Forecast Densities for Economic Aggregates from Disaggregate Ensembles[*]

Francesco Ravazzolo[†] (Norges Bank and BI)

Shaun P. Vahey[‡] (ANU)

June 8, 2012

Abstract

We extend the 'bottom up' approach for forecasting economic aggregates with disaggregates to probabilistic forecasting. Our methodology utilises a Linear Opinion Pool to combine the forecast densities from many disaggregate forecasting specifications, using weights based on the Continuous Ranked Probability Score. We also adopt a post-processing step prior to forecast combination. These methods are adapted from the meteorology literature; see, the discussions in Gneiting and Thorarinsdottir (2010). In our application, we use our approach to forecast US Personal Consumption Expenditure inflation from 1990q1 to 2009q4. Our ensemble combining the evidence from 16 disaggregate PCE series outperforms an integrated moving average specification for aggregate inflation in terms of density forecasting.

Keywords: Ensemble forecasting, density combinations, disaggregates, opinion pool
JEL codes: C11; C32; C53; E37; E52

[*] We benefited greatly from discussions with Todd Clark, Anthony Garratt, Kirstin Hubrich, Christian Kascha, James Mitchell, Kalvinder Shields, Tara Sinclair, Michael Smith and Simon van Norden. We thank conference and seminar participants at Melbourne University, Oslo University, the European Central Bank, the Veissmann European Research Centre, the Society for Nonlinear Dynamics and Econometrics 17th Annual Symposium, and the 6th Eurostat Colloquium. The views expressed in this paper are our own and do not necessarily reflect those of Norges Bank. We thank the ARC, Norges Bank, the Reserve Bank of Australia and the Reserve Bank of New Zealand for supporting this research (LP 0991098).

[†] Norges Bank and BI Norwegian Business School. francesco.ravazzolo@norges-bank.no

[‡] *Corresponding author*: Shaun Vahey, ANU. spvahey@gmail.com

1 Introduction

Policymakers regularly combine the leading evidence in disaggregate variables to perform probabilistic assessments of aggregate behaviour; see, for example, Greenspan (2004), and the discussions by Feinstein, King, and Yellen (2004). However, the techniques used by central banks in practice to incorporate disaggregate information into probabilistic assessments for monetary policy purposes remain informal.

The scope for producing density forecasts for economic aggregates based on disaggregate information has not been explored in previous economics studies. This is surprising given the widespread recognition that evaluations of point forecast accuracy are only relevant for highly restricted loss functions. More generally, complete probability distributions over outcomes provide information helpful for making economic decisions; see, for example, the discussions in Granger and Pesaran (2000), Timmermann (2006) and Gneiting (2011). Accordingly, several central banks, including the Bank of England, Norges Bank, Sveriges Riksbank have committed to publishing density or interval forecasts for macroeconomic aggregates in recent years.

In contrast to the informal methods currently utilised by central bankers to incorporate disaggregate information into probabilistic assessments, many practitioners within central banks favour a particular methodology for producing point forecasts known as the 'bottom-up' approach; see Lütkepohl (2009) for a survey of methods for forecasting economic aggregates. The 'bottom-up' approach is a two-step procedure, in which a system of equations is used to forecast the disaggregate series in the first step, and then the aggregate forecast is constructed by feeding in the disaggregate forecasts (augmented with an assumption about the time series behaviour of the index weights).

In this paper, we propose an ensemble approach to build up the evidence in disaggregate series to make probabilistic forecasts for an economic aggregate. We formulate the forecasting problem as one in which a forecaster (recursively) selects a linear combination of component forecast densities to produce a forecast density for the aggregate. Each component forecast is produced from an autoregressive linear time series model for a single disaggregate series. The resulting ensemble approximates the many unknown relationships between the disaggregates and the aggregate using time-varying weights across the disaggregate forecast densities.

A key insight of our paper is that the 'bottom-up' approach to aggregate forecasting commonly favoured by practitioners in central banks constitutes a form of ensemble forecasting. By this approach, the researcher considers a model space comprising a large number of forecasting models, generated by varying measurements and/or model specifications. The predictive densities from the many misspecified forecasting models can be combined in many ways. In this paper, we utilise the Linear Opinion Pool, following (among others) Jore, Mitchell and Vahey (2010) in their analysis of forecasting with vector autoregressions. Regardless of the technology utilised for combination in practice, the aim of ensemble forecasting is to approximate the unknown process with a large number of misspecified forecasting specifications. Hence the methodology is based on a Bayesian perspective, although the component models can be estimated and combined by either frequentist or Bayesian methods.

Exploiting the close correspondence between the 'bottom-up' approach and ensemble forecasting, we consider a number of computational techniques adapted from the meteorology literature; see, for example, Gneiting and Thorarinsdottir (2010). These include: using time-varying weights (that differ from the index weights) based on the continuously ranked probability score; and a post-processing step to adjust the location of the forecast densities prior to construction of the ensemble predictive densities.

In our application based on US Personal Consumption Expenditure deflator data, we assess the forecast performance of the disaggregate ensemble utilising both of these techniques. Our ensemble densities for inflation, based on 16 disaggregate series, outperform densities from both a first-order moving-average process for the change in aggregate inflation, and from a simple aggregate autoregressive model, over the out of sample period 1990q1 to 2009q4.

The remainder of this paper is structured as follows. In Section 2, we describe our methods for modelling the relationship between the economic aggregate and the disaggregates. In Section 3, we apply our methodology to US data to produce aggregate inflation forecast densities from an ensemble system utilising disaggregate information. We compare and contrast the predictive densities with those resulting from our alternative specifications which ignore disaggregate information. In the final section, we conclude with some suggestions for further research.

2 Disaggregate Ensemble Forecasting Methodology

We begin with a characterisation of the 'bottom-up' approach commonly used by practitioners in central banks for one step ahead point forecasting.

2.1 The 'Bottom-up' Approach

Consider an economic aggregate, y_τ, defined as the weighted arithmetic mean of the N disaggregates, $x_{i,\tau}$:

$$y_\tau \equiv \sum_{i=1}^{N} \omega_{i,\tau}\, x_{i,\tau}, \qquad i = 1, \ldots, N, \qquad \tau = \underline{\tau}, \ldots, \overline{\tau} \tag{1}$$

with weights $\omega_{i,\tau}$ that are between 0 and 1, $0 < \omega_{i,\tau} < 1$, and sum to one, $\sum_i \omega_{i,\tau} = 1$, across the disaggregates, indexed $i = 1, \ldots, N$. Given the weights, $\omega_{i,\tau}$, and the disaggregate series, $x_{i,\tau}$, equation (1) defines the economic aggregate, y_τ over the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$. Forecasting the aggregate y_τ conditional on the information set dated $\tau - 1$, the researcher faces two unknowns: the disaggregate variables, $x_{i,\tau}$ and the weights, $\omega_{i,\tau}$.

The disaggregate forecasts are typically badly behaved in practice because the forecasting equations are misspecified. To illustrate, suppose the $N \times 1$ vector of disaggregates $\mathbf{x}_\tau$ follows the vector autoregression (VAR) process:

$$\mathbf{x}_\tau = \mathbf{a} + \mathbf{b}_1\mathbf{x}_{\tau-1} + \mathbf{b}_2\mathbf{x}_{\tau-2} + \cdots + \mathbf{b}_p\mathbf{x}_{\tau-p} + \epsilon_\tau \qquad \tau = \underline{\tau}, \ldots, \overline{\tau} \tag{2}$$

with $\epsilon_\tau \sim N(\mathbf{0}, \Xi_\tau)$. Since economic theory restricts neither the interdependence between the disaggregates, nor the number of own lags for each disaggregate, the dimension of the VAR has the potential to be very large. And as van Garderen et al (2000) and Lütkepohl (2010) point out, there is no reason to restrict disaggregate relationships to be linear (or even Gaussian). Even small dimensional VARs often produce weak forecasting performance in macroeconomic time series applications, afflicted by a variety of misspecification issues; see (among others) Clark and McCracken (2010). These include unknown non-linearities, structural breaks and measurement errors.

Given the prevalence of uncertain instabilities in macro time series, 'bottom up' practitioners abstract from the dependence between disaggregates entirely, restricting attention to an autoregessive prediction model for each disaggregate series. For example, the first order autoregression, AR(1), for disaggregate i:

$$x_{i,\tau} \;=\; c_i + d_i x_{i,\tau-1} + \eta_{i,\tau} \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{3}$$

with $\eta_i \sim N(0, \Psi_i)$. Equation (3) is a single equation from a restricted version of the VAR given by equation (2). Although this equation is of the type typically deployed by 'bottom-up' practitioners, the restrictions cannot be motivated by economic theory, and the linear Gaussian autoregressive forecasting equation is misspecified.

Nevertheless, to 'bottom-up' a point forecast for the economic aggregate, a system of N equations of this form is used to generate the disaggregate point forecasts, $x^e_{i,\tau}$. These are passed through equation (1) (with an assumption about the index weights, $\omega_{i,\tau}$) to provide the point forecast for the aggregate y_τ. Practitioners in central banks typically build-up a point forecast in this manner for an economic aggregate based on one step ahead disaggregate projections, using the previous period's weights, $\omega_{i,\tau-1}$; see, for example, Marcellino, Stock and Watson (2003), and the discussions in Lütkepohl (2009, 2010).[1] Forecast performance deteriorates if the weights change through time, or if the disaggregate forecasts are inaccurate. Lütkepohl (2010) emphasises that the weights $\omega_{i,\tau}$ are generally time-varying and ex ante unknown in real-time economic applications. Problems include changes to the definition of a disaggregate, chain-linking, periodic rebasing

[1] Marcellino, Stock and Watson (2003) refer to the 'bottom-up' approach as forecast 'pooling'. Others prefer the term 'build-up'.

of indexes, benchmark revisions, and changes in the number of disaggregates. Many of these effects are more pronounced at longer forecasting horizons. Many of these effects are more substantial when forecasting over long horizons. Perhaps with these concerns in mind, central banks generally use the approach for one step ahead forecasting in practice

In summary, 'bottom-up' practitioners use the following approach. Ignore the dependence between disaggregate time series, utilise simple autoregressive time series models to produce one step ahead forecasts for each disaggregate, and produce an aggregate real-time forecast by assuming the weights are unchanged from the previous period.

The use of 'bottom up' point forecasting by practitioners is controversial. For example, assuming the weights are known and fixed ex ante, Hendry and Hubrich (2011) argue that the 'bottom up' strategy cannot approximate a 'true' multivariate model. They advocate a simple aggregate autoregressive model, augmented with selected lagged disaggregates, in preference to the 'bottom up' approach based on root mean squared forecast error. Lütkepohl (2010) suggests a similar strategy for the case with time-varying and unknown weights.

Although the finding by Hendry and Hubrich (2011) is important for 'bottom up' practitioners concerned with point forecasting, as argued above, central banks require probabilistic information for policy. Root mean squared forecast error can be misleading about the accuracy of forecast probabilities. In contrast to Hendry and Hubrich (2011), our methodology is aimed at density forecasting.

2.2 An Extension of the 'Bottom-up' Approach Based on the Linear Opinion Pool

Equation (1) provides no guidance on how to produce aggregate densities. Regardless of whether the weights are known ex ante, a 'bottom-up' approach uses ad hoc exclusion restrictions to incorporate disaggregate information, and the model space is incomplete by construction. Timmermann (2006) discusses options for forecast aggregation from the perspective of the forecast combination literature. In this section, we propose using the

Linear Opinion Pool (LOP) to aggregate.

Opinion pools have a long tradition in management science for expert combination problems, where the framework is sometimes referred to as a 'mixture of experts'. As emphasised by Wallis (2005), the approach is particularly useful for the combination of survey information since no in-sample information is required about the model used by each expert; see also Mitchell and Hall (2005). Geweke (2009) discusses the differences between LOP and mixture models, and argues that the former is more appropriate if the model space is incomplete—where all the models considered are misspecified.

More formally, given $i = 1, \ldots, N$ disaggregates (where N could be a large number), the forecaster constructs a predictive density for the economic aggregate by taking a convex combination, sometimes referred to as a LOP, of the disaggregate densities. The disaggregate ensemble (DE) density is defined as:

$$DE \;=\; g(y_\tau) \;=\; \sum_{i=1}^{N} w_{i,\tau}\, h(y_\tau \mid I_{i,\tau}), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{4}$$

where $h(y_\tau \mid I_{i,\tau})$ are the one step ahead forecast densities from component model (based on disaggregate information), indexed $i = 1, \ldots, N$, for the economic aggregate y_τ, conditional on the information set $I_{i,\tau}$, which contains information dated $t-1$ and earlier. The non-negative weights, $w_{i,\tau}$, in this finite mixture sum to unity, are positive, and vary by recursion in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$. For convenience, we write the forecast densities to be aggregated as $h(y_{i,\tau})$.

Notice that the predictive density for the aggregate given by the LOP will be a mixture of the forecast densities produced by the component disaggregate forecasting specifications. If all the disaggregate forecasting specifications are misspecified, there is no reason to restrict attention to a forecast density for the aggregate, $g(y_\tau)$, to be from the same distributional family as the components themselves. (Kascha and Ravazzolo (2010) discuss alternative opinion pools which satisfy this restriction, in which case the methodology is said to be 'externally Bayesian'.) Hence, a useful feature of our LOP approach is that aggregate forecast densities can capture non-Gaussian behaviour, even if the disaggregate forecast densities are Gaussian.

An implication of the LOP approach is that the index weights, $\omega_{i,\tau}$, will not generally be useful for forecasting. There is no reason to believe that applying the index weights (or if these are unknown, lagged index weights) to forecast densities produced from misspecified linear disaggregate forecasting specifications will provide a suitable approximation.

2.3 Disaggregate Component Model Space

Having set out the construction of the aggregate forecast densities, $g(y_\tau)$, we now describe how to generate the disaggregate forecast densities $p(x_{i,\tau})$, and the predictives for the aggregate $h(y_{i,\tau})$ which enter the LOP, equation (4). To construct the forecast densities $p(x_{i,\tau})$, we adopt the disaggregate component model space commonly utilised in the 'bottom-up' approach to point forecasting. Namely exactly the same N univariate autoregressive models described by equation (3).

As we noted earlier, we interpret the N forecasting specifications used in conventional 'bottom-up' analyses as a 'perturbed' ensemble model space. In ensemble forecasting applications, researchers consider perturbations to a single basic misspecified model.[2] The perturbations might be to the measurements and/or the model space; see (among others) Raftery et al (2005), Bao et al (2010) and Doblas-Reyes et al (2009). In our economic application, the N disaggregate forecasting models can be trivially rewritten as an autoregressive forecasting model for a single variable, where the variable of interest is systematically perturbed to consider each of the candidate disaggregate variables in turn. That is, equation (3) can be rewritten:

$$z_\tau \; = \; c_z + d_z z_{\tau-1} + \eta_{z,\tau} \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \qquad z_\tau = x_{1,\tau}, \ldots, x_{N,\tau}. \tag{5}$$

To construct the forecast densities for the aggregate, $h(y_{i,\tau})$, which are inputs into the LOP, equation (4), we use a post-processing step for the disaggregate densities, $p(x_{i,\tau})$. This procedure adjusts the location of the disaggregate forecast densities prior to con-

[2] Bache et al (2010) note that the technologies for ensemble density construction differ across applied statistics fields, as does the model space. For example, in weather forecasting applications, the models are typically chaotic.

structing the ensemble density forecast; see, for example, the discussions in (among others) Atger (2003), Stensrud and Yussouff (2007), Bao et al (2010) and Gneiting and Thorarinsdottir (2010). With the disaggregate forecast densities (approximately) correctly located, the weights in the LOP are governed by shape considerations. Although more flexible approaches are feasible, a simple bias-correction step is often sufficient to ensure well-calibrated ensemble densities in practice; see, for example, Stensrud and Youssoff (2007).

To implement this post-processing step, we recursively estimate the Linear Gaussian model:

$$y_s \;=\; a + (x^e_{i,s}) + \varepsilon_s, \qquad s = \underline{s}, \ldots, \tau - 1 \tag{6}$$

where $x^e_{i,s}$ is the expected value (median) of the predictive density $p(x_{i,s})$ from equation (5), for all i. Then, we define the bias-corrected disaggregate forecast density for the aggregate:

$$h(y_{i,\tau}) \;=\; \widehat{a} \;+\; p(x_{i,\tau}) \tag{7}$$

where $\widehat{a}$ is the estimate of a in equation (6).

We note that it is not necessary to utilise this two-step approach with the LOP. For example, in our application below we experimented with ensemble specifications that predict the aggregate directly with the disaggregates. However, we found the additional flexibility afforded by the two step approach gave considerable improvements in forecasting performance.

2.4 Ensemble Weights

Finally, to construct the ensemble forecast density for the aggregate, $g(y_\tau)$ via the LOP, equation (4), we require weights, $w_{i,\tau}$. Since our methodology is motivated by the assumption that the disaggregate forecasting equations are misspecified and that the ensemble methods approximate the unknown 'true' specification, we propose recursively updating the weights, $w_{i,\tau}$, according to the density forecasting performance of the (bias-corrected) forecast predictives, $h(y_{i,\tau})$.[3] Hersbach (2000), Gneiting and Raftery (2007), Panagiotelis and Smith (2008), Groen, Paap and Ravazzolo (2009) and Arora et al (2013) (among others) have argued that the Continuous Ranked Probability Score (CRPS), which rewards predictive densities with high probabilities near (and at) the outturn, provides a robust metric of density forecast performance. Gneiting and Raftery (2007) refer to the concentration of a forecast density about its central location as 'sharpness', and the location as 'distance'. The CRPS metric favours densities with small distance and high sharpness.

The CRPS is measured as the difference between the predicted and actual cumulative distribution. Figure 1 provides an illustrative example for a particular observation: the CRPS measures the area between the predictive (for this example, assumed to be Gaussian) and the actual cumulative distribution (marked by shading). The (positive) score approaches zero as the predictive density converges on the true (but unobserved) density.

More formally, following Panagiotelis and Smith (2008), the CRPS of a component density for a particular observation can be defined as:

$$CRPS = E_h|y - Y| - 0.5E_h|y - y'| \tag{8}$$

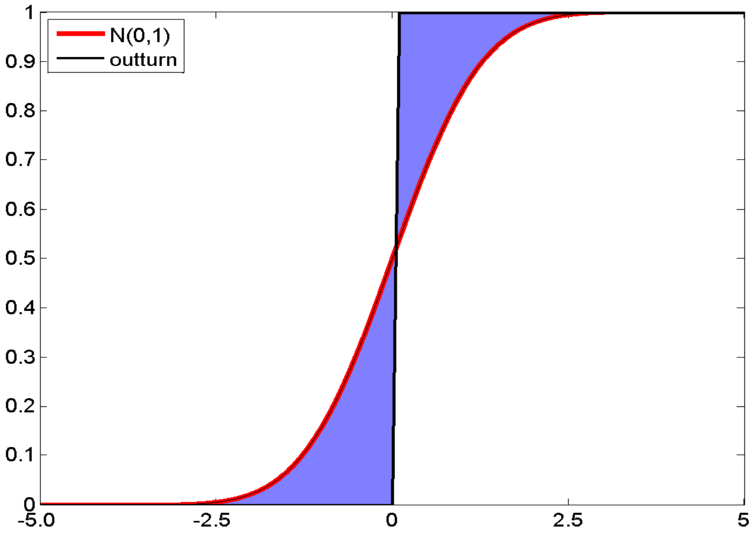
where E_h is the expectation for the predictive $h(Y_\tau)$, y and y' are independent random draws from the predictive, and Y is the observed outturn. The expectation terms can be approximated using Monte Carlo draws from the component forecast density; Panagiotelis and Smith (2008, equation 4.5) provide the computational steps required.

For each forecast density, $h(y_{i,\tau})$, we construct the cumulative CRPS over the evaluation period. The weight on an individual component density i in each observation of the evaluation period is then calculated by:

$$w_{i,\tau} = \frac{\left[\sum_{\underline{s}}^{\tau-1} \Gamma(h(Y_{i,\tau}))\right]}{\sum_{i=1}^{N}\left[\sum_{\underline{s}}^{\tau-1} \Gamma(h(Y_{i,\tau}))\right]}, \qquad \tau = \underline{s}, \ldots, \underline{\tau}, \ldots, \overline{\tau}. \tag{9}$$

[3] In macro-econometric studies of point forecast combination with US data, it is often observed that equal-weight combinations compare favourably with recursive-weight combinations. Jore, Mitchell and Vahey (2010) and Garratt, Mitchell, Vahey and Wakerly (2011) show that result does not generalise to forecast densities.

Figure 1: CRPS



Note: The figure shows the cumulative distribution of a normal density with zero mean and unit variance, N(0,1), and the cumulative distribution of the realised value 0. The coloured area measures the CRPS.

with Γ is the inverse of the CRPS, $0 \leq \Gamma \leq \infty$, and higher scores are preferred.

Geweke (2009) argues for optimal combinations based on maximising the logarithmic score of the ensemble density. Although, in this paper, we use an alternative scoring rule based on the CRPS to construct the weights, optimised CRPS weights are feasible. But given the short samples used in macroeconomic forecasting applications, the scope for improvement is modest, and we leave this avenue for subsequent research, utilising longer samples of data.[4]

[4] The weights in the LOP in our application do not converge to zero for any specification. Our proposed weights do not lead to an outcome close to model selection on our sample.

2.5 Methodological Summary

Our disaggregate ensemble methodology can be summarised as follows. For each observation in the forecaster's evaluation period, we estimate N univariate time series representations, one for each disaggregate. The 'fit' of each bias-corrected component forecast density is assessed with the CRPS, and used to construct weights for the ensemble forecast density. These weights vary through the evaluation period. In this manner, we approximate the forecast densities for the true, but unknown, relationships between the disaggregates and the aggregate. An appendix to this paper provides some simulations to illustrate our approach further.

3 Application: Forecasting PCE Inflation for the US

In our forecasting US inflation application, we consider US Personal Consumption Expenditure deflator (PCE) data. We construct a disaggregate ensemble using an evaluation period from 1975q1 to 2009q3, and then examine the calibration of the ensemble aggregate forecast densities using probability integral transforms, *PITS*, at the end of the evaluation. We also examine forecast performance relative to a number of aggregate benchmarks. We stress that our focus in this example is the predictive performance of the ensemble. We do not aim to select a preferred single disaggregate predictor of aggregate inflation from the (likely) misspecified disaggregate components.

We begin our analysis by describing the US data. Then we describe our disaggregate ensemble, aggregate benchmarks, density evaluation methods, and results.

3.1 Data

The dataset contains time series for the disaggregate components of the PCE. The data are available on the Bureau of Economic analysis http://www.bea.gov/national/nipaweb.[5]

[5] To our knowledge, the disaggregate data used in this study are not available on a real-time basis, although Croushore (2009) discusses the revisions in aggregate PCE.

The PCE data permit breakdowns at various levels of disaggregation. We emphasise that, in principle, our methodology could be applied to any level of disaggregation. In our application, we illustrate our technique with 16 disaggregates. These are: Motor vehicles and parts, Furnishings and durable household equipment, Recreational goods and vehicles, Other durable goods, Food and beverages, Clothing and footwear, Gasoline and other energy goods, Other nondurable goods, Housing and utilities, Health care, Transportation services, Recreation services, Food services and accommodations, Financial services and insurance, Other services, and Final consumption expenditures of nonprofit institutions serving households. For all inflation series, the PCE aggregate and its disaggregates, we work with the quarterly growth rates (calculated as 100 times the log difference in the price levels). Clark (2006) documents both the considerable variation in the mean and volatility of disaggregate PCE variables through time, and the heterogeneity across disaggregates.

3.2 Disaggregate Ensemble and Aggregate Specifications

We start our in-sample estimation with 1975q1 and end in 2009q4. With our out-of-sample evaluation period (τ) from 1990q1 ($\underline{\tau}$) to 2009q4 ($\overline{\tau}$), the period from 1985q1 to 1989q4 comprises a 'training period' to initialise the ensemble weights. The bias-correction step and ensemble combinations are based on a rolling window of 20 quarters, denoted $s = \tau - 20, \ldots, \tau - 1$, for the results reported below. (Using an expanding window for bias-correction and combination gave some degradation in relative performance but the qualitative results are unchanged.)

The 16 disaggregate (component) forecasting equations each utilise an AR(2) specification, estimated by Bayesian methods, using non-informative priors and a rolling window of 40 observations. (Using an AR(4) disaggregate forecasting equation gives qualitatively similar performance.) The predictive densities follow the t-distribution, with mean and variance equal to OLS estimates; see, for example, Koop (2003, chapter 3) for details.

In addition to our disaggregate ensemble, DE, we also evaluate the predictive densities from two time series models for aggregate inflation. The first uses a linear model for the aggregate, that is, using a linear autoregressive model for aggregate PCE inflation, with two lags, AR(2). We use noninformative priors for the AR(2) parameters with a rolling window of 40 observations for parameter estimation.[6]

The second aggregate variant uses a first-order moving average process for the change in inflation—that is, an Integrated Moving Average (IMA) process for aggregate inflation. Clark (2011) reports that this model outperforms AR benchmarks in terms of US inflation density forecasting. Following Clark (2011), we use a 40 observation moving window for in-sample estimation of the parameters.[7]

3.3 Density Evaluation

Following (among others) Jore, Mitchell and Vahey (2010), we evaluate the ensemble predictive densities using a battery of (one-shot) tests of absolute forecast accuracy, relative to the 'true' but unobserved density. Like Rosenblatt (1952) and Diebold, Gunther and Tay (1998), we utilize the probability integral transforms, *PITS*, of the realisation of the variable with respect to the forecast densities. A forecast density is preferred if the density is correctly calibrated, regardless of the forecasters loss function. The *PITS* are:

$$z_\tau = \int_{-\infty}^{y_\tau} g(u)du.$$

The *PITS* should be both uniformly distributed, and independently and identically distributed if the forecast densities are correctly calibrated. Hence, calibration evaluation requires the application of tests for goodness-of-fit and independence.

The goodness-of-fit tests employed include the Likelihood Ratio (LR) test proposed by Berkowitz (2001), the Anderson-Darling test, and the Pearson (χ^2) test used by Wallis

[6] We experimented with aggregate AR models of order one through four but found little variation in performance.

[7] In a sequel paper, Ravazzolo and Vahey (2012) we consider additional forecasting methodologies including a density forecasting extension of the Hendry and Hubrich (2011) approach, and a dynamic factor model. Neither approach outperformed the IMA benchmark in terms of density forecasting.

(2003). Our Berkowitz test is a three degrees of freedom variant, with a test for independence, where under the alternative z_τ follows an AR(1) process. The Anderson-Darling (AD) test for uniformity, a modification of the Kolmogorov-Smirnov test, gives more weight to the tails of the forecast density. The Pearson (χ^2) tests divides the range of the z_τ into eight equiprobable classes and tests for uniformity in the histogram. We also test directly for independence of the *PITS* using a Ljung-Box (LB) test, based on autocorrelation coefficients up to four. A well-calibrated ensemble should give high probability values for all four of these tests—implying the null hypothesis of no calibration failure cannot be rejected.

Turning to our analysis of relative predictive accuracy, we consider a Kullback-Leibler Information Criterion (KLIC) based test, utilising the expected difference in the Logarithmic Scores of the candidate forecast densities; see, for example, Bao, Lee and Saltoglu (2007), Mitchell and Hall (2005) and Amisano and Giacomini (2007). Suppose there are two density forecasts, $g(y_\tau \mid I_{1,\tau})$ and $g'(y_\tau \mid I_{2,\tau})$, and consider the loss differential $d_\tau = \ln g(Y_\tau \mid I_{1,\tau}) - \ln g'(Y_\tau \mid I_{2,\tau})$. The null hypothesis of equal accuracy is $\mathcal{H}_0 : E(d_\tau) = 0$. The sample mean, $\overline{d}_\tau$, has under appropriate assumptions the limiting distribution: $\sqrt{T}(\overline{d}_\tau - d_\tau) \rightarrow N(0, \Omega)$. The Logarithmic Score of the i^{th} density forecast, $\ln g(Y_\tau \mid I_{i,\tau})$, is the log of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn Y_τ. In our LS test of relative forecast performance, we abstract from the estimation procedure used to generate the forecast densities. Mitchell and Wallis (2011) discuss the value of information-based methods for evaluating forecast densities that are well-calibrated on the basis of *PITS* tests.

3.4 Results

Before considering the density evaluations for our disaggregate ensemble, we summarize the point forecast performance. The disaggregate ensemble (DE) outperform both aggregate specifications, the AR(2), and the IMA in terms of Root Mean Squared Forecast Error (RMSFE). For the AR(2) benchmark, the raw RMSFE is 0.314. The IMA gives

Table 1: Forecast density performance, 1990q1 - 2009q4

	LR3	AD	χ^2	LB	LS	LS_test
DE	**0.537**	**0.095**	**0.517**	**0.527**	**0.107**	**0.000**
			Individual models			
AR(2)2	0.000	0.000	0.000	**0.633**	-0.466	
IMA	0.000	0.000	0.000	**0.885**	-0.361	**0.000**

Note: The column LR is the Likelihood Ratio p-value of the test of zero mean, unit variance and independence of the inverse normal cumulative distribution function transformed *PITS*, with a maintained assumption of normality for transformed *PITS*. AD is the p-value for the Anderson-Darling test for uniformity of the pits. The small-sample (simulated) p-values are computed assuming independence of the *PITS* for the Anderson-Darling test. χ^2 is the p-value for the Pearson chi-squared test of uniformity of the *PITS* histogram in eight equiprobable classes. LB is the p-value from a Ljung-Box test for independence of the *PITS*. A bold number indicates that the null hypothesis of a correctly specified model cannot be rejected at 5% significance level for LR, AD, χ^2 and LB. LS is the average Logarithmic Score over the evaluation period. A bold number in the final column indicates that the null of the LS test of equal density predictive accuracy relative to the AR(2) benchmark is rejected at the 5% significance level.

a four percent improvement over the AR(2), and the disaggregate ensemble gives an 8 percent improvement. Stock and Watson (2007) discuss the difficulty of outperforming simple benchmarks in terms of RMSFE with data that include the Great Moderation data; see also Groen, Paap and Ravazzolo (2009) for similar results.

The evaluation of the forecast densities are presented in table 1. The three rows refer to the disaggregate ensemble, DE, the aggregate autoregressive benchmark, AR(2), and the IMA, respectively. The six columns of table 1 report the p-values for the four PITS tests (Berkowitz LR test, the Anderson-Darling AD test, the χ^2, the LB test), together with the Logarithmic Scores (averaged over the evaluation period), and the Logarithmic Score test for density forecasting performance, relative to the AR(2) benchmark.

Looking at the DE results shown in the top row, we see that the null hypothesis of no calibration failure cannot be rejected at the 5 percent significance level for all of the four individual diagnostic tests, marked in bold. We note that each of these diagnostic tests for calibration is conducted on an individual basis. A 5 percent significance level on each individual test would imply a Bonferroni-corrected p-value of 5/4=1.25 percent (reported as 0.0125 in the table).

The aggregate specifications, shown in the remaining two rows of table 1, display a number of instances of calibration failure. The AR(2) benchmark, first row, fails three diagnostic tests, all with p-values below 1 percent. The more flexible aggregate specification, IMA, also fails three tests at the 1 percent level.

Figure 2 plots the *PITS* histograms for the three candidates, the DE, the AR(2) and the IMA. The histograms for the AR(2) and the IMA display severe departures from uniformity. The DE histogram is more evenly spread across the decile counts, although visual inspection suggests calibration could be improved.

Turning to the Logarithmic Scores of the forecast densities, shown in the sixth column of table 1, we see that the disaggregate ensemble DE records the best relative performance, followed by the IMA. The LS test p-value (marked in bold) indicates that the null hypothesis of equal forecast performance can be rejected at the 1 percent significance level for the DE relative to the (rolling window) AR(2) aggregate benchmark. The IMA aggregate specification also improves on the AR(2) benchmark at the one percent level. An LS test of the DE relative to the (rolling window) IMA confirms the superiority of the DE at the 10 percent significance level.

To shed further light on the contribution of disaggregate information, figure 3 plots the weights in the disaggregate ensemble DE. As we might expect, given the univariate nature of the components, there is uncertainty about the relative importance of disaggregate components through the evaluation. The weights lie in the (approximate) interval [0.01, 0.11] throughout the evaluation, with frequent changes in the identity of the most important disaggregate.

In the three panels of figure 4, we plot the one step ahead density forecasts from the DE, the AR(2) and the IMA through our out of sample evaluation, together with the median forecast in each case, and the outturn. Considering first the DE approach, the central mass of the predictive density declines steadily from around 1.0 percent in 1990 to around 0.5 percent in 2000. Thereafter, the median progressively increases (with some reversals) to peak (locally) in 2008. The recent slump sees the median forecast drop to the levels seen in 1990 again, but still misses the 2008q4 observation. The DE performs better for several spikes in aggregate inflation which occurred between 2000q1 and the recent economic crisis. We emphasise that the relatively strong performance of the DE owes much to the inclusion of the disaggregate forecasts for food and energy. Forcing the weights to zero on these disaggregates causes forecast performance to drop substantially.

The difference between the DE percentiles shown varies a little through the evaluation, with greater dispersion in the pre-2000 forecast densities. Furthermore, the probability that (quarterly) inflation is less than zero is rarely more than 5 percent, with 2008q4 an obvious exception.

Turning to the IMA and AR(2) specifications, it is immediately apparent that the forecast densities are much more diffuse than for the DE in general. Regardless of the percentile considered, the distance from the median is much greater than in the DE case. The dispersion of the percentiles shows greater variation for the AR(2) than the IMA suggesting that the volatility in the IMA is not very responsive to the recent data. The probability that inflation lies below zero for either model typically exceeds 5 percent, even though there are very few outturns of negative inflation. This confirms the poor density forecasting performance of the aggregate models suggested by evaluation based on the *PITS*. It is also worth noting the disappointing median forecast performance of both models. For example, neither model responds substantially to the slump of late 2008. The poor responsiveness reflects the absence of disaggregate information in the aggregate forecasting specifications.

We summarise the results from our forecast density evaluations as follows. First, the disaggregate ensemble performs well in both tests of absolute and relative density forecasting performance. Second, as Jore, Mitchell and Vahey (2010) and Clark (2011) emphasise, although simple autoregressive models of aggregate inflation produce reasonable point forecasts, the benchmark can be bettered considerably in terms of forecast densities. Third, the disaggregate ensemble approach outperforms an IMA specification in density forecasting performance.

4 Conclusions

In this paper, we have proposed a methodology for constructing forecast densities for economic aggregates based on disaggregate evidence using an ensemble predictive system. In our application, we have shown that the disaggregate ensemble approach delivers well-calibrated forecast densities for US PCE aggregate inflation from 1990q1 to 2009q4. Alternative forecasting specifications using only aggregate information failed to match the density forecasting performance of our disaggregate ensemble.

Our applied work indicates that including disaggregate information via an ensemble system improves probabilistic forecasts for US aggregate inflation. Our results also confirm formally the view endorsed by many economic policymakers that disaggregate information can be helpful for forecasting. However, our methodology differs markedly from the standard ‘bottom-up’ approach in providing probabilistic information to policymakers, rather than point forecasts. Future work should investigate the robustness of this performance advantage using data from other countries.

5 References

Altger, F. (2003) “Spatial and Interannual Variability of the Reliability of Ensemble-based Probabilistic Forecasts: Consequences for Calibration”, *Monthly Weather Review*, 131, 1509-1523.

Amisano, G. and R. Giacomini (2007) “Comparing Density Forecasts via Likelihood Ratio Tests”, *Journal of Business and Economic Statistics*, 25, 2, 177-190.

Arora, S.M., M.A. Little and P.E. McSharry (2013) “Nonlinear and Nonparametric Modelling Approaches for Forecasting the US GNP”, *Studies in Nonlinear Dynamics and Control*, forthcoming.

Bache, I.W., J. Mitchell, F. Ravazzolo and S.P. Vahey (2010) “Macro modeling with many models”, in D. Cobham, Ø. Eitrheim, S. Gerlach, and J. Qvigstad (Eds.), *Twenty Years of Inflation Targeting: Lessons Learned and Future Prospects*, Cambridge University Press, 398-418.

Bao, Y., T-H. Lee and B. Saltoglu (2007) “Comparing Density Forecast Models”, *Journal of Forecasting*, 26, 203-225.

Bao, L., T. Gneiting, E.P. Grimit, P. Guttop, and A.E. Raftery (2010) “Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction”, *Monthly Weather Review*, 138, 1811-1821.

Bates, J.M. and C.W.J. Granger (1969) “Combination of Forecasts”, *Operational Research Quarterly*, 20, 451-468.

Berkowitz, J. (2001) “Testing Density Forecasts, with Applications to Risk Management”, *Journal of Business and Economic Statistics*, 19, 4, 465-474.

Clark, T.E. (2006) “Disaggregate Evidence on the Persistence of Consumer Price Inflation, *Journal of Applied Econometrics*, 21, 563-587.

Clark, T.E. (2011) “Real-time Density Forecasts from VARs with Stochastic Volatility”, *Journal of Business and Economic Statistics*, 29, 3, 327-341.

Clark T.E. and M.W. McCracken (2010) “Averaging Forecasts from VARs with Uncertain Instabilities", *Journal of Applied Econometrics*, 25, 5-29.

Croushore, D. (2009) "Revisions to PCE Inflation Measures: Implications for Monetary Policy", FRB Philadelphia Working Paper 08-8, revised July 2009.

Doblas-Reyes, F.J., A. Weisheimer, M. Déqué, N. Keenlyside, M. McVean, J. M. Murphy, P. Rogel, D. Smith, T. N. Palmer (2009) "Addressing Model Uncertainty in Seasonal and Annual Dynamical Ensemble Forecasts", *Quarterly Journal of the Royal Meteorological Society*, 135, 1538-1559.

Diebold, F.X., T.A. Gunther, and A.S. Tay (1998) "Evaluating Density Forecasts; with Applications to Financial Risk Management", *International Economic Review*, 39, 863-83.

Feinstein, M., M.A. King, and J. Yellen (2004) "Innovations and Issues in Monetary Policy: Panel Discussion", *American Economic Review*, Papers and Proceedings, May, 41-48.

Garratt, A., J. Mitchell, S.P. Vahey and Wakerly (2011) "Real-time Inflation Forecast Densities from Ensemble Phillips Curves", *North American Journal of Economics and Finance*, 22, 77-87.

Geweke, J. (2009) *Complete and Incomplete Econometric Models'*, Princeton University Press.

Gneiting, T. "Making and Evaluating Point Forecasts", *Journal of the American Statistical Association*, 106, 746-762.

Gneiting, T. and A.E. Raftery (2007) "Strictly Proper Scoring Rules, Prediction and Estimation", *Journal of the American Statistical Society*, 102, 477, 359-378.

Gneiting, T., and T. Thorarinsdottir (2010) "Predicting Inflation: Professional Experts versus No-change Forecasts", http://arxiv.org/abs/1010.2318.

Granger, C. and M.H. Pesaran (2000) "Economic and Statistical Measures of Forecast Accuracy", *Journal of Forecasting*, 19, 537-560.

Greenspan, A. (2004) "Risk and Uncertainty in Monetary Policy", *American Economic Review*, Papers and Proceedings, May, 33-40.

Groen, J.J.J., R. Paap and F. Ravazzolo (2009) "Real-time Inflation Forecasting in a

Changing World", Federal Reserve Bank of New York Staff Reports, 388.

Hendry, D. F. and K. Hubrich (2011) "Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate", *Journal of Business and Economic Statistics*, 29, 2, 216-227.

Hersbach, H. (2000) "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems", *Weather and Forecasting*, 15, 559-570.

Huurman C., F. Ravazzolo and C. Zhou (2012) "The power of weather", *Computational Statistics and Data Analysis*, forthcoming.

Jore, A.S., J. Mitchell and S.P. Vahey (2010) "Combining Forecast Densities from VARs with Uncertain Instabilities", *Journal of Applied Econometrics*, 25, 621-634.

Kascha, C. and F. Ravazzolo (2010) "Combining Inflation Density Forecasts", *Journal of Forecasting*, 29, 231-250.

Koop, G. (2003) *Bayesian Econometrics*, Wiley.

Lütkepohl, H. (2009) "Forecasting Aggregated Time Series Variables: a Survey", Economics Working Papers ECO2009/17, European University Institute.

Lütkepohl, H. (2010) "Forecasting Nonlinear Aggregates and Aggregates with Time-varying Weights", Economics Working Papers ECO2010/11, European University Institute.

Marcellino, M., J. Stock and M. Watson (2003) "Macroeconomic Forecasting in the Euro area: Country Specific versus Euro Wide Information", *European Economic Review*, 47, 1-18.

Mitchell, J. and S.G. Hall (2005) "Evaluating, Comparing and Combining Density Forecasts using the KLIC with an Application to the Bank of England and NIESR Fan Charts of Inflation", *Oxford Bulletin of Economics and Statistics*, 67, 995-1033.

Mitchell, J. and K.F. Wallis (2011) "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness", *Journal of Applied Econometrics*, 26, 1023-1040.

Panagiotelis, A. and M. Smith (2008) "Bayesian Density Forecasting Intraday Electricity Prices using Multivariate Skew t Distribution", *International Journal of Forecasting*, 24, 710-727.

Raftery, A.E., T. Gneiting, F. Balabdaoui and M. Polakowski, (2005) "Using Bayesian Model Averaging to Calibrate Forecast Ensembles", *Monthly Weather Review*, 133, 1155-1174.

Ravazzolo, F. and S.P. Vahey (2012) "Combining Disaggregate Forecasts to Predict Deflation", mimeo.

Rosenblatt, M. (1952) "Remarks on a Multivariate Transformation", *The Annals of Mathematical Statistics*, 23, 470-472.

Stensrud, D.J. and N. Yussouf (2007) "Bias-corrected Short-range Ensemble Forecasts of Near Surface Variables", *Meteorological Applications*, 12, 217-230.

Stock, J.H. and M.W. Watson (2007) "Why has US Inflation Become Harder to Forecast?", *Journal of Money, Credit and Banking*, 39, 3-34.

Timmermann, A. (2006) "Forecast Combination", G. Elliot, C. Granger, C. and A. Timmermann (eds.) *Handbook of Economic Forecasting,* North-Holland, 197-284.

van Garderen, K.J, K. Lee and M.H. Pesaran (2000) "Cross-sectional Aggregation of Non-linear Models", *Journal of Econometrics*, 95, 285-331.

Wallis, K.F. (2003) "Chi-squared Tests of Interval and Density Forecasts, and the Bank of England's Fan Charts", *International Journal of Forecasting*, 19, 165-175.

Wallis, K.F. (2005) "Combining Density and Interval Forecasts: a Modest Proposal", *Oxford Bulletin of Economics and Statistics*, 67, 983-994.

Appendix

To illustrate how the ensemble system reacts to time variation in the weights, $\omega_{i,\tau}$, and the parameters of the disaggregate forecasting equations, equation (3), we describe eight simulation exercises.[8]

We begin by describing the basic case, exercise 1. We simulate two disaggregate variables, each of which follows a first order autoregressive model, AR(1) with Gaussian error,

[8] In these simulations, we work with a small number of disaggregates for illustrative purposes.

given by equation (3). The aggregate index, y_t, satisfies equation (1) for two disaggregates ($i = 1, 2$), with index weights $\omega_i = 0.5$. Each simulation has 1000 replications. Using a total sample of 120 observations (indexed by $t = 1, \ldots, 120$) in each simulation, we construct out of sample disaggregate forecasts for $t = 41, \ldots, 120$. We estimate the disaggregate models using a Bayesian AR(1) with non-informative priors, and an expanding window of observations for in-sample estimation. (The predictive densities follow the t-distribution, with mean and variance equal to OLS estimates; see, for example, Koop (2003, chapter 3) for details.) Out of sample forecast densities for $t = 41, \ldots, 120$ are passed through the LOP, using a 20-period training window to initialise the ensemble weights. A moving window of 20 observations is used to both bias-correct the disaggregate densities and to construct the ensemble weights for LOP. Hence, the out of sample evaluation for the ensemble starts in $t = \underline{\tau} = 61$ and ends in $\overline{\tau} = 120$. We forecast the aggregate using an aggregate AR(1) specification as a benchmark forecasting model.

In the seven subsequent simulation exercises, we explore the implications of introducing specification errors to the forecasting system. These include evolving index weights, and various forms of structural breaks in the disaggregate forecasting specifications. In each simulation, the disaggregate ensemble and the benchmark aggregate AR(1) model ignores the time variation in the 'true' specification so that we can study the impacts of unknown specification errors.

2. The index weight ω_1 follows an autoregressive process, such that the weight is bounded between [0.25,0.75], and the weights sum to one.

3. As exercise 2 except that each disaggregate has a single break in the mean at observation $t = 20$.

4. As exercise 3 except that each disaggregate has two breaks in the mean, the first at observation $t = 20$, the second at $t = 60$.

5. As exercise 2 except that each disaggregate has a single break in the error variance at observation $t = 20$.

6. As exercise 5 except that each disaggregate has two breaks in the error variance, the first at observation $t = 20$, the second at $t = 60$.

7. As exercise 2 except that each disaggregate has a single break both the mean and the error variance at observation $t = 20$.

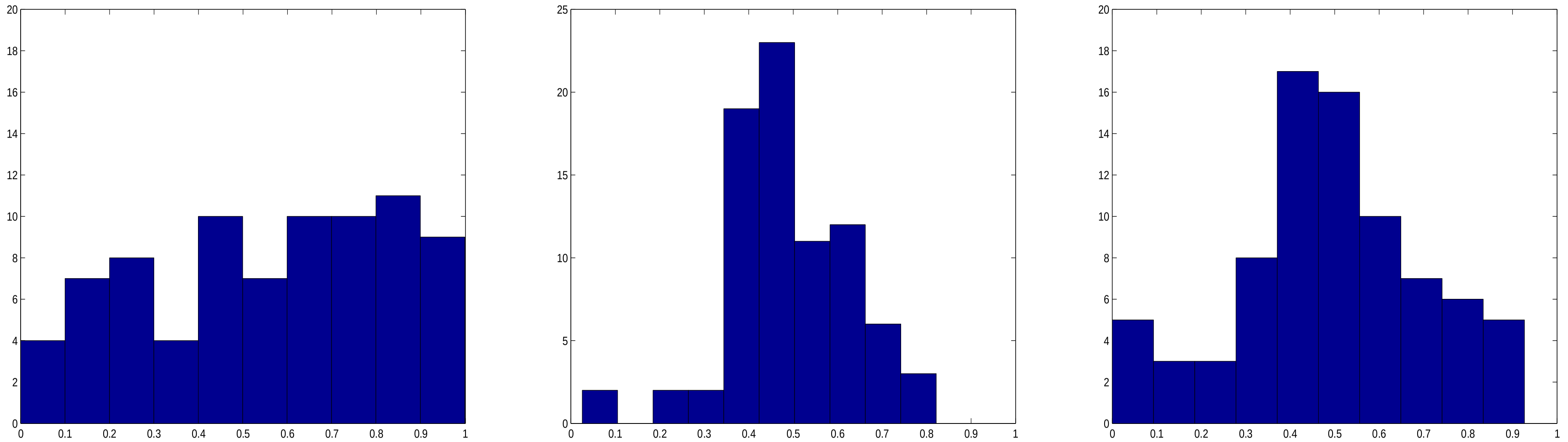
8. As exercise 7 except that each disaggregate has two breaks in both the mean and the error variance, the first at observation $t = 20$, the second at $t = 60$.

To check that our results in exercises 2 through 8 are not sensitive to the assumption that the index weights are time-varying, we repeated exercises 2-8 with constant weights. The results of these simulations are quantitatively similar to exercises 2-8 and so are not reported. That is, the time variation in the index weights has negligible impacts on the performance of the disaggregate ensemble relative to the aggregate benchmark.

To judge forecasting performance, we use the average Logarithmic Score over the evaluation period, $\underline{\tau} = 61$ to $\overline{\tau} = 120$. The Logarithmic Score of the i^{th} density forecast, $\ln g(Y_\tau \mid I_{i,\tau})$, is the log of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn Y_τ. Mitchell and Wallis (2011) provide a recent discussion of scoring rules and the justification for testing relative density forecasting performance from the perspective of the Kullback-Leibler Information Criterion (KLIC). Gneiting and Raftery (2007) analyse the relationships between scoring rules and Bayes factors. A higher average Logarithmic Score denotes better density forecasting performance. We provide histograms based on the 1000 repetitions for each simulation exercise.

There are two striking features from our simulations. First, regardless of which case we consider, the disaggregate ensemble (DE) is never inferior to the aggregate benchmark forecasting model (AR) in terms of the average Logarithmic Score across the 1000 replications. Second, the biggest differences in density forecasting performance arise in cases where the disaggregate forecasting specifications exhibit multiple structural breaks (especially in the means). In particular, in exercises 4 and 8.

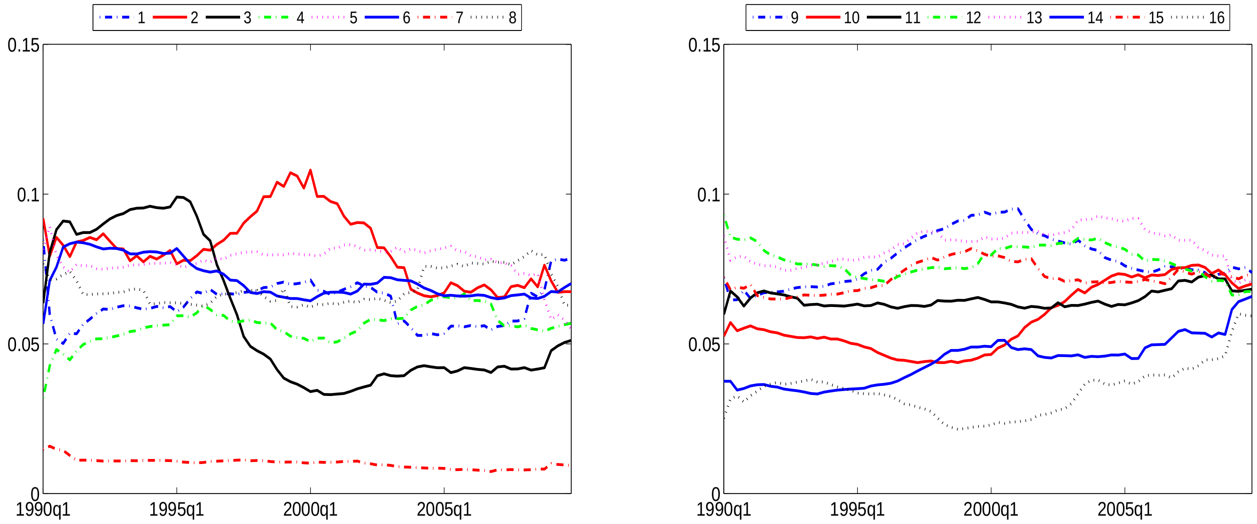
Figure 2: *PITS* histograms



(a) DE - AR(2) - IMA

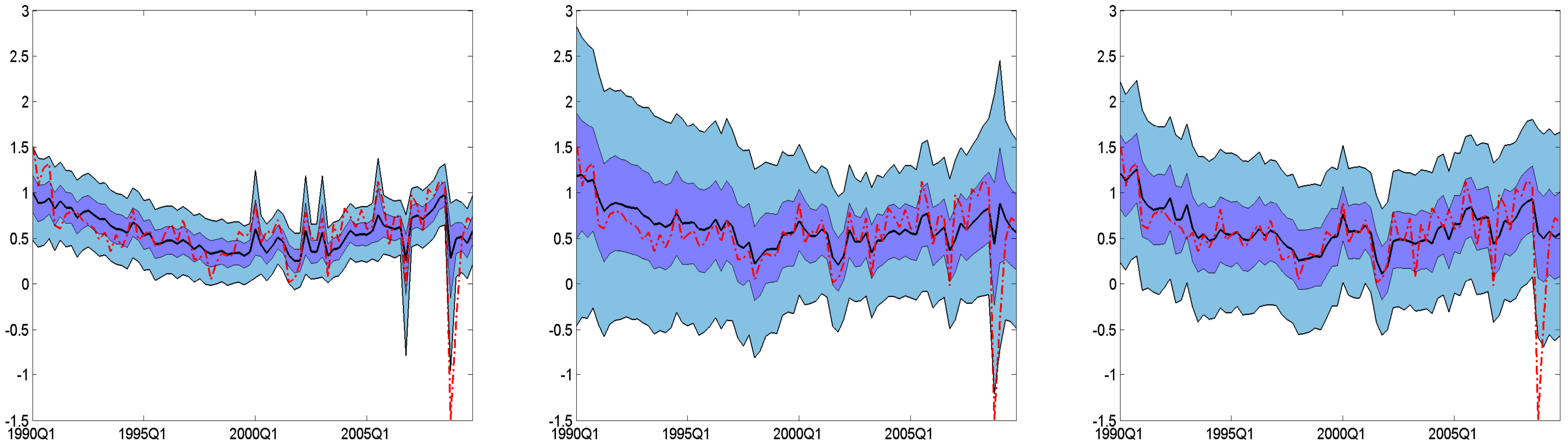
Note: The histogram shown are the decile counts of the *PITS* transforms.

Figure 3: DE weights



Note: The figures plot the weights given by disaggregate ensemble DE. The disaggregate order 1-16 for DE corresponds to the text in Section 3.1.

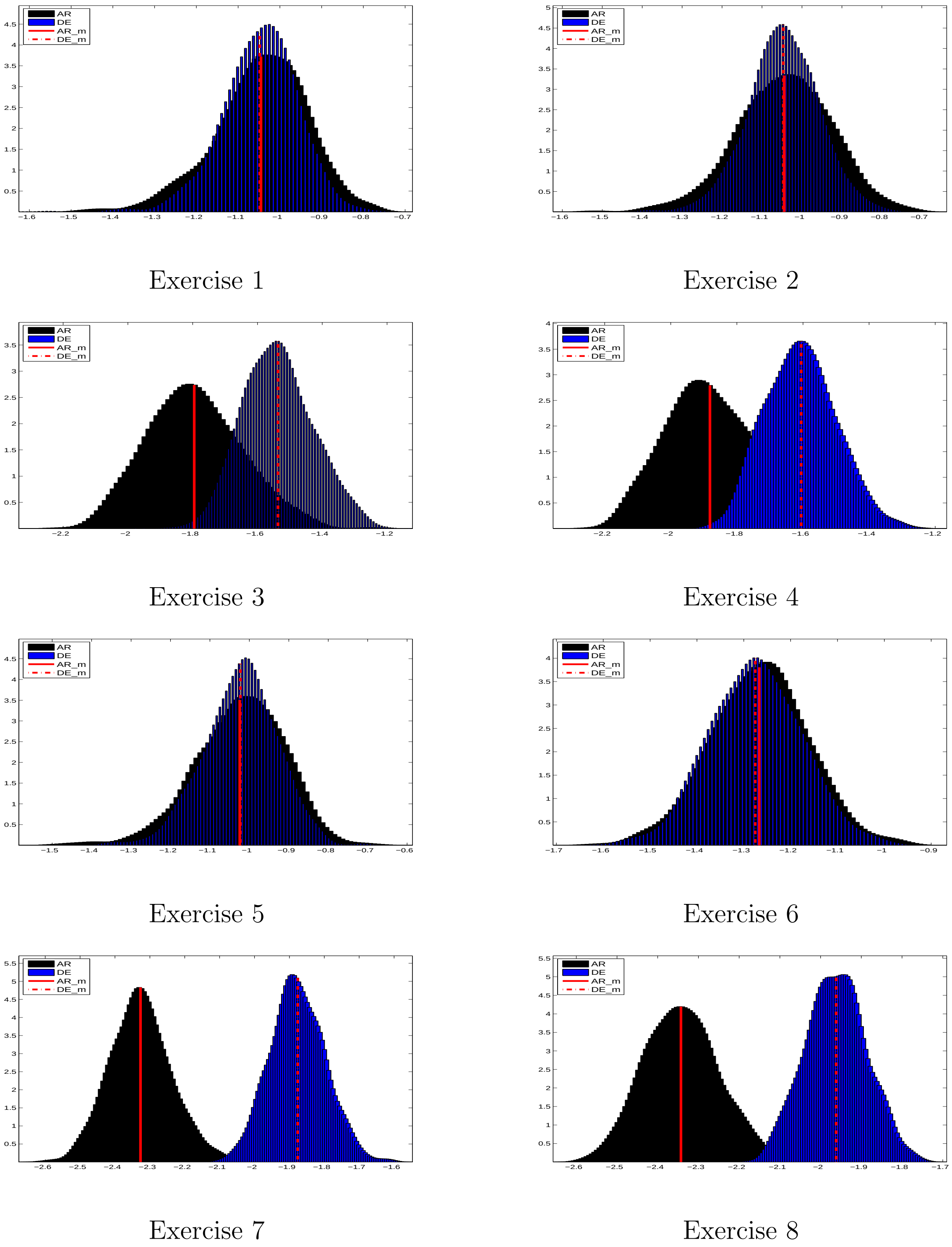
Figure 4: Interval forecasts



(a) DE- AR(2) - IMA

Note: The black solid lines represent the 5%, 25%, 50%, 75%, and 95% percentiles of the of the predictive densities given by three forecasting methods and the red dashed line shows the actual inflation.

Figure 5: Simulation results



Exercise 1

Exercise 2

Exercise 3

Exercise 4

Exercise 5

Exercise 6

Exercise 7

Exercise 8

Note: The figures show histograms of the LS for the AR model and the DE, the mean of the LS for the AR model (red lines) and the LS for the DE (red dashed lines) for our simulation exercises.