AR(p) procedures), in order to be able to perform the dynamic prediction and obtain the forecast errors.

The ARIMA and VAR models do not suffer from this missing-data problem, because both are closed systems, so no missings are created and, therefore, a dynamic prediction and forecast can be performed.

3.2 Forecast

The method used for the forecast and model performance testing was a rolling forecast, where there is a fixed length window in all the procedures applied to the data, from the estimation to the forecasting for each of the three models tested in this paper. As explained before, the window length is equal to 120 periods (months), spanning 10 years of data for each model to be estimated.

Once the data for all the periods to forecast is available, as explained in the last section, forecasts for 1 to 12 months ahead are constructed. As there were 70 different rolling windows to which each of the models was estimated,

and a different forecast was constructed for each model, the results for this process, as for the factor selection and OCMT model construction, are shown in the following section.

3.3 Robustness considerations

Three different approaches to handle the missing observations in every window of the rolling exercise are used[10]. First, a *naive* behavior for the variables is assumed where, after the last available observation of each window, the last available value is assumed for all the missing observations. Then, a difference between these values and the forecast of each of the three models is calculated and squared to find the forecast error for each of the models. Second, a *Mean* approach is implemented for the missing observations where, from the last available observation, the mean of the variable during each window replaces the missing values and, therefore, the forecast error is calculated with the difference between the forecast of each model and his mean value. Finally and, following the same pattern, a *AR(1)* behavior is assumed, where the values for the missings are generated from an AR(1) process from the variable calculated during each window.

On the other hand, to measure the significance of the *Root Mean Forecast Error (RMFE)* differences between the models, a Diebold-Mariano (1995) test is implemented between each of the three models, following Diebold and Mariano (1995). This test is used to determine whether the forecasts made by the OCMT model and the two benchmark models are significantly different. The sample for which the DM Test was performed is the full sample of the dataset, finding that the OCMT model outperforms significantly, both the ARIMA and VAR models. All of the three Diebold Mariano tests performed in this paper were performed assuming the MSE criteria and an

[10]The missing observations are the out of sample observations, where an assumption for each variable's behavior in these periods have to be made and implemented to test the forecast error against them.

uniform kernel, following Diebold and Mariano (1995). Appendix D shows the complementary tables for the Mean and ARIMA apporaches for missing observations.

4 Results

Before estimating and constructing the OCMT model for this paper, a Bai and Ng (2002) criteria was implemented to determine the optimal number of factors for the model. For the most part of the rolling procedure, five factors were selected, which will enter the OCMT model as pre-selected variables that belong to the true model and as a specification for the inflation variable. Something remarkable is that the number of factors increase from four to five twice; but, the second time it happens, it remains in five factors for the rest of the rolling exercise. Despite the enormous COVID-19 effects in overall world economy, the number of factors did not increase when the pandemic information entered the active dataset.

Following the Bai and Ng (2002) criteria for optimal factor selection, the OCMT procedure was performed with the factors generated and estimated, they entered as a pre-selected variable set that explains inflation alongside the variables selected by the OCMT econometrical approach. Table 5 in Appendix B summarizes these results, where an average of 2.8 variables were selected in each window of the rolling exercise, and an average of 4.6 factors were also selected; accounting for a total of 7.5 predictors (average) per window to explain inflation.

The variables selected in the OCMT procedure are summarized in Table 4, where the first lag of the inflation is always selected as a regressor for the model, and the annual lag of the inflation, alongside the commercial expectatives for next semester, are selected for more than 50 % of the times to explain inflation's behavior. Finally, Producer Price Index (PPI) and CPI enter as predictors for 8.5 % and 2.8 % of the windows, respectively.

Variable name	Times selected / 70
1st lag dependent variable	1
12th lag dependent variable	0.7
Expectations about the economic situation for next semester (commercial)	0.543
2nd lag dependent variable	0.486
PPI (total).	0.086
CPI.	0.029

Table 1: Summary of variable selection in the OCMT procedure.

Table 2 shows us the average root mean forecast error for each of the models in the full-sample analysis (2006-2021) and the pre-COVID analysis (2006-2019), where we can see that the OCMT model has a better performance than both the VAR and ARIMA models in the full sample analysis; yet, in the pre-COVID analysis, the VAR model outperforms, marginally, the OCMT and ARIMA models.

Evaluation sample:	Full sample: 2006m2-2021m11	Pre-COVID: 2006m2-2019m11
Model:	RMFE	
OCMT	0.038	0.045
ARIMA	0.046	0.044
VAR	0.046	0.044

Table 2: RMFE comparison between models and samples.

To see this dynamic in 1 to 12 months ahead out-of-sample forecasts, figure 1 shows how the OCMT performs better when high-volatility periods take place compared to both benchmark models. It also shows that the RMFE for both ARIMA and VAR models are very similar. In order to determine if these differences in RMFE between models are significant, a Diebold-Mariano (1995) test was implemente between the OCMT and ARIMA model and, also, other DM Test for the OCMT and the VAR model was performed. As Table 3 shows, for the naive approach[11], the OCMT model has not only a lower MSE than both benchmark models but, with a p-value equal to 0.011 and

[11]Explained in more detail at the end of Section 3.

0.0130 for the OCMT-ARIMA and OCMT-VAR comparison, respectively, it rejects the null hypothesis of equal forecasting power between the models; therefore, the results and the difference in RMFE is significant between both OCMT and ARIMA model and also between the OCMT and VAR model.

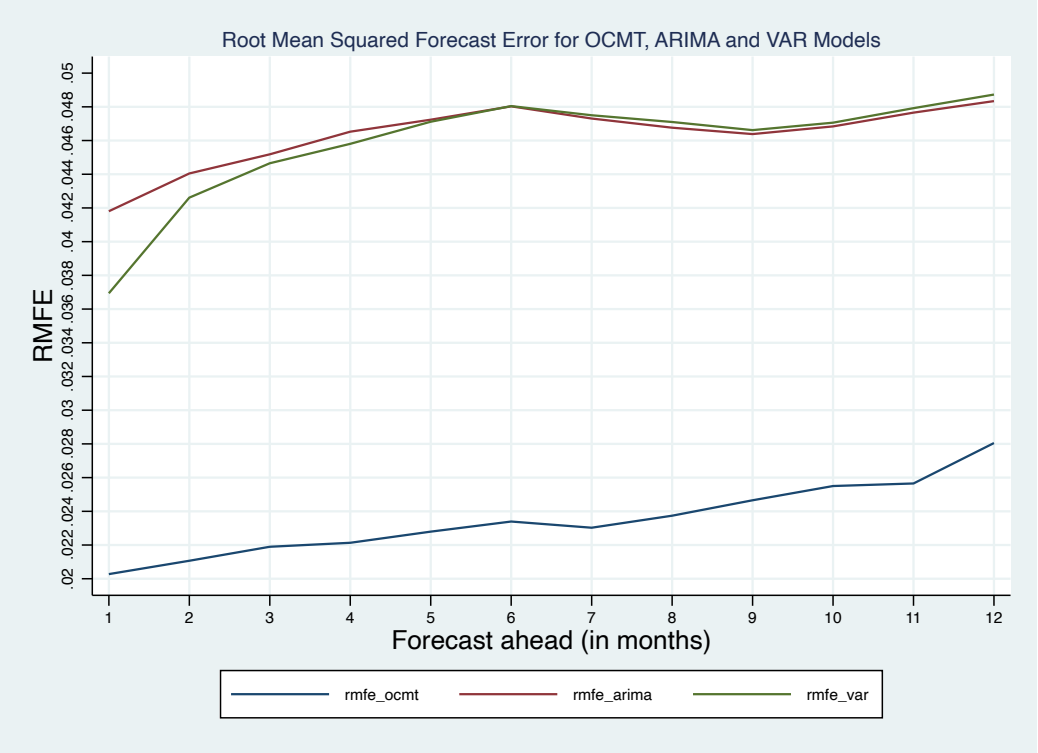

Figure 1: RMFE comparison for 1 to 12 months ahead out-of-sample forecasts (naive approach).

When a mean approach for the missing data is implemented, RMFE's distribution over time is shown in figure 2 in Appendix C, where the main difference, compared to figure 1, is that, in the one-month ahead forecast, a big drop in the RMFE can be seen. This drop is explained by the mean of the inflation throughout the entire window, which is fairly different from the last observable value (which was implemented in the naive approach). In this approach, the difference between the mean and the last observation explains the big drop in the RMFE medition. Nevertheless, this affects mainly the

OCMT model (and, in less magnitufe, the VAR model), this is due to the increased ammount of regressors used by this model, where an average of 7.5 variables were selected (including different variables, inflation lags and factors) to model inflation's behavior. The VAR model uses unemployment rate alongside inflation, nevertheless the dimension of the model is considerably smaller. As shown in the figure 2, both ARIMA and VAR models converge to the same RMFE in time, and OCMT's RMFE is significantly smaller than both benchmark models RMFE. The DM Test performed show a p-value of 0.017 and 0.008 when compared against the ARIMA and VAR respectively, rejecting that the OCMT and benchmark models are equally powerful when forecasting inflation (see Table 4.).

Finally, when an AR(1) behavior is modeled for all variables to measure the models against it, as shown in figure 3 in Appendix C, and when forecasting five to six periods ahead, there is and inflection point for all the model RMFE's behavior. For more than 6 months-ahead forecasts, both the ARIMA and VAR models reduce it's RMFE, yet the OCMT's RMFE increases, being consistent with the hypothesis that, as an AR(1) behavior was assumed for all variables after the estimation period, then the ARIMA and VAR models are closer to the Data Generating Process (DGP) in the long run. The DM Test results for this approach also supported the OCMT model as the most appropiate for forecasting inflation, with a p-value of 0.023 when compared to the ARIMA model, and 0.013 when compared to the VAR model[12].

A Diebold-Mariano (1995) test was implemented between the ARIMA and VAR Benchmark models; resulting in 0.078 and 0.079 MSE respectively, and a p-value of 0.312, not being able to reject the null hypothesis, thus, neither of the benchmark models is superior between them. Also, a complementary analysis was implemented for all twelve horizons, testing the dynamic

[12]This result might change when the forecasting horizon is increased considerably.

Naive	Full sample: 2006m2-2021m11		
Diebold-Mariano test	OCMT	ARIMA	VAR
MSE	0.064	0.07902	0.07904
P-Value: OCMT against	-	0.0118	0.0130

Table 3: Diebold-Mariano (1995) Tests: Naive approach.

of the Diebold-Mariano (1995) test when the forecast horizon for the models is changed. Table 6 in Appendix E shows this dynamic and reflects that the OCMT procedure produces the model with the lowest MSE out of the three models analyzed[13]. Also, the p-value dynamic of these tests show that with a confidence level of 90%, the OCMT procedure produces the model with best forecast accuracy in all twelve forecast horizons, with the ARIMA and VAR model losing forecasting power when the forecasting horizon increases.

5 Concluding remarks

This paper investigates the implementation of a OCMT variable selection procedure to model Colombian inflation. Under a high-volatility context (as it is the Colombian case), the implementation of the information criteria proposed by Bai and Ng (2002) helps this model to perform better when a hidden signal is implicitly given in the data. In other words, the OCMT variable selection algorithm performs better than the benchmark models when latent signals are recognized in the data by the criteria. The number of factors used to model inflation where somewhere in between four and five; considerably more than what Stock and Watson (2002) found in their implementation of Bai and Ng (2002) criteria for the US case, where only a single factor was selected alongside some lags of the inflation, to explain it's behavior from more than 200 variables dataset. Nevertheless, different from Stock and Watson

[13]This analysis was implemented only for the naive approach for missing data.

(2002), this paper complements the factor model with another set of predictors obtained by the OCMT procedure (Chudik et al., 2018), finding this model and it's forecast a significantly better approach than a Philips curve-based VAR model and an ARIMA model when high volatility period takes place, based on Diebold and Mariano (1995).

Finally, OCMT variable selection process really puts into perspective the relevance of expectations when macroeconomic modelling, being commercial expectations the third most selected variable in this procedure (54 % of the times), which represents a good indicator and predictor of overall inflation movements. This last conclusion remarks the importance of talking about different argent's perspectives of the economy's future, as shown in Carlos and Gabriel (2010).

This paper aims to stimulate further development of forecasting techniques and it's implementation. For instance, a recursive window may be implemented to study the dynamic of variables and to construct a more powerful model to forecast with. Arruda et al. (2011) study the brazilian case and compares two different models, a linear Phillips curved based VAR model and a non-linear VAR model with a threshold; therefore, an interesting road from this study might involve studying different approaches to implement and compare the OCMT procedure and an alternative outside the benchmark models studied here. Further research may include non-traditional variables inclusion, despite focusing only on official variables and information, non-traditional information (e.g.,Google Trends) could make the model more accurate.

Declarations

Availability of data and material: The data used in the paper will be made available for replication purposes. To get access to it please e-mail the author of this paper.

Code availability: Stata ado-files will be made available for replication purposes.

References

Arruda, E. F., R. T. Ferreira, and I. Castelar (2011). Modelos lineares e não lineares da curva de phillips para previsão da taxa de inflação no brasil. *Revista Brasileira de Economia - RBE* (3).

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191 – 221.

Bai, J. and S. Ng (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics 25*(1), 52–60.

Box, G. E. P. (2008). *Time series analysis: forecasting and control.* Wiley series in probability and statistics. John Wiley.

Bvuchete, M., S. S. Grobbelaar, and J. van Eeden (2021). A network maturity mapping tool for demand-driven supply chain management: A case for the public healthcare sector. *Sustainability (2071-1050) 13*(21), 11988.

Carlos, C. and L.-M. Gabriel (2010). Las expectativas macroeconÓmicas de los especialistas: Una evalución de pronósticos de corto plazo en méxico. *El Trimestre Económico 77*(306(2)), 275 – 312.

Castro, C. (2003). Sistema de modelos multivariados para la proyecciÃ³n del producto interno bruto. *Archivos de EconomÃa*.

Chudik, A., G. Kapetanios, and H. Pesaran (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Chudik , A , Kapetanios , G Pesaran , H 2018 , ' A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models ' ECONOMETRICA*.

Diebold, F.X. (1, . . and R. . . . Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*(3), 253–263.

Espinoza, R., F. Fornari, and M. J. Lombardi (2012). The role of financial variables in predicting economic activity. *Journal of Forecasting 31*(1), 15 – 46.

Gordon, R. J. (1982). Price inertia and policy ineffectiveness in the united states, 1890-1980. *Journal of Political Economy 90*(6), 1087 – 1117.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

Serge, D., de Mattos Daiane, and P. G. C. Ferreira (2019). Nowcasting: An r package for predicting economic variables using dynamic factor models. *R J. 11*, 230.

Stock, J. H. and M. W. Watson (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual 4*, 351 – 394.

Stock, J. H. and M. W. Watson (1999). Forecasting inflation. *Journal of Monetary Economics 44*(2), 293–335.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*(2), 147 – 162.

Appendixes

Appendix A: variable description

No.	Dataset name	Source	Code	Variable
1	zx2	DANE	8	Indicador de Seguimiento a la Economia.
2	zx3	DANE	3	Tasa de desempleo.
3	zx4	DANE	6	Desocupados.
4	zx5	BANREP	5	Cartera Comercial.
5	zx6	BANREP	5	Cartera Consumo.
6	zx7	BANREP	5	Cartera Microcredito.
7	zx8	BANREP	5	Cartera Hipotecaria (sin ajustar).
8	zx9	BANREP	5	Cartera Bruta Total (sin ajustar).
9	zx10	BANREP	5	Cartera Neta Total (sin ajustar).
10	zx11	DANE	6	Índice de produccion real de la industria manufacturera.
11	zx12	Fedesarollo	2	Índice de Confianza del Consumidor.
12	zx13	Fedesarollo	2	Índice de Expectativas del Consumidor.
13	zx14	Fedesarollo	2	Índice de Condiciones Economicas.
14	zx15	Fedesarollo	2	Índice de Confianza Comercial.
15	zx16	Fedesarollo	2	Percepcion de la situacion economica actual de la empresa (comercial).
16	zx17	Fedesarollo	2	Nivel de existencias (comercial).
17	zx18	Fedesarollo	1	Expectativas sobre la situacion economica para el proximo semestre (comercial).
18	zx19	Fedesarollo	7	Índice de Confianza Industrial.
19	zx20	Fedesarollo	7	Volumen actual de pedidos (industrial).
20	zx21	Fedesarollo	7	Nivel de existencias (industrial).
21	zx22	Fedesarollo	3	Expectativas de produccion para los proximos tres meses (industrial).
22	zx23	DANE	6	Ventas reales del comercio minorista (sin vehiculos ni combustibles).
23	zx24	DANE	5	Índice de Precios al Consumidor.
24	zx25	DANE	5	Índice de Precios al Productor (total nacional).
25	zx26	BANREP	5	Índice de la tasa de cambio real (Comercio total, IPP deflactor).
26	zx27	BANREP	5	Índice de la tasa de cambio real (Comercio total, IPC deflactor).
27	zx28	DANE	6	Índice de numero de licencias aprobadas para construccion (empalme DANE)
28	zx29	DANE	6	Índice de area aprobada para consturccion (empalme DANE)
29	zx30	DANE	6	Produccion de cemento gris (toneladas)
30	zx31	DANE	6	Despachos de cemento gris (toneladas)
31	zx32	BANREP	5	Efectivo en circulacion en poder del sector real (entidades y agentes que no son establecimientos de credito)
32	zx33	BANREP	5	Efectivo + Reserva Bancaria
33	zx34	BANREP	5	Ahorro + CDT. No incluye los CDTs en poder del Banco de la Republica ni los CDT emitidos por Findeter en el marco de la resolucion 1318 de 2020.
34	zx35	BANREP	5	Efectivo + Depositos en cuenta corriente.
35	zx36	BANREP	5	M1 + Cuasidineros. No incluye los CDTs en poder del Banco de la Republica ni los CDT emitidos por Findeter en el marco de la resolucion 1318 de 2020.
36	zx37	BANREP	5	Efectivo + Total Depositos en poder del publico.
37	zx38	XM	6	Demanda comercial nacional (Gwh).
38	zx39	DANE	5	Exportaciones totales segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
39	zx40	DANE	5	Exportaciones de Agropecuarios, alimentos y bebidas segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
40	zx41	DANE	5	Exportaciones de Combustibles y Prod de industrias extractivas segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
41	zx42	DANE	8	Exportaciones manufactureras segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
42	zx43	DANE	8	Exportaciones de otros sectores segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
z 43	zx44	DANE	8	Importaciones totales segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
44	zx45	DANE	6	Importaciones de Agropecuarios, alimentos y bebidas segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
45	zx46	DANE	6	Importaciones de Combustibles y Prod de industrias extractivas segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
46	zx47	DANE	6	Importaciones manufactureras segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
47	zx48	DANE	6	Importaciones de otros sectores segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.
48	zx49	FNC	6	Miles de sacos de 60 Kg de cafe verde equivalente
49	zx50	BANREP	2	Tasa representativa del mercado mensual promedio
50	zx51	BANREP	2	Tasa representativa del mercado fin de mes
51	zx52	DIAN	6	Recaudo total por impuestos
52	zx53	XM	6	Demanda de energia no regulada (Gwh)
53	zx54	XM	6	Demanda de energia de industrias manufactureras (Gwh)
54	zx55	XM	5	Demanda de energia de explotacion de minas y canteras (Gwh)
55	zx56	BANREP	2	Reservas bancarias
56	zx57	BANREP	5	Miles de millones COP
57	zx58	BANREP	5	Millones de USD
58	zx59	FRED	5	USD
59	zx60	BANREP	5	COP
60	zx61	DANE	5	Centavos de dolar por libra de 453.6 gr de Cafe Excelso

Appendix B: OCMT variable selection - results

Long description	# Of times selected	Inclusion frequency	2016	2017	2018	2019	2020	2021
1st lag dependent variable	70	1	1	1	1	1	1	1
2nd lag dependent variable	34	0.486	1	1	1		1	1
12th lag dependent variable	49	0.7	1	1	1	1	1	1
Indicador de Seguimiento a la Economia.	0	0						
Tasa de desempleo.	0	0						
Desocupados.	0	0						
Cartera Comercial.	0	0						
Cartera Consumo.	1	0.014	1					
Cartera Microcredito.	0	0						
Cartera Hipotecaria (sin ajustar).	0	0						
Cartera Bruta Total (sin ajustar).	0	0						
Cartera Neta Total (sin ajustar).	0	0						
Índice de produccion real de la industria manufacturera.	0	0						
Índice de Confianza del Consumidor.	0	0						
Índice de Expectativas del Consumidor.	0	0						
Índice de Condiciones Economicas.	0	0						
Índice de Confianza Comercial.	0	0						
Percepcion de la situacion economica actual de la empresa (comercial).	0	0						
Nivel de existencias (comercial).	0	0						
Expectativas sobre la situacion economica para el proximo semestre (comercial).	38	0.543	1	1	1		1	1
Índice de Confianza Industrial.	0	0						
Volumen actual de pedidos (industrial).	0	0						
Nivel de existencias (industrial).	0	0						
Expectativas de produccion para los proximos tres meses (industrial).	0	0						
Ventas reales del comercio minorista (sin vehiculos ni combustibles).	0	0						
Índice de Precios al Consumidor.	2	0.029						1
Índice de Precios al Productor (total nacional).	6	0.086						1
Índice de la tasa de cambio real (Comercio total, IPP deflactor).	0	0						
Índice de la tasa de cambio real (Comercio total, IPC deflactor).	0	0						
Índice de numero de licencias aprobadas para construccion (empalme DANE)	0	0						
Índice de area aprobada para consturccion (empalme DANE)	0	0						
Produccion de cemento gris (toneladas)	0	0						
Despachos de cemento gris (toneladas)	0	0						
Efectivo en circulacion en poder del sector real (entidades y agentes que no son establecimientos de credito)	0	0						
Efectivo + Reserva Bancaria	0	0						
Ahorro + CDT.	0	0						
Efectivo + Depositos en cuenta corriente.	0	0						
M1 + Cuasidineros.	0	0						
Efectivo + Total Depositos en poder del publico...	0	0						
Demanda comercial nacional (Gwh).	0	0						
Exportaciones totales segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Exportaciones de Agropecuarios, alimentos y bebidas segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Exportaciones de Combustibles y Prod de industrias extractivas segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Exportaciones manufactureras segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Exportaciones de otros sectores segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Importaciones totales segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Importaciones de Agropecuarios, alimentos y bebidas segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Importaciones de Combustibles y Prod de industrias extractivas segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Importaciones manufactureras segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Importaciones de otros sectores segun grupos de productos OMC a partir de la agregacion CUCI Rev.3.	0	0						
Miles de sacos de 60 Kg de cafe verde equivalente	0	0						
Tasa representativa del mercado mensual promedio	0	0						
Tasa representativa del mercado fin de mes	0	0						
Recaudo total por impuestos	0	0						
Demanda de energia no regulada (Gwh)	0	0						
Demanda de energia de industrias manufactureras (Gwh)	0	0						
Demanda de energia de explotacion de minas y canteras (Gwh)	0	0						
Reservas bancarias	0	0						
Reservas Internacionales Brutas (Millones de USD)	0	0						
Spot WTI (USD)	0	0						
Precio Interno del Cafe (COP)	0	0						
Precio Externo del Cafe (Centavos de dolar por libra de 453.6 gr de Cafe Excelso)	0	0						
First Common Factor	70	1	1	1	1	1	1	1
Second Common Factor	70	1	1	1	1	1	1	1
Third Common Factor	70	1	1	1	1	1	1	1
Fourth Common Factor	70	1	1	1	1	1	1	1
Fifth Common Factor	43	0.614	1	1	1	1	1	1
Sixth Common Factor	0	0						
Number os selected factors by Bai and Ng criterion:		4.6	4.455	4.083	4.167	5	5	5
Number of selected regressors by OCMT procedure (excludes PCs, which are always conditioned on)		2.8	2.091	3.833	2.667	2	2.5	4.091
Number of variables in the forecasting model apart from the intercept (includes PCs)		7.5	6.545	7.917	6.833	7.9	7.9	9.091

Appendix C: RMFE for Mean and AR(1) approach

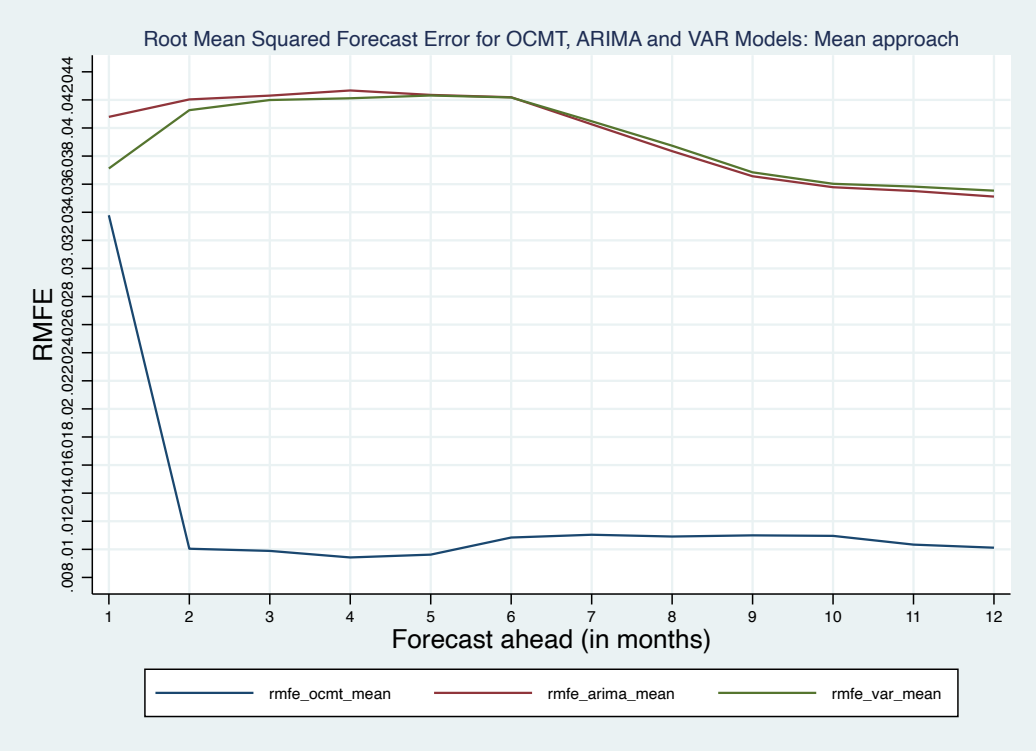

Figure 2: RMFE comparison for 1 to 12 months ahead out-of-sample forecasts (mean approach).

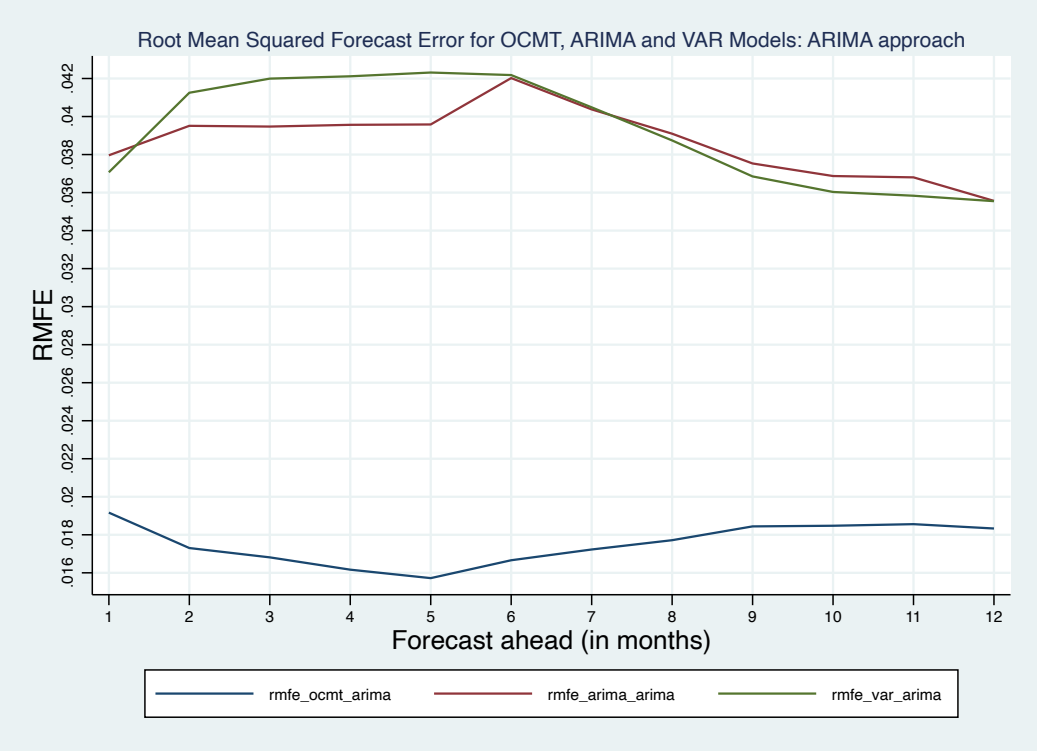

Figure 3: RMFE comparison for 1 to 12 months ahead out-of-sample forecasts (AR(1) approach).

Appendix D: Diebold-Mariano Test results

Mean	Full sample: 2006m2-2021m11		
Diebold-Mariano test	OCMT	ARIMA	VAR
MSE	0.059	0.074	0.079
P-Value: OCMT against	-	0.017	0.008

Table 4: Diebold-Mariano (1995) Tests: Mean approach.

Arima	Full sample: 2006m2-2021m11		
Diebold-Mariano test	OCMT	ARIMA	VAR
MSE	0.0624	0.074	0.0793
P-Value: OCMT against	-	0.023	0.013

Table 5: Diebold-Mariano (1995) Tests: AR(1) approach.

Appendix E: Expanded Diebold-Mariano Test results

Naive	Full sample: 2006m2-2021m11		
Diebold-Mariano test	OCMT	ARIMA	VAR
MSE (h=1)	0.06114	0.08092	0.07903
P-Value: OCMT against	-	0.0927	0.0028
MSE (h=2)	0.06099	0.08288	0.07908
P-Value: OCMT against	-	0.0868	0.0013
MSE (h=3)	0.06082	0.08485	0.07912
P-Value: OCMT against	-	0.0740	0.002
MSE (h=4)	0.06065	0.08682	0.07916
P-Value: OCMT against	-	0.0633	0.003
MSE (h=5)	0.06086	0.08878	0.0792
P-Value: OCMT against	-	0.0535	0.003
MSE (h=6)	0.06237	0.09073	0.07925
P-Value: OCMT against	-	0.0392	0.005
MSE (h=7)	0.06248	0.09266	0.07929
P-Value: OCMT against	-	0.0262	0.009
MSE (h=8)	0.06258	0.09457	0.07933
P-Value: OCMT against	-	0.0249	0.0114
MSE (h=9)	0.06281	0.09646	0.07937
P-Value: OCMT against	-	0.0167	0.087
MSE (h=10)	0.0627	0.09833	0.07941
P-Value: OCMT against	-	0.0085	0.0055
MSE (h=11)	0.0626	0.1002	0.07944
P-Value: OCMT against	-	0.0035	0.0096
MSE (h=12)	0.06305	0.102	0.07948
P-Value: OCMT against	-	0.0032	0.0087

Table 6: Diebold-Mariano (1995) Tests for all forecasting horizons: Naive approach.