Conference on

Forecasting and Monetary Policy

Berlin, 23-24 March 2009

Simon Price
Bank of England and City University

Jan Groen
Federal Reserve Bank of New York

George Kapetanios
Queen Mary College

„Real-Time Evaluation of Inflation Report Forecasts for Inflation and Growth“

Real time evaluation of *Inflation Report* forecasts for UK inflation and growth *

Jan J. J. Groen† George Kapetanios ‡ Simon Price§

5 June 2008

Abstract

We compare the Bank of England's *Inflation Report* (*IR*) quarterly forecasts for growth and inflation to real-time forecasts using a variety of statistical forecasting models that have previously been found useful as forecasting benchmarks. These include linear and non-linear univariate models, and VARs. The results reveal the well-known difficulty of forecasting in a stable macroeconomic environment, and the *IR* forecasts for GDP growth are inferior to forecasts from linear and non-linear univariate models. However, for the inflation forecast the *IR* strongly dominates. The explanation for these inflation results appears to be that the official forecasts incorporate judgements about monetary regime changes, leading to inflation mean shifts, which improve forecast performance markedly. The joint use of rolling windows and simple unweighted forecast combination helps improve the statistical models' performance, although where they outperform the *IR*, the improvement is not significant.

Keywords: real-time data, forecast performance, inflation, growth.

JEL: C530

1 Introduction

Forecasts of inflation and output growth are central to the practical operation of monetary policy by central banks.[1] In the case of the Bank of England, 'fan charts' published in the quarterly *Inflation Report* (*IR*) show the whole distribution of forecasts of inflation and growth. These

*The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England, the Federal Reserve Bank of New York or the Federal Reserve System. The authors wish to thank an Associate Editor, two anonymous referees, David Hendry, and seminar participants at the 2007 International Symposium on Forecasting in New York, the 2008 Annual Conference of the Royal Economic Society at Warwick University, the Bank of England and City University for helpful comments.

†Federal Reserve Bank of New York, jan.groen@ny.frb.org

‡Queen Mary and Westfield College, g.kapetanios@qmul.ac.uk

§Bank of England and City University, simon.price@bankofengland.co.uk

[1] Indeed, it is sometimes suggested that inflation-targeting central banks should target inflation forecasts (Svensson (1997)), and sometimes that some do (Leitemo (2003)). There is a case to be made against this practice, for example in Woodford (2000): but there is no doubt that even if central banks do not target inflation forecasts, they certainly take great care to produce and publish them.

encapsulate the judgement of the policy-setting Monetary Policy Committee (MPC) for the prospects for inflation and growth, conditioned on specific assumptions, including interest and exchange rates and some exogenous variables, as well as on general judgements about the future. From the *IR*, we read that these 'represent the MPC's best collective judgement about the most likely paths for inflation and output, and the uncertainties surrounding those central projections.' Moreover, it should be understood that 'the projections .... [are not] mechanically produced by models: they reflect the judgments of the Committee'.[2] It is often argued (e.g., Wallis and Whitley (1991)) that judgements generally improve forecasts. So it would be interesting to know whether this is the case in the UK.

This has been particularly pertinent over the recent past, during which the economy was relatively stable. Forecasting many macroeconomic variables is hard: beyond some horizon, we cannot outperform the unconditional mean. These horizons can be quite short.[3] Data typically have some obvious short-run cyclical variation which has to be accounted for, but it is often possible to capture this with a simple AR process. The reasons for this are not hard to understand. Stock and Watson (2007) point out that the well-documented move towards macroeconomic stability, sometimes referred to as the 'Great Moderation', has made forecasting easier in the sense that macroeconomic variables stray less far from the unconditional mean than in the past, but more difficult in the sense that it is hard to outperform naïve models.

Recent experience in the UK is consistent with this view. The Bank of England has constructed a 'suite' of purely statistical forecasting models[4] as statistical benchmarks to help inform the forecast and policy processes, which the MPC considers as part of the wide range of information that it examines. An interesting question, then, is to examine how official forecasts for growth and inflation compare with such benchmark models. Thus in this paper we do precisely that, including autoregressive models, a random walk (RW), some non-linear univariate models and a small range of vector autoregressive (VAR) models, the latter estimated by both OLS and Bayesian methods (BVARs). Naturally, there are other forecasting models that could have been included in this comparison. But our main aim is simply to assess how official forecasts measure against sensible econometric benchmarks, rather than find which specific, tailormade model performed best in this particular period for this particular data set.[5]

One complicating issue is the conditionality of the forecasts. The Bank of England's forecasts considered here are conditioned on a particular interest rate (in this case held constant at the most recent value).[6] The point is that the conditioning path for interest rates may not be the best forecast: see the arguments made by Woodford (2000) on some pitfalls facing forward-looking

[2] Bank of England (1999), *Economic models at the Bank of England.*

[3] Galbraith and Tkacz (2006) present some recent evidence for the United States and Canada.

[4] Described in Kapetanios, Labhard, and Price (2008). The statistical suite was constructed partly in response to the Pagan (2003) report on the use of models at the Bank of England. The notion that the Bank of England uses a wide variety - a suite - of models in its analysis predates this purely statistical set: see Bank of England (1999).

[5] We note that were the official forecasts found to be no better than or inferior to statistical forecasts, that would not be grounds for replacing them. As we alluded to earlier, central banks' forecasts are published as a part of the policy process, and reflect views about the workings of the economy and appropriate policy judgements, as well as offering opportunities to explain policy decisons. Hence the MPC's model and judgement based forecast could never be replaced by purely econometric models.

[6] Faust and Wright (2006) discuss how to evaluate such forecasts. They conclude that the *IR* forecasts (and the *Greenbook* forecasts also examined in the working paper version of this paper, Groen, Kapetanios, and Price (2008b)) may be inefficient.

policymakers. However, the Bank of England's own published evaluations of the *IR* forecasts unambiguously refer to the 'forecasting record' (*IR* August 2004, pp 50-51). Thus we treat the conditional projections as if they were forecasts. Nevertheless, this conditionality must bias to some extent the tests against the official forecasts.[7] It is also crucial in this assessment that we condition on the data set that was available at the time the *IR* forecasts were generated, and we do so using a (limited) real-time data base that has been compiled at the Bank of England.

There has been very little evaluation of Bank of England forecasts, in contrast to the close attention that has been paid to the Federal Reserve Board (FRB) *Greenbook* forecasts. A very recent example is Edge, Kiley, and Laforte (2008). Their primary aim is to asses the use of an estimated DSGE model in forecasting, but they include the staff forecasts using real-time data in their comparisons to a 12 variable VAR(1) estimated by OLS and a Bayesian method, as well as to internal forecasts and the DSGE model. They restrict attention to the late 1990's (September 1996 to March 2001) for the *Greenbook* comparisons. At the horizons they report, (1-4 and 8 quarters) the *Greenbook* beats[8] the benchmark VAR(1) only at the highest horizon for GDP growth, and then by a small margin. By contrast, for GDP inflation, the *Greenbook* beats the benchmark VAR(1)and BVAR(1) at all horizons by substantial margins. The flavour of their results is broadly similar to that in Sims (2002), who examined the *Greenbook* record over the period 1979 to 1995. He concluded the inflation forecasts were very respectable,[9] while for output growth the record was less outstanding. More recently, Gamber and Smith (2007) argue that the FRB has an inflation forecasting edge over the private sector, although the advantage has eroded, especially after the FRB shifted towards more transparency in 1994. Finally, Adolfson, Andersson, Lindé, Villani, and Vredin (2007) conduct a similar exercise to Edge, Kiley, and Laforte (2008) (albeit not with real-time data) for the Sveriges Riksbank's inflation projections (as published in the Riksbank's *Inflation Report*) since its independence in 1999. Their results indicate that a BVAR model performs better than (or at least as well as) the Riksbank forecasts beyond the one quarter horizon.

There is also more to the Bank of England's forecasts than the central estimates; an estimate of the entire forecast probability distribution is published. We do not attempt to assess the forecasts in terms of their density, but others have done so. Clements (2004) evaluates the inflation density forecasts (conditioned on constant interest rates) published between August 1997 and February 2002. The point forecast tests show these forecasts are unbiased in the periods examined (the first two quarters and the one year horizon), and are at least as accurate as naïve forecasts in the first two quarters. Tests on the entire fan chart distribution show the same pattern; the first two quarters perform well compared to a naïve distribution, but it is found lacking at the one-year-ahead point because it overestimates the probability of high inflation outturns.

In the remainder of this paper, we first summarise the forecasting exercise we undertake and describe our methodology in Section 2. In Section 3, we first briefly describe the data and our sources and then report our first pass results for output growth and inflation forecasts. Section 4 focusses in more detail on the inflation forecast comparison and provides a rationale for the added value of judgements when it comes to forecasting inflation. The final section

[7] The Bank of England also publishes forecasts conditioned on the market path, and has recently given more emphasis to these. However, they are not available for as long a period as the constant rate forecasts. Neither can it be presumed that the market path outperforms a constant rate.

[8] Using a mean absolute error criteria, but the authors also report RMSE results that are similar.

[9] This conclusion depends on the sample period. For the post 1984 period, naïve models performed as well.

concludes.

2 Methodology

In this section we describe the models we use and our evaluation methodology. Our aim is not to run a forecast horse race and find the best model, but to compare the official forecasts to sensible statistical benchmarks. The performance of some such models used at the the Bank of England is described in Kapetanios, Labhard, and Price (2008), and our choices are largely selected from among those. Simple univariate models are often found to have good performance, and this appears to be the case for the UK in the recent data. Marcellino (2006) assesses a large number of linear and non-linear specifications in a growth and inflation forecasting exercise, and finds that although some non-linear models are good benchmarks some of the time, univariate linear models are robust. One constraining factor that we face in the current exercise is the lack of a comprehensive real-time data base. This effectively prevents us from estimating data-rich models, such as those described and evaluated in Kapetanios, Labhard, and Price (2008).

Allowing for mean-shifts is particularly pertinent for inflation, where there is clear evidence of breaks after recent changes in monetary regime (namely, the shift to inflation targeting in 1992 and Bank of England independence in 1997). One way to handle this is by rolling regressions, as emphasised in Pesaran and Timmermann (2007). The approach initially adopted here is instead to pre-test for structural breaks using the method of Bai and Perron (1998), and then to forecast the de-meaned series if breaks are detected. Implicitly, it assumes the mean shift is not associated with changing dynamics. This has the advantage of allowing the use of longer samples in estimation, and will be more effective the less frequent are structural breaks. But we subsequently examine rolling regressions and forecast combination.

2.1 Models

Turning to the specific models we examine, the two linear univariate models we include are the autoregressive (AR) model,

$$\Delta y_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \Delta y_{t_j} + \epsilon_t^{\Delta y} \tag{1}$$

and the random walk (RW) model

$$\Delta y_t = \Delta y_{t-1} + \epsilon_t^{\Delta y} \tag{2}$$

where $\epsilon_t^{\Delta y} \sim$ i.i.d.$(0, \sigma^2)$. Δy_t is inflation or output growth. The lag order p is determined using the Akaike information criterion (AIC), starting at $p^{max} = 4$. The random walk is a standard naïve benchmark. Although it may work well for some variables and horizons, in practice it is outperformed by low order AR models.

Equations (1) and (2) are used, after estimation up to the forecasting date t_0 in case of (1), to generate forecasts for horizons $h = 1, \ldots, 8$ in an indirect, iterative manner.[10] For the estimation of 1 up to forecast date t_0, we use the actual 'vintage' or snapshot of Δy data that

[10] For the RW model the forecast will be $\Delta y^e_{t+8} = \Delta y^e_{t+7} = \Delta y^e_{t+6} = \Delta y^e_{t+5} = \Delta y^e_{t+4} = \Delta y^e_{t+3} = \Delta y^e_{t+2} = \Delta y^e_{t+1} = \Delta y_t$.

were available at time t_0. So we are not substituting our current observations of Δy for the period up to t_0.[11]

We also use two non-linear univariate models: an exponential smooth transition autoregressive (exponential STAR) model and a Markov-Switching (MS) autoregressive model. Kapetanios, Labhard, and Price (2008) find the former often performs well on UK data. The STAR model is given by

$$\Delta y_t = c + \sum_{i=1}^{p} \delta_i(\Delta y_{t-i} - c) + [(1 - e^{-\theta(\Delta y_{t-d} - c)^2}) \sum_{i=1}^{p} \gamma_i(\Delta y_{t-i} - c)] + \epsilon_t^{\Delta y} \tag{3}$$

where c, d, δ_i and γ_i are parameters to be estimated. c can be interpreted as the mean of the process Δy_t. This model essentially implies that there is one autoregressive model for Δy_t when Δy_{t-1} is close to c and another when Δy_{t-1} is far away from c. In the case of inflation, if c is viewed as the inflation target then this means that policy becomes more active when y_{t-1} is away from the target than otherwise. Estimation of the model is by nonlinear least squares (NLS). In practice, the parameters c and θ are difficult to obtain. We use a grid search to obtain some ideas on their values prior to using NLS with the outcome of the grid search as initial conditions. We specify the MS model as

$$\Delta y_t = c_{s_t} + \sum_{i=1}^{p} \gamma_{i,s_t}(\Delta y_{t-i} - c_{s_t}) + \epsilon_t^{\Delta y} \tag{4}$$

where s_t is a Markov Chain taking values in the set $\{1, ..., M\}$ with transition matrix P. This model essentially implies that there are M regimes in the economy regulated by an unobserved Markov Chain. The model is estimated by maximum likelihood using the filter suggested by Hamilton (1989).

The multivariate models we consider are vector autoregressive (VAR) models, estimated by OLS and Bayesian methods. In the latter case we employ the standard Minnesota prior, a flat prior that the specification is a random walk. The first set of VAR models comprise 'macroeconomic' models,

$$\begin{pmatrix} \Delta y_t \\ \Delta x_{1t} \\ \vdots \\ \Delta x_{5t} \end{pmatrix} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_6 \end{pmatrix} + \sum_{j=1}^{p} \Gamma_j \begin{pmatrix} \Delta y_{t-j} \\ \Delta x_{1,t-j} \\ \vdots \\ \Delta x_{5,t-j} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{6t} \end{pmatrix} \tag{5}$$

where Γ_j has dimension 6×6 and

$$\begin{pmatrix} \Delta x_{1t} \\ \vdots \\ \vdots \\ \Delta x_{5t} \end{pmatrix} = \begin{pmatrix} \text{Annual inflation or output growth} \\ \text{Oil price inflation} \\ \text{Annual change effective exchange rate} \\ \text{Short-Term Interest Rate} \\ \text{Real Factor} \end{pmatrix}$$

estimated by OLS (MacroVAR) or Bayesian methods (BMacroVAR). The 'real factor' variable in VAR model (5) is the first principal component extracted from the following 5-variable set of

[11] The data that we use to evaluate these and our other forecasts is discussed below.

real-time real activity series:

$$\begin{pmatrix} z_{1t} \\ \vdots \\ \vdots \\ z_{5t} \end{pmatrix} = \begin{pmatrix} \text{Annual growth rate real household consumption} \\ \text{Annual growth rate real whole economy investment} \\ \text{Annual growth rate real imports} \\ \text{Annual growth rate real exports} \\ \text{Change in inventories (quarterly, in units)} \end{pmatrix}. \tag{6}$$

The principal component is extracted by first taking the original vintages of the variables in (6) and standardise them (denoted $\tilde{Z}$, $T \times 5$), then taking the eigenvector that corresponds with the first eigenvalue of $\tilde{Z}'\tilde{Z}$ (denoted $v_{\tilde{Z}}$). The 'real factor' variable in (5) then equals $\tilde{Z}v_{\tilde{Z}}$. This process is repeated for each vintage of real activity data (so we obtain vintages of principal components).

One interesting recent exercise for the UK suggests that money can be helpful in some periods: see Garratt, Koop, Mise, and Vahey (2007). We therefore also consider 'monetary' VAR models,

$$\begin{pmatrix} \Delta y_t \\ \Delta x_{1t} \\ \Delta x_{2t} \end{pmatrix} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix} + \sum_{j=1}^{p} \Gamma_j \begin{pmatrix} \Delta y_{t-j} \\ \Delta x_{1,t-j} \\ \Delta x_{2,t-j} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{pmatrix} \tag{7}$$

where

$$\begin{pmatrix} \Delta x_{1t} \\ \Delta x_{2t} \end{pmatrix} = \begin{pmatrix} \text{Annual change narrow money} \\ \text{Annual change broad money} \end{pmatrix},$$

again estimated by OLS (MonVAR) or Bayesian methods (BMonVAR).

Forecast combinations have been used with great success across a wide array of applications. Timmermann (2006) provides a number of potential reasons that can explain their success. One is diversification across information sets when it is not feasible to pool the underlying individual information sets to construct a single, overall model. Also, in case of unknown misspecification bias combining individual forecasts can be seen as a way to make the forecast robust again this bias. Finally, if structural breaks occur, forecasts from individual models may be affected differently, possibly due to differences in the aforementioned misspecification bias across the models. It is exactly for this reason that Hendry and Clements (2002) claim that forecast combinations might be able to outperform forecasts from individual models when the underlying data generating process is subject to deterministic shifts. We therefore also compare the *IR* projections with a combination of the forecasts from the eight individual statistical models. In doing that, we have to take a stand on how to combine these individual models. Estimates of the individual model weights can often be substantially biased, which essentially reflects a 'generated regressor' problem as the forecast combination combines recursive generated forecasts from individual models. This may explain why simple, unweighted, averages of individual forecasts are often found to perform well empirically. Additionally, Hendry and Clements (2002) find that in the case of deterministic breaks, an unweighted average of forecasts often perform well, even when some of the underlying models are nested. This has been explored in the context of unstable VARS by Clark and McCracken (2008) and Jore, Mitchell, and Vahey (2008) (the latter focussing on density forecasts). Hence, our forecast combination will be a simple, unweighted average of the projections from our statistical models.

The forecasting models are recursively updated in the following sequence.

1. The first official forecast considered is generated at t_0.

2. Apply the Bai-Perron multiple break test on the forecast variable Δy_t up to t_0 and demeaning Δy_t if breaks are detected, using the vintage of data available in t_0. In practice, for GDP growth no breaks are detected.

3. Estimate statistical models on the *demeaned* (if applicable) forecast variable Δy_t over the sample $t = 1, \ldots, t_0$.

4. Generate forecast for $h = 1,\ 4,\ 8$ using the estimates up to t_0; for $h > 1$ construct iterative forecasts.

5. Remean the forecasts to construct a final mean estimate up to t_0.

6. Forecast errors are computed relative to the final vintage of Δy_t data.

7. Repeat for $t_0, \ldots, T - h$ for each h.

2.2 Forecast evaluation: inference

It is important to establish which release of data against which the forecasts should be evaluated. In the UK, as in many other countries, there are substantial revisions to data, often for several years after the initial release.[12] Some of these revisions flow from new information collected by the statistical agency. Others - often quantitatively important - arise from methodological changes, which it may be thought would be hard for the forecaster to predict. Consequently it is often argued that policymaking forecasters cannot be expected to forecast final releases of data. Thus Clark and McCracken (2008), for example, evaluate their models against the second release. But the Bank of England does in fact aim to forecast final releases of data. There is abundant textual evidence that this is the case. For example, in the footnote to Chart 5.1 on page 39 of Bank of England (2008) it is made clear that the forecasts of GDP growth reported there are for final releases of data ('the mature estimate'). Recently, the *IR* forecasts make this explicit, by publishing error distributions around the past data that arise from data revisions. This is not at all a change in practice, as can be inferred from Elder, Kapetanios, Taylor, and Yates (2005), an official examination of the Bank of England's forecast errors. On page 340 we read that the 'MPC is well aware of the likelihood of data revisions and takes account of that when making forecasts.' The forecast evaluations carried out in that article are therefore carried out with the latest available data. Thus we proceed on that basis in our own work, and use the most recent vintages available.

We evaluate forecasts exclusively on the basis of the square root of the mean squared forecast error (RMSE). As is well known, this is justified if forecasters are minimising quadratic loss. In an interesting paper, Patton and Timmermann (2007) examine the Federal Reserve's forecasts of output growth and find that they would be optimal if overpredictions of output growth were costlier than underpredictions. However, the Bank of England's inflation remit[13] has been

[12] Faust, Rogers, and Wright (2005) quantify the revisions to GDP for the G-7 economies and they also find that for some countries, including the UK, these revisions are predictable.

[13] Specified in a succession of letters from the Chancellor of the Exchequer: see HMT (1999) for a statement of the overall policy.

explicitly symmetric over the evaluation period, and consequently quadratic loss is arguably a good approximation. It is certain that an asymmetric loss function would be a misspecification.

The forecast evaluation criterion is therefore the relative root mean square error:

$$\frac{\sqrt{MSE_S^h}}{\sqrt{MSE_{IR}^h}} \tag{8}$$

with

$$MSE^h = \frac{1}{T - t_0 - h} \sum_{s=t_0}^{T-h} e_{s,s+h}^2.$$

We test for significant differences using the Diebold-Mariano-West test[14] for the MSE difference:

$$z_{MSE} = \sqrt{T - t_0 - h} \left(\frac{MSE_S^h - MSE_{IR}^h}{\sqrt{Var(u_{t+h} - (MSE_S^h - MSE_{IR}^h))}} \right) \tag{9}$$

with $u_{t+h} = e_{S,s,s+h}^2 - e_{IR,s,s+h}^2$ and $Var(u_{t+h} - (MSE_S^h - MSE_{IR}^h)$ estimated as in Newey and West (1987) for $h > 1$.

Inference based on the Diebold-Mariano-West statistic z_{MSE} can be problematic. If the *IR* and statistical models are non-nested, then we have in (9) $z_{MSE} \sim \mathrm{N}(0, 1)$. However, if they are nested then z_{MSE} in (9) converges to a Brownian motion functional; see Clark and McCracken (2005). In practice, the more judgement dominates the *IR* projections the more these projections become non-nested in our statistical benchmark models.[15] However, as it is hard to assess how much in each period the *IR* forecasts are driven by judgement, it is equally hard to take a stand on the asymptotic distribution of the z_{MSE} statistic in (9). Moreover, there may be finite sample bias, autocorrelation, (autoregressive) heteroskedasticity, and so on. We therefore use a bootstrap distribution for the z_{MSE} statistic in (9) to construct a *one-sided* test for the null hypothesis that the models have lower RMSE than the *IR* benchmark: $H_0 : MSE_S^h - MSE_{IR}^h = 0$ *versus* $H_1 : MSE_S^h - MSE_{IR}^h < 0$.

The algorithm for the bootstrap is as follows.

1. Center $u_{t+h} = e_{S,t,t+h}^2 - e_{IR,t,t+h}^2$ around 0 resulting in $\breve{u}_{t+h}$; this will impose $H_0 : MSE_S^h - MSE_{IR}^h = 0$.

   (a) For $h = 1$: Re-sample $\breve{u}_{t+h}$'s through the 'wild bootstrap' (Liu (1988)):

   keep the $\breve{u}_{t+h}$'s fixed and $sign(\breve{u}_{t+h}) \begin{cases} + & \text{if } \iota < 0.5 \\ - & \text{otherwise} \end{cases}$ where $\iota \sim \mathrm{U}(0 \quad 1)$

   (b) For $h > 1$: Re-sample blocks of $\breve{u}_{t+h}$'s assuming, as in Politis and Romano (1994), time-varying block sizes based on a Geometric Distribution with average block size equal to bm quarters. For $h = 4 :\ bm = 2$ and $h = 8 :\ bm = 3$.

[14] Diebold and Mariano (1995), West (1996).

[15] Many structural models, including DSGE models, can be written as restricted VAR models, which are nested in many of our benchmark models. In very simplified terms, the *IR* projections can therefore been seen as the outcome of some weighted average of such a restricted VAR model and judgement, where the weights are practically unknown to the outsider.

Table 1: Annual Growth of GDP - Extract From the Real-time Database

		Release date						
		2003 Q1	2003 Q2	2003 Q3	...	2005 Q4	2006 Q1	2006 Q2
Reference date	2002 Q4	2.25	2.18	2.31	...	2.07	2.07	2.33
	2003 Q1		2.26	2.31	...	2.20	2.20	2.30
	2003 Q2			1.80	...	2.21	2.21	2.35
	⋮				⋱	⋮	⋮	⋮
	2005 Q3					1.69	1.89	1.75
	2005 Q4						1.82	1.82
	2006 Q1							2.35

2. Compute and save z_{MSE} test statistics.

3. Repeat 10,000 times and compute p-value based on the empirical distribution of the z_{MSE} statistics.

3 Results for UK Growth and Inflation

The forecast evaluation sample is 1997Q3 - 2006Q2, spanning the period when official Bank of England forecasts were published in its *IR*. We forecast annual growth rates at horizons $h = 1$, 4 and 8, as this is this how these official forecasts are presented. Our benchmark forecasts are the year-to-year changes in *IR* RPIX and GDP projections.

3.1 Data

Details of the data, of which only GDP are ever revised, are listed in Appendix A. Briefly, we have data on RPIX[16] annual inflation (the inflation target variable between 1992q4 and 2003Q4), annual oil price inflation, the proportional change in the effective exchange rate, the three-month T-bill interest rate, annual changes in narrow and broad money and annual growth in GDP. We also use data on annual growth in household consumption, real investment, export and imports in volume terms as well as the quarterly unit change in inventories to construct, via principal components, a real factor series. For this annual GDP growth and the five series that underly the real factor variable we use for each quarter the vintage (or snapshot) of original data at that time.

The structure of the real-time data base from which these data are taken is illustrated in Table 1, in this case for UK investment. The real-time data set for each variable of interest is an upper-triangular data matrix with publication dates ordered horizontally and reference dates vertically down. Each column represents a new release of data published by the statistical agency, and each release includes observations of differing maturities. Table 2 shows the maturity of the various observations.

[16]Retail price index excluding mortgage interest costs.

Table 2: Stylised Real-time Database - Maturity of Observations

		Release date					
		2003 Q1	2003 Q2	2003 Q3	...	2006 Q1	2006 Q2
Reference date	2002 Q4	1	2	3	...	13	14
	2003 Q1		1	2	...	12	13
	2003 Q2			1	...	11	12
	⋮				⋱	⋮	⋮
	2005 Q4					1	2
	2006 Q1						1

3.2 Forecasting analysis relative to *IR* GDP growth projections

We have vintages of annual GDP growth data available for each quarter starting in 1993Q1, and all of these vintages consistently have data available from 1962Q1 onwards. Hence the total sample length is 1962Q1 - 2006Q2. The first *IR* projections were published in August 1997 (shortly after Bank of England independence in May of that year), which implies that for the purpose of a comparison with the *IR* projections, the first vintage of GDP data to be used for estimating the univariate models should be the 1997Q2 vintage. Thus we employ 1997Q3 - 2006Q2 as the forecast evaluation period.

The results are reported in Table 3. Both the AR model and the forecast combination outperform the *IR* at all the forecast horizons considered by a substantial relative RMSE margin. Moreover, in both cases this performance is significant at the 10% level at each horizon. The monetary VARs also do well and the Macro VARs are the best performers at the 1-quarter horizon. Thus simple statistical models are able to perform as well or better than the official Bank of England forecasts.

3.3 Forecasting analysis relative to *IR* RPIX inflation projections

Unlike GDP, the RPIX price index is not revised. Hence, a forecasting analysis using the 'latest' vintage of RPIX data suffices to conduct a forecasting exercise in real time for this forecast variable. The RPIX price index is the retail price index (RPI) excluding mortgage interest rate payments, which was introduced in 1975 when mortgage interest rate payments became a component of the RPI price index. Therefore the pre-1975 RPI data would have been identical to RPIX data, which justifies combining data on post-1975 RPIX and pre-1975 RPI price indices into a single price index series. This results in an annual RPIX inflation series spanning 1964Q1 - 2006Q2. Over this period, inflation clearly exhibited large changes in mean. As described above, we accommodate these mean shifts by running a Bai-Perron structural break test on the inflation series, and using the identified break dates to demean the series, which is then used in the estimation of the models with data that start in 1964Q1. The inflation forecasts are thus remeaned inflation projections from the models. We follow the same procedure for each statistical RPIX inflation projection, except, of course, for the random walk model (2). The forecast evaluation relative to *IR* RPIX inflation projections are based on these remeaned forecasts from our statistical benchmarks. The results can be found in Table 4.

The results are in striking contrast to those for growth. In no case does a model outperform

Table 3: Relative RMSE forecast evaluation GDP growth: 1997Q3-2006Q2

h	AR	RW	STAR	MS	MacroVAR	BMacroVAR	MonVAR	BMonVAR	AVERAGE
1	0.81	0.86	0.73	0.73	0.72	0.70	0.87	0.82	0.71
	(-1.34)	(-0.95)	(-1.93)	(-1.97)	(-1.85)	(-2.02)	(-0.84)	(-1.22)	(-2.09)
	0.10	*0.18*	*0.02*	*0.03*	*0.04*	*0.03*	*0.21*	*0.12*	*0.02*
4	0.76	1.09	0.74	0.74	1.20	1.20	0.78	0.77	0.67
	(-1.38)	(0.40)	(-1.32)	(-1.37)	(0.64)	(0.63)	(-1.50)	(-1.51)	(-1.72)
	0.09	*0.61*	*0.09*	*0.08*	*0.71*	*0.71*	*0.08*	*0.07*	*0.05*
8	0.88	1.23	0.81	0.83	1.50	1.50	0.79	0.79	0.86
	(-1.87)	(1.08)	(-1.72)	(-2.71)	(2.58)	(2.50)	(-2.06)	(-2.19)	(-1.62)
	0.05	*0.85*	*0.03*	*0.00*	*0.99*	*0.99*	*0.01*	*0.01*	*0.04*

The table reports the ratio of the RMSE based on the out-of-sample forecast errors models relative to *IR* GDP growth projections conditioned on a constant interest rate path for a forecast horizon equal to h quarters. The forecast errors are computed over the subsample 1997Q3-2006Q2, whereas the model parameters used for the forecasts are estimated with a recursively growing sample that starts in 1962Q1. The distribution of the statistics are estimated using the bootstrap techniques described in the text. t-statistics against the hypothesis that the test statistic is below unity are reported in parentheses and p-values in italics. Statistics significantly less than unity are indicated in red.
AR indicates the autoregressive model (1); RW the random walk (2); STAR the smooth transition AR model (3); MS the Markov-switching AR model (4); MacroVAR the VAR model (5) employing macroeconomic variables estimated by OLS; BMacroVAR as for MacroVAR estimated using the Minnesota prior; MonVAR the VAR model (7) employing monetary variables estimated by OLS; BMonVAR as for MonVAR estimated using the Minnesota prior; AVERAGE for the unweighted average of forecasts from the eight individual models.

Table 4: Relative RMSE forecast evaluation RPIX inflation; 1997Q3-2006Q2

h	AR	RW	STAR	MS	MacroVAR	BMacroVAR	MonVAR	BMonVAR	AVERAGE
1	1.11	1.07	1.74	1.24	1.83	1.85	1.36	1.32	1.17
	(0.57)	(0.37)	(3.48)	(1.30)	(4.36)	(4.35)	(1.58)	(1.44)	(1.03)
	0.71	*0.64*	*1.00*	*0.89*	*1.00*	*1.00*	*0.93*	*0.92*	*0.84*
4	1.24	1.25	3.23	2.23	3.21	3.07	1.18	1.17	1.60
	(1.40)	(1.30)	(7.59)	(4.53)	(2.91)	(2.88)	(1.07)	(0.99)	(2.40)
	0.86	*0.87*	*1.00*	*1.00*	*1.00*	*1.00*	*0.83*	*0.81*	*0.98*
8	1.45	1.34	4.70	3.77	1.98	1.74	1.29	1.31	1.98
	(2.45)	(2.09)	(9.29)	(6.10)	(3.10)	(2.50)	(1.91)	(2.11)	(4.87)
	0.97	*0.98*	*1.00*	*1.00*	*1.00*	*0.99*	*0.97*	*0.98*	*1.00*

The table reports the ratio of the RMSE based on the out-of-sample forecast errors models relative to *IR* RPIX inflation projections for a forecast horizon equal to h quarters. In each case we *recursively* apply the Bai-Perron mean break date test and subsequently demean before estimating and remean when forecasting the RPIX inflation data based on that break date procedure. The forecast errors are computed over the subsample 1997Q3-2006Q2, whereas the model parameters used for the forecasts are estimated with a recursively growing sample that starts in 1964Q1. The distribution of the statistics are estimated using the bootstrap techniques described in the text. t-statistics against the hypothesis that the test statistic is below unity are reported in parentheses and p-values in italics. Statistics significantly less than unity are indicated in red.

Model mnemonics as in Table 3.

the *IR* forecasts. It is clear from the reported *p*-values that the *IR* forecasts are significantly better than the statistical forecasts in several cases. In general, *IR* RPIX inflation projections have a markedly better RMSE performance than the statistical models. Hence, in contrast to the finding in Stock and Watson (2007) that in the present era of the 'Great Moderation' univariate forecasts for inflation are hard to beat, *IR* inflation projections seem to be the unbeatable forecasting benchmark for UK inflation.

4 Revisiting the forecasting analysis for UK inflation

Why are the statistical inflation forecasts so poor, relative to the *IR*, especially given the relative strengths for output? The answer may be that judgement is especially important here. Judgements can be particularly useful if data are unstable.[17] In this case, there is a clear shift in the unconditional mean of inflation after a shift in monetary regime. Clements and Hendry argue forcefully (in, e.g., 1998a,b) that the main source of forecast error is structural change; Hendry (2000) argues that the dominant cause of these failures is the presence of deterministic shifts.

4.1 *Ex Post* analysis

The *Inflation Report* inflation forecasts tend to converge to the inflation target at medium term horizons. The relevant judgement here is that there is a strong prior on the long-run attractor, the unconditional mean. What makes this pertinent is that there is *a priori* information available to the Bank of England about shifts in this mean; namely, when the inflation-targeting regime changed in 1997Q3 after Bank of England independence. We can simulate this application of judgement in our statistical models by supposing that break dates are known. One way of implementing this is simply to undertake whole-sample Bai-Perron tests. Chart 1 shows that the whole sample test picks up a mean shift with uncanny precision at 1997Q3. By contrast, the *ex ante* recursively estimated shift is not detected until three years later - although this is precisely what would be expected with this test, which trims the sample by 15%.

Table 5 reveals that when this *ex post* demeaning is applied, the overwhelming superiority of the *IR* forecasts disappears. In several cases the statistical models outperform the *IR*'s, although only significantly in two cases (the 1-quarter horizon MS and the 8-quarter horizon AR). Naturally, this exercise is loaded against the *IR* as a crucial piece of full-sample information (regarding the unconditional mean) is used in the statistical models. Thus we do not in any way imply that statistical models outperform the *IR* forecasts over our sample. Nevertheless, this clearly suggests that the superiority of judgement in this case lies in the prior applied to the long-run attractor.

This can also be seen in Chart 2, which shows the *IR* forecasts of inflation and the real-time AR forecasts for the first nine quarters, for the first two years after Bank of England independence in 1997. The AR forecasts all tend towards a 9-quarter attractor around 2.8%. The *IR* forecasts are more dispersed but more accurate; averaged over all horizons, the relative RMSE is 1.34. The reason for this good performance appears to follow from the obvious tendency towards an increasingly lower attractor.

[17] For an interesting discussion of this issue, see Franses (2006).

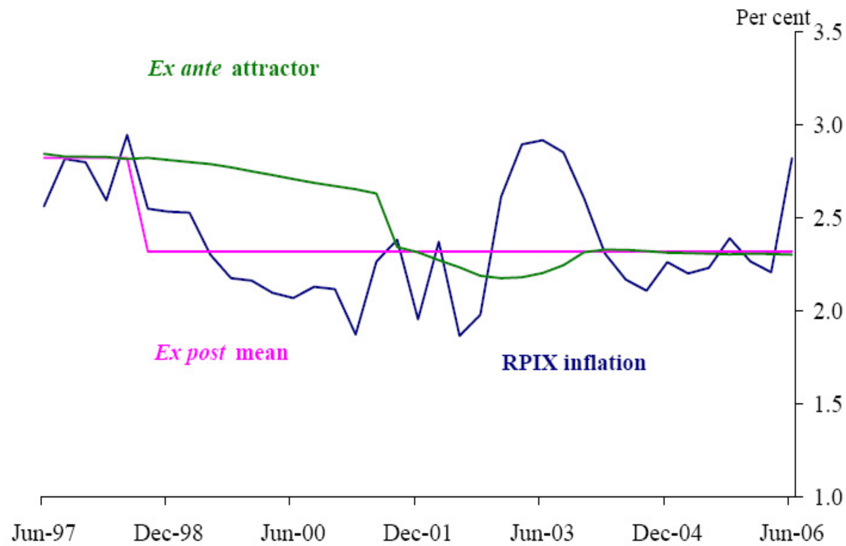
Table 5: Relative RMSE forecast evaluation RPIX inflation revisited using full-sample information on timing of breaks; 1997Q3-2006Q2

h	AR	RW	STAR	MS	MacroVAR	BMacroVAR	MonVAR	BMonVAR	AVERAGE
1	0.78	0.84	0.80	0.76	1.44	1.46	1.03	0.99	0.86
	(-0.92)	(-0.93)	(-1.18)	(-1.42)	(2.33)	(2.41)	(0.18)	(-0.05)	(-0.82)
	0.19	*0.18*	*0.13*	*0.09*	*0.99*	*0.99*	*0.56*	*0.48*	*0.21*
4	0.87	1.17	0.93	0.83	3.14	3.04	0.99	0.94	1.21
	(-0.84)	(0.82)	(-0.37)	(-0.97)	(3.43)	(3.42)	(-0.06)	(-0.36)	(1.02)
	0.25	*0.78*	*0.40*	*0.24*	*1.00*	*1.00*	*0.48*	*0.37*	*0.82*
8	0.78	1.07	0.80	0.76	1.39	1.23	0.95	0.92	0.83
	(-2.13)	(0.38)	(-1.41)	(-1.56)	(2.31)	(1.70)	(-1.33)	(-0.58)	(-1.39)
	0.08	*0.65*	*0.21*	*0.20*	*0.97*	*0.92*	*0.41*	*0.36*	*0.23*

The table reports the ratio of the RMSE based on the out-of-sample forecast errors models relative to *IR* RPIX inflation projections for a forecast horizon equal to h quarters. In each case we initially apply the Bai-Perron mean break date test using the *entire data-set* subsequently demean before estimating and remean when forecasting the RPIX inflation data based on that break date procedure. The forecast errors are computed over the subsample 1997Q3-2006Q2, whereas the model parameters used for the forecasts are estimated with a recursively growing sample that starts in 1964Q1. The distribution of the statistics are estimated using the bootstrap techniques described in the text. t-statistics against the hypothesis that the test statistic is below unity are reported in parentheses and p-values in italics. Statistics significantly less than unity are indicated in red.

Model mnemonics as in Table 3.

Figure 1: Recursive and whole-sample Bai-Perron tests for breaks in UK RPIPX inflation



A similar argument is made in Villani (2008). In that paper, a methodology is developed for setting priors on the long-run unconditional mean (the steady state) in Bayesian VAR models. Villani finds that using informative priors for a number of variables including growth and inflation improves forecast accuracy in a seven-variable BVAR for the Swedish economy.

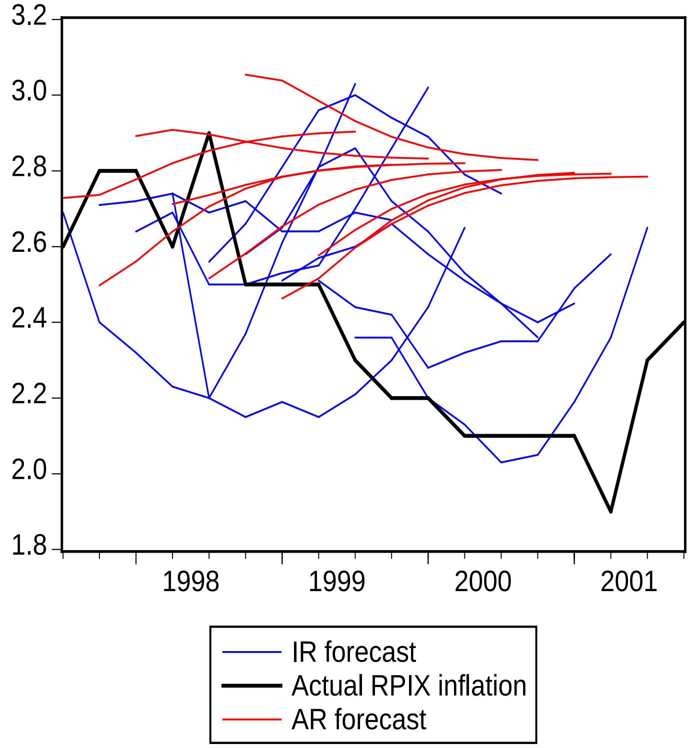
4.2 The real-time forecasting record for inflation revisited

As mentioned briefly in the previous subsection, the Bai and Perron (1998) structural break test has as a feature that one needs to trim 15% (at the quarterly frequency) at the start and the end of the sample to guarantee a properly sized break test. And as we have shown, applying this test recursively will substantially impede the models' ability to adapt to a inflation mean shift in the forecast evaluation period. So are there alternative techniques that allow our statistical models better to accommodate the breaks in the inflation mean in real time?

Often, as discussed in Section 2, a simple way to deal with instabilities in the data when generating forecasts, is to estimate a model on a rolling window of data of fixed size. We therefore repeat our forecasting analysis for inflation by updating the model parameters on rolling, fixed-sized sample of historical data, *without* demeaning and then remeaning the inflation series. We ran the models using window sizes of 7, 8, 9 and 10 years worth of quarterly data. The models generally performed best with a 7-year data window; the resulting forecasting analysis can be found in Table 6. Compared to the results in Table 4 some improvement *vis-à-vis IR* inflation projections is noticeable, in particular for the STAR model, the VAR models and the forecast combination. However, in most cases, except the STAR model, the RMSE ratios are still substantially larger than one.[18]

[18]For the STAR model the improvement in RMSE at the 2-year horizon relative to that for the *IR* is not significant.

Figure 2: *Inflation Report* and real-time AR forecasts of RPIX inflation



One issue with using rolling windows to update the model parameters is the choice of the window size. In the vicinity of a structural break one wants to use as little historical data as possible, as a too large sample of data will be dominated by pre-break information and bias the resulting forecasts. However, once one moves away from the break date, it may well be more appropriate to use a longer sample of historical data, as the risk of 'pre-break contamination' becomes less and the usage of more data will make the parameter estimates more efficient. Pesaran and Timmermann (2007), therefore, advocate basing forecasts on model estimations using rolling windows with a *time-varying* window size. One way to implement this is to construct a forecast combination for the same model estimated over different window sizes. Moreover, as argued by Hendry and Clements (2002), combining forecasts across different models can in itself insulate the forecast from the adverse effects of an unknown break.

We follow Clark and McCracken (2008) and combine the latter two forecast combination strategies: for each of the statistical models, except the random walk, we generate forecasts with model updating using rolling windows of sizes 7, 8, 9 and 10 years. This results, at each horizon, in 29 inflation forecasts, which we then combine based on an unweighted average. The performance of this combination strategy is reported in the last column of Table 6. At the

Table 6: Relative RMSE forecast evaluation RPIX inflation revisited using 7-year rolling windows; 1997Q3-2006Q2

h	AR	RW	STAR	MS	MacroVAR	BMacroVAR	MonVAR	BMonVAR	AVERAGE	$(\text{AVERAGE})^2$
1	1.09	1.07	1.02	1.12	1.32	1.25	1.05	1.01	1.04	1.01
	(0.53)	(0.37)	(0.11)	(0.66)	(1.62)	(1.43)	(0.27)	(0.04)	(0.25)	(0.05)
	0.69	*0.64*	*0.54*	*0.74*	*0.94*	*0.92*	*0.60*	*0.51*	*0.59*	*0.52*
4	1.18	1.25	1.07	1.63	1.40	1.35	1.22	1.20	1.20	1.23
	(0.93)	(1.30)	(0.37)	(2.94)	(1.74)	(1.55)	(1.11)	(1.11)	(1.19)	(1.38)
	0.80	*0.89*	*0.66*	*0.99*	*0.94*	*0.91*	*0.83*	*0.82*	*0.84*	*0.88*
8	1.10	1.34	0.99	2.44	1.23	1.24	1.33	1.25	1.27	1.62
	(0.74)	(2.09)	(-0.02)	(4.11)	(2.02)	(1.92)	(1.68)	(1.45)	(1.74)	(1.94)
	0.75	*0.97*	*0.52*	*1.00*	*0.96*	*0.94*	*0.93*	*0.90*	*0.94*	*0.95*

The table reports the ratio of the RMSE based on the out-of-sample forecast errors models relative to *IR* RPIX inflation projections for a forecast horizon equal to h quarters. The forecast errors are computed over the subsample 1997Q3-2006Q2, whereas the model parameters used for the forecasts are now estimated using a rolling window of seven years (i.e. twenty eight quarters) without demeaning and remeaning the inflation series.

The distribution of the statistics are estimated using the bootstrap techniques described in the text. t-statistics against the hypothesis that the test statistic is below unity are reported in parentheses and p-values in italics. Statistics significantly less than unity are indicated in red.

Model mnemonics as in Table 3 and in addition $(\text{AVERAGE})^2$ indicates the average across the individual models with model parameters updated using rolling windows of, respectively, 28, 32, 36 and 40 quarters of historical data.

1-quarter horizon the combination strategy yields an improvement and results in an inflation forecast performance that is on average similar to that of the *Inflation Report*. Beyond this short-term horizon, however, the statistical inflation forecasts remain inferior. Hence, the usage of rolling windows is not enough to make our statistical inflation forecasts as adaptable to a shift in the inflation mean as the *IR* projections. While updating our statistical models with rolling windows goes some way to enhance their real-time adaptability to breaks, the adjustment to breaks is probably remains too smooth, as it was in the case of the demeaning-remeaning strategy (see Figure 1). One obvious restriction on the rolling window scheme is that there is a lower bound on the window size. In our case 7 years (28 quarters) is already a small sample to estimate some of our statistical models.

In the previous subsection we highlighted the role of judgement regarding changes in the unconditional mean of inflation and how this was likely to have explained the superiority of *IR* inflation forecasts. Mechanical adjustments to our statistical models, such as model updating based on rolling windows or demeaning-remeaning the inflation series based on recursive break testing, appear not to be capable of replicating the effect of this judgement. These approaches simply retain too much information from the previous steady state, which thus results in biased forecasts. One possible work-around is to augment these mechanical approaches with a variable that can approximate this particular judgement about the inflation attractor that is deployed by policy makers at the Bank of England. Business surveys on inflation expectations are an obvious candidate for such a variable. Respondents to such surveys usually comprise of forecasters at forecasting institutions and (investment) banks, economists at universities, as well as practitioners such as trade union representatives and finance directors. Their outlook on inflation very often involves some kind of judgement, whether this is based on their own beliefs about the long-run, or a perception that a newly announced inflation target is a credible one.

In the UK, in contrast to the US, there are not many long running surveys of inflation expectations available. One useful survey is the Barclays BASIX Survey of Inflation Expectations, which has been run by Barclays Capital on a quarterly basis since 1986Q4. As part of this survey, finance directors, business and academic economists and trade union representatives, are asked about their expectation regarding the future annual change 'in retail prices' (so the price index is not specified in this survey). More specifically, respondents have to provide two inflation forecasts: the rate of inflation 12 months ahead and inflation for the next 12 months in 12 months time.[19] It is the latter, the one-year ahead inflation in a year's time, that we will focus on, as it is more likely that this forecast is driven by the respondents' judgement about the inflation attractor.

We now proceed by taking the AR model (1), the 'macro' VAR model (5) and the monetary VAR model (7) from our range of statistical benchmark models, and replacing them with variants in which we add the expected one-year ahead inflation rate in a year's time from the BASIX survey. Thus the AR model (1) is replaced with a bivariate VAR model

$$\begin{pmatrix} \Delta y_t \\ \Delta y_{basix,t} \end{pmatrix} = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} + \sum_{j=1}^{p} \Gamma_j \begin{pmatrix} \Delta y_{t-j} \\ \Delta y_{basix,t-j} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix} \tag{10}$$

where $\Delta y_{basix,t}$ is the BASIX survey expectation at t for the inflation rate between $t+4$ and $t+8$.[20] Model (10) is estimated both with OLS and in a Bayesian manner based on the Minnesota prior. Similarly, we replace the VAR models (5) and (7) with versions where we add $\Delta y_{basix,t}$ as, respectively, a sixth and a fourth variable in (5) and (7) respectively. Again, we estimate these expanded versions with both OLS and a Minnesota prior. In generating the inflation forecasts with these VAR models we opt for updating the model parameters with rolling windows, partly because of the relatively short span of the BASIX survey measures (1986Q4 - 2006Q2), and partly because to some extent it proved useful in our previous exercises. The resulting forecasting performance *vis-à-vis IR* inflation projections using 7-year windows of historical data is summarised in the first six columns of Table 7.

Relative to the comparable columns in Table 6, the inclusion of the inflation expectations measure does not add much in terms of improved forecasting power for these statistical models. Only at longer term horizons there seem to be an improvement in the performance of the monetary VAR models. Table 7 also reports the results for a forecast combination, which consists of an unweighted average of the forecasts from the aforementioned VAR models that include our inflation expectations measure together with those from the STAR model, the MS-AR model and the random walk. In comparison with the first forecast combination in Table 6,

[19]So they are providing a forecast for the expected change in prices from t to $t+4$ and the expected change in prices from $t+4$ to $t+8$, where t is in quarters.

[20]This measure is a weighted average of this expectation for each of the finance directors, business economists, academic economists, and trade union representatives, with weights determined by the relative size of each group.

this forecast combination improves, in particular at longer term horizons, but it still fails to match the *IR* forecasting performance.

Recall that earlier we explored how the *IR* inflation projections performed when compared to a forecasting combination that simultaneously combined forecasts across our statistical models, each updated along a sequence of updating windows. The last column in Table 7 is based on a similar forecast combination; we generate inflation forecasts based on the four earlier described VAR models that incorporate information from the BASIX inflation expectations survey (using both OLS and the Minnesota prior) and the STAR, MS-AR and random walk models. Each model is updated using 7-year, 8-year, 9-year and 10-years windows of data and thus delivers four inflation forecasts for each horizon (except for the parameter-free random walk model). The resulting 33 inflation forecasts for each horizon are combined with an unweighted average. Relative to the one in Table 6, the Clark and McCracken (2008) forecast combination strategy in Table 7 delivers a similar performance at the 1-quarter horizon, but does yield RMSE measures beyond that horizon that are lower than for the *IR* inflation projections. And although inference indicates that this improvement is not significant, the bootstrap p-values make it clear that this forecasting approach has on average a similar performance for UK inflation as the *Inflation Report*.

Hence, an approach that combines a range of statistical models that incorporates *ex ante* forward-looking information regarding the unconditional inflation mean as well as time-variation in the underlying information sets is able to mimic the performance of *IR* projections. Thus the analysis in this subsection confirms the conclusions of the *ex post* analysis in the previous subsection.

Table 7: Relative RMSE forecast evaluation RPIX inflation revisited using 7-year rolling windows and inflation expectations survey data; 1997Q3-2006Q2

h	AR+1	BAR+1	MacroVAR+1	BMacroVAR+1	MonVAR+1	BMonVAR+1	AVERAGE	(AVERAGE)^2
1	1.12	1.13	1.37	1.35	1.08	1.04	1.05	1.01
	(0.64)	(0.70)	(1.46)	(1.48)	(0.47)	(0.25)	(0.28)	(0.09)
	0.73	*0.75*	*0.92*	*0.92*	*0.67*	*0.59*	*0.60*	*0.53*
4	1.15	1.12	1.37	1.35	1.09	1.07	1.06	0.95
	(0.83)	(0.64)	(1.49)	(1.52)	(0.49)	(0.37)	(0.36)	(-0.25)
	0.77	*0.72*	*0.92*	*0.92*	*0.67*	*0.65*	*0.63*	*0.45*
8	1.25	1.20	1.23	1.31	1.17	1.21	1.08	0.99
	(1.22)	(1.08)	(1.56)	(2.31)	(0.78)	(0.97)	(0.58)	(-0.05)
	0.88	*0.85*	*0.93*	*0.98*	*0.81*	*0.86*	*0.71*	*0.50*

The table reports the ratio of the RMSE based on the out-of-sample forecast errors models relative to *IR* RPIX inflation projections for a forecast horizon equal to h quarters. The forecast errors are computed over the subsample 1997Q3-2006Q2, whereas the model parameters used for the forecasts are now estimated using a rolling window of seven years (i.e. twenty eight quarters) without demeaning and remeaning the inflation series.

The distribution of the statistics are estimated using the bootstrap techniques described in the text. t-statistics against the hypothesis that the test statistic is below unity are reported in parentheses and p-values in italics. Statistics significantly less than unity are indicated in red.

AR+1 indicates the bivariate VAR model (10) employing the current BASIX survey-based expectation for inflation between $t+4$ and $t+8$ estimated with OLS; BAR+1 as for AR+1 estimated using the Minnesota prior; MacroVAR is the VAR model (5) but now in addition also includes the BASIX measure of longer term inflation expectations and is estimated by OLS; BMacroVAR+1 as for MacroVAR+1 estimated using the Minnesota prior; MonVAR+1 the VAR model (7) employing monetary variables but now in addition also includes the BASIX measure of longer term inflation expectations and is estimated by OLS; BMonVAR+1 as for MonVAR estimated using the Minnesota prior; AVERAGE for the unweighted average of forecasts from the six aforementioned models plus STAR model (3), MS-AR model (4) and the random walk (2); (AVERAGE)^2 is the unweighted average of the forecasts from the nine models in AVERAGE updated using respectively windows with 28 quarters, 32 quarters, 36 quarters and 40 quarters of data.

5 Conclusions

Simple statistical models including autoregressive processes can provide forecasts that are comparable to or better than official post-independence Bank of England *IR* forecasts. This reflects the well-known difficulty of adding forecast value-added in a time of macroeconomic stability. By contrast, for inflation the *IR* forecasts are clearly dominant. This latter inflation result appears to be a reflection of the importance of judgement. Specifically, their forecasts incorporate the knowledge that they are targeting a particular value of inflation.

The fact that the real-time performance of the *IR* inflation forecasts are so good relative to the statistical benchmarks is particularly impressive given the period, characterised as it is by low volatility in macroeconomic outcomes (characteristic of the Great Moderation). As Stock and Watson (2007) observe, in such an environment it may be very difficult to beat simple forecasts; but official UK inflation forecasters appear able to do so. There is evidence that this superiority is due to the ability of policymakers to identify mean-shifts in the inflation process, as when the full-sample is used to identify such shifts, the statistical models tend to perform better than the official forecast, albeit with *ex-post* information about the unconditional mean. Thus one conclusion that emerges very strongly from this paper is that early mean-shift detection is crucial. Arguably, emphasis should be put on work in this area. One approach is to extend univariate methods of monitoring for structural change to panels of series, as in Groen, Kapetanios, and Price (2008a).

We also establish that statistical forecast performance can be brought closer to the official forecasts by the joint use of two methods that previous research has found to be useful in the presence of structural change. The first of these is the use of rolling estimation windows, reflecting the importance of time variation in parameters. The second is a simple average of models, often argued to be useful in combatting the effects on forecast failure of structural change. However, the official *Inflation Report* inflation forecast performance is extremely good. It is only when rolling estimation over a number of window lengths and forecast combination are jointly used that the statistical forecasts can beat the official forecasts at the four and eight quarter horizons, and then not by a significant amount.

A Data Appendix

- Inflation: Year-to-year change in the long-run RPIX price index series 1964Q1 - 2006Q2 (RPI pre-1976). Source: Bank of England.
- Oil price inflation: Year-to-year change in the Brent oil spot price 1964Q1 - 2006Q2. Source: Global Financial Data.
- Proportional change effective exchange rate: Year-to-year change in the narrow trade-weighted sterling index 1964Q1 - 2006Q2. Source: Bank for International Settlements.
- Short-term interest rate: Three-month T-bill interest rate 1964Q1 - 2006Q2. Source: Global Financial Data.
- Proportional change narrow money: Year-to-year change in the Notes & Coins series 1964Q1 - 2006Q2. Source: Bank of England.
- Proportional change broad money: Year-to-year change in the M4 series 1964Q1 - 2006Q2. Source: Bank of England.
- Output growth: Year-to-year change in GDP 1961Q1 - 2006Q2. For each quarter use the vintage (or snapshot) of original data at that time. Source: Real Time Data Base, Bank of England.
- Household consumption growth: Year-to-year change in household consumption in volume terms 1961Q1 - 2006Q2. For each quarter use the vintage (or snapshot) of original data at that time. Source: Real Time Data Base, Bank of England.
- Investment growth: Year-to-year change in whole economy investment in volume terms 1961Q1 - 2006Q2. For each quarter use the vintage (or snapshot) of original data at that time. Source: Real Time Data Base, Bank of England.
- Export growth: Year-to-year change in exports in volume terms 1961Q1 - 2006Q2. For each quarter use the vintage (or snapshot) of original data at that time. Source: Real Time Data Base, Bank of England.
- Import growth: Year-to-year change in imports in volume terms 1961Q1 - 2006Q2. For each quarter use the vintage (or snapshot) of original data at that time. Source: Real Time Data Base, Bank of England.
- Inventories change: Quarterly change in units of inventories 1962Q1 - 2006Q2. For each quarter use the vintage (or snapshot) of original data at that time. Source: Real Time Data Base, Bank of England.

References

Adolfson, M., M. K. Andersson, J. Lindé, M. Villani, and A. Vredin (2007): "Modern Forecasting Models in Action: improving Macroeconomic Analyses at Central Banks," *International Journal of Central Banking*, 3, 111–44.

Bai, J., and P. Perron (1998): "Testing for and Estimation of Multiple Structural Breaks," *Econometrica*, 66, 47–79.

Bank of England (1999): *Economic Models at the Bank of England*. London: Bank of England.

——— (2008): *Inflation Report*.

Clark, T., and M. W. McCracken (2005): "The Power of Tests of Predictive Ability in the Presence of Structural Breaks," *Journal of Econometrics*, 124, 1–31.

Clark, T. E., and M. W. McCracken (2008): "Averaging forecasts from VARs with uncertain instabilities," *Journal of Applied Econometrics*.

Clements, M. P. (2004): "Evaluating the Bank of England Density Forecasts of Inflation," *Economic Journal*, 114, 855–77.

Clements, M. P., and D. F. Hendry (1998a): *Forecasting economic time series*. Cambridge: CUP.

Clements, M. P., and D. F. Hendry (1998b): "Intercept corrections and structural change," *Journal of Applied Econometrics*, 11, 475–94.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–63.

Edge, R. M., M. T. Kiley, and J.-P. Laforte (2008): "A Comparison of Forecast Performance Between Federal Reserve Staff Forecasts, Simple Reduced-Form Models, and a DSGE Model," *Board of Governors of the Federal Reserve System.*

Elder, R., G. Kapetanios, T. Taylor, and T. Yates (2005): "Assessing the MPCs fan charts," *Bank of England Quarterly Bulletin*, pp. 326–48.

Faust, J., J. H. Rogers, and J. H. Wright (2005): "News and Noise in G-7 GDP Announcements," *Journal of Money, Credit, and Banking*, 37, 403–19.

Faust, J., and J. H. Wright (2006): "Efficient forecast tests for conditional policy forecasts," *Federal Reserve Board.*

Franses, P. H. (2006): "Formalizing judgemental adjustment of model-based forecasts," *EI Report 2006-19, Econometric Institute, Erasmus University.*

Galbraith, J., and G. Tkacz (2006): "How far can we forecast? Forecast content horizons for some important macroeconomic time series," *Department of Economics, McGill University, Departmental Working Papers.*

Gamber, E. N., and J. K. Smith (2007): "Has the Fed's Forecasting Advantage Eroded?," *RPF Working Paper No. 2007-002.*

Garratt, A., G. Koop, E. Mise, and S. P. Vahey (2007): "Real-time Prediction with UK Monetary Aggregates in the Presence of Model Uncertainty," *Birkbeck Working Papers in Economics and Finance 0714.*

Groen, J. J. J., G. Kapetanios, and S. Price (2008a): "Multivariate Methods for Monitoring Structural Change," *Bank of England Working Paper, forthcoming.*

Groen, J. J. J., G. Kapetanios, and S. Price (2008b): "Real time evaluation of *Inflation Report* and *Greenbook* forecasts for inflation and growth," *Bank of England Working Paper, forthcoming.*

Hamilton, J. (1989): "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.

Hendry, D. F. (2000): "On detectable and non-detectable structural change," *Structural Change and Economic Dynamics*, 11, 45–65.

Hendry, D. F., and M. P. Clements (2002): "Pooling of Forecasts," *Econometrics Journal*, 5, 1–26.

HMT (1999): *The New Monetary Framework.*

Jore, A. S., J. Mitchell, and S. P. Vahey (2008): "Combining forecast densities from VARs with uncertain instabilities," *Norges Bank Working Paper 2008/01.*

Kapetanios, G., V. Labhard, and S. Price (2008): "Forecast combination and the Bank of England's suite of statistical forecasting models," *Economic Modelling,* forthcoming.

Leitemo, K. (2003): "Targeting inflation by constant-interest-rate forecasts," *Journal of Money, Credit and Banking*, 35, 609–26.

Liu, R. Y. (1988): "Bootstrap Procedure under Some Non-i.i.d. Models," *Annals of Statistics*, 16, 1696–708.

Marcellino, M. (2006): "A Simple Benchmark for Forecasts of Growth and Inflation," *CEPR Discussion Paper no. 6012.*

Newey, W., and K. D. West (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Pagan, A. (2003): *Report on Modelling and Forecasting at the Bank of England.* London: Bank of England.

Patton, A. J., and A. Timmermann (2007): "Testing Forecast Optimality Under Unknown Loss," *Journal of the American Statistical Association*, 102, 1172–84.

Pesaran, M. H., and A. Timmermann (2007): "Selection Of Estimation Window In The Presence Of Breaks," *Journal of Econometrics*, 137, 134–61.

Politis, D. N., and J. P. Romano (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–13.

Sims, C. A. (2002): "The role of models and probabilities in the monetary policy process," *Brookings Papers on Economic Activity*, 2002, 1–40.

Stock, J., and M. Watson (2007): "Why Has Inflation Become Harder to Forecast?," *Journal of Money, Credit and Banking*, pp. 3–33.

Svensson, L. E. O. (1997): "Inflation Forecast Targeting: implementing and Monitoring Inflation Targets," *European Economic Review*, 41, 1111–46.

Timmermann, A. (2006): "Forecast Combinations," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann, pp. 135–196. North Holland, Amsterdam.

Villani, M. (2008): "Steady state priors for vector autoregressions," *Journal of Applied Econometrics, forthcoming.*

Wallis, K. F., and J. D. Whitley (1991): "Sources of Error in Forecasts and Expectations: UK Economic Models, 1984-88," *Journal of Forecasting*, 10, 231–253.

West, K. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–84.

Woodford, M. (2000): "Pitfalls of Forward-Looking Monetary Policy," *The American Economic Review*, 90, 100–4.