Ditch The Ex! Measure Core Inflation with a Disaggregate Ensemble*

Francesco Ravazzolo† (Norges Bank)

Shaun P. Vahey‡ (Melbourne Business School)

May 21, 2009

Abstract

We propose a methodology for producing core inflation predictive densities based on disaggregate inflation evidence. The densities reflect the ability of the disaggregates to predict measured inflation. We construct ensemble predictive densities based on a linear mixture of experts. In our US application, we show that the well-known 'Ex' measure of core inflation produces inaccurate predictive densities for measured US inflation from 1997Q2 to 2008Q1. Our preferred disaggregate ensemble core measure delivers significantly better density forecasting performance. We conclude that the disaggregate information excluded in the 'Ex' measure–food and energy–has substantive impacts on the probability of inflation events of interest to the Federal Reserve.

Keywords: core inflation; underlying inflation; forecast density combination; ensemble forecasting

JEL codes: C11; C32; C53; E37; E52

---

*We benefited greatly from discussions with Todd Clark, Christian Kascha, and Michael Smith. The views expressed in this paper are our own and do not necessarily reflect the views of Norges Bank.

†Norges Bank, Research Department. francesco.ravazzolo@norges-bank.no

‡*Corresponding author*: Shaun Vahey, Melbourne Business School, 200 Leicester St, Carlton, Victoria, Australia, 3053. Telephone: +61 3 9349 8174. Fax: +61 3 9349 8271. s.vahey@mbs.edu.

1 Introduction

The commonly-offered justification for stripping out particular disaggregates to provide a measure of core inflation labels the discarded information as idiosyncratic. The movements of the excluded disaggregates—food and energy prices in the case of the 'Ex' core—are presumed to reflect relative prices, rather than the general price level.

Monetary policymakers assess routinely and repeatedly the probability of inflation outcomes; see, for example, Greenspan (2004) and the discussion by Feinstein, King, and Yellen (2004). It is surprising, therefore, that earlier studies have not assessed the probabilistic implications of core inflation measures. The papers by, for example, Smith (2004) and Kiley (2008) restrict attention to the evaluation of point forecasts, and provide no guidance on the usefulness of the Ex core for general (but unknown) loss functions with non-Gaussian inflation predictives.

In this paper, we propose a methodology for policymakers concerned with the complete forecast density for measured inflation. We formulate the measuring core inflation problem as one in which a policymaker (recursively) selects a linear combination of disaggregate predictive densities to produce an ensemble density for measured inflation. Each component of the ensemble is produced from a time series model using a single disaggregate series. The construction of the ensemble forecast uses out-of-sample density combination methods; see, for example, Jore, Mitchell and Vahey (2009). We emphasize that our measure of underlying inflation, which we refer to as the 'Disaggregate Ensemble' (DE) core, comprises a forecast density for measured inflation. Our technique provides an estimate of the entire probability distribution of the possible future values of measured inflation, allowing for uncertainty about both the level of disaggregation, and the importance of particular disaggregates.

In an application based on US Personal Consumption Expenditure deflator data, we assess the forecast accuracy of the DE and Ex core measures over the out of sample period 1997Q2 to 2008Q1. We find the DE predictives to be better calibrated to those from a simple autoregressive model for measured inflation. In contrast, predictives from the traditional Ex core fail to outperform the same benchmark. In contrast to the point forecast case discussed by Stock and Watson (2007), simple autoregressive models of inflation can be beaten in terms of forecast density performance using US Great Moderation data.

We draw three conclusions from our US application. First, the practice of zero-weighting particular disaggregates in the construction of core inflation discards important information about the inflationary process; see, for example, Quah and Vahey (1995) and

Cecchetti and Wynne (2003). Despite the advocacy of (among others) Mishkin (2007, Speech, HEC Montreal, October 20), the Ex core approach is deficient as a basis for assessing the probability of inflation events of interest to policymakers. Second, our DE core produces well-calibrated forecast densities for measured inflation by using efficiently the disaggregate information. Policymakers that ignore the possibility that food and energy shocks feed through understate the risk of inflationary and deflationary events. Third, since the Federal Reserve has recently moved beyond point forecasts to publishing forecast intervals for inflation, it should adopt a core inflation measure that matters for entire probability distribution of measured inflation. The Federal Reserve should drop the Ex core measure.

The remainder of this paper is structured as follows. In Section 2 we describe our methods for ensemble modeling of the relationship between measured inflation and the disaggregates. In Section 3, we apply our methodology to US data to produce core inflation predictive densities from an ensemble system. We compare and contrast the ensemble predictives for measured inflation with those resulting from our autoregressive benchmark. In the final section, we conclude.

2 Ensemble Forecast Methodology

In this section, we describe our methodology for measuring core inflation utilizing an ensemble prediction system. By building ensemble systems, researchers aim to combine out of sample forecasts from component models. This provides a pragmatic framework to produce macroeconometric forecasts robust to instabilities of uncertain timing and nature—unknown structural changes. Recent studies include (among others) Stock and Watson (2003), Clark and McCracken (2009), Jore, Mitchell and Vahey (2009) and Garratt, Mitchell and Vahey (2009). The first two studies find that simple combinations of component models produce accurate point forecasts. Jore, Mitchell and Vahey (2009) and Garratt, Mitchell and Vahey (2009) use recursively updated weights to produce well-calibrated predictive densities.

We begin by describing how the ensemble predictive densities are constructed (from times series component models). Then we discuss our methodology for measuring DE core inflation.

2.1 Ensemble Construction

Monetary policymakers examine regularly disaggregate inflation series for leading evidence of the measured inflationary process. A number of central banks (including the Federal Reserve, the Bank of England, Norges Bank and Sveriges Riksbank) now publish predictive densities or intervals for measured inflation. With these innovative tools aimed at communicating probabilistic information in mind, we construct ensemble predictives based on the (out of sample) forecast performance of (potentially many) component models. As we shall discuss subsequently, each component model will use a particular disaggregate series.

To summarize the density combination approach, for each observation in the policymaker's evaluation period, we use forecast density combination to compute the weight on each disaggregate component model, based on the 'fit' of the component predictive densities for measured inflation. Given these weights, we will construct an ensemble forecast density for measured inflation.

More formally, consider a monetary policymaker aggregating forecasts supplied by 'experts'. Each expert uses a model based around the relationship between measured inflation and a preferred disaggregate. The experts are differentiated by the disaggregate of interest, where each expert examines a unique disaggregate to produce a forecast density for measured inflation. (Among others) Timmermann (2006 p177) discusses expert combination. Recent applications include Wallis (2004), aggregating survey information, and Mitchell and Hall (2005), combining forecasts from two institutions.

Given $i = 1, \ldots, N$ disaggregates (where N could be a large number), we define disaggregate ensemble (DE) core inflation by the convex combination (sometimes referred to as a linear opinion pool):

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

where $g(\pi_{\tau,h} \mid I_{i,\tau})$ are the h-step ahead forecast densities from component model i, $i = 1, \ldots, N$ of measured inflation π_τ, conditional on the information set I_τ.

Each component model produces h-step ahead forecasts for measured inflation. Hence, the disaggregate variable used to produce an h-step ahead forecast density for τ are dated $\tau - h$ or earlier. The non-negative weights, $w_{i,\tau,h}$, in this finite mixture sum to unity, are positive, and vary by recursion in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$.

We emphasize that the ensemble forecast density could be non-Gaussian even if the component models produce Gaussian predictives.

From a Bayesian perspective, ensemble modeling approach has many similarities with a predictive likelihood approach. Given a definition of density fit, component density combination respects Bayes' rule, with equal (prior) weight attached to each component. In Bayesian language, a non-informative prior. Koop (2003, chapter 11) provides a recent general discussion of Bayesian model averaging methods.

To implement our ensemble construction using equation (1) requires a measure of component density fit to provide the weights. A number of recent applications in the economics literature have used density scoring rules. In the application that follows, we shall utilize the Continuous Ranked Probability Score (CRPS), which as (among others) Hersbach (2000), Gneiting and Raftery (2007) and Panagiotelis and Smith (2008) discuss, rewards predictive densities from components with high probabilities near (and at) the outturn.

2.2 Interpretation

Since we define the DE core as the ensemble predictive density for measured inflation, by construction, this measure combines the information embodied in the inflation disaggregates into an (h-step ahead) predictive density for measured inflation. The (recursive) weights on the disaggregate component models are determined by their relative forecasting performance for measured inflation. Bounded on the (0,1) interval, the weights provide a direct indication of the importance of each disaggregate in forecasting measured inflation. The higher the weight, the more important the disaggregate. Given that the Ex core approach to measuring core inflation labels food and energy shocks as idiosyncratic—not part of the core inflation process—the weight on this component should be zero. The weights on the ensemble provide an indication of whether the data support the Ex core.

We emphasize that the flexible shape of the ensemble predictives represents an important feature of our methodology—we allow for non-Gaussian behaviour by construction.[1] Hence, we can examine whether excluding the food and energy disaggregate affects the assessment of the probability of upper tails events. From a theoretical perspective, a food and energy shock has the potential to affect both the central mass of the probability distribution for measured inflation, and the higher moments. Our flexible methodology for measuring core inflation allows for complex and time varying responses by the central

[1]Kascha and Ravazzolo (2009) compare and contrast log and linear pooling. Log opinion pools force the ensemble predictives to be unimodal and symmetric.

bank to the disaggregate shock.

3 Application: Core inflation for the US

This application of our methodology uses US Personal Consumption Expenditure deflator (PCE) data. We assess the accuracy of the DE and other core measures, including the well-known Ex core, using an evaluation period from 1997Q2 to 2008Q1. In addition to examining the weights on the disaggregate component models, we test the calibration of the measured inflation predictives using the probability integral transforms, *PITS*, and test forecast performance of the candidates relative to an autoregressive benchmark.

We begin our analysis by describing the US data. Then we describe our component model space, the ensembles and our results.

3.1 Data

The data set contains time series for the disaggregate components of the PCE.[2] These data permit breakdowns at various levels of disaggregation, ranging from a simple decomposition of the PCE into two series, core PCE, and food and energy, to a breakdown of the PCE into over 100 disaggregate series.

In our application, we disaggregate the PCE in two ways. First, we consider the core PCE excluding food and energy prices (Ex core), and the residual PCE component. The second form of disaggregation breaks the PCE into three disaggregates: durables, nondurables and services. There is uncertainty about the appropriate form of the disaggregation.

$$\text{PCE} \begin{array}{l} \nearrow \ \text{Ex core PCE, and Food and energy} \\ \searrow \ \text{Durables, Non-durables, and Services} \end{array}$$

We emphasize that, in principle, our methodology could be applied to any level of disaggregation. Tables AI-AIII in Clark (2006) provide further details on levels of disaggregation in the US PCE data.

[2]Data are available on the Bureau of Economic analysis http://www.bea.gov/national/nipaweb.

For all inflation series, PCE and its disaggregates, we work with quarterly growth rates; see figure 1. Restricting our attention to Great Moderation data, we start our sample with 1984Q1 and end with 2008Q1. Hence, we restrict our analysis to inflation during the US Great Moderation—when conventional wisdom has it that inflation is difficult to forecast well. The evaluation period for the predictives is 1997Q2 to 2008Q1; the period 1993Q2 to 1997Q1 we use as a 'training period' to initialize the ensemble weights. We note, however, that the volatility and the mean of PCE measured inflation do vary through the sample as figure 1 makes clear. Of the disaggregate series, the Food and Energy sectors sees a marked increase in volatility towards the end of the period, as does Non-durables. In contrast, the Ex core, Durables and Services exhibit declining levels from the mid-1990s.

3.2 Component models

A number of studies have noted the scope for parameter change to improve forecasting performance. Therefore, we consider component models that allow for structural breaks of unknown number and timing. More precisely, we specify the following mixture innovation model for each disaggregate inflation series:

$$\pi_\tau = \beta_{0\tau} + \textstyle\sum_{p=1}^{k} \beta_{p\tau}\pi_{\tau-p} + \sigma_\tau \varepsilon_\tau$$

$$\beta_{j\tau} = \beta_{j,\tau-1} + \kappa_{j\tau}\eta_{j\tau}, \quad j = 0, \ldots, k \tag{2}$$

$$ln\sigma_\tau^2 = ln\sigma_{\tau-1}^2 + \kappa_{k+1,\tau}\eta_{k+1,\tau}$$

where $\varepsilon_\tau \sim N(0,1)$, $\eta_\tau = (\eta_{0\tau}, ..., \eta_{k+1,\tau})' \sim N(0,Q)$ with Q a diagonal matrix and elements $q_0^2, \ldots, q_{k+1}^2$, and $\kappa_\tau = (\kappa_{0\tau}, \ldots, \kappa_{k+1,\tau})'$ is a $((k+2) \times 1)$ vector of unobserved uncorrelated 0/1 processes with $\Pr[\kappa_{j\tau} = 1] = p_j$ for $j = 0, \ldots, k+1$.

Hence, each of the regression parameters $\beta_{j\tau}$ and the residual variance σ_τ^2 remain the same as their previous values $\beta_{j,\tau-1}$ and $\sigma_{\tau-1}^2$ unless $\kappa_{j\tau} = 1$ and $\kappa_{k+1,\tau} = 1$ in which case $\beta_{j\tau}$ changes with $\eta_{j\tau}$ and $ln(\sigma_\tau)^2$ changes with $\eta_{k+1,\tau}$ respectively. See, for example, Koop and Potter (2007) and Giordani, Kohn, van Dijk (2007) for similar approaches. As the changes in the variance parameters $\ln\sigma_\tau^2$ are stochastic we allow for a form of stochastic volatility; see Giordani and Kohn (2008). The flexibility of the specification in (2) stems from the fact that the parameters $\beta_\tau = (\beta_{0\tau}, \ldots, \beta_{k\tau})'$ and σ_τ^2 are allowed to change every time period, but they need not change. The occurrence of a change is described by the latent binary random variable $\kappa_{i\tau}$, while the magnitude of the change is determined by $\eta_{i\tau}$, which is assumed to be normally distributed with mean zero. Another attractive property of (2) is that the changes in the individual regression parameters are not restricted to coincide but rather are allowed to occur at different points in time.

Recall that we construct the DE core by combining the predictive densities from all of the disaggregate component models. We describe these disaggregate predictives in an Appendix. Although the component models described in this section forecast disaggregate inflation, we construct the ensemble predictives for measured inflation by evaluating the disaggregate forecasts for measured inflation. In each recursion, we center the component forecasts on measured inflation as described in the Appendix. In effect, this step restricts the ensemble to be uni-modal.[3]

In this application, we restrict attention to 1-step ahead forecasts for measured inflation; applications for longer horizon counterparts pose no conceptual difficulties.[4]

3.3 Ensembles

As noted above, the ensemble forecast densities for measured inflation use equation (1). To gauge density fit, and to derive the component weights for the ensemble, we use a scoring rule.[5] In our application, we utilize the Continuous Ranked Probability Score (CRPS) which rewards component models with high probabilities near (and at) the outturn.

The weights for the h-step ahead DE core densities are:

$$w_{i,\tau,h} = \frac{\left[\sum_{\underline{\tau}}^{\tau-1-h} X(g(y_{\tau,h} \mid I_{i,\tau}))\right]}{\sum_{i=1}^{N}\left[\sum_{\underline{\tau}}^{\tau-1-h} X(g(y_{\tau,h} \mid I_{i,\tau}))\right]}, \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}. \tag{3}$$

The computation of X, the CRPS-based measure of density performance is described in the Appendix, with $0 \leq X \leq \infty$, and higher scores preferred.

To assess relative forecast performance from the core ensembles, we apply the test of equal accuracy given in Mitchell and Hall (2005). Suppose there are two density forecasts, $g(\pi_{\tau,h} \mid I_{1,\tau})$ and $g(\pi_{\tau,h} \mid I_{2,\tau})$, and consider the loss differential $d_{\tau,h} = \ln g(\pi_{\tau,h} \mid I_{1,\tau}) - \ln g(\pi_{\tau,h} \mid I_{2,\tau})$. The null hypothesis of equal accuracy is $\mathcal{H}_0 : E(d_{\tau,h}) = 0$.

[3]Removing this step does not affect the qualitative evidence against the Ex measure. However, the resulting multi-modal predictives have weaker calibration properties.

[4]We note, however, that probability of keeping measured inflation within the target at the horizon of interest to policymakers reflects the credibility of the monetary regime.

[5]Mitchell and Wallis (2009) provide a recent discussion of scoring rules and the KLIC justification for testing relative calibration performance. Gneiting and Raftery (2007) analyze the relationships between scoring rules and Bayes factors.

The sample mean, $\overline{d}_{\tau,h}$, has under appropriate assumptions the limiting distribution: $\sqrt{T}(\overline{d}_{\tau,h} - d_{\tau,h}) \rightarrow N(0, \Omega)$. We note that the logarithmic score of the i-th density forecast, $\ln g(\pi_{\tau,h} \mid I_{i,\tau})$, is the logarithm of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn $\pi_{\tau,h}$. Hence, the log score approach evaluates the predictives at the outturn only.

We also compute *PITS*, probability integral transforms, as in, for example, Little, McSharry and Taylor (2008) and apply the Berkowitz (2001) likelihood ratio test for independence, zero mean and unit variance of the *PITS*. The test statistic is distributed $\chi^2(3)$, chi-square with 3 degrees of freedom, with the null of no calibration failure. Tests based on the *PITS* provide an indication of whether the null of no calibration failure can be rejected. That is, the test is for absolute, rather than relative forecast performance.

3.4 Results

In this section, we report results for the DE core corresponding to both methods of disaggregation. That is, with measured inflation split into two disaggregates (Ex core, and food and energy) and into three disaggregates (durables, non-durables, and services). For each type of disaggregation, we construct a DE core measure, which we refer to as DE2 and DE3 core, respectively. We also consider a variant which combines both types of disaggregation, which we refer to as DE5. A further core measure has weights restricted to be fixed for the four quarters of each calender year, DE5-ANN.[6]

In addition to these DE core measures, we also evaluate the predictives from two additional variants. The first uses a linear model to forecast measured inflation without disaggregate information. That is, using a linear autoregressive model for aggregate measured inflation, with two lags, AR(2).[7] We use this AR model as our benchmark in tests of relative forecast performance.

The second variant uses the mixture innovations model for a single disaggregate series, namely the Ex core PCE (which excludes food and energy), labeled Ex PCE in the tables and figures.

Before turning to the density evaluations for our various ensembles, we summarize the point forecast performance in table 1 for our four DE core measures (DE2, DE3, DE5,

[6]The corresponding benchmarks DE2-ANN and DE3-ANN have slightly weaker forecast performance than DE5-ANN. Results can be obtained on request.

[7]We use uninformative priors for the AR(2) parameters with an expanding window. The predictive densities follow the t-distribution, with mean and variance equal to OLS estimates; see, for example, Koop (2003) for details. We also experimented with an AR(4), but the forecasting performance was weaker than for the AR(2).

DE5-ANN) and the AR(2) and the Ex PCE, shown in the rows. The first column displays the root mean squared prediction error (RMSPE) of each ensemble (benchmark) relative to the AR(2). The RMSPE of the AR(2) benchmark is 0.163. All of the ensembles give a larger RMSPE than the AR(2), with minor differences between them. The second column provides p-values for Harvey, Leybourne and Newbold (1997) test for equal predictive accuracy for each ensemble and the AR(2). The null lies outside the 95 percent confidence interval in all cases. We conclude that an AR(2) model performs better than all of the core measures, including the Ex, from the perspective of point forecast accuracy. Smith (2004) and Kiley (2008) discuss the point forecasting properties of various core inflation measures. Given the unknown loss function of the monetary policymaker, and the likelihood of non-Gaussian behavior, we prefer to focus on the forecast performance for the entire probability distribution of measured inflation.

The evaluation of the forecast densities presented in table 2 contrasts strongly with the point forecast results in table 1. The six rows again refer to the four DE core measures, and the two additional variants. The columns report the p-values for the Mitchell-Hall predictive density accuracy test (based on the log scores), the log scores themselves (averaged over the evaluation period), and the Berkowitz likelihood ratio test (based on the *PITS*).

Looking at the Mitchell-Hall LS-test, two of the core ensembles outperform the AR(2) benchmark based on 95 percent confidence intervals. Namely, DE5 and DE5-ANN. In contrast, the ensembles DE2 (based on the Ex core, and food and energy disaggregation), and DE3 (based on durables, service and nondurables disaggregation) fail to outperform the AR(2) with 80 percent confidence intervals. It appears that it helps to consider both types of disaggregation in the PCE. Turning to the Ex core, Ex PCE, the null of equal forecast performance relative to the AR(2) cannot be rejected for an 85 percent confidence interval. The second column of table 2 reveals that the DE5 and DE5-ANN ensembles give the lowest log scores, with approximately a 40 percent improvement over the AR(2). The Ex core, DE2 and DE3 all beat the log score of the AR benchmark, up to almost 30 percent, although the Mitchell-Hall test implies that the improvements are statistically insignificant.[8]

Finally, turning to the third column three of table 2, we note that the *PITS* tests indicate all the predictives to be well-calibrated at the 85 percent level. The *PITS* of the

[8]The ensembles and Ex core give broadly similar CRPS values to the AR(2) benchmark, averaged over the whole evaluation period. DE5 gives modest improvements, with the DE5-ANN very close to (but slightly worse than) the AR(2).

various forecast densities do not allow us to distinguish between the candidates for our Great Moderation sample.[9]

We draw the following conclusions from the forecast density evaluations in table 2. First, the core measures using all of the disaggregate information, DE5 and DE5-ANN, perform well. Second, the restriction that the weights are fixed intra year has little impact on forecast performance—the forecast performance of DE5-ANN compares well with DE5. Finally, the inferiority of the Ex core approach reflects the inefficiency of zero-weighting some disaggregate information.

To shed further light on the contribution of disaggregate information, figure 2 provides plots of the weights in the DE5 and DE5-ANN core through the evaluation period. Although there is uncertainty about which disaggregate component is the most important, it seems that Ex measure always receives a high weight. However, the weight attached to the Ex never exceeds 40 percent, regardless of whether we allow the weights to vary with each recursion, DE-5 core, or if we restrict the weights to be fixed intra-year, DE5-ANN. Apparently the services component carries considerable information, receiving a weight that varies between 20 and 30 percent, increasing towards the end of the evaluation. In contrast, the food and energy sector receives a small weight in both core measures. Although this weight is less than 10 percent for all observations, it exhibits very little variation through evaluation period. We conclude that the data do not support the exclusion of the food and energy disaggregate from the DE core.

In figure 3, we plot the median from our preferred DE core measure, DE5, and the sample data for measured inflation, together with the 25 and 75 percentiles from the ensemble density. The plot shows that the median of the DE5 core is considerably less volatile than the actual measured inflation series. This measure of the central mass drifts upwards through the evaluation period, from around 0.2 percent in 1997 to roughly 0.6 by 2007. The difference between the two percentiles varies very little through the evaluation, typically remaining close to 0.5 percentage points. Furthermore, the probability that (quarterly) inflation is less than zero is rarely more than 25 percent, even at the start of the period. And considerably less than that towards the end of the evaluation.

To provide further insight into the probability of tail events for inflation, figure 4 provides the ensemble predictive densities at particular observations, namely 1997Q2 and 2008Q1. We see that the DE5 predictives contain more mass in the tails than those

[9] Jore, Mitchell and Vahey (2009) and Garratt, Mitchell and Vahey (2009) distinguish successfully between competing ensemble predictives using tests based on the *PITS* for US data that include the Great Inflation period.

resulting from the Ex core measure. In so far as the predictives exhibit relatively minor departures from symmetry, we conclude that the deficiency of the Ex core measure as a predictor of measured inflation stems largely from the dispersion of the predictives. That is, the Ex core attributes too much mass to the center of the distribution. The disaggregate evidence suggests that there is a risk that food and energy shocks will feed through into the inflationary process.

4 Conclusions

In this paper, we have proposed an approach to measuring core inflation based on utilizing the information in disaggregates in the presence of uncertainty about the form of disaggregation. Our approach constructs an ensemble for measured inflation from disaggregate forecasting components. In our US application, we showed that the DE core produces well-calibrated predictives for measured inflation. The Ex core approach, popular with policymakers, was not supported by the data. In particular, the Ex core strategy delivered too much mass in the center of the predictive densities. There is a risk that food and energy shocks affect the inflationary process.

5 References

Berkowitz, J. (2001) "Testing density forecasts, with applications to risk management", *Journal of Business and Economic Statistics*, 19, 465-474.

Carter, C. and R. Kohn (1994) "On Gibbs sampling for state-space models", *Biometrika*, 81, 541-553.

Carter, C. and R. Kohn (1997) "Semiparametric Bayesian inference for time series with mixed spectra", *Journal of the Royal Statistical Society*, Series B, B, 255-268.

Cecchetti, S.G. and M.A. Wynne (2003) "Defining price stability", *Economic Policy*, October 397-434.

Clark, T.E. (2006) "Disaggregate evidence on the persistence of consumer price inflation, *Journal of Applied Econometrics*, 21, 563-587.

Clark T.E. and M.W. McCracken (2009) "Averaging forecasts from VARs with uncertain instabilities", *Journal of Applied Econometrics*, forthcoming.

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 and S.P. Vahey (2009) "Measuring output gap uncertainty", unpublished manuscript, Birkbeck College, London, April.

Geman, S. and D. Geman (1984) "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", IEEE Transaction on Pattern Analysis and Machine Intelligence, 6, 721-741.

Gerlach, R., C. Carter, and R. Kohn (2000) "Efficient Bayesian inference for dynamic mixture models", *Journal of the American Statistical Association*, 95, 819-828.

Giordani, P. and R. Kohn (2008) "Efficient Bayesian inference for multiple change-point and mixture innovation models", *Journal of Business and Economic Statistics*, 26, 66-77.

Giordani, P., R. Kohn, and D. van Dijk (2007) "A united approach to nonlinearity, outliers and structural Breaks", *Journal of Econometrics*, 137, 112-137.

Gneiting, T. and A.E. Raftery (2007) "Strictly proper scoring rules, prediction and estimation", *Journal of the American Statistical Society*, March, vol 102, 477, 359-378.

Greenspan, A. (2004) "Risk and uncertainty in monetary policy", *America Economic Review*, Papers and Proceedings, May, 33-40.

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

Hersbach, H. (2000) “Decomposition of the continuous ranked probability score for ensemble prediction systems”, *Weather and Forecasting*, 15, 559-570.

Jore, A.S., J. Mitchell and S.P. Vahey (2009) “Combining forecast densities from VARs with uncertain instabilities”, *Journal of Applied Econometrics*, forthcoming.

Kascha, C. and F. Ravazzolo (2009) “Combining inflation density forecasts”, Norges Bank Working Paper 08-22, revised March.

Kiley, M. (2008) “Estimating the common trend rate of inflation for consumer prices and consumer prices excluding food and energy prices”, Federal Reserve Board, FEDS 2008-38, July.

Kim, S., S. Shephard, and S. Chib (1998) “Stochastic volatility: Likelihood inference and comparison with ARCH models, *Review of Economic Studies*, 65, 361-393.

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

Koop, G. and S. Potter (2007) “Estimation and forecasting in models with multiple breaks”, *Review of Economic Studies*, 2007, 74, 763-789.

Little, M.A., P.E. McSharry and J. W. Taylor (2008) “Parsimonious modeling of UK daily rainfall for density forecasting”, in Geophysical Research Abstracts, EGU General Assembly, Vienna 2008, Volume 10.

Mishkin, F. (2007) “Headline versus core inflation in the conduct of monetary policy”, speech given at the Business Cycles, International Transmission and Macroeconomic Policies Conference, HEC Montreal, October 20.

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 (2009) “Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness”, unpublished manuscript, University of Warwick.

Panagiotelis, A. and M. Smith (2008) “Bayesian density forecasting if intraday electricity prices using multivariate skew t distribution”, *International Journal of Forecasting*, 24, 710-727.

Quah, D.T. and S.P. Vahey (2005) “Measuring core inflation”, *Economic Journal*, 105, 1130-44.

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

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.

Shephard, N. (1994), “Partial non-Gaussian state-space models”, *Biometrika*, 81, 115-131.

Smith, J.K. (2004) “Weighted median inflation: is this core inflation?”, *Journal of Money, Credit and Banking*, 36, 253-263.

Tanner, M. and W. Wong (1987) “The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, 82, 528-550.

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

Wallis, K.F. (2005) “Combining density and interval forecasts: a modest proposal”, *Oxford Bulletin of Economics and Statistics*, 67, 983994.

A Time-Varying parameter model

A.1 Prior Specification and Posterior Simulation

We specify the following mixture innovation model for a given time series π:

$$\pi_\tau = \beta_{0\tau} + \sum_{p=1}^{k} \beta_{p\tau}\pi_{i\tau-p} + \sigma_\tau\varepsilon_\tau$$

$$\beta_{j\tau} = \beta_{j,\tau-1} + \kappa_{j\tau}\eta_{j\tau}, \quad j = 0, ..., k \tag{A-1}$$

$$ln\sigma^2_\tau = ln\sigma^2_{\tau-1} + \kappa_{k+1,\tau}\eta_{k+1,\tau}$$

where $\varepsilon_t \sim N(0,1)$, $\eta_\tau = (\eta_{0\tau}, ..., \eta_{k+1,\tau})' \sim N(0,Q)$ with Q a diagonal matrix and elements $q_0^2, ..., q_{k+1}^2$, and $\kappa_\tau = (\kappa_{0\tau}, ..., \kappa_{k+1,\tau})'$ is a $((k+2)\times 1)$ vector of unobserved uncorrelated 0/1 processes with $\Pr[\kappa_{j\tau} = 1] = p_j$ for $j = 0, ..., k+1$. The model parameters are the structural break probabilities $p = (p_0, ..., p_{k+1})'$ and the vector of variances of the size of the breaks $q = (q_0, ..., q_{k+1})'$. We collect the model parameters in a $(2(k+1)\times 1)$ vector $\theta = (p_0, ..., p_{k+1}, q_0, ..., q_{k+1})'$.

To facilitate the posterior simulation we make use of independent conjugate priors. For the structural break probability parameters we take Beta distributions

$$p_j \sim Beta(a_j, b_j) \tag{A-2}$$

The parameters a_j and b_j can be set according to our prior belief about the occurrence of structural breaks. For the variance parameters we take the inverted Gamma-2 prior

$$q_j^2 \sim \text{IG-2}(\nu_j, \delta_j) \tag{A-3}$$

where ν_j, δ_j are parameters which can be chosen to reflect the prior beliefs about the variances. Realistic values of the parameters in the different prior distributions depend on the problem at hand. In general, we suggest to assign to ν_j high values. This means to have strong believes that the magnitude of a break at time τ for parameter $\beta_{j\tau}$ (σ_t^2) associated to $\Pr[\kappa_{j\tau} = 1] = 1$ is equal to δ_j. The prior on (A-2) can consequently be chosen to limit the number of these breaks. As the posterior probability $\Pr[\kappa_{j\tau} = 1]$ is lower than 1, prior information is weak on breaks with magnitude lower than δ_j or situations of not changes.

Posterior results are obtained using the Gibbs sampler of Geman and Geman (1984) combined with the technique of data augmentation of Tanner and Wong (1987). The latent variables $B = \{\beta_\tau\}_{\tau=1}^T$, $R = \{\sigma_\tau^2\}_{\tau=1}^T$ and $K = \{\kappa_\tau\}_{\tau=1}^T$ are simulated alongside the model parameters θ.

The complete data likelihood function is given by

$$\begin{aligned} p(\pi, B, K, R|\theta) = \textstyle\prod_{\tau=1}^T p(\pi_\tau|\beta_\tau, \sigma_\tau^2) \prod_{j=0}^k p(\beta_{j\tau}|\beta_{j,\tau-1}, \kappa_{j\tau}, q_j^2) \\ p(\sigma_\tau^2|\sigma_{\tau-1}^2, \kappa_{k+1,\tau}, q_{k+1}^2) \textstyle\prod_{j=0}^{k+1} p_j^{\kappa_{j\tau}}(1-p_j)^{1-\kappa_{j\tau}}, \end{aligned} \tag{A-4}$$

where $\pi = (\pi_1, \ldots, \pi_T)$. The terms $p(\pi_\tau|\beta_{j\tau}, \sigma_\tau^2)$ and $p(\beta_{j\tau}|\beta_{j,\tau-1}, \kappa_{j\tau}, q_j^2)$ are normal density functions which follow directly from (A-1) and $p(\sigma_\tau^2|\sigma_{\tau-1}^2, \kappa_{k+1,\tau}, q_{k+1}^2)$ is an exponential normal density function. If we combine (A-4) together with the prior density $p(\theta)$, which follows from (A-2)-(A-3), we obtain the posterior density

$$p(B, K, R, \theta|\pi) \propto p(\theta)p(\pi, B, K, R|\theta). \tag{A-5}$$

For the Gibbs sampling procedure we employ the efficient sampling algorithm of Gerlach, Carter and Kohn (2000) to handle the (occasional) structural breaks. If we define $K_\beta = \{\kappa_{0t}, \ldots, \kappa_{kt}\}_{t=1}^T$ and $K_\sigma = \{\kappa_{k+1,t}\}_{t=1}^T$, the sampling scheme can be summarized as follows:

1. Draw K_β conditional on R, K_σ, θ and π.

2. Draw B conditional on R, K, θ, and π.

3. Draw K_σ conditional on B, K_β, θ, and π.

4. Draw R conditional on B, K, θ, and π.

5. Draw θ conditional on B, K, and π.

The (occasional) structural breaks, measured by the latent variable $\kappa_{j\tau}$, are drawn using the algorithm of Gerlach, Carter and Kohn (2000), which derives its efficiency from generating $\kappa_{j\tau}$ without conditioning on the states $\beta_{j\tau}$ (σ_τ^2). The conditional posterior density for $\kappa_{*,\tau}$, $\tau = 1, \dots, T$ unconditional on B is

$$\begin{aligned} p(\kappa_{*,\tau}|K_{*,-\tau}, K_{k+1}, R, \theta, \pi) &\propto p(\pi|K_*, K_{k+1}, R, \theta) p(\kappa_{*,t}|K_{*,-\tau}, \theta) \\ &\propto p(\pi_{\tau+1}, \dots, \pi_T|\pi_1, \dots, \pi_\tau, K, R, \theta) \\ &\qquad p(\pi_\tau|\pi_1, \dots, \pi_{\tau-1}, \kappa_1, \dots, \kappa_\tau, R, \theta) p(\kappa_{*,\tau}|K_{*,-\tau}, \theta), \end{aligned} \quad \text{(A-6)}$$

where $K_{*,-\tau} = \{\kappa_{*,s}\}_{s=1,s\neq\tau}^T$. Note that the term $p(\kappa_{*,\tau}|K_{*,-\tau}, \theta)$ is simply given by $\prod_{j=0}^k p_j^{\kappa_{jt}}(1-p_j)^{1-\kappa_{jt}}$. The two remaining densities $p(\pi_{\tau+1}, \dots, \pi_T|\pi_1, \dots, \pi_\tau, K, R, \theta)$ and $p(\pi_\tau|\pi_1, \dots, \pi_{\tau-1}, \kappa_1, \dots, \kappa_\tau, R, \theta)$ can be evaluated as shown in Gerlach, Carter and Kohn (2000). Because $\kappa_{*,\tau}$ can take a finite number of values, the integrating constant can easily be computed by normalization.

The full conditional posterior density for the latent regression parameters B is computed using the simulation smoother as in Carter and Kohn (1994). The Kalman smoother is applied to derive the conditional mean and variance of the latent factors.

To draw K_σ and R in steps 3 and 4 we want to follow the same approach. As the model for $\ln \sigma_t^2$ does not result in a linear state space model the Kalman filter cannot be applied. Therefore, we apply the approach of Giordani and Kohn (2008) and rewrite the model (A-1) as

$$\begin{aligned} \ln(\pi_\tau - \beta_{0\tau} - \sum_{p=1}^k \beta_{p\tau}\pi_{i\tau-p})^2 &= \ln \sigma_\tau^2 + u_\tau \\ \beta_{j\tau} &= \beta_{j,\tau-1} + \kappa_{j\tau}\eta_{j\tau}, \qquad j = 0, \dots, k, \\ \ln \sigma_\tau^2 &= \ln \sigma_{\tau-1}^2 + \kappa_{k+1,\tau}\eta_{k+1,\tau} \end{aligned} \quad \text{(A-7)}$$

where $u_\tau = \ln \varepsilon_\tau^2$ has a log χ^2 distribution with 1 degree of freedom. We follow Carter and Kohn (1994), Carter and Kohn (1997), Shephard (1994) and Kim, Shephard, Chib (1998) who show that the $\ln \chi^2(1)$ distribution can be approximated very accurately by a finite mixture of normal distributions. We consider a mixture of five normal distributions such that the density of u_τ is given by

$$f(u_\tau) = \sum_{s=1}^{5} \varphi_s \frac{1}{\omega_s} \phi((u_t - \mu_s)/\omega_s). \quad \text{(A-8)}$$

with $\sum_{s=1}^{5} \varphi_s = 1$. The appropriate values for μ_s, ω_s^2 and φ_s can be found in Carter and Kohn (1997, Table 1). In each step of the Gibbs sampler we simulate a component of the mixture distribution from the distribution of the mixing distribution. Given the value of the mixture component we can apply standard Kalman filter techniques. Hence, the variables K_σ and R can be sampled in a similar way as K_β and B in step 1 and 2.

To sample the parameters θ we can use standard results in Bayesian inference. Hence, the probabilities π_j are sampled from Beta distributions, and the variance parameters q_j^2 are sampled from inverted Gamma-2 distributions.

A.2 Predictive density

The one-step ahead predictive density of $\pi_{\tau,1}$ at time τ conditional on I_τ is given by

$$\begin{aligned} p(\pi_{\tau,1}|I_\tau) = \iint \sum_{K} \sum_{\kappa_{\tau+1}} p(\pi_{\tau+1}|S, \beta_{\tau+1}, \sigma_{\tau+1}^2) \prod_{j=0}^{k} p(\beta_{j,\tau+1}|\beta_{j,\tau}, \kappa_{j,\tau+1}, q_j^2) \\ p(\sigma_{\tau+1}^2|\sigma_\tau^2, \kappa_{k+1,\tau+1}, q_{k+1}^2) \prod_{j=0}^{k+1} p_j^{\kappa_{j,\tau+1}} (1-p_j)^{1-\kappa_{j,\tau+1}} p(B, K, R, S, \theta|\pi) dB dR d\theta, \quad \text{(A-9)} \end{aligned}$$

where $p(\pi_{\tau+1}|\beta_{\tau+1}, \sigma_{\tau+1}^2)$ and $p(\beta_{j,\tau+1}|\beta_{j,\tau}, \kappa_{j,\tau+1}, q_j^2)$ and $p(\sigma_{\tau+1}|\sigma_\tau, \kappa_{\tau+1}, q_{k+1}^2)$ follow directly from (A-1) and where $p(B, K, R, S, \theta|\pi)$ is the posterior density in (A-5) using information I_τ. Computation of this predictive density is straightforward using the Gibbs draws. In each Gibbs step, we simulate the $\pi_{\tau+1}$ using (A-1) as data generating process, where we replace the parameters and the latent variables by the draw from the posterior distribution. As point estimate we use the posterior median. The procedure can be easily extended to h-step ahead forecasts.

The procedure can be applied to each of the N disaggregates to derive the predictive density $p(\pi_{\tau,h} \mid I_{i,\tau})$ for each disaggregate i, $i = 1, \ldots, N$.

A.3 Centering component forecast densities

The density $p(\pi_{\tau,h}|I_{i\tau})$ forecasts future values of the disaggregate i, $i = 1, \ldots, N$. We (recursively) center the predictives from the disaggregate component models on the outturns of measured inflation, π, using the following two-step process.

First, we estimate with Ordinary Least Squares (OLS):

$$\pi_s = a + p(\pi_{s,h}|I_{i,s})_m + \varepsilon_s \quad s = \underline{\tau}_1, \ldots, \tau - 1 \tag{A-10}$$

where $p(\pi_{s,h}|I_{i,s})_m$ is the point forecast implied by the predictive density $p(\pi_{s,h}|I_{i,s})$, derived by taking the median value.

Second, we adjust the predictive density:

$$g(\pi_{\tau,h} \mid I_{i,\tau}) = p(\pi_{\tau,h} \mid I_{i,\tau}) - \widehat{a} \tag{A-11}$$

where $\widehat{a}$ is the OLS estimates of a in (A-10). We use $g(\pi_{\tau,h} \mid I_{i,\tau})$ to score the disaggregate forecast densities, and to construct the ensemble predictives.

B CRPS computation for components

The Continuous Ranked Probability Score is a proper scoring rule for forecast densities; see Gneiting and Raftery (2007). It rewards forecasts that assign high probability to events that are realized, and also rewards high probabilities near the event. The concentration of a density about its central location is sometimes referred to as "sharpness", and the location as distance. The CRPS favours densities with small distance and high sharpness.

The CRPS is measured as the difference between the predicted and actual cumulative distribution. Figure 5 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.

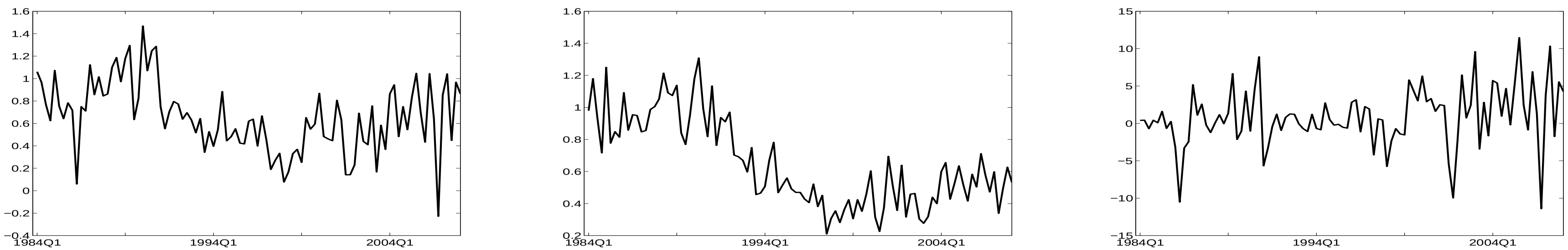
More formally, following Panagiotelis and Smith (2008), the CRPS of a component density for a particular observation can be defined as:

$$CRPS = E_F|\pi - \pi_o| - 0.5E_F|\pi - \pi'| \tag{A-12}$$

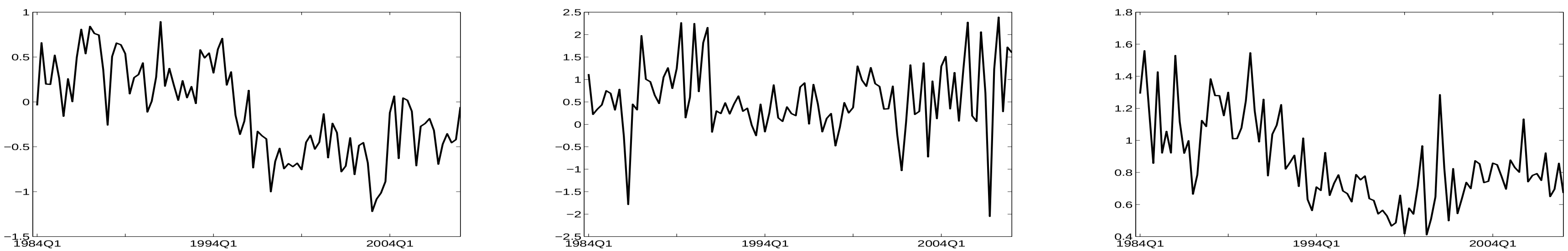
where E_F is the expectation for the predictive F, π and π' are independent random draws from the predictive, and π_o is the observed outturn. We approximate the expectation terms using the Monte Carlo draws from the component forecast density; Panagiotelis and Smith (2008, equation 4.5) provide the computational steps required.

For each component density, we construct the mean CRPS at each observation, averaged over the evaluation period. The weight on an individual component density i in each period is then calculated using equation (2), where X is the inverse of the mean CRPS.

Figure 1: Data



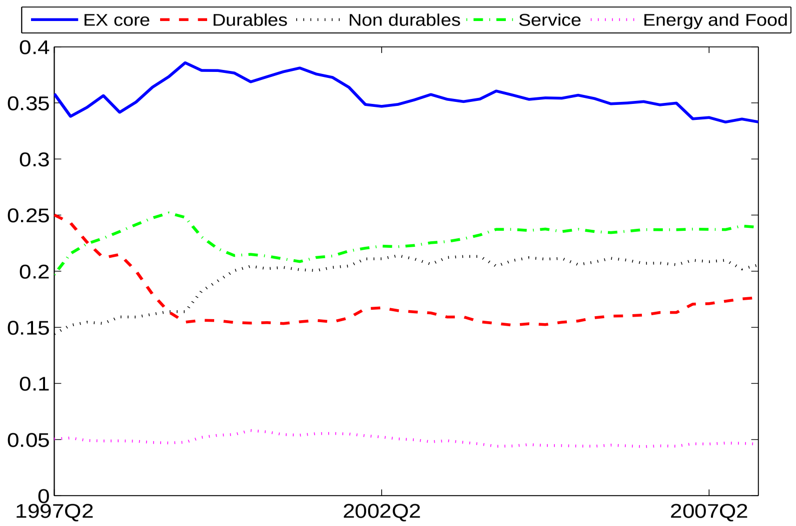
(a) PCE - Ex core PCE - Food & Energy

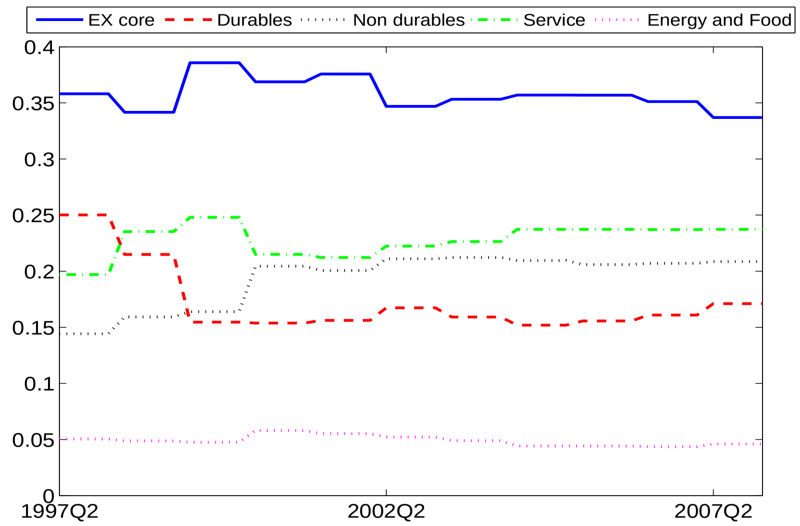


(b) DUR - ND -SERVICES

Note: The graphs in these figures show PCE inflation and its disaggregate over the sample period 1984:Q1-2008Q1.

Figure 2: Core inflation weights, DE5 and DE5-ANN

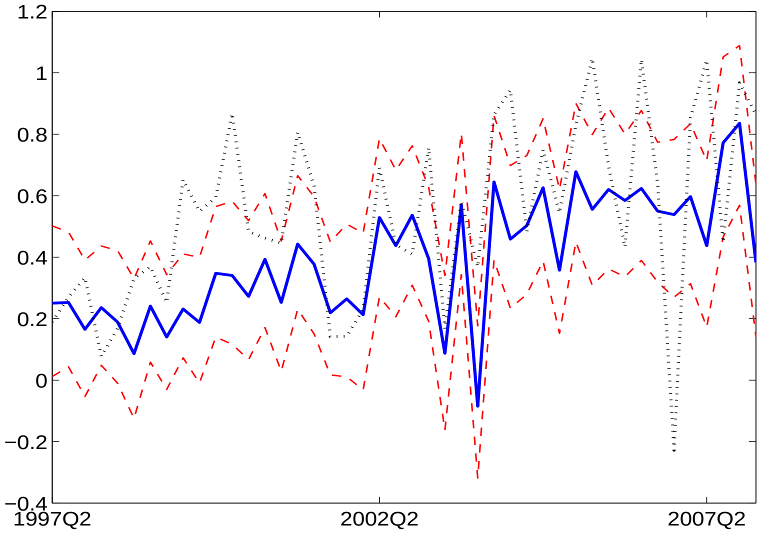




(a) DE5 - DE5-ANN

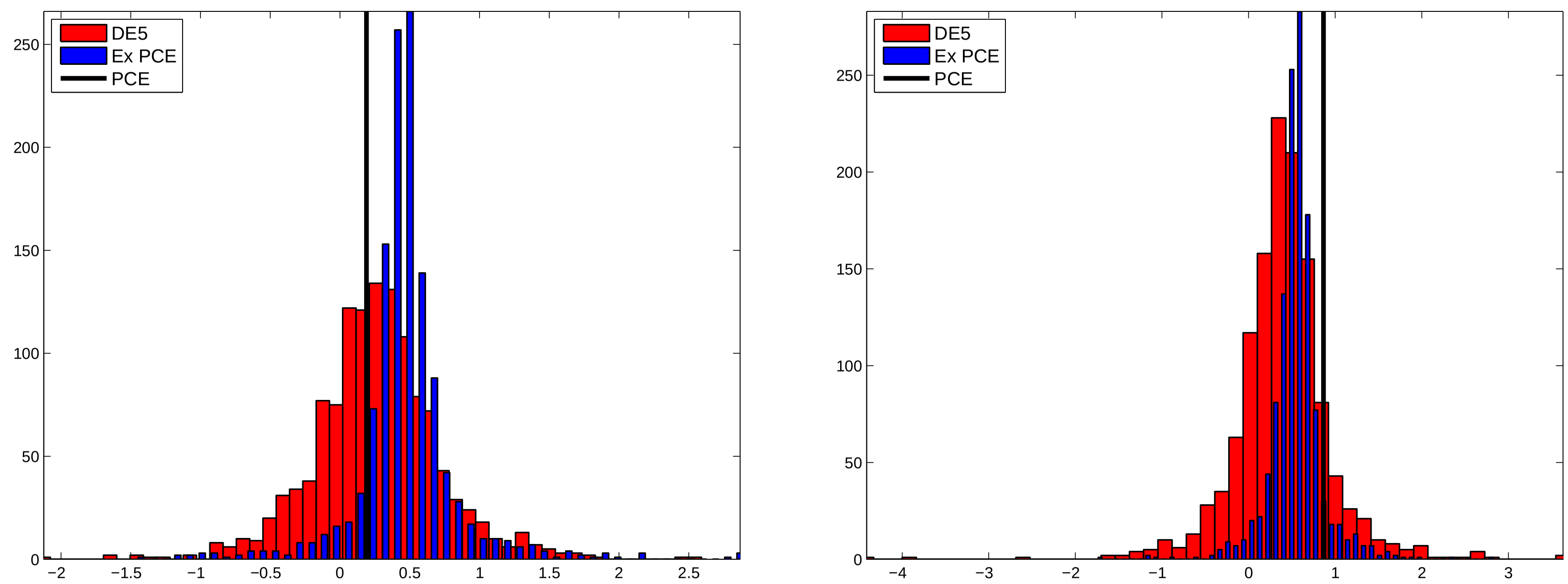
Note: The figures plot in the left panel weights given by disaggregate ensemble DE5, in the right panel weights given by disaggregate ensemble DE5-ANN.

Figure 3: Core inflation interval forecasts



Note: The figure shows the posterior median (solid line) of the predictive density given by disaggregate ensemble DE5 and the actual inflation (in dashed line). The red dashed lines in the graphs are the 25th and 75th percentiles of the predictive density.

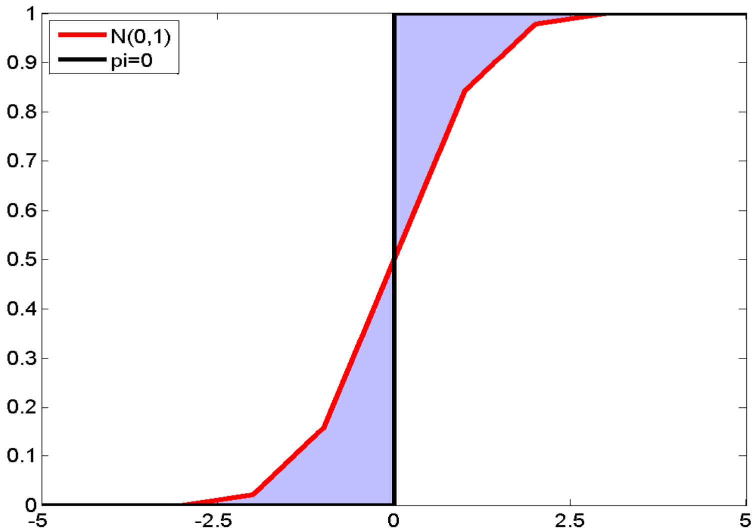
Figure 4: Ex Core and DE5 density forecasts



(a) 1997Q2-2008Q1

Note: The figures plot the histogram of the density forecasts given by Ex Core measure and by the disaggregate ensemble DE5 for two different periods, the first and last forecasts. The realized value for PCE is also provided.

Figure 5: 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 realized value 0. The colored area measure the CRPS.

Table 1: Point forecast performance, various ensembles

	RMSPE	HLN test
	Individual benchmarks	
AR(2)	*0.163*	
Ex PCE	1.354	0.000
	Disaggregate Ensembles	
DE2	1.199	0.006
DE3	1.615	0.000
DE5	1.191	0.002
DE5-ANN	1.197	0.002

Note: The columns "RMSPE" report the RMSPE to forecast PCE over the sample 1997Q2-2008Q1 using AR(2) models for PCE, a TVP for Ex PCE and disaggregate ensembles. The first cell in the table reports in *italics* the value using an AR(2) for PCE, while all the other numbers report statistics relative to those of the AR(2). The column "HLN test" reports the p-values of the Harvey, Leybourne and Newbold (1997) test for equal predictive accuracy of the AR(2) for PCE versus all the other models.

Table 2: Density forecast performance, various ensembles

	LS-test	LS	LR
	Individual benchmarks		
AR(2)		*-0.427*	0.123
Ex PCE	0.199	0.705	0.215
	Disaggregate Ensembles		
DE2	0.257	0.957	0.162
DE3	0.714	0.862	0.169
DE5	0.014	0.597	0.174
DE5-ANN	0.017	0.615	0.174

Note: The column LS-test is the p-value of Mitchell and Hall (2005) logarithmic Scores test of equal density predictive accuracy of the AR(2) models versus all the other methods (Ex PCE, and disaggregate ensembles) over the sample 1997Q2-2008Q1. LS is the average logarithmic scores, average over the evaluation period. The cell for AR gives the value for the AR(2) model, while all the other LS's report statistics relative to those of the AR(2). Numbers lower than 1 indicates that the competitor provide a mean LS lower than the AR(2) and therefore outperform it. LR is the Likehihood Ratio p-value of the test of zero mean and unit variance of the inverse normal cumulative distribution function transformed *PITS*, with a maintained assumption of normality for transformed *PITS*.