Working Paper no. 323

Forecast combination and the Bank of England's suite of statistical forecasting models

George Kapetanios, Vincent Labhard and Simon Price

May 2007

Bank of England

Forecast combination and the Bank of England's suite of statistical forecasting models

George Kapetanios [*]

Vincent Labhard[**]

and

Simon Price[†]

Working Paper no. 323

[*] Queen Mary, University of London.
Email: g.kapetanios@qmul.ac.uk

[**] EU Countries Division, European Central Bank.
Email: vincent.labhard@ecb.int

[†] Bank of England.
Email: simon.price@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England. We are grateful to colleagues who have contributed to the project, including Luca Benati (who estimated the QMA model described in Section 2.5), Giovanni Caggiano (Glasgow University) and Christoph Schleicher, and for advice and comments from Todd Clark, Phil Evans, Domenico Giannone, Adrian Pagan, Lucrezia Reichlin and Tony Yates as well as participants at a Bank workshop in August 2005. This paper was finalised on 9 February 2007.

The Bank of England's working paper series is externally refereed.

Information on the Bank's working paper series can be found at
www.bankofengland.co.uk/publications/workingpapers/index.htm.

Publications Group, Bank of England, Threadneedle Street, London, EC2R 8AH; telephone +44 (0)20 7601 4030, fax +44 (0)20 7601 3298, email mapublications@bankofengland.co.uk.

ISSN 1749-9135 (on-line)

Contents

Abstract

The Bank of England has constructed a 'suite of statistical forecasting models' (the 'Suite') providing judgement-free statistical forecasts of inflation and output growth as one of many inputs into the forecasting process, and to offer measures of relevant news in the data. The Suite combines a small number of forecasts generated using different sources of information and methodologies. The main combination methods employ weights that are equal or based on the Akaike information criterion (using likelihoods built from estimation errors). This paper sets a general context for this exercise, and describes some features of the Suite as it stood in May 2005. The forecasts are evaluated over the period of Bank independence (1997 Q2 to 2005 Q1) by a mean square error criterion. The forecast combinations generally lead to a reduction in forecast error, although over this period some of the benchmark models are hard to beat.

Key words: Forecasting, forecast combining.

JEL classification: C530.

Summary

Monetary policy at the Bank of England and at many other central banks is forward looking. So it is essential to be able to forecast accurately the future evolution of the economy. Consequently, the Bank of England maintains a large number of models, ranging from the purely statistical to data-free theoretical models, which we call upon to answer not only forecast but also other questions. As part of this general philosophy, the Bank has developed a range of purely statistical forecasting models (referred to hereafter as the 'Suite') which can be used to construct judgement-free statistical forecasts of inflation and output growth and which form one of many inputs into the Monetary Policy Committee's (MPC's) forecast process. This process culminates in the forecast fan charts reported in the *Inflation Report* which show a range of possible outcomes. These encapsulate the MPC's collective judgement of the prospects for inflation and growth, and are conditioned on specific assumptions, including interest and exchange rates and some exogenous variables, as well as on general views about the future.

We describe the Suite as it stood when it was first created in May 2005. Naturally, this is merely a snapshot, as the Suite continues to evolve; models or model combinations may be added or dropped, and the data continually change. On the evidence of the data and models that we examine in this paper, combinations of statistical forecasts generate good forecasts of the key macroeconomic variables, which can serve as judgement-free benchmarks to compare with the policymaker's projections. Moreover, changes in forecasts as new data arrive provide a summary measure of the relevant news in the data, giving a natural indicator of changing inflationary pressure over the horizons of policy interest.

We use two broad types of models. The first uses only univariate models (using only the variable to be forecast), which capture information solely in the forecast variable's history. Within this broad class we include linear and non-linear models of various types, including ones which may be more robust to some types of structural change. The second comprises multivariate models (including more than one variable), which capture a wider range of information. The data sets here vary in size, the largest using over 60 variables. Here too we include models which may be robust to structural change.

One important issue is the 'attractor', the value to which the forecast tends in the long run. If

models fit the data well they will tend to produce a long-run forecast close to the average of the past. In the case of inflation, the monetary regime has changed over the sample period: the recent average inflation rate is substantially lower than over the whole sample period, reflecting the success in meeting the inflation targets in place since 1992. We test for structural breaks in the mean, and then forecast the inflation rate less this mean.

Individual forecasts are then combined to produce a single forecast. Forecast combination has a good track-record of improving forecasts. The combinations we use are a simple average of all the forecasts in the Suite, where all individual forecasts have an equal weight, which has been shown to work well in practice; and our preferred method based on goodness-of-fit, which we have shown may have a superior forecast ability.

This exercise is essentially practical, and success is measured by improved forecasts. Data typically has some obvious short-run cyclical variation that has to be accounted for, but it is often possible to capture this with a simple autoregressive (AR) process (where the model is a combination of past values of the variable being forecast). So we assess the forecasts since Bank independence in 1997 Q2 to 2005 Q1 relative to a benchmark AR forecast. Over our sample the AR forecasts are hard to beat, especially for inflation, with most of the models doing worse for most periods, although two non-linear models do better at most horizons. However, the benchmark combinations can beat the AR at many horizons for both growth and inflation. Thus the Suite appears to be fit for its intended purpose, as a statistical benchmark forming one of many inputs into the MPC's forecast process.

1 Introduction

Monetary policy at the Bank of England and elsewhere is forward looking: policy is set with an eye to what we expect to happen in the future. So it is essential to be able to forecast the future evolution of the economy as accurately as possible. Consequently, the Bank of England maintains a large number of models, ranging from the purely statistical to data-free theoretical models, which we can call upon to answer not only forecast but also other questions in a number of contexts. The various uses to which these models are put is described in *Economic models at the Bank of England* .(1) In 2003, Pagan conducted a review(2) of modelling and forecasting processes at the Bank, and concluded that the existing range of models, although well suited to providing policy analysis, was less suited to providing alternative forecasts of inflation and GDP growth.(3) In particular, he recommended that more attention should be paid to models that exploited data and information which were not currently used in the Bank of England Quarterly Model (BEQM),(4) and that were based on modelling approaches not currently emphasised. Building on this suggestion, the Bank decided to develop a suite of judgement free, statistical models designed specifically to forecast, spanning the range of potential specifications, and to systematically evaluate their forecasting performance for the two key variables published in the *Inflation Report*; namely, inflation and GDP growth. The forecasts in this suite of statistical forecasting models (the 'Suite') would then be combined to provide a single best statistical forecast for inflation and a single best statistical forecast for GDP growth.

In terms of the Bank's forecasting and policy process, such a forecast is clearly only one of many inputs into the wider forecast process. This process ultimately delivers the 'fan charts' published in the *Inflation Report*, which show the whole distribution of the forecast, encapsulating the MPC's judgement of the prospects for inflation and growth at any moment, and conditioned on specific assumptions, including interest and exchange rates and some exogenous variables, as well as on general judgements about the future. On a narrow forecasting-performance front, there is evidence (eg, Wallis and Whitley (1981)) that judgements generally improve forecasts. But the crucial point is that the model is designed to aid the policy process. From the *Inflation Report*, 'the fan charts represent the MPC's best collective judgement about the most likely paths for inflation

(1) Bank of England (1999, 2000).
(2) Pagan (2003) and Bank of England (2003).
(3) See also the Bank's response, Bank of England (2003).
(4) Prior to the introduction of BEQM, described in Bank of England (2005), the model used was the Medium Term Macroeconometric Model described in Bank of England (1999, 2000).

and output, and the uncertainties surrounding those central projections', and from *Economic models at the Bank of England*, 'the projections in the fan charts [are not] mechanically produced by models: they reflect the judgments of the Committee'.

So while policymakers cannot use an automatically generated statistical forecasting model for their core projections, such a model may nevertheless have a role to play as a simple summary of the information in the data about the forecast variable of interest.(5) The problem is then that there are many competing models, using different data and methods. We need to find ways of filtering the disparate outputs in an informative way. The econometric literature discussed below helps here, by suggesting individual models that may have good forecast performance and by highlighting ways of combining forecasts.(6) There are both classical and Bayesian arguments that support the combining of different forecasts, and these are discussed in Section 4.

Regarding models, simple linear autoregressive (AR) models are often found to perform well in practice, where the variable of interest is forecast using only information from its own history. However, there may be evidence of non-linearity in particular series, and some simple univariate non-linear models may be improvements on the AR. With multivariate models, there is an obvious gain from the use of more information in extra variables, but a loss in precision comes with the corresponding rise in parameters. Linear vector autoregressions (VARs) are the benchmark multivariate model, but suffer from both over parameterisation and a limited use of information. Methods of circumventing the over parameterisation problem include the use of Bayesian techniques to reduce the parameter space. Problems of structural change can be addressed by state-space modelling of parameter variation or recursive estimation of VARs, or by using techniques that are robust to a class of structural change (as in the double-differenced approach). Incorporation of more information from large amounts of data can be achieved with various types of factor models. All the models we use are described in more detail in Section 2.

This exercise is essentially practical, and success is measured by improved forecasts. It has to be noted, however, that forecasting macroeconomic variables is hard: beyond a few quarters, it is difficult to beat the unconditional mean. Data typically has some obvious short-run cyclical

---

(5) See for example a similar tool recently developed at the Bank of Canada: Demers and Marci (2005).
(6) While we are primarily concerned with forecasts of the first moment, we are also interested in the distribution round the mean, just as the *Inflation Report* projections are presented in a fan chart. This can be generated by bootstrap methods.

variation which has to be accounted for, but (as already observed) it is often possible to capture this with a simple AR process. This is easy to understand. Stock and Watson (2005) point out that the well-documented move towards macroeconomic stability, sometimes referred to as the 'Great Stability', has made forecasting more easy 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. Stock and Watson (2005) examine this for US inflation.

> On the one hand, inflation ... has become much less volatile, so the root mean squared error of even naïve or relatively poor forecasts had declined since the mid-1980s. ... Inflation has become easier to forecast. On the other hand, the relative improvement of standard multivariate forecasting models, such as the backwards-looking Phillips curve, over a univariate benchmark has been smaller ... since the mid-1980s than before. ... It has become much more difficult for an inflation forecaster to provide value added beyond a univariate model.

The message is that a good test of a forecasting model is whether it can beat a simple regression. Thus we assess the forecast over the post-Bank of England independence period (1997 Q2 to 2005 Q1, the last sample point available at the time of this assessment) relative to a benchmark AR forecast, using relative root mean square errors. We also look at standard model diagnostics, using them as an indication of unmodelled information in the series.

In Section 2 we present the range of models we consider. Then we discuss the treatment of structural breaks in the inflation process in Section 3. In Section 4 we discuss some technical details of forecast combinations. In Section 5 we report the forecast evaluations. The final section concludes. The appendices outline some alternative forecast combinations explored in the project, list the mnemonics we use to refer to models and combinations, and define the data.

2 Models

In this section we briefly describe the models we examine. For convenience, we list the mnemonics in Table A. They were selected to span the range of models that are commonly used in forecasting, including some (non-linear specifications and standard Bayesian VARs) that were suggested by Pagan (2003). Part of our motivation for model choice was that we wanted to include standard models that are well understood and widely used. These models can be divided into different categories. One basic category is the benchmark: easy to estimate models that are known to do well in practice. It is well known that extremely simple models can forecast extremely well, and that is the case here. We also make distinctions between uni and multivariate, linear and non-linear, and monetary and non-monetary models. The first allows us to assess whether broad information sets have forecast information beyond that in the series itself. The second is important because it is often argued that non-linear models forecast well, perhaps at particular periods. And the third may be practically useful if it is thought monetary data are particularly important.

Table A: List of model mnemonics

UC	Unconditional mean (benchmark only)
RW	Random walk
AR	Autoregressive model
V	Vector autoregressive model
VM	Vector autoregressive model, monetary
DDV	Double-differenced vector autoregressive model
DDVM	Double-differenced vector autoregressive model, monetary
RV	Recursively estimated vector autoregressive model: small data set
RVGEN	Recursively estimated vector autoregressive model: large data set
MS	Markov-switching model
STAR	Smooth-transition autoregressive model
FW	Factor model
BVM	Bayesian vector autoregressive model
BVMM	Bayesian vector autoregressive model, monetary
QMA	Time-varying parameter model (inflation only)
QMAM	Time-varying parameter model, monetary (inflation only)

2.1 Benchmark forecasts

2.1.1 Unconditional mean (UC)

The first benchmark model in the Suite is that the variable of interest is equal to the unconditional mean over the recent past,

$$y_t = \alpha + \epsilon_t \tag{1}$$

where y_t is the variable of interest. This requires little discussion, although see Section 3 for a description of how we handle structural change. The forecast from this model is simply

$$E(y_{t+h|t}) = \alpha \tag{2}$$

where $E(y_{t+h|t}) = E(y_{t+h}|y_t)$ is the h-step ahead forecast. Clearly, all forecasts are the same for all horizons. Although this model is unlikely to do well at short horizons, in the long run it might be expected to be a powerful alternative to more complex models. The reasoning behind this is twofold. First, in a stationary world series are mean-reverting, and short-term dynamics become irrelevant at long horizons. Second, there are good theoretical reasons for holding strong priors that in long-run variables are determined by simple and invariant factors (underpinning stationarity). For growth, these are technical progress and population growth: for inflation, in the current UK institutional framework it is the inflation target.

2.1.2 Random walk (RW)

Another simple benchmark is the random walk or no-change model. Random walks are often found to forecast surprisingly well. They have also been argued to be robust to common forms of structural change (Clements and Hendry (2002)), namely, intercept shifts. The form of this model is given by

$$y_t = y_{t-1} + \epsilon_t \qquad \textbf{(3)}$$

where y_t is the variable of interest.

The h-step ahead forecast from this model is simply

$$E(y_{t+h|t}) = y_t \qquad \textbf{(4)}$$

where $E(y_{t+h|t}) = E(y_{t+h}|y_t, y_{t-1}, ...)$ is the h-step ahead forecast.

2.1.3 Autoregression (AR)

The main benchmark is the autoregressive model. Low-order autoregressive (AR) processes are close to the simplest forecasting tools available. In practice, univariate representations can often be captured by low-order systems. While not robust to structural change, they are robust to misspecification following incorrect choice of explanatory variables.

The form of the model is given by

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \epsilon_t \qquad \textbf{(5)}$$

where y_t is the variable of interest. The lag order p is chosen by information criteria, described in Section 3.

The forecast from this model is

$$E(y_{t+h|t}) = \alpha_0 + \alpha^t y_t \quad \textbf{(6)}$$

where $E(y_{t+h|t}) = E(y_{t+h}|y_t)$ is the h-step ahead forecast.

2.2 Forecasts from linear vector autoregressions

Linear vector autoregressions (VARs) are linear relationships between a small set of variables. Much of the economic interest in them relates to identification (structural VARs, SVARs), but unless the restrictions overidentify the model, which is not usually the case, for forecasting purposes identification is not important. Despite the small number of variables typically included the number of parameters is often large in relation to the sample size. For forecasting purposes, this makes it desirable to reduce the parameter space.

2.2.1 Basic VARs

The standard linear reduced-form VAR takes the form

$$\mathbf{y}_t = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t \quad \textbf{(7)}$$

where $\mathbf{y}_t = (y_{1,t}, \ldots, y_{m,t})'$ is the vector of variables in the model. Again, the lag order, p has to be selected and that is usually performed using information criteria.

It is possible to parameterise these VARs as vector-equilibrium correction models (VECMs), ie including cointegrating relationships. The advantage of this approach is that the forecast is pinned down by the long-run equilibrium; however, these models are not suitable to forecasting in the presence of structural change, because in this case the forecast is pinned down by a long-run relationship which may no longer be appropriate.

Overall, VARs are among the least accurate forecasting model classes available, largely due to overparameterisation. It is sometimes suggested that increased focus should therefore be placed on the forecasting performance of the chosen models over the recent past and weight should be placed on specification tests, and robustness of forecasting performance with respect to lag order and forecasting evaluation period. We do not examine this explicitly, although we do estimate VARs recursively.

The forecasts from the VAR model are computed recursively by

$$E(\mathbf{y}_{t+h|t}) = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t+h-i} \qquad \textbf{(8)}$$

where $\mathbf{y}_{t+h-i} = E(y_{t+h-i|t})$ if $t + h - i > t$ and $\mathbf{y}_{t+h-i}$ otherwise. The lag order, p, is selected using the AIC.

There are two VARs in the Suite of Models. The first is a standard VAR which includes an output growth measure (alternatively GDP or private sector output although we report only those forecasts using GDP in this paper), CPI inflation, oil price inflation, the return on the nominal effective sterling exchange rate and a three-month interest rate. The second VAR falls into the monetary category, and also includes two monetary aggregates (growth rates of M0 and M4).

2.2.2 Double-differenced VARs

Motivated by the difficulties structural breaks present for forecasting, Clements and Hendry (2002) advocate a double-differencing methodology. The rationale is that while in an environment where the DGP is constant a structural model, such as a VECM, dominates, if there are deterministic shifts a VECM will be thrown off in a profound way; it will try to equilibrate towards a long run that is no longer appropriate. By contrast, a double-differenced model will be robust to such shocks.

The form of this model is

$$\Delta \mathbf{y}_t = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \Delta \mathbf{y}_{t-i} + \mathbf{u}_t \qquad \textbf{(9)}$$

where $\Delta \mathbf{y}_t = (\Delta y_{1,t}, \ldots, \Delta y_{m,t})'$ is the vector of variables in the model. Note that the term double differencing applies because usually the variables $\mathbf{y}_t$ are differences of (logs of) some variables in levels such as output. The lag order, p, is selected using the AIC.

The forecasts from the DDVAR model are

$$E(\mathbf{y}_{t+h|t}) = \mathbf{y}_t + \sum_{j=1}^{h} E(\Delta \mathbf{y}_{t+j|t}) = \mathbf{y}_t + \sum_{j=1}^{h} \left[\mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \Delta \mathbf{y}_{t+h-i} \right] \qquad \textbf{(10)}$$

There are two double-differenced VARs in the Suite. They use the same variables as the standard VARs described above.

2.2.3 Forecasts from Bayesian VARs

Bayesian methods have proved useful in the estimation of straightforward reduced-form VARs. Classical VAR methodology suffers from overparameterisation. The noise in the data can obscure the signal: confidence intervals can be very wide. The Bayesian approach uses priors with a small number of parameters to extract the signal parsimoniously. Unlike much Bayesian analysis, the priors tend to be atheoretical; for example, that the processes are random walks (the Minnesota prior).

The BVAR models that are in the Suite are in the spirit of Doan, Litterman and Sims (1984) and Litterman (1986) based on the Minnesota prior, which fixes the prior mean of the VAR parameters to zero, and with a prior variance dependent on two hyperparameters, θ and λ. More specifically, let $\phi_{i,jk}$ denote the (j,k)th element of the ith lag VAR coefficient matrix. The Minnesota prior variance matrix for the matrices of VAR coefficients is diagonal with $var\{\phi_{i,jk}\} = (\lambda/i)^2$ if $j = k$ and $(\lambda\theta\sigma_j/i\sigma_k)^2$ if $j \neq k$, $j, k = 1, ..., m+n$, where σ_j^2 is the variance of the innovation error in the jth equation of the VAR model, for which the unrestricted VAR estimator is substituted, $j = 1, ..., m+n$. Possible values of θ are 0.2 or 0.8 while the value of λ can be taken to be between 0.1 to 0.9. While some practitioners regard BVAR models as ideally suited for forecasting, others point to the risks associated with choosing the wrong priors as a major drawback. It is this latter aspect of BVAR models which would need to be addressed in the forecasting analysis. For the purposes of our forecasts we have used the values $\lambda = 0.2$ and $\theta = 0.5$. In the above specification the posterior mean of the VAR coefficient matrices have a closed-form solution. These posterior means can be used in the place of standard coefficient matrix estimates and then forecasts can be produced in the standard way. The same variables are used in this case as for the standard VAR models.

2.2.4 Recursively estimated VARs

The problem of structural change has led Pesaran and Timmermann (2000) to suggest a recursive approach to estimation, where it is considered best to use only recent data. A data-dependent method for determining the amount of data to use has also been proposed. We use an approach where a VAR is recursively estimated with a selection from a large set of variables. This falls into the robust category. The form and forecast of this model is as in the case of the standard VAR, so

$$\mathbf{y}_t = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t \tag{11}$$

and

$$E(y_{t+h|t}) = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t+h-i} \tag{12}$$

The main difference is that the set of variables included (apart from the variable of interest) changes from period to period depending on a criterion. Possible criteria are an information criterion or some measure of out-of-sample performance. We consider two recursive VAR models in the Suite. The first starts with a set of eight variables which (apart from the variables of interest) are ten-year interest rates, one-year interest rates, growth in real households disposable income, the return on the nominal effective sterling exchange rate, growth in M0, central bank reserves and the MORI Economic Optimism index. Then, all VAR models of dimension 2 to 4 are evaluated out-of-sample over a window (which we currently choose to be made up of 32 periods). The VAR model with the lowest forecast root mean square error is chosen as the forecasting model for that period. The whole process is repeated every period. The second VAR model in the Suite considers a much larger set of variables. This set is the same as that used for the factor model described below. For this set of VAR models we choose the Akaike information criterion as an in-sample measure of fit of the equation relating to the variable of interest to be the evaluation criterion. The number of VAR models under investigation is too large for each one of them to be evaluated separately. We therefore, use the approach of Kapetanios (2005) and use simulated annealing to minimise the Akaike information criterion. The model that minimises the criterion according to the minimisation algorithm is used for forecasting in that period. The whole procedure is repeated every period.

2.3 Forecasts from univariate non-linear models

There is a large number of non-linear models that are used to model univariate processes, including models from the smooth-transition autoregressive (STAR), the threshold autoregressive (TAR) and the Markov-switching (MS) family. What these models have in common is that the univariate process switches between regimes: in the MS model the switch is based on a probability of transition; in the STAR model it is based on a threshold value. While there is evidence that non-linear models can fit better than linear specifications, it is unclear how adequate they are for forecasting, although Pagan (2003) suggested they may be useful at long horizons. Two possibilities, which are included in the Suite, are a Hamilton Markov-switching model and a model selected from a range of STAR possibilities. A range of test procedures and specification searches are required.

For one step ahead, the forecasts for the non-linear models are

$$y_{t+1|t} = E(y_{t+1}|y_t) = E(f(y_t)|y_t) \tag{13}$$

where $f(.)$ denotes the functional form of the conditional mean of the non-linear model. Except for one step ahead though, the forecasts for non-linear models differ from those for linear models. To illustrate, consider the two step ahead forecasts. These are

$$y_{t+2|t} = E(y_{t+2}|y_t) = E(f(y_{t+1})|y_t) = E(f(f(y_t) + \epsilon_{t+1})|y_t) \tag{14}$$

where ϵ_{t+1} is the error term of the non-linear model at time $t+1$. The problem is that

$$E[f(.)] \neq f(E[.]) \tag{15}$$

Consequently, to produce multi-step forecasts generally requires numerical integration to solve $y_{t+2|t} = E(f(y_{t+1})|y_t)$. It is also possible to use stochastic simulations instead, and this is the approach we adopt for the Suite.

2.3.1 Markov-switching model (MS)

The first non-linear model in the Suite is a Markov-switching model. This is given by

$$y_t = c_{s_t} + \sum_{i=1}^{p} \gamma_{i,s_t} y_{t-i} + \epsilon_t \tag{16}$$

where $\epsilon_t \sim iid(0, \sigma^2_{s_t})$ and 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 can be estimated via maximum likelihood using the filter suggested by Hamilton (1989). The MS model in the Suite has two regimes for both mean and volatility.

The forecast from this model is

$$y_{t+h|t} = \sum_{j=1}^{m} \left(c_j + \sum_{i=1}^{p} \gamma_{i,j} y_{t+h-i} \right) \Pr(s_{t+h} = j|t) \tag{17}$$

where $\Pr(s_{t+h} = j|t)$ denotes the probability that $s_{t+h} = j$ conditional on information available at time t. Note that for this model no stochastic simulations are needed.

2.3.2 Smooth-transition autoregression (STAR)

The second model is the smooth-transition autoregression (STAR). Given the policy mandate of holding inflation rate at 2% an ESTAR model appears appropriate for reasons that will become obvious below. The model is given by

$$y_t = c + \sum_{i=1}^{p} \delta_i y_{t-i} + [(1 - e^{-\theta(y_{t-d}-c_1)^2}) \sum_{i=1}^{p} \gamma_i y_{t-i}] + \epsilon_t \qquad \textbf{(18)}$$

where c, c_1, d, δ_i and γ_i are parameters to be estimated. c_1 can be interpreted as an attractor for process y_t . This model essentially implies that there is one autoregressive model for the forecast variable when y_{t-1} is close to c_1 and another when y_{t-1} is far away from c_1. If c_1 is viewed as the 2% target then this means that policy becomes more active when y_{t-1} is away from 2% than otherwise. Estimation of the model is by non-linear least squares. The parameters c_1 and θ are sometimes difficult to obtain, in which case a grid search may be used to obtain some ideas on their values before using non-linear least squares with the outcome of the grid search as initial conditions. The forecast does not have a closed-form solution for multi-step forecasts. We therefore use stochastic simulations to obtain multi-step forecasts.

2.4 Forecasts from factor models

Factor models aim to summarise large bodies of information in an essentially atheoretical way. Although standard, static, factor analysis is inappropriate for dynamic models, in large samples it is consistent. Stock and Watson (2002) suggested using standard models in a forecast context, and their method is simple to implement, works relatively well in practice and is straightforward to interpret. Stock and Watson (2002) apply this and other methods to US inflation. Recently, Forni, Hallin, Lippi and Reichlin (2000) have generalised the factor approach to exploit dynamic information using spectral methods, although one-sided methods are needed to generate forecasts. Another way of addressing the dynamic factor approach would be to use an unobserved component model specified in state space. But with large data sets, this is not feasible using maximum likelihood methods. An alternative is to use subspace algorithms, and this has been implemented by Kapetanios (2005). Recently, attention has been paid to models which augment VARs with factors, as another way of confronting the over parameterisation problem: the factor-augmented vector autoregressive (FAVAR) model of Bernanke, Boivin and Eliasz (2005) is an example. However, Boivin and Ng (2005) find the static model performs well in realistic situations. In the Suite we use static principal components, and a complex dynamic model.

Construction of forecasts from principal components is very straightforward, although care needs to be applied at the data-preparation stage. Outliers, seasonality and other data features can dominate and lead to undesirable outcomes. Seasonally adjusted data is preferred. Essentially, data need to be transformed to a stationary form, which normally requires either a log-transformation, first or second differencing, or first differencing the log-transform. Once this preliminary part of the analysis is completed, one applies principal components. Denote the matrix of observations by X. Then the factors may be obtained in one of two ways: either get the k first ordered (with respect to the eigenvalues) eigenvectors of XX' and use these as the factors; or get the k first ordered (with respect to the eigenvalues) eigenvectors of $X'X$, denote them by c and use Xc as factors.

Once the factors from the data set have been obtained, they are treated as any other variable and used together with the variable of interest in a VAR model to produce forecasts as usual. We use a single factor associated with the largest eigenvalue of $X'X$ in the forecast model. Thus the form as identical to the VAR considered above:

$$\mathbf{y}_t = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t \qquad \textbf{(19)}$$

where $\mathbf{y}_t = (y_{1,t}, \ldots, y_{m,t})'$ includes the factor and the variable of interest. The forecasts are given recursively by

$$E(\mathbf{y}_{t+h|t}) = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t+h-i} \qquad \textbf{(20)}$$

where $\mathbf{y}_{t+h-i} = E(y_{t+h-i|t})$ if $t + h - i > t$ and $\mathbf{y}_{t+h-i}$ otherwise. The lag order, p, is selected using the AIC.

2.5 Forecasts from time-varying coefficient models

Bayesian methods offer an alternative methodology for estimating time-varying and unobserved factor models, and there is an active research program pursuing this. One general model specification is offered by Cogley and Sargent (2002), who estimate a time-varying parameter Bayesian VAR with stochastic volatility.

The Bank has constructed a closely related model to forecast inflation, albeit using frequentist methods, the *Quarterly Monetary Assessment* (QMA) model. In the spirit of Stock and Watson

(1999),

$$\pi_{t+k} - \mu_{t+k} = \sum_{j=1}^{J} \left[\phi_{j,t}(\pi_{t-j} - \mu_{t-j}) + \alpha_{j,t} ra_{t-j} + \beta_{j,t} m_{t-j} + \gamma_{j,t} cr_{t-j} \right] + \epsilon_{t+k} \quad \textbf{(21)}$$

$$\epsilon_t | I_{t-1} \sim N(0, \sigma^2_{\epsilon,t}) \quad \textbf{(22)}$$

$$\sigma^2_{\epsilon,t} = \alpha_0 + \alpha_1 \epsilon^2_{t-1} + (1-\alpha_1)\sigma^2_{\epsilon,t-1} \quad \textbf{(23)}$$

$$\mu_t = \mu_{t-1} + u_t \quad \textbf{(24)}$$

where π_{t+k} is the rate of inflation prevailing between periods $t+k$-1 and $t+k$, μ_t is a drift term, designed to capture low-frequency shifts in the equilibrium level of inflation, which evolves as a random walk, ra_t, m_t and cr_t are the real activity, money growth and credit growth factors, ϵ_t is a reduced-form shock to the rate of inflation whose distribution at time t conditional on information at time t-1 is, according to **(22)**, normal with conditional variance $\sigma^2_{\epsilon,t}$, and $\sigma^2_{\epsilon,t}$, in turn, is postulated to evolve according to an IGARCH(1,1) specification. Finally, the time-varying loadings of the cyclical component of inflation onto itself, and of the three factors onto cyclical inflation — the $\phi_{j,t}$,$\alpha_{j,t}$,$\beta_{j,t}$ and $\gamma_{j,t}$ — are all postulated to evolve according to random walks.

Inflation projections at time $T+k$, where T is the sample length, are computed by first estimating **(21)** shifted back in time by k periods. In other words, we estimate

$$\pi_t = \mu_t + \sum_{j=1}^{J} [\phi_{j,t-k}(\pi_{t-k-j} - \mu_{t-k-j}) + \alpha_{j,t-k} ra_{t-k-j} +$$
$$+ \beta_{j,t-k} m_{t-k-j} + \gamma_{j,t-k} rer_{t-k-j}] + \epsilon_t \quad \textbf{(25)}$$

This gives us smoothed (two-sided) estimates $\mu_{T|T}$, ..., $\mu_{T-k-j|T}$, $\phi_{j,T-k|T}$, $\alpha_{j,T-k|T}$, $\beta_{j,T-k|T}$, and $\gamma_{j,T-k|T}$. Then, we simulate $\mu_{T|T}$, ..., $\gamma_{j,T-k|T}$ k periods ahead based on the MLE estimates of the standard deviations and of their innovation variances,(7) and we forecast π_{T+k} based on **(21)**. Such an approach presents the crucial advantage of eliminating the need to forecast future values of the two factors, and of the rate of change of the real exchange rate (rer_t), and of only requiring time-T observations on the variables of interest in order to form k-step ahead projections. On the other hand, a clear drawback is that the model has to be re-estimated for each forecasting horizon, thus markedly increasing the computational burden. In practice, however, the extent of variation of parameters' estimates across k is extremely small, so that in practice, at least for $1 \leq k \leq 8$, estimates for $k = 1$ can be regarded as a good approximation to estimates for $k > 1$.

Given that the ϵ_t's are not observed, we augment the state vector to include ϵ_t, and we replace, in

(7) Stochastic simulations take into account parameter uncertainty, by drawing from the distribution of the estimated parameters.

(23), ϵ_{t-1}^2 with its estimate conditional on information at time t-1, $E_{t-1}[\epsilon_{t-1}^2]$, thus obtaining the following approximate expression for $\sigma_{\epsilon,t}^2$:

$$\sigma_{\epsilon,t}^2 \simeq \alpha_0 + \alpha_1 \left[\epsilon_{t-1|t-1}^2 + E_{t-1}\left(\epsilon_{t-1}\text{-}\epsilon_{t-1|t-1}\right)^2\right] + \alpha_2 \sigma_{\epsilon,t-1}^2 \qquad \textbf{(26)}$$

Both $\epsilon_{t-1|t-1}$ and its estimated precision, $E_{t-1}\left(\epsilon_{t-1}\text{-}\epsilon_{t-1|t-1}\right)^2$, can then be easily recovered from the resulting approximated Kalman filter.

By defining the state vector, ξ_t, as ξ_t =[μ_t , μ_{t-1} , ..., μ_{t-j}, $\phi_{1,t}$, ..., $\phi_{J,t}$, $\alpha_{1,t}$, ..., $\alpha_{J,t}$, $\beta_{1,t}$, ..., $\beta_{J,t}$, $\gamma_{1,t}$, ..., $\gamma_{J,t}$, ϵ_t]$'$, the model can be cast into state-space form. Running the Kalman filter requires a time-varying matrix of loadings computed based on the linearised version of the model. We linearise the observation equation, **(25)**, by taking a first-order Taylor expansion around $\xi_{t|t-1}$, thus getting

$$\pi_t = \mu_t + \sum_{j=1}^{J}\left[\phi_{j,t}(\pi_{t-j} - \mu_{t-j|t-1}) + \phi_{j,t|t-1}(\mu_{t-j|t-1} - \mu_{t-j})\right] + \qquad \textbf{(27)}$$

$$+\sum_{j=1}^{J}\left[\alpha_{j,t} ra_{t-j} + \beta_{j,t} m_{t-j} + \gamma_{j,t} rer_{t-j}\right] + \epsilon_t$$

We compute k-step ahead forecasts based on the non-linear equation **(21)**.

Finally, the transition equation is given by $\xi_t = F\xi_{t-1} + \upsilon_t$ where

$$F = \left[\begin{array}{cc|c} 1 & 0_{1\times 20} & \\ \hline I_4 & 0_{4\times 17} & 0_{21\times 1} \\ \hline 0_{16\times 5} & I_{16} & \\ \hline \multicolumn{2}{c|}{0_{1\times 21}} & 0 \end{array}\right] \qquad \textbf{(28)}$$

with time-varying covariance matrix of the vector of innovations $Q_t \equiv E\left[\upsilon_t \upsilon_t'\right] \simeq$ diag([σ_u^2, 0, 0, ..., 0, $\sigma_{\alpha,1}^2$, ..., $\sigma_{\gamma,J}^2$, $\sigma_{\epsilon,t}^2$]$'$).

Following Stock and Watson (1999), we extract the real activity and money growth factors as the first principal components of matrices of HP-filtered indicators (the complete list of series is given in Appendix E). For the real activity factor we consider twelve series. We log and HP-filter. [8]

For the money growth factor we consider four series. We compute the quarter-to-quarter rate of growth of each series (quoted at an annual rate), HP-filter it, and extract the money growth factor as the first principal component of the matrix of indicators. The key reason for HP-filtering the

(8) In contrast to Stock and Watson, we perform two-sided HP-filtering.

rates of growth of monetary aggregates is that, within the present model, the low-frequency component of inflation is entirely captured by the inclusion of the random walk term μ_t in **(21)**. What the time-varying loadings (the $\beta_{j,t}$) capture is therefore the time-varying relationship between the *cyclical* components of inflation and money growth. Such an approach presents the key advantage of automatically controlling for shifts in the velocity of monetary aggregates, which otherwise should be properly modelled. Finally, for the credit growth factor we consider five series.

The model is estimated via maximum likelihood. The log-likelihood function is computed via the previously described approximated Kalman filter, and is maximised numerically with respect to unknown parameters by means of the MATLAB subroutine fminsearch.m, based on the Nelder-Mead simplex algorithm. In performing MLE estimation, we impose the following restrictions on the parameters. First, that all the standard deviations of the innovations to the random coefficients (the $\phi_{j,t}$, $\alpha_{j,t}$, $\beta_{j,t}$ and $\gamma_{j,t}$) be positive. By defining the standard deviation of the innovation to the generic random coefficient $x_{j,t}$ as $\sigma_{x,j,t}$ (with $x = \phi, \alpha, \beta$, or γ), such a restriction is implemented by reparameterising the log-likelihood function, setting $\sigma_{x,j,t} = exp(\hat{\sigma}_{x,j,t})$, and optimising with respect to the auxiliary parameters, the $\hat{\sigma}_{x,j,t}$. Second, we impose the restriction that the unconditional expectation of $\sigma^2_{\epsilon,t}$, equal to $\alpha_0(1 - \alpha_1 - \alpha_2)$, always be positive. Specifically, we set $\alpha_0 = exp(\hat{\alpha}_0)$; $\alpha_1 = exp(\hat{\alpha}_1)/[1 + exp(\hat{\alpha}_1)]$; and $\alpha_2 = \{exp(\hat{\alpha}_2)/[1 + exp(\hat{\alpha}_2)]\} \times \{1 - exp(\hat{\alpha}_1)/[1 + exp(\hat{\alpha}_1)]\}$, and we optimise with respect to $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$. Such a reparameterisation guarantees that $\alpha_0 > 0$; $0 < \alpha_1$; $\alpha_2 < 1$ and $\alpha_1 + \alpha_2 < 1$, which, in turn, guarantees that not only $\sigma^2_{\epsilon,t}$, but also its unconditional expectation, are always positive.

Standard deviations for the ML estimates of both the structural and the auxiliary parameters (marked ˆ) are computed by taking the square root of the diagonal elements of the Hessian.

3 Modelling issues

There are a number of modelling issues we need to address, foremost among which is the treatment of structural breaks.

3.1 Lag selection

In several models we need to select lag lengths. The lag order p is chosen by information criteria. This involves maximising the criterion

$$-ln(\hat{\sigma}_p^2) - IC_p \quad \textbf{(29)}$$

over p, where $\hat{\sigma}_p^2$ is the residual variance for a given p and IC_p is a penalty term depending on the criterion used and p. Specifically, we use the Akaike information criterion (AIC), an asymptotically unbiased measure of minus twice the model log likelihood:

$$AIC = -2\ln[pr(D_t|M_i, D_{t-1})] + 2\,p$$

where p is the number of parameters.

3.2 Structural breaks in the unconditional mean

There is considerable evidence that breaks have occurred in the evolution of UK inflation in the period we use for estimation; primarily, the mean level has declined. Such a decline is not compatible with a stationary process for inflation. However, there is little consensus about the exact nature of this break. It is important that we take into account possible breaks in inflation, and in this subsection we describe how we do this.

Since we wish to model the process of inflation using a wide variety of models it is important that any adjustment for breaks is as simple as possible. We assume that there have been a number of breaks in the mean of inflation and model these as follows

$$y_t = \sum_{i=1}^{k} a_i I(t_{i-1} < t < t_i) + \epsilon_t \quad \textbf{(30)}$$

where k is the number of breaks and $I(A)$ is the indicator function taking the value 1 if A holds and zero otherwise. Bai and Perron (1998) show that for a wide class of stationary processes ϵ_t, the number of breaks k, the timings of the breaks t_i and the coefficients a_i can be estimated consistently using a search algorithm suggested by Bai and Perron (1998). We use this approach to demean inflation and then apply our forecasting models to the residuals $\hat{\epsilon}_t$.

3.3 Imposing an attractor in the forecast

Modelling past breaks is one thing, but allowing for recent or future breaks another. Here we describe how we deal with such an assumed break in the forecast.

In the United Kingdom, the inflation target changed from 2.5% retail prices index (RPI) inflation to 2.0% consumer prices index (CPI) inflation in November 2003. It is reasonable to suggest that following this change in the target, there has been a break in the mean of any inflation measure. (9) It is therefore conceivable that one would want to impose this break in the forecast and set a specific attractor. In order to illustrate our approach, assume a VAR(p) model

$$\mathbf{y}_t = \mathbf{A}_0 + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{u}_t \tag{31}$$

where $\mathbf{y}_t$ is the vector of variables. We wish to impose that in the forecast there has been a mean shift in the forecast variable. We also make the assumption that this shift appears in the constant vector $\mathbf{A}_0$. We know the magnitude of the shift in the mean vector and we wish to translate this to $\mathbf{A}_0$.

Let m be the old mean vector. Let the new mean vector be $m^* = m + d$ where $d = (d_1, 0, ..., 0)'$, That is, we maintain the assumption that the variable with the mean shift is the first one. Let the new constant vector we look for be $\mathbf{A}_0^*$. Define $\mathbf{C} = I - \sum_{i=1}^{p} \mathbf{A}_i$. Then we have $m = \mathbf{C}^{-1}\mathbf{A}_0$ or $\mathbf{A}_0 = \mathbf{C}m$, and also $\mathbf{A}_0^* = Cm^*$. Then

$$\mathbf{A}_0^* = \mathbf{C}m^* = \mathbf{C}(m + d) = \mathbf{C}(\mathbf{C}^{-1}\mathbf{A}_0 + d) = \mathbf{A}_0 + \mathbf{C}d \tag{32}$$

This procedure can be carried out for all linear models. For the non-linear models, we have to adopt a slightly different strategy. For the Markov-switching models we need to apply this transformation to the autoregressive polynomials of both regimes. For the STAR model there is no closed-form solution but we can obtain the new constant that guarantees that the forecast tends towards the desired mean by numerical simulation.

4 Forecasting using model averaging

Given our set of models, is anything to be gained from combining or averaging them? Recent surveys of forecast combination from a classical perspective are to be found in Newbold and Harvey (2002) and Clements and Hendry (1998)). Clements and Hendry (2002) have a recent paper which may provide the state of the classical art in this area. If it were possible to identify the correctly specified model and the data generating process (DGP) is unchanging, then one might think that combining forecasts could only worsen performance. But the weight of evidence dating back to Bates and Granger (1969) and Newbold and Granger (1974) reveals that combinations of

(9) This does not in itself imply that the underlying monetary stance changed.

forecasts often outperform individual forecasts. Models may be incomplete, in different ways; they employ different information sets.[10] Forecasts might be biased, and biases can offset each other. Even if forecasts are not biased, they will differ in their variance. It is tempting to think that the best forecast is the minimum variance forecast, but this will not generally be optimal as there will be covariances between forecasts which should be taken into account. Thus combining misspecified models may improve the forecast. Note that despite this, combining forecasts will not in general deliver the optimal forecast, while combining information will. Clements and Hendry (1998) therefore argue that combining is opposed to their notion of an encompassing research strategy. Nevertheless, it may not be practicable to estimate the fully encompassing model, not least because the set of variables is vast. Thus we have a justification for combining forecasts. One could call this the classical misspecification case. An obvious and commonly used method of constructing a combination is to run a linear regression of the variable of interest on the forecasts as in Granger and Ramanathan (1984), chosen for a suitable horizon, and possibly with time-varying parameters. This allows us to exploit covariances between the forecasts.[11] But in our case we have an insufficiently large sample of forecast observations relative to the set of forecasts to implement this.

There is an alternative classical argument. Clements and Hendry (eg 1998) have forcefully argued that the main practical forecasting problem is parameter change, and specifically deterministic shifts. The implication is that forecasts should be combined because different models are affected differently by the break. Moreover, there is not necessarily a need to include only non-encompassed models. However, estimating optimal weights by regression may not be optimal because models that have not yet suffered a structural break are selected; but that is no guarantee they will not break down in the future. The problem then is that very poor forecasts may drag the combined performance down; they recommend using trimmed means or medians. Thus we have a classical structural break case.

(10) Another way of looking at forecast combination is that it combines many different sources of information. Lars Svensson described what central bankers do in practice in Svensson (2004). 'Large amounts of data about the state of the economy and the rest of the world ... are collected, processed, and analyzed before each major decision.' In an effort to assist in this task, econometricians began assembling large macroeconomic data sets and devising ways of forecasting with them: James Stock and Mark Watson (eg Stock and Watson (1999)) were in the vanguard of this campaign, pioneering the use of factor models which summarise large bodies of information in an essentially atheoretical way. Stock and Watson (1999) suggested using standard models in a forecast context, and their method is simple to implement, works relatively well in practice and is straightforward to interpret. Boivin and Ng (2005) find this method outperforms alternatives in the realistic case when the dynamic structure is unknown and the error process is complex.

(11) Diebold and Pauly (1987) advocate weighting by inverse discounted h-step ahead forecasts, so that the near future is given more weight but information is obtained about several horizons.

In the classical, 'frequentist', view of the world, there is a true model, albeit one that may be changing through time. The difficulty is how to estimate it. Given this model, there is uncertainty over the data, and over the estimated parameters. This uncertainty leads to an uncertain forecast, over which a probability distribution is defined.

But from a Bayesian perspective, probabilities measure the degree of belief that an agent has in an event. Parameters themselves are random variables with a probability distribution, rather than the estimated parameters being distributed around a given value. Bayes' law describes how new information can be used to update the conditional probability of a state occurring. In our context, the information comes from economic data; the state is a future value of a variable of interest. The belief in the forecast is conditional on the past data and our initial, prior beliefs. Moreover, there is uncertainty over models. We have a range of models, none of which is the 'true' model. Instead, we might characterise our views by means of probabilities associated with each model. The higher the probability, the stronger our belief in the model. Given these probabilities, we can construct the mean forecast, and the distribution around that mean.

Bayesian methods have been applied to related situations, with some success. We describe the methodology in Appendix A. For example, Jacobson and Karlsson (2004) hunt over a huge range of models, and evaluate forecasting performance. They then find that optimally weighted combination of the best ten outperforms any of the individual models. Their method is based on a very simple class of models, the ARDL, but is nevertheless extremely intensive computationally. Application requires the specification of likelihoods. These are easily defined in a regression context; other models will typically need numerical approximation methods. One pragmatic approach is to compute the weights as an average of equal weighting and the Granger-Ramanathan regression based method. This 'shrinkage' (towards equal weighting) method has been interpreted in a Bayesian framework by Diebold and Pauly (1990).

A key notion in Bayesian model averaging is the conditional probability of a model M_i being the true model, given the data, D_t, $pr(M_i|D_t)$. But there is a frequentist analogue, and a weight scheme based on this has been implied in a series of papers by Akaike and others.[12] Akaike's suggestion derives from the Akaike information criterion (AIC). AIC is an asymptotically unbiased measure of minus twice the log likelihood of a given model. It contains a term in the

(12) See, eg, Akaike (1978, 1979, 1981, 1983) and Bozdogan (1987) and expounded further by Burnham and Anderson (1998).

number of parameters in the model, which may be viewed as a penalty for overparameterisation. Akaike's original frequentist interpretation[13] relates to the classic mean-variance trade-off. In finite samples, when we add parameters there is a benefit (lower bias), but also a cost (increased variance). More technically, from an information theoretic point of view, AIC is an unbiased estimator of the Kullback and Leibler (1951) (KL) distance of a given model where the KL distance is given by

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x|\hat{\theta})}\right) dx \quad \textbf{(33)}$$

Here $f(x)$ is the unknown true model generating the data, $g(x|.)$ is the entertained model and $\hat{\theta}$ is the estimate of the parameter vector for $g(x|.)$. The KL distance is an influential concept in the model selection literature and forms the basis of the development of AIC. Within a given set of models, the difference of the AIC for two different models can be given a precise meaning. It is an estimate of the difference between the KL distance for the two models. Further, $exp\,(-1/2\Psi_i)$ is the relative likelihood of model i where $\Psi_i = AIC_i - \min_j AIC_j$ and AIC_i denotes the AIC of the ith model in $\mathcal{M}$. Thus, $exp\,(-1/2\Psi_i)$ can be thought of as the odds for the ith model to be the best KL distance model in $\mathcal{M}$. In other words this quantity can be viewed as the weight of evidence for model i to be the KL best model given that there is some model in $\mathcal{M}$ that is KL best as a representation of the available data. Note that there is no assumption made here about the true model belonging to $\mathcal{M}$. We are only considering the ranking of models in terms of KL distance. This may be viewed as a crucial difference from a Bayesian analysis, in which it is assumed that a model in $\mathcal{M}$ or a weighted average of the models in $\mathcal{M}$ is the true model, as in the Bayesian probabilistic view the models must span the complete set.

It is natural to normalise $exp\,(-1/2\Psi_i)$ so that

$$w_i = \frac{exp\,(-1/2\Psi_i)}{\sum_{i=1}^{N} exp\,(-1/2\Psi_i)} \quad \textbf{(34)}$$

where $\sum_i w_i = 1$. We refer to these as *AIC* weights.

We note w_i are not the relative frequencies with which given models would be picked up according to AIC as the best model given $\mathcal{M}$. Since the likelihood provides a superior measure of data based weight of evidence about parameter values compared to such relative frequencies (see, eg, Royall (1997)), it is reasonable to suggest that this superiority extends to evidence about a best model given $\mathcal{M}$. The w_i can be thought of as model probabilities under non-informative priors

(13) Akaike (1979) offers a Bayesian interpretation.

giving a parallel to Bayesian analysis. However, this analogy should not be taken literally as these model weights are firmly based on frequentist ideas and do not make explicit reference to prior probability distributions about either parameters or models. Also, the criterion is only one such which can form the basis of such weights.

This approach is explored in Kapetanios, Labhard and Price (2007) using Monte Carlo experiments and with UK inflation data, and is found to perform as well as or better than Bayesian averaging. In the light of this result, it currently forms our preferred weighting method. We found that alternative information criteria, specifically the Schwarz, also work well, and this could be used as an alternative. Equally, the Bayesian scheme works very similarly in practice. We have also implemented an alternative based on the predictive likelihood, described in Appendix B.(14)

5 Forecast evaluations

We now turn to the results. The forecast variables we consider are those published in the Bank's *Inflation Report*, namely GDP growth and CPI inflation. GDP growth is measured and forecast as a percentage change on a year earlier. CPI inflation is measured in the same way, but exhibits shifts in the mean over the sample period (is non-stationary). Consequently, we transform it into a stationary series first as described above. Once the forecast has been obtained, the mean is then added back on to convert the series back into the original units.

We report all the individual forecasts described above, listed above in Table A and briefly described in Appendix C for convenience. Although we do not include it in the Suite, we report the unconditional mean in the tables as an additional benchmark. The unconditional mean is a poor forecaster at short horizons, but at long horizons should in principle be hard to beat in the absence of structural change. However, forecast comparisons between the unconditional mean and the Suite forecasts for inflation use different information sets and are therefore invalid. This is particularly acute in the inflation case where the time-varying unconditional mean is estimated using the whole sample.

For completeness, we report various combinations, listed in Table B and briefly described in Appendix C. The main combinations of interest are the simple and the information theoretic

(14) The Bayesian and predictive likelihood methods were not implemented in the Suite at the time of writing.

average. Beyond these two, other combinations might let policymakers focus on different selections of the data and models. The 'Monetary' forecast combination is based on models which exploit the information content of monetary variables. It is obtained by giving equal weight to the forecasting models in the monetary category. The 'Robust' is based on the four models which are robust to structural change. It is obtained by giving equal weights to the models in the robust category. The 'Multivariate' simply averages the nine multivariate forecasts, and therefore emphasises news in a wide range of data. The 'Univariate' is similarly an average of the four univariate forecasts and emphasises news in the variable itself. In practice these combinations are almost invariably inferior to the main averages.

Table B: List of combination mnemonics

WEQ	Simple model average
WM	Average of models belonging to the monetary category
WMN	Average of models not belonging to the monetary category
WR	Average of models belonging to the robust category
WRN	Average of models not belonging to the robust category
WV	Average of multivariate models
WVN	Average of univariate models
WITMA	Information-theoretic model average

The evaluation period covers the entire inflation-targeting period from 1997 Q2 to 2005 Q2 (32 periods). The evaluation covers all 21 models currently in the Suite, and all forecast horizons from 1 to 12 steps ahead. In total, we therefore evaluate 672 projections (32 projections for 21 models) and 7,272 individual forecasts (for example, 32 for 1 step ahead, 31 for 2 steps ahead, 22 for 11 steps ahead and 21 for 12 steps ahead).[15]

Regarding forecast assessments, there are broad two categories of tests: of forecast accuracy (ie the distance between the forecasts and the outturns); and of forecast rationality (ie whether forecast errors are zero on average and whether forecasts can be improved using additional information). We concentrate on the former. There are several formulae which could be used, including the sum of forecast errors, the mean forecast error, the sum of absolute errors, the mean

(15) There is clearly a trade-off between sample estimation and evaluation. In an interesting paper, Clark and McCracken (2004) report that the accuracy of forecasts from a given model can be improved by combining forecasts from a model estimated with the full sample of available data with forecasts from a model estimated with a rolling sample of data. Clark and McCracken (2006) argue that out-of-sample evidence is weak because, even when the models are stable over time, forecast performance measures are less powerful than in-sample tests, and the situation is worsened by structural breaks. As always, however, we are limited by availablity of data, especially as we are keen to avoid forecasts over a period with a known structural break.

absolute error, the sum of squared errors, Theil's U-statistic or the root mean squared error (RMSE). We use a relative RMSE statistic (RRMSE). This statistic is computed as the square root of the sum of squared forecast errors, relative to the same expression for the benchmark forecast, which we choose to be the autoregressive (AR) forecast; for the AR forecast therefore, the resulting number is equal to 1. Because it is expressed in relative terms, the RRMSE measure has the advantage of being comparable across forecasts. It also has the advantage of being robust to positive and negative forecast errors and of large forecast errors being penalised, due to the quadratic form.

5.1 Growth: individual forecasts

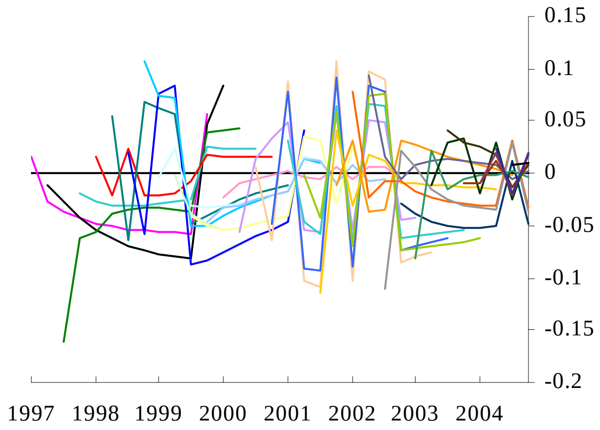
Table C reports relative performance for the models in the Suite. There is a wide range of performance. The unconditional mean should do reasonably well at long horizons, and as Table C shows this is the case, although it is worse than the AR at all horizons except at 12 quarters. It is very poor at horizons 1 and 2. As expected, the AR benchmark is generally hard to beat. However, there are models which provide better forecasts for some horizons, notably the STAR and Markov-switching models at all horizons for GDP growth. Other models that outperform the AR at least for one horizon are the monetary VAR, the RV model for the first quarter and the BVMM model for three quarters ahead. And several models outperform the simple random walk forecast. It should be noted that even models that outperform the AR on average do not necessarily do so for all horizons. This can be seen in Chart 1, which plots the absolute difference in forecast errors for all forecast evaluation periods for the well-performing Markov-switching model, relative to the AR benchmark.[(16)] Another characteristic is that there is persistence in forecast performance: models that perform well at any point are also likely to do so in the next period.[(17)]

5.2 Growth: forecast combinations

While it is hard to beat the AR when forecasting growth with individual models, it is quite possible to do so with the forecast combinations (see Table D). And even where the forecast combinations are not outperforming the AR, they have RRMSEs which are much closer to one than for most of the individual forecasts, indicating that combining forecasts does help forecast performance.

(16) A negative value means that the MS model is performing well.
(17) The chart also makes it clear that the good performance of the Markov-switching model is not due to an ability to forecast a small number of outlying observations. This is also true for the STAR forecasts.

Chart 1: Absolute difference in forecast errors for GDP growth (model MS versus model AR)



However, the MS and STAR models are generally preferred to the combinations. For GDP growth, our preferred information criteria based method (WITMA) beats the AR at all horizons (peaking at a 7% RMSE reduction at twelve quarters). On the whole, the restricted combinations perform worse than the simple average, although the combination of univariate models does relatively well.

5.3 Inflation: individual forecasts

Turning to inflation as the forecast variable, we also find a similar picture in terms of the best-performing models (see Table E) but overall performance is poorer. Most models outperform the random walk, but the AR is hard to beat. The models which outperform the AR at least once include the standard VAR and the two recursive VARs. The best forecasting models for inflation are the MS model up to 11 quarters ahead, STAR model for up to 9 quarters ahead and the factor model forecast for up to 6 quarters ahead. An illustration of the pattern of forecast errors is again given for the MS case in Chart 2. The QMA model outperforms the RW in all but one horizon, but only outperforms the AR at one horizon. It also beats some of the other models at some horizons (notably at 9 quarters), but at the same time it performs poorly at other horizons (notably 3-7 quarters ahead as well as 12 quarters ahead). The version emphasising monetary variables performs similarly, although it does very well at horizons 9 and 10. For horizons greater than 4 periods the unconditional mean does very well indeed. However, as observed above, the (time-varying) unconditional mean (which is not included in the Suite) has an advantage over the

Table C: Root mean square error relative to AR benchmark for individual models (GDP growth)

h	UC	RW	AR	V	VM	DDV	DDVM	RV	RVGEN	MS	STAR	FW	BVM	BVMM
1	2.06	1.03	1.00	1.26	1.54	1.57	1.60	1.20	1.33	0.97	0.95	1.08	1.16	1.18
2	1.49	1.05	1.00	1.23	1.23	1.55	1.70	1.15	1.17	0.94	0.92	1.06	1.16	1.03
3	1.23	1.11	1.00	1.27	0.94	1.69	1.80	1.23	1.37	0.92	0.90	1.06	1.18	0.95
4	1.12	1.16	1.00	1.26	1.07	1.71	1.89	1.32	1.62	0.90	0.89	1.06	1.21	1.01
5	1.11	1.19	1.00	1.36	1.14	1.79	1.94	1.48	1.53	0.89	0.88	1.01	1.32	1.08
6	1.07	1.24	1.00	1.46	1.20	1.94	2.02	1.51	1.56	0.88	0.90	1.05	1.41	1.15
7	1.04	1.27	1.00	1.55	1.20	2.01	2.10	1.45	1.47	0.88	0.90	1.01	1.47	1.14
8	1.02	1.32	1.00	1.63	1.15	2.16	2.31	1.41	1.33	0.88	0.91	1.02	1.53	1.10
9	1.01	1.33	1.00	1.64	1.02	2.28	2.48	1.30	1.35	0.88	0.91	1.00	1.52	1.00
10	1.01	1.30	1.00	1.66	0.97	2.29	2.61	1.24	1.34	0.89	0.91	0.98	1.52	0.95
11	1.00	1.26	1.00	1.59	1.04	2.23	2.51	1.15	1.35	0.90	0.90	0.95	1.44	1.02
12	0.98	1.24	1.00	1.43	1.17	2.24	2.47	1.03	1.36	0.90	0.89	0.92	1.30	1.12

mnemonics as in Table A
h = forecast horizon
Values below 1.0 highlighted in red

other models as it has been estimated using data to the end of the sample and comparisons are invalid. An unconditional mean defined over the whole period performs very poorly.

5.4 Inflation: forecast combinations

We find that forecast combining helps to improve forecast performance for inflation. While the individual forecasts were generally able to outperform the benchmark AR at only a few horizons, the information theoretic forecast combination systematically outperforms the AR at all horizons, in several cases by a large margin. Moreover, WITMA beats or matches the MS and STAR forecasts at almost all horizons. The restricted combinations are largely dominated by the WITMA.

5.5 Diagnostics for the forecasting equations

A desirable property of a forecasting model is that the residuals are well-behaved, primarily because diagnostic failure indicates the presence of unmodelled information. We therefore present a number of standard diagnostics for the residuals of the forecasting equation in the various forecasting models. These diagnostics are based on standard tests for normality, ARCH effects, serial correlation, and non-linearity. In the case of ARCH effects and serial correlation, the tests consider orders 1 and 4. We also present the error variance.

Table D: Root mean square error relative to AR benchmark for combined models (GDP growth)

h	AR	WEQ	WM	WMN	WR	WRN	WV	WVN	WITMA
1	1.00	0.92	1.19	0.85	0.98	1.02	0.94	0.97	0.97
2	1.00	0.92	1.10	0.87	1.04	0.96	0.94	0.96	0.95
3	1.00	0.93	1.05	0.90	1.16	0.92	0.96	0.96	0.95
4	1.00	0.96	1.09	0.93	1.22	0.93	1.00	0.96	0.95
5	1.00	0.99	1.07	0.98	1.31	0.94	1.04	0.96	0.95
6	1.00	1.04	1.12	1.03	1.40	0.98	1.11	0.97	0.96
7	1.00	1.06	1.12	1.06	1.44	0.99	1.13	0.97	0.97
8	1.00	1.08	1.14	1.07	1.51	1.00	1.15	0.98	0.98
9	1.00	1.07	1.14	1.06	1.54	0.97	1.14	0.98	0.98
10	1.00	1.05	1.14	1.04	1.52	0.96	1.11	0.96	0.96
11	1.00	1.01	1.10	1.00	1.46	0.96	1.07	0.95	0.95
12	1.00	0.99	1.12	0.96	1.38	0.97	1.04	0.93	0.93

mnemonics as in Table B
h = forecast horizon
Values below 1.0 highlighted in red

Chart 2: Absolute difference in forecast errors for CPI inflation (model MS versus AR)

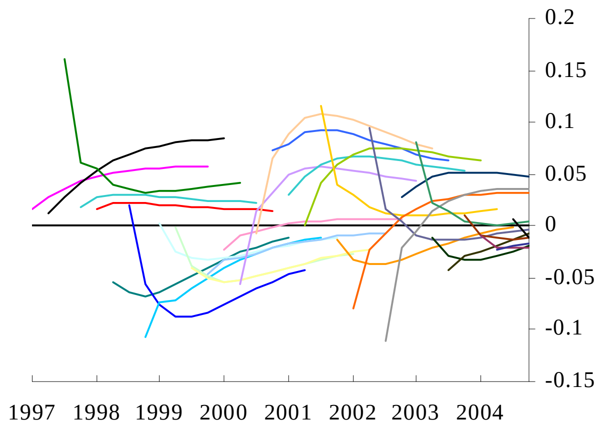


Table E: Root mean square error relative to AR benchmark for individual models (inflation, using GDP as predictor)

h	UC	RW	AR	V	VM	DDV	DDVM	RV	RVGEN	MS	STAR	FW	BVM	BVMM	QMA	QMAM
1	1.20	1.11	1.00	1.18	1.73	1.47	2.15	1.21	1.57	0.99	0.98	0.97	1.03	1.15	1.07	1.09
2	1.10	1.18	1.00	1.19	1.88	1.14	1.45	1.21	1.38	0.97	0.96	0.98	1.02	1.05	1.04	1.06
3	1.06	1.17	1.00	1.22	1.60	1.40	1.98	1.16	1.69	0.94	0.93	0.97	1.01	1.03	1.22	1.26
4	0.93	1.34	1.00	1.22	1.49	1.45	1.64	1.04	1.26	0.92	0.92	0.99	1.05	1.07	1.29	1.37
5	0.90	1.46	1.00	1.19	1.36	1.42	1.78	1.03	1.63	0.92	0.93	0.97	1.03	1.06	1.25	1.27
6	0.92	1.41	1.00	1.24	1.42	1.64	1.92	0.94	1.27	0.92	0.94	0.98	1.10	1.09	1.27	1.37
7	0.87	1.62	1.00	1.27	1.35	1.85	2.35	1.02	1.10	0.91	0.94	1.00	1.13	1.12	1.28	1.39
8	0.88	1.72	1.00	1.27	1.39	1.91	1.91	0.97	0.93	0.92	0.96	1.01	1.13	1.11	1.11	1.11
9	0.91	1.75	1.00	1.25	1.31	2.00	2.26	0.96	1.10	0.95	0.98	1.03	1.15	1.15	0.97	0.85
10	0.91	1.94	1.00	1.20	1.19	2.22	2.86	0.99	0.86	0.95	1.02	1.03	1.13	1.14	1.03	0.83
11	0.91	2.02	1.00	1.03	1.30	2.36	2.73	1.03	1.08	0.98	1.05	1.00	1.03	1.04	1.19	1.15
12	0.94	1.91	1.00	0.93	1.05	2.23	2.41	1.03	0.99	1.02	1.08	1.01	1.02	1.01	1.42	1.47

mnemonics as in Table A
h = forecast horizon
Values below 1.0 highlighted in red

Table F: Root mean square error relative to AR benchmark for combined models (inflation, using GDP as predictor)

h	AR	WEQ	WM	WMN	WR	WRN	WV	WVN	WITMA
1	1.00	0.89	1.21	0.84	1.08	1.00	0.92	1.00	0.90
2	1.00	0.96	1.11	0.94	0.97	1.04	0.98	1.00	0.93
3	1.00	1.03	1.21	0.99	1.26	1.02	1.08	0.97	0.96
4	1.00	0.99	1.15	0.96	1.06	1.04	1.02	1.00	0.92
5	1.00	0.97	1.12	0.93	1.08	1.02	0.98	1.03	0.90
6	1.00	0.97	1.09	0.95	1.16	1.02	1.00	1.00	0.89
7	1.00	1.05	1.21	0.99	1.32	1.06	1.08	1.05	0.90
8	1.00	1.06	1.21	1.01	1.18	1.08	1.08	1.09	0.91
9	1.00	1.09	1.24	1.04	1.37	1.09	1.11	1.10	0.93
10	1.00	1.17	1.39	1.08	1.56	1.10	1.18	1.17	0.95
11	1.00	1.16	1.40	1.07	1.46	1.09	1.15	1.21	0.95
12	1.00	1.05	1.15	1.02	1.41	1.04	1.02	1.17	0.97

mnemonics as in Table B
h = forecast horizon
Values below 1.0 highlighted in red

Table G: Diagnostic tests (GDP growth)

	RW	AR	V	VM	DDV	DDVM	RV	RVGEN	MS	STAR	FW	BVM	BVMM
Va	0.20	0.18	0.19	0.19	0.21	0.20	0.12	0.12	0.18	0.17	0.18	0.19	0.18
N	0.99	0.50	0.40	0.69	0.60	0.25	0.70	0.43	0.33	0.49	0.83	0.35	0.68
SC (1)	0.76	0.44	0.86	0.10	0.23	0.44	0.26	0.30	0.37	0.48	0.94	0.34	0.22
SC (4)	0.27	0.41	0.75	0.56	0.26	0.54	0.92	1.00	0.47	0.34	0.59	0.94	0.59
ARCH (1)	0.67	0.60	0.34	0.87	0.96	0.64	0.54	0.59	0.57	0.57	0.42	0.94	0.99
ARCH (4)	1.00	0.97	1.00	1.00	0.43	0.30	0.71	0.98	0.98	1.00	1.00	0.87	0.98
NONL	0.67	0.77	0.07	0.98	0.23	0.19	0.49	0.88	0.76	0.51	0.38	0.75	0.95

mnemonics as in Table B
Va = variance
N = normality (p-value)
SC (1) = serial correlation 1 lag (p-value)
SC (4) = serial correlation 4 lags (p-value)
ARCH (1) = autoregressive heteroscedasticity 1 lag (p-value)
ARCH (4) = autoregressive heteroscedasticity 4 lags (p-value)
NONL = reset test for non-linearity (p-value)
p-values below 0.10 highlighted in red

Table H: Diagnostic tests (inflation)

	RW	AR	V	VM	DDV	DDVM	RV	RVGEN	MS	STAR	FW	BVM	BVMM
Va	0.11	0.09	0.09	0.09	0.10	0.11	0.13	0.08	0.09	0.09	0.08	0.09	0.09
N	0.79	0.95	0.92	0.89	0.22	0.46	0.42	0.92	0.93	0.92	0.92	0.92	0.91
SC (1)	0.07	0.68	0.36	0.27	0.69	0.79	0.85	0.60	0.20	0.24	0.80	0.45	0.36
SC (4)	0.87	0.83	0.73	0.90	0.98	0.97	0.98	0.68	0.90	0.92	0.76	0.76	0.90
ARCH (1)	0.33	0.04	1.00	0.78	0.56	0.84	0.43	0.18	0.08	0.07	0.18	0.89	0.85
ARCH (4)	0.98	0.82	0.74	0.97	0.33	0.95	0.99	0.61	0.82	0.77	0.99	0.67	0.93
NONL	0.15	0.04	0.49	0.22	0.12	0.19	0.38	0.11	0.23	0.18	0.51	0.51	0.23

mnemonics as in Table B
Va = variance
N = normality (p-value)
SC (1) = serial correlation 1 lag (p-value)
SC (4) = serial correlation 4 lags (p-value)
ARCH (1) = autoregressive heteroscedasticity 1 lag (p-value)
ARCH (4) = autoregressive heteroscedasticity 4 lags (p-value)
NONL = reset test for non-linearity (p-value)

In all cases, the null hypothesis is that the residuals are well-behaved: that is, the residuals are normal, there is no serial correlation, and so on. The normality test is a joint test for skewness and excess kurtosis. Failure implies that the distribution of the estimates is non-normal, but does not necessarily have further implications. However, non-normality is often used as an indicator of unmodelled features of the data, including unmodelled outliers. The test for serial correlation picks up systematic relationships between residuals and their lags, and is again an indicator of unmodelled information. Similarly, the ARCH test, based on the between squared residuals, might give evidence of heteroscedasticity due to excluded variables. The non-linearity test picks up systematic relationships between the residuals and powers of the lagged residuals, and is a diagnostic for functional form misspecification.

Starting with the forecasting equations for growth, the p-values associated with each of these diagnostics are given in Table G, together with the error variance. The table shows that across all models, the residuals are well-behaved. There is only one case (the non-linearity test in the basic VAR) where there is 10% significance and nothing at 5%. In the case of inflation, there are two cases at the 5% level (ARCH and non-linearity in the AR model) and two instances where the diagnostic is significant at the 10% level (first-order serial correlation in the RW model and ARCH(1) in the STAR model). So for both forecast variables, the test violations are no more than would be expected by chance.

5.6 Persistence in forecast errors

There is some evidence that the Bank's main forecast errors have been persistent, as discussed in a recent *Bank of England Quarterly Bulletin* article (Elder, Kapetanios, Taylor and Yates (2005)). The evidence is based on only five years' worth of data (too little to draw strong conclusions) but there have been periods of up to two years during which outturns have been consistently higher or lower than forecasts. As the article explains, this is what we would expect to happen, even if the forecast were the best possible given the available information. This is because the forecaster cannot observe all of her past forecast errors and hence cannot learn from past mistakes. While one-period ahead forecast errors can be observed relatively quickly, after one quarter, others are observed much later; multiple-step forecast errors can be observed after a lag corresponding to the number of steps ahead the forecast was made. As a consequence, forecast errors may be persistent. That is, if the outcome exceeds the forecast in one period, this may also happen in the subsequent quarter. *A priori*, we may not expect this to be the case in the Suite. The Suite is conceptually quite different, as it is not based on judgement and does not condition on a pre-set path for interest rates. The basis for Suite forecasts is provided exclusively by the information in the data, and the difference between the various Suite forecasts is in the amount of information that is used, and how it is extracted from the data. The evidence indeed suggests that neither the Suite (in-sample) residuals nor the forecast errors for inflation are very persistent.

6 Conclusions

We have constructed a suite of statistical forecasting models that spans the space of commonly used models. This can be used to generate forecasts of output growth and inflation which can then be combined to create single 'best-guess' statistical forecasts.

Evaluation of the forecasts over the inflation-targeting period reveals that several individual forecasts outperform the simple AR forecast at various forecast horizons. Most fail to outperform the AR consistently, although the Markov-switching and STAR forecasts for GDP growth and the Markov-switching, STAR and factor model forecasts for inflation perform well for several horizons. These are interesting and practically useful results. But it is striking that forecast performance relative to the AR model is improved when forecasts are combined, and the best forecast combinations for both growth and inflation are those based on the information criteria based on in-sample fit of the forecasting models. These combined forecasts incorporate information from the entire range of models and data, and so to some extent they are robust to model misspecification. As another performance test, diagnostics based on the residuals of the forecasting equations show that the residuals are well-behaved in the large majority of cases.

On the evidence of the data we examine in this paper, combinations of statistical forecasts generate good forecasts of the key macroeconomic variables we are interested in, which can serve as a judgement-free benchmark forecast to compare with the policymaker's projections (which are conditional on judgements and assumptions about policy paths, and use a wide range of inputs). Moreover, the impact of new data on these forecasts provides a summary measure of the relevant news in the data, giving a natural indicator of changing inflationary pressure over the horizons of policy interest.

Appendix A: Bayesian model averaging

Model averaging reflects the need to account for model uncertainty in carrying out statistical analysis. From a Bayesian perspective, model uncertainty is straightforwardly handled using posterior model probabilities. The use of posterior model probabilities for forecasting has been suggested, discussed and applied by, among others, Min and Zellner (1993), Koop and Potter (2003), Draper (1995) and Wright (2003a,b). Briefly, under Bayesian model averaging a researcher starts with a set of models which have been singled out as useful representations of the data. We denote this set as $\mathcal{M} = \{M_i\}_{i=1}^{N}$ where M_i is the ith of the N models considered. The focus of interest is some quantity of interest for the analysis, denoted by Δ. This could be a parameter, or a forecast, such as inflation h quarters ahead. The output of a Bayesian analysis is a probability distribution for Δ given the set of models and the observed data at time t. Let us denote the relevant information set at time t by D_t. We denote the probability distribution as $pr(\Delta|D, \mathcal{M})$. This is given by

$$pr(\Delta|D_t, \mathcal{M}) = \sum_{i=1}^{N} pr(\Delta|M_i, D_t) pr(M_i|D_t) \quad \textbf{(A-1)}$$

where $pr(\Delta|M_i, D_t)$ denotes the conditional probability distribution of Δ given a model M_i and the data D_t and $pr(M_i|D_t)$ denotes the conditional probability of the model M_i being the true model given the data. It is clear that implementation requires two quantities to be obtained at each point in time. First, $pr(\Delta|M_i, D_t)$ which is easily obtained from standard model-specific analysis. Second, the weights, $pr(M_i|D_t)$. It is easy to see that the weights are formed as part of a stochastic process where $pr(M_i|D_t)$ is obtained from $pr(M_i|D_{t-1})$ via a number of intermediate steps. This implies the need of a prior distribution $pr(M_i|D_0) = pr(M_i)$ and for $pr(\theta_i|M_i, D_{t-1})$ to be specified.

Thus we need to obtain a number of expressions for **(A-1)** to be operational. First, using Bayes' theorem

$$pr(M_i|D_t) = \frac{pr(D_t|M_i, D_{t-1}) pr(M_i|D_{t-1})}{pr(D_t|D_{t-1})} = \frac{pr(D_t|M_i, D_{t-1}) pr(M_i|D_{t-1})}{\sum_{i=1}^{N} pr(D_t|M_i, D_{t-1}) pr(M_i|D_{t-1})} \quad \textbf{(A-2)}$$

where $pr(D_t|M_i, D_{t-1})$ denotes the conditional probability distribution of the data given the model M_i and the previous period's data, $pr(M_i|D_{t-1})$ denotes the conditional probability of the model M_i being true, given the previous period's data.

$$pr(D_t|M_i, D_{t-1}) = \int pr(D_t|\theta_i, M_i, D_{t-1})pr(\theta_i|M_i, D_{t-1})d\theta_i \quad \textbf{(A-3)}$$

(A-3) is the likelihood of model M_i and θ_i are the parameters of model M_i. Given this, the quantity of interest is

$$E(\Delta|D_t) = \sum_{i=1}^{N} \hat{\Delta}_i pr(M_i|D_t) \quad \textbf{(A-4)}$$

In theory (see, eg, Madigan and Raftery (1994)) when Δ is a forecast, this sort of averaging provides better average predictive ability than single model forecasts.

Appendix B: Predictive likelihood model averaging

An extension to the information theoretic approach is to use forecast errors from regression models in the construction of L_i, rather than in-sample residuals. To fix ideas consider the regression model

$$y_t = \alpha' x_t + \epsilon_t \qquad \textbf{(B-1)}$$

The concentrated log-likelihood of this model is given by $-T/2ln(\hat{\sigma}^2)$ where $\hat{\sigma}^2 = 1/T \sum_{t=1}^{T} \hat{\epsilon}_t^2$, $\epsilon_t = y_t - \hat{\alpha}^{(1,T)} x_t$ and $\hat{\alpha}^{(1,T)}$ denotes the estimate of α using data from $t = 1$ to $t = T$. The predictive likelihood measure replaces $\hat{\epsilon}_t$ with $\tilde{\epsilon}_t$ for $t = t_0, \dots, T$, where $\tilde{\epsilon}_t = y_t - \hat{\alpha}^{t-t_0,t-1} x_t$. In other words we use out-of-sample forecast errors rather than residuals. Interestingly, this implies that the predictive likelihood measure will change depending on the forecast horizon. Clearly, due to the recursive nature of the scheme there are fewer out-of-sample errors than residuals since one has to have an original sample for the first estimate of α, $\hat{\alpha}_{1,t_0}$, where t_0 has to be chosen *a priori*. Note that if we set $t_0 = bT$, where $0 < b < 1$, the model selection consistency properties of the various information criteria are retained. This approach is examined in Kapetanios *et al* (2007), where it is found to perform as well as or better than the Akaike-based method.

Appendix C: Forecast descriptions and mnemonics

Individual forecasts

1. UC (unconditional mean).
2. RW (random walk or no-change model). The first benchmark model in the Suite. The forecast is equal to the last data point.
3. AR (autoregressive model). The second benchmark model in the Suite. The forecast depends linearly on the lag(s) of the forecast variable.
4. V (vector autoregressive model). The forecast depends linearly on lag(s) of the forecast variable and other key macro variables.
5. VM (vector autoregressive model belonging to the monetary category). The forecast depends linearly on lag(s) of the forecast variable and other key macro variables including money.
6. DDV (double-differenced vector autoregressive model). The forecast depends linearly on lag(s) of (second differences of) the forecast variable and other key macro variables.
7. DDVM (double-differenced vector autoregressive model belonging to the monetary category). The forecast depends linearly on lag(s) of (second differences of) the forecast variable and other key macro variables, including money.
8. RV (recursively estimated vector autoregressive model). The forecast depends linearly on lag(s) of macro variables which are selected according to their forecasting ability. The selection is made from a small set of potential variables, exploiting all possible dimensions and combinations.
9. RVGEN (recursively estimated vector autoregressive model). The forecast depends linearly on lag(s) of macro variables which are selected according to their forecasting ability. The selection is made from a large set of potential variables, using a search algorithm.
10. MS (Markov-switching model). The forecast depends on the lag(s) of the forecast variable and the properties of two regimes. In each regime the dependence is linear but overall it is non-linear.
11. STAR (smooth-transition autoregressive model). The forecast depends on the lag(s) of the forecast variable and a threshold. Close to the threshold, the dependence is linear, otherwise non-linear.
12. FW (factor model). The forecast depends on the lag(s) of the forecast variable and a common

factor, extracted by means of static principal components.

13. BVM (Bayesian vector autoregressive model). The forecast depends linearly on lag(s) of the forecast variable and other key macro variables. The coefficients are estimated using Bayesian techniques.
14. BVMM (Bayesian vector autoregressive model belonging to the monetary category). The forecast depends linearly on lag(s) of the forecast variable and other key macro variables, including money. The coefficients are estimated using Bayesian techniques.
15. QMA (Quarterly Monetary Assessment). A dynamic factor model with a high degree of time variation in parameters, including volatility. The model is estimated by maximum likelihood, using the Kalman filter.
16. QMAM (Quarterly Monetary Assessment belonging to the monetary category). A dynamic factor model with a high degree of time variation in parameters, including volatility. The model includes several monetary indicators. The model is estimated by maximum likelihood, using the Kalman filter.

Combined forecasts

17. WEQ (simple model average). The benchmark forecast combination. The forecast is a weighted average of all individual forecasts.
18. WM (average of models belonging to the monetary category). The forecast is a combination of forecasts from the models which include money.
19. WMN (average of models not belonging to the monetary category). The forecast is a combination of forecasts from the models which do not include money.
20. WR (average of models belonging to the robust category). The forecast is a combination of forecasts from the models robust to structural change.
21. WRN (average of models not belonging to the robust category). The forecast is a combination of forecasts from the models not robust to structural change.
22. WV (average of multivariate models). The forecast is a combination of forecasts which depend on lag(s) of macro variables including the forecast variable.
23. WVN (average of univariate models). The forecast is a combination of forecasts which depend on lag(s) of only the forecast variable.
24. WITMA (information-theoretic model average). The forecast is a combination of forecasts based on the in-sample fit of the models as measured by the Akaike criterion.

Appendix D: Factor data set

In this appendix, we provide a list of the series used in the factor models. These series come from a data set which has been constructed to match the set used by Stock and Watson (2002). In total, this data set has 131 series, comprising 20 output series, 25 labour market series, 9 retail and trade series, 6 consumption series, 6 series on housing starts, 12 series on inventories and sales, 8 series on orders, 7 stock price series, 5 exchange rate series, 7 interest rate series and 6 monetary aggregates, 19 price indices and an economic sentiment index. We retained the 58 series with at least 90 observations. For each series the list gives an alias (the ONS code where available, **emboldened**), a brief description, seasonal adjustment (SA), the transformation applied to the series to ensure stationarity and the first available observation. The transformations applied to the series are: 1 = no transformation; 2 = first difference; 3 = second difference; 4 = logarithm; 5 = first difference of logarithm; 6 = second difference of logarithm. Series 4, 5, 10, 11, 12, 13, 21 and 32 are derived series, described in the list. The series are grouped into 10 categories.

Series 1 to 8: Real output and income

1. **ABMI**: Gross Domestic Product: chained volume measures: SA 5 Q1:1955
2. **CKYY** Index of production (IOP): Manufacturing SA 5 Q1:1948
3. IOP: Durable Manufacturing SA 5 Q1:1948
4. IOP: Semi-durable Manufacturing SA 5 Q1:1948; constructed as **CKZB** (IOP: Industry DB: Manuf of textile & textile products) plus **CKZC** (IOP: Industry DC: Manuf of leather & leather products) plus **CKZG** (IOP: Industry DG: Manuf of chemicals & man-made fibres) plus **CKZH** (IOP: Industry DH: Manuf of rubber & plastic products)
5. IOP: Non-durable Manufacturing SA 5 Q1:1948; constructed as **CKZA** (IOP: Industry DA: Manuf of food, drink & tobacco) plus **CKZE** (IOP: Industry DE: Pulp/paper/printing/publishing industries) plus **CKZF** (IOP: Industry DF: Manuf coke/petroleum prod/nuclear fuels)
6. **CKYX** IOP: Mining & quarrying SA 5 Q1:1948
7. **CKYZ** IOP: Electricity, gas and water supply SA 5 Q1:1948
8. **NRJR**: Real households disposable income SA 5 Q1:1955

Series 9 to 21: Employment and hours

9. **DYDC**: UK Workforce jobs: Total SA 5 Q2:1959
10. Employed, Nonagric. Industries SA 5 Q2:1978; constructed as **DYDC** (UK Workforce jobs

(SA) : Total) minus **LOLI** (UK Workforce jobs (SA): Total - A,B Agriculture & fishing) minus **LOMJ** (UK Workforce jobs (SA): Total - G-Q Total services)

11. Employment Rate: All NSA 1 Q1:1971; concatenate **MGRZ** and **MGRZ**_EXP (LFS: In employment: UK: All: Aged 16), concatenate **MGSL** and **MGSL_EXP** (LFS: Population aged 16+: UK: All), then compute 1-**MGRZ**/**MGSL**
12. Employees on nonag. Payrolls: Total SA 5 Q2:1978; constructed as **BCAJ** (UK Employee jobs: Total (SA)) minus **YEHU** (UK Employee jobs (SA): All jobs Agriculture, hunting, forestry & fishing)
13. Employees nonag. Payrolls: Total: private SA 5 Q2:1978; constructed as Series 12 minus **LOKS** (UK Employee jobs (SA): Public admin. & defence)
14. **YEJF** Employee jobs: All jobs: Production Inds. SA 5 Q2:1978
15. **YEHX** Employee jobs: All jobs: Construction SA 5 Q2:1978
16. **YEHW** Employee jobs: All jobs: Manufacturing SA 5 Q2:1978
17. **LOKL** Employee jobs: Wholesale & retail trade SA 5 Q2:1978
18. **YEIA** Employee jobs: Banking, finance & ins. SA 5 Q2:1978
19. **YEID** Employee jobs: Total services SA 5 Q2:1978
20. **LOKS** Employee jobs: Public admin. & defence SA 5 Q2:1978
21. Avg. weekly hrs. prod. wkrs.: manuf. SA 1 Q1:1971; constructed from **YBUS** and **YBUS_EXP** (LFS: Total actual weekly hours worked (millions): UK: All), **MGRZ** and **MGRZ_EXP** (LFS: In employment: UK: All: Aged 16+ SA), as **YBUS**/**MGRZ**

Series 22 to 23: Trade

22. **BOKI** BOP: Balance: Total Trade in Goods SA 5 Q1:1955
23. **ELBJ** BOP: Balance: Manufactures SA 5 Q1:1970

Series 24 to 29: Consumption

24. **ABJR** Household final consumption expenditure SA 5 Q1:1955
25. **UTID** Durable goods: Total SA 5 Q1:1964
26. **UTIT** Semi-durable goods: Total SA 5 Q1:1964
27. **UTIL** Non-durable goods: Total SA 5 Q1:1964
28. **UTIP** Services: Total SA 5 Q1:1964
29. **TMMI** Purchase of vehicles SA 5 Q1:1964

Series 30 to 35: Real inventories and inventories sales

30. **CDQN** Change in Inventories: Manufacturing SA 5 Q4:1954
31. **CDQZ** Change in Inv: Manuf: Textiles & Leather SA 5 Q4:1954

32. Manuf & Trade Invent: Nondurable Goods SA 5 Q4:1954; constructed as **CDQP** (Change in Inventories: Manufacturing: Fuels) plus **CDQX** (Change in Inventories: Manufacturing: Food, Drink & Tobacco) plus **CDQT** (Change in Inventories: Manufacturing: Chemicals)
33. **FAJX** Change in Inventories: Wholesale SA 5 Q1:1959
34. **FBYN** Change in Inventories: Retail SA 5 Q1:1955
35. **FAPF** Ratio for Mfg & Trade: Inventory/Output SA 2 Q1:1955

Series 36 to 38: Stock prices

36. **FTALLSH_PI** FTSE All-Share Price Index 5 Q1:1980
37. **FTSE100_PI** FTSE 100 5 Q1:1980
38. **FTALLSH_DY** FTSE All-Share Dividend Yield 1 Q1:1980

Series 39 to 43: Exchange rates

39. **A_GBG** Sterling - Effective SA 5 Q1:1979
40. **A_ERS** EURO / £SA 5 Q1:1979; constructed from **A_DMS** (MTH AVE - DEUTSCHEMARK /£) and fixed conversion rate of 1.95583
41. **A_SFS** SWISS FRANC /£SA 5 Q1:1979
42. **A_JYS** JAPANESE YEN /£SA 5 Q1:1979
43. **A_USS** UNITED STATES DOLLAR /£SA 5 Q1:1979

Series 44 to 47: Interest rates

44. Spread 6-months 1
45. Spread 1-year 1
46. Spread 5-years 1
47. Spread 10-years 1

Series 48 to 50: Monetary and quantity credit aggregates

48. **AUYN** Money stock: M4 SA 6 Q2:1963
49. **AVAE** M0 wide monetary base SA 6 Q2:1969
50. **AEFI** BOE: reserves & other accounts outstanding NSA 6 Q1:1975

Series 51 to 57: Price indices

51. **PLLU** PPI: Output of manufactured products NSA 6 Q1:1974
52. **LCPI** Long Run CPI NSA 6 Q1:1975
53. **ABJS** Implicit Price Deflator (IDEF): Household final consumption expenditure SA 6 Q1:1955
54. **UTKT** Durable goods IDEF SA 6 Q1:1964
55. **UTLB** Semi-durable goods IDEF SA 6 Q1:1964
56. **UTKX** Non-durable goods IDEF SA 6 Q1:1964

57. **UTKZ** Services IDEF SA 6 Q1:1964

Series 58: Surveys

58. **MORI** MORI General Economic Optimism index SA 1 Q3:1979

Appendix E: QMA data set

The CPI (acronym is CHVJ) is from the *Office for National Statistics*.

The twelve real activity indicators used to construct the real activity factor are (mnemonics are in parentheses): Households' final domestic consumption expenditure at constant 1995 market prices (ABJR); General government's final domestic consumption expenditure at constant 1995 market prices (NMRY); Gross value added, chained volume measures, construction (GDQB); Exports at constant 1995 market prices (IKBK); Imports at constant 1995 market prices (IKBL); Gross fixed capital formation at constant 1995 market prices (NPQT); Gross domestic product at constant 1995 market prices (ABMI); Gross value added, chained volume measures, manufacturing (CKYY); Gross value added, chained volume measures, transport, storage and communications (GDQH); Gross value added, chained volume measures, all service industries (GDQS); Gross value added, chained volume measures, distribution, hotels and catering (GDQE); and Gross value added, chained volume measures, all production industries (CKYW). All series are available from 1955:1 to 2003:2.

The two series we use to construct the money growth factor are: notes & coin in circulation outside the Bank of England: level, seasonally adjusted (AVAB; sample period: 1969:3-2003:2), and quarterly break-adjusted M4 level, seasonally adjusted (M4SA; sample period: 1963:2–2003:2). The five series we use to construct the credit growth factor are: quarterly amounts outstanding of monetary financial institutions' sterling net lending to other financial corporations (in sterling millions) seasonally adjusted (VQSI); quarterly amounts outstanding of monetary financial institutions' sterling net lending to private non-financial corporations (in sterling millions) seasonally adjusted (VQSG); quarterly amounts outstanding of monetary financial institutions' sterling net secured lending to individuals (in sterling millions) seasonally adjusted (VQSL); quarterly amounts outstanding of monetary financial institutions' sterling net unsecured lending to individuals (in sterling millions) seasonally adjusted (VQSK); quarterly amounts outstanding of monetary financial institutions' sterling net lending to unincorporated businesses and non-profit institutions serving households (in sterling millions) seasonally adjusted (VQSP).

References

Akaike, H (1978), 'A Bayesian analysis of the minimum AIC procedure', *Annals of the Institute of Statistical Mathematics*, Vol. 30.

Akaike, H (1979), 'A Bayesian extension of the minimum AIC procedure of autoregressive model fitting', *Biometrika*, Vol. 66.

Akaike, H (1981), 'Modern development of statistical methods', in Eykhoff, P (ed), *Trends and progress in system identification*, Pergamon Press Paris, pages 169–84.

Akaike, H (1983), 'Information measures and model selection', *International Statistical Institute*, Vol. 44.

Bai, J and Perron, P (1998), 'Testing for and estimation of multiple structural breaks', *Econometrica*, Vol. 66, pages 47–79.

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

Bank of England (2000), *Economic models at the Bank of England: September 2000 update*, London: Bank of England.

Bank of England (2003), 'Bank's response to the Pagan Report', *Bank of England Quarterly Bulletin*, Spring, pages 89–91.

Bank of England (2005), *The Bank of England Quarterly Model*, London: Bank of England.

Bates, J M and Granger, C W J (1969), 'The combination of forecasts', *Operations Research Quarterly*, Vol. 20, pages 451–68.

Bernanke, B S, Boivin, J and Eliasz, P (2005), 'Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach', *Quarterly Journal of Economics*, Vol. 120, pages 387–422.

Boivin, J and Ng, S (2005), 'Understanding and comparing factor-based forecasts', *International Journal of Central Banking*, Vol. 1, pages 117–52.

Bozdogan, H (1987), 'Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions', *Psychometrika*, Vol. 52, No. 3, pages 345–70.

Burnham, K P and Anderson, D R (1998), *Model selection and inference*, Berlin: Springer Verlag.

Clark, T E and McCracken, M W (2004), 'Improving forecast accuracy by combining recursive and rolling forecasts', Federal Reserve Bank of Kansas City, *Working Paper no. 04-10*.

Clark, T E and McCracken, M W (2006), 'The predictive content of the output gap for inflation: resolving in-sample and out-of-sample evidence', *Journal of Money, Credit, and Banking,* forthcoming.

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

Clements, M P and Hendry, D F (2002), 'Pooling of forecasts', *Econometrics Journal*, Vol. 5, pages 1–26.

Cogley, T and Sargent, T (2002), 'Drifts and volatilities: monetary policies and outcomes in the post WWII US', *mimeo*, Arizona State University.

Demers, F and Marci, P (2005), 'Econometric forecasts package: short-run forecasting models for the current analysis of the Canadian economy', *mimeo*, Bank of Canada.

Diebold, F X and Pauly, P (1987), 'Structural change and the combination of forecasts', *Journal of Forecasting*, Vol. 6, pages 21–40.

Diebold, F X and Pauly, P (1990), 'The use of prior information in forecast combination', *International Journal of Forecasting*, Vol. 6, pages 503–08.

Doan, T, Litterman, R B and Sims, C A (1984), 'Forecasting and conditional projection using realistic prior distributions', *Econometric Reviews*, Vol. 3, pages 1–144.

Draper, D (1995), 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society, Series B*, Vol. 57, pages 45–97.

Elder, R, Kapetanios, G, Taylor, T and Yates, T (2005), 'Assessing the MPC's fan charts', *Bank of England Quarterly Bulletin*, Autumn, pages 326–48.

Forni, M, Hallin, M, Lippi, M and Reichlin, L (2000), 'The generalised factor model: identification and estimation', *Review of Economics and Statistics*, Vol. 82, pages 540–54.

Granger, C W J and Ramanathan, R (1984), 'Improved methods of combining forecasting', *Journal of Forecasting*, Vol. 3, pages 197–204.

Hamilton, J (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica*, Vol. 57, pages 357–84.

Jacobson, T and Karlsson, S (2004), 'Finding good predictors for inflation: a Bayesian model averaging approach', *Journal of Forecasting*, Vol. 23, page 479.

Kapetanios, G (2005), 'Variable selection using non-standard optimisation of information criteria', Queen Mary and Westfield College, *Working Paper no. 533*.

Kapetanios, G, Labhard, V and Price, S (2007), 'Forecasting using Bayesian and information theoretic model averaging: an application to UK inflation', *Journal of Business Economics and Statistics*, forthcoming.

Koop, G and Potter, S (2003), 'Forecasting in large macroeconomic panels using Bayesian model averaging', *Federal Reserve Bank of New York Report 163*.

Kullback, S and Leibler, R A (1951), 'On information and sufficiency', *Annals of Mathematical Statistics*, Vol. 22, pages 79–86.

Litterman, R B (1986), 'Forecasting with Bayesian vector autoregressions: five years of experience', *Journal of Business and Economic Statistics*, Vol. 4, pages 25–38.

Madigan, D and Raftery, A E (1994), 'Model selection and accounting for model uncertainty in graphical models using Occam's window', *Journal of the American Statistical Association*, Vol. 89, pages 1,535–46.

Min, C and Zellner, A (1993), 'Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates', *Journal of Econometrics*, Vol. 56, pages 89–118.

Newbold, P and Granger, C W J (1974), 'Experience with forecasting univariate time series and the combination of forecasts', *Journal of the Royal Statistical Society, Series A*, Vol. 137, pages 131–65.

Newbold, P and Harvey, D I (2002), 'Forecast combination and encompassing', in Clements, M and Hendry, D F (eds), *A companion to economic forecasting*, Oxford: Basil Blackwell, chap. 12.

Pagan, A (2003), 'Report on modelling and forecasting at the Bank of England', *Bank of England Quarterly Bulletin*, Spring, pages 60–88.

Pesaran, M H and Timmermann, A (2000), 'A recursive modelling approach to predicting UK stock returns', *Economic Journal*, pages 159–91.

Royall, R M (1997), *Statistical evidence: a likelihood paradigm*, New York: Chapman and Hall.

Stock, J and Watson, M (1999), 'Forecasting inflation', *Journal of Monetary Economics*, Vol. 44, pages 293–335.

Stock, J and Watson, M (2002), 'Macroeconomic forecasting using diffusion indices', *Journal of Business and Economic Statistics*, Vol. 20, pages 147–62.

Stock, J and Watson, M (2005), 'Has inflation become harder to forecast?', unpublished.

Svensson, L E O (2004), 'Monetary policy with judgment: forecast targeting', unpublished.

Wallis, K F and Whitley, J D (1981), 'Sources of error in forecasts and expectations: UK economic models, 1984-88', *Journal of Forecasting*, Vol. 10, pages 231–53.

Wright, J H (2003a), 'Bayesian model averaging and exchange rate forecasts', Board of Governors of the Federal Reserve System, *International Finance Discussion Papers no. 779*.

Wright, J H (2003b), 'Forecasting US inflation by Bayesian model averaging', Board of Governors of the Federal Reserve System, *International Finance Discussion Papers no. 780*.