Forecasting Annual UK Inflation using an Econometric Model over 1875–1991

Michael P. Clements
Department of Economics,
University of Warwick,
Coventry CV4 7AL.

David F. Hendry*
Economics Department,
Oxford University

March 29, 2006

Abstract

In recent work, we have developed a theory of economic forecasting for empirical econometric models when there are structural breaks. This shows that well-specified models may forecast poorly, whereas it may be possible to design forecasting models or devices which are more immune to the effects of breaks. In this chapter, we provide an empirical illustration of some of these developments when the goal is to generate sequences of inflation forecasts over a long historical period. The empirical model we take as our starting point is the model of annual inflation in the UK over 1875–1991 in Hendry (2001b). This model includes excess demand forces from all sectors of the economy (goods and services, factors of production, money, financial assets, foreign exchange, and government deficits), represented by equilibrium-correction terms. During the last century or so there have been a multiplicity of structural breaks, many of huge magnitude, emanating from a wide range of sources. We present a comparison of the one through 4-year ahead in-sample inflation ‘forecasts’ of the final two decades for the econometric model against a number of competitors. These competitors are all suggested by, or designed to bring out, various aspects of the theory of forecasting in the face of breaks. We consider forecasts from the original model, a general model, models selected for each horizon, a naive forecast, and a joint model, as well as intercept corrected forecasts, and a number of ‘pooling’ approaches. Of these possible models and methods, we find that forecast combination improves accuracy, especially over longer horizons, but forecast failure is all too common.

1 Introduction

The centerpiece of UK economic policy over the last five years has been control of inflation based on the interest rate set by the Bank of England’s Monetary Policy Committee. The Bank seeks to forecast inflation two years’ ahead, and raise or lower interest rates as that forecast is above or below the target (which was 2.5% pa using the RPI, and is now 2% based on the CPI). Thus, the accuracy with which inflation can be forecast from their model, conditional on policy variables, is of considerable importance to the success of that policy. Several authors have sought to evaluate how well the Bank

*Financial support from the ESRC under a Professorial Research Fellowship, RES051270035, is gratefully acknowledged by the second author. We are indebted to Toni Espasa and Robert Fildes for helpful comments on an earlier draft.
has done, with the more recent contributions focusing on the Bank’s density forecasts (see e.g., Wallis (2003), Clements (2004) and Hall and Mitchell (2005)). Our purpose is two-fold. To examine inflation forecasts from a longer-run time series to gauge the success—or otherwise—of inflation forecasting during turbulent periods of economic change, and in so doing, to provide an empirical illustration of some aspects of our recent work on the development of a general theory of economic forecasting in the face of structural breaks (see, e.g., Clements and Hendry (1998, 1999a, 2002b, 2005b)). Our earlier work shows that well-specified models may forecast poorly when there are structural breaks, or when the parameters drift, although it may be possible to design forecasting models or devices whose forecasts are less affected.

The ‘well-specified model’ is that of Hendry (2001b). This is an eclectic model of price inflation in the UK over 1875–1991. Inflation responds to excess demands from all sectors of the economy, namely goods and services, factors of production (labour and capital), money, financial and foreign exchange markets, and government. The model also allows for historical contingencies, such as major wars and oil crises. The research agenda was to investigate which of the many competing explanations available in economics ‘caused’ inflation. Additionally, the roles of non-stationarities due to both unit roots from the integrated-cointegrated nature of the economy, and ‘structural breaks’, such as economic-policy regime shifts, world wars and commodity crises, major legislative and technological changes were of concern in developing a sustainable model for such a long run of data. In this chapter our interest is squarely on the ‘forecasting’ performance1 of the selected model and some related specifications over the final two decades of that sample, when inflation first rose dramatically, then fell sharply, and remained at a lower, more stable level, before rising towards the end of the forecast period. The relative forecasting performance of the models is compared with what might have been anticipated on the basis of the general theory of economic forecasting referred to above. Specifically, the present exercise evaluates the empirical forecasting performance of various approaches in terms of five dichotomies:

1. parsimonious against profligate parametrizations;
2. selecting different models for each horizon as against lagging a fixed specification;
3. robust versus non-robust devices;
4. pooled forecasts compared to individual models; and
5. intercept-corrections relative to a (mimic of a) ‘factor’ model.

The forecasting theory predicts that the first member of each of these five dichotomies should deliver some improvement over the second when the data generation process is subject to location shifts and the models are not correctly specified, as seems realistic in practice. In section 2 we present a review of the theory that supports these assertions. Before doing so, we briefly summarise the forecasting models and methods indicated by these five dichotomies.

---

1The lag of the information used in every model is at least as long as the forecast horizon, so all models could be used to generate ex ante forecasts, in this regard. The inverted commas around forecasting indicate in-sample forecasts over the period on which the model was selected and estimated, as opposed to true out-of-sample forecasts. More subtly, the forecasts are not ‘real time’. The effects of data vintage on model specification and the production and evaluation of forecasts have been addressed in a number of recent papers: Croushore and Stark (2001, 2003), Orphanides (2001), Koenig, Dolmas and Piger (2003), Patterson (1995, 2003), Robertson and Tallman (1998), Faust, Rogers and Wright (2003).
For 1., profligate parametrizations both increase parameter estimation uncertainty and induce the inclusion of ‘unreliable’ variables whose coefficients can alter over the forecast period. Thus, we will compare forecasting from:

\[ \hat{y}_{t+h|t} = \hat{\beta}'_h z_t, \]  

(1)
as against:

\[ \tilde{y}_{t+h|t} = \tilde{\gamma}'_h z_{1,t}, \]  

(2)

where \( z_{1,t} \) is a suitably reduced subset of \( z_t \), the set of explanatory variables dated period \( t \) and earlier (so possibly including lags). \( y \) is the variable being forecast, and the \( h \) subscript on the parameter vector indicates the dependence on the forecast horizon — the parameter estimates change with the forecast horizon.

For 2., selection should more closely match the model to the step-ahead to be forecast, adding flexibility over a fixed specification. We denote this by:

\[ \hat{y}_{t+h|t} = \hat{\delta}'_h z_{h,t}, \]  

(3)
to indicate that the elements of \( z_{h,t} \) may change with the horizon, as against (2).

For 3., robustness matters for avoiding systematic forecast errors when structural breaks occur—and there is evidence for location shifts over the forecast periods considered here. Here, we compare the various ‘causal’ specifications above with the naive predictor:

\[ \Delta^2 y_{t+h|t} = 0. \]  

(4)

For 4., pooling has long been known to perform successfully relative to individual forecasting devices (see inter alia, Bates and Granger, 1969, Newbold and Granger, 1974, Clemen, 1989, Diebold and Lopez, 1996, Stock and Watson, 1999a, Newbold and Harvey, 2002, and Fildes and Ord, 2002). A formal analysis of why pooling will outperform when all models are mis-specified representations of processes subject to location shifts is provided in Hendry and Clements (2004a). The average forecast from sets of competing individual specifications, \( \hat{y}_{t+h|t}^{(i)} \), is given by:

\[ \bar{y}_{t+h|t} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{t+h|t}^{(i)}. \]  

(5)

Finally, for 5., the application of that theory to factor models as proposed in Stock and Watson (1999a) implies that intercept corrections will have value added, as the former are not robust to location shifts: perhaps surprisingly, pooling forecasts may be more productive than pooling information when the factor model data base is not fully comprehensive and location shifts occur. Thus, we will compare:

\[ \hat{y}_{t+h|t} = \hat{\phi}_h \left( \hat{\beta}'_1 z_t \right), \]  

(6)

against intercept correcting (say) (2).

The structure of the paper is as follows. After reviewing forecasting when there are breaks in section 2, as part of a more general theory of forecasting, section 3 explains the data series, the equilibrium-correction measures of the various excess demands, the general model specification, and the finally selected model. Next, section 4 presents the individual forecasting methods, and their ‘forecasts’ over the decades 1982–1991 and 1972–1981, for 1 through 4 years ahead. Five individual forecasting device are investigated, all based on multi-step estimation (see e.g., Weiss, 1991, Clements and Hendry, 1996c, Chevillon, 2000, and Bhansali, 2002):
a. the originally chosen model with appropriately lagged values of its regressors;
b. re-selecting from the general model when all variables are lagged the horizon length;
c. that general model itself;
d. a robust (naive) ‘zero-acceleration’ predictor;
e. and jointly forecasting UK and world inflation.

The impact of intercept corrections is also investigated.

Then section 4.2 examines four ‘pooling’ approaches, namely an average of the individual forecasts (see Hendry and Clements, 2004a); a mimic of the ‘factor forecasts’ approach in Stock and Watson (1999a) where the available information is pooled using a linear combination of regressors lagged; a simplification of that method using lagged fitted values of the original model; and an overall average of the unpooled forecasts. Section 5 compares these forecasts and considers the relative performance of forecast pooling as against pooling information. Section 6 discusses the implications of these simple forecasting models for inflation targeting at the Bank of England, and section 7 concludes. All the empirical analyses and graphics use PcGive (see Hendry and Doornik, 1996) and PcGets (see Doornik, 1998, and Hendry and Krolzig, 1999).

2 Aspects of a general theory of forecasting

If we assume that the model to be used for forecasting closely matches the actual process generating the data, both in the past, and during the future period for which we require forecasts, then there are two main sources of forecast error: the random error terms in the model that represent unmodelled factors that affect the variable to be forecast; and the fact that the model’s parameters are not known and have to be estimated. These sources of error form the staple of many standard textbook accounts, and follow naturally from the assumption that the model is ‘correctly-specified’ for a constant process. These sources of forecast error or uncertainty can be described as predictable, as opposed to unpredictable, uncertainty (a distinction formally drawn by Clements and Hendry, 1998, 1999a, and Ericsson, 2002) and can be readily quantified. Unpredictable uncertainty is more difficult to analyse and quantify, although three main sources can be discerned. One is when the forecast model is mis-specified in the sense that it omits relevant variables that in reality influence the variable being forecast. Another is when the relationships on which the forecasting model is based are subject to shifts, where the structural change may be precipitated by various factors, be they institutional, political, financial, or technological in nature. Thirdly, forecast errors will result from the data that is being used in the forecast model being mis-measured.

Most econometrics texts consider only the predictable sources of forecast error, as they are primarily concerned with the construction of ex ante measures of forecast uncertainty. For example, Wooldridge (2003, esp. p. 630-1) considers the impact of estimation and the random disturbances in deriving prediction intervals, as do Stock and Watson (2003, p. 450-1), Enders (1995, p. 99-101) and Harris and Sollis (2003, p. 10-15). Mittelhammer, Judge and Miller (2000, p. 88-91) justify OLS as an estimator that avoids large forecast errors (in correctly specified models). Greene (2000, Section 7.11) considers estimation uncertainty and also refers to the impact of uncertainty over the values of the variables that the
forecast is conditioned on (the data mis-measurement source).\(^2\) Nevertheless, most standard textbooks detail tests of the stability of regression coefficients, and in particular tests of forecast/predictive failure, which are explicitly based on the recognition that parameter non-constancy (i.e., structural breaks) may cause large forecast errors, so that this source of forecast error is clearly recognised (e.g., Patterson, 2000).

Our analyses of sources of forecast error indicate that in practice, structural breaks are likely to be the main cause of large, unexpected forecast errors. In sub-section 2.1, we show the effects of location shifts on the forecast performance of the popular equilibrium-correction (formerly, ‘error-correction’) class of models, and in sub-section 2.2, we outline models and methods that might be expected to generate less inaccurate forecasts when there are breaks. We also explain how an acknowledgement of the ‘unpredictable’ sources of forecast error suggests possible gains from: selecting different models for each horizon versus a fixed specification, and pooling or combining forecasts from different sources (sub-section 2.3).

2.1 Structural breaks

We illustrate the effects of a location shift on the forecasts from an otherwise well-specified model. Many economic and financial time series can be characterised as being unit-root processes (i.e., integrated of order 1, \(I(1)\)), yet must satisfy cointegration for meaningful long-run relationships between them to exist. Consequently, we consider a first-order VAR (first-order for simplicity), where the vector of \(n\) variables of interest is denoted by \(x_t\):

\[
x_t = \tau + \Upsilon x_{t-1} + \epsilon_t \quad \text{where} \quad \epsilon_t \sim \mathcal{N}_n [0, \Omega],
\]

(7)

and where \(\Upsilon\) is an \(n \times n\) matrix of coefficients and \(\tau\) is an \(n\) dimensional vector of constant terms. The specification in (7) is assumed constant in-sample, and the system is taken to be \(I(1)\), satisfying the \(r < n\) cointegration relations:

\[
\Upsilon = I_n + \alpha \beta'.
\]

(8)

where \(\alpha\) and \(\beta\) are \(n \times r\) full-rank matrices. Then (7) can be reparameterized as a vector equilibrium-correction model (VEqCM):

\[
\Delta x_t = \tau + \alpha \beta' x_{t-1} + \epsilon_t,
\]

(9)

where both \(\Delta x_t\) and \(\beta' x_t\) are \(I(0)\), but may have non-zero means. Let \(E[\Delta x_t] = \gamma\) where \(\beta' \gamma = 0\), and \(E[\beta' x_t] = \mu\), then:

\[
\tau = \gamma - \alpha \mu
\]

(10)

so that we can write the VEqCM in ‘mean-deviation’ form as (see Clements and Hendry, 1996a):

\[
(\Delta x_t - \gamma) = \alpha (\beta' x_{t-1} - \mu) + \epsilon_t.
\]

(11)

The breaks we consider focus on shifts in the equilibrium mean, for the reasons detailed in Clements and Hendry (1999a, 2005b). At \(T + 1\) the DGP becomes:

\[
\Delta x_{T+1} = \gamma + \alpha (\beta' x_T - \mu^*) + \epsilon_{T+1}.
\]

\(^2\)An exception is Kennedy (1998, p. 289), who has specification error, conditioning error, sampling error and random error. This matches the sources distinguished here, except that Kennedy’s ‘specification error’ conflates the separate mis-specification and structural break sources, which transpire to have potentially very different effects.
Letting: 
\[ \hat{\Delta x}_{T+1|T} = E [\gamma + \alpha (\beta' x_T - \mu) + \epsilon_{T+1} | x_T] = \gamma + \alpha (\beta' x_T - \mu) \]
denote the conditional 1-step ahead forecast of \( \Delta x_{T+1} \) based on the model prior to the mean shift (and ignoring parameter estimation certainty to focus the analysis), then the forecast error is given by:
\[ \Delta x_{T+1} - \hat{\Delta x}_{T+1|T} = \epsilon_{T+1} - \alpha \nabla \mu^* . \tag{12} \]
where \( \nabla \mu^* = \mu^* - \mu \). The expected forecast error is:
\[ E [\Delta x_{T+1} - \hat{\Delta x}_{T+1|T}] = -\alpha \nabla \mu^* \neq 0 \tag{13} \]
so that the forecasts of the 1-step ahead changes (and levels) are biased.

Consider forecasting \( T + j \) based on \( T + j - 1 \), for \( j > 1 \): it is immediately apparent that the biases in the resulting sequence of 1-step ahead forecasts are given by (13). The VEqCM will produce biased 1-step forecasts ever after as a consequence of the shift in the equilibrium mean that occurred at period \( T + 1 \). There are a number of possible solutions to avoiding systematic errors. We discuss these in the following sub-section.

### 2.2 Robustifying forecasts against structural breaks

#### 2.2.1 Differencing the VEqCM

Consider forecasting not from (11) itself, but from a variant thereof, given by differencing (11):
\[ \Delta x_t = \Delta x_{t-1} + \alpha \beta' \Delta x_{t-1} + \Delta \epsilon_t = (I_n + \alpha \beta') \Delta x_{t-1} + u_t \tag{14} \]
or:
\[ \Delta^2 x_t = \alpha \beta' \Delta x_{t-1} + u_t. \tag{15} \]
(14) is a restricted version of a ‘VAR in differences – DVAR’ (see Clements and Hendry, 1995, 1996b) because the form of the parameter matrix linking \( \Delta x_t \) and \( \Delta x_{t-1} \) is given by \( \Upsilon \). To trace the behaviour of (14) one period after a break in \( \mu \), let:
\[ \tilde{\Delta x}_{T+2|T+1} = (I_n + \alpha \beta') \Delta x_{T+1}, \]
where from (15):
\[ \Delta x_{T+2} = \Delta x_{T+1} + \alpha (\beta' \Delta x_{T+1} - \Delta \mu^*) + \Delta \epsilon_{T+2} = \Delta x_{T+1} + \alpha \beta' \Delta x_{T+1} + \Delta \epsilon_{T+2} \]
since only at time \( T \) does \( \Delta \mu^* = \nabla \mu^* \). Then:
\[ E [\Delta x_{T+2} - \hat{\Delta x}_{T+2|T+1}] = E [\Delta x_{T+1} + \alpha \beta' \Delta x_{T+1} - (I_n + \alpha \beta') \Delta x_{T+1}] = 0. \]
Thus, the differenced VEqCM (DVEqCM) ‘misses’ for forecasting period \( T + 1 \), but the subsequent 1-step forecasts are unbiased. However, there is a cost in terms of the forecasts being less precise. The 1-step forecasts errors are \( e_{T+i} = \Delta x_{T+i} - \hat{\Delta x}_{T+i|T+i-1} \), so ignoring parameter estimation uncertainty as being \( O_p (T^{-1/2}) \), for \( i = 1, 2 \) we have:
\[ e_{T+1} = -\alpha \nabla \mu^* + \Delta \epsilon_{T+1} \]
\[ e_{T+2} = \Delta \epsilon_{T+2}. \]
As the error is \( \{ \epsilon_t \} \) is assumed white noise, then differencing doubles the 1-step error variance.
2.2.2 The DDV

We next consider second differencing. Because most economic time series do not continuously accelerate, the second difference has a zero unconditional expectation:

\[ E[\Delta^2 x_t] = 0. \]  

(16)

This suggests the well-known ‘minimal information’ or ‘constant-change’ forecasting rule:

\[ \Delta \tilde{x}_{T+1|T} = \Delta x_T. \]  

(17)

Hendry (2005) considers a more general DGP which incorporates non-modelled explanatory variables \( \{ \tilde{w}_t \} \):

\[ \Delta x_t = \gamma_0 + \alpha_0 (\beta_0' x_{t-1} - \mu_0) + \Psi_0 \tilde{w}_t + \epsilon_t \]  

(18)

to explain why forecasts of the form of (17) may sometimes perform well. We assume that \( \epsilon_t \sim \mathrm{IN}_n [0, \Omega_\epsilon] \) (independently of included variables and their history), with population parameter values denoted by the subscript 0. Here, \( \tilde{w}_t \) denotes potentially many omitted \( l(0) \) effects, possibly all lagged, so that they could in principle be included in a forecasting model if they were known. The assumption that they are \( l(0) \) may hold because of ‘internal’ cointegration, or because they have been differenced an appropriate number of times, or because they are intrinsically stationary. The postulated econometric model used for forecasting is the VEqCM for \( x_t \) as described above, so that the \( \{ \tilde{w}_t \} \) variable are omitted:

\[ \Delta x_T = \gamma + \alpha (\beta' x_{T-1} - \mu) + \nu_T. \]

We again neglect parameter estimation uncertainty as being \( O_p(T^{-1}) \), so that the forecasts are given by:

\[ \Delta x_{T+i|T+i-1} = \gamma + \alpha (\beta' x_{T+i-1} - \mu). \]  

(19)

Over the forecast horizon, the DGP is:

\[ \Delta x_{T+i} = \gamma^*_0 + \alpha^*_0 ((\beta^*_0)' x_{T+i-1} - \mu^*_0) + \Psi^*_0 \tilde{w}_{T+i} + \epsilon_{T+i}. \]  

(20)

Denote the forecast error by \( \tilde{u}_{T+i} = \Delta x_{T+i} - \Delta \tilde{x}_{T+i|T+i-1} \):

\[ \tilde{u}_{T+i} = \gamma^*_0 + \alpha^*_0 ((\beta^*_0)' x_{T+i-1} - \mu^*_0) + \Psi^*_0 \tilde{w}_{T+i} + \epsilon_{T+i} - \gamma - \alpha (\beta' x_{T+i-1} - \mu). \]  

(21)

Conditional on \((x_{T+i-1}, \tilde{w}_{T+i-1}) \), \( \tilde{u}_{T+i} \) has an approximate mean forecast error of:

\[ E[\tilde{u}_{T+i} | x_{T+i-1}, \tilde{w}_{T+i-1}] = (\gamma^*_0 - \gamma) - (\alpha^*_0 \mu^*_0 - \alpha \mu) + [\alpha^*_0 (\beta^*_0)' - \alpha \beta'] x_{T+i-1} + \Psi^*_0 E[\tilde{w}_{T+i} | x_{T+i-1}, \tilde{w}_{T+i-1}], \]  

(22)

and an approximate conditional error-variance matrix:

\[ \Psi[\tilde{u}_{T+i} | x_{T+i-1}, \tilde{w}_{T+i-1}] = \Psi^*_0 \Psi[\tilde{w}_{T+i} | x_{T+i-1}, \tilde{w}_{T+i-1}] \Psi^{*'}_0 + \Omega_\epsilon. \]  

(23)

The conditional mean-square forecast error matrix is the sum of:

\[ E[\tilde{u}_{T+i} | x_{T+i-1}, \tilde{w}_{T+i-1}] E[\tilde{u}_{T+i} | x_{T+i-1}, \tilde{w}_{T+i-1}]' \]  

and (23), where the former includes terms in the shifts in the parameters.
errors to the situation where the investigator uses the sequence of $\Delta x_{T+i-1}$ to forecast $\Delta x_{T+i}$, as in the DDV given by equation (17):

$$\Delta \tilde{x}_{T+i|T+i-1} = \Delta x_{T+i-1}. \tag{24}$$

From (20), for $i > 1$, $\Delta x_{T+i-1}$ is given by:

$$\Delta x_{T+i-1} = \gamma_0^* + \alpha_0^* \left( (\beta_0') x_{T+i-2} - \mu_0^* \right) + \Psi_0^* w_{T+i-1} + \epsilon_{T+i-1}. \tag{25}$$

Let $\tilde{u}_{T+i|T+i-1} = \Delta \tilde{x}_{T+i} - \Delta x_{T+i|T+i-1}$, then:

$$\tilde{u}_{T+i|T+i-1} = \gamma_0^* + \alpha_0^* \left( (\beta_0') x_{T+i-1} - \mu_0^* \right) + \Psi_0^* w_{T+i-1} + \epsilon_{T+i}
- \left[ \gamma_0^* + \alpha_0^* \left( (\beta_0') x_{T+i-2} - \mu_0^* \right) + \Psi_0^* w_{T+i-1} + \epsilon_{T+i-1} \right]
= \alpha_0^* (\beta_0') \Delta x_{T+i-1} + \Psi_0^* \Delta w_{T+i} + \Delta \epsilon_{T+i}. \tag{26}$$

The forecasts will be unbiased, so that no systematic failure should result from the ‘constant-change’ forecast:

$$E \left[ \tilde{u}_{T+i|T+i-1} \right] = \alpha_0^* E \left[ (\beta_0') \Delta x_{T+i-1} \right] + \Psi_0^* E \left[ \Delta w_{T+i} \right] + E \left[ \Delta \epsilon_{T+i} \right] = \alpha_0^* (\beta_0') \gamma_0^* = 0. \tag{27}$$

We can also derive an approximate expression for the conditional forecast-error variance, if we neglect covariances, as:

$$V \left[ \tilde{u}_{T+i|T+i-1} \right] = V \left[ \alpha_0^* (\beta_0') \Delta x_{T+i-1} \right] + V \left[ \Psi_0^* \Delta w_{T+i} \right] + V \left[ \Delta \epsilon_{T+i} \right]
= \alpha_0^* (\beta_0') V \left[ \Delta x_{T+i-1} \right] \beta_0^* \alpha_0^* + \Psi_0^* V \left[ \Delta w_{T+i} \right] \Psi_0^* + 2 \Omega_e \tag{27}$$

which is also the mean-square error matrix, given that the forecasts are unbiased.

We have shown that unbiased forecasts result without the forecaster knowing the causal variables or the structure of the economy, or whether there have been any structural breaks or shifts: $\Delta x_{T+i-1}$ reflects all the effects in the DGP. However, there are two drawbacks: the unwanted presence of $\epsilon_{T+i-1}$ in (25) doubles the innovation error variance term in (27); and all variables are lagged one extra period, which adds the ‘noise’ of many $l(-1)$ effects. Comparing (27) to the VEqCM forecast-error variance (23), $\Omega_e$ enters (27) with a factor of two relative to (23), although $V \left[ \Delta w_{T+i} \right]$ in (27) will typically be less than twice $V \left[ w_{T+i} \right]$ unless $\{ w_i \}$ is not positively autocorrelated. Nevertheless, (27) also contains a term in $V \left[ \Delta x_{T+i-1} \right]$, suggesting a clear bias/variance trade-off between using the carefully modelled (19) and the ‘naive’ predictor (24). In forecasting competitions across many states of nature with structural breaks and complicated DGPs, it is easy to see why $\Delta x_{T+i-1}$ may win on MSFE.

### 2.2.3 Intercept corrections

When there are structural breaks, forecasting methods which adapt quickly following the break will often fare best in sequential real-time forecasting. We have reviewed a number of related models that ensure rapid adaptation (after the break has occurred) via some form of differencing. We can show that the widespread use of some macro-econometric forecasting practices, such as intercept corrections (ICs, or residual adjustments), can also be justified by structural change, in that their use works in a similar fashion to differencing: see Clements and Hendry (1996b, 1999a). The importance of adjusting purely model-based forecasts has been recognized by a number of authors: see, *inter alia*, Marris (1954), Theil

Other ways of mitigating the effects of structural change believed to be of a more continual nature are to employ moving windows of data in model estimation, and so exclude distant observations. Alternatively, the model’s parameters may be allowed to ‘drift’. An assumption sometimes made in the empirical macro literature is that VAR parameters evolve as driftless random walks (with zero-mean, constant-variance gaussian innovations) subject to constraints that rule out the parameters drifting into non-stationary regions (see Cogley and Sargent, 2001, 2005, for recent examples).

2.3 Other implications

2.3.1 Model mis-specification and the pooling or combination of forecasts

The obvious way to ameliorate forecast errors emanating from model mis-specification is to adopt better forecasting models – making sure that models use all relevant information. Nevertheless, given the large number of variables that are potentially important and the relatively short historical samples on which to identify stable forecasting models, uncertainty over the appropriate model to use for forecasting is likely to persist. Combinations of forecasts from different models have often been found to outperform the individual forecasts, in the sense that the combined forecast delivers a smaller mean-squared forecast error. This attests to the difficulty of incorporating all the useful information in a single model, as well as being a way in which more accurate forecasts can sometimes be obtained. In addition, Hendry and Clements (2004b) show that pooling may also be justified by the constituent models’ forecasts being differentially susceptible to structural breaks. Recently there has been some interest in Bayesian model averaging (BMA) for forecasting. BMA is a combination approach that combines models’ forecasts with weights given by the posterior probabilities that those models are the true model.

2.3.2 Selecting models for different horizons

The practice of using the same model specification for generating forecasts at all horizons of interest requires that the model is ‘unconditional’ in the sense that all the explanatory variables of the variable of interest are modelled (e.g., as in a vector autoregression). As it may be difficult to model some of the explanatory variables, it may be better to specify models that relate the variable of interest at period $T + h$ directly to the values of explanatory variables dated $T$ or earlier, if the aim is to forecast $h$-steps

---

5 On forecast combination, see inter alia Diebold and Lopez (1996) and Newbold and Harvey (2002) for recent surveys, and Clemen (1989) for an annotated bibliography, as well as contributions by Barnard (1963), Bates and Granger (1969), Newbold and Granger (1974), Chong and Hendry (1986), Granger (1989), Harvey, Leybourne and Newbold (1998), Stock and Watson (1999a), Stock and Watson (1999b) and Clements and Galvão (2006). An alternative route to combining the predictions from a number of models would be to use a factor model (see e.g., Forni, Hallin, Lippi and Reichlin (2000)). Factor models can be interpreted as a particular form of ‘pooling of information’ in contrast to the ‘pooling of forecasts’.

4 As model mis-specification will not necessarily show up as forecast failure (Clements and Hendry, 2002c), comparing a model’s forecasts with those of rival models, or a benchmark, can be a valuable way of evaluating the model’s forecast performance.

Questions then arise as to whether the specification should be kept fixed as \( h \) changes, or whether the model should be re-selected in some way for each \( h \) (as in, say, Phillips, 1995, 1996).

If it were possible to write down a model that was correctly-specified for the data generation process, including the specification of models for the explanatory variables, then these issues clearly would not arise. Given the ‘text-book’ paradigm of forecast errors emanating only from the model’s random disturbance terms and estimation error, the appropriate strategy would be to generate multi-step forecasts from that model by iterating forward the 1-step forecasts to obtain 2-step forecasts, and so on. When the models are poor approximations to the phenomena being modelled, these and related issues need to be considered (see also Clements and Hendry, 2005a).

3 The model of UK inflation

3.1 The data series

Our data set of annual UK time series comprises constant-price GDP (denoted \( Y \)); prices (\( P \), the implicit deflator of \( Y \)); nominal broad money (\( M \)); interest rates (Treasury-bill rate \( R_s \), bond rate \( R_l \), and an opportunity-cost measure, \( R_{os} \), from Ericsson, Hendry and Prestwich, 1998); nominal national debt (\( N \)); employment (\( L \)), unemployment (\( U \)); working population (\( W_{pop} \)); real capital stock (\( K \)); nominal wages (\( W \)); normal hours (\( H \)); external prices (\( P_e \)); and a nominal effective exchange rate (\( E \)): these data series are defined more precisely in the appendix. Capital letters denote the original variables, and lower-case letters the corresponding logarithms (so \( y = \log Y \)). The estimation sample period is usually 1875–1991 after creating lags (the time series for \( W \) ended in 1990). Individual-year indicators were used for turbulent periods, namely the major wars and their after effects (1900–01, 1914–22, 1939–46), the breakdown of Bretton Woods and the two oil crises (1970–71, 1973–75, 1979–80), different intercepts pre and post 1945 (denoted \( I_{2,t} \)), and a switch between 1880–81 (probably due to recording errors): all indicators were zero other than in the years noted.

Although there must be substantial errors of measurement, both conceptual and numerical, in most of these time series over such a long historical period, we treat inflation as an \( I(0) \) variable subject to intermittent breaks, measured with an \( I(0) \) error. However, as annual inflation has varied from \( +25\% \) to almost \(-20\% \), primarily due to wars and their aftermaths, the \( I(0) \) measurement-error effects should be swamped by the large signal.

3.2 Feedback relations

Rather than merely include many levels of non-stationary time series, excess-demand measures were constructed as equilibrium corrections from individual-equation analyses, based on the relevant economic theories. These were then incorporated in the general inflation model described in section 3.3 below. A detailed description of their formulation is provided in Hendry (2001b).

3.2.1 Excess demand for goods and services

The output ‘gap’ (between aggregate demand and supply) plays a central role in macroeconomic analyses of inflation. We used deviations from a measure of potential per capita capacity (\( cap_t \)) which allowed

---

6The data are available on the Journal of Applied Econometrics data archive at http://qed.econ.queensu.ca/jae/
for secular changes in the rate of (non-embodied) technical progress, namely: \(^7\)

\[
cap_t = \begin{cases} 
2.650 + 0.0052t + 0.36(k_t - wpop_t) & 1870 - 1944 \\
1.875 + 0.0135t + 0.36(k_t - wpop_t) & 1945 - 1991.
\end{cases}
\]

Then \(y^d_t = y_t - l_t - \cap_t\) defined the excess demand for goods and services. For convenience, Figure 1 records the resulting time series (panel a).

![Figure 1: Excess demands for output, money, debt, currency, markup and labour](image)

### 3.2.2 Money, interest rates and debt

The excess money demand relation in panel b was based on that in Ericsson et al. (1998):

\[
m^d_t = (m_t - p_t - y_t) + 7.3R_{t,n} + 0.38. \tag{28}
\]

The ‘short-long spread’ \(S_t = R_{s,t} - R_{l,t} + 0.006\) measured relative demands in financial markets, whereas the excess-demand relation for national debt was:

\[
n^d_t = n_t - p_t - y_t + 40S_t + 0.46, \tag{29}
\]

shown in panel c.

\(^7\)These are similar to static regressions of \((y - l)\) on \((k - wpop)\), constant and trend over 1867–1918 and 1945–1991 respectively.
3.2.3 World prices and exchange rates

Let \( p_{\varepsilon,t} = p_{t} - e_{t} \) denote world prices in sterling, then the ‘real exchange rate’ measuring deviations from purchasing-power parity in panel d was:

\[
e_{r,t} = p_{t} - p_{\varepsilon,t} + 0.52. \tag{30}
\]

3.2.4 Unit labour costs

The profit markup, \( \pi_{t}^{*} \), incorporating real commodity prices \( (p_{o} - p)_{t} \), was measured by:

\[
\pi_{t}^{*} = 0.25e_{r,t} - 0.675 (c - p)^{*}_{t} - 0.075 (p_{o} - p)_{t} + 0.11l_{2,t} + 0.25, \tag{31}
\]

where \( c_{t} - p_{t} = w_{t} - p_{t} + l_{t} - y_{t} \) measured real unit labour costs, and \( (c - p)^{*}_{t} \) incorporated falling ‘normal hours’: see panel e.

3.2.5 Wages and unemployment

Disequilibrium unemployment, or the excess-demand for labour measure, \( U_{d}^{d} \), was:

\[
U_{d}^{d} = U_{r,t} - 0.05 - 0.77(R_{l,t} - \Delta p_{t} - \Delta y_{t}),
\]

so \( U_{r,t} \) rises when the real interest rate \( (R_{l,t} - \Delta p_{t}) \) exceeds the real growth rate \( (\Delta y_{t}) \), and falls otherwise: panel f records the resulting series.

3.2.6 Commodity prices

Finally, commodity prices in £ \( (p_{o,t} = p_{o\$,t} - e_{t}) \) were included in the general model.

3.3 The selected model

The initial general unrestricted model (GUM) of \( \Delta p_{t} \) included one lag of each of \( y_{d}^{d}, m^{d}, n^{d}, U^{d}, e_{r}, c, (R_{s} - R_{l}), p, p_{o}, R_{l}, \Delta p_{o,t}, \Delta p_{w}, \Delta U_{r}, \Delta w, \Delta c, \Delta m, \Delta n, \Delta R_{s}, \Delta R_{l}, \Delta p_{o} \), as well as the contemporaneous value of \( \Delta p_{e,t} \) to all the indicators noted above, and a linear trend. This yielded an equation with a standard error of \( \hat{\sigma} = 1.21\% \) for 52 variables, \( SC = -7.3 \) (the Schwarz criterion: see Schwarz, 1978) and no significant mis-specification tests. That model was simplified using PcGets (see Hendry and Krolzig, 2001) and the author’s decisions on how to combine variables (primarily the indicators, which were reduced to an overall index, \( I_{d} \)). The finally selected model was:

\[
\hat{\Delta}p_{t} = 0.27 \Delta p_{t-1} + 0.18 y_{t-1}^{d} + 0.038 I_{d,t} + 0.19 \Delta m_{t-1} - 0.83 S_{t-1} \tag{32}
\]

\[
R^{2} = 0.975 \quad \hat{\sigma} = 1.14\% \quad SC = -8.66 \quad V = 0.20 \quad J = 1.26
\]

\[
\chi_{\text{red}(2)}^{2} = 2.67 \quad F_{\text{arch}(2,104)} = 0.20 \quad F_{\text{reset}(1,107)} = 0.27
\]

\[
F_{\text{het}(18,89)} = 1.03 \quad F_{\text{Chow}(10,98)} = 1.71 \quad F_{\text{red}(42,65)} = 0.72 \quad T = 1875–1991.
\]
In (32), $R^2$ is the squared multiple correlation, $\hat{\sigma}$ is the residual standard deviation, coefficient standard errors are shown in parentheses, $V$ and $J$ are the variance-change and joint parameter-constancy tests from Hansen (1992). The diagnostic tests are of the form $F_j(k, T - l)$ which denotes an approximate $F$-test against the alternative hypothesis $j$ for: $k^{th}$-order serial correlation ($F_{ar}$: see Godfrey, 1978), $k^{th}$-order autoregressive conditional heteroscedasticity ($F_{arch}$: see Engle, 1982), heteroscedasticity ($F_{het}$: see White, 1980); parameter constancy ($F_{Chow}$: see Chow, 1960) over $k^{th}$ periods; all reductions ($F_{red}$: see Hendry and Krolzig, 2001); and a chi-square test for normality ($\chi^2_{nd}(2)$: see Doornik and Hansen, 1994). All diagnostic and constancy tests remained insignificant in (32), and re-estimation using only post-1945 data delivered similar estimates (but a smaller coefficient on $y_{d,t}$.)

Almost all excess demands, or variables associated with every sector or market, matter, often highly significantly. The model suggests rapid adjustment of inflation to shocks, but with all the main turbulent periods remaining as ‘outliers’, captured by $I_{d,t}$.

4 Individual forecasting methods

To generate forecasts from (32) would require forecasts of the many explanatory variables involved, a task not to be lightly undertaken. Thus, we consider five simpler approaches to forecasting the final decades, 1982–1991, and 1972–1981, using multi-step estimation, namely projecting $\Delta p_t$ on lagged values of the relevant set of the regressors (i.e., $h$-periods lagged for $h$-step forecasts). The first model uses the set of regressors in (32) at the longer lags: this is equivalent to (2). The second re-selects a forecasting model by $PcGets$ from the variables in the GUM lagged $h$ periods: see (3) above. The third retains all variables in the GUM lagged $h$ periods—(1). The fourth is the naive predictor $\hat{\Delta} p_{T+h|T} = \Delta p_T$ which has long been a useful and demanding benchmark: see (4). The final approach is to jointly forecast UK and world inflation. We consider these in turn. Note that estimation always ends at 1981 (1971), after which we compute a sequence of $h$-step ahead forecasts for the available horizon: recursive updating might improve over the outcomes recorded here (as discussed in section 2).

4.1 1982–1991

We consider the more quiescent period first and in greater detail, and simply summarize the results for the second period, which was particularly turbulent, and hence (ex post) bound to favour intercept correction methods.

4.1.1 Multi-step estimation of (32) for forecasting

Suitably lagging the regressors in (32), then re-estimating, generated the four sets of estimates shown in appendix 9 equations (40)–(43) (where $\Delta p_{e,t}$ was lagged to formulate (40)). The unmodelled residual autocorrelation created by the multi-step nature of the estimation invalidates conventional standard errors, but those shown for the estimated equations are autocorrelation and heteroscedasticity consistent (see e.g., Andrews, 1991), signified by the brackets. There is a steady increase in the root mean-square error (RMSE) with the length of the horizon for the initial specification, but $\hat{\sigma}$ provides an accurate pre-estimate of RMSE, consistent with the predictive-failure statistic values of $\chi^2_f(10)$ being close to 10 (see e.g., Kiviet, 1986). Had $I_{d,t}$ been omitted, $\hat{\sigma}$ would have substantially over-estimated the RMSE, since no major unanticipated shocks occurred. There is a deterioration in significance for most coefficients.
as the lag length increases, other than the indicator index (so the outliers persist), the spread, and the
growth rate of money. Although $\Delta p_t$ could be zero in principle, this is most unlikely in practice (and
did not occur here), so the mean absolute percentage error (MAPE) is viable, and reflects that only a
little over half of the absolute percentage change in prices is forecast (see Clements and Hendry, 1993,
and the associated discussion for an evaluation of forecast-accuracy measures: the sample here is too
small to sustain their invariant generalized forecast error measure).

Figure 2 records the outcomes for the one through four steps ahead forecasts from (40)–(43) on
a common scale (all error bars show $\pm 2\widehat{\sigma}$). There is a notable increase in that crude measure of un-
certainty, and a deterioration in the closeness of the forecasts to outturns as the horizon grows, but no
evidence of forecast failure. However, the 3- and 4-year ahead forecasts systematically underpredict.

![Figure 2: Forecasts and outcomes with standard error bars for multi-step estimation of (32)](image)

### 4.1.2 Selecting a forecasting model by PcGets

In principle, there is no formal justification for applying PcGets to a non-congruent GUM, since the
selection criteria are invalidated by the unmodelled residual serial correlation inherent in multi-step
procedures. Nevertheless, as an empirical issue, one might learn whether re-selecting improves over the
lagged variants in the previous section. The indicators in the initial GUM were replaced by $I_{d,t}$, since
interest does not centre on modelling the outliers – the forecasts are over a period where $I_{d,t} = 0$. The
selected equations are shown in (44)–(48) in the appendix.

There is a small improvement in the in-sample fit from re-selecting, with several variables dropped
and two others included. The RMSE is smaller, but the MAPE is larger than (40). Going 2-steps ahead
the improvement in RMSE is more marked over (41) suggesting a greater benefit in forecasting after re-selecting than simply retaining the optimal 1-step model lagged.

However, that last effect changes radically at 3-steps ahead. An increasing role attributed to a deterministic trend is responsible for the preservation of fit, but the loss of forecast accuracy. Re-selection when the deterministic trend is excluded from the candidate set delivers a model similar to (42). Consequently, the deterministic trend was now eliminated from the GUM, resulting in (47), which stabilizes the RMSE. This empirical finding is consistent with other evidence on the potentially pernicious nature of deterministic trends in forecasting (see e.g., Clements and Hendry, 2001b, 2001c).

Finally, at 4-steps ahead there is an improvement over (43), but not by a substantial margin. Interestingly, money growth does not enter the models for either 3 or 4-steps ahead, despite being prominent in (40)–(43). Figure 3 reports the forecasts from the four selected models, where the last two excluded the deterministic trend from the initial GUM.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3}
\caption{Forecasts and outcomes with standard error bars for \textit{PcGets} selection}
\end{figure}

### 4.1.3 Forecasting by the GUM

Because of the size of the equations, only summary statistics are reported. At one-step ahead:

\[
\text{RMSE} = 2.41\% \quad \hat{\sigma} = 1.50\% \quad \text{MAPE} = 43.9 \quad \text{SC} = -7.60.
\] (33)

Although the value of \(\hat{\sigma}\) is smaller than either earlier 1-step forecasting models, that for RMSE is much larger, reflecting the disadvantages of retaining a substantial over-parametrization (24 parameters). The value of MAPE has more than doubled compared to (40) and SC is considerably larger.
Next, for 2-step forecasts:

\[
\text{RMSE} = 4.0\% \quad \hat{\sigma} = 1.94\% \quad \text{MAPE} = 56.3 \quad \text{SC} = -7.09. \quad (34)
\]

Now \(\chi^2_f(10) = 42.5\), so the forecasts are very poor relative to the in-sample fit, as the second panel in fig. 4 confirms.

This pattern becomes increasingly clear—the 3-step and 4-step ahead GUM forecasts are dreadful:

\[
\text{RMSE} = 11.2\% \quad \hat{\sigma} = 2.53\% \quad \text{MAPE} = 186.8 \quad \text{SC} = -6.56. \quad (35)
\]

Even if the deterministic trend is dropped, we find:

\[
\text{RMSE} = 4.74\% \quad \text{MAPE} = 69.7 \quad (36)
\]

which, while better, remains awful. At 4-steps ahead again without the trend:

\[
\text{RMSE} = 12.6\% \quad \text{MAPE} = 216.3 \quad (37)
\]

Such evidence is strongly against using the multi-step estimated GUM for multi-step forecasting, consistent with a great deal of evidence on forecast performance of alternative methods: see e.g., Makridakis, Andersen, Carbone, Fildes et al. (1982) reviewed by Fildes and Makridakis (1995) Makridakis and Hibon (2000), Clements and Hendry (2001a), and Fildes and Ord (2002). Figure 4 records the forecast performance for the general GUMs (excluding trends). The poor quality of GUM forecasts is obvious to the eye. The first panel reports both the error bars of \(\pm 2\hat{\sigma}\) and corresponding bands which also incorporate the conventionally-measured parameter estimation uncertainty, showing its large contribution (the unmodelled residual autocorrelation at longer horizons precludes the use of the latter). By way of comparison, there is essentially no difference between such bars and bands for the selected models.

### 4.1.4 Zero-acceleration forecasts

One simple robust predictor is based on the non-accelerating inflation hypothesis, namely \(\hat{\Delta p}_{T+h|T} = \Delta p_T\). It is stretching language to call such a naive extrapolation a ‘forecast’: see Clements and Hendry (1999b) and Hendry (2001a). Nevertheless, when location or trend breaks occur (‘deterministic shifts’ in the language of Clements and Hendry, 1999a), \(\hat{\Delta p}_{T+h|T}\) adapts rapidly, so rarely suffers forecast failure, whereas equilibrium-correction models almost always fail badly in such a setting. The outcomes are as follows:

<table>
<thead>
<tr>
<th>Horizon</th>
<th>RMSE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-step</td>
<td>1.66%</td>
<td>28.6</td>
</tr>
<tr>
<td>2-step</td>
<td>3.89%</td>
<td>49.1</td>
</tr>
<tr>
<td>3-step</td>
<td>4.59%</td>
<td>60.7</td>
</tr>
<tr>
<td>4-step</td>
<td>5.06%</td>
<td>76.9</td>
</tr>
</tbody>
</table>

As might be expected, performance is best at shortest horizons, and deteriorates fairly rapidly as the horizon grows.
4.1.5 Joint forecasting

The original model (32) included the contemporaneous value of $\Delta p_{e,t}$, which was substituted by its lagged value $\Delta p_{e,t-1}$ in (40), and even longer lags for longer horizons. In a sense, (40) could be interpreted as a ‘reduced form’, where a model of the type explored in section 4.1.4 is used for $\Delta p_{e,t}$. Since that device proved relatively successful, little may be thought to be lost by such a substitution. However, macro-econometric modellers tend to forecast ‘external variables’ off-line and then condition ‘endogenous’ forecasts on them. Thus, we also considered jointly forecasting $\Delta p_t$ and $\Delta p_{e,t}$ using (32) – with appropriate longer lags of the other regressors for the former, and a single-order autoregression with an intercept for the latter. The outcomes were as follows:

1-step RMSE = 1.80%  MAPE = 31.5
2-step RMSE = 2.59%  MAPE = 40.2
3-step RMSE = 3.79%  MAPE = 55.6
4-step RMSE = 3.39%  MAPE = 44.8

The outcomes are close to those of the ‘solved’ estimates, but there is a small improvement at the longest horizon, consistent with the hypothesis that a location-shift occurred near the start of the forecast period (see e.g., Clements and Hendry, 1999a, ch. 3).
4.2 Combined forecasting methods

4.2.1 Pooling individual forecasts

Combining or ‘pooling’ individual forecasts of the same event – as suggested by Bates and Granger (1969) – has often been found to be beneficial on a RMSE basis: see Newbold and Harvey (2002) for a recent survey. A simple rule for combination, such as the average, seems to work well: see Fildes and Ord (2002). Hendry and Clements (2004a) consider the theory of ‘pooling’ forecasts when the data generation process is non-stationary from both unit roots and structural breaks, and all available forecasting models are mis-specified. They show that averages of the individual forecasts can outperform the best, and that other forms of weighting can be inferior when location shifts occur. Such results suggest that pooling information – as in (say) Stock and Watson (1999a) – might also deliver benefits. Thus, we consider both forms of combining.

First, the arithmetic average of the five methods reported in section 4 (see table 1 below, denoted ‘pooled’) dominates the best of the individual forecasts for the first two horizons on both RMSE and MAPE, comes second at 3-steps ahead, and only loses at the final horizon. This last outcome is entirely due to the dreadful GUM forecasts: dropping the GUM (a ‘trimmed’ pool, not shown in the table) delivers RMSE = 2.52% and MAPE = 32.8 for \( h = 4 \), which produces significant outperformance relative to any individual method. Indeed, at 3-years ahead, the trimmed mean has RMSE = 2.08% and MAPE = 28.8 so is an outright winner at all horizons. An apparent recommendation is, therefore, to use a trimmed mean, eliminating outlier forecasts.

Hendry and J. (2004) suggest that model averaging in general is prone to potentially serious problems if care is not taken about the set that is averaged across. They offer a ‘counter example’ where the pooled outcome does very badly if all models are included (as happens e.g., in approaches like Eklund and Karlsson, 2004, or BMA as noted above). Selecting by eliminating poor models can substantively improve forecast performance, consistent with the outcome in the previous paragraph.

4.2.2 ‘Factor’ forecasts

The main alternative to combining forecasts is to combine the underlying information, perhaps using a factor model as in Stock and Watson (1999a). To help elucidate the roles of the many different assumptions underlying such an approach, we consider the most obvious combination of all the regressor variables, \( z_{t-1} \), in the present context, namely the coefficient estimates in the GUM using only data up to 1981, \( \hat{\beta}_1 \), say, where all variables are lagged one period. Then, \( \hat{y}_{T+1|T} = \hat{\phi}_h \left( \hat{\beta}_1' z_{T-1} \right) \), holding fixed the coefficient estimates at their 1-step values, where \( \hat{\phi}_h \) is the least-squares in-sample estimate. Given the dreadful performance of the GUM forecasts based on \( \hat{\beta}_h' z_T \) (where the coefficients are re-estimated as the horizon changes), \( \hat{y}_{T+1|T} \) may not seem a promising candidate, but appearances can be deceptive – these are the dominating forecasts for horizons 2–4 (other than the trimmed mean). The main lesson is, perhaps, that if information is to be ‘averaged’, the choice of weights is vital, and those based on inefficient lagged estimates like \( \hat{\beta}_h \) may lose badly.

To investigate whether the benefits of the ‘factor’ approach accrue simply from using the lagged fitted values of the GUM as the forecasting variable, the same procedure was applied using the simplified
model from (40), and fixing its coefficients, \( \tilde{\gamma}' \) say, such that:

\[
\tilde{y}_{T+1|T} = \tilde{\psi}_{h}(\tilde{\gamma}'_1 z_{1,T})
\]

(39)

where all the simplification zero restrictions have been imposed. These are denoted ‘fitted’, and are superior for the first two horizons, then distinctly poorer thereafter, a manifest instance where a more ‘complicated’ model (38) outperforms a simpler (39).

Figure 5 records the forecast performance of the factor model.

Finally, we averaged all seven methods (the previous five with ‘factor’ and ‘fitted’), as well as a ‘trimmed’ mean, and comment on these ‘averages’ in the next section.

5 Summary comparisons

Over the relatively stable sub-period 1982–1991, inflation fell from 7% to 2.5% then rose back to 7.5%. In contrast, over the very unstable period 1972–1981, inflation rose from 9.5% to 24% then fell back to 9.5%. The aim of this section is to contrast the fortunes of the different devices under these different circumstances.
5.1 ‘Forecasting’ 1982–1991

Collecting together all the RMSEs and MAPEs for the various methods and horizons over 1982–1991 yields table 1.

<table>
<thead>
<tr>
<th>Horizon</th>
<th>RMSEs</th>
<th>MAPEs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>(32)</td>
<td>1.8%</td>
<td>2.7%</td>
</tr>
<tr>
<td>selected</td>
<td>1.7%</td>
<td>2.3%</td>
</tr>
<tr>
<td>GUM</td>
<td>2.4%</td>
<td>4.0%</td>
</tr>
<tr>
<td>$\Delta^2_p$</td>
<td>1.7%</td>
<td>3.9%</td>
</tr>
<tr>
<td>joint</td>
<td>1.8%</td>
<td>2.6%</td>
</tr>
<tr>
<td>'pooled'</td>
<td>1.6%</td>
<td>1.6%</td>
</tr>
<tr>
<td>'factor'</td>
<td>2.4%</td>
<td>1.8%</td>
</tr>
<tr>
<td>'fitted'</td>
<td>1.8%</td>
<td>1.7%</td>
</tr>
<tr>
<td>'average'</td>
<td>1.7%</td>
<td>1.6%</td>
</tr>
<tr>
<td>'trimmed'</td>
<td>1.6%</td>
<td>1.5%</td>
</tr>
<tr>
<td>AR($h$)</td>
<td>1.4%</td>
<td>2.2%</td>
</tr>
<tr>
<td>'trim+AR'</td>
<td>1.4%</td>
<td>1.4%</td>
</tr>
</tbody>
</table>

Of the individual methods, the PcGets selection dominates at all horizons in terms of RMSEs, but at 1-year ahead only for MAPEs. The naive predictor is also a strong performer for 1- and 2-years ahead, but fades as the horizon grows: the same outcome has been found using large econometric systems by Eitrheim, Husebø and Nymoen (1999), and is consistent with the theory in Clements and Hendry (1999a).

Of the ‘combination’ methods, the ‘factor’ forecasts do best at the longest two horizons, with the ‘pooled’ outcomes of the individual forecasts better up to 2-years ahead: these dominate overall at those horizons (other than the ‘pooled’ and ‘trimmed’ averages). Averaging all the individual and the two ‘factor’ approaches improves over the ‘pooled’ forecast at longer horizons, and even dominates the better of ‘factor’ and ‘fitted’ up to 2-years ahead.

Finally, ‘trimming’ the overall average, by omitting the GUM, again delivers a substantive improvement: the trimmed mean of the six constrained forecasts is essentially undominated at every horizon on both RMSE and MAPE measures. This evidence is consistent with the explanation of the success of pooling in Hendry and Clements (2004a). If only one forecast is to be used, the overall trimmed average is the most reliable, especially on MAPE. We consider the implications of this evidence in section 6.

Commenting on these, the GUM forecasts are uniformly the worst, yet the ‘factor’ model which uses precisely the same variables does well: instead of estimating the lagged coefficients, it fixes the coefficients at those of the one-lagged model, then applies them to the lagged values. The ‘fitted’ model is similar in formulation, but fixes the coefficients from (40); that delivers an improvement at the first two horizons over the ‘factor’ model, but not at longer horizons. Thus, the GUM seems to capture information of value to longer-term forecasts, which is ‘lost’ in the welter of inefficient estimation at longer lags. However, the success of the ‘factor’ model points to an alternative explanation for the
outcomes in Stock and Watson (1999a): each of averaging information and using 1-step estimates seems beneficial, so their combination is better still. Such an explanation would not depend on ‘a few factors’ actually explaining the bulk of data variance. However, the trimmed mean still dominates overall, and the turbulent period discussed in the next sub-section delivers a very different ranking.

The fitted values from the GUM are close to the actual inflation rate, $\Delta p_t$, so the ‘factor’ forecast values must be close to $\Delta p_{t-h}$. To test that conjecture, an AR($h$) model of the form: $\hat{\Delta} p_{t|t-h} = \gamma_0 + \gamma_1 \Delta p_{t-h}$ was also investigated, as shown on the last line of table 1. At one-step, AR($h$) outperforms on both criteria, but is not particularly impressive otherwise. However, it does ‘beat’ the factor forecasts in several periods. More interestingly, including AR($h$) in the pooling (still trimmed) improves performance in every period yet again.

Figure 6 records the forecast errors of all these devices (except AR($h$) and the resulting pooled forecast) at all horizons, with the trimmed average shown as a broader line. That averaging will be beneficial can be seen from the mix of over and under forecasting across the methods, despite their being merely different functions of the same information set.

5.2 ‘Forecasting’ 1972–1981

The specification of the GUM and the selection of the model in (32) are bound to have been influenced by the historical experience of the 1970s, when inflation was perhaps the major economic problem. Oil prices were included in part as a response to that episode. Consequently, it is stretching the phrase to

![Figure 6: Forecast errors over 1982–1991 from multi-step estimation of all models](image)
call this an investigation of ‘forecasting’ behaviour: consequently, only summary statistics are provided. Nevertheless, the very poor performance of some methods even after the event provides an upper bound on how they might perform *ex ante* in similar circumstances. Equally, a different ranking of methods relative to the previous sub-section provides an illustration of the lack of robustness of some methods to unanticipated location shifts. Note that the indicator variables for later periods were set to zero, but the specifications were otherwise unchanged, all forecasts being based on estimates up to 1971.

Table 2 records the RMSEs and MAPEs for the various methods and horizons over 1972–1981, including one example of an intercept correction. The joint forecasts were not included.

Table 2: RMSEs and MAPEs over 1972–1981

<table>
<thead>
<tr>
<th>Horizon</th>
<th>RMSEs</th>
<th>MAPEs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>(32)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>selected</td>
<td>3.1%</td>
<td>5.9%</td>
</tr>
<tr>
<td>GUM</td>
<td>5.3%</td>
<td>5.3%</td>
</tr>
<tr>
<td>$\Delta^2 p$</td>
<td>5.8%</td>
<td>7.3%</td>
</tr>
<tr>
<td>AR($h$)</td>
<td>6.0%</td>
<td>11.3%</td>
</tr>
<tr>
<td>‘pooled’</td>
<td>5.2%</td>
<td>6.7%</td>
</tr>
<tr>
<td>‘factor’</td>
<td>5.3%</td>
<td>7.0%</td>
</tr>
<tr>
<td>‘fitted’</td>
<td>3.4%</td>
<td>4.9%</td>
</tr>
<tr>
<td>‘average’</td>
<td>3.4%</td>
<td>4.8%</td>
</tr>
<tr>
<td>‘trimmed’</td>
<td>3.5%</td>
<td>5.4%</td>
</tr>
<tr>
<td>‘facIC’</td>
<td>3.6%</td>
<td>4.0%</td>
</tr>
</tbody>
</table>

Of the *individual* methods, the *PcGets* selection no longer always dominates, although it does so in several periods. The most surprising result is how well the GUM performs, dominating for 3-steps ahead on both measures. This is in complete contrast to the quieter period, and difficult to understand given both the lack of robustness and the profligate parameterization. Moreover, despite its robustness, the no-acceleration predictor is relatively poor even for 1- and 2-years ahead. These anomalous results could reflect specific features of the period, but they are certainly food for thought.

In this turbulent period, it can be no surprise that AR($h$) is very poor. However, the ‘combination’ methods (the 5 individual forecasts ‘pooled’, ‘factor’ forecasts and ‘fitted’) also all do poorly, and do not dominate individual methods. Averaging all the individual and factor-type approaches does, however, deliver a more competitive forecast, although it is not always dominant.

As before, ‘trimming’ the overall average, by omitting the worst performer – which here is AR($h$) – again delivers an improvement at some horizons, but is not always dominant even over the untrimmed average.

We now face the opposite of the previous subsection: the GUM forecasts are fair, they form the ‘factors’, yet the latter and the ‘fitted’ model deliver poor forecasts, especially at longer horizons. Most interestingly, the last two both act like AR($h$), consistent with the notion above that factors mimic the lagged dependent variable.

The theory in Clements and Hendry (1999a) suggests that intercept corrections (ICs) should do well when location shifts occur, and that is borne out here (the last line of the table). Only one example
is shown namely for the factor forecasts, but a dramatic improvement results, delivering an almost undominated outcome. (The IC was set in 1971 and not repeated as the forecast intervals increased.)

6 Policy implications

The Monetary Policy Committee (MPC) of the Bank of England has the task of hitting an inflation target of 2.5% p.a. two-years ahead, with a latitude of ±1%. Their main policy tool is the level of the short-term interest rate. The target inflation rate is well below the range observed over the forecast horizon considered here (and tiny compared to the values of the late 1970s), and has a much tighter band than the best of the RMSEs in table 1. However, if the standard deviation of inflation is related to its (absolute) level, so is smaller at lower levels, then all is not lost. For example, if the MAPEs from the ‘trimmed’ forecast remained constant over time for one and two years ahead at around 25–30%, then a mean inflation rate of 2.5% would be consistent with average errors of around 0.6% and 0.8% respectively. Thus, in future research it seems worth testing for such a dependence between level and standard deviation, perhaps using models in the class applied by Richard and Zhang (1996).

The model in (32) has a relatively constant equation standard error, but in part that is because data-based indicators removed the main outliers. If, instead, a first-order autoregression is fitted for $\Delta p_t$ with a GARCH(1,1) error (see e.g., Bollerslev, 1986), highly significant conditional heteroscedasticity is found. The log of the conditional standard deviation is in turn significantly positively related to $\Delta p_{t-1}$. Even so, the solution for $\Delta p_t = 0$ yields a conditional variance in excess of 2.5%.

Conversely, the original model formulation was based on a sample that included the current forecast period, so the results may be somewhat more favourable than can be achieved in an ex ante setting (although the indicators were not re-selected after eliminating $\Delta p_{e,t}$).

The Bank uses a suite of models to inform its policy decisions (see Bank of England, 1999), and while it is unclear if they use a trimmed mean for forecasting inflation, the MPC may ‘average’ across the many (and potentially conflicting) signals they receive. Thus, a test worth conducting is whether a trimmed mean of their different forecasts would have dominated, and if so, what its operating characteristics were.

7 Conclusions

In this chapter, we have provided a detailed illustration of forecasting when there are structural breaks in a substantive context: that of forecasting UK inflation. The forecasting exercise has been focused on five ‘predictions’, which derive from a general theory of forecasting briefly reviewed in section 2. The five predictions were corroborated on the first sub-sample, but the second delivered several anomalous results. First, the empirical forecasting performance of the parsimonious parametrizations strongly dominated that of the profligate GUM. This partly reflects estimation uncertainty, but also could occur due to the inclusion of ‘irrelevant’ variables which suffer location shifts. Secondly, selecting different models for different horizons from the lagged regressor set also led to gains compared to using the whole-sample dominant model with the same lag structure. For forecasting, where there are breaks and models are misspecified, greater adaptability of the forecasting model may often prove beneficial. Thirdly, the simple robust forecasting device did at least as well at 1-step ahead as the (non-robust) whole-sample dominant model, suggesting that there were location shifts in some of the variables immediately before, or over the early part, of the forecast period. This also points to the benefits of an adaptable model, whereby
forecasts soon go ‘back on track’ after a shift. Fourthly, pooling forecasts dominated individual approaches, especially when a trimmed mean was used. Fifthly, the ‘factor’ model forecasts based on the 1-period GUM performed well, suggesting that pooling information can be valuable as well, especially since those dominated over using the fitted values from the simplified model. The dreadful forecasts from the GUM suggest that multi-step re-estimation is too inefficient a method for utilizing additional information. No ‘prediction’ was made about the relative ranking of the joint forecasting device, which in the event proved similar to the ‘solved’ variant of (40). All of these findings are consistent with the hypothesis that a location-shift occurred at or near the start of the first forecast period.

Indicator variables for previous outliers should be included in forecasting equations when there is unlikely to be another exemplar, and excluded otherwise, if the equation standard error is to provide an accurate estimate of the forecast period RMSE, and is to be used in the construction of interval and density forecasts. Although this advice may seem counterintuitive at first sight, excluding indicators leaves a large equation standard error, which will over-estimate the RMSE when there are no further breaks, and conversely for incorrect inclusion when an unanticipated break occurs. Thus, any decision on retaining/excluding indicators should reflect an appraisal of the probability of a shift, even when its sign is unknown so intercept correction is precluded. The optimal decision would be to include an indicator for a future break, but that really requires a crystal ball.

The present MPC target of 2.5% p.a. inflation ±1% would require much more precise forecasts 2-years ahead than any found here on their RMSE performance. However, if the MAPE remained constant as the level of inflation fell, the required accuracy could be attainable using the trimmed mean forecast.

There are a number of recommendations that follow from our results. At a general level, the results serve as a reminder that when there are breaks the ‘best’ model in-sample need not provide particularly accurate forecasts out-of-sample. Specific recommendations are: (a) use a trimmed mean, eliminating outlier forecasts; (b) avoid deterministic trends in forecasting as these are potentially pernicious; (c) carefully appraise the retention of in-sample indicators in forecasting models; (d) for inflation forecasts, check for dependence between the level and standard deviation, and (e) investigate the past performance of the trimmed mean of the real-time forecasting models.

References


lake Consultants Press.


8 Appendix: Data definitions

\[ Y_t = \text{real GDP, £million, 1985 prices} \]  

\[ P_t = \text{implicit deflator of GDP, (1985=1)} \]  

\[ M_t = \text{nominal broad money, £million} \]  
[1], [2]

\[ R_{s,t} = \text{three-month treasury bill rate, fraction p.a.} \]  
[1], [2]

\[ R_{l,t} = \text{long term bond interest rate, fraction p.a.} \]  
[1], [2]

\[ R_{n,t} = \text{opportunity-cost of money measure} \]  
[3]

\[ N_t = \text{nominal National Debt, £million} \]  
[8]

\[ U_t = \text{unemployment} \]  

\[ W_{pop,t} = \text{working population} \]  

\[ U_{r,t} = \frac{U_t}{W_{pop,t}} \text{ (unemployment rate, fraction)} \]  
[4], [5]

\[ L_t = \text{employment (= } W_{pop,t} - U_t) \]  
[4], [5]

\[ K_t = \text{Gross capital stock} \]  

\[ W_t = \text{wages} \]  
[4], [5]

\[ H_t = \text{normal hours (from 1920)} \]  
[6], p.148, [9]

\[ P_{e,t} = \text{world prices, (1985=1)} \]  
[1], [2], [10]

\[ E_t = \text{annual-average effective exchange rate} \]  
[1], [2], [10]

\[ P_{o8,t} = \text{commodity price index, $} \]  
[11]

\[ \Delta x_t = (x_t - x_{t-1}) \text{ for any variable } x_t \]  
–

\[ \Delta^2 x_t = \Delta x_t - \Delta x_{t-1} \]  
–

\[ I_{d,t} = \text{Indicator index} \]  
–

Sources

Notes
Wage index: hourly wage rates prior to 1946, then weekly wage rates afterwards, so the latter were standardized by dividing by normal hours. However, GDP is an annual flow, and the trend rate of decline of hours is about 0.5% p.a. (based on a drop from 56 to 40 between 1913 and 1990, with an additional increase in paid holidays), so unit labour costs were adjusted accordingly.

Exchange rate: annual £/$ rate till 1954, then an annual aggregate of quarterly data on the trade-weighted effective exchange rate, spliced to the £/$ rate in 1955.

World prices: US prices till 1954, then a trade-weighted annual aggregate of quarterly data on the corresponding PPP values, from which the price data were derived and spliced to US prices in 1955.
9 Appendix: Model estimates

9.1 Lagged variants of the selected full-sample model

\[
\Delta \hat{p}_t = 0.25\Delta p_{t-1} + 0.041I_{d,t} + 0.22y_{t-1}^d - 0.14\pi_{t-1}^* - 1.0S_{t-1} \\
+ 0.29\Delta m_{t-1} + 0.62\Delta R_{s,t-1} + 0.06\Delta p_{e,t-1} + 0.06\Delta p_{o,t-1} \\
\]

(40)

\[
\chi^2(10) = 11.9 \quad \text{RMSE} = 1.77\% \quad \hat{\sigma} = 1.60\% \quad \text{MAPE} = 32.4 \quad SC = -7.96
\]

\[
\Delta \hat{p}_t = 0.04\Delta p_{t-2} + 0.049I_{d,t} + 0.33y_{t-2}^d - 0.12\pi_{t-2}^* - 1.6S_{t-2} \\
+ 0.42\Delta m_{t-2} + 0.36\Delta R_{s,t-2} + 0.01\Delta p_{e,t-2} + 0.05\Delta p_{o,t-2} \\
\]

(41)

\[
\chi^2(10) = 12.1 \quad \text{RMSE} = 2.65\% \quad \hat{\sigma} = 2.41\% \quad \text{MAPE} = 42.4 \quad SC = -7.14
\]

\[
\Delta \hat{p}_t = 0.09\Delta p_{t-3} + 0.054I_{d,t} + 0.25y_{t-3}^d - 0.03\pi_{t-3}^* - 1.8S_{t-3} \\
+ 0.37\Delta m_{t-3} + 0.08\Delta R_{s,t-3} - 0.06\Delta p_{e,t-3} + 0.06\Delta p_{o,t-3} \\
\]

(42)

\[
\chi^2(10) = 11.6 \quad \text{RMSE} = 3.54\% \quad \hat{\sigma} = 3.28\% \quad \text{MAPE} = 42.0 \quad SC = -6.53
\]

\[
\Delta \hat{p}_t = 0.12\Delta p_{t-4} + 0.055I_{d,t} + 0.13y_{t-4}^d - 0.08\pi_{t-4}^* - 1.6S_{t-4} \\
+ 0.25\Delta m_{t-4} + 0.36\Delta R_{s,t-4} + 0.02\Delta p_{e,t-4} + 0.05\Delta p_{o,t-4} \\
\]

(43)

\[
\chi^2(10) = 7.57 \quad \text{RMSE} = 3.43\% \quad \hat{\sigma} = 3.97\% \quad \text{MAPE} = 41.0 \quad SC = -6.15
\]

9.2 PcGets forecasting models

\[
\Delta \hat{p}_t = 0.26\Delta p_{t-1} + 0.039I_{d,t} + 0.19y_{t-1}^d - 0.84S_{t-1} + 0.25\Delta m_{t-1} \\
+ 0.62\Delta R_{s,t-1} + 0.08\Delta p_{o,t-1} + 0.002 (p_o - p)_{t-1} + 0.030 (t - \bar{t}) \\
\]

(44)

\[
\text{RMSE} = 1.72\% \quad \hat{\sigma} = 1.55\% \quad \text{MAPE} = 30.5 \quad SC = -8.03
\]

\[
\Delta \hat{p}_t = 0.043I_{d,t} + 0.36y_{t-2}^d - 1.1S_{t-2} + 0.22\Delta m_{t-2} + 0.09\Delta p_{o,t-2} \\
+ 0.05\Delta n_{t-2} + 0.004 (p_o - p)_{t-2} + 0.055 (t - \bar{t}) \\
\]

(45)

\[
\text{RMSE} = 2.34\% \quad \hat{\sigma} = 2.07\% \quad \text{MAPE} = 42.3 \quad SC = -7.48
\]
\[ \tilde{\Delta}p_t = 0.37 \Delta p_{t-3} + 0.046 I_{d,t} + 0.36 y_{t-3}^d - 1.1 S_{t-3} + 0.08 \Delta p_{o,t-3} + 0.65 R_{t,t-3} \\
- 0.030 p_{t-3} - 0.023 (p_o - p)_{t-3} - 0.23 \Delta c_{t-3} + 0.072 (t - \bar{t}) \]

RMSE = 4.71% \( \hat{\sigma} = 2.57\% \) MAPE = 66.4 SC = -6.98 \( (46) \)

\[ \tilde{\Delta}p_t = 0.39 \Delta p_{t-3} + 0.041 I_{d,t} + 0.29 y_{t-3}^d + 0.08 \Delta p_{o,t-3} - 0.22 \pi^*_t \\
- 0.06 (p_o - p)_{t-3} - 0.30 \Delta c_{t-3} - 0.025 n_{t-3}^d + 0.10 \Delta n_{t-3} + 0.26 \]

RMSE = 2.44% \( \hat{\sigma} = 2.62\% \) MAPE = 44.4 SC = -6.94 \( (47) \)

\[ \tilde{\Delta}p_t = 0.63 \Delta p_{t-4} + 0.045 I_{d,t} - 0.036 n_{t-4}^d - 0.20 \pi^*_t + 0.27 \\
- 0.06 (p_o - p)_{t-4} - 0.37 \Delta c_{t-4} + 0.86 \Delta S_{t-4} \]

RMSE = 2.50% \( \hat{\sigma} = 2.93\% \) MAPE = 40.4 SC = -6.78 \( (48) \)