8/RT/10 **August 2010**

Research Technical Paper

Understanding and Forecasting Aggregate and Disaggregate Price Dynamics

Colin Bermingham **Antonello D'Agostino***

Central Bank and Financial Services Authority of Ireland
P.O. Box 559, Dame Street
Dublin 2
Ireland
http://www.centralbank.ie

*The views expressed in this paper are the personal responsibility of the authors. They are not necessarily held either by the CBFSAI or the ESCB. All errors are our own. The authors would like to thank for helpful comments. Email: colin.bermingham@centralbank.ie, antonello.dagostino@centralbank.ie

Abstract

The issue of forecast aggregation is to determine whether it is better to forecast a series directly or instead construct forecasts of its components and then sum these component forecasts. Notwithstanding some underlying theoretical results, it is generally accepted that forecast aggregation is an empirical issue. Empirical results in the literature often go unexplained. This leaves forecasters in the dark when confronted with the option of forecast aggregation. We take our empirical exercise a step further by considering the underlying issues in more detail. We analyse two price datasets, one for the United States and one for the Euro Area, which have distinctive dynamics and provide a guide to model choice. We also consider multiple levels of aggregation for each dataset. The models include an autoregressive model, a factor augmented autoregressive model, a large Bayesian VAR and a time-varying model with stochastic volatility. We find that once the appropriate model has been found, forecast aggregation can significantly improve forecast performance. These results are robust to the choice of data transformation.

JEL classification: E17, E31.
Keywords: Aggregation, Forecasting, Inflation.
Original Version: August 2010
This Version: February 2011

Non-technical summary

The issue of forecast aggregation is to determine whether it is better to forecast a series directly or instead construct forecasts of its components and then sum these component forecasts. Notwithstanding some underlying theoretical results, it is generally accepted that forecast aggregation is an empirical issue. We conduct empirical exercises and relate our findings back to the properties of the dataset and the models used. The exercise is conducted on both United States (US) and Euro Area (EA) inflation. Although both datasets relate to inflation, these datasets have distinct characteristics and we tailor the model to the properties of the data. In all the empirical exercises in this paper, forecast aggregation leads to better forecasts. The aggregate forecast often has the least satisfactory performance and this makes the argument for aggregation more compelling given that multiple levels of aggregation are used.

The performance of the aggregated forecasts depends on the type of model used. In particular, the model must capture the key characteristics of the data. There is strong comovement in US inflation. Simple AR models do not perform very well in this context but multivariate models such as factor models and BVAR models that can capture this common movement or pick up feedback between the series have more accurate forecasts. For the Euro Area inflation rate, there is far less commonality and the series have more individual dynamics. Simple AR models tend to work well for this type of dataset. They have more accurate forecasts than both the benchmark and their multivariate counterparts.

The exercises are mainly based on multistep forecasts of year-on-year inflation rates. For US inflation, we forecast the h-quarter price change for $h = 1, ..., 8$ and find the results are robust to this change in the target forecast variable. We also introduce a time-varying model with stochastic volatility where forecasts are constructed iteratively. The time-varying model in conjunction with forecast aggregation leads to further improvements in forecast power. These robustness checks corroborate the main results in favour of forecast aggregation. The paper provides a substantive endorsement of the forecast aggregation approach, particularly in terms of inflation. The key to realising gains in terms of forecast aggregation lies in the ability to uncover the appropriate model for a particular dataset.

1 Introduction

When forecasting economic variables, one is often faced with the choice of either forecasting an aggregate directly or forecasting its components and then summing the component forecasts. This is frequently encountered when forecasting inflation, where prices are commonly available for a large number of components series in addition to the aggregate price index. The aggregation issue is a major practical consideration when it comes to forecasting key economic indicators but frequently forecasters are in the dark in terms of which approach is likely to yield the best results. There is a considerable set-up cost when estimating models on disaggregate data if there are a large number of component series so researchers are understandably reluctant to pursue this strategy unless it is likely to yield benefits.

Arguably, the literature on forecast aggregation is at an impasse. The early contributions focussed on deriving theoretical results but this approach was eventually abandoned as the underlying assumptions were too restrictive. Empirical papers tend to focus on a specific application. Competing sets of forecasts are constructed for a given country or set of countries to see whether forecast aggregation helps. With the exception of Hubrich (2003), few papers offer potential explanations of why the forecast aggregation strategy was a success or failure so there is little guidance to forecasters faced with the option of combining disaggregate forecasts.

We conduct empirical exercises but relate our findings back to the properties of the dataset and the models used. The exercise is conducted on both United States (US) and Euro Area (EA) inflation. Although both datasets relate to inflation, these datasets have distinct characteristics and we tailor the model to the properties of the data. In contrast to most previous studies, we consider multiple levels of aggregation for each dataset.[1] We find that, once the appropriate model is found for a dataset, forecast aggregation always leads to improvements in forecast accuracy - the critical issue is to find the appropriate model. Frequently, the forecast based on the aggregate results in the worst forecast performance. This story is consistent with the theoretical literature. By providing a detailed explanation for main factors driving results for both datasets, we provide a greater understanding of the key issues relative to other empirical papers.

In the next section, we provide a summary of the main contributions in both the theoretical and empirical side of the literature. Section 3 briefly considers the issue of aggregation of forecasts in terms of an AR model. We do this from an heuristic perspective with a view to highlighting the main issues. Section 4 describes the data. Section 5 describes the models used in the paper with the results reported in section 6. Section 7 outlines some robustness checks while section 8 provides a summary and concludes the paper.

[1]One exception is Duarte and Rua (2005), who consider a 5-item and 59-item breakdown of the CPI in Portugal. We examine four different levels of aggregation for the US and three for the EA in this study.

2 Literature Review

Early contributions in the area of forecast aggregation were mainly confined to theoretical results based on an assumed data generating process (DGP). Assuming that the components are ARIMA processes, Rose (1977) examines the DGP and forecasts for an aggregate of these models. Others including Tiao and Guttman (1980), Kohn (1982) and Luetkepohl (1984a, 1984b) followed this approach with the DGP or forecast performance of the aggregated process related to an assumed structure for the DGP of the components. Based on asymptotic theory, it is possible to state that the disaggregate forecast will have a lower forecast error if the DGPs of all components are known. Luetkepohl (1984a) acknowledges that the superiority of the disaggregate forecast is no longer assured if the DGPs aren't known and instead must be estimated. In practice, DGPs are not known to forecasters so the results of these studies provide the starting point to investigations of forecast aggregation but the question of forecast aggregation has a strong empirical element.

European Monetary Union (EMU) revived interest in the topic of forecast aggregation but given the limited success of the theoretical approach, the literature changed direction and empirical exercises became much more common. There have been two distinct approaches adopted. The traditional approach, which is followed in this paper, is to construct forecasts of the disaggregates and combine them. In a couple of recent papers, Hendry and Hubrich (2006, 2010) suggest the alternative route of including disaggregates directly in the model of the aggregate. These two papers consider both predictability in population and forecastability in sample through both analytical and empirical work. They consider a number of practical issues such as changing coefficients, specification error and estimation uncertainty. They find that including disaggregate information in the aggregate models helps to improve forecasts. Luetkepohl (2010) includes practical exercises that also follow this approach. Systems of VARs are estimated which include both aggregate and disaggregate information for employment and inflation in the Euro Area. Luetkepohl finds that although taking disaggregate information into account should theoretically improve forecasts, the inclusion of too many disaggregates can result in estimation error and specification error which ultimately leads to an efficiency loss.

This paper is concerned with the traditional approach, which is the focus of much of the literature. Hubrich (2003) and Benalal et al (2004) both examine HICP inflation for the euro area (Benalal et al also consider the four largest coutries) and find that there are no significant benefits to forecast aggregation. In country specific studies of HICP inflation, Duarte and Rua (2005), Bruneau et al (2007) and Moser et al (2007) all find forecast aggregation leads to improved forecasts for inflation for Portugal, France and Austria respectively. The results for the EA papers contrast with the country specific studies but all papers employ different models and are estimated over different time spans.

Forecast aggregation has also been examined in the context of output forecasting. Zellner and Tobias (2000) forecast the aggregate growth rate of 18 industrial countries using an aggregate and disaggregate approach. They report improved forecasts from the disaggregate approach. Marcellino, Stock and Watson (2003) forecast prices and three activity measures for the euro area directly and by aggregating country specific models. They find forecasts are more accurate when country specific models are aggregated. With the exception of Hubrich (2003) and Benalal et al (2004), the results of the empirical papers generally support forecasting disaggregates. Hubrich (2003) and Benalal et al (2004) are both reliant on short spans of data, as they were conducted shortly after the beginning of monetary union. This suggests that estimation error may have been a significant problem, particularly in the case of highly parameterised models such as VARs. They also focus on a small number of disaggregates - five in each case. We find that it is preferable to use a more detailed breakdown and our results support forecast aggregation for EA inflation.

The aim of this paper is to look into the issue of forecast aggregation in greater detail than the existing literature. In contrast to the standard approach, we utilise multiple levels of disaggregate data for each dataset and a number of different models (autoregressive (AR), factor augmented AR, Bayesian Vector Autoregression (VAR) with Minnesota priors and time varying VAR). This allows us to explore the properties of the data which lead to benefits in terms of forecast aggregation. By considering two separate dataset with different characteristics, we are also able to highlight the importance that the selection of the correct model type has on the results. These insights are valuable to other forecasters contemplating the aggregation approach.

3 Factors Affecting Forecast Performance

In this section, we discuss the factors that are likely to impact on forecast performance. We do not provide conclusive theoretical results which determine the results of our empirical exercise. The theoretical models are not sufficiently rich to capture the interplay of all relevant factors in a unified framework. But a discussion of some of the relevant issues here helps to provide a more intuitive understanding of the empirical work. We frame the discussion in this section around the AR model as this is the most basic model that we use in the empirical exercise.

3.1 Specification of Aggregate Process

In the case examined in the paper, the price aggregate is a weighted average of all the other sub-components, and a as consequence, its dynamic are likely to be quite complex. For example, theoretical results tell us that the aggregate of two AR(1) processes will be an ARMA(2,1) process. More generally, the aggregate of an AR(p_1) and AR(p_2) process will be an ARMA(($p_1 + p_2$),max(p_1, p_2)) process. Thus, when aggregating a large number, say n, of component AR process, the theoretical AR lag length is $\sum_{i=1}^{n} p_i$, which may be even greater than the number of data points available when n is large. The theoretical MA lag is simply the longest MA lag found among the individual series. The estimated aggregate process will represent an approximation to this theoretical model, with a lot of the theoretical coefficients set equal to zero. All theortically relevant coefficients will not be statistically significant so some will be excluded in practice. The exclusion of relevant parameters is balanced against the need for parsimony. Amongst others, Enders (2010) points out that forecasts may be better from overly parsimonious models relative to those that exactly fit the theoretical model, as the former may benefit from low estimation error / parameter uncertainty. Therefore, in forecasting work, we might not worry about estimating all coefficients as long as the key parameters are tightly estimated.

3.2 Forecast Variance

If we again consider an AR process, the variance of the forecast error is known to depend on both the variance of the disturbance term and the amount of estimation error. Suppose that $y_t = ay_{t-1} + \epsilon_t$. In the absence of parameter uncertainty, the one-step forecast is given by $E_t y_{t+1} = ay_t$ with the corresponding mean squared forecast error (MSFE) given by $E_t(y_{t+1} - ay_t)^2 = E_t\epsilon^2_{t+1} = \sigma^2_\epsilon$. To take account of parameter uncertainty, the known quantity a is replaced with $\hat{a}$ in the calculation of the MSFE:

$$\begin{aligned} MSFE &= E_t(y_{t+1} - \hat{a}y_t)^2 \\ &= E_t[(ay_t - \hat{a}y_t)^2) + \epsilon^2_{t+1}] \\ &= E_t[(a-\hat{a})^2](y_t)^2 + \sigma^2_\epsilon \end{aligned} \quad (1)$$

with the later equalities holding due to independence assumptions. Clearly, the parameter estimation error is strictly positive and contributes to overall forecast variance. It is decreasing in terms of the sample size (and disappears asymptotically) but increasing in terms of the number of parameters. Thus, in a model with a short data span and a lot of parameters, it can prove a significant obstacle to forecasting. This underpins the need for parsimony in forecasting.

A serious shortcoming of the AR model is that a univariate specification takes no account of missing information. Prices can be influenced by a range of factors so the model will be miss-specified through omitted variable basis. The resulting coefficient bias will hurt forecast performance. The influence of the missing factors will show up in the residuals of each equation so that the fit of the model will be lower than desired. In the context of aggregating forecasts, there is a second implication. One would hope that the forecast errors of the individual series will offset each other to some extent. If, however, the missing information impacts the individual series in the same direction[2], then all the individual forecasts will tend to either overshoot or undershoot the correct value. This could seriously disadvantage the forecast aggregation approach. The key to overcoming omitted variable bias is by including extra information in the individual regressions in as parsimonious a manner as possible. We will return to comment on these issues in the results section.

[2] An positive oil price shock would cause most of the inflation series to increase.

4 Data

4.1 US Data

The analysis in this paper draws on both US data and EA data. The US series are NIPA data from the Bureau of Economic Analysis (BEA).[3] The price series are personal consumption expenditures available quarterly from 1959Q1 - 2009Q4. The data are already available at different levels of aggregation. This paper considers four different levels of aggregation for the dataset. The first is a three item breakdown which includes the prices of durable goods, non-durable goods and services. We next consider a fifteen item breakdown. The price categories are still quite broad at this level of aggregation and examples include food, housing and transport. A full list of price series for all levels of aggregation is provided in the Table 1. The third breakdown consists of fifty different price series The categories here are quite narrowly defined and again are presented in Table 1. The final breakdown is based on 169 series. The series are too numerous to list in the Table but a list of included items is available upon request.

As we wish to compare aggregated individual forecasts with the forecasts from the overall PCE inflation rate, we must be able to construct the PCE inflation rate from the individual inflation rates as a first step. This requires the weights of each item for each level of aggregation. All data are taken from Tables 2.4.4U, 2.4.5U and 2.4.6U on the BEA website. The price series are chained index values and their weights are calculated according to the approximation provided in Dolmas (2006):

$$w_{i,t+1} = \frac{1}{2}\frac{Q_{i,t}P_{i,t}}{\sum Q_{i,t}P_{i,t}} + \frac{1}{2}\frac{Q_{i,t+1}P_{i,t}}{\sum Q_{i,t+1}P_{i,t}}$$

The weight at time $t+1$ is equal to an average of the expenditure share of the product at time t and its expenditure share had consumers bought the $t+1$ quantity at time t prices. In each case, the accuracy of this approximation was checked by constructing the aggregate inflation rate from the components. The aggregate inflation rate was recovered with a high level of precision, which ensures the validity of the empirical exercise.

Figure 1 graphs the Year on Year (YoY) PCE inflation rate and its component inflation rates for each of the four different levels of aggregation used in the paper. In each graph, the thick blue line is the aggregate inflation rate. For the graph of the 3 items, the individual items move in tandem with the PCE inflation rate. As the number of items in each breakdown increases, the series obviously have more individual dynamics although there is still quite noticeable comovement with the PCE rate, indicated by the tight bunching of series around the PCE inflation rate.

[3] Available at: http://www.bea.gov/national/nipaweb/SelectTable.asp

4.2 EA Data

The euro area data are price series for the Harmonized Index of Consumer Prices (HICP). The series along with their weights are available on the Eurostat website.[4] The series are disaggregated at three different levels with a 5-item, 12-item and 32-item breakdown. The items included in each level of aggregation are also presented in Table 1. The series are monthly and the sample period is from January 1996 to December 2009. Although it is possible to get data at a more detailed level over the latter part of the sample, it is not possible to do so for the entire sample so the 32-item breakdown represents the most detailed available for our purposes. There is a strong seasonal pattern in some of the euro area data when month-on-month growth rates are calculated. Seasonally adjusted data are not available. In addition, the seasonal pattern is not stable over the sample and so it not possible to estimate a consistent seasonally adjusted series. To mitigate this problem, estimation is conducted using year-on-year growth rates. Seasonality is not an issue with the US data as all series are seasonally adjusted.

Figure 2 graphs the YoY HICP inflation rate and the component rates at the three levels of aggregation used. At the five item level, the series display more heterogeneous dynamics relative to the US data. This pattern is repeated with the 12 and 32 item datasets. Although there is bunching around the aggregate, these series have stronger individual characteristics than the US data. This is probably due to the fact that the US data are from one country whereas the EA data combines the inflation rates of several difference countries.

[4] Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

Figure 1: PCE Inflation and its Component Inflation Rates

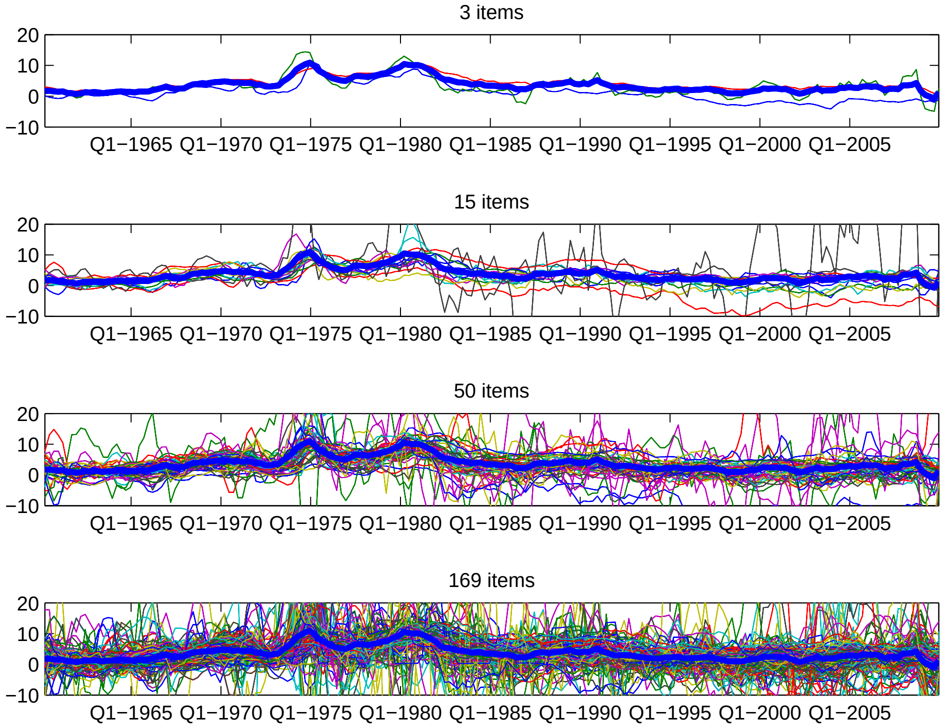
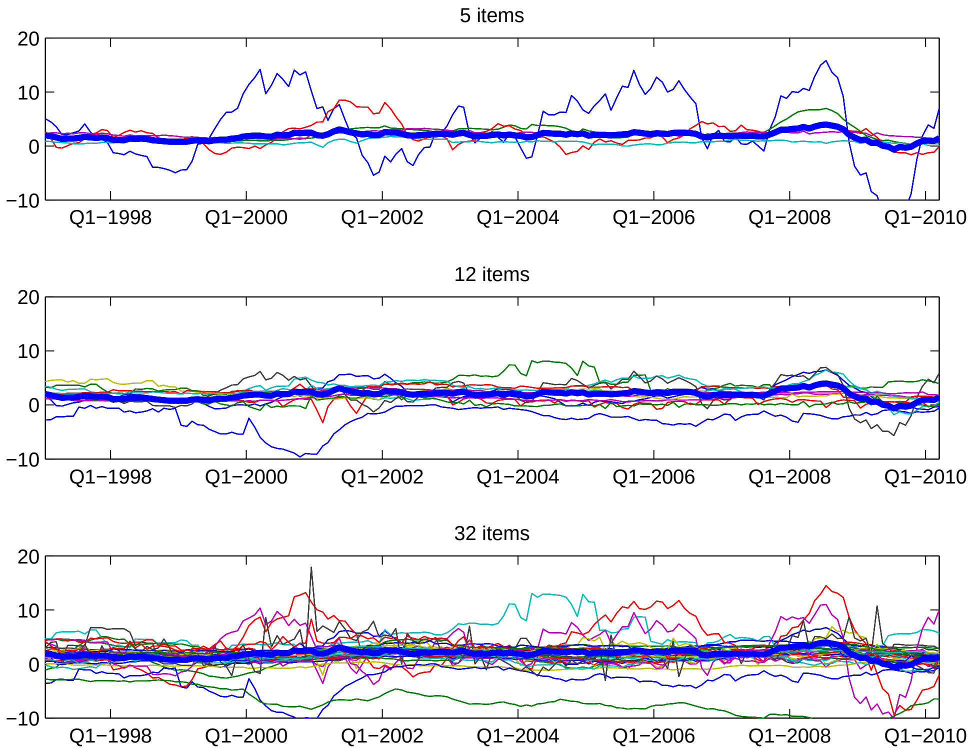


Figure 2: HICP Inflation and its Component Inflation Rates



5 Data Transformation and Models Specification

The empirical exercise in this paper is addressed in the following way: we construct one set of forecasts by estimating models on the aggregate series and a second set by using the same model to forecast the individual series prior to aggregation, then we compare the accuracy of both approaches. The target variable is the aggregate, annualized h period inflation, defined as $\pi_t^h = k\log(\frac{P_t}{P_{t-h}})$, where the constant k is the normalization term.[5] P_t is the aggregate level of price index. Given a model m, we perform a pseudo out-of-sample forecasting simulation. At time t, we estimate the parameters of the model and compute the forecasts of the aggregate and disaggregate inflation series at horizon h, then we update the sample with a new observation and, at time $t+1$, we re-estimate the parameters of the model and compute again the forecasts for time $t+1+h$. This pseudo out-of-sample forecasting exercise is iterated up to the end of the sample for each type of model. This exercise is not a real time exercise. However, as we only use inflation rates for each dataset, the series are all published with the same lag. In addition, the inflation rates in both datasets are not subject to a significant degree of revision. From this perspective, we argue that this exercise provides a realistic appraisal of aggregate versus disaggregate forecasting.

Forecasts of the target variable at horizon h are denoted as $\hat{\pi}^{h,m}_{a,t+h|t}$, when they are computed directly on the aggregate inflation series, and as $\hat{\pi}^{h,m}_{d,t+h|t} = \sum_{j=1}^{N_s} w_{j,t}\hat{\pi}^{h,m}_{j,t+h|t}$ when they are computed by aggregating forecasts of disaggregate inflation series j.[6] The first subscripts a or d denote if the forecast of the target variables is computed with the aggregate or disaggregate inflation series respectively, while $t+h|t$ refers to the fact that, for horizon h, forecasts are computed by using information up to time t. Finally, the first superscript h denotes the transformation adopted for prices, while m refers to the model employed.

We will use the following models:

Atkeson Ohanian Model (AO) (2001)

$$\pi^h_{t+h} = \pi^4_t + \omega^h_{t+h} \tag{2}$$

The forecast at $t+h$ is computed as:

$$\hat{\pi}^{h,AO}_{a,t+h|t} = \pi^4_t \tag{3}$$

AutoRegressive Model (AR)

For a generic inflation series j:

$$\pi^h_{j,t+h} = \alpha^h_j + B^h_j(L)\pi^h_{j,t} + \varepsilon^h_{j,t+h} \tag{4}$$

[5] It is $\frac{400}{h}$ in the case of quarterly data and $\frac{1200}{h}$ in the case of monthly data.

[6] N_s is the number of series in the sth set of disaggregate series; $s = \{1,2,3,4\}$ in the case of US dataset and $s = \{1,2,3\}$ in the case of Euro area dataset.

where $B_j^h(L) = B_{j,0}^h + ... + B_{j,s}^h L^s$ is a polynomial in the lag operator L. Parameters are estimated by Ordinary Least Squares (OLS). The forecast for a given horizon h is computed as:

$$\hat{\pi}_{j,t+h|t}^{h,AR} = \hat{\alpha}_j^h + \hat{B}_j^h(L)\pi_{j,t}^h \tag{5}$$

Factor Augmented AutoRegressive Model (FAAR)

$$\pi_{j,t+h}^h = \nu_j^h + C_j^h(L)\pi_{j,t}^h + \gamma F_t + \zeta_{j,t+h}^h \tag{6}$$

This is the AR model of *eq.*(4) augmented with one factor. The factor is estimated with the first principal component (Stock and Watson, 2002) computed on the most detailed data set available. For example, the factor for the US dataset is computed on the dataset of 169 series, while that for the Euro area is computed on the dataset of 32 series. Parameters of *eq.*(6) are estimated by OLS. The forecast at horizon h is given by:

$$\hat{\pi}_{j,t+h|t}^{h,FAAR} = \hat{\nu}_j^h + \hat{C}_j^h(L)\pi_{j,t}^h + \hat{\gamma}\hat{F}_t \tag{7}$$

Bayesian VAR (BVAR)

This is a Bayesian VAR with the Minnesota prior as proposed by Banbura, Giannone and Reichlin (2010). Let's denote with $P_{j,t}$, the price level for series $j \in S_i, where S_i = \{1, ..., j, ..., n_i\}$; the model is estimated on the log-level of the series denoted as $p_{S_i,t}$:

$$p_{S_i,t} = c + A_1 p_{S_i,t-1} + ... + A_p p_{S_i,t-p} + v_t \tag{8}$$

where $p_{S_i,t}$ is a $(S_i \times 1)$ vector of variables, c is a $(S_i \times 1)$ vector of constants, $A_1...A_p$ are $(S_i \times S_i)$ matrices of coefficients and v_t is a $(S_i \times 1)$ vector of disturbances. The estimation of the model for a large set of variables is unfeasible due to the curse of dimensionality.

One solution is to impose restrictions (prior beliefs) on the parameters of the system. Following Banbura, Giannone and Reichlin (2010) we impose Litterman (1986) priors. The coefficients of a matrix A_i, $i = 1, ..., p$ are normally distributed random variables with the mean of the coefficient matrix on the first lag (matrix A_1) equal to an identity matrix I_{S_i} and the mean of all the other coefficients equal to zero. The variance of the parameters depends on a parameter τ which defines the tightness of the priors. A value of τ equal to zero exactly imposes the random walk with drift model on the variables, while a value of τ bigger than zero allows for some variability around the mean of the coefficients and the random walk prior is not exactly imposed.[7] We impose also another type of prior, on the sum of coefficients of the matrices $A_1...A_p$. This prior is imposed by means of another parameter μ. If $A_1 + ... + A_p = I_{n_i}$ the prior is imposed exactly and the specification is equivalent to a VAR in first differences. This will imply that the forecasts will converge to the variable's growth rate.

[7] A scale parameter, to fix the variance of the coefficients, is set by estimating the variance of the residuals from a univariate model of order p on the single variables $p_{j,t}$.

A forecast of the log level of series j at horizon h is then computed as:

$$\hat{p}_{S_i,t+h|t} = \hat{c} + \hat{A}_1 \hat{p}_{S_i,t+h-1|t} + ... + \hat{A}_p \hat{p}_{S_i,t+h-p|t} \tag{9}$$

where $\hat{p}_{S_i,t+h-i|t} = p_{S_i,t+h-i}$ if $i >= h$. In practice, the forecast at time $t+h$ is computed recursively from the forecast at time $t+1$. The estimates of the parameters correspond to the median of the posterior distributions. For each series $j \in S_i$, the annualized h period inflation rate is computed as:

$$\hat{\pi}^{h,BVAR}_{j,t+h|t} = (\hat{p}_{j,t+h|t} - p_{j,t})k \tag{10}$$

The maximum lag length for the AR model is specified ex ante and the actual lag length is chosen according to the Bayes Information Criterion (BIC), while the lag specification for the BVAR is selected by choosing the number of lags that minimize the squared forecast errors of the previous period, including a minimum of 5 lags for the quarterly dataset and a minimum of 13 lags for the monthly dataset. The values of the hyperparamters, μ and τ are chosen on a grid search so that the fitted model has an R^2 as close as possible to 50%. This ensures a reasonable in-sample fit but guards against over-fitting, which leads to poor out-of-sample forecasts.

Recursive Strategy for all Models

The first estimation sample (prior to data transformations) for US data is 1959:Q1-1995:Q2 with forecasts beginning from 1995:Q2+1 to 1995:Q2+h. This is the earliest sample for which weights are available. Final recursive forecasts are computed up to 2009:Q4. For the EA data, the first estimation sample is 1996:M1-2001:M12 and the forecast begins in 2001:M12+1 to 2001:M12+h. The final recursive forecasts end in 2009:M12. Forecast accuracy is evaluated through the Mean Square Forecast Error (MSFE) statistic, however, to facilitate the comparison, the accuracy of a model m is compared (ratio) with that obtained by the Atkeson-Ohanian random walk model, used as the benchmark. Finally, as already mentioned above, some series in the EA dataset are characterized by strong seasonal patterns. To mitigate seasonality the exercise is performed on the year-on-year data transformation. *Eq.*(4) for example is modified as follows:

$$\pi^4_{j,t+h|t} = \alpha^4_j + B^4_j(L)\pi^4_{j,t} + \varepsilon^4_{j,t+h}$$

where the superscript 4 refers to the data transformation.[8] It is not possible to seasonally adjust the Euro Area dataset. The weights that are provided by Eurostat could not be used to construct the aggregate from the component series if the series are transformed. In addition, the seasonaly pattern of the series appears to change over the sample and so a portion of the residuals of the seasonally adjusted series fail standard statistical tests.

[8] As robustness check, in the last section of the paper, we report a set of results for an alternative price transformation (for the US dataset).

6 Results

6.1 AR

The numbers in Table 2 are ratios of the RMSE from the AR model relative to the AO benchmark, with a value less than one indicating the forecast model outperforms the benchmark at the specified horizon. The one and two star superscripts denote a statistically significant improvement in forecast performance relative to the benchmark at the 5% and 10% levels respectively. Due to the nested nature of the models, we use the test statistic proposed by Clark and West (2007). The first part of the table shows the results for the United States. When an AR model is used to forecast the aggregate directly, it is only possible to improve upon the benchmark at the one-quarter horizon. Forecast performance relative to the benchmark gets relatively worse as the horizon increases. When AR models are applied to the disaggregates and combined, the forecasts are improved. The AR model based on the 15-item aggregation provides the best forecasts, as they outperform the benchmark up to 5 quarters ahead. This information is also presented in graphical format in Figure 3. It clearly shows how poorly the aggregate forecast performs relative to the disaggregates over most of the horizon. Another feature evident from the graph is that the greatest gains in forecast accuracy are given by the 3-item and 15-item breakdown. Forecast accuracy decreases, particularly over longer horizons, when the 50-item and 169-item breakdown are used. Considering the US AR model in isolation, any gains in forecasting the disaggregates have already been exploited when the 15-item breakdown is used.

Figure 3: RMSFE of AR Models versus AO (US-YoY)

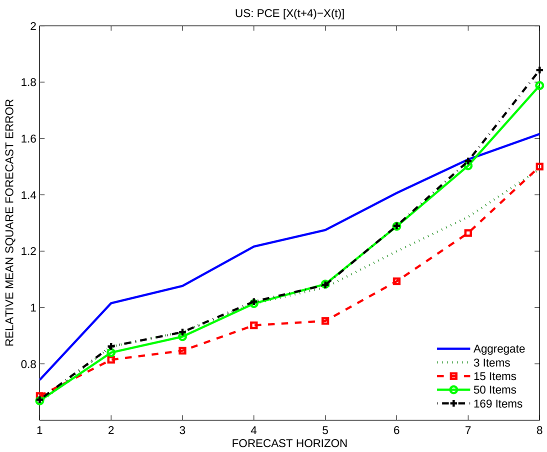
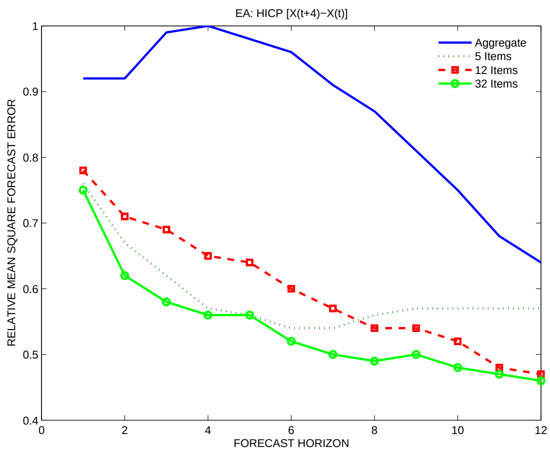


Figure 4: RMSFE of AR Models versus AO (EA-YoY)



The second part of Table 2 presents the results for the Euro Area. The horizon is now twelve months rather than eight quarters and there are only three levels of aggregation rather than four. The results here differ from the US results in a number of ways. The forecasts here are better than the US forecasts in the sense that improvements relative to the benchmark are much greater. Furthermore, the greatest gains relative to the benchmark are at the longer horizons for the Euro Area whereas the

US model forecasts have their greatest gains at the shorter horizons. This behavior is largely explained by the performance of the benchmark. The HICP is not as persistent as the PCE inflation rate. Consequently, the AO benchmark is not as good for the Euro Area, particularly at longer horizons. The key messages are the same as for the US however. The results in the table again show that the aggregate performs poorly relative to the disaggregates. The 32-item breakdown results in the best forecasts. The results are also graphed in Figure 4 and the difference in the performance of the aggregate relative to the disaggregates is quite stark.

6.2 FAAR

The first section of Table 3 documents the forecast performance for the US when a factor is included in the forecast equation. The aggregate now improves upon the benchmark when forecasting up to three quarters in the future. The aggregate forecast still has the least satisfactory performance however, which is also evident graphically in Figure 5. The model based on 169 disaggregates now has the best forecast performance over most horizons. The factor forecasts outperform the simple AR model so that the best forecasts overall for US inflation come from the 169-item model. The EA results with the FAAR are remarkably similar to those for the AR model. The aggregate forecast is strongly outperformed by the disaggregates, with the 32-item breakdown yielding the best results and this pattern is clearly evident in Figure 6. The results of Tables 2 and 3 are supportive of forecast aggregation, with the most accurate forecasts coming from disaggregate models.

Figure 5: RMSFE of FAAR Model versus AO (US-YoY)

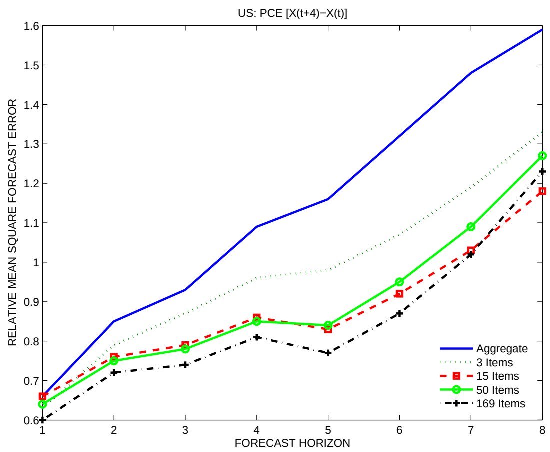


Figure 6: RMSFE of FAAR Model versus AO (EA-YoY)

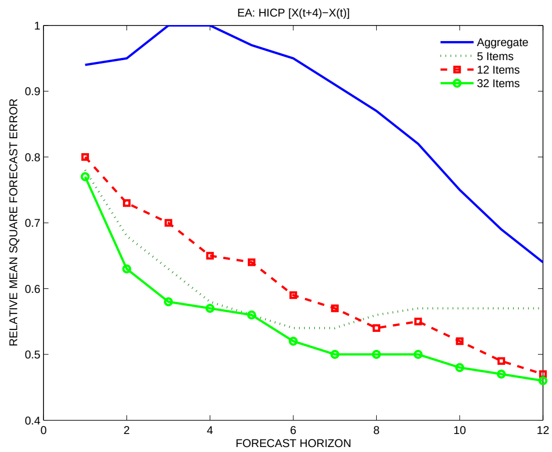


Table 4 documents the change in forecast accuracy when the factor is added to the forecasting equation. There are universal improvements in forecast accuracy for the US data. In addition, the forecasts based on the most detailed breakdown enjoy the greatest improvements in forecast performance. In contrast, for the EA models, the inclusion of the factor leads to virtually no change in the forecasts. To examine the reasons for this, we examine the structure of the dataset to see if there is a large common element to the series. Firstly, we regress the individual series on the factor alone and report the average R^2 in Table 5. We also calculate the average correlation between the series. There is strong commonality in the US PCE dataset. One factor explains 85% of the variation in the aggregate series. The average R^2 declines in line with the number of disaggregates but even at the 169-item level, the average R^2 is 30%. Similarly, the average correlation between the series is high at this level. Commonality is much lower for the euro area inflation series. One factor only explains 37% of the variation in the aggregate series. This drops to about 20% for the disaggregates. Similarly, the average correlation between the series is low.

The strong common element in the US dataset is picked up by the factor model. The simple AR model which excludes the factor is, therefore, mis-specified via the omission of a relevant variable. This will have the usual effect of creating a bias in the coefficients, which will obviously impact the forecasts. Although not reported, the correlation amongst the residuals for the US AR models was found to be far higher than their EA counterparts, as the common factor was captured by the residuals and this imparted much stronger correlation. As described earlier, higher correlation amongst the disaggregate residuals is not good for forecasting, as it leads to forecast errors bunching in a particular direction. This underlying structure in the datasets explains why the factor needs to be included for the US models but not for the EA models.

The one outstanding issue is why the forecasts based on more detailed data improve to a greater extent. The aggregate and 3-item US FAAR models have very modest improvements in forecast power relative to their AR counterparts. The PCE aggregate and the 3-item inflation rates are a weighted average of a large number of underlying inflation rates. Similarly, the factor is a weighted average of all inflation rates. Although the factor weights are calculated to satisfy a maximum variance criterion, the aggregate inflation rate and the 3-item inflation rates are much like factors as they pick up a lot of the commonality in the data. Consequently, the aggregate and 3-item AR models are effectively modelling the common component. Thus, improvements over the AR model are relatively modest when the factor is added as the factor and AR series both model the same component of the data. At the more detailed levels of aggregation, the AR component can pick up the stronger individual dynamics while the factor picks up the common element, which is still meaningful even for the 169-item breakdown. The AR component and the factor are now modelling different behaviour. This is why forecast improvements are greater for the detailed breakdown. This demonstrates the interplay between model choice and the level of aggregation in the data.

6.3 BVAR

Table 6 presents the results of the BVAR model. The US results show that the BVAR is also a fruitful way to exploit the dynamic properties of the data, with the forecast errors again much smaller than those from the standard AR model, particularly for short horizon forecasts. In comparison with the factor model, the BVAR tends to perform better for the short horizons and the factor model does better over the longer horizons. By averaging first by horizon and then by level of aggregation, we find that the BVAR forecasts are equally accurate to the factor model forecasts. The exception to this pattern is the 15-item BVAR, which outperforms the 15-item factor forecast over all horizons. The best individual forecast of all forecast methods considered to date is also the 15-item BVAR, which is 6% more accurate than the best FAAR model, when averaged by horizon. This is the only model which is more accurate than the benchmark at all horizons, although the improvement is not statistically significant at quarter 8. Figure 7 graphs the performance of the BVAR models for the US and the superiority of the 15-item specification is clear.

The best BVAR model for the EA is the 5-item model, which is depicted in Figure 8. However, the results for the EA show that the BVAR fails to improve on the simple AR model. The strong individual dynamics of the series for the HICP means that the simple AR model provides the best forecast. Any attempt to capture common comovement or feedback between the series does not improve the forecasts. The BVAR forecasts are also weaker than the factor model, due to a drop in accuracy for the longer horizons but this is of less significance here as both are outperformed by the AR model.

Figure 7: RMSFE of BVAR Model) versus AO (US-YoY)

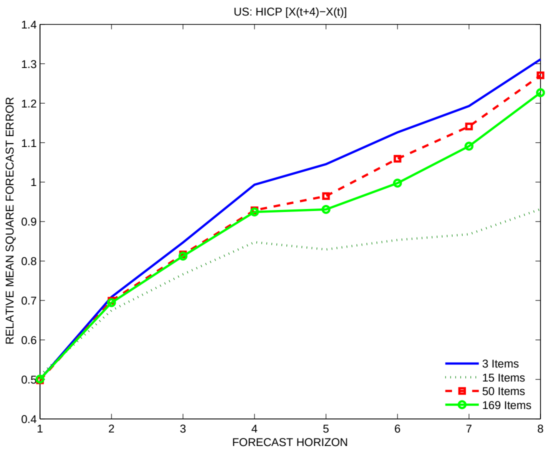
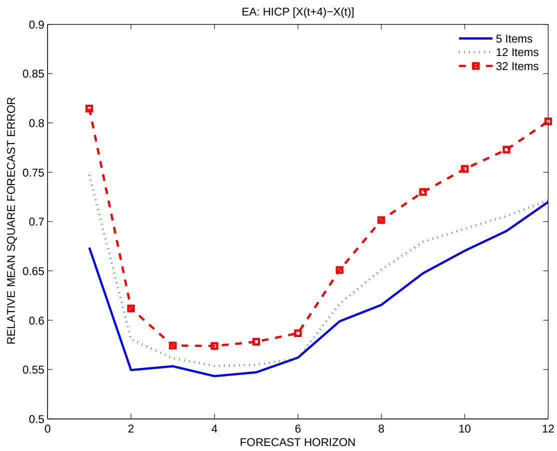


Figure 8: RMSFE of BVAR Model Model versus AO (EA-YoY)



In summary, we have found that the 32-item AR model works best for the Euro Area. In the absence of strong comovement among the series, the AR model is appropriate as there is no omitted variable bias or tendency for the forecast errors to group in the positive or negative directions. For the US, there is a need to model the common component of the series. Without this, there is omitted variable bias and forecast errors will cluster, reducing the chances of getting off-setting errors. This is why the factor model and BVAR models work. For low levels of aggregation, the factor models don't improve upon the AR model greatly given the collinearity between the factors and the AR component. Once we allow for more a detailed breakdown of inflation, the AR components and factor model different components and the common element is modelled in a parsimonious way.

6.4 Alternative Model for US Data

As a robustness check, we consider one alternative model for the US. Given the seasonality issues in the Euro Area dataset, our attention is limited to the US because this model constructs forecasts iteratively using quarter-on-quarter growth. The final type of model considered is a time-varying parameter AR model with stochastic volatility (TV-AR). D'Agostino et al (2009) estimate a TV-AR model for three macro variables in the U.S and find it does particularly well at forecasting inflation. The computational cost of estimating the TV-AR model means that it is only likely to be applied to a small number of items in practice. For this reason, we only conduct a partial exercise in which we estimate the model for the 15-item breakdown. The BVAR with 15 items is the most accurate model so it is instructive to use this as a comparator. We assume that:

$$\pi^1_{j,t} = \delta_{j,t} + \rho_{1,t}\pi^1_{j,t-1} + ... + \rho^1_{p,t}\pi_{j,t-p} + e^1_{j,t} \tag{11}$$

where $\delta_{j,t}$ is the time varying intercept, $\rho_{i,t}$ with $i = 1, ..., p$ are time varying coefficients and $e^1_{j,t}$ is a Gaussian white noise with zero mean and time-varying variance σ^2_t. We assume that σ_t evolves as geometric random walk, belonging to the class of models known as stochastic volatility.

$$log(\sigma_t) = log(\sigma_{t-1}) + u_t \tag{12}$$

Forecasts at time $t + h$ are computed iteratively:

$$\hat{\pi}^1_{j,t+h|t} = \hat{\delta}_{j,t} + \hat{\rho}_{1,t}\hat{p}^1_{j,t+h-1} + ... + \hat{\rho}_{p,t}\hat{p}^1_{j,t+h-p} \tag{13}$$

where $\hat{\pi}^1_{j,t+h-i} = \pi^1_{j,t+h-i}$ if $i >= h$. The estimates of the parameters correspond to the median of the posterior distributions.[9]

A technical issue arises when we generate multi-step expectations; we have to evaluate the future path of drifting parameters. We follow the literature and treat those parameters as if they had remained constant at the current level. See Sbordone and Cogley (2008) for a discussion of the implications of this simplifying assumption.

For each series j, forecasts are first cumulated to recover the h period inflation:

$$\hat{\pi}^{h,TV-AR}_{j,t+h|t} = \frac{1}{h}\sum_{s=1}^{h}\hat{\pi}^{1,TV-AR}_{j,t+s|t} \tag{14}$$

and are then aggregated to recover the forecast for the aggregate index:

$$\hat{\pi}^{h,TV-AR}_{d,t+h|t} = \sum_{j=1}^{N_s} w_{j,t}\hat{\pi}^{h,TV-AR}_{j,t+h|t} \tag{15}$$

The results for this exercise are presented in the first two columns of Table 10. As before, the first results column of the table shows the RMSE of the AO benchmark. The second column shows the forecasts errors of the TV-AR relative to the AO, with

[9]We fix $\lambda_1 = \lambda_2 = 10e^{-02}$. These are the tightness parameters governing the amount of time-variation in the coefficients and volatility respectively.

a value less than one indicating that the TV-AR has the better forecast. The third column compares the TV-AR to the BVAR. The results in the second column show that the forecast errors compare favourably to the benchmark over the entire forecast horizon. When compared to the BVAR, the TV-AR has more accurate forecasts over most horizons. The TV-AR does well for the short-term forecasts but its edge relative to the BVAR steadily declines to the point where the BVAR does better for quarters 7 and 8. Taken on average however, the TV-AR has the better performance with forecasts that are 6% more accurate on average over all horizons. The results demonstrate that combining forecast aggregation with time variation in the parameters and allowing for stochastic volatility can lead to even greater improvements in forecast performance. As the comparison for the TV-AR models is based on 15-items, we graph the results of the 15-item breakdown for all models in Figure 9.[10] It shows that the AR model is not appropriate for a dataset with these properties. There is a big improvement moving to the factor model and further improvements when the BVAR and TV-AR models are used.

7 Summary and Conclusions

In this paper, we conduct an empirical exercise to test if it is possible to achieve gains in forecast accuracy by forecasting the individual components of inflation and aggregating the individual forecasts relative to forecasting the aggregate inflation rate directly. The empirical exercise uses data on both United States and Euro Area inflation. These datasets are quite distinct and require a different modelling approach. We consider four levels of disaggregation for the United States and three for the Euro Area. In all the empirical exercises in this paper, forecast aggregation leads to better forecasts. The aggregate forecast often has the least satisfactory performance and this makes the argument for aggregation more compelling given that multiple levels of aggregation are used.

The performance of the aggregated forecasts also depends on the type of model used. In particular, the model must capture the key characteristics of the data. There is strong comovement in US inflation. Simple AR models do not perform very well in this context but multivariate models such as factor models and BVAR models that can capture this common movement or pick up feedback between the series have more accurate forecasts. For the Euro Area inflation rate, there is far less commonality and the series have more individual dynamics. Simple AR models tend to work well for this type of dataset. They have more accurate forecasts than both the benchmark and their multivariate counterparts. We also discuss the issue of estimation error. We suggest that persistent series should be subject to low estimation error as they can be parameterised parsimoniously.

The exercises are mainly based on multistep forecasts of year-on-year inflation rates. For US inflation, we forecast the h-quarter price change for $h = 1, ..., 8$ and find the

[10] Figure 9 is contained with other graphs at the end of the appendix.

results are robust to this change in the target forecast variable. We also introduce a time-varying model with stochastic volatility where forecasts are constructed iteratively. The time-varying model in conjunction with forecast aggregation leads to further improvements in forecast power. These robustness checks corroborate the main results in favour of forecast aggregation. The paper provides a substantive endorsement of the forecast aggregation approach, particularly in terms of inflation. The key to realising gains in terms of forecast aggregation lies in the ability to uncover the appropriate model for a particular dataset.

References

[24] Atkeson A. and L. Ohanian (2001): "Are Phillips Curve useful for forecasting Inflation", Reserve Bank of Minneapolis, Quarterly Review, 25(1), Pages 2-11.

[24] Banbura M., D. Giannone and L. Reichlin (2010): "Large Bayesian VARs", Journal of Applied Econometrics, Vol. 25(1), pages 71-92,.

[24] Benalal, N., Diaz del Hoyo, J. L., Landau, B., Roma, M. and Skudelny, F. (2004), "To aggregate or not to aggregate? euro area inflation forecasting", *ECB Working Paper 374*.

[24] Bruneau, C., De Bandt, O., Flageollet, A. & Michaux, E. (2007), "Forecasting inflation using economic indicators: The case of France", *Journal of Forecasting*, Vol. 26, pp. 1 - 22.

[24] Cogley, T., and A. M. Sbordone, (2008),"Trend Inflation, Indexation, and Inflation Persistence in the New Keynesian Phillips Curve", *American Economic Review*, Vol. 98(5), pp. 2101-26.

[24] Clark, T., and K. West, (2008),"Approximately normal tests for equal predictive accuracy in nested models", *Journal of Econometrics*, Vol. 138, pp. 291-311.

[24] D'Agostino, A., Gambetti, L. and Giannone, D. (2009), "Macroeconomic Forecasting and Structural Change", *European Central Bank Working Paper No. 1167.*

[24] Domlas, J. (2005), Commentary, *Federal Reserve Bank of Dallas Working Paper 0506.*

[24] Duarte, C. and Rua, A. (2007), "Forecasting inflation through a bottom-up approach: How bottom is bottom?", *Economic Modelling*, Vol. 24, No. 6, November 2007, pp. 941-953.

[24] Enders, W. (2010), *Applied Econometric Time Series, 3rd Edition*, Wiley.

[24] Hendry, D. F. and Hubrich, K. (2006), "Forecasting aggregates by disaggregates", *European Central Bank Working Paper 589.*

[24] Hendry, D. F. and Hubrich, K. (2010), "Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate", *Journal of Business and Economic Statistics*.

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

[24] Kohn, R. (1982), "When is an aggregate of a time series efficiently forecast by its past?", *Journal of Econometrics*, Vol. 18, pp. 337-349.

[24] Luetkepohl, H. (1984a), "Forecasting contemporaneously aggregated vector ARMA processes", *Journal of Business and Economic Statistics*, Vol.2, No.3, pp. 201-214.

[24] Luetkepohl, H. (1984b), "Linear transformations of vector ARMA processes", *Journal of Econometrics*, Vol.26, pp. 283-293.

[24] Luetkepohl, H. (2010), "Forecasting Non-linear Aggregates and Aggregates with Time-Varying Weights", *EUI Working Paper*, No. 2010/11.

[24] Marcellino, M., Stock, J. H. and Watson, M. W. (2003), "Why Has US Inflation Become Harder to Forecast?", *Journal of Money, Credit and Banking*, Vol. 39, pp. 3-33.

[24] Moser, G., Rumler, F. & Scharler, J. (2007), "Forecasting Austrian inflation", *Economic Modelling*, Vol. 24, Issue 3, May 2007, pp. 470-480.

[24] Rose, D.E. (1977), "Forecasting aggregates of independent ARIMA processes", *Journal of Econometrics* Vol. 5, pp.323-345.

[24] Stock, J. H. and Watson, M. W. (2002), "Macroeconomic forecasting using Diffusion Indexes ", *Journal of Business and Economic Statistics*, Vol. 20, No.2, pp. 147-162.

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

[24] Tiao, G. C. and Guttman, I. (1980), "Forecasting contemporal aggregates of multiple time series", *Journal of Econometrics*, Vol. 12, pp. 219-230.

[24] Zellner, A. and Tobias, J. (2000), "A note on aggregation, disaggregation and forecasting performance", *Journal of Forecasting* Vol. 19, pp. 457-469.

Table 1: List of Items in each Aggregate

	HICP Inflation Aggregates		
5 Item List	**12 Item List**	**32 Item List**	
Processed Food	Food + beverages	Food	Health
Unprocessed Food	Alcohol + Tobacco	Non-alcoholic beverages	Purchase of vehicles
Non-Energy Goods	Clothing + Footwear	Alcoholic beverages	Vehicles operation
Energy	Housing	Tobacco	Transport services
Services	Furnishing	Clothing	Postal services
	Health	Footwear	Telephone and telefax
	Transport	Rents for Housing	Electronic Equipment
	Communications	Housing Maintenance	Other durables for recreation
	Recreation + culture	Water supply + misc. services	Recreation, garden and pets
	Education	Electricity, gas and fuels	Recreation services
	Restaurants + hotels	Furniture and furnishings	Reading and stationary
	Miscellaneous	Textiles	Holidays
		Appliances	Education
		Ware and Utensils	Catering services
		Tools and Equipment	Accommodation services
		Routine Maintenance	Miscellaneous
	PCE Inflation Aggregates		
3 Item List	**13 Item List**	**50 Item List**	
Durables	Motor vehicles and parts	New motor vehicles	Exp. abroad by US residents
Non-Durables	Durable household equipment	Used motor vehicles	Less remittances to nonresidents
Services	Rec. goods and vehicles	Vehicle parts	Housing
	Other durable goods	Furniture and furnishings	Household utilities
	Food and bev off-premises	Household appliances	Outpatient services
	Clothing and footwear	Household utensils	Hospital and nursing homes
	Gas and other energy goods	Equipment for house and garden	Motor vehicle services
	Other nondurable goods	Video, audio and IT equipment	Public transportation
	Housing and utilities	Sporting equipment	Parks, theaters,museums etc
	Health care	Sports and recreational vehicles	Audiovisual + IT services
	Transportation services	Recreational books	Gambling
	Recreation services	Musical instruments	Other recreational services
	Food service + accomm	Other durable goods	Food services
	Financial services	Food+ non-alc. bev. off-premises	Accommodations
	Other services	Alcoholic beverages off-premises	Financial services
		Food produced + consumed on farm	Insurance
		Garments	Communication
		Other clothes and footwear	Education services
		Gas + other energy goods	Professional and other services
		Pharmaceutical + medical products	Personal care and clothing services
		Recreational items	Social serv + religious activities
		Household supplies	Household maintenance
		Personal care products	Foreign travel by US Residents
		Tobacco	Less Exp in US by nonresidents
		Newspapers and magazines	Nonprofit Institution Exp.

Note: Some categories have been abbreviation. The list for the 169-item breakdown is available upon request.

Table 2: Forecast Errors for Standard AR Models

	United States					
Quarter	**AO**	**Aggregate**	**3 Items**	**15 Items**	**50 Items**	**164 Items**
1	0.35	0.74*	0.66*	0.69*	0.67*	0.67*
2	0.73	1.02	0.86**	0.81**	0.84**	0.86**
3	1.16	1.08	0.91**	0.85**	0.90**	0.91**
4	1.63	1.22	1.01	0.94**	1.01	1.02
5	1.69	1.28	1.07	0.95**	1.08	1.08
6	1.56	1.41	1.20	1.09	1.29	1.29
7	1.48	1.53	1.32	1.26	1.50	1.52
8	1.32	1.62	1.49	1.50	1.79	1.84
	Euro Area					
Month	**AO**	**Aggregate**	**5 Items**	**12 Items**	**32 Items**	
1	0.07	0.92*	0.76**	0.78**	0.75**	
2	0.18	0.92*	0.67**	0.71**	0.62**	
3	0.33	0.99	0.62**	0.69**	0.58**	
4	0.49	1.00	0.57**	0.65**	0.56**	
5	0.68	0.98	0.56**	0.64**	0.56**	
6	0.83	0.96	0.54**	0.60**	0.52**	
7	0.99	0.91*	0.54*	0.57**	0.50**	
8	1.18	0.87*	0.56*	0.54*	0.49*	
9	1.34	0.81*	0.57*	0.54*	0.50*	
10	1.51	0.75*	0.57*	0.52*	0.48*	
11	1.67	0.68*	0.57*	0.48*	0.47*	
12	1.84	0.64*	0.57*	0.47*	0.46*	

Note: The table presents ratios of RMSE for each model relative to the benchmark. A value less than one indicates that the model has more accurate forecasts than the benchmark. The RMSE of the Atkeson-Ohanian benchmark is in the first column. ** denote forecasts are statically better than benchmark at 10% level with * denoting improvement at the 5% level.

Table 3: Forecast Errors for FAAR Model

Quarter	United States AO	Aggregate	3 Items	15 Items	50 Items	164 Items
1	0.35	0.66*	0.63*	0.66*	0.64*	0.60*
2	0.73	0.85**	0.79**	0.76**	0.75**	0.72**
3	1.16	0.93**	0.87**	0.79**	0.78**	0.74**
4	1.63	1.09	0.96**	0.86**	0.85**	0.81**
5	1.69	1.16	0.98*	0.83**	0.84**	0.77**
6	1.56	1.32	1.07	0.92*	0.95*	0.87*
7	1.48	1.48	1.19	1.03	1.09	1.02
8	1.32	1.59	1.33	1.18	1.27	1.23

Month	Euro Area AO	Aggregate	5 Items	12 Items	32 Items
1	0.07	0.94	0.78**	0.80**	0.77**
2	0.18	0.95	0.68**	0.73**	0.63**
3	0.33	1.00	0.63**	0.70**	0.58**
4	0.49	1.00	0.58**	0.65**	0.57**
5	0.68	0.97	0.56**	0.64**	0.56**
6	0.83	0.95*	0.54**	0.59**	0.52**
7	0.99	0.91*	0.54**	0.57**	0.50**
8	1.18	0.87*	0.56*	0.54**	0.50**
9	1.34	0.82*	0.57*	0.55*	0.50*
10	1.51	0.75*	0.57*	0.52*	0.48*
11	1.67	0.69*	0.57*	0.49*	0.47*
12	1.84	0.64*	0.57*	0.47*	0.46*

See Notes for Table 2 above.

Table 4: Change in RMSE by Including Factor

	United States				
Quarter	**Aggregate**	**3-item**	**15-item**	**50-item**	**169-item**
1	0.89**	0.95**	0.96**	0.95**	0.90**
2	0.83**	0.92**	0.93**	0.89**	0.83**
3	0.86**	0.95*	0.93**	0.87**	0.81**
4	0.90**	0.95**	0.92**	0.84**	0.79**
5	0.91**	0.91**	0.88**	0.78**	0.71**
6	0.94*	0.89**	0.84**	0.74**	0.67**
7	0.97	0.90	0.82**	0.73**	0.67**
8	0.98	0.89	0.79**	0.71**	0.67**
	Euro Area				
Month	**Aggregate**	**5 Items**	**12 Items**	**32 Items**	
1	1.02	1.02	1.02	1.03	
2	1.03	1.02	1.03	1.03	
3	1.01	1.02	1.01	1.01	
4	1.00	1.01	1.01	1.01	
5	0.99	1.01	1.00	1.00	
6	0.99	1.00	1.00	1.00	
7	1.00	1.00	1.00	1.00	
8	1.00	1.01	1.00	1.00	
9	1.01	1.00	1.00	1.00	
10	1.01	1.00	1.00	1.00	
11	1.01	1.01	1.01	1.01	
12	1.01	1.01	1.01	1.01	

Note: The table presents ratios of RMSE for AR models which include a factor to those that don't. It's a measure of the change in forecast accuracy as a result of including the factor in the model. A value less one means the model with the factor more accurate forecasts. Star superscripts have same meaning as before.

Table 5: Commonality within Datasets

PCE	**Aggregate**	**3-item**	**15-item**	**50-item**	**169-item**
R^2	0.85	0.66	0.50	0.37	0.30
Ave. Corr.	n/a	0.73	0.61	0.47	0.39
HICP	**Aggregate**	**5-item**	**12-item**	**32-item**	
R^2	0.37	0.21	0.18	0.21	
Ave. Corr.	n/a	0.16	0.12	0.11	

Note: The individual inflation rates are regressed on the factor only. R^2 is the average for level of aggregation. The second row is average correlation between the inflation rates.

Table 6: Forecast Errors for BVAR Model

	United States				
Quarter	**AO**	**3-item**	**15-item**	**50-item**	**169-item**
1	0.35	0.50**	0.51**	0.50**	0.50**
2	0.73	0.71**	0.67**	0.70**	0.69**
3	1.16	0.85**	0.77**	0.82**	0.81**
4	1.63	0.99*	0.85**	0.93**	0.92**
5	1.69	1.05	0.83**	0.96**	0.93**
6	1.56	1.13	0.85**	1.06	1.00
7	1.48	1.19	0.87**	1.14	1.09
8	1.32	1.31	0.93	1.27	1.23
	Euro Area				
Month	**AO**	**5 Items**	**12 Items**	**32 Items**	
1	0.07	0.67**	0.75**	0.81**	
2	0.18	0.55**	0.58**	0.61**	
3	0.33	0.55**	0.56**	0.57**	
4	0.49	0.54**	0.55**	0.57**	
5	0.68	0.55**	0.55**	0.58**	
6	0.83	0.56**	0.56**	0.59**	
7	0.99	0.60**	0.62**	0.65**	
8	1.18	0.62*	0.65**	0.70*	
9	1.34	0.65*	0.68**	0.73*	
10	1.51	0.67*	0.69*	0.75*	
11	1.67	0.69*	0.71*	0.77**	
12	1.84	0.72*	0.72*	0.80**	

Note: The table presents the performance of the BVAR. As this is a is a multivariate model, there are no results to report for the aggregate alone. Star superscripts have same meaning as before.

Table 7: US Time-Varying AR Model Based on 15 Items

	$P_{t+h} - P_t$		
Quarter	**AO**	**TV-AR/AO**	**TV-AR/BVAR**
1	0.35	0.78	0.88
2	0.73	0.75	0.87
3	1.16	0.74	0.87
4	1.63	0.73	0.89
5	1.69	0.69	0.90
6	1.56	0.72	0.94
7	1.48	0.79	1.03
8	1.32	0.88	1.11

Note: The second results column shows the RMSE of the 15-item TV-AR while the third column is the ratio of the RMSE of the 15-item TV-AR to the 15-item BVAR.

Appendix 1: Alternative Data Transformation

As a robustness check, we perform the forecasting exercise for a different price transformation, the h-level change in log prices $p_{t+h} - p_t$. This is analyzed only for the US dataset, given that the seasonal issues with the EA data makes it difficult to look at alternative data transformations beyond the year-on-year inflation rate, which is analyzed in the main body of the paper. At a given horizon h, the AR and FAAR forecasting equations are exactly those of $eq.(4)$ and $eq.(5)$ respectively. Forecasts with the BVAR model are computed exactly as before, but the log-level of price at time t is then subtracted by the forecast of p_{t+h} in order to recover the h-level change of log-prices.

Appendix Table A1 presents the results for the standard AR model. The aggregate does not perform well, as it beats the benchmark only for the one-period forecast. The 3-item and 15-item forecasts are both more accurate than the benchmark up to five quarters. The results are presented graphically in Figure 9. The pattern mimics the year-on-year results, where the 3-item and 15-item forecasts are far better than the benchmark. Table A2 presents the results with the factor included. In further agreement with the year-on-year results, the aggregate still has the worst average performance and the 169-item model now has the most accurate forecasts. Figure 10 plots the results and it demonstrates how quickly the performance of the aggregate deteriorates over the forecast horizon from a strong starting position. Table A3 shows that the BVAR models also improve significantly on the AR specification, especially in relation to the long-range forecasts. The performance of the BVAR models is graphed in Figure 11. For the multistep forecasts in the previous section, the 15-item BVAR is the only breakdown which is clearly better than its factor comparator. For the forecasts considered here, the 3-item BVAR also outperforms the 3-item factor model due to good performance at the short horizons. Thus, when we compare the two model types averaged by all their forecasts, the BVAR is more accurate than the factor model by approximately 7%. As before, the best BVAR model is still based on 15 items. Overall, these results strongly mirror those of the year-on-year specification. In this exercise, the aggregate model never provides the best forecasts and often provides the worst. The key properties of the data affect the disaggregate forecasts in the same way irrespective of the data transformation.

Table A1: Errors for AR model Forecasts of $P_{t+h} - P_t$

	United States					
Quarter	**AO**	**Aggregate**	**3 Items**	**15 Items**	**50 Items**	**164 Items**
1	3.09	0.90	0.92	0.92	0.90	0.93
2	2.27	1.05	0.96	0.92	0.93	0.96
3	1.82	1.16	0.98	0.96	0.99	1.04
4	1.63	1.17	0.98	0.99	1.06	1.09
5	1.42	1.14	0.99	0.98	1.09	1.15
6	0.98	1.29	1.10	1.15	1.35	1.48
7	0.77	1.45	1.23	1.37	1.65	1.85
8	0.62	1.54	1.35	1.65	2.02	2.35

Note: The table presents ratios of RMSE for each model relative to the benchmark. Only applies to US data.

Table A2: Errors for FAAR Model Forecasts of $P_{t+h} - P_t$

	United States					
Quarter	**AO**	**Aggregate**	**3 Items**	**15 Items**	**50 Items**	**164 Items**
1	3.09	0.85	0.93	0.91	0.88	0.89
2	2.27	0.97	0.96	0.90	0.87	0.84
3	1.82	1.09	1.00	0.92	0.90	0.85
4	1.63	1.11	1.01	0.93	0.91	0.84
5	1.42	1.09	1.01	0.88	0.88	0.81
6	0.98	1.28	1.09	0.91	0.94	0.86
7	0.77	1.46	1.19	1.00	1.05	0.97
8	0.62	1.54	1.27	1.09	1.17	1.10

Note: The table presents ratios of RMSE for each model relative to the benchmark. Only applies to US data.

Table A3: Errors for BVAR Model Forecasts of $P_{t+h} - P_t$

	United States				
Quarter	**AO**	**3-item**	**15-item**	**50-item**	**169-item**
1	3.09	0.84	0.89	0.90	0.91
2	2.27	0.81	0.86	0.92	0.92
3	1.82	0.81	0.86	0.94	0.95
4	1.63	0.80	0.82	0.92	0.94
5	1.42	0.78	0.76	0.91	0.93
6	0.98	0.88	0.76	0.92	0.92
7	0.77	0.97	0.76	0.93	0.93
8	0.62	1.10	0.79	0.96	0.93

Note: The table presents the performance of the BVAR. As this is a is a multivariate model, there are no results to report for the aggregate alone.

Figure A1: RMSFE of AR Model versus AO (US)

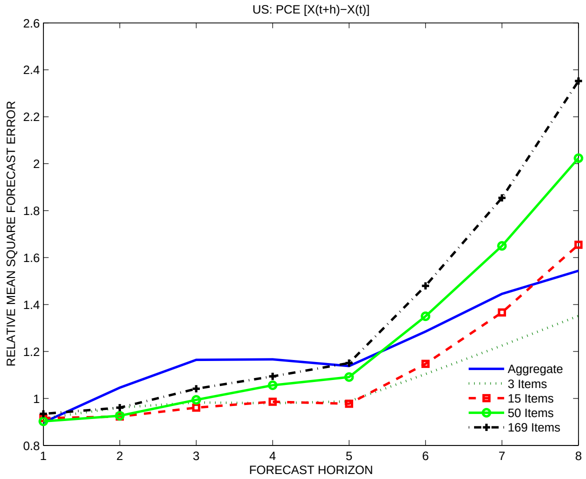


Figure A2: RMSFE of FAAR Model versus AO (US)

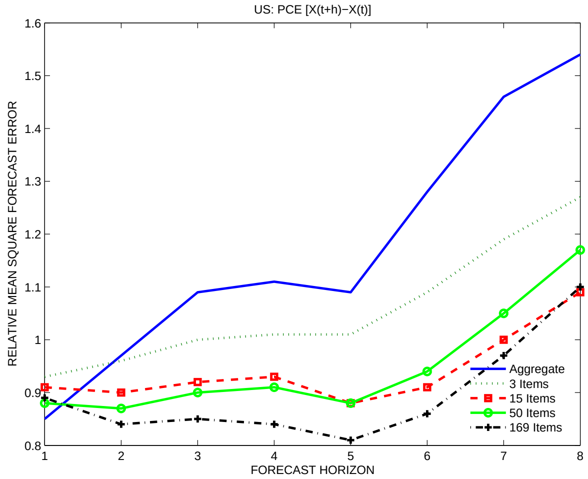


Figure A3: RMSFE of BVAR Model versus AO (US)

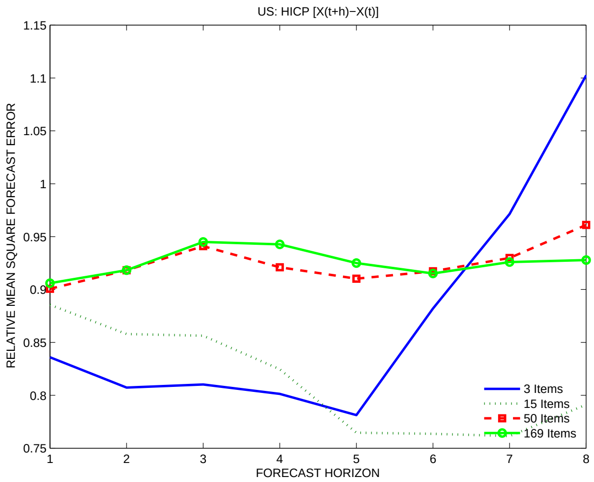


Figure 9: RMSFE of the AR, FAAR, BVAR and TV-AR Models versus AO (US - 15 items)