Weigh(t)ing the Basket: Aggregate and Component-Based Inflation Forecasts for the Euro Area[☆]

Jakub Chalmovianský[1], Mario Porqueddu[2,3], and Andrej Sokol[4]

[1]Masaryk University
[2]European Central Bank
[3]Bank of Italy
[4]Centre for Macroeconomics

February 11, 2022

Abstract

We compare direct forecasts of HICP and HICP excluding energy and food in the euro area and five member countries to aggregated forecasts of their main components from large Bayesian VARs with a shared set of predictors. We focus on conditional point and density forecasts, in line with forecasting practices at many policy institutions. Our main findings are that point forecasts perform similarly using both approaches, whereas directly forecasting aggregate indices tends to yield better density forecasts. In the aftermath of the Great Financial Crisis, relative forecasting performance was typically only affected temporarily. Inflation forecasts made by Eurosystem/ECB staff perform similarly or slightly better than those from our models for the euro area.

JEL classification: C11, C32, C53, E37
Keywords: inflation; forecasting; forecast evaluation; Bayesian VAR

[☆]Corresponding author: Andrej Sokol, sokolandrej@gmail.com.

Competing interests statement

Jakub Chalmovianský. Declaration of interests: none.

Mario Porqueddu. Declaration of interests: none.

Andrej Sokol. Declaration of interests: none.

1 Introduction

Inflation forecasts play a crucial role in both public and private sector decisions, and considerable effort is devoted to both their production and ex-post evaluation[1]. One interesting feature of the aggregate price indices that are commonly used to measure inflation, such as the HICP in the euro area or the PCE in the United States, is that they can be represented as weighted sums of their component indices.[2] This raises the natural question of whether it is better to forecast the aggregate index directly (direct or top-down approach), or to aggregate forecasts for its underlying components using appropriate weights (indirect or bottom-up approach).

When the data-generating process is known, theory predicts that aggregating component forecasts can improve forecasts of the aggregate relative to the direct approach (see for example Luetkepohl, 1984), but in practice the true data-generating process is unknown and has to be estimated. Thus, the relative performance of the two approaches is an empirical issue, and one to which the existing literature does not provide a univocal answer in the case of inflation. The question is not purely academic: many forecasting tools and processes at central banks and other institutions lean towards one or the other approach, and those choices warrant regular scrutiny in light of new empirical evidence.

In this paper we revisit the merits of bottom-up versus top-down forecasts of price inflation. We compare both point and density forecasts of HICP inflation (headline and excluding energy and food) from a joint model of the components, augmented with a number of additional predictors, to those obtained from a model of aggregate inflation augmented with the same set of predictors. We adopt a Bayesian VAR (BVAR) framework, as VARs provide a natural framework to model the joint dynamics of a relatively large number of time series such as ours. We focus on euro area forecasts and on forecasts for the five largest euro area economies, that is, France, Germany, Italy, Spain and the Netherlands. As in Bańbura et al. (2015), we evaluate forecasts conditioned on paths for some of the predictors, to be more directly comparable to real-world forecasting setups in central banks and beyond[3].

For both measures of inflation, point forecasts from the two models perform similarly, which is in line with the inconclusive findings of the existing literature. However, we also find that for density forecasts, the aggregate model performs slightly better at most horizons, confirming previous results for the US. Our main result for individual countries is that the aggregate model seems to perform as well as or slightly better than the component model for both inflation measures and both point and density forecasts. One notable exception are very short-term point forecasts, where the bottom-up model mostly yields better results. Overall, these results suggest

[1] For a more detailed discussion, see for example Nakamura (2005), Stock and Watson (2007), Koop and Korobilis (2012), Faust and Wright (2013), or Bermingham and D'Agostino (2014).

[2] This holds exactly for HICP, which is a Laspeyres index, while for PCE it involves an approximation, as the latter is a Fisher index, and thus non-additive in the component indices.

[3] Results for unconditional forecasts are available upon request.

that except perhaps for very short-term forecast (a few months ahead), there's no real loss in forecasting ability from a simpler forecasting approach that focuses directly on the aggregate indices.

We also assess the impact of the aftermath of the Great Financial Crisis on the forecasting performance of both models: for headline inflation and most forecast horizons, the aggregate model only becomes obviously better for a short set of rolling forecast windows centered around the 2011 period, whereas for HICP excluding energy and food, there is no clear change in ranking following the crisis, although longer-term forecasts from the component model appear to deteriorate relative to the top-down model during the so-called "missing inflation" period.

Finally, we also find that point forecasts for both euro area HICP and HICP excluding food and energy made by Eurosystem/ECB staff, which employ the same conditioning assumptions we use for our projections, are at least as accurate, and in the near term more accurate, than those from our models. That Eusystem/ECB staff projections are competitive with state-of-the art model-based forecasts based on a similar information set is in itself an important counter to the criticism they have sometimes been subjected to (see for example Darvas, 2018). Nevertheless, past inflation forecast errors were at times undoubtedly large and persistent (see Koester et al., 2021). Since our model-based density forecasts clearly highlight the large uncertainty surrounding inflation projections (even before the COVID-19 shock), a greater emphasis on such uncertainty in their communication might help better understand Eurosystem/ECB staff and other publicly available inflation forecasts going forward.

1.1 Related literature

The idea that aggregation of forecasts of components of a series could improve accuracy relative to a direct forecast has a long tradition. Theil (1954) argues that if the data-generating process (DGP) is known, the components contain at least as much information as the aggregate measure, and because the forecast errors of the components partially cancel out, forecast accuracy should improve; Rose (1977), Tiao and Guttman (1980), Kohn (1982), Luetkepohl (1984) provide corroborative results for a range of DGPs. However, as shown in Luetkepohl (1987), if the DGP is unknown and needs to be estimated, a common situation in practice, then the relative forecast accuracy of direct and bottom-up approach depends on the specifics of the underlying component series and aggregation method[4], and is thus an empirical rather than a theoretical issue.

[4]Hendry and Hubrich (2006) and Hendry and Hubrich (2011) investigate additional options for exploiting disaggregated information to forecast the aggregate measure: instead of aggregating forecasts for the components, they propose to include disaggregated information directly in the model for the aggregate using factor models. This idea has found fruitful ground in the GDP nowcasting literature (see for example Bok et al., 2018, and references therein).

Recent contributions focusing on inflation[5] usually compare the top-down approach with forecasts aggregated either over geographical space (e.g. individual country forecasts) or components of the aggregate price index, with mixed findings. A prominent study of spatial aggregation is Marcellino et al. (2003), where various AR, VAR, and dynamic factor models for inflation and other macroeconomic variables are evaluated for 11 EMU countries and the resulting aggregates, finding that aggregation of country forecasts can improve the accuracy of aggregate forecasts. Benalal et al. (2004) investigate both dimensions of aggregation for four large euro area countries, namely France, Germany, Italy, and Spain, and the five main HICP components. They conclude that in the longer run, the direct approach provides better forecasts for the euro area as a whole and for individual countries as well. On the other hand, for shorter horizons, the results are mixed: for the euro area as a whole, the bottom-up approach for HICP components is better, but spatial aggregation does not improve forecasting accuracy.

Hubrich (2005) finds that for the euro area, direct forecasts are more accurate for some horizons, while studies focusing on single European countries tend to find evidence in favour of the bottom-up approach, at least for some forecasting horizons (Reijer and Vlaar, 2006, Duarte and Rua, 2007, Moser et al., 2007). Bermingham and D'Agostino (2014) argue that such inconclusive findings could be due to short data samples for the euro area, but they themselves claim to find clear evidence in favour of the bottom-up approach for both the euro area and the United States. On the other hand, Espasa and Mayo-Burgos (2013) qualify such a claim, arguing that disaggregation as such does not improve aggregate forecasts, unless the common features shared amongst the components are also taken into account, for example by means of a factor structure. More recently, Des and Guentner (2016) also find that bottom-up forecasts of sector-level value-added deflators dominate direct forecasts, and attribute the difference to the competing models' behaviour during the Great Financial Crisis.

Ravazzolo and Vahey (2014) is one of the first papers to also evaluate the density forecast performance of direct and bottom-up approaches, finding that for the period preceding the Great Financial Crisis, bottom-up approaches performed invariably better in forecasting US PCE, a finding later also confirmed by Tallman and Zaman (2017). Further evidence favouring bottom-up approaches in terms of density forecast accuracy is provided by Mazur (2016) for Poland, and Cobb (2019) for France, Germany and the UK.

The rest of the paper is structured as follows: in Section 2 we describe our estimation strategy, the forecast evaluation tests that we run and the data we use. In Section 3 we discuss our results, and in Section 4 we conclude. We report results for HICP excluding food and energy in Appendix A, while detailed results for individual countries and some additional charts for the euro area can be found in a separate Online Appendix, available upon request.

[5]Forecast aggregation is of course not confined to price indices. For example, a classic issue are the relative merits of bottom-up versus top-down GDP forecasting approaches (see for example Anesti et al., 2017, Heinisch and Scheufele, 2018).

2 Methodology

To evaluate the relative merits of bottom-up versus top-down inflation forecasts, we compare joint models of high-level price index components (suitably aggregated) to models of the aggregate price index itself, both augmented with a set of additional predictors. We model the components jointly to avoid the loss of useful information for the aggregate forecast (as also argued in Espasa and Mayo-Burgos, 2013)[6]. VARs provide a natural framework to model the joint dynamics of a relatively large number of time series such as ours: Bayesian estimation addresses the ensuing curse of dimensionality[7], and Kalman filtering and smoothing techniques allow us to study conditional as well as unconditional forecasts (see for example Bańbura et al., 2015). We then apply a battery of standard tests to evaluate both point and density forecasts, and we also compare point forecasts to those made by Eurosystem/ECB staff.

2.1 Model specification

The baseline model we estimate, similar to Giannone et al. (2014), is a $VAR(p)$:

$$X_{i,t} = A_0 + A_1 X_{i,t-1} + A_2 X_{i,t-2} + \cdots + A_p X_{i,t-p} + e_{i,t} \tag{1}$$

where X_t collects the observations of N scalar variables at time t, A_0 is a vector of N constants, A_k are, $\forall k = 1, \ldots, p$, square matrices of size N collecting the other parameters of the model, and e_t is a vector of residuals of size $N \times 1$.

We estimate three versions of (1) that differ in the set and number of variables. An overview of each specification can be found in Table 1. The largest model, which we refer to as the Component model, features $N = 14$ variables and contains all five main HICP components. The Aggregate models feature $N = 11$ variables. Data are mostly monthly, with a few quarterly variables interpolated to monthly frequency. We estimate all models in (log) levels, and then compute the implied inflation rates.

We set the number of lags to $p = 13$, and we implement Bayesian shrinkage using the standard Minnesota prior centered on a random walk model (Litterman, 1986), complemented with the sum-of-coefficients modification for the autoregressive coefficients initially introduced by Doan et al. (1984). Thus, we shrink the parameters of our VAR system by controlling the scale of the prior covariance matrix through a hyperparameter, λ, and we also introduce a restriction on the A_k matrices that, for a general $VAR(p)$ model, shrinks a model specified in levels towards one

[6] We have also estimated separate VAR models of each individual price index component augmented with the same set of additional predictors, and found that the performance was invariably worse than that of the joint component model. This is in line with Cobb (2019) and Ravazzolo and Vahey (2014), who argue that such an approach ignores potentially useful cross-correlations among components.

[7] For a discussion, see Bańbura et al. (2010).

specified in first differences. This is referred to as inexact differencing, and the tightness of this prior is controlled by a second hyperparameter, μ.[8] Based on a grid search and on long-standing in-house use of the model, we set $\lambda = 1/22$ and $\mu = 1/(22 * 40)$.[9] For estimation, we rely on the BEAR Toolbox version 4.2 (see Dieppe et al., 2016), and forecasts conditional on assumptions about future paths of selected variables in the system (see Section 2.3) are obtained with the algorithm developed in Waggoner and Zha (1999).

2.2 Forecast evaluation

The evaluation of conditional forecasts presented in Section 3 is based on both point and density forecasts for data vintages from June 2005 to March 2019 and forecast horizons up to 36 months. As point forecasts we use the medians of the predictive distributions. For the component models, the implied aggregate index is computed for each draw using appropriate weights. We evaluate all forecasts against data outturns of the aggregate of interest (headline HICP or HICP excluding food and energy) from the latest vintage available in our dataset, i.e. the one corresponding to the ECB's March 2019 Macroeconomic Projection Exercise (more details on the data we use are provided in Section 2.3).

The main measures we use to compare forecasting performance are the root mean squared error (RMSE) for point forecasts, and average log predictive scores for density forecasts, defined as follows:

$$RMSE_h = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{y}_{r+h} - y_{r+h})^2} \tag{2}$$

$$l_h(y_{t+h}) = \frac{1}{R}\sum_{r=1}^{R}\log \hat{p}_r(y_{t+h} \mid y_t) \tag{3}$$

where R denotes the total number of vintages, h is the forecast horizon, $\hat{y}_{r+h}$ the predicted value at horizon h forecasted as of the last available observation for vintage r, y_{r+h} the corresponding data outturn, and $\hat{p}_r(y_{t+h} \mid y_t)$ denotes the estimated predictive density for the r-th data vintage and forecast horizon h, evaluated at the realization of the variable of interest y_{t+h}. For both measures we also evaluate the statistical significance of their differences, using the Diebold-Mariano (Diebold and Mariano, 1995) and Amisano-Giacomini tests (Amisano and Giacomini,

[8] For $\lambda = 0$, the data information is ignored and the posterior equals the prior. On the other hand, for $\lambda \to \infty$, the posterior is equivalent to the ordinary least square estimates; therefore, no prior information is taken into account. For $\mu = 0$, $I_n = A_1 + A_2 + \cdots + A_p$ and the specification of the model is equivalent to a VAR in first differences, whereas for $\mu \to \infty$ the prior becomes uninformative (diffuse).

[9] Optimising the values of λ and μ, as in Giannone et al. (2015), each time the model is re-estimated yields forecasts that are very similar to our main results.

2007).[10][11]

To further assess the calibration of the predictive densities, we also compute probability integral transforms (PITs), defined as

$$u_{r,t+h} = \int_{-\infty}^{y_{t+h}} \hat{p}_r(x \mid y_t)dx \equiv P_r(y_{t+h} \mid y_t) \tag{4}$$

and formally test for some of their 'desirable' attributes, using a similar set of tests as Rossi and Sekhposyan (2014), Ravazzolo and Vahey (2014), Korobilis (2017), or Cobb (2019). Specifically, we run uniformity tests (Kolmogorov-Smirnov and Anderson-Darling), independence tests (Ljung-Box test of serial correlation) and Berkowitz (2001)'s joint test of zero mean, unit variance and independence. Since for $h > 2$ independence is violated by construction, we follow suggestions from Diebold et al. (1998), Clements and Smith (2000), and Rossi and Sekhposyan (2014), split our sample of 56 recursive estimates into non-overlapping sub-samples and carry out inference separately for each such sub-sample, and then consolidate results using Bonferroni bounds.

2.3 Data

The price series we use are based on the Harmonized Index of Consumer Prices (HICP). Specifically, we use monthly series for headline HICP, HICP excluding food and energy prices, and the five main components of HICP: prices of unprocessed food, processed food, non-energy industrial goods, energy, and services. To compute forecasts for the aggregate measures from the main components we sum their weighted values using HICP weights available on an annual basis from the ECB's Statistical Data Warehouse (SDW) portal[12].

As for the additional variables used in our BVAR models, we include the producer price index (PPI), non-energy commodity prices (food and overall), the oil price, the EUR/USD and the nominal effective exchange rate, euro area real GDP, unit labor costs and compensation per employee[13], as detailed in Table 1. These variables were taken from the Eurosystem/ECB staff projection database, but are also publicly available. HICP data (excluding energy prices) are seasonally (not calendar) adjusted up to the March 2016 vintage, the more recent vintages are both seasonally and calendar adjusted. The rest of the monthly variables are not seasonally adjusted, while all quarterly variables are seasonally and calendar adjusted. We take logarithms

[10] In our computations we use both tests with the correction introduced by Harvey et al. (1997), which improves the small-sample properties of the underlying likelihood ratio test.

[11] Strictly speaking, both testing frameworks are not designed for recursive estimation schemes such as ours; however, Clark and McCracken (2013) show that disregarding this fact has negligible practical implications.

[12] For more information, see: https://sdw.ecb.europa.eu/browse.do?node=9691207

[13] GDP, compensation per employee and unit labor costs are only available at quarterly frequency. Since our model is specified at monthly frequency, we interpolated these series prior to estimation using Kalman filtering and smoothing techniques.

Table 1 Different BVAR specifications

Aggregate Models	HICP Overall index HICP All-items excl. food and energy
Component Model	HICP Unprocessed food HICP Processed food (incl. alcohol and tobacco) HICP Non-energy industrial goods HICP Services
All models	HICP Energy PPI (domestic sales, consumer goods industry) Unit labor costs (whole economy) Non-energy commodity prices: Food (in USD) Non-energy commodity prices (in USD) Nominal effective exchange rate Oil price (in USD) EUR/USD Exchange rate Compensation per employee Real GDP

of all variables entering our models.

In order to mimic the Eurosystem/ECB staff projections, we use real-time data as they were available to forecasters at the time of each projection exercise[14] from June 2005 to March 2019, yielding a total of 56 vintages. We adopt a recursive estimation strategy, so observed values for all variables and all vintages start in January 1997, and the last observation advances over time. Thus, we use data with at least 100 observations (for the oldest vintage), and up to 264 observations for the most recent vintage. Since availability of data vintages for HICP excluding food and energy and PPI consumer goods is limited (real-time data are available only starting from the March 2009 and December 2015 exercises, respectively), we use ex-post revised historical data for all the previous unavailable data vintages.[15]

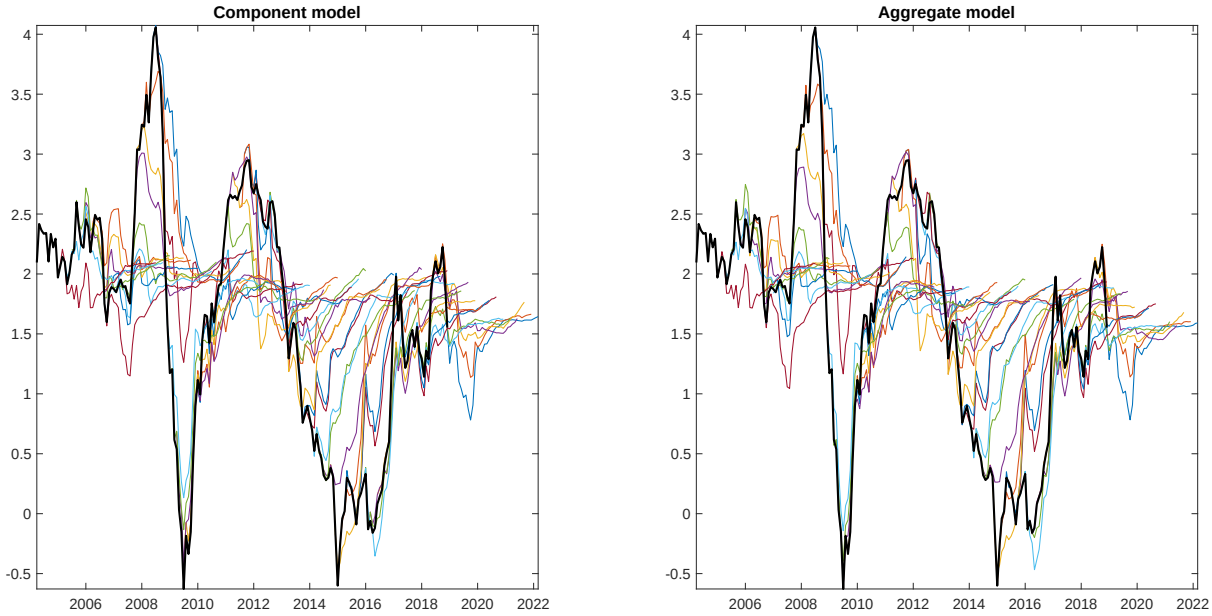
Conditional forecasts are based on assumptions about the future development of the following variables: HICP energy[16], non-energy food commodity prices, non-energy commodity prices, the nominal effective exchange rate, the oil price and the EUR/USD exchange rate. For each vintage, these assumptions match exactly those made in the corresponding Eurosystem/ECB staff projection exercise. Moreover, for certain vintages, flash estimates of HICP data are available and treated as data.

[14]The cut-off dates for data availability are therefore as published in the Eurosystem/ECB staff macroeconomic projections.

[15]This should not significantly affect our results since these data tend to be revised very little. For a detailed discussion of data revisions in the euro area, see Giannone et al. (2012)

[16]By conditioning to Eurosystem/ECB projections we incorporate information about indirect tax changes, refining and distribution margins and administrative prices.

Figure 1 Conditional forecasts of headline inflation over time



Note: The figures show consecutive forecast vintages of year-on-year inflation (colored lines) against the latest data vintage (solid black).

3 Results

Our main results for headline HICP inflation are reported in Section 3.1, with some additional charts reported in the Online Appendix. Results for HICP inflation excluding food and energy are discussed in Section 3.2, with charts and tables reported in Appendix A. Results for the rolling exercise are presented in Section 3.3, while results for individual countries are briefly discussed in Section 3.4, with most charts and tables relegated to the Online Appendix.

3.1 Headline inflation

Figure 1 provides an overview of conditional point forecasts of headline inflation up to 36 months ahead for all vintages in our sample. The Component and Aggregate models yield very similar projections, with a strong tendency to mean-reversion. The similarity is reflected also in their root mean squared errors (Table 2): RMSEs on year-on-year inflation for the Component model increase monotonically from 0.10 for 1-month-ahead forecasts to 1.18 for forecasts three years ahead, and the direct (Aggregate) and bottom-up (Component) projections are not statistically different according to Diebold-Mariano tests.

We also compare the point forecasts from both models to the projections made by Eurosystem/ECB staff as part of the (B)MPE projection exercises. These projections of course incorporate much more information than captured by the 11 to 14 time series in our models, as well

Table 2 Component vs Aggregate model conditional forecasts for headline inflation

Months ahead	RMSE *Component model*	relative RMSE *Component/Aggregate*	relative RMSE *Component/(B)MPE*
1	0.10	0.98	–
3	0.31	0.96	**1.53**
6	0.58	0.97	1.08
12	0.93	0.99	1.15
24	1.06	0.99	1.06
36	1.18	1.02	–

Note: A relative RMSE < 1 indicates that Component model forecasts are more accurate. Bold text denotes statistical significance of the difference at the 5% level, based on Diebold–Mariano tests with Harvey et al. (1997) correction. The comparison with Eurosystem/ECB staff (B)MPE projections (last column) is made on quarterly, rather than monthly, year-on-year inflation rates.

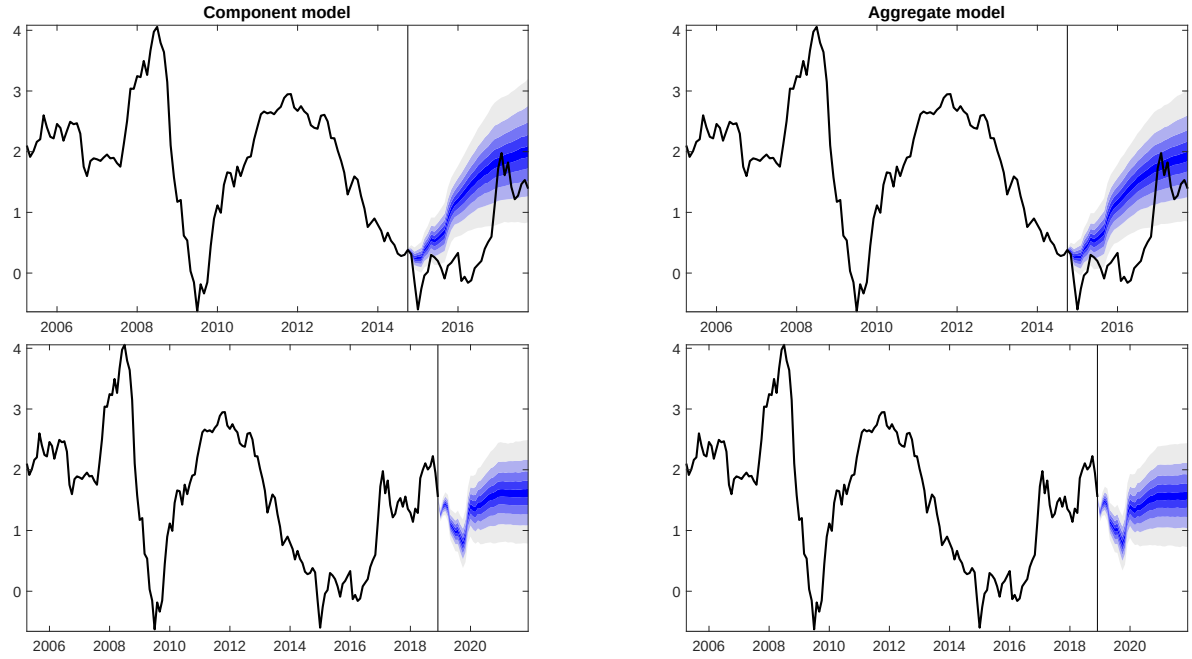
as expert judgement. Nevertheless, they are made conditional on the same set of assumptions about the evolution of certain variables over the forecast as our models, and are therefore a natural and interesting benchmark[17]. The last column of Table 2 shows relative RMSEs for the Component model and (B)MPE projections. Three months ahead, (B)MPE projections perform markedly better, and the difference is statistically significant. This should not come as a surprise, given the rich set of additional information and expert judgement that informs especially near-term projections. However, by the second quarter, this advantage seems to be much reduced, and the performance is similar, statistically speaking. The result that BVAR forecasts are competitive with professional forecasts that also incorporate expert judgement, such as those made by central banks, chimes with earlier findings by Angelini et al. (2019) (also for the euro area, but based on quarterly VARs for the four largest EA countries), Domit et al. (2019) (for the United Kingdom) and Iversen et al. (2016) (for Sweden).

Predictive densities from the two models can appear qualitatively similar for selected vintages. Figure 2 plots two examples: forecasts from the December 2014 vintage, at the beginning of a prolonged period of low inflation in the euro area (for a discussion, see Bobeica and Sokol, 2019), and from the last vintage in our dataset, i.e. based on the same assumptions as the March 2019 ECB staff projection exercise. However, a comparison of average log predictive scores (Figure 3) shows that up to 18 months ahead, predictive densities from the aggregate model outperform those from the bottom-up one by a margin, and the differences are statistically significant based on Amisano and Giacomini (2007) tests. The differences largely disappear for longer horizons. That is not surprising, given that conditioning assumptions tend to revert to the historical averages of the respective variables over longer horizons[18], and in their absence, the projections would converge to the respective models' unconditional predictive distributions, which we would expect to be similar. For shorter horizons, probability integral transforms for the aggregate

[17] For a more comprehensive analysis of Eurosystem/ECB projections, see Kontogeorgos and Lambrias (2019).

[18] Except for variables such as the exchange rate, which is simply held flat over the whole projection period, or the oil price, which reflects option prices.

Figure 2 Fan charts with conditional forecasts of headline inflation



Note: Top panel - December 2014 vintage; bottom panel - March 2019 vintage. The fan charts depict the evolution of selected quantiles of the predictive distributions of year-on-year inflation over the projection horizon: the darkest band is centered around the median, and the outer edges of the lightest bands correspond to the 0.05 and 0.95 quantiles.

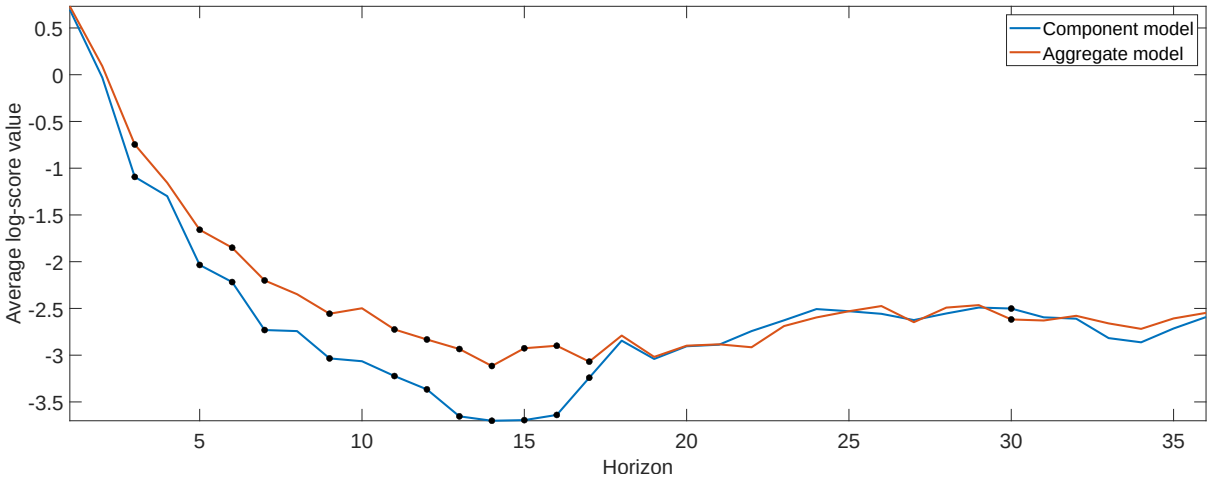
model also suggest somewhat better calibration of the predictive densities compared to the component model. Formal tests (Table 3) tend to reject null hypotheses of correct calibration more often for the component model than for the aggregate model, although rejections become the norm for both models at longer horizons.

3.2 HICP excluding energy and food inflation

For HICP excluding food and energy, a frequently used measure of underlying inflation that excludes some of the most volatile components from the headline index (see for example Ehrmann et al., 2018), point and density forecasts from the direct and bottom-up approaches are also qualitatively similar (Figures A.1 and A.2). Quantitatively, Table A.1 shows that point forecasts from the Aggregate model perform somewhat better than their component model counterparts, but the differences are typically not statistically significant[19]. The component model also performs similarly or worse than Eurosystem/ECB staff projections, although differences are invariably statistically insignificant in this case. Log predictive scores indicate better forecasting performance of the Aggregate model across all forecast horizons; the differences are small, but statistically significant for most horizons (Figure A.3). As for headline inflation, formal tests on

[19]Differences are statistically significant in a few selected months, not shown in Table A.1 to conserve space; the full results are available upon request.

Figure 3 Average log predictive scores for conditional forecasts - headline inflation



Note: A higher (less negative) log predictive score indicates better forecast accuracy. Black squares mark horizons for which the two scores are statistically different from each other based on Amisano and Giacomini (2007) with Harvey et al. (1997) correction.

Table 3 Calibration tests on Probability Integral Transforms for headline inflation models

	Aggregate model				**Component model**			
	Uniformity		*Independence*	*Joint* H_0	*Uniformity*		*Independence*	*Joint* H_0
h	*KS*	*AD*	*LB*	*Ber*	*KS*	*AD*	*LB*	*Ber*
1	0.670	0.442	0.964	0.525	**0.039**	**0.009**	0.832	**0.004**
3	**0.003**	**<0.001**	0.316	**<0.001**	**0.009**	**<0.001**	0.458	**<0.001**
6	**0.004**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
12	**0.003**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
24	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
36	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

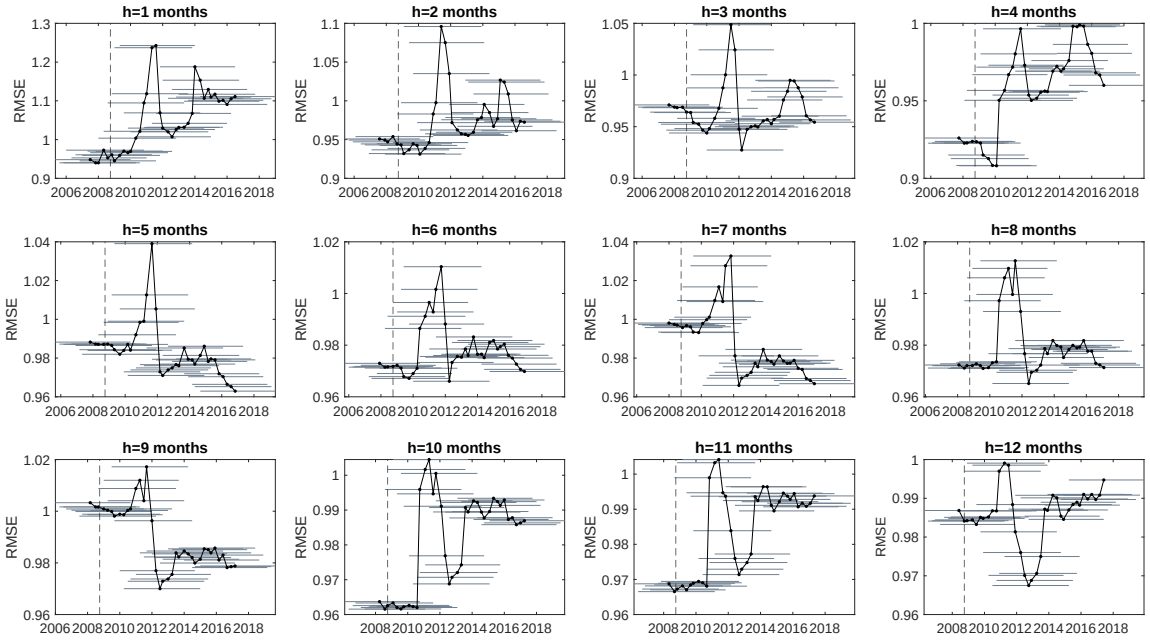
Note: P-values (for horizons $h > 2$, minimum p-values) of the respective test: Kolmogorov-Smirnov (KS), Anderson-Darling (AD), Ljung-Box of the mean (LB), Berkowitz (Ber). Bold text indicates rejection at the 5% significance level (for horizons $h > 2$, using Bonferroni bounds). See Section 2 for more details and references.

PITs (Table A.2) suggest somewhat better calibration of the aggregate model at short horizons, while at longer horizons the null hypothesis of correct calibration tends to be rejected for both models.

3.3 Rolling exercise

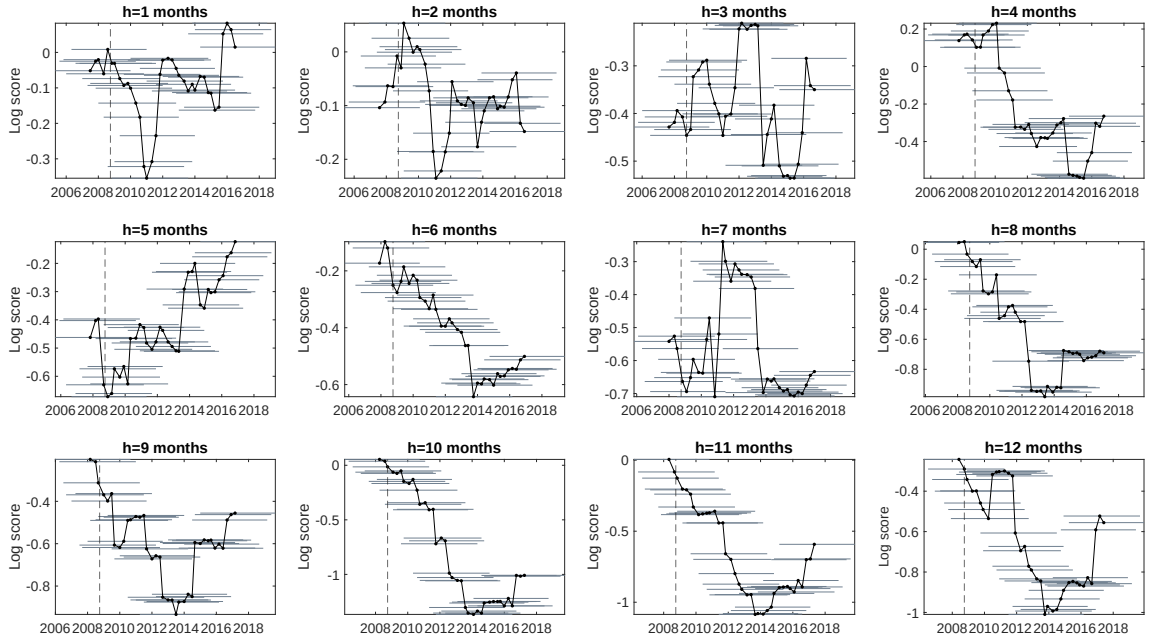
In order to investigate whether large aggregate shocks have any bearing on the performance of the direct vs. bottom-up approaches, we also report relative RMSE and log predictive scores over a rolling window of 5 years. Figure 4 shows the evolution of the relative RMSE between the component and aggregate models for headline inflation over our forecast evaluation sample. The most apparent feature that stands out across horizons is a marked, though temporary, improvement of the relative performance of the aggregate model in the windows centered around 2011. This is driven by a faster fall in the rolling RMSE of the aggregate model following the

Figure 4 Rolling relative RMSE (component vs aggregate model) by forecast horizon for headline inflation



Note: The black markers denote the center of the 5 year rolling window spanned by the light grey bars. The vertical dashed line denotes the onset of the Great Recession. A relative RMSE$> 1 (< 1)$ indicates better performance of the aggregate (component) model.

Figure 5 Rolling relative log scores (component vs aggregate model) by forecast horizon for headline inflation



Note: The black markers denote the center of the 5 year rolling window spanned by the light grey bars. The vertical dashed line denotes the onset of the Great Recession. A negative (positive) value indicates better performance of the aggregate (component) model.

Table 4 Relative RMSEs of Component vs Aggregate model conditional forecasts for headline inflation in individual countries

Months ahead	France	Germany	Italy	Netherlands	Spain
1	0.76	**0.73**	**0.53**	0.38	0.49
3	**0.86**	**0.87**	**0.65**	**0.85**	**0.59**
6	0.94	0.94	0.82	0.95	**0.87**
12	1.00	1.01	0.94	0.98	0.96
24	1.00	0.97	0.99	1.02	0.98
36	1.00	1.00	1.00	1.03	1.00

Note: A relative RMSE < 1 indicates that Component model forecasts are more accurate. Bold text denotes statistical significance of the difference at the 5% level, based on Diebold–Mariano tests with Harvey et al. (1997) correction.

large errors made during the Great Financial Crisis, and also indicates a somewhat better ability of the aggregate model to forecast inflation during the so-called "missing disinflation" episode (see for example Bobeica and Jarociński, 2019). A similar story holds for rolling logarithmic scores (Figure 5). Rolling results for HICP inflation excluding food and energy, reported in the Appendix, show two main features: a swift alignment of the performance of the two models following the Great Financial Crisis at short horizons; and a gradual deterioration of the relative performance of the component model at horizons longer than 9 months in the windows starting from those centered in early 2010, that is, shortly before the onset of, as well as during the "missing *inflation*" period that was one of the salient features of the euro area economy over the last decade (see Bobeica and Sokol, 2019).

3.4 Individual countries

Results for the five largest euro area economies are mixed, although some common threads across countries can be found. The bottom up approach usually yields better headline inflation point forecasts in the very short run, while at longer horizons the models again perform similarly (Table 4). On the other hand, the aggregate model delivers better, or very similar, density forecasts throughout (see Online Appendix). For HICP excluding energy and food inflation, point forecasts are typically not statistically different for the two approaches, with very few exceptions, where the aggregate model is superior (Table A.3). Log scores for HICP excluding energy and food inflation are typically very similar across models, with the notable exception of Italy, where the aggregate model clearly dominates the other approach for horizons beyond a year (see Online Appendix).

4 Conclusions

We have provided a comparison of direct conditional forecasts of HICP and HICP excluding energy and food in the euro area and five member countries to weighted sums of forecasts of their main components from large Bayesian VARs with a shared set of predictors. Our main finding is that point forecasts perform similarly using both approaches, whereas direct forecasts tend to yield better density forecasts. For individual countries, the aggregate model also tends to perform somewhat better, with a few exceptions, most notably short-term point forecasts for headline inflation. Where such a comparison is possible, we find that inflation forecasts made by Eurosystem/ECB staff perform similarly or slightly better than those from our models. The aftermath of the Great Financial Crisis seem to mostly have had temporary effects on the ranking between the aggregate and component models for either measure of HICP inflation.

Acknowledgements

The views expressed in this paper are those of the authors, and do not necessarily reflect those of the European Central Bank or the Bank of Italy. Chalmovianský was a PhD Trainee in the Prices and Costs Division at the European Central Bank during part of this project. Sokol was a member of staff of the European Central Bank during part of this project. We are grateful to Michele Lenza, seminar participants at the European Central Bank and Bank of England, and participants to the Cergy/Skema New Directions in Inflation Forecasting 2021 workshop for helpful comments and discussions.

References

Amisano, G. and R. Giacomini (2007): "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business & Economic Statistics*, 25, 177–190.

Anesti, N., S. Hayes, A. Moreira, and J. Tasker (2017): "Peering into the present: the Banks approach to GDP nowcasting," *Bank of England Quarterly Bulletin*, 57, 122–133.

Angelini, E., M. Lalik, M. Lenza, and J. Paredes (2019): "Mind the gap: A multi-country BVAR benchmark for the Eurosystem projections," *International Journal of Forecasting*, 35, 1658 – 1668.

Bańbura, M., D. Giannone, and M. Lenza (2015): "Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections," *International Journal of Forecasting*, 31, 739–756.

Bańbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92.

Benalal, N., J. L. D. del Hoyo, B. Landau, M. Roma, and F. Skudelny (2004): *To aggregate or not to aggregate? Euro Area inflation forecasting*, Working paper series / European Central Bank 374, Frankfurt am Main: European Central Bank.

Berkowitz, J. (2001): "Testing Density Forecasts, with Applications to Risk Management," *Journal of Business & Economic Statistics*, 19, 465–474.

Bermingham, C. and A. D'Agostino (2014): "Understanding and forecasting aggregate and disaggregate price dynamics," *Empirical economics : a journal of the Institute for Advanced Studies, Vienna, Austria*, 46, 765–788.

Bobeica, E. and M. Jarociński (2019): "Missing Disinflation and Missing Inflation: A VAR Perspective," *International Journal of Central Banking*, 15, 199–232.

Bobeica, E. and A. Sokol (2019): "Drivers of underlying inflation in the euro area over time: a Phillips curve perspective," *Economic Bulletin Articles*, 4.

Bok, B., D. Caratelli, D. Giannone, A. M. Sbordone, and A. Tambalotti (2018): "Macroeconomic Nowcasting and Forecasting with Big Data," *Annual Review of Economics*, 10, 615–643.

Clark, T. and M. McCracken (2013): "Chapter 20 - Advances in Forecast Evaluation," in *Handbook of Economic Forecasting*, ed. by G. Elliott and A. Timmermann, Elsevier, vol. 2 of *Handbook of Economic Forecasting*, 1107 – 1201.

Clements, M. P. and J. Smith (2000): "Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment," *Journal of Forecasting*, 19, 255–276.

Cobb, M. P. A. (2019): "Aggregate density forecasting from disaggregate components using Bayesian VARs," *Empirical Economics*.

Darvas, Z. (2018): "Forecast errors and monetary policy normalisation in the euro area. Bruegel Policy Contribution Issue n?24 | December 2018," .

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39, 863–883.

Diebold, F. X. and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Dieppe, A., B. van Roye, and R. Legrand (2016): "The BEAR toolbox," Working Paper Series 1934, European Central Bank.

Doan, T., R. Litterman, and C. Sims (1984): "Forecasting and conditional projection using realistic prior distributions," *Econometric Reviews*, 3, 1–100.

Domit, S., F. Monti, and A. Sokol (2019): "Forecasting the UK economy with a medium-scale Bayesian VAR," *International Journal of Forecasting*, 35, 1669 – 1678.

Duarte, C. and A. Rua (2007): "Forecasting inflation through a bottom-up approach: How bottom is bottom?" *Economic Modelling*, 24, 941–953.

Des, S. and J. Guentner (2016): "Forecasting Inflation Across Euro Area Countries and Sectors: A Panel VAR Approach: Forecasting Inflation: A Panel VAR Approach," *Journal of Forecasting*.

Ehrmann, M., G. Ferrucci, M. Lenza, and D. O'Brien (2018): "Measures of underlying inflation for the euro area," *Economic Bulletin Articles*, 4.

Espasa, A. and I. Mayo-Burgos (2013): "Forecasting aggregates and disaggregates with common features," *International Journal of Forecasting*, 29, 718–732.

Faust, J. and J. Wright (2013): "Forecasting Inflation," *Part A of Handbook of Economic Forecasting*, 2, 2–56.

Giannone, D., J. Henry, M. Lalik, and M. Modugno (2012): "An Area-Wide Real-Time Database for the Euro Area," *The Review of Economics and Statistics*, 94, 1000–1013.

Giannone, D., M. Lenza, D. Momferatou, and L. Onorante (2014): "Short-term inflation projections: A Bayesian vector autoregressive approach," *International Journal of Forecasting*, 30, 635–644.

Giannone, D., M. Lenza, and G. E. Primiceri (2015): "Prior Selection for Vector Autoregressions," *The Review of Economics and Statistics*, 97, 436–451.

Harvey, D., S. Leybourne, and P. Newbold (1997): "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 13, 281 – 291.

Heinisch, K. and R. Scheufele (2018): "Bottom-up or direct? Forecasting German GDP in a data-rich environment," *Empirical Economics*, 54, 705–745.

Hendry, D. and K. Hubrich (2006): "Forecasting economic aggregates by disaggregates," Working Paper Series 589, European Central Bank.

——— (2011): "Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate," *Journal of Business & Economic Statistics*, 29, 216–227.

Hubrich, K. (2005): "Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy?" *International Journal of Forecasting*, 21, 119–136.

Iversen, J., S. Lasen, H. Lundvall, and U. Sderstrm (2016): "Real-Time Forecasting for Monetary Policy Analysis: The Case of Sveriges Riksbank," Working Paper Series 318, Sveriges Riksbank (Central Bank of Sweden).

Koester, G., E. Lis, C. Nickel, C. Osbat, and F. Smets (2021): "Understanding low inflation in the euro area from 2013 to 2019: cyclical and structural drivers," Occasional Paper Series 280, European Central Bank.

Kohn, R. (1982): "When is an aggregate of a time series efficiently forecast by its past?" *Journal of Econometrics*, 18, 337–349.

Kontogeorgos, G. and K. Lambrias (2019): "An analysis of the Eurosystem/ECB projections," Working Paper Series 2291, European Central Bank.

Koop, G. and D. Korobilis (2012): "Forecasting Inflation Using Dynamic Model Averaging," *International Economic Review*, 53, 867–886.

Korobilis, D. (2017): "Quantile regression forecasts of inflation under model uncertainty," *International Journal of Forecasting*, 33, 11 – 20.

Litterman, R. (1986): "Forecasting with Bayesian Vector Autoregressions-Five Years of Experience," *Journal of Business & Economic Statistics*, 4, 25–38.

Luetkepohl, H. (1984): "Forecasting Contemporaneously Aggregated Vector ARMA Processes," *Journal of Business & Economic Statistics*, 2, 201–214.

——— (1987): *Forecasting Aggregated Vector ARMA Processes*, vol. 284 of *Lecture Notes in Economics and Mathematical Systems,*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Marcellino, M., J. Stock, and M. Watson (2003): "Macroeconomic forecasting in the Euro area: Country specific versus area-wide information," *European Economic Review*, 47, 1–18.

Mazur, B. (2016): "Density Forecasts Based on Disaggregate Data: Nowcasting Polish Inflation," *Dynamic Econometric Models*, 15.

Moser, G., F. Rumler, and J. Scharler (2007): "Forecasting Austrian inflation," *Economic Modelling*, 24, 470–480.

Nakamura, E. (2005): "Inflation forecasting using a neural network," *Economics Letters*, 86, 373–378.

Ravazzolo, F. and S. Vahey (2014): "Forecast densities for economic aggregates from disaggregate ensembles," *Studies in Nonlinear Dynamics & Econometrics*, 18.

Reijer, A. and P. Vlaar (2006): "Forecasting Inflation: An Art as Well as a Science!" *De Economist*, 154, 19–40.

Rose, D. E. (1977): "Forecasting aggregates of independent ARIMA processes," *Journal of Econometrics*, 5, 323–345.

Rossi, B. and T. Sekhposyan (2014): "Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set," *International Journal of Forecasting*, 30, 662 – 682.

Stock, J. H. and M. W. Watson (2007): "Why has US inflation become harder to forecast?" *Journal of Money, Credit and Banking*, 39, 3–33, conference on Quantitative Evidence on Pricw Determination, Feder Reserve Board Gemgf, Washington, DC, SEP 29-30, 2005.

Tallman, E. W. and S. Zaman (2017): "Forecasting inflation: Phillips curve effects on services price measures," *International Journal of Forecasting*, 33, 442–457.

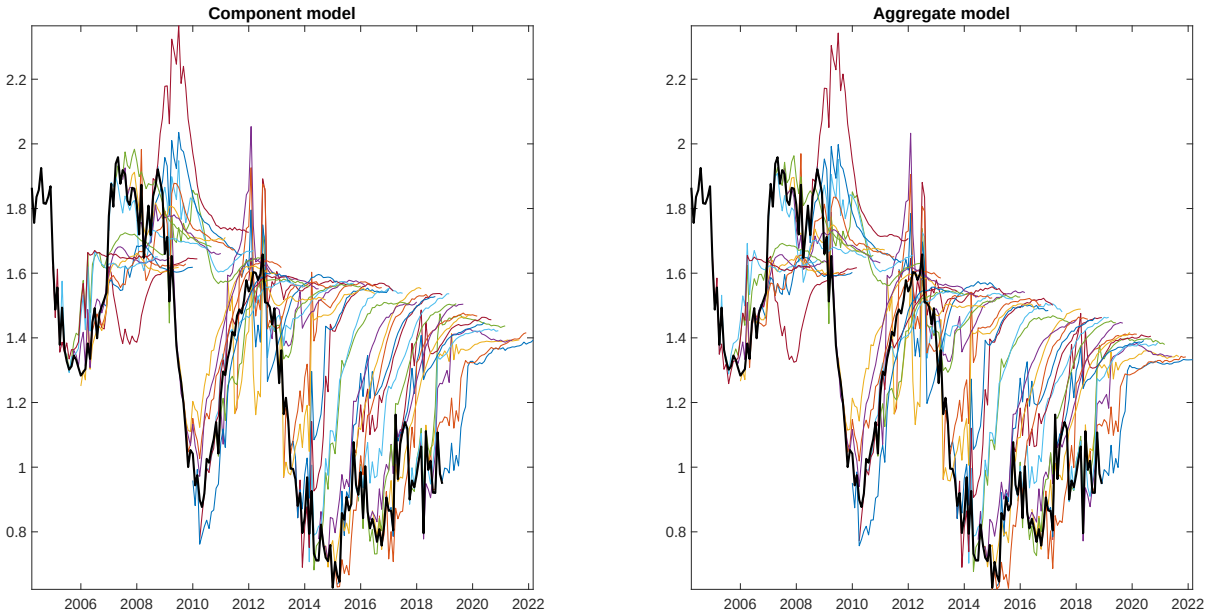
Theil, H. (1954): *Linear aggregation of economic relations*, Contributions to economic analysis, North-Holland Pub. Co.

Tiao, G. C. and I. Guttman (1980): "Forecasting contemporal aggregates of multiple time series," *Journal of Econometrics*, 12, 219–230.

Waggoner, D. and T. Zha (1999): "Conditional Forecasts In Dynamic Multivariate Models," *The Review of Economics and Statistics*, 81, 639–651.

A Euro Area Results: HICP Excluding Food and Energy

Figure A.1 Conditional forecasts of HICP inflation excluding food and energy over time



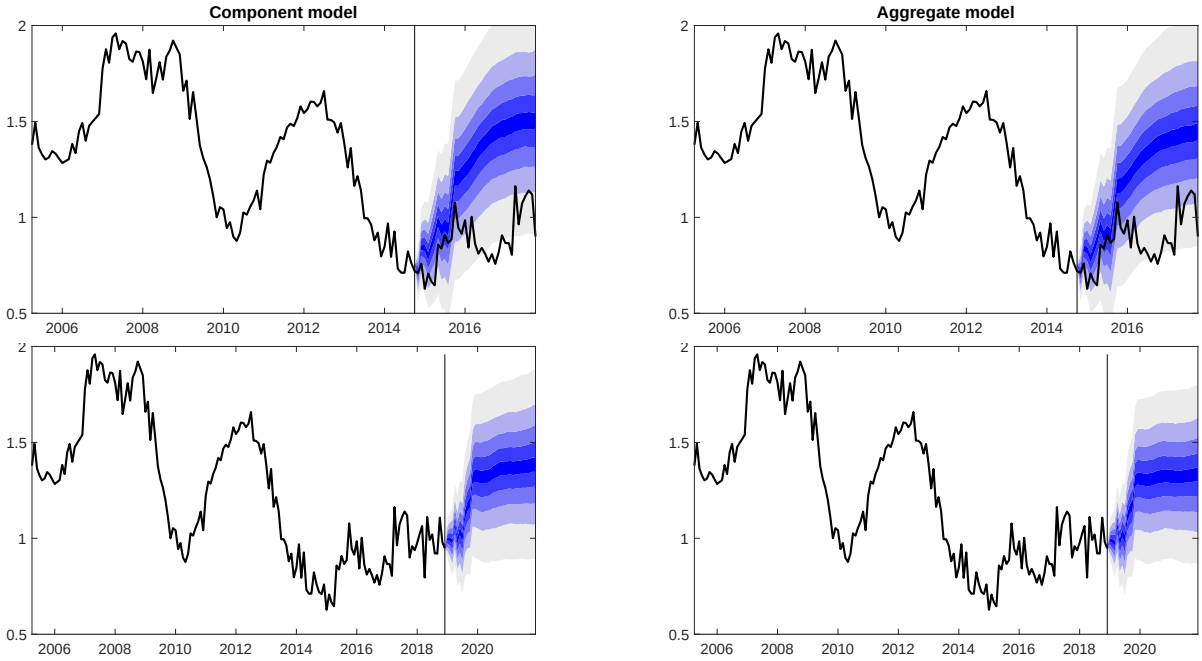
Note: The figures show consecutive forecast vintages of year-on-year inflation (colored lines) against the latest data vintage (solid black).

Table A.1 Component vs Aggregate model conditional forecasts for HICP inflation excluding food and energy

Months ahead	RMSE *Component model*	relative RMSE *Component/Aggregate*	relative RMSE *Component/(B)MPE*
1	0.10	1.03	–
3	0.16	1.03	1.64
6	0.25	1.03	1.00
12	0.41	1.04	1.00
24	0.51	1.06	1.09
36	0.55	1.07	–

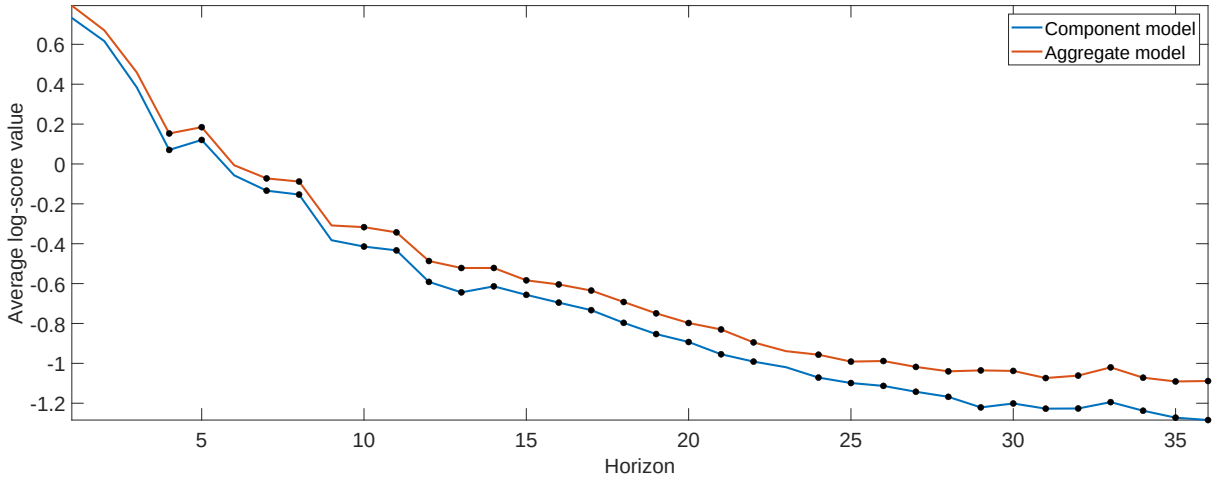
Note: A relative RMSE < 1 indicates that Component model forecasts are more accurate. Bold text denotes statistical significance of the difference at the 5% level, based on Diebold–Mariano tests with Harvey et al. (1997) correction. The comparison with Eurosystem/ECB staff (B)MPE projections (last column) is made on quarterly, rather than monthly, year-on-year inflation rates.

Figure A.2 Fan charts with conditional forecasts of HICP inflation excluding food and energy (top panel - vintage December 2014, bottom panel - vintage March 2019)



Note: Top panel - December 2014 vintage; bottom panel - March 2019 vintage. The fan charts depict the evolution of selected quantiles of the predictive distributions of year-on-year inflation over the projection horizon: the darkest band is centered around the median, and the outer edges of the lightest bands correspond to the 0.05 and 0.95 quantiles.

Figure A.3 Average log predictive scores for conditional forecasts - HICP inflation excluding food and energy



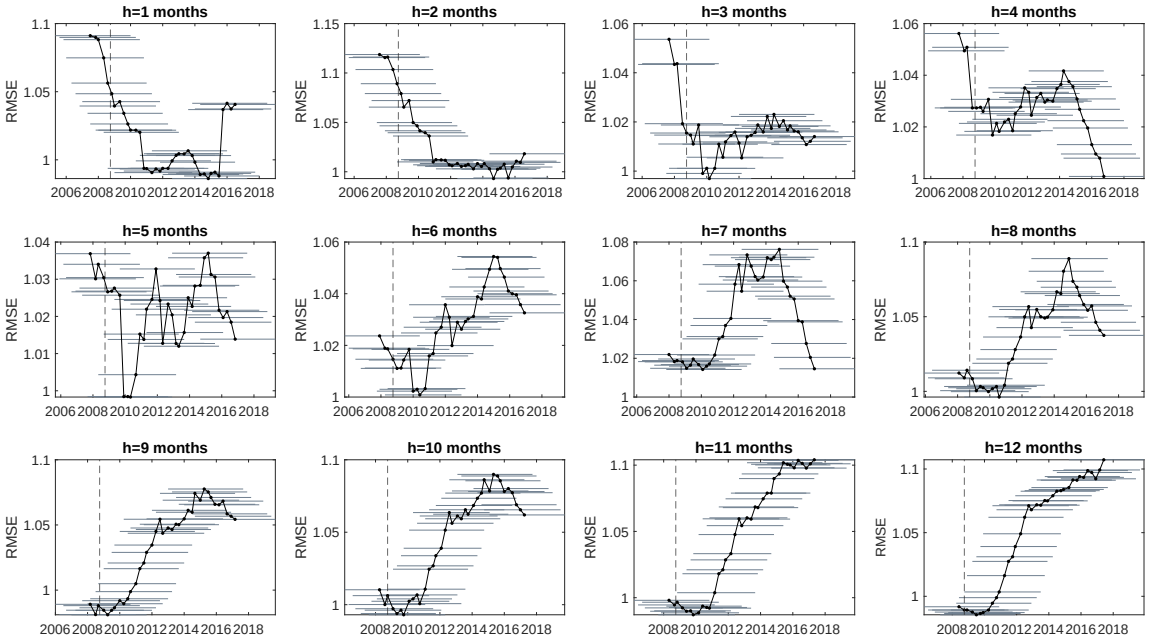
Note: A higher (less negative) log predictive score indicates better forecast accuracy. Black squares mark horizons for which the two scores are statistically different from each other based on Amisano and Giacomini (2007) with Harvey et al. (1997) correction.

Table A.2 Calibration tests on Probability Integral Transforms for HICP inflation excl. food and energy

	Aggregate model				**Component model**			
	Uniformity		*Independence*	*Joint* H_0	*Uniformity*		*Independence*	*Joint* H_0
h	*KS*	*AD*	*LB*	*Ber*	*KS*	*AD*	*LB*	*Ber*
1	0.434	0.215	0.454	0.111	0.140	0.068	0.453	**0.045**
3	0.292	0.083	0.892	0.122	0.112	**0.020**	0.949	0.026
6	0.039	**0.004**	**<0.001**	**<0.001**	**0.001**	**<0.001**	**<0.001**	**<0.001**
12	0.004	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
24	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
36	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

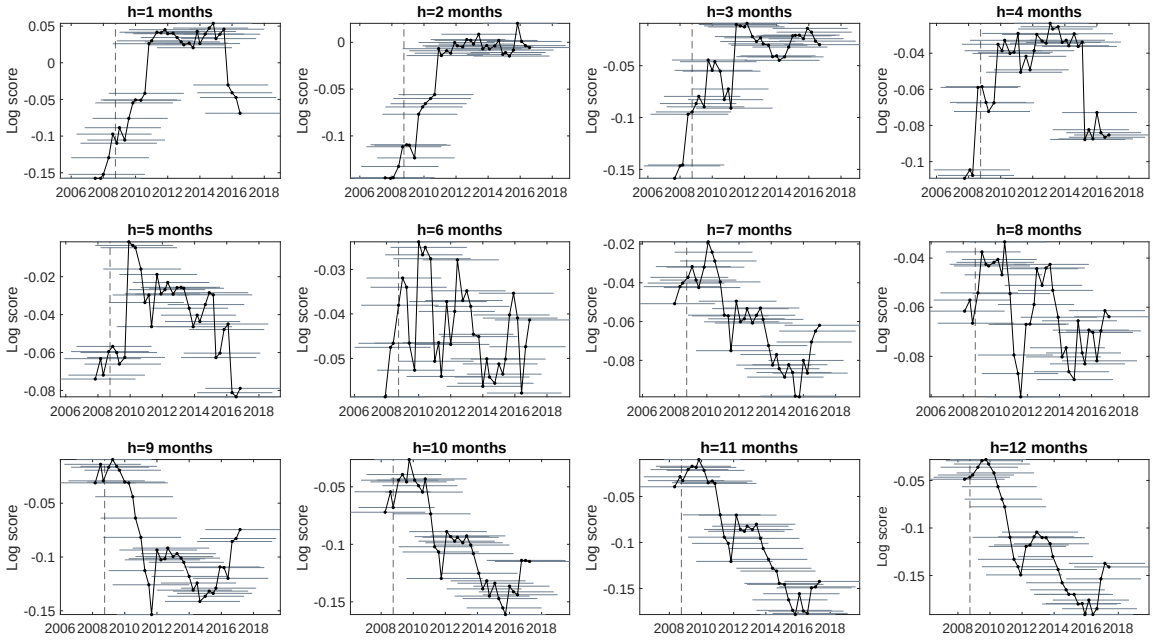
Note: P-values (for horizons $h > 2$, minimum p-values) of the respective test: Kolmogorov-Smirnov (KS), Anderson-Darling (AD), Ljung-Box of the mean (LB), Berkowitz (Ber). Bold text indicates rejection at the 5% significance level (for horizons $h > 2$, using Bonferroni bounds). See Section 2 for more details and references.

Figure A.4 Rolling relative RMSE (component vs aggregate model) by forecast horizon for HICP inflation excluding food and energy



Note: The black markers denote the center of the 5 year rolling window spanned by the light grey bars. The vertical dashed line denotes the onset of the Great Recession. A relative RMSE$> 1 (< 1)$ indicates better performance of the aggregate (component) model.

Figure A.5 Rolling relative log scores (component vs aggregate model) by forecast horizon for HICP inflation excluding food and energy



Note: The black markers denote the center of the 5 year rolling window spanned by the light grey bars. The vertical dashed line denotes the onset of the Great Recession. A negative (positive) value indicates better performance of the aggregate (component) model.

Table A.3 Relative RMSEs of Component vs Aggregate model conditional forecasts for HICP inflation excluding food and energy in individual countries

Months ahead	France	Germany	Italy	Netherlands	Spain
1	**1.21**	1.04	0.95	**1.07**	1.22
3	0.96	0.95	0.97	1.05	1.01
6	0.91	0.94	0.97	1.02	1.00
12	0.98	0.92	0.98	1.05	1.01
24	1.05	0.97	1.02	1.04	**1.03**
36	1.04	0.95	1.03	1.04	**1.03**

Note: A relative RMSE < 1 indicates that Component model forecasts are more accurate. Bold text denotes statistical significance of the difference at the 5% level, based on Diebold–Mariano tests with Harvey et al. (1997) correction.