Erasmus School of Economics

Multi-horizon comparison of multivariate inflation forecasting

Robert V. Janssen,
July 2019

Abstract

This paper applies the multi-horizon comparison methodology from Quaedvlieg (2019) to assess the forecasting performance of direct and iterative multivariate inflation forecasts, with both high and low lag orders. We use various macroeconomic indicators in a GETS restricted estimation to forecast US inflation and show that high order VARs on average prefer iterative forecasts, while low order VARs on average prefer the direct forecasts. Finally, we provide evidence that the best high order multivariate forecasts outperform the best low order multivariate forecasts on every individual horizon (uniform superior predictive abilities). This implies that in this setting, inflation forecasts are most accurately forecasted with a high order VAR using an iterative approach.

Keywords: Inflation, Multi-horizon, VAR forecasting, Superior Predictive Ability

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis Financial Economics

Multi-horizon comparison of multivariate inflation forecasting

Name student: Robert Janssen
Student ID number: 386097

Supervisor: Rogier Quaedvlieg
Second assessor: [To be determined]

Date final version: 25th July 2019

Erasmus University Rotterdam

1. Introduction

Inflation has a vital impact on the modern-day economy. It is included in most signed contracts and is used by federal agencies to determine monetary policy. Therefore, inflation forecasting has been extensively covered in the literature, but researchers still aim to reach a consensus on which forecasting methodologies currently available, have superior forecasting abilities using which set of parameters. The forecasting horizon (h) is one parameter which might influence model selection and forecasting methodologies, as a nowcast ($h = 0$) is different from short-term ($h = 1,2$) or long-term forecasts ($h = 8,12,24$). To assess which forecast is optimal, methodologies can be compared on various horizons individually, but the performance of a forecast on the entire forecasting path (i.e. from h=1 to h=12) is then discarded.

This study will use the concepts of average superior predictive ability (aSPA) and uniform superior predictive ability (uSPA) as introduced by Quaedvlieg (2019) to evaluate multi-horizon forecasting performance. In his paper, the author shows the contribution of his multi-horizon forecast comparison on the study by Marcellino, Stock and Watson (2006), who compare the performance of iterative forecasting (one-period ahead forecast iterated forward to desired horizon) versus direct forecasting (horizon-specific model estimation with the dependent equal to multiperiod ahead value being forecasted). Unlike Quaedvlieg (2019) and Marcellino et al. (2006), this study will consider the implications of direct and iterative multivariate multi-horizon forecasting for one macroeconomic indicator only; inflation. With this focus, we intend to provide a multivariate extension on studies with univariate inflation indicators and an extension on studies which focus on exhibiting the forecasting methodology. We will combine the latest forecasting performance statistic with the most relevant macroeconomic indicators for inflation.

These inflation indicators will be based on the categorization of macroeconomic timeseries by Marcellino et al. (2006), who divide 170 timeseries into five categories, named category A to E. Pesaran, Pick and Timmermann (2011) base their multivariate forecasts on the same dataset by Marcellino et al. (2006) to test how certain parameters, like lag selection based on the Akaike Information Criterion (AIC) or the Bayes Information Criterion (BIC), can affect direct and iterative forecasting performance. Just like Marcellino et al. (2006), the authors find that Category E, which includes inflation, wages and money, considerably differentiates in forecasting preferences in addition to the distinctions caused by lag selection. Therefore, this paper will distinguish lag selection through AIC and BIC and distinguish two categories; Category E[1] and what this paper will address as Category M, including one indicator from each other category (A-D).

Overall, the objective of this study is to assess if direct or iterative forecasting methodologies have superior predictive abilities in a multi-horizon comparison of multivariate inflation forecasts. To

[1] These categories A-D and E refer to the similarly named categories in Marcellino et al. (2006)

hypothesize on the conclusion of the objective, it is useful to separate under which circumstances direct or iterative forecasting is preferred. Marcellino et al. (2006) show in a bivariate vector autoregressive (VAR) with low lag order (*p*) that direct forecasting is outperforming the iterative approach, while with a high lag iterative forecasting appears to excel. Considering different forecasting horizons (*h*), the authors find that for short horizons a direct approach is optimal, while for a long horizon an iterative approach is more suitable. Such findings actually confirm the presence of the misspecification bias of short lag models. When Marcellino et al. (2006) consider the categories included in their bivariate forecasts, they find that when variables from Category E are excluded, iterative outperforms direct in most forecasts. Also when one of the two variables is included in Category E, iterative forecasts appear to be more accurate. However, when the bivariate model contains two variables from Category E, direct forecasts are suddenly more accurate. Pesaran and Timmermann (2011) find that the multivariate iterative forecasts for variables from category A-D outperform the direct forecasts regardless of specific horizons, estimation windows or lag-selection methods. For Category E it is exact opposite, as the authors find that direct forecasts outperform iterative forecast regardless of the parameters. These findings are slightly different than those from Marcellino et al. (2006), as they argued that the lag order did have an effect on the performance of forecasting for most indicators. However, as both papers show that direct and iterative preference also depends on uni-, bi- or multivariate modelling and as Pesaran et al. (2011) contains multivariate modelling, it is decided to follow their results for the hypothesis. Therefore, we would expect for the Category E inflation forecast that the direct approach will outperform the iterative approach, regardless of the lag structure, while for the forecast with Category M, including variables from all categories, we would expect the iterative approach to outperform the direct approach.

We will construct multivariate vector autoregressions (VARs) with, next to US CPI, five macroeconomic indicators from Category M and Category E to estimate inflation. After transforming the selected timeseries, an initial assessment of correlations between the macroeconomic indicators will be provided with some background on the relations between some of these indicators. The AIC and the BIC will be evaluated to decide on optimal lag orders. Next, general-to-specific (GETS) restriction will be applied, where insignificant coefficients will be removed from the VARs. Then the specific estimations for both categories and both lag orders will be forecasted according to an iterative and direct approach. These forecasts will be evaluated with respect to each other and a univariate autoregressive (AR) benchmark in a multi-horizon comparison to determine which forecasts have superior predictive abilities.

This paper continues as follows, section 2 provides a brief overview of inflation forecasting literature of the past two decades, followed by section 3 explaining the data and descriptive statistics and section 4 describing the estimation and restriction methodology. Section 5 continues with forecasting methods and discusses the results from the multi-horizon comparison and section 6 discusses limitations and implications for future research. Finally, section 7 will conclude.

2. Literature Review

Due to the vital role of inflation in the US economy, the inflation forecasting literature is abundant. James H. Stock and Mark W. Watson have published several studies since 1997 on forecasting inflation in the US and their studies serve as a good example of the development of the inflation forecasting literature over time, of which a concise overview will be provided below. Later, Faust and Wright (2013) construct a horse-race to observe the performance of Stock and Watson models with respect to other models available at that time. In addition, mixed frequency, big data and machine learning models will be assessed, before moving on to a few examples of other variable types which can be used to forecast inflation. Finally, parameters affecting forecasting performances and other forecasting approaches will be discussed.

Staiger, Stock and Watson (1997) start with studying the relationship of the natural unemployment rate (NAIRU) with respect to inflation in combination with a Phillips curve. Eventually, they find that real aggregate activity with an index of 168 economic indicators provides the best short run forecasts of inflation instead of unemployment (Stock and Watson, 1999). A few years later, Stock and Watson (2002) focused on forecasting macroeconomic indicators in general and used a multivariate VAR to forecast CPI with industrial production (as proxy for real activity) in an iterative forecast using the BIC to determine the lag order. They find that the VAR outperforms the AR on a 12- and 24-month horizon, but not on a 6-month horizon. In 2003, Stock and Watson find that real asset prices and spreads have little to no effect on forecasting inflation. Just before the financial crisis, Stock and Watson (2007) published a paper on the difficulties in forecasting inflation, where they conclude that inflation should be well defined by time-varying parameters in an integrated moving average process. Multivariate forecasts have difficulties to beat the time-varying univariate model in this study. After the financial crisis, the authors aim to provide an example to support inflation forecasting after the recession. They propose a model with a time-varying moving average parameter and a single Phillips curve coefficient, where the deviation of short-term inflation reacts stably to the unemployment recession gap (Stock and Watson 2010). One issue with all these methodologies applied over the years, is that it is difficult to compare them due to different sample sizes of inflation timeseries (Stock and Watson, 2008). Faust and Wright (2013) provide a summary of the performance by the ‘explosion’ of new models with respect to the more traditional models. They conclude despite the surprisingly hard to beat univariate autoregressive model, that the judgmental survey forecasts are considerably more accurate than model-based forecasts. With respect to multivariate VARs, Faust and Wright (2013) include a term structure VAR(1) and a VAR(2) with time-varying parameters for inflation, unemployment rate and treasury bill yields, which is unable to outperform an AR(1).

One of the reasonably well performing models, unobserved components with stochastic volatility model, is used in combination with random-walk-plus-white-noise by Stock and Watson (2016) to study a multivariate extension. The authors find that incorporating core or sectoral inflation data improves the accuracy of headline inflation forecasting. Nonetheless, one caveat of their study is that quarterly averages of monthly rates have been used, which can be resolved with mixed frequency forecasting. For example, Schorfheide and Song (2015) construct a mixed frequency VAR where data from different frequencies are combined into one estimation to forecast inflation. Harchaoui and Janssen (2018) use a MIDAS estimation to forecast monthly inflation based on daily online inflation as constructed by Cavallo and Rigobon (2016), also known as the Billion Prices Project (BPP). This project uses web-scraping software to extract online prices on a daily basis into a big data set which is converted towards an index.

Big data applications like the BPP have become available due to the development of computational power since the beginning of this century. Inoue and Kilian (2008) were one of the first to argue that machine learning models could outperform univariate inflation benchmarks, based on the bootstrap aggregation of forecasts (bagging). Medeiros, Veiga, Vasconcelos and Zilberman (2019) study multiple machine learning methods using extensive data sets and find that all machine learning methods in their study are able to outperform univariate models on several horizons. The Random Forest model, which is a collective of fully-grown regression trees estimated on different bootstrap samples of original data, appears to be the best performing machine learning model. A decomposition of the indicators selected by this model show that prices, employment, housing and interest exchanges have the most relative importance.

The models and approaches mentioned sofar, provide an indication of how diverged and broad the inflation forecasting literature is. With respect to the inputs for inflation forecasts both surveys and macroeconomic indicators have been discussed. Checchetti, Chu and Steindel (2000) include commodity prices and exchange rates in addition to these two in their discussion on the unreliability of inflation indicators. In their univariate comparison the survey-based forecasts also appear to be the best, followed by the price of gold and oil and the exchange rates. The authors argue that although single indicators might be unreliable in a simple estimation, when they are incorporated in a multivariate model, they might actually become reliable. With indicators like commodities and exchange rates, a Global VAR (GVAR) can be used to include connections between countries and variables (Pesaran, Schuermann and Weiner, 2004). In a later study, Pesaran, Schuermann and Smith (2009) use this approach as they acknowledge that macroeconomic policies need to consider the increasing interdependencies across markets and countries. They construct a VAR with GDP, inflation, interest, equity prices and exchange rates and benchmark their GVAR with respect to univariate autoregressive and random walk models. In addition to the ability of GVARs to outperform their benchmarks, they find that for developed economies incorporating financial variables, like long term interest rates and real equity prices, improves the accuracy of inflation forecasts.

Forecasting approaches

The papers discussed above are primarily focused on the model and variables to apply for the estimation, yet various parameters can have substantial effects on forecasting performances of estimations. Pesaran et al. (2011) use the same 170 macroeconomic indicators as Marcellino et al. (2006), to study whether direct or iterative forecasts provide more accurate forecasts under changing forecasting parameters. They distinguish univariate and multivariate estimations, AIC and BIC lag selection methodologies, short and long estimation windows, short and long horizons and even Category E and Category M variables. The authors conclude that there is no dominant approach for all macroeconomic indicators, but that for instance, univariate iterative models are more accurate in forecasting for Category E variables. Moreover, Binner, Elger, Jones and Nilsson (2010) show that especially long horizon forecasts are improved if a measure of skewness is included in the forecast equation. McElroy (2015) adds that in a univariate setting, if the model and estimation techniques are similar, the direct multi-step and iterative forecasts will be identical.

Another approach to forecast macroeconomic indicators is through forecasting combinations, as introduced by Aiolfi and Timmermann (2006), where models clustered on their performance are followed by pooled forecasts within each cluster. Forecast combination weights, which are assigned to the forecasts, can either be set to equal or determined by estimation. Chan and Pauwels (2018) later propose a framework to assess forecast combinations and they find that increasing the number of models included in a forecast combination, decreases the MSFE, by which these forecasts are evaluated and that selecting models based on the MSFE also results in selection bias. One caveat in their paper is that they cluster forecasts based on past performance of one period ahead forecasting. This would imply that the performance on multiple horizons is not considered at all, which might therefore bias the selection as well. The authors could use the aSPA and uSPA statistic by Quaedvlieg (2019) to cluster the forecasts based on their multi-horizon forecasting accuracy. In addition to the pairwise multi-horizon test, Quaedvlieg also extends upon the Model Confidence Set (MCS) of Hansen, Lunde and Nason (2011) which allows to select the most accurate methodologies of a larger set of models, which could then be clustered for forecasting combinations as well. Hansen et al. (2011) tested their methodology on the inflation forecasting approaches stated in Stock and Watson (1999) and also find that pre-1984 the Phillips curve provides the most suitable forecast. Post-1984 however, the MCS selects the autoregressive no-change forecasts (month-over-month) as most accurate forecasting methodology.

In short, inflation forecasting literature has evolved noticeably over the past two decades, moving from the Phillips curve with unemployment rates to a Global VARs and random-walk-plus-white-noise unobserved components with stochastic volatility models. Still, there is a vast amount of literature occupied with determining which of the existing models forecast more accurately using which set of parameters. This paper will contribute to the inflation forecasting literature by demonstrating how a multi-horizon comparison for a multivariate VAR with macroeconomic indicators can help to select optimal forecasting models.

3. Data and descriptive statistics

In this section, a more detailed description of the macroeconomic timeseries will be provided in addition with a motivation for the choice to use revised headline CPI as inflation measure. Then, the descriptive statistics of the transformed timeseries will be discussed along with the correlations between the various indicators.

Other than Pesaran et al. (2011) and Quaedvlieg (2019), we do not use the entire dataset of 170 macroeconomic timeseries from Marcellino et al. (2006). In addition to CPI, we will select five timeseries from Category E to form the Category E VAR and one timeseries from each other Category (A-E) to form the Category M VAR The timeseries included in Category E of Marcellino et al. (2006) can actually be divided into six indicators: Consumer Prices Indices (CPI), Produces Price Indices (PPI), Personal Consumption Expenditures (PCE), Average Hourly Earnings, Money Stock and Monetary Bases. Each indicator contains several timeseries which provide industry or product specific information. To estimate and forecast inflation, it would be ideal if the timeseries representing an indicator cover similar industries and products as well, but most importantly the timeseries included should have the most impact on the economy with respect to others.

For inflation there has been an ongoing debate related to whether it should be estimated and forecasted using headline inflation (all items) or core inflation (without food and energy). Faust and Wright (2013) argue that food and energy could be treated as pure noise and that when core inflation is forecasted it could be considered as if it were a prediction of total inflation. They support this by showing that when core inflation is forecasting total inflation, it is mostly outperforming the same methodologies using total inflation as input for total inflation forecasts. Bullard (2011), however, argues that a simple model with core inflation would be misspecified and acting as a proxy for other variables, like developments in the real economy and monetary policy, that should actually be included in predicting headline inflation. Eventually, Thornton (2011) concludes that headline inflation should be targeted by policymakers as it is a widely used index for both public and private contracts, despite his own argument that neither core nor headline inflation is consistent or reliable enough to be helpful for policymakers. We decide that because of its wide-spread use in contracts and the large economic influence to use headline inflation (CPI all items) in this paper. Furthermore, the use of revised or vintage data in academic research is debatable, but this paper follows Komunjer and Owyang (2012) by assuming that forecasters aim to forecast a true value and that a revision is a more precise reflection of the true value. This implies that for the other five indicators the more general, influential and revised timeseries have been chosen as well, as can be seen in Table 1.

The macroeconomic indicators which will compose Category M, includes the same inflation measure (CPI all items) and combines it with an indicator from every category, including Category E as well. From the various categories it is decided to retrieve the Industrial production index (A[2]); Civilian unemployment rate (B); Manufacturing purchasing managers index (C); the 10-year treasury constant maturity rate (D) and the producer price index (E), as these are the indicators which are most often discussed in economic news compared to the other indicators or timeseries in those categories.

Data transformation and descriptive statistics

For this study, the monthly data is extracted from the FRED, Federal Reserve bank of St. Louis for all timeseries except the PMI[3]. We use the eldest to the most recent data available and then reduce the sample to the timeseries with the most missing data, resulting in a range from January 1972 until December 2018. Since this sample period is different from the one used by Marcellino et al. (2006) and Pesaran et al. (2011), the results will not be directly comparable, but it will include the latest sixteen years of macroeconomic data. Table 1 shows an overview of the timeseries extracted and their abbreviations which will be used throughout this paper.

Table 1 – Time series and transformations

Notes: NSA depicts non-seasonally adjusted; SA depicts seasonally adjusted. The St. Louis codes allow to match the timeseries applied here.

Variable	Abbreviation	Trans.	(N)SA	Category	St. Louis code	Description
CPI	CPI	LnΔ	NSA	E, M	CPIAUCNS	Consumer Price Index: all urban consumers all items
PPI	PPI	LnΔ	NSA	E, M	WPUFD49207	Producer price index: finished goods
Gov. 10y Rate	GovR	Δ	NSA	M	GS10	10Y Treasury constant maturity rate
IP Index	IPI	LnΔ	NSA	M	IPB50001N	Industrial production index
Unemployment	UnemR	Δ	NSA	M	UNRATENSA	Civilian unemployment rate
PMI	PMI	Δ	SA	M		Manufacturing purchasing managers index
PCE	PCE	LnΔ	SA	E	PCE	Personal consumption expenditures
Avg. Hr. Earn	AHE	LnΔ	NSA	E	CEU0500000008	Average hourly earnings: private nonagricultural
M1	M1	LnΔ	NSA	E	M1NS	Money stock: M1
Monetary Base	MB	LnΔ	NSA	E	BOGMBASE	Monetary base adjusted for reserve req. changes

For PMI and PCE we extract timeseries which are already corrected for seasonality, as non-seasonally adjusted data is unavailable. For the remaining ten timeseries, non-seasonally adjusted data is extracted, which is corrected for seasonality by computing the twelve-month variation compared to the same month a year before. The transformations applied after seasonality corrections are similar to Marcellino et al. (2006) except for PMI, which they keep at its level as they perform univariate and bivariate autoregressions. However, the first order difference is required to allow the PMI trend to contribute to forecasting inflation.

Table 2 shows the descriptive statistics of the timeseries, which includes 551 monthly observations from 1973:01 until 2018:12, since the data for 1972 is dropped because their annual variations could not be

[2] The letters refer to the same categories as in Marcellino et al (2006)
[3] PMI was extracted through the economic trends database of AeroWeb.

established due to missing data. Due to the transformations, the mean and median of most indicators are small and close to each other, but the considerable levels of kurtosis and skewness still present keep these timeseries non-normally distributed. Monetary base (MBAS) is compared to the other indicators even more skewed and has a larger peak in its distribution, which is most likely caused by the consistent monetary policy in the US until the early 2000s followed by substantial monetary expansion until now (Appendix Figure A.1). The considerable maxima and minima of government rate (GovR), unemployment rate (UnemR) and PMI compared to their means and medians are not surprising as these variables were not transformed on a logarithmic scale.

Table 2 – Descriptive statistics after transformations

Notes: Descriptive statistics based on 1973:01-2018:12 sample after transformations

	CPI	PPI	GovR	IPI	UnemR	PMI	PCE	AVGHE	M1	MBAS
Category	*E,M*	*E,M*	*M*	*M*	*M*	*M*	*E*	*E*	*E*	*E*
Mean	-0.000031	-0.000058	-0.000145	-0.000116	0.001452	-0.032305	0.005178	-0.000028	-0.000089	-0.000316
Median	-0.000060	-0.000008	0.010000	-0.000533	0.000000	0.000000	0.004691	0.000203	-0.000316	-0.000138
Maximum	0.020046	0.041785	1.840000	0.048310	1.200000	10.500000	0.027510	0.009416	0.057786	0.231985
Minimum	-0.025261	-0.047384	-2.690000	-0.051674	-1.300000	-9.400000	-0.020767	-0.013094	-0.051191	-0.193739
Std. Dev.	0.003901	0.008522	0.435569	0.011509	0.259296	2.231569	0.005275	0.003529	0.010437	0.026991
Skewness	-0.365506	0.007048	-0.211442	0.160491	-0.036021	-0.039609	0.082463	-0.402081	0.304833	1.243448
Kurtosis	7.985885	6.851395	7.566677	5.607179	5.130987	5.062777	5.958162	3.159290	7.935987	31.594830
Jarque-Bera	583	341	483	158	104	98	202	15	568	18914
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000446	0.000000	0.000000
Sum	-0.016925	-0.032162	-0.080000	-0.063698	0.800000	-17.800000	2.852966	-0.015189	-0.049282	-0.174288
Sum Sq. Dev.	0.008370	0.039947	104.34600	0.072856	36.978840	2738.9450	0.015307	0.006851	0.059915	0.400684
Observations	551	551	551	551	551	551	551	551	551	551

To increase the understanding of the potential relations between the macroeconomic indicators and the indicators of the different categories, Table 3 provides the correlation matrix. The high correlation of PPI with CPI compared to other indicators of Category E confirms that it is the right indicator to include in the Category M VAR as well. Basic economic theory suggests that changes in producer prices would be reflected in consumer prices and therefore the correlation with CPI would be straightforward. Clark (1995) initially shows that this is not necessarily the case, but he argues that despite the weak causal link, PPI might prove helpful in forecasting CPI. Caporale, Katsimi and Pittis (2002) later provide some empirical evidence, which confirms that causality is running from PPI towards CPI.

The government 10-year interest rate is correlated with both CPI and PPI, which is in line with Lee (1992), who shows in his multivariate approach that interest rates explain a substantial fraction of inflation. However, the correlation with CPI and PPI is positive, while according to basic economic theory one would expect the opposite, as lower government rates should increase spending and therefore inflation. Another explanation for this is that the timeseries used for this indicator is the long-term treasury rate which is very subjective to global market dynamics. This is different for the Federal Fund Rate, set by the FED, which is initially more in line with basic theory.

With respect to the other indicators in Category M (IPI, Unemployment, PMI), there appears to be no correlation with CPI. Meanwhile the IPI is significantly correlated with almost all other indicators from both categories. As the IPI is generally considered an indicator for GDP it is not surprising that it is correlated with many other indicators. Unfortunately, this correlation between Category E indicators and IPI from Category M might lead to difficulties with statistically differentiating the forecasts later. The four indicators from Category A-D (shown in the table 3 as M), show correlations with each other, which suggests that significant relations amongst these categories exists which should not be discarded.

Table 3 – Correlation matrix

Notes: Correlation matrix based on 1973:01-2018:12 sample. *, **, *** indicate respectively p<0.1, p<0.05, P<0.01.

	CPI	**PPI**	**GovR**	**IPI**	**UnemR**	**PMI**	**PCE**	**AVGHE**	**M1**	**MBAS**
Category	*E,M*	*E,M*	*M*	*M*	*M*	*M*	*E*	*E*	*E*	*E*
CPI	----									
p-value	----									
PPI	0.7367	----								
p-value	0.000***	----								
Gov. 10y Rate	0.1958	0.1622	----			*Category M*				
p-value	0.000***	0.0001***	----							
IP Index	0.0344	0.0022	0.1780	----						
p-value	0.4207	0.9583	0.000***	----						
Unemployment	-0.0626	-0.0407	-0.1886	-0.3950	----					
p-value	0.1425	0.3400	0.000***	0.000***	----					
PMI	-0.0201	0.0640	0.2405	0.2229	-0.1756	----				
p-value	0.6386	0.1336	0.000***	0.000***	0.000***	----				
PCE	0.1099	0.1187	0.0325	0.1481	-0.0658	0.0646	----			*Category E*
p-value	0.0098***	0.0053***	0.4461	0.0005***	0.1228	0.1300	----			
Avg. Hr. Earn	0.0634	0.0084	0.0220	0.1412	-0.0357	0.0754	-0.0349	----		
p-value	0.1369	0.8447	0.6071	0.0009***	0.4034	0.0769*	0.4138	----		
M1	-0.1506	-0.0842	0.0024	-0.1451	-0.0320	0.0603	-0.0516	-0.0016	----	
p-value	0.0004***	0.0483**	0.9552	0.0006***	0.4533	0.1573	0.2270	0.9697	----	
Monetary Base	-0.2953	-0.2813	-0.0522	-0.1281	0.0754	-0.0995	-0.1118	-0.0156	0.3275	----
p-value	0.000***	0.000***	0.2216	0.0026***	0.0771*	0.0194**	0.0086***	0.7141	0.000***	----

Although the correlations of most indicators within Category E are not significant (see Category E box in table 3), PCE, M1 and Monetary base are significantly correlated with CPI. As PCE is actually another inflation measure it might act as an additional autoregressive factor in the estimation. Nonetheless, the discrepancies between the timeseries, which originate from PCE being based on what business are selling, while CPI is based on what households are buying, might be able to provide some additional predictive power (Haubrich and Millington, 2014). The fact that Monetary Base and M1 have similar correlations with respect to CPI is not surprising, as they could also be explained as two different timeseries providing an indication of money supply. Monetary Base captures all currency in circulation (which can be directly increased by the FED) and the reserve balances which banks have in their FED accounts. M1 captures all currency held by the public and at their banks. Theoretically an increase in money supply would push inflation. The FED also argues that they have observed historically close relationships between money supply and CPI, but in recent decades, the relevance of money supply as guide for monetary policy has weakened (FED, 2019).

Finally, IPI, Unemployment, PMI and Average Hourly Earnings are the four inflation indicators which are uncorrelated with CPI. Stock and Watson (1999) initially argued that in the Phillips curve the unemployment rate might be indicative for inflation in the short term. However, Checchetti et al. (2000) find that this indicator does not perform well in a univariate setting and Atkeson and Ohanian (2001) even state the conventional wisdom is wrong. Eventually, the uncorrelation between unemployment and inflation appears to be in line with the literature. Average Hourly Earnings is also included in Checchetti et al. (2000) and despite its outperformance with respect to the benchmark, a causality issue is also addressed. Arguably, this indicator arguably follows inflation instead of predicting it, which might make it less suitable as inflation indicator. In the same paper, the PMI (NAPM), is also covered, but with debatable performance. The PMI might actually be a more economy wide indicator, as it is evaluated by analysts when they sense a change in the direction of the economy. Checchetti et al. (2000) conclude that even when economic indicators do not always suit as an inflation indicator, they might still provide useful information in a multivariate setting. Koenig (2002) shows for example, that the PMI conveys useful real GDP information and that together with inflation and unemployment it is able to predict general Federal Reserve policy.

Despite not all indicators being correlated with CPI, this paper will still include these in the multivariate estimation, because of several reasons. First, the correlations above are crude and indicative, as they do not consider certain timeframes (e.g. rolling window correlations). Second, some uncorrelated indicators might provide some explanatory power through their relations with other variables. Third, the correlation estimation is contemporary instead of lagged. The variance decomposition of the restricted estimations will shed some more light on the relations between certain macroeconomic indicators and CPI in the next section.

4. Estimations and restrictions

This section will elaborate on the multivariate VAR estimation equations, the lag order selection and the restrictions which will be imposed on the specific coefficients of the multivariate VARs. After restrictions, the variance decomposition of the VARs will demonstrate which indicators are substantially affecting CPI in our regressions.

After transformations and including a maximum lag order of twelve months, the monthly data is available from 1974:01 until 2018:12. The estimation period has been randomly set until 1999:12, which implies that out-of-sample forecasting will start from 2000:01 until 2018:12.

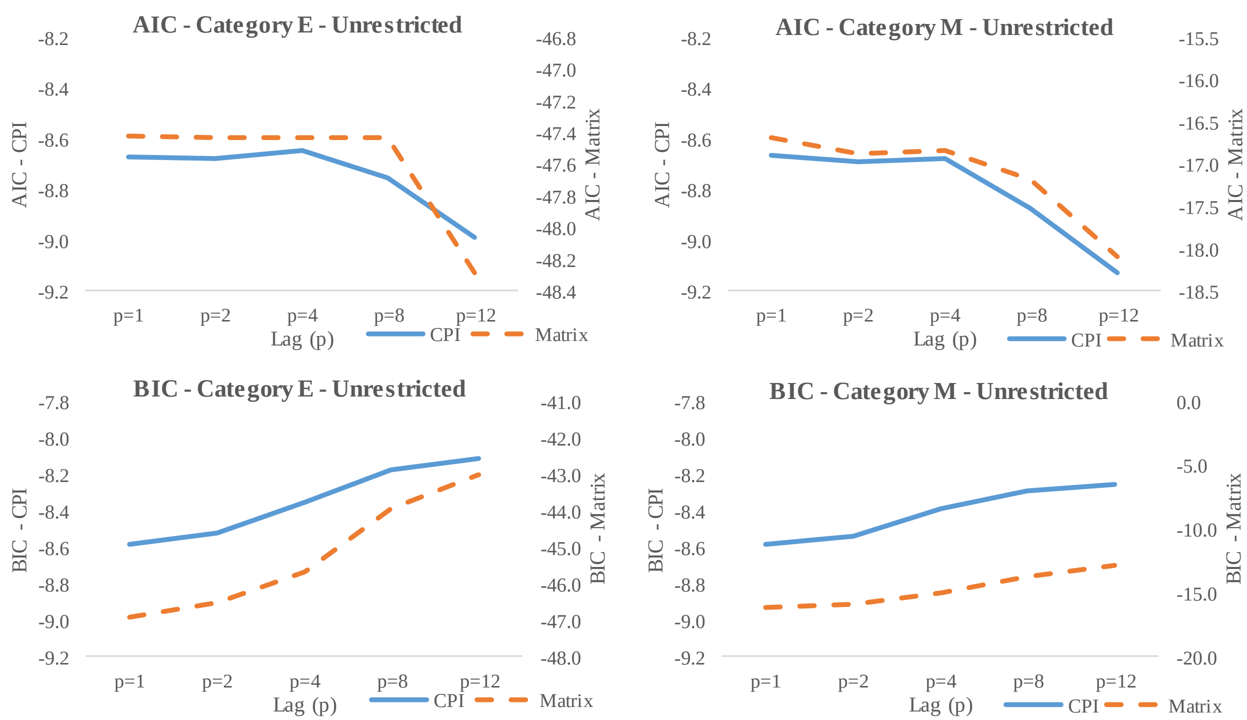
To combine the various relations between the macroeconomic indicators, multivariate vector autoregressions (VAR(p)) with an OLS estimation approach provide the ability to estimate the 6 x 6 matrix for p lags for each category into one regression as follows,

$$Y_{c,t} = \alpha + \sum_{i=1}^{p} \Phi_{c,i} Y_{c,t-i} + \varepsilon_t, \qquad (1)$$

where $Y_{c,t}$ is a vector of indicators at time t for category c, $Y_{c,t-i}$ denotes the indicator matrix for category c for the respective lag p at time t and $\Phi_{c,i}$ denotes the coefficient matrix for the respective indicators at lag p. The constant and the error term are respectively reflected by α and ε_t. Finally, for lag order p currently holds $p = 1,2,4,8,12$ and category $c = e, m$, where e represents Category E indicators and m represents Category M.

Figure 1 – Akaike and Bayes Information Criteria for unrestricted VAR(*p*) estimations

Notes: Matrix refers to the AIC and BIC of the entire VAR(p), where CPI refers to the AIC and BIC of inflation only.



One of the parameters in this study is the determination of the lags included in the estimation. We partially follow Marcellino et al. (2006), who use the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) to select the optimal lag order with $0 \leq p \leq 12$. However, the authors recompute the AIC and BIC at every step of their recursive forecast, which implies that lag orders might differ every step. This study will use the AIC and BIC to determine two different lag orders for the entire study. Figure 1 depicts the AIC and BIC of CPI within the multivariate estimation and the information criteria for the entire VAR(p) for both categories. One can observe that the AIC for CPI has a similar downward sloping trend towards a twelve-month lag. Although the absolute value of the AIC for the Category M VAR(p) is substantially lower than for Category E, they both visualize a similar trend, with a considerable drop at p = 12. This drop is most likely related the twelfth lag specifically, as it includes seasonality information of the year before, which is an effect of the year-on-year seasonality correction that cannot be ruled out. Since the AIC consistently points towards p = 12, it is selected as high order lag. With respect to the BIC, the CPI and the VAR(p) follow a similar trend for both categories, but away from high order lags to low order lags and evidently a one-month lag order is selected.

General-to-specific restriction

As argued by George, Sun and Ni (2008), researchers studying forecasting continuously have to balance the inclusion of enough information to analyze sophisticated issues with being restrictive enough to extract sharp results. In addition to these authors, Korobilis (2013) shows that Bayesian (stochastic) variable selection is able to improve forecasting accuracy with respect to their unrestricted counterparts while maintaining parsimonious models. The BIC selected models for Category E and M currently entail one lag of 6 x 6 = 36 coefficients. As mentioned by Krolzig (2000) every additional lag enlarges the estimation with 36 coefficients, which means that the AIC selected model includes 432 coefficients, most of which insignificant. The Granger-Causality (GC) test can be applied to impose restrictions, however in this research this method is surpassed because of two disadvantages. First, the GC test measures the probability that at least one specific lag coefficient of a variable is significant, but it does not consider how many lags are significant. Second, if the restrictions are imposed based on the test, many insignificant lags will be included in the estimation because one lag was significant.

An alternative of the GC test, is the more precise General-to-specific (GETS) approach for a VAR by Krolzig (2000), where the coefficients of insignificant variables for each lag are restricted to zero. The restricted multivariate models are then estimated again and if necessary additional restrictions are imposed. This implies that for both categories the coefficients in all lags individually, have been manually assessed if their t-statistics were within the -2.0 and 2.0 range. The significance is based on the estimation sample ranging from 1973:01 to 1999:12. Based on their statistical significance, restriction matrices have been manually constructed for every specific lag and every category. The multivariate estimation can now be described as follows;

$$Y_{c,t} = \alpha + \sum_{i=1}^{p} R_{c,i} \Phi_{c,i}^{R} Y_{c,t-i} + \varepsilon_t, \tag{2}$$

where $R_{c,i}$ denotes restriction matrix for lag i of category c and $\Phi_{c,i}^{R}$ denotes the restricted estimated coefficients for lag p of category c. As mentioned above, lag lengths of 1 and 12 have been selected based on the information criteria which implies $p = 1, 12$. So, for $p = 12$, every lag until 12 has its own restriction matrix capturing the coefficients of that specific lag which are restricted or not. Appendix Tables A.1 show the restriction matrices for the low order VAR and Tables A.2 show the restriction matrices of the first lag of the high order VAR only[4].

Table 4A – Restricted CPI VAR Category E

Note: all empty cells contain zeros in the restriction matrices

Lag	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
CPI	1	1					1		1	1		1
PPI	1									1		
PCE				1								
AVGHE			1									
M1												
MBAS										1		

Table 4B – Restricted CPI VAR Category M

Note: all empty cells contain zeros in the restriction matrices

Lag	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
CPI	1						1		1	1		1
PPI	1									1		
GovR						1						
IPI										1		
UnemR			1								1	
PMI			1									

Tables 4A and 4B are a visual representation of many restriction matrices (examples in the Appendix) to provide an overview on which specific variable is included in which specific lag to estimate CPI using a high order VAR. A 1 indicates that this specific coefficient is included in the estimation, while all other coefficients are not included and therefore restricted to zero. After restrictions, CPI includes 6 and 5 autoregressive lags in respectively Category E and M, which confirms the strong autoregressive component in inflation forecasting. This confirms the findings by Faust and Wright (2013) as well, who show that simple autoregressive inflation models are surprisingly difficult to beat. As the autoregressive and PPI lags are very similar for both categories, one can carefully conclude that the relation of PPI with respect to CPI is robust to the inclusion of other variables in the VAR. Simultaneously, this also raises the question whether all these indicators in both categories should still be included. M1 for instance, is completely restricted and Monetary Base is only included with one lag, while their correlations initially appeared to be strong. Especially for the money supply indicators an explanation might be the difference between the randomly selected estimation sample (until 1999:12) and the correlation sample (until 2018:12), as inflation was not affected by the increased money supply by the Federal Reserve after 2007 (Appendix A.1). Binner, Tino, Anderson, George and Kendall (2010) also do not provide support for using monetary aggregates to forecast inflation until the mid-2000s. Lastly, if lags in general are considered, one notices that the first and the tenth lag are included often, while lags five and eight are not included at all. Since the first lag captures most recent information it is straightforward that it is included, but the question related to the rationale behind the inclusion or exclusion of some of the other lags remains to be answered for now.

To ensure that the low order VAR(1) is treated similarly, the insignificant coefficients of this estimation are restricted as well. The lag variables included after restrictions are similar to the ones in Table 4, CPI and PPI. Although this would suggest that for the low lag order estimations the differences between both categories would be minimal, both categories will be used for forecasting, since their estimation is initially based on different indicators.

[4] The restriction matrices for the residual lags are available upon request (also in format of tables 4A and 4B).

Variance decomposition

Besides CPI, PPI and Unemployment rate, all other indicators provide only one lag coefficient in the estimation of CPI and therefore the explanatory power of these indicators might be questioned. To determine the relative importance, the Cholesky variance decomposition of CPI is shown in Table 5 for twelve periods (horizons). Figure 2 denotes a visual representation of table 5, including a line for the SE of the low order variance decomposition in table 6.

Table 5 CPI Variance decomposition (high lag order) – Category E and M, respectively

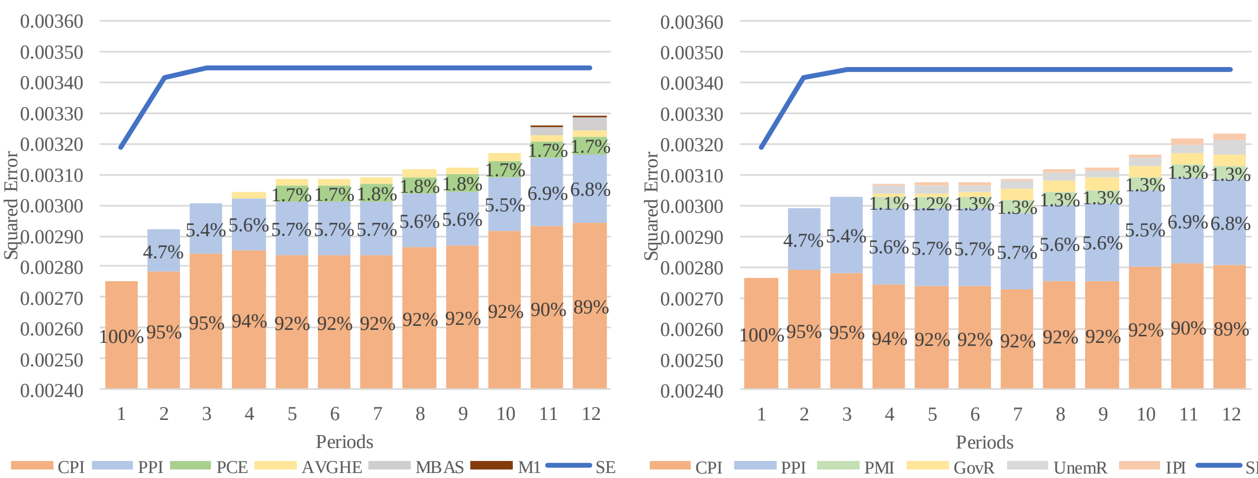
Note: SE represents the total forecasting error and the percentages provide a decomposition of the SE.

Period	SE	CPI	PPI	PCE	AVGHE	M1	MBAS
1	0.00275	100.0%					
2	0.00292	95.3%	4.7%				
3	0.00300	94.6%	5.4%				
4	0.00304	93.7%	5.6%		0.6%		
5	0.00308	92.0%	5.7%	1.7%	0.6%		
6	0.00308	92.0%	5.7%	1.7%	0.6%		
7	0.00309	91.8%	5.7%	1.8%	0.7%		
8	0.00312	91.9%	5.6%	1.8%	0.7%		
9	0.00312	91.9%	5.6%	1.8%	0.7%		
10	0.00317	92.1%	5.5%	1.7%	0.7%		
11	0.00326	89.9%	6.9%	1.7%	0.7%	0.1%	0.8%
12	0.00329	89.5%	6.8%	1.7%	0.7%	0.2%	1.3%

Period	SE	CPI	PPI	GovR	IPI	UnemR	PMI
1	0.00277	100.0%					
2	0.00299	93.2%	6.8%				
3	0.00303	91.9%	8.1%				
4	0.00307	89.3%	8.2%	0.4%	0.2%	0.8%	1.1%
5	0.00308	89.0%	8.2%	0.4%	0.3%	0.8%	1.2%
6	0.00308	88.9%	8.2%	0.4%	0.3%	0.8%	1.3%
7	0.00309	88.3%	8.1%	1.2%	0.3%	0.8%	1.3%
8	0.00312	88.3%	8.0%	1.3%	0.3%	0.8%	1.3%
9	0.00312	88.2%	8.1%	1.3%	0.3%	0.8%	1.3%
10	0.00317	88.5%	7.9%	1.3%	0.3%	0.8%	1.3%
11	0.00322	87.4%	8.7%	1.2%	0.6%	0.8%	1.3%
12	0.00324	86.8%	8.6%	1.2%	0.7%	1.4%	1.3%

Figure 2 CPI Variance decomposition – Category E and M, respectively

Note: the SE line refers to the total SE of low order lag, the stacked bar charts represent the decomposed SE of high order lags



Basically, these tables and figures show how each innovation or shock of a certain indicator is affecting the forecasting errors of CPI[5]. One can observe that while the forecasting error increases with the horizon, the relevance of CPI remains high and stable on an absolute level, but decreases relatively as the other indicators are included as well. The fact that the forecasting errors of both categories are quite similar, provides another indication that statistically differentiating Category E and M forecasts might become difficult. Nonetheless, one could argue that the other indicators of Category M provide more information

[5] The variance decompositions of the other indicators are available upon request.

relative to the other indicators of Category E. The relative importance of PPI is an example of that, as it accounts for 4.8%-6.9% of the variation in combination with Category E, while it accounts for 6.8%-8.7% of the variation with Category M. Finally, this decomposition also shows the effect indicators have on CPI through other indicators. For example, M1 is completely excluded in the Category E estimation, yet in periods 11 and 12 it is marginally responsible for the variation of CPI.

With respect to the Cholesky variance decomposition of the low lag order estimation in Table 6 and Figure 2, one can observe that the CPI forecasting error is increasing sharply after the first horizon and then becomes constant after a few horizons. On every horizon, the forecasting errors of the low lag order estimations are higher than those from the high lag order, which is most likely related to the forecasting approach used in the Cholesky variance decomposition. The other indicators of the estimations are restricted in the only lag included. As a result, restricted indicators cannot affect the CPI variance through other indicators, like they could in the high lag order estimation. Finally, where in the high lag order specification the other indicators of Category M have more relative importance with respect to Category E, in the low lag order specification the relative importance of the other indicator of Category E is larger.

Table 6 CPI Variance decomposition (low lag order)

Note: SE represents total forecasting error, indicators without influence on CPI are left out of this table.

	Category E			Category M		
Period	SE	CPI	PPI	SE	CPI	PPI
1	0.003191	100.0%		0.003190	100.0%	
2	0.003416	96.4%	3.6%	0.003416	96.6%	3.4%
3	0.003446	95.6%	4.4%	0.003443	96.0%	4.0%
4	0.003449	95.5%	4.5%	0.003446	95.9%	4.1%
5	0.003450	95.5%	4.5%	0.003446	95.9%	4.1%
6	0.003450	95.5%	4.5%	0.003446	95.9%	4.1%
7	0.003450	95.5%	4.5%	0.003446	95.9%	4.1%
8	0.003450	95.5%	4.5%	0.003446	95.9%	4.1%
9	0.003450	95.5%	4.5%	0.003446	95.9%	4.1%
10	0.003450	95.5%	4.5%	0.003446	95.9%	4.1%
11	0.003450	95.5%	4.5%	0.003446	95.9%	4.1%
12	0.003450	95.5%	4.5%	0.003446	95.9%	4.1%

Overall, this section has been able to establish four restricted multivariate estimations for forecasting; one high and low order VAR for both Category E and Category M. As is common in the forecasting literature, the forecasts will also be benchmarked to an autoregressive (AR) model. Like Faust and Wright (2013), both high and low lag orders could be benchmarked to an AR(1) only, but as one of the objectives of this paper is to test forecasting approaches for both lag orders, an unrestricted AR(12) will be included as well. The specification, following the notation of Marcellino et al. (2006) is as follows,

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \varepsilon_t, \qquad (3)$$

where y_t denotes the CPI at time t, ϕ_i denotes the coefficient for lag p, and α and ε_t denote the constant and the error term, respectively.

5. Forecasts and Multi-horizon comparison

Where the previous section focuses on the estimation component of inflation forecasting, this section will build upon that and focus on the actual forecasting approaches and more importantly, the forecast evaluations. The forecasting results will be presented in a pairwise multi-horizon comparison, which provides a statistical test on the superior predictive abilities forecast might have with respect to each other.

Like Marcellino et al. (2006), this paper will use an iterative and a direct approach to forecast inflation, but we will expand on their research by imposing a multi-horizon comparison. In addition, where the authors use an expanding window size, we will use fixed rolling window estimation, since that is required for the validity of the multi-horizon comparison (Quaedvlieg, 2019). The starting estimation window ranges from 1974:01 until 1999:12, equal to 312 observations. The estimation window is then iterated forward in steps equal to one month until the end of the sample at 2018:12, which is equal to 229 steps. For every estimation window, the same restriction matrix, as discussed above, will be used to restrict the coefficients. In forecasting multivariate estimations, the other indicators are forecasted as well, but since the focus of this paper is on CPI only, these results will be discarded after forecasting. Lastly, CPI will be forecasted for monthly horizons *h = 1, ..., 12*, in accordance with Breitung and Knüppel (2018), who show in their paper that key macroeconomic indicator forecasts are hardly informative when forecasted beyond two to four quarters.

The iterative approach is a one period ahead forecast, which is iterated forward until the desired horizon. Equation (2) captures the rolling window estimation for the iterative forecast. The estimated constant $\hat{\alpha}$ and coefficients $\widehat{\Phi}^{R}_{c,i}$ from this equation are plugged into the iterative forecast in combination with the forecasted value of the previous horizon, as shown in equation (4);

$$\hat{Y}^{I}_{c,t+h} = \hat{\alpha} + \sum_{i=1}^{p} \widehat{\Phi}^{R}_{c,i} \hat{Y}^{I}_{c,t+h-i}, \qquad (4)$$

where $\hat{Y}^{I}_{c,t+h}$ denotes the iterative forecasted indicator based on category c for horizon h, $\hat{Y}^{I}_{c,t+h-i}$ denotes the forecasted value of the previous horizon, which for $h = 1$ is equal to the actual values.

The direct forecast on the other hand, requires a horizon-specific model to be estimated, where the dependent variable is equal to the desired horizon;

$$Y^{h}_{c,t+h} = \beta + \sum_{i=1}^{p} R_{c,i} \mathrm{P}^{h,R}_{c,i} Y_{c,t+1-i} + \varepsilon_{t+h}, \qquad (5)$$

where $Y^{h}_{c,t+h}$ denotes the estimation of the indicator at horizon h for category c, $R_{c,i}$ denotes the restriction matrix, which is the same matrix as for the iterative estimation forecast in equation (2), $\mathrm{P}^{h,R}_{c,i}$ denotes the restricted estimated coefficient matrix for horizon h and $Y_{c,t+1-i}$ denotes the indicator at time $t + 1$ on which the estimation is based. Finally, β and ε_{t+h} are respectively the constant and the error term. The estimated constant $\hat{\beta}$ and coefficients $\hat{\mathrm{P}}_{c,i}^{h,R}$ are then put into the direct forecast model for this specific horizon as follows,

$$\hat{Y}_{c,t+h}^{D,h} = \hat{\beta} + \sum_{i=1}^{p} \hat{\mathrm{P}}_{c,i}^{h,R}\, Y_{c,t+1-i}, \tag{6}$$

where $\hat{Y}_{c,t+h}^{D,h}$ denotes the direct forecasted indicator for horizon h based for category c.

Finally, the univariate autoregressive estimation and iterative and direct forecasting specifications follow the approach of Marcellino et al. (2006). Equation (3) shows the estimation of coefficient ϕ_i, which is implemented in the following iterative forecast;

$$\hat{y}_{t+h}^{I} = \hat{\alpha} + \sum_{i=1}^{p} \hat{\phi}_i\, \hat{y}_{t+h-i}^{I}, \tag{7}$$

where $\hat{y}_{t+h}^{I}$ denotes the iterative forecasted value of CPI at horizon h and $\hat{y}_{t+h-i}^{I}$ denotes the forecasted values of the previous horizon(s), which for $h = 1$ is equal to the actual value of CPI. For the direct approach, the univariate horizon specific specification is as follows;

$$y_{t+h}^{h} = \beta + \sum_{i=1}^{p} \rho_i^{h}\, y_{t+1-i} + \varepsilon_{t+h}, \tag{8}$$

where y_{t+h}^{h} denotes CPI at horizon h, ρ_i^{h} denotes the coefficient for horizon h at lag p and β and ε_{t+h} are respectively the constant and the error term at horizon h. Next the estimated coefficient $\hat{\rho}_i^{h}$ is implemented in the following direct forecast;

$$\hat{y}_{t+h}^{D,h} = \hat{\beta} + \sum_{i=1}^{p} \hat{\rho}_i^{h}\, y_{t+1-i}, \tag{9}$$

where $\hat{y}_{t+h}^{D,h}$ denotes the direct forecasted value of CPI at horizon h, based on the estimated coefficient and the actual value of CPI, at that specific lag.

Multi-horizon comparison

Various macroeconomic indicators, estimation models and forecasting approaches are combined in the forecasting literature and then evaluated on individual horizons. However, as these factors both individually and combined have their effect on forecasting performance, establishing solid conclusions can be difficult. Quaedvlieg (2019) has developed a statistic which evaluates the performance of a forecast (A) with respect to another forecast (B) on all horizons. He introduces average superior predictive abilities (aSPA), where forecast A is outperforming B on average on all horizons, and uniform superior predictive abilities (uSPA), where forecast A is outperforming B at each individual horizon. By applying this statistic this study might be able to show that with certain parameters a few approaches can be discarded as there is another approach which will forecast more accurately on all horizons. This paper uses the pairwise superior predictive ability test in MATLAB as made available by Quaedvlieg on his website.

Forecasting different horizons unfortunately results in an unbalanced sample. Following the intentions to use the most recent data available, the estimated forecasts from 2000:01 to 2000:11 will be discarded, which leaves an out-of-sample forecasting period ranging from 2000:12 to 2018:12, equal to 217 monthly observations for all twelve horizons. Following Quaedvlieg (2019), the squared forecasting error for a forecast on every horizon can now be computed with respect to actual CPI as follows;

$$l_{i,t}^{h} = (y_t - \hat{y}_{i,t}^{h})^2, \tag{10}$$

where, $l_{i,t}^{h}$ captures the squared error for horizon h at time t for forecast i. Eventually, when one combines the time series of squared errors for every horizon of a forecast, a matrix of 217 x 12 (t x h) is constructed which is denoted as $L_{i,t}$. Table 7 provides an overview of the twelve matrices which have been constructed, where the univariate AR CPI forecasts will serve as benchmark to the multivariate VAR CPI forecasts. To compare two forecasts, their loss differential is determined as follows;

$$d_{ij,t} = L_{i,t} - L_{j,t}, \tag{11}$$

where, $d_{ij,t}$ captures the loss differential of forecast j with respect to forecast i. As both aSPA as uSPA are one-sided tests, the p-value of loss differential $d_{ij,t}$ can only determine whether forecast j has SPA over forecast i. If the p-value is insignificant, reversed loss differential $d_{ji,t}$ might show that it is actually forecast i which has SPA over j. Since this would imply 12 x 12 = 144 loss differentials and p-values to be estimated and evaluated, we decide to compare the forecasting performance with similar lag orders first.

Table 7 – Overview of forecasts for comparison

Low lag orders	High lag orders
AR(1) - Iterative	AR(12) - Iterative
AR(1) - Direct	AR(12) - Direct
VAR(1) - Iterative Category M	VAR(12) - Iterative Category M
VAR(1) - Iterative Category E	VAR(12) - Iterative Category E
VAR(1) - Direct Category M	VAR(12) - Direct Category M
VAR(1) - Direct Category M	VAR(12) - Direct Category E

Like Quaedvlieg (2019), this paper sets the block length to $\ell = 3$ and uses $B = 999$ bootstrap re-samples. The statistic provides the possibility for the aSPA to add a weighting scheme to customize horizon importance to an individual's preferences, but for now equal weights will be applied to all. Tables 8A and 8B show the high lag order p-values of the one-sided tests for respectively aSPA and uSPA and tables 9A and 9B show the same, but for the low lag order. To illustrate the table and how the p-values should be interpreted, an example from the high lag order will be discussed. The aSPA p-value of Direct Category M with respect to Iterative Category M is 0.9920, which would initially imply that Direct Category M is definitely no worse than the benchmark Iterative Category M. When the reversed loss differential is considered for these forecasts, the p-value is equal to 0.0080, which implies that Iterative has aSPA over

Direct for Category M. Next, we move to the same spot in the uSPA table and observe a p-value of 0.0641, implying uSPA is absent. Ultimately, one can argue that when using high lag orders for Category M, the iterative approach is on average outperforming the direct approach.

Table 8A – p-values aSPA – High lag order

Note: A p-value < 0.025 indicates that forecast *j* has aSPA w.r.t. *I*

i \ *j*		Iterative			Direct		
		AR	**E**	**M**	**AR**	**E**	**M**
Iterative	**AR**	--	0.0150	0.0160	0.1882	0.9890	0.9700
	E	0.9710	--	0.8599	0.9670	0.9900	0.9930
	M	0.9790	0.1391	--	0.9710	0.9960	0.9920
Direct	**AR**	0.7798	0.0200	0.0290	--	0.9880	0.9840
	E	0.0140	0.0100	0.0070	0.0100	--	0.0891
	M	0.0160	0.0090	0.0080	0.0130	0.9209	--

Table 8B – p-values uSPA – High lag order

Note: A p-value < 0.025 indicates that forecast *j* has uSPA w.r.t. *I*

i \ *j*		Iterative			Direct		
		AR	**E**	**M**	**AR**	**E**	**M**
Iterative	**AR**	--	0.4965	0.0751	0.6066	0.9359	0.9399
	E	0.9299	--	0.7868	0.9019	0.9600	0.9560
	M	0.8879	0.8258	--	0.9309	0.9520	0.9610
Direct	**AR**	0.8128	0.4895	0.0893	--	0.9399	0.9469
	E	0.3403	0.0010	0.0000	0.0080	--	0.6567
	M	0.0691	0.8238	0.0641	0.2392	0.8348	--

Table 9A – p-values aSPA – Low lag order

Note: A p-value < 0.025 indicates that forecast *j* has aSPA w.r.t. *I*

i \ *j*		Iterative			Direct		
		AR	**E**	**M**	**AR**	**E**	**M**
Iterative	**AR**	--	0.6036	0.8619	0.0390	0.0350	0.1011
	E	0.3874	--	0.9790	0.0250	0.0220	0.0801
	M	0.1361	0.0180	--	0.0190	0.0180	0.0691
Direct	**AR**	0.9590	0.9730	0.9790	--	0.5936	0.9770
	E	0.9650	0.9800	0.9930	0.4214	--	0.9940
	M	0.8819	0.9099	0.9339	0.0190	0.0110	--

Table 9B – p-values uSPA – Low lag order

Note: A p-value < 0.025 indicates that forecast *j* has uSPA w.r.t. *I*

i \ *j*		Iterative			Direct		
		AR	**E**	**M**	**AR**	**E**	**M**
Iterative	**AR**	--	0.8569	0.8649	0.9499	0.8899	0.9139
	E	0.9429	--	0.9409	0.9590	0.9209	0.9289
	M	0.9009	0.8529	--	0.9550	0.9119	0.9299
Direct	**AR**	0.8569	0.8208	0.8238	--	0.8158	0.9309
	E	0.8839	0.8949	0.8849	0.8509	--	0.8969
	M	0.8719	0.8589	0.8729	0.9009	0.7918	--

First, the univariate autoregressive forecasts will be discussed. Marcellino et al. (2006) argue for this category that direct forecasts with low lag orders should perform better, while iterative forecasts should perform better with high lag orders. With respect to the horizon, the authors argued that for short horizons direct forecasts should be used and for long horizons iterative forecasts should be used. When the direct and iterative AR forecasts are compared in the tables above, one can observe that on a multi-horizon comparison both iterative and direct forecasts do not perform worse with respect to each other. Therefore, one could argue that the difference between shorter and longer horizons, as observed by Marcellino et al. (2006), has some substantial influence on the optimal forecasting approach.

Second, the multivariate forecasts with high lag orders will be considered. Marcellino et al. (2006) found for their bivariate VARs that the iterative approach is favored with high lag orders and long horizons, while with short horizons the direct approach is preferred. In the multi-horizon comparison of this study one can observe that the multivariate iterative forecasts are performing very well, since they have aSPA over almost all other forecasts, including their benchmark (Iterative AR). Additionally, one can observe that the multivariate iterative forecasts have uSPA with respect to the direct Category E forecast, but not to the Category M forecast. Furthermore, although the forecasts from Category E and M do not show statistically significant outperformances in pairwise comparisons, the uSPA other models have over Direct Category E and not over Category M, would argue that the forecasting abilities of Direct Category M are slightly worse. Overall, one can conclude that if CPI is forecasted with a high lag order, an iterative approach will provide superior predictive abilities on average compared to similar direct forecasts.

Third, the multivariate forecasts with low lag orders will be considered. Marcellino et al. (2006) concluded with respect to their bivariate VARs that the direct approach is favored with low lag orders and short horizons, while the iterative approach is favored for long horizons. With respect to the high lag orders in table 8, one can clearly observe that the direct forecasts now have aSPA compared to most iterative forecasts, but evidence for uSPA is absent. Furthermore, the direct forecast for Category E, which was the weakest forecast for the high lag orders, now appears to be one of the strongest. This direct multivariate forecast also has aSPA with respect to the direct Category M forecast, which is actually quite surprising as both estimations have been restricted to CPI and PPI only. Even in the iterative forecasts, Category E has aSPA compared to Category M. Apparently, the slight difference in coefficient estimation caused by the initial estimations were sufficient to establish these aSPA. Overall, one can conclude for low lag orders that a univariate AR(1) or a multivariate with Category E inflation forecast should on average be executed using a direct approach. Additionally, one can argue that in forecasting inflation with Category E variables, a direct approach is more suitable.

Although this paper might provide useful insights for selecting a forecasting approach to forecast inflation, what the optimal lag length to forecast inflation should be, has not been assessed yet. Therefore, the two strongest forecasts from the high and low lag orders will be selected to compute their loss differentials and to test if superior predictive abilities are present in a multi-horizon comparison. The low lag orders will be represented by the direct univariate AR(1) and multivariate Category E forecasts and the high lag orders will be represented by both iterative multivariate forecasts. The results, shown in tables 10A and 10B, are striking, as multivariate iterative inflation forecasts with high lag orders have uSPA over the best low lag order direct forecasts. This finding is in line with the Cholesky variance decomposition of Figure 2, where the low lag order VAR has higher squared errors for both categories on all horizons. So, if you would like to forecast inflation through a multivariate VAR, you would be advised to use a high lag order with iterative forecasting.

Table 10A – p-values aSPA- High vs low lag

Note: A p-value<0.025 indicates that forecast *j* has aSPA w.r.t. *i*

i \ *j*		Low		High	
		AR	**E**	**E**	**M**
Low	**AR**	--	0.5936	0.0050	0.0080
Low	**E**	0.4214	--	0.0090	0.0040
High	**E**	0.9890	0.9930	--	0.8599
High	**M**	0.9920	0.9870	0.1391	--

Table 10B – p-values uSPA – High vs low lag

Note: A p-value<0.025 indicates that forecast *j* has uSPA w.r.t. *i*

i \ *j*		Low		High	
		AR	**E**	**E**	**M**
Low	**AR**	--	0.8158	0.0030	0.0010
Low	**E**	0.8509	--	0.0230	0.0030
High	**E**	0.9790	0.9860	--	0.7868
High	**M**	0.9720	0.9830	0.8258	--

6. Discussion

Despite some striking results which have been presented, several limitations to this study should be explained, some of which might be interesting for future research to consider. Related to seasonality there are two caveats. First, it should be noted that not all data has been seasonally corrected in the exact same way, as for PMI and PCE non-seasonally adjusted data was not available. Nonetheless, both are included in the high lag order VAR with one lag, which provides some assurance that these were not discarded because of this difference. Second, this study has not been able to fully extract CPI seasonality, because year-on-year growth rates combined with a twelve-month lag causes the actual CPI value a year ago to be partially included again. Eventually, this might provide an explanation on why CPI is the only indicator with a twelve-month lag, but since the ten-month lag is included as well it might also be an informative lag for other reasons.

Moreover, when using rolling regressions every new sample is regressed and restricted, but in this paper the restriction matrix is always the same, based on the starting sample until 1999:12. As these restrictions had to be manually imposed on every coefficient and in multiple steps to follow the GETS approach it was not feasible to do this for every estimation. As a result, insignificant coefficients might be estimated for various rolling windows, for which this paper had the intention to restrict them. Nonetheless, this study is confident that the effect of including a few of these insignificant estimators will be marginal. For example, the CPI and PPI lags were selected for both categories regardless of the other indicators and these two indicators explain 95-100% of the CPI variance when it is decomposed. Future research could consider to apply Bayesian variable selection using stochastic search and to explore whether forecasting accuracy improves if every step specific step of the rolling regression is individually restricted.

Additionally, like in Marcellino et al. (2006), one could include, restrict and forecast an estimation which ranges to $p = 4$. The current restrictions in this paper on the low lag order with $p = 1$ might be interpreted as quite stringent. It would be interesting to observe if direct forecasts for Category E keep their relative strength with slightly higher lag orders. Future research could also look into the performance of unrestricted multivariate estimations, as restricting can be a time-consuming approach depending on the number of lags.

Furthermore, in addition to the pairwise comparisons, Quaedvlieg (2019) provides an extension on the Model Confidence Set (MCS) as introduced by Hansen, Lunde and Nason (2011). It is decided not to apply this methodology here, since it will show if the best selected models have aSPA or uSPA with respect to the other models with a given level of confidence. Yet through pairwise testing it has already been established that the high lag order multivariate iterative forecasts perform the best and that Category E or M are not significantly better in forecasting with respect to each other. The additional value of performing an MCS, which requires strong computational power is therefore relatively low.

Finally, as argued by Pesaran et al. (2009) it appears that financial variables like interest rates and real equity prices have the ability to improve forecasting accuracy. A limitation of the approach with one indicator from each category, is that the real equity prices have not been included in the multivariate estimation, as it was decided to use interest rates. McCracken and Ng (2016) provide an overview of the other classifications of macroeconomic indicators, where timeseries are classified into eight or fourteen categories for instance. Future research could use the results from this paper and replace a variable like M1 for real equity prices, as M1 is entirely restricted and has marginal impact on CPI. Same holds for GDP, but since it is not available monthly, it would impose a sampling frequency issue. One could decide to execute the entire study with quarterly data or use a MIDAS model to combine these mixed frequencies. Additionally, one could use this approach to forecast macroeconomic indicators like GDP in order to evaluate if indicators from other categories than E (as in Marcellino et al. (2006)), show different forecasting preferences. With respect to inflation forecasting, one could also perform a multi-horizon comparison on all models from Faust and Wright (2013) and the more recent machine learning models of Medeiros et al. (2019), to determine which models have superior predictive abilities and which ones can be discarded under certain conditions.

7. Conclusion

By combining the empirical forecasting problem stated by Marcellino, Stock and Watson (2006) and the recently introduced multi-horizon comparison by Quaedvlieg (2019), this study has been able to provide a contribution to the inflation forecasting literature. Although, the strong self-explanatory power of CPI is still clearly present in a multivariate setting, including macroeconomic indicators improves the forecasting performances considerably. Nonetheless, this paper is convinced that the indicator sets which are included are still suboptimal in forecasting inflation. These two categorical VARs served to show the various forecasting preferences caused by the macroeconomic indicators included in these categories. Differences between the inflation forecasts based on categories E and M exist, yet this paper shows that these differences are only significant using suboptimal forecasting approaches and that with optimal forecasts these differences are not present.

Finally, this is not the first and not the last paper written on inflation forecasting and hopefully the literature can move towards a conclusive set of macroeconomic indicators which should be included in a multivariate inflation forecast. The uniform superior predictive abilities with respect to the other forecasts, allow us to conclude the following; when one would like to forecast inflation with a restricted multivariate VAR, we would suggest a high lag order and an iterative forecasting approach. With that, we provide our contribution in this debate and would suggest that an optimal set of macroeconomic indicators to include should be the next contribution.

8. References

AeroWeb Database National Economic Trends (2019) Producer Managers Index – Manufacturing. Link to webpage, extracted May 21st 2019.

Aiolfi, M., Timmermann, A. (2006) Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, Vol 135, Vol 1-2, pp. 31-53

Atkeson, A., Ohanian, L.E. (2001) Are Phillips curves useful for forecasting inflation? *Federal Reserve bank of Minneapolis quarterly review*, Vol 25, Issue 1, pp. 2-11

Binner, J.M., Tino, P.J, Anderson, R., Jones, B.E., Kendall, G. (2010a) Does money matter in inflation forecasting? *Physica A: Statistical Mechanics and its Applications.* Vol. 389, Issue 21, pp. 4793-4808

Binner, J.M., Elger, T., Jones, B.E., Nilsson, B. (2010b) Inflation forecasting, relative price variability and skewness. *Applied Economics Letters* Vol. 17, Issue 6, pp. 593-596

Bullard, J. (2011) Measuring Inflation: The Core is Rotten. *Federal Reserve Bank of St. Louis Review* Vol.93 Issue 4, pp. 223-233

Breitung, J., Knüppel, M., (2018) How far can we forecast? Statistical tests of the predictive content *Deutsche Bundesbank Discussion Paper* Vol. 7, pp. 1 – 39.

Caporale, M.G., Katsimi, M., Pittis, N. (2002) Causality links between consumer and producer prices: some empirical evidence. *Southern Economic Journal* Vol. 68, Issue 3, pp. 703-771

Cavallo, A., Rigobon, R. (2016) The billion prices project: using online prices for measurement and research. *Journal of Economic Perspectives* Vol. 30, Issue 2, pp. 151-178

Chan, F., Pauwels, L. (2018) Some theoretical results on forecast combinations. *International Journal of Forecasting*, Vol. 34, Issue 1, pp. 64-74

Checchetti, S.G., Chu, R.S., Steindel, C. (2000) The unreliability of inflation indicators. *Current issues in economic and finance – Federal Reserve Bank of New York* Vol. 6 Issue 4, pp. 1-6

Clark, T.E. (1995) Do producer prices lead consumer prices? *Economic Review – Federal Reserve Bank of Kansas City* Vol. 80, pp. 25-39.

Faust, J., Wright, J.H. (2013) Forecasting Inflation. *Handbook of economic forecasting* Vol. 2, pp. 2-56. Elsevier.

Federal Reserve (2019) What is the money supply? Is it important? Link to webpage, extracted July 4th 2019.

George, E.I, Sun, D., Ni, S. (2008) Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, Vol. 142, Issue 1, pp. 553-580

Hansen, P.R., Lunde, A., Nason, J.M. (2011) The Model Confidence Set *Econometrica*, Vol. 79, No. 2, pp. 453 – 497

Harchaoui, T.M., Janssen, R.V. (2018) How can big data enhance the timeliness of official statistics? The case of the US consumer price index. *International Journal of Forecasting*, Vol. 34, Issue 2, pp. 225-234

Haubrich, J.G., Millington, S. (2014) PCE and CPI inflation, what's the difference? *Economic Trends – Federal Reserve Bank of Cleveland.* Link to webpage, extracted July 4th 2019.

Inoue, A., 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* Vol. 103, Issue 482, pp. 511-522

Komunjer, I., Owyang, M.T. (2012) Multivariate forecast evaluation and rationality testing. *The review of economics and statistics* Vol. 94 (4), pp. 1066-108

Koenig, E.F. (2002) Using the purchasing managers' index to assess the economy's strength and the likely direction of monetary policy. *Federal Reserve Bank of Dallas Economic and Financial Policy Review*, Vol. 1, Issue 6, pp. 1-14.

Korobilis, D. (2013) VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics* Vol. 28, Issue 2, pp. 204-230

Krolzig, H.M. (2000) General-to-specific reductions of vector autoregressive processes, *Econometric Studies: A Festschrift in Honour of Joachim Frohn*, pp. 129-157.

Lee, B.S. (1992) Causal relations among stock returns interest rates, real activity and inflation. *The Journal of Finance* Vol. 47 Issue 4, pp. 1591-1603.

Marcellino, M., Stock, J.H., Watson, M.W. (2005) A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics* Vol. 135, pp. 499-526

McCracken, M.W., Ng, S. (2016) FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics,* Vol. 34, Issue 4, pp. 574-589

McElroy, T. (2015) When are Direct Multi-step and Iterative Forecasts Identical? *Journal of Forecasting* Vol. 34, Issue 4, pp. 315-336

Medeiros, M.C., Vasconcelos, G.F., Veiga, A., Zilberman, E. (2019) Forecasting inflation in a data-rich environment: the benefits of machine learning methods, Journal of Business & Economic Statistics, (just-accepted), pp. 1- 45

Pesaran, M.H., Schuermann, T., Weiner, S.M. (2004) Modeling regional interdependencies using a global error-correcting macroeconometric model. *Journal of Business and Economic Stastics*, Vol. 22, Issue 2, pp. 129-162

Pesaran, M.H., Schuermann, T., Smith, L.V. (2009) Forecasting economic and financial variables with global VARs. *International journal of forecasting*, Vol. 25, Issue 4, pp. 642-675

Pesaran, M. H., Pick, A., Timmermann, A. (2011) Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics* 164, pp. 173-187

Quaedvlieg, R. (2019) Multi-Horizon Forecast Comparison. Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2019.1620074

Schorfheide, F., Song, D. (2015) Real-time forecasting with a mixed-frequency VAR. *Journal of Business and Economic Statistics*, Vol. 33, Issue 3, 366-380

Staiger, D., Stock, J.H., Watson, M.W. (1997) The NAIRU, unemployment and monetary policy. *Journal of economic perspectives*, Vol. 11, Issue 1, pp. 33-49

Stock, J.H., Watson, M.W. (1999) Forecasting Inflation *Journal of monetary economics*. Vol. 44, pp. 293-335

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

Stock, J.H., Watson, M.W. (2003) Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, Vol. 41, Issue 3, pp. 788-829

Stock, J.H., Watson, M.W. (2007) Why has US inflation become harder to forecast? *Journal of Money, Credit and banking*, Vol. 39, pp. 3 -33

Stock, J.H, Watson, M.W., (2008) *Phillips curve inflation forecasts* (No. w14322). National Bureau of Economic Research

Stock, J.H., Watson, M.W. (2010) *Modeling inflation after the crisis* (No. w16488). National Bureau of Economic Research

Stock, J.H., Watson, M.W. (2016) Core inflation and trend inflation. *Review of Economics and Statistics*, Vol. 98, Issue 4, pp.770-784

Thornton, D.L. (2011) Core Versus Headline Inflation Again. *Economic Synopses* No. 16

9. Appendix

Figure A.1 – Non-seasonally corrected Monetary Base and CPI

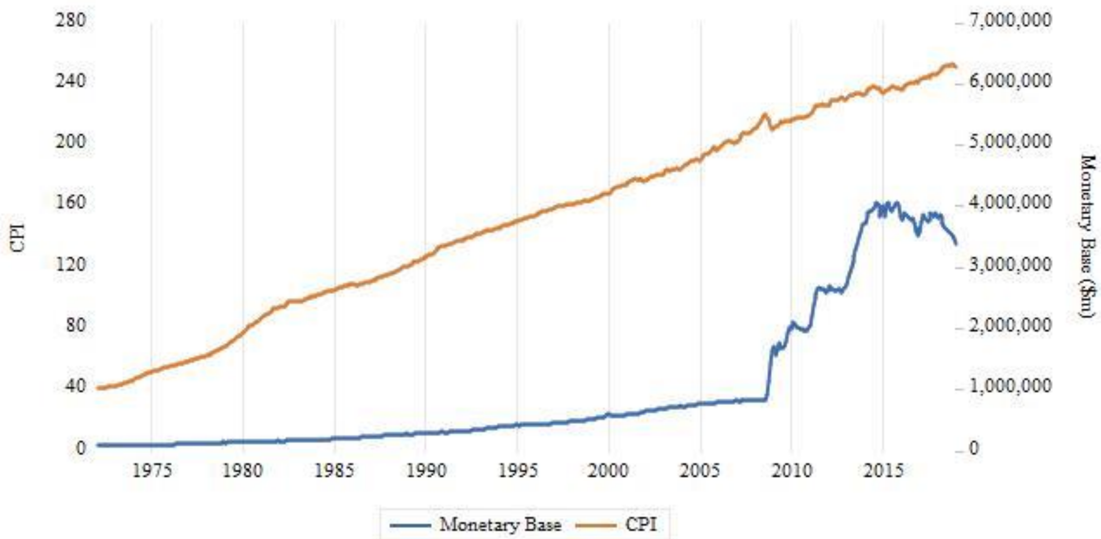


Table A.1A – Restriction Matrix – Low order VAR – Category M

Lag 1	CPI(-1)	PPI(-1)	GovR(-1)	IPI(-1)	UnemR(-1)	PMI(-1)
CPI	1	1	0	0	0	0
PPI	0	1	0	0	0	0
GovR	0	0	1	0	0	1
IPI	0	0	1	0	1	1
UnemR	1	1	0	1	0	1
PMI	1	0	0	1	0	0

Table A.1B – Restriction Matrix – Low order VAR – Category E

Lag 1	CPI(-1)	PPI(-1)	PCE(-1)	AVGHE(-1)	M1(-1)	MBAS(-1)
CPI	1	1	0	0	0	0
PPI	0	1	0	0	0	0
PCE	0	0	1	0	0	0
AVGHE	0	0	0	1	0	0
M1	1	0	0	1	1	0
MBAS	0	0	0	0	0	1

Table A.2A – Restriction Matrix – High order VAR – Category M

Lag 1	CPI(-1)	PPI(-1)	GovR(-1)	IPI(-1)	UnemR(-1)	PMI(-1)
CPI	1	1	0	0	0	0
PPI	0	1	0	0	0	0
GovR	0	0	1	0	0	1
IPI	0	0	0	0	1	1
UnemR	0	0	0	0	1	1
PMI	0	0	0	1	0	0

Note: CPI restrictions are the same as shown in Table 4, but just for the first lag.

Table A.2B – Restriction Matrix – High order VAR – Category E

Lag 1	CPI(-1)	PPI(-1)	PCE(-1)	AVGHE(-1)	M1(-1)	MBAS(-1)
CPI	1	1	0	0	0	0
PPI	0	1	0	0	0	0
PCE	0	0	1	0	0	0
AVGHE	0	0	0	1	0	0
M1	0	0	0	1	1	0
MBAS	0	0	0	0	0	1

Note: CPI restrictions are the same as shown in Table 4, but just for the first lag.