Economic Forecasting: Some Lessons from Recent Research

David F. Hendry
Department of Economics, and
Nuffield College, Oxford

Michael P. Clements*
Department of Economics,
University of Warwick

October 8, 2001

Abstract

This paper describes some recent advances and contributions to our understanding of economic forecasting. The framework we develop helps explain the findings of forecasting competitions and the prevalence of forecast failure. It constitutes a general theoretical background against which recent results can be judged. We compare this framework to a previous formulation, which was silent on the very issues of most concern to the forecaster. We describe a number of aspects which it illuminates, and draw out the implications for model selection. Finally, we discuss the areas where research remains needed to clarify empirical findings which lack theoretical explanations.

Contents

1 Introduction

A forecast is any statement about the future, so economic forecasting is a vast subject. To be really successful at forecasting, one requires a 'crystal ball' that reveals the future: unfortunately, these appear to be unavailable – as the *Washington Times* headlined in relation to the probability of a recession in the USA, 'Never a crystal ball when you need one'.[1] Consequently, we focus on 'extrapolating' from present information using systematic forecasting rules. While many such extrapolative methods do at least exist, they face the difficulty that the future is uncertain – for two reasons. The first is uncertainty where we understand the probabilities involved, so can incorporate these in (say) measures of forecast uncertainty. The second is uncertainties we do not currently understand, and is the more serious problem, particularly in economics where non-stationary behaviour is the norm – as Clements and Hendry (1999a) quote:

> Because of the things we don't know we don't know, the future is largely unpredictable.
> Singer (1997, p. 39)

Empirical models can take into account the effects of earlier events – even though these were unanticipated at the time – and so 'explain' the past quite well. However, new unpredictable events will occur in the future, so the future will always appear more uncertain than the past. Any operational theory of economic forecasting must allow for such contingencies, where any of the data moments (especially levels and variability) of I(0) transformations of economic variables might alter because of changes in technology, legislation, politics, weather, and society: Stock and Watson (1996) document the pervasiveness of structural change in macroeconomic time-series.

Regular persistent changes are now modelled by stochastic trends, so unit roots are endemic in econometric and forecasting models. Structural breaks – defined as sudden large changes, invariably unanticipated – are a major source of forecast failure, namely a significant deterioration in forecast performance relative to the anticipated outcome, usually based on the historical performance of a model. To date, no generic approaches to modelling such breaks have evolved, although considerable effort is being devoted to non-linear models, many of which primarily select rare events. Thus, in practice, economic forecasts end up being a mixture of science—based on econometric systems that embody

---

[1] Robert J. Samuelson, 16 June, 2001, p.A23.

consolidated economic knowledge and have been carefully evaluated—and art, namely judgments about perturbations from recent unexpected events.

The theme of our paper is that recommendations about model types for forecasting, and associated methods, need to be based on a general theory of economic forecasting that has excess empirical content. First, section 2 sketches an earlier theoretical background which can loosely be equated to the 'textbook' treatment. Unfortunately, despite its mathematical elegance and the simplicity of its prescriptions, the evidence against it providing a useful theory for economic forecasting cannot be ignored: see section 2.1. Section 3 then proposes a more viable framework based on Clements and Hendry (1999a), and section 3.1 outlines the underlying forecast-error taxonomy. Proposals based on inducing 'principles' from the experience of forecast successes and failures are discussed in section 4. Ten areas where the new theoretical framework appears to account for the evidence are investigated in section 5: the basis for their selection is that we do not anticipate major changes in those areas. The implications of that theory for model selection are then drawn in section 6. Section 7 considers ten areas where further research remains a high priority, in many instances, already ongoing. Finally, section 8 provides some concluding remarks. The results reported below draw on a number of published (or forthcoming) papers and books. However, the paper does not claim to be complete in any sense, partly because the subject is now advancing rapidly on many fronts.

2 Background

Historically, the theory of economic forecasting has relied on two key assumptions (see e.g., Klein, 1971):

(1) the model is a good representation of the economy, and

(2) the structure of the economy will remain relatively unchanged.

Given these assumptions, several important theorems can be proved, each with many testable implications: see Clements and Hendry (1998) for details and proofs. We refer to this as 'optimality theory' following Makridakis and Hibon (2000).

First, forecasts from such models will closely approximate the conditional expectation of the data, so the 'best' model generally produces the best forecasts. This entails that an in-sample congruent encompassing model will dominate in forecasting. Moreover, for example, the only judgements that should improve forecasts are those based on advance warnings of events to come (such as notice of future tax changes or strikes). Further, it should not pay to pool forecasts across several models – indeed, pooling refutes encompassing – and adding biased forecasts or those from a badly-fitting model should merely serve to worsen (say) mean-square forecast errors (MSFEs).

Second, forecast accuracy should decline as the forecast horizon increases because more innovation errors accrue and predictability falls. Interval forecasts calculated from in-sample estimates reflect this property.[2]

Third, in-sample based interval forecasts should be a good guide to the likely variations in the forecast errors. Monte Carlo simulation evidence from studies embodying the two assumptions corroborate this finding (see *inter alia*, Calzolari, 1981, and Chong and Hendry, 1986).

Given such a strong foundation, one might anticipate a successful history of economic forecasting. The facts are otherwise.

---

[2]Such confidence intervals from dynamic models need not be monotonically non-decreasing in the horizon, but this is a technical issue – see Chong and Hendry (1986).

2.1 The failure of 'optimality theory'

Unfortunately, empirical experience in economic forecasting has highlighted the poverty of these two assumptions. Such an outcome should not be a surprise: all econometric models are mis-specified, and all economies have been subject to important unanticipated shifts: for example, Barrell (2001) discusses six examples of endemic structural change since the 1990s. Also, Clements and Hendry (2001e) seek to ascertain the historical prevalence of forecast failure in output forecasts for the UK, and any association of such poor forecasts with major 'economic events'. Since the future is rarely like the past in economics, forecast failure has been all too common.

There is a vast literature evaluating the forecast performance of models. Early forecast-evaluation exercises compared econometric model forecasts to those of naive time-series models such as 'no-change' predictors: see, e.g., Theil (1966), Mincer and Zarnowitz (1969), Dhrymes *et al.* (1972), and Cooper and Nelson (1975) with findings that were not favourable to the large econometric systems. More recently, Wallis (1989) and McNees (1990) survey UK and US evidence respectively, although the former concludes that 'published model forecasts generally outperform their time series competitors' (p.46).[3] In the assessment of the track record of the UK Treasury by a long-serving Chief Economic Advisor, Burns (1986) saw little improvement in forecast accuracy over time, despite substantive improvements in the underlying models.

The major empirical forecasting competitions, such as Makridakis, Andersen, Carbone, Fildes *et al.* (1982) reviewed by Fildes and Makridakis (1995), produce results across many models on numerous time series that are inconsistent with the implications of the two assumptions above: see Clements and Hendry (1999b) and section 5.7. Although which model does best in a forecasting competition depends on how the forecasts are evaluated and what horizons and samples are selected, 'simple' extrapolative methods tend to outperform econometric systems, and pooling forecasts often pays. While Stock and Watson (1999a) strongly confirm the latter finding, they report that 'simple' methods, such as non-change forecasts' do relatively poorly.

Even within the present generation of equilibrium-correction economic forecasting models, there is no evidence that the 'best' in-sample model is the best at forecasting, as shown by the results in Eitrheim, Husebø and Nymoen (1999). Those authors find that at short horizons (up to four quarters), badly-fitting extrapolative devices nevertheless outperform the Norges Bank econometric system, although the Norges Bank model 'wins' over longer horizons (12 quarters ahead) because the greater forecast-error variances of the simpler devices offset their smaller biases.

The final conflicting evidence is that 'judgement' has value added in economic forecasting (see Turner, 1990, and Wallis and Whitley, 1991). One might surmise that forecasters have 'fore-knowledge' which contributes to that finding, but the wide-spread use of intercept corrections which 'set the model back on track' (i.e., ensure a perfect fit at the forecast origin) suggests that existing estimated macro-economic models do not provide a good approximation to the conditional expectation over the forecast horizon. The next section explores the consequences of abandoning these two assumptions, and instead allowing that models are mis-specified for the data generation process (DGP), and that the DGP itself changes.

3 A more viable framework

The forecasting theory in Clements and Hendry (1999a) makes two matching, but far less stringent, assumptions:

---

[3]Notice that 'published forecasts' often embody judgemental adjustments.

(1) models are simplified representations which are incorrect in many ways, and

(2) economies both evolve and suddenly shift.

In this more realistic setting, none of the theorems discussed in section 2 hold. Section 3.1 lists the set of potential sources of forecast error and their likely consequences, but concludes that shifts in deterministic terms (intercepts and linear trends) are the major source of forecast failure. When such shifts occur, the best model in-sample need not produce the best forecasts. Further, pooling of forecasts may pay dividends by averaging offsetting biases. Also, longer-term forecasts may be more accurate than short-term ones (however, see Stock, 1996, for an analysis where unit-root effects dominate this effect). Judgement – or at least one class of intercept corrections – can improve forecasting performance. Finally, calculated confidence intervals can be seriously misleading about actual forecast uncertainty.[4] Thus, almost the opposite implications hold compared to the previous theory – and these now do match empirical findings. In particular, since differencing lowers the degree of a polynomial in time by one degree, intercepts and linear trends are eliminated by double differencing, so such devices might be expected to perform well in forecasting despite fitting very badly in-sample.

The following simple example illustrates the virtues of differencing.[5] Suppose the mean of a process changes at $t = \tau$ in an otherwise stationary model, where ε_t may be serially correlated:

$$y_t = \mu_1(1 - 1_T^\tau) + \mu_2 1_T^\tau + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{1}$$

where the indicator variable $1_{\tau+j}^\tau = 1$ for $t \in [\tau, \tau + j]$ and is zero otherwise. If the change at time τ is not modelled, then in terms of first differences:

$$\Delta y_t = \mu_1 \Delta(1 - 1_T^\tau) + \mu_2 \Delta 1_T^\tau + \Delta\varepsilon_t$$
$$= (\mu_2 - \mu_1)1_\tau^\tau + \Delta\varepsilon_t, \tag{2}$$

and the first term will add to the residual, so over the sample as a whole, there will not be a redundant common factor of $(1 - L)$. The residuals are likely to be negatively autocorrelated in the absence of any dynamic modelling, offset by any original positive autocorrelation in the $\{\varepsilon_t\}$. So, the expected level of y_t changes from μ_1 to μ_2 at time τ, but the break produces only one non-zero blip of $\mu_2 - \mu_1$ at τ in the first difference. It is easy to see that forecasts based on the first difference specification will be robust to shifts. Write the h-step ahead forecast of the level of $\{y\}$ from period t as $\widehat{y}_{t+h|t} = \widehat{\Delta y}_{t+h|t} + \widehat{y}_{t+h-1|t}$, i.e., the forecast of the change plus the forecast level in period $t+h-1$. Suppose $h = 1$ so that $\widehat{y}_{|t} \equiv y_t$, then for $t \geq \tau$, $\mathsf{E}[\widehat{y}_{t+1|t}] = \mu_2 = \mathsf{E}[y_{t+h|t}]$ (because $\mathsf{E}[\widehat{\Delta y}_{t+1|t}] \simeq 0$) proving unbiasedness. This result generalizes for $h > 1$ by a recursive argument. As Osborn (2001) notes, the non-invertibility of the error term in the first-differenced model suggests that empirically lags are likely to be added to mop up the serial correlation, which will lessen the 'adaptability' of the model. Nevertheless, it is evident that estimating (1) with an assumed constant mean will generate biased forecasts to an extent that depends upon $\mu_2 - \mu_1$ and the timing of τ relative to the forecast origin.

Fundamentally, causal variables (variables that actually determine the outcome) cannot be proved to help a model's forecasts. After a shift, a previously well-specified model may forecast less accurately than a model with no causal variables. This result helps explain the rankings in forecast competitions. The best causal description of the economy may not be robust to sudden shifts, so loses to more adaptive models for forecasting over periods when shifts occurred. Also, pooling can be beneficial because different models are differentially affected by unanticipated shifts. Further, a levels shift can temporarily

---

[4]Nevertheless, we welome the rapid increase in reporting of interval forecasts as an addition to point forecasts, and the use of (e.g.) fan charts to represent uncertainty: Ericsson (2001) provides an exposition.

[5]Clements and Hendry (1997) and Osborn (2001) provide analyses of differencing for seasonal data.

contaminate a model's short-term forecasts, but the effects wear off, so earlier longer-term forecasts of growth rates can be more accurate than 1-step ahead forecasts made a few periods after a shift. Thus, explanations of the empirical results are provided by the more general framework. By itself that does not preclude alternative explanations, so section 3.1 investigates whether other potential sources of forecast errors could account for the evidence.

3.1 A forecast-error taxonomy

Clements and Hendry (1998, 1999a) derive the following nine sources of forecast error as a comprehensive decomposition of deviations between announced forecasts and realized outcomes:

  (1) shifts in the coefficients of deterministic terms,
  (2) shifts in the coefficients of stochastic terms,
  (3) mis-specification of deterministic terms,
  (4) mis-specification of stochastic terms,
  (5) mis-estimation of the coefficients of deterministic terms,
  (6) mis-estimation of the coefficients of stochastic terms,
  (7) mis-measurement of the data,
  (8) changes in the variances of the errors, and
  (9) errors cumulating over the forecast horizon.

Any one – or combinations – of these nine sources could cause serious forecast errors. However, theoretical analyses, Monte Carlo simulations, and empirical evidence all suggest that the first source is the most pernicious, typically inducing systematic forecast failure. Clements and Hendry interpret shifts in the coefficients of deterministic terms as shifts in the deterministic terms themselves, so the next most serious problems are those which are equivalent to such deterministic shifts, including the third, fifth and seventh sources. For example, omitting a linear trend or using a biased estimate of its coefficient are equivalent forms of mistake, as may be data mis-measurement at the forecast origin in models where such a measurement error mimics a deterministic shift.

Conversely, the other sources of forecast error have less damaging effects. For example, even quite large shifts in the coefficients of mean-zero stochastic variables have small effects on forecast errors: see Hendry and Doornik (1997), Hendry (2000b) and section 5.6. The last two sources in the taxonomy certainly reduce forecast accuracy, but large stochastic perturbations seem needed to precipitate systematic forecast failure.

The 'optimality' paradigm discussed in section 2 offers no explanation for observed forecast failures. At various stages, bad forecasts have been attributed (especially in popular discussions, such as the Press) to 'mis-specified models', 'poor methods', 'inaccurate data', 'incorrect estimation', 'data-based model selection' and so on, without those claims being proved: the research in Clements and Hendry (1999a) demonstrate the lack of foundation for most such 'explanations', whereas the sources follow as discussed above.

4 'Principles' based on empirical forecast performance

Allen and Fildes (2001) thoroughly review the empirical evidence on the practical success of different approaches to economic forecasting based on econometric models. They find that models which are admissible reductions of VARs that commenced with relatively generous lag specifications, estimated by least squares, and tested for constant parameters do best on average. Thus, their conclusions are consistent with the theory implications of the previous section.

They regard the following as issues that remain unresolved from past performance:

(1) the role of causal variables, particularly when such variables are forecast by auxiliary models;
(2) whether congruent models outperform non-congruent, and hence:
(3) whether there is value-added in mis-specification testing when selecting forecasting models; and
(4) whether cointegration restrictions improve forecasts.

However, all four of their unresolved issues have no generic answer: Clements and Hendry (1999a) show that under the assumptions of section 3, causal variables cannot be proved to dominate non-causal; that congruent models need not outperform non-congruent, so rigorous mis-specification testing need not help for selecting forecasting models; and that equilibrium-mean shifts induce forecast failure, so cointegration will improve forecasting only if the equilibrium means remain constant. Conversely, if an economy were reducible by transformations to a stationary stochastic process, so unconditional moments remained constant over time, then well-tested, causally-relevant congruent models which embodied valid restrictions would both fit best, and by encompassing, dominate in forecasting on average. Depending on the time periods examined, and the behaviour of the data therein, either state of nature might hold, so ambiguous empirical findings can emerge.

Against this background, we now evaluate ten areas where explanations can be offered consistent with the empirical evidence.

5 Ten areas of understanding

Here we consider:

(1) the occurrence of forecast failure (5.1);
(2) the role of causal models in forecasting (5.2);
(3) using intercept corrections to offset deterministic shifts (5.3);
(4) unit roots and cointegration (5.4);
(5) model selection and 'data mining' (5.5);
(6) deterministic shifts versus other breaks (5.6);
(7) explaining the outcomes of forecasting competitions (5.7);
(8) the role of simplicity in forecasting (5.8);
(9) evaluating forecasts (5.9);
(10) the behaviour of difference-stationary versus trend-stationary models (5.10).

We consider these in turn in the following sub-sections.

5.1 Accounting for forecast failure

The ingredients have all been laid out above: in the language of Clements and Hendry (1999a), deterministic shifts or their equivalent are the primary culprit. In fact, the widespread use of cointegration-based equilibrium-correction models (EqCMs) in macro-econometric forecasting may have increased their sensitivity to deterministic shifts, particularly in equilibrium means. An upward shift (say) in such a mean induces a 'disequilibrium' which the model is designed to remove, by adjusting in the opposite direction, hence the forecasts will be for a decline precisely when the data show a jump up, and conversely. An example is provided in Hendry and Doornik (1997). Clements and Hendry (1996a, 2001b) provide comprehensive discussions.

5.2 The role of causal models in forecasting

In part, section 6 will address this issue, but here we record the two most salient aspects. In 'normal times' when there are no large deterministic shifts or their equivalent, then a congruent encompassing model will dominate both in-sample and over the forecast period. Unfortunately, as the old joke goes, the last 1000 years have been an exceptional period, and there is every likelihood that the future will see more large, unanticipated shocks – indeed, the recent collapse of the telecoms industry is a reminder that new uncertainties occur. Consequently, causal models cannot be relied on to dominate out of sample. Application of the forecast-error taxonomy to a vector EqCM, VAR and VAR in differences in Clements and Hendry (1999a) reveals that they suffer equally on average when a break occurs after forecasts are produced, so the possibility of future unanticipated breaks is not an argument against causal models or in favour of more adaptive devices. Thus, causal model could maintain a major role if they could be made more adaptive to breaks, a theme explored in the section 5.3.

5.3 Intercept corrections

A potential solution to deterministic shifts is intercept correction (IC), adjusting an equation's constant term when forecasting, usually based on realized equation errors immediately prior to the forecast origin. Historically, IC has been heavily criticized, sometimes named con factor, cheat term, or *ad hoc* adjustment. However, one basis for the value-added from ICs is the result that models with no causal variables might outperform those with correctly included causal variables: ICs are certainly non-causal (though they might proxy for unmodelled causal factors), so the issue is whether they are the right type of non-causal factor. Clements and Hendry (1996a, 1999a) formally establish that when the DGP is susceptible to structural breaks, forecasts made in ignorance of such changes having recently occurred can be improved by ICs which reflect, and so offset, deterministic shifts that would otherwise swamp useful information from causal factors. Clements and Hendry (1996a) also show that ICs can reduce the biases of forecasts from vector EqCMs (VEqCMs) when there are shifts in deterministic terms, provided the change has occurred prior to the forecast origin.

Since ICs offset deterministic shifts after they have occurred, they should almost always be implemented if such shifts are suspected, especially following recent forecast failure. Thus, although a pessimistic result, the refutation of the claim that causal models should outperform non-causal models is an important step towards understanding the actual behavior of economic forecasts, and the value-added of judgement therein. Nevertheless, any forecast-error bias reduction is typically achieved only at the cost of an increase in the forecast-error variance, so that in an unchanged world, for example, the indiscriminate use of ICs may adversely affect accuracy measured by squared-error loss. This suggests a more judicious use of ICs, and in particular, making the decision conditional on the outcome of a pre-test for structural change, the topic of section 7.1.

5.4 Unit roots and cointegration

Current best practice in econometrics uses the technique of cointegration to remove another major source of non-stationarity, that due to stochastic trends or unit roots; see Hendry and Juselius (2000, 2001) for recent expositions. Unfortunately, cointegration makes the resulting models sensitive to shifts in their equilibrium means. Hansen and Johansen (1998) describe tests for constancy in a VEqCM, and Johansen, Mosconi and Nielsen (2000) consider cointegration analysis in the presence of structural breaks in the deterministic trend.

There are several potential solutions to offsetting the detrimental impact of equilibrium-mean shifts on forecast accuracy, although there is little hard evidence on their relative efficacy to date.

First, Clements and Hendry (1995) show that neglecting possible long-run relations between the variables should be relatively benign, unless one wishes to forecast linear combinations of the variables given by those long-run relations. The existence of cointegration matters, since some combinations become I(0), but its imposition seems less important. Instead, ignoring cointegration but analyzing differences may be beneficial if the means of the cointegrating relations are non-constant: see Clements and Hendry (1996a). Thus VARs in growth rates offer some protection, or 'robustification', against non-constancy relative to vector equilibrium-correction models.

Second, ICs can be used in VEqCMs to make them more robust to such shifts, and given the pernicious consequences of equilibrium-mean shifts, doing so becomes a priority. Thus, when I(0) transformations need not be stationary, cointegration is most useful as a modelling device, rather than a method of improving *ex ante* forecasting.

5.5 Model selection or 'data mining'

Clements and Hendry (2001f) investigate the impact of model-selection strategies on forecast performance. They examine both constant and non-constant processes, using restricted, unrestricted and selected models, when the DGP is one of the first two. Thus, they avoid biasing the outcome in favour of always using the simplest model, which happens to work because it coincides with the DGP. Moreover, the non-constancies can occur in an irrelevant variable that was nevertheless included through 'data mining'. Their selection strategy is general-to-specific (*Gets*), and they find no evidence that *Gets* induces significant over-fitting, nor thereby causes forecast-failure rejection rates to greatly exceed nominal sizes. Parameter non-constancies put a premium on correct specification, but in general, model-selection effects appear to be relatively small, and progressive research is able to detect the misspecifications considered.

5.6 Deterministic shifts versus other breaks

There exists a vast literature on testing for structural breaks or non-constancies: see for example, Hansen (2001). From a forecasting perspective, breaks that occur towards the end of the estimation period are of primary interest, although unmodelled breaks anywhere in the series may affect the ability to detect more recent breaks.

Hendry (2000b) finds that structural breaks which leave the unconditional expectations of the I(0) components unaltered in I(1) cointegrated time series are not easily detected by conventional constancy tests. Thus, dynamics and adjustment speeds may alter without detection. However, shifts in long-run means are generally easy to detect. Using a VEqCM model class, he contrasts the ease of detection of 'deterministic' and 'stochastic' shifts by Monte Carlo.

5.7 Explaining the results of forecasting competitions

The major forecasting competitions involve many hundreds of time series and large numbers of forecasting models. Makridakis and Hibon (2000) record the latest in a sequence of such competitions, discussed by Clements and Hendry (2001c). Fildes and Ord (2001) consider the role such competitions have played in improving forecasting practice and research, and conclude that the four main implications of forecasting competitions are:

(a) 'simple methods do best';

    (b) 'the accuracy measure matters';

    (c) 'pooling helps'; and

    (d) 'the evaluation horizon matters'.

The explanation for the four findings in (a)–(d) has three facets. The first facet is that economies are non-stationary processes which are not reducible to stationarity by differencing, thereby generating moments that are non-constant over time. The second facet is that some models are relatively robust to deterministic shifts, either by transforming their impact into 'blips' or by adapting rapidly to them. The third facet is that most measures of forecast accuracy are not invariant under data transformations (see Clements and Hendry, 1993).

We comment in section 5.8 on (a), but the combination of the first two facets is the key. The third directly explains (b). However, (c) remains to be analytically modelled for general settings: see section 7.5. Finally, the evaluation horizon matters for all three reasons, because no method can be robust to breaks that occur after forecasts are announced, so the shorter the horizon when breaks are intermittent, the more that favours robust devices. This also appears to explain the findings in Eitrheim *et al.* (1999).

5.8 Simplicity in forecasting

An unfortunate confusion which has resulted from the findings of forecasting competitions is that 'simpler models do better': see e.g., Kennedy (2001). The source of the successful approaches is their adaptability (primarily to shifts in intercepts and trends), not their simplicity *per se*: Clements and Hendry (1999b, 2001c) explain why. It just happens that, to date, many adaptive models have been simple. Important examples include exponentially-weighted moving averages (EWMAs), or double-differenced devices ('same change' forecasts, or $\Delta^2 \widehat{y}_{T+1} = 0$). A linear deterministic trend $\widehat{y}_{T+1} = a + b\,(T+1)$ is a simple model which does badly in forecasting (see section 5.10), so simplicity alone is not the relevant criterion. An important implication of the finding that adaptability dominates verisimilitude in forecasting competitions is that *ex ante* forecasting comparisons should not be used to evaluate models (except for forecasting): see section 6.1.

5.9 Evaluating forecasts

Forecast evaluation has long been based on statistical criteria, from examining moments – such as forecast biases and variances – through tests for efficiency and the related literature on forecast encompassing, to comparisons between different forecasting devices as discussed above. More recently, attention has turned to the evaluation of forecasts when they are instrumental in decision taking, so an explicit loss function for forecast errors defines the costs: see Granger (2001), Granger and Pesaran (2000a, 2000b), and Pesaran and Skouras (2001). Consequently, the choice of forecasts depends on their purpose, as represented by the loss function, rather than just a statistical criterion: it seems natural that a stock broker measures the value of forecasts by their monetary return, not their MSFE. This development also removes the ambiguity of evaluation based on (say) MSFE measures due to their lack of invariance under linear transformations when the outcome is from a multi-variate or multi-horizon forecasting exercise (see Clements and Hendry, 1993).

A related topic is the increased focus on density forecasting, where the complete probability distribution of possible future outcomes is forecast: see Tay and Wallis (2001) for a survey, Clements and Smith (2000b) for a multi-step application comparing linear and non-linear models, and Diebold, Gunther and Tay (1998) for the role of density forecasting in decision taking. Earlier reporting of forecast-error means and variances only corresponded to a complete characterization of their density for

normal distributions. Most calculated 'fan charts' correspond to that scenario, but that is not an inherent feature, and asymmetric confidence intervals are discussed in Hatch (2001) and Tay and Wallis (2001) (which they call 'prediction intervals').

The final aspect we note is that conditional heteroscedasticity may entail changing widths of forecast confidence intervals, induced by (say) autoregressive (ARCH: see Engle, 1982), general, or related error processes (GARCH: see e.g., Bollerslev, Engle and Nelson, 1994), stochastic volatility (see *inter alia*, Kim, Shephard and Chib, 1998), or inherent in the model specification (see e.g., Richard and Zhang, 1996). Granger, White and Kamstra (1989) and Christoffersen (1998) consider forecast confidence interval evaluation for 'dynamic' intervals (that reflect the changing volatility of the ARCH-type process) and Clements and Taylor (2000) consider methods appropriate for high-frequency data that exhibit periodic volatility patterns.

5.10 Difference-stationary versus trend-stationary models

Difference-stationary (DS) and trend-stationary (TS) models have markedly different implications for forecasting when the properties of each are derived as if it were the DGP: see Sampson (1991). However, such a state of nature can never be actualized: only one model can be the DGP. Consequently, Clements and Hendry (2001d) examine forecasting with the two models when the DGP is in turn either DS or TS, so that the other model is then mis-specified. They consider known and estimated parameters, letting the relation between the estimation sample, T, and the forecast horizon h vary. For known parameters, when a TS process is the DGP, the forecast-error variances of both models are $O(1)$; and when the DS process is the DGP, both are $O(h)$. Thus, the apparently very different properties of the models is due purely to the behaviour of the DGPs: given the DGP, the models actually have similar behaviour. With parameter-estimation uncertainty in the TS DGP, both models' forecast-error variances increase as the square of the horizon for fixed T, the DS/TS variance ratio goes to infinity as T increases – but less quickly than h – whereas for faster rates of increase of T, the ratio converges to 2. For the DS DGP, both the TS and DS models' variances are of the same order: only when T increases at a faster rate than h does the order of the TS model variance exceed that of the DS model. Their Monte Carlo simulations corroborated these results, as well as the serious mis-calculation of the forecast confidence intervals when the other model is the DGP.

In terms of section 3, when deterministic shifts occur, the DS model is considerably more adaptive than the TS, which rapidly produces systematic forecast failure, exacerbated by the calculation of its interval forecasts being far too narrow in I(1) processes: see Hendry (2001).

6 Implications for model selection

A number of important implications follow from the corroboration of the general forecasting theory in section 3 by the evidence presented above. Here we focus on the role of forecasts in selecting econometric models in section 6.1; the implications of forecast-error taxonomies in section 6.2, the role of forecasts in selecting policy models in section 6.3, and impulse-response analyses in section 6.4.

6.1 The role of forecasts in econometric model selection

Forecasting success is not a good index for model selection (other than for forecasting), and certainly should not be used for selecting policy models, a theme explored further in section 6.3. The *raison d'être* of developing rigorously tested, congruent and encompassing econometric systems is for policy

analysis, not forecasting. Second, forecast failure is equally not a ground for model rejection (with the same *caveat*). Consequently, a focus on 'out-of-sample' forecast performance to judge models, usually because of fears over 'data-mining', is unsustainable (see, e.g., Newbold, 1993, p.658). In any case, as section 5.5 shows, data-based model selection does not seem likely to explain forecast failure. Thus, although some failures are due to bad models, and some successes occur despite serious mis-specification, the observation of failure *per se* merely denotes that something has changed relative to the previous state, with no logically valid implications for the model of that state.

Nor do the above findings offer any support for the belief that a greater reliance on economic theory will help forecasting models (see, e.g., Diebold, 1998), because that does not tackle the root source of forecast failure. Instead, a realistic alternative is to construct forecasting models which adapt quickly after any shift is discovered, so that systematic forecast failure is avoided. This involves re-designing econometric models to capture some of the robustness characteristics of the models that win forecasting competitions. As presaged above, one possible approach is to intercept correct a 'causal' model's forecasts, an issue also addressed in section 7.1.

6.2 Implications of the forecast-error taxonomy

The general 'non-parametric' forecast-error taxonomy presented in Hendry (2000a) formalizes that in section 3.1, and confirms the conclusions reached in that section. Since causally-relevant variables cannot be proved to out-perform non-causal in forecasting, the basis is removed for asserting that agents expectations should be 'rational', namely coincide with the conditional expectation of the variable at the future date. While agents may well have access to all the relevant information, they cannot know how every component will enter a future joint data density which changes in unanticipated ways.

An obvious alternative is that agents use the devices that win forecasting competitions. If so, by automatically adjusting to movements in the policy variables, their forecasts may be invariant to changes in policy rules, matching the absence of empirical evidence supporting the Lucas (1976) critique (see Ericsson and Irons, 1995). Conversely, econometric models which embodied data-based proxies for such agents' prediction rules would also prove resilient to policy-regime shifts.

6.3 The role of forecasts in selecting policy models

Hendry and Mizon (2000b) note that 'the policy implications derived from any estimated macro-econometric system depend on the formulation of its equations, the methodology used for the empirical modelling and evaluation, the approach to policy analysis, and the forecast performance'. They criticize current practice in all four areas, but in this section, we are primarily concerned with the role of forecast performance in selecting policy models, about which they draw two main conclusions:

- being the 'best' forecasting model does not justify its policy use; and
- forecast failure is insufficient to reject a policy model.

The first holds because the class of models that 'wins' forecasting competitions is usually badly mis-specified in econometric terms, and rarely has any implications for economic-policy analysis, lacking both target variables and policy instruments. Moreover, intercept corrections could improve forecast performance without changing policy advice, confirming their argument. The second holds because forecast failure reflects unanticipated deterministic shifts, which need not (but could) affect policy conclusions. Thus, neither forecast success nor failure entails either good or bad policy advice: policy models need policy evaluation.

Since shifts in policy regimes correspond to post-forecasting breaks for extrapolative devices, Hendry and Mizon (2000a) note that neither econometric models nor time-series predictors alone are adequate, and provide an empirical illustration of combining them.

6.4 Impulse-response analyses

The difficulty of detecting shifts in policy-relevant parameters has adverse implications for impulse-response analyses. Many vector autoregressions (VARs) are formulated in the differences of economic variables, so changes in their intercepts and coefficients may not be detected even if tested for. In such a state of nature, full-sample estimates become a weighted average of the different regimes operating in sub-samples, so may not represent the correct policy outcomes. Thus, the very breaks that are least harmful in forecasting are most detrimental for policy advice. Since Hendry and Mizon (2000b) also list a range of well-known problems with impulse-response analyses, it is clear that more reliable approaches are urgently required.

7 Ten areas in need of improved understanding

Ten inter-related areas where analytical insights may yield substantial benefits are:

(1) pre-testing for the inclusion of intercept corrections (7.1);
(2) modelling shifts (7.2);
(3) inter-forecast smoothing (7.3);
(4) the role of survey information in forecasting (7.4);
(5) pooling of forecasts and forecast encompassing (7.5);
(6) discriminating measurement errors from innovation shifts (7.6);
(7) multi-step estimation for multi-step forecasting (7.7);
(8) the advantages of explicitly checking co-breaking for forecasting (7.8);
(9) attempts to forecast rare events (7.9); and the closely related issue of:
(10) leading indicators in forecasting (7.10).

We consider these in turn.

7.1 Pre-testing for intercept corrections

In real time, sequences of forecasts are made from successive forecast origins, for each of which the constancy of the model may be questioned, and various actions taken in the event of rejection. In this section we consider 'model adaptation' using the simple expedient of an intercept correction, where the issue of interest is whether pre-testing for a break can yield gains relative to their blanket application or, conversely, no intervention.

Clements and Hendry (2001g) take just this set up, that is, forecasting is an ongoing venture, and series of 1 to h-step ahead forecasts are made at each of a sequence of forecast origins. The 'historical sample' lengthens by one observation each time the forecast origin moves forward, so the possibility of testing for structural change, and the action to be taken if it is detected, arises afresh. One testing strategy is the repeated application of one-off tests for structural change. Alternatively, the sequential testing procedures of Chu, Stinchcombe and White (1996) monitor for structural change as new observations accrue. The overall size of a sequence of repeated tests will approach unity as the number of applications (i.e., forecast origins) goes to infinity, whereas the Chu *et al.* (1996) sequential CUSUM test has the correct asymptotic size by construction. Whether or not it is costly to falsely reject constancy will in

part depend on the form of the intervention to be made, but it is also likely that the sequential tests will lack power when breaks do occur. A full investigation needs to be undertaken – here we report an example based on the repeated application of one-off tests. .

One possible strategy is automatic IC, whereby at each forecast origin, the forecasts are set 'back on track', by making a correction to the equations' intercepts based on the most recently observed errors. Such a strategy is implemented by augmenting the model with a dummy variable which takes values of unity in the last l periods. Thus, for a forecast origin T, and setting $l = 1$, this form of intervention is equivalent to estimating the model on data up to time $T - 1$. There are then two possibilities: a constant adjustment and an impulse adjustment, depending on whether the dummy variable is assumed to take the value of unity or zero over the period $\{T + 1, \ldots, T + h\}$. In the first case, forecasts are generated from a model corrected by the vector of in-sample errors at period T (again, assuming $l = 1$). In the second case, when the dummy variable is zero over the forecast period, the correction only affects the estimated parameters (by ignoring the period T observation when $l = 1$).

The *form* of the recommended correction will partly depend on the model, namely whether the model is in levels, differences, or is a VEqCM (e.g., constant adjustments are likely to be a better response to equilibrium-mean shifts in VEqCMs), and partly on the 'permanence' of the break. The *timing*, l, of the first unit value in the dummy will depend on the point at which the break occurred, the trade-off between forecast-error bias reduction and variance increases, and the type of shock. In particular, the last choice needs to reflect that an end-of-sample outlier may be a measurement error, or an impulse, rather than a step shift.

Clements and Hendry (2001g) consider two strategies that employ pre-tests of parameter constancy. In the first (Test$_1$), at each forecast origin, either the purely model-based forecasts or the intercept-corrected forecasts – based on whether or not a test on h 1-step forecasts up to the forecast origin is significant – is selected. If a break is signalled, a correction is applied based on the last l errors. The second strategy retains information on all previous constancy rejections (Test$_2$), and a dummy variable is added for each forecast origin at which constancy was rejected. As doing so should improve in-sample fit, the constancy test should be more likely to reject, in which case, a dummy is added for the last l periods up to the forecast origin as with the other strategy.

They examine the performance of a four-lag VAR for output growth Δy and the spread S (between 1 year and 3 month Treasury Bill rates) from 1959:3 to 2001:1. Their first forecast origin is 1969:4 with a maximum horizon of 8 quarters which generates 118 sequences of 1 to 8-step ahead forecasts. Figure 1 reports the MSFEs for S across the various strategies when $l = 4$ (similar results hold when $l = 1$). The constant adjustment does much less well than an impulse; using a more stringent significance level has little effect; it is slightly better to test than always correct; and Test$_1$ is somewhat better than Test$_2$, but the difference is marginal. Never correcting is worse than always using an impulse, but much better than a constant adjustment. A similar pattern was found for Δy, but with a more marked improvement of Test$_1$ over Test$_2$.

Figure 2 records the rejection frequencies for three forms of Chow (1960) test on the constancy of the VAR equation for S, and one system test, all scaled by their 5% 1-off critical values. As can be seen, the forecast errors seem to be drawn from a 2-regime process, switching in 1982, after which it enters a much more quiescent state. In the early period, 'outliers' proliferate, hence the benefit of impulse over constant adjustments, but after 1982 (the bulk of the evaluation sample) no breaks occur, which helps explain the poor performance of always using a constant adjustment.

Investigation of the *form* the IC might take could prove useful. An example of a way of restricting the ICs is suggested by Bewley (2000), who considers implementing ICs on the lines discussed in Clements and Hendry (1999a), using an alternative parameterization of the deterministic terms in the VAR. We
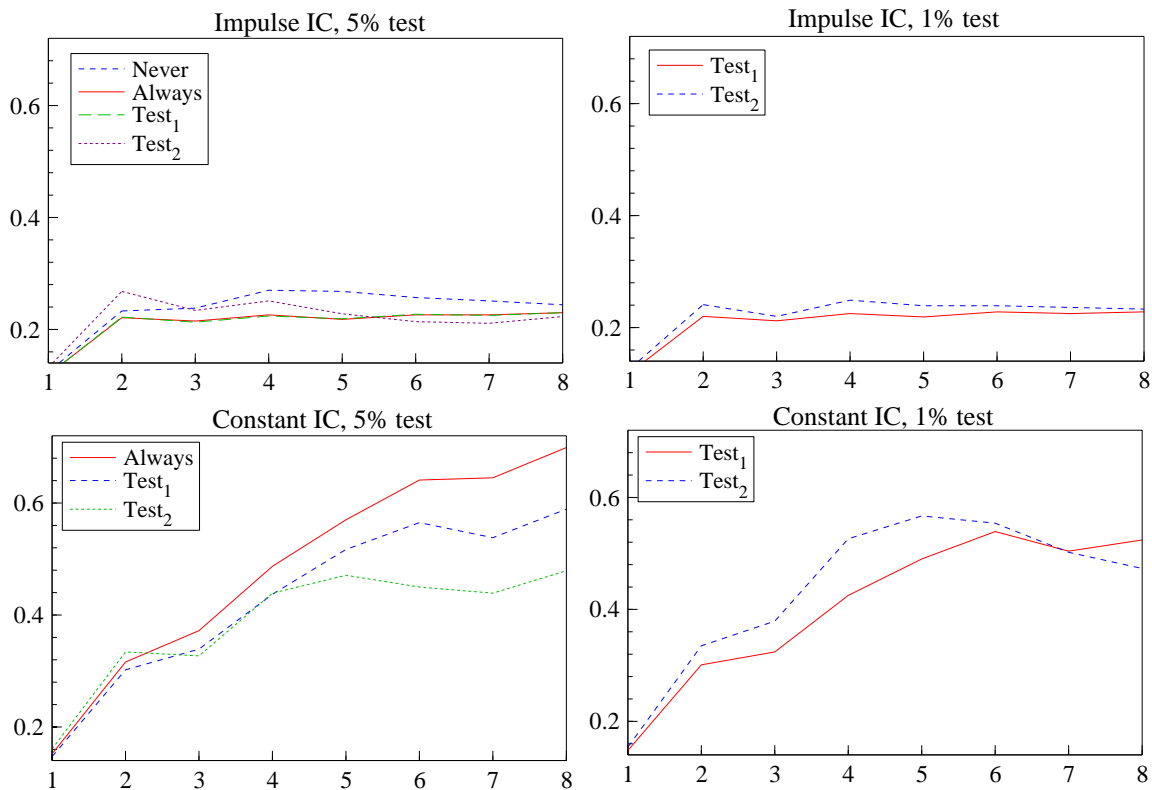
Figure 1 MSFEs for forecasting the spread.
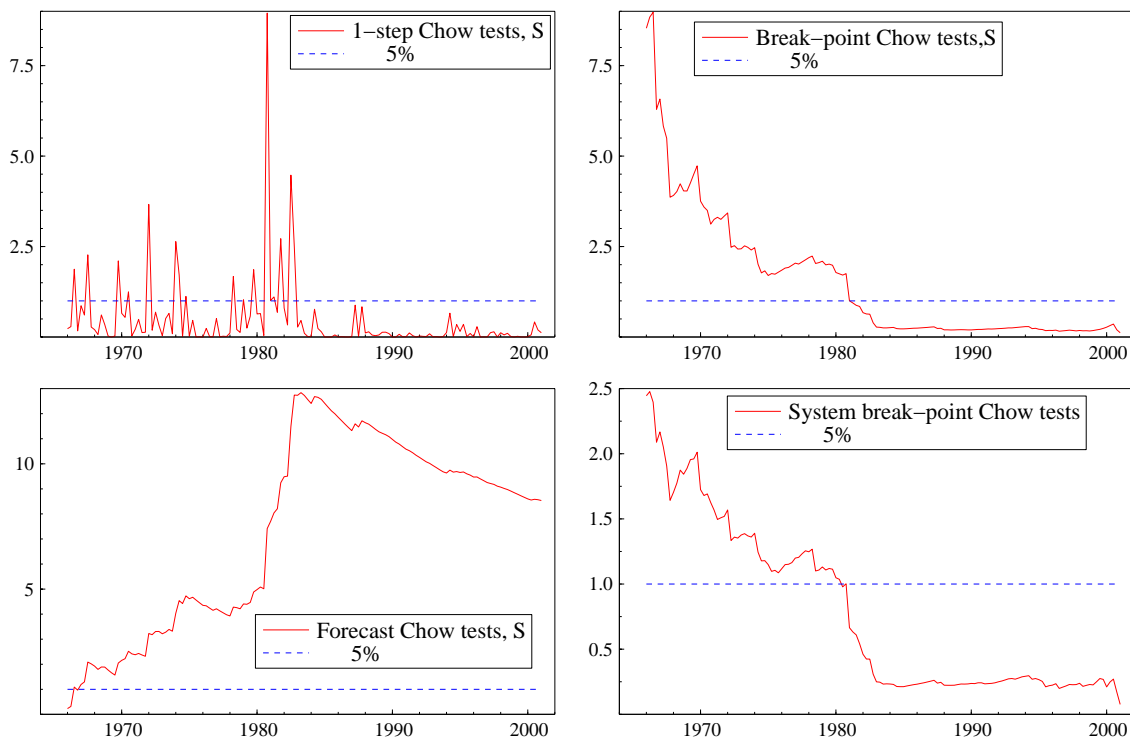
Figure 2 Chow test rejection frequencies for S and the VAR.

note the advantages of the alternative parameterisation below, but note that he does not consider the cointegrated case, where the corresponding re-parameterization is more awkward. The idea is to isolate the long-run growth in the system, given by the vector $\boldsymbol{\delta}$, as the vector of intercepts, so that shifts in growth rates are more easily discerned: a second advantage of is that zero restrictions can be placed on

specific elements of $\boldsymbol{\delta}$. For simplicity, consider an n-dimensional VAR in levels with a maximum lag of $p = 2$:

$$\mathbf{x_t} = \boldsymbol{\tau} + \mathbf{A_1}\mathbf{x_{t-1}} + \mathbf{A_2}\mathbf{x_{t-2}} + \boldsymbol{\epsilon_t} \tag{3}$$

where $\boldsymbol{\epsilon_t} \sim \mathsf{IN_n}\left[\mathbf{0}, \boldsymbol{\Omega_\epsilon}\right]$. In VEqCM form with r cointegrating vectors $\boldsymbol{\beta}'\mathbf{x_t}$, (3) becomes:

$$\Delta\mathbf{x_t} = \boldsymbol{\tau} + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x_{t-1}} + \mathbf{B_1}\Delta\mathbf{x_{t-1}} + \boldsymbol{\epsilon_t} \tag{4}$$

where $\boldsymbol{\alpha}\boldsymbol{\beta}' = \mathbf{A_1} + \mathbf{A_2} - \mathbf{I_n}$, and $\mathbf{B_1} = -\mathbf{A_2}$. In the Clements and Hendry (1999a) formulation:

$$\Delta\mathbf{x_t} - \boldsymbol{\delta} = \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\mathbf{x_{t-1}} - \boldsymbol{\mu}\right) + \mathbf{B_1}\left(\Delta\mathbf{x_{t-1}} - \boldsymbol{\delta}\right) + \boldsymbol{\epsilon_t} \tag{5}$$

where $\boldsymbol{\tau} = \left(\mathbf{I_n} - \mathbf{B_1}\right)\boldsymbol{\delta} - \boldsymbol{\alpha}\boldsymbol{\mu}$ with $\mathsf{E}[\Delta\mathbf{x_t}] = \boldsymbol{\delta}$ and $\mathsf{E}\left[\boldsymbol{\beta}'\mathbf{x_t}\right] = \boldsymbol{\mu}$ when there are no breaks. Forecasts of $\mathbf{x_t}$ will approach linear time trends with slope $\boldsymbol{\delta}$ as the horizon increases: thus, it is important to accurately estimate $\boldsymbol{\delta}$. When $\boldsymbol{\alpha} = \mathbf{0}$, Bewley (2000) sets to zero the elements of $\boldsymbol{\delta}$ for variables that do not exhibit drift, which can often be based on economic analysis (see Hendry and Doornik, 1997, for an illustration). Such restrictions are non-linear in (5), and infeasible on $\boldsymbol{\tau}$ in (3) or (4). However, applying the Bewley (1979) transform to (4) delivers:

$$\Delta\mathbf{x_t} = \boldsymbol{\delta} + \mathbf{D}\left(\boldsymbol{\beta}'\mathbf{x_{t-1}} - \boldsymbol{\mu}\right) + \mathbf{C_0}\boldsymbol{\Delta}^2\mathbf{x_t} + \mathbf{v_t} \tag{6}$$

where $\mathbf{C_0} = -\left(\mathbf{I} - \mathbf{B_1}\right)^{-1}\mathbf{B_1}$, $\mathbf{D} = \left(\mathbf{I} - \mathbf{B_1}\right)^{-1}\boldsymbol{\alpha}$. Equivalent forecasts to (4) are obtained if, given a super-consistent estimate of $\boldsymbol{\beta}$, (6) is estimated using $\Delta\mathbf{x_{t-1}}$ as an instrument for $\Delta^2\mathbf{x_t}$. Once $\boldsymbol{\alpha} \neq \mathbf{0}$, so the cointegration rank is non-zero, the relevant restrictions include $\boldsymbol{\beta}\boldsymbol{\delta} = \mathbf{0}$, not just $\delta_i = 0$. Moreover, tests for deterministic shifts involve $\boldsymbol{\delta}$ and $\boldsymbol{\mu}$, whereas only the combined intercept $\boldsymbol{\delta} - \mathbf{D}\boldsymbol{\mu}$ is available. A test focusing specifically on shifts in $\boldsymbol{\mu}$ would be valuable.

7.2 Modelling shifts

Again from a real time perspective, Phillips (1994, 1996) proposes a formal procedure for re-selecting and re-estimating a model as the sample changes.[6] This amounts to a more substantial revision to the equations than adjustments to the equations' intercepts, and he finds improved forecasts.

Other authors have focused on the possibility of modelling intercept shifts using a variety of regime-switching models. The idea behind the residual-based method of intercept corrections is that the structural change occurs close to the end of the sample but is unknown to the forecaster. However, in some instances a time series may have exhibited a sudden change in mean over the sample period. For example, consider the time series depicted by Hamilton (1993), figures 2–4, pp.232–234. Then a number of possibilities arise, one of which is to include appropriate dummy variables (impulse or shift, depending on whether the change is immediately reversed) to capture the effects of outliers or 'one-off' factors, without which the model may not be constant over the past. This strategy is popular in econometric modeling: see, for example, Clements and Mizon (1991) . However, to the extent that these 'one-off' factors could not have been foreseen *ex ante* and may occur again, the model standard error is an under-estimate of the true uncertainty inherent in explaining the dependent variable (1-step ahead), and prediction intervals derived from the model may be similarly misleading. Thus, a more accurate picture of the uncertainty surrounding the model predictions may be obtained by explicitly building into the probabilistic structure of the model the possibility that further regime changes may occur. Hamilton (1989) suggested using Markov switching regression (MS–R) models in these circumstances, where the

---

[6]This section draws on ch. 10 of Clements and Hendry (1999a).

temporal dependence in time series suggested the use of autoregressions (hence, MS–AR), building on the work of, e.g., Goldfeld and Quandt (1973). However, forecast confidence intervals (with a reasonably high nominal coverage level) even from models that omit this additional source of uncertainty are often found to be alarmingly wide, so that a greater benefit would appear to be any reductions in bias that might be achieved.

In this regard, consider the model:

$$\Delta y_t - \mu(s_t) = \alpha \left(\Delta y_{t-1} - \mu(s_{t-1}) \right) + \epsilon_t, \tag{7}$$

where $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$, and the conditional mean $\mu(s_t)$ switches between two states:

$$\mu(s_t) = \begin{cases} \mu_1 > 0 & \text{if } s_t = 1 \text{ ("expansion" or "boom")}, \\ \mu_2 < 0 & \text{if } s_t = 2 \text{ ("contraction" or "recession")}, \end{cases} \tag{8}$$

and the states s_t are determined by an ergodic Markov chain with transition probabilities:

$$\mathsf{p}_{\mathsf{ij}} = \Pr(\mathsf{s}_{\mathsf{t}+1} = \mathsf{j} \mid \mathsf{s}_{\mathsf{t}} = \mathsf{i}), \quad \sum_{\mathsf{j}=1}^{2} \mathsf{p}_{\mathsf{ij}} = 1 \quad \forall \mathsf{i}, \mathsf{j} \in \{1, 2\}. \tag{9}$$

Clements and Krolzig (1998) show that the forecast function for this model can be written as:

$$\widehat{\Delta y}_{T+h|T} - \mu_y = \alpha^h \left(\Delta y_T - \mu_y \right) + (\mu_2 - \mu_1) \left[(\mathsf{p}_{11} + \mathsf{p}_{22} - 1)^{\mathsf{h}} - \alpha^{\mathsf{h}} \right] \widehat{\zeta}_{T|T}.$$

where μ_y is the unconditional mean of Δy_t, and $\widehat{\zeta}_{T|T}$ is the filtered probability of being in regime 2 corrected for the unconditional probability. Thus, the conditional mean of Δy_{T+h} equals the optimal prediction rule for a linear model (the first term), plus the contribution of the Markov regime-switching structure, which is given by the term multiplied by $\widehat{\zeta}_{T|T}$, where $\widehat{\zeta}_{T|T}$ contains the information about the most recent regime at the time the forecast is made. The contribution of the non-linear part to the overall forecast also depends on the magnitude of the regime shifts, $|\mu_2 - \mu_1|$, and on the persistence of regime shifts $\mathsf{p}_{11} + \mathsf{p}_{22} - 1$ relative to the persistence of the Gaussian process, given by α. In their empirical model of post War US GDP growth, the predictive power of detected regime shifts is extremely small, $\mathsf{p}_{11} + \mathsf{p}_{22} - 1 \simeq \alpha$, so the conditional expectation collapses to a linear prediction rule. In general, then, the persistence of regimes and the degree of precision with which the current regime can be determined are important factors.

A number of other studies have reached fairly negative conclusions from a forecasting perspective – at least, there appears to be no clear consensus that allowing for non-linearities of these types leads to an improved forecast performance (see, e.g., De Gooijer and Kumar, 1992). Clements and Smith (2000b, 2000a) examine forecast performance from various non-linear specifications: see Granger and Teräsvirta (1993), and Franses and Van Dijk (2000) for more extensive discussions of forecasting with non-linear models. Swanson and White (1997) consider a 'flexible specification' of linear and non-linear models where the latter is linked to shifts, and Koop and Potter (2000) seek to differentiate between non-linearity, structural breaks, and outliers: Stock and Watson (1999a) conclude that non-linear models do not substantively outperform linear.

New classes of model are almost certainly required, perhaps variants of the switching class proposed by Engle and Smith (1998). Osborn, Sensier and Simpson (2001) claim that Markov switching models with leading indicators to help predict the regime may fare better. The improvements result from being better able to predict entry and exit to the 'rare event' of recessions: see section 7.9. The simple algebra above shows how this might help. However, as yet there does not seem to be a consensus on the advantages of any given approach for DGPs with deterministic shifts.

7.3 Forecast smoothing

It is unclear whether forecasting agencies should regard 'accuracy' as their dominant goal, relative to, say, 'plausibility'. When forecasts of the same outcome are made at different times, the implicit cost function may penalize sharp changes between adjacent forecasts: Nordhaus (1987) presents evidence that such inter-forecast smoothing occurs. Indeed, Don (2001) rejects the role of statistical criteria in judging 'forecast quality', and favours 'logical and economic coherence, and stability'. The third of these entails smoothing the announced forecasts towards previous statements when changes in available information entail more substantive revisions. Clements (1995) examined judgemental adjustments introduced to reduce 'high frequency' fluctuations in forecasts, but found no significant positive first-order serial correlation in the revisions to fixed-event forecasts for either the judgemental or mechanical forecasts from the *Oxford Economic Forecasting* model for the UK in the late 80s and early 90s. He concluded that their forecasts were not excessively smooth in the Nordhaus (1987) sense, although ICs reduced the dispersion of purely model-based forecasts. Further work in this area is reported in Clements (1997) and Clements and Taylor (2001).

Analytical results are needed of the impact of smoothing behaviour by forecasting agencies on the various sources of forecast error in section 3.1, not just when there are substantial white-noise measurement errors. For example, smoothing is antithetical to using ICs based on the latest forecast errors, and must induce delayed responses to deterministic shifts.

7.4 Role of surveys in forecasting

Survey information is possibly causal (in that the reported findings alter the behaviour of some economic agents), but there does not seem much evidence on that. Consequently, we regard surveys as a non-causal input to the forecasting processes. Such information could be entered as a regressor in forecasting systems, but that seems subject to the same problems as Emerson and Hendry (1996) found for leading indicators (see section 7.10). Alternatively, surveys might inform the estimate of the variables at the forecast origin (see section 7.6), perhaps guiding the choice of IC. Clements and Hendry (1998) suggest using signal extraction across all the available measures of the forecast origin to obtain better estimates.

7.5 Pooling of forecasts

There is a vast theoretical and empirical literature on pooling of forecasts (see the survey in Clemen, 1989), but as yet few results within the general framework of section 3.[7] The pooled forecasts across many methods reported in Stock and Watson (1999a) do well using either the mean or median forecast.

If two forecasts are differentially biased (one upwards, one downwards) it is easy to see why pooling would be an improvement over either. It is less easy to prove that a combination need improve over the best of a group, particularly as most forecasts will fail in the same direction after a deterministic shift – and all must do so if forecasting over a period where such a break unexpectedly occurs. Averaging does reduce variance, but only to the extent that separate sources of information are used. An alternative interpretation is that, relative to a 'baseline' forecast, additional forecasts act like ICs, which we know can improve forecasting performance not only if there are structural breaks, but also if there are deterministic mis-specifications. For example, Clements and Hendry (1999a) interpret the cross-country

---

[7]This paper was written before Hendry and Clements (2001g), who present formal explanations of why pooling will outperform when deterministic shifts occur in mis-specified models, and confirm their analysis by empirical and simulation illustrations.

pooling in Hoogstrate, Palm and Pfann (1996) as a specific form of IC, although such pooling can also be viewed as an application of Stein–James 'shrinkage' estimation (see e.g., Judge and Bock, 1978).

The need to pool violates encompassing (see Lu and Mizon, 1991), so reveals non-congruence, but it was shown above that congruence *per se* could not be established as a necessary feature for good forecasts. Hendry and Clements (2001g) show that no proof is possible of whether or not to include purely non-encompassed models when pooling, since all others are inferentially redundant only over the sample, not necessarily over the forecast period when deterministic shifts occur.

7.6 Measurement errors versus innovation shifts

Measurement errors in the latest available data on the forecast origin are bound to impinge adversely on ICs: see Hillmer (1984). Other sources of information, such as surveys as noted above, or the lapse of time, are needed to determine whether 'anomalous' readings on the state of the economy represent a shift or a mistake. Revisions to 'first-release' data are often substantial relative to the growth of the variables being forecast, confirming the benefits of appraising all available sources of information about the forecast origin, and suggesting 'smoothing' ICs, but a formal analysis is not available as yet. Wallis (1986) considers the related issue of the 'ragged edge' problem of missing data at the forecast origin.

7.7 Multi-step estimation

The general forecast-error taxonomy in Hendry (2000a), and special cases thereof discussed by Clements and Hendry (1998, 1999a), do not accord a major role to parameter estimation uncertainty or estimation biases. There are many reasons for such a result: biased parameter estimates need not entail biased forecasts; estimation uncertainty is of $O(T^{-1/2})$ in stationary, and $O(T^{-1})$ in integrated, processes, relative to other error sources of $O(1)$; and even correct in-sample specification with fully-efficient estimation is no guarantee of good forecasts in processes with breaks. Nevertheless, 'better' estimation has remained a topic of interest in the literature, including multi-step estimators which match the model estimation criterion with the forecast horizon: Bhansali (2001) provides a comprehensive review. One important reason may be (in)accurate estimates of deterministic terms.

Consider an h-step forecast from (3) when $p = 1$ commencing at a forecast origin at time T. Since:

$$\mathbf{x_{T+h}} = \sum_{i=0}^{h-1} \mathbf{A_1^i}\boldsymbol{\tau} + \mathbf{A_1^h}\mathbf{x_T} + \sum_{i=0}^{h-1} \mathbf{A_1^i}\boldsymbol{\epsilon_{T+h-i}} \tag{10}$$

the postulated multi-step system is:

$$\mathbf{x_{T+h}} = \boldsymbol{\phi_h} + \boldsymbol{\Gamma_h}\mathbf{x_T} + \mathbf{e_{T+h}}. \tag{11}$$

Thus, forecasts after estimation from minimizing e.g., $\left|\sum_{t=h+1}^{T} \mathbf{e_t}\mathbf{e_t'}\right|$ in (11) are given by:

$$\widetilde{\mathbf{x}}_{T+h} = \widetilde{\boldsymbol{\phi}}_h + \widetilde{\boldsymbol{\Gamma}}_h\mathbf{x_T},$$

rather than from estimating the parameters of (3) and using the analog of (10):

$$\widehat{\mathbf{x}}_{T+h} = \sum_{i=0}^{h-1} \left(\widehat{\mathbf{A_1}}\right)^i \widehat{\boldsymbol{\tau}} + \left(\widehat{\mathbf{A_1}}\right)^h \mathbf{x_T}. \tag{12}$$

When the process is stationary:

$$\sum_{i=0}^{h-1} \mathbf{A_1^i} = (\mathbf{I_n} - \mathbf{A_1})^{-1}\left(\mathbf{I_n} - \mathbf{A_1^h}\right),$$

so:

$$\sum_{i=0}^{h-1} \mathbf{A_1^i} \boldsymbol{\tau} = (\mathbf{I_n} - \mathbf{A_1})^{-1} \left(\mathbf{I_n} - \mathbf{A_1^h}\right) \boldsymbol{\tau} = \boldsymbol{\phi_h}.$$

It is difficult to see how multi-step estimation could offer more than minor gains in stationary processes. Despite biased parameter estimates, the long-run mean $\mathsf{E}[\mathbf{x_t}] = \boldsymbol{\mu}$ will be estimated consistently; and if the error process has a symmetric distribution, the forecasts from (12) will be unbiased, even if both the systematic and error dynamics are mis-specified.

However, when the process is non-stationary, intercepts partly represent drift terms, so mis-estimation could have more serious consequences. In the special case $\mathbf{A_1} = \mathbf{I_n}$, $\boldsymbol{\tau} = \boldsymbol{\delta}$ from (5), so letting $\widehat{\boldsymbol{\epsilon}}_{T+h} = \mathbf{x_{T+h}} - \widehat{\mathbf{x}}_{T+h}$:

$$\mathsf{E}\left[\widehat{\boldsymbol{\epsilon}}_{T+h} \mid \mathbf{x_T}\right] = \left(\mathsf{h}\boldsymbol{\delta} - \mathsf{E}\left[\sum_{i=0}^{h-1} \widehat{\mathbf{A_1}}^{i}\widehat{\boldsymbol{\delta}}\right]\right) + \left(\mathbf{I_n} - \mathsf{E}\left[\widehat{\mathbf{A_1}}^{h}\right]\right)\mathbf{x_T}. \tag{13}$$

Let the 'average error' in $\widehat{\mathbf{A_1}}$ as an estimator of $\mathbf{I_n}$ at a fixed sample of size T be $\boldsymbol{\Lambda}$:

$$\mathsf{E}\left[\widehat{\mathbf{A_1}}\right] = \mathbf{I_n} - \boldsymbol{\Lambda} \text{ approximating by } \mathsf{E}\left[\widehat{\mathbf{A_1}}^{i}\right] \simeq (\mathbf{I_n} - \boldsymbol{\Lambda})^{i} \simeq \mathbf{I_n} - i\boldsymbol{\Lambda},$$

and similarly, the 'average error' in $\widehat{\boldsymbol{\delta}}$ as an estimator of $\boldsymbol{\delta}$ be $\boldsymbol{\eta}$ ($\mathsf{E}\left[\widehat{\boldsymbol{\delta}}\right] = \boldsymbol{\delta} - \boldsymbol{\eta}$), then neglecting interactions and powers:

$$\mathsf{E}\left[\widehat{\boldsymbol{\epsilon}}_{T+h} \mid \mathbf{x_T}\right] \simeq \mathsf{h}\left(\left[\mathbf{I_n} - \frac{(\mathbf{h} - 1)}{2}\boldsymbol{\Lambda}\right]\boldsymbol{\eta} + \boldsymbol{\Lambda}\left[\frac{(\mathbf{h} - 1)}{2}\boldsymbol{\delta} + \mathbf{x_T}\right]\right). \tag{14}$$

A term like (14) could become large as h increases, especially as under-estimating unit roots converts $\boldsymbol{\delta}$ from a 'drift term' in an integrated process to an 'equilibrium mean' in the resulting (pseudo-stationary) estimated process. For example, an unmodelled negative moving-average error in (3) would induce such an outcome: see Hall (1989). However, Clements and Hendry (1996b) find that serious mis-specification of a mean-zero dynamic model is needed to ensure any gain from multi-step estimators even in integrated processes: the simulation evidence in Bhansali (2001) matches theirs, even though he also considers processes with non-zero means. Chevillon (2000) provides an analytic explanation for such Monte Carlo results in a scalar process, and shows that (e.g.) the biases and MSFEs are not monotonic functions of the DGP parameters or the horizon. He also considers DGPs with deterministic shifts just prior to the forecast origin (within h periods), and suggests that multi-step estimation does not ensure advantages in that setting either.

7.8 Co-breaking in forecasting

Co-breaking investigates whether shifts in deterministic terms in individual series cancel under linear combinations (see Hendry, 1995, Hendry and Massmann, 2000, and Clements and Hendry, 1999a). At first sight, a finding of co-breaking might seem invaluable for forecasting, since improved accuracy for the co-breaking combinations must result. Unfortunately, in an *ex ante* context, some of the series where the break itself occurs will still fail to be forecast well, so other combination will continue to suffer forecast failure. Nevertheless, both for econometric modelling and for forecasting some important variables, co-breaking seems likely to help.

As with cointegration, the feature which brings benefits is the existence of co-breaking, rather than its imposition on a model, although the latter will help in efficiency terms, and perhaps understanding.

An estimation algorithm for conditional co-breaking (in a dynamic model) has been proposed by Krolzig and Toro (2000); and for unconditional co-breaking (in the underlying process) by Massmann (2001), whose Monte Carlo experiments suggest reasonable power properties for tests of co-breaking rank, although the break points were assumed known a priori. An operational algorithm would have to jointly diagnose breaks and find co-breaking vectors, along the lines of Johansen (1988).

7.9 Forecasting rare events

The analysis above has primarily been concerned with post-break corrections, where the specification of the indicator variable to represent an intercept correction can be seen as determining the estimate of the magnitude and timing of any putative break. Forecasts made before a break and in ignorance of its impending occurrence are bound to suffer its full effects. Consequently, attempts to forecast future 'rare events' which entail deterministic shifts must be considered.

Environmental rare events such as hurricanes, earthquakes and volcano eruptions usually issue 'advance signs' that are harbingers of impending problems. Recent advances in (say) earth sciences for forecasting volcanic eruptions have focused on leading indicators (e.g., the temperature of the vented steam, where rises indicate increased activity), so we reconsider that avenue in section 7.10. If economic counterparts have corresponding attributes, then a search for 'early-warning signals' is merited. As noted above, Osborn *et al.* (2001) treat recessions as sufficiently rare that leading indicators in a regime-shift model might help in their prediction, and claim some success.

Unfortunately, many other rare events are not part of a sequence like business cycles on which even a small sample of observations is available: examples include the 1984 Banking Act and the 1986 Building Societies Act in the UK. Even so, 'rare events' should be partly predictable since they have causes, and some of those causes may be discernible in advance. One route may be monitoring high-frequency data, which should reflect deterministic shifts much sooner in real time, although there is the corresponding drawback that such data tend to be noisier. Nevertheless, 'early-warning' signals merit serious consideration, and we believe that high-frequency readings on the state of the economy must play a role in this area.

7.10 Leading indicators

Emerson and Hendry (1996) found that in theory and practice composite leading indicators (CLIs) by themselves were not likely to prove good at forecasting relative to robustified devices. Moreover, adding a leading indicator to a VAR, as in Artis, Bladen-Hovell, Osborn, Smith and Zhang (1995), might even jeopardize the latter's robustness for little gain (but see Marsland and Weale, 1992). More recently, Camba-Mendez, Kapetanios, Weale and Smith (2001) compare the performance of CLIs against a set of 'benchmark' VARs, but find that they are out-performed by 'naive predictors'. They attribute this outcome to the choice of leading indicators, and suggest improved measures can be found.

Another purpose of CLIs might be to 'forecast' a possible deterministic shift. However, it is difficult to see why present approaches to selecting such indicators would be optimal for that task, and recent experience remains somewhat discouraging: see Stock and Watson (1989, 1993).

8 Conclusions

A theory of economic forecasting that allows for structural breaks and mis-specified models (*inter alia*) has radically different implications from one that assumes stationarity and well-specified models. It can

be shown that theorems that can be readily established assuming stationarity and correct specification do not carry over to the more realistic setting, where 'realistic' denotes consonance with the empirical evidence on forecast failure and from forecasting competitions. Proposals for 'improving' forecasting need to be examined and judged within this setting. Doing so suggests ten areas where empirical performance can be understood, and ten that deserve greater research. Moreover, there are important implications from the revised theory about selecting models for forecasting and economic policy analysis.

References

Allen, P. G., and Fildes, R. A. (2001). Econometric forecasting strategies and techniques. In Armstrong, J. S. (ed.) , *Principles of Forecasting*, pp. 303–362. Boston: Kluwer Academic Publishers.

Artis, M. J., Bladen-Hovell, R. C., Osborn, D. R., Smith, G. W., and Zhang, W. (1995). Turning point prediction in the UK using CSO leading indicators. *Oxford Economic Papers*, **47**, 397–417.

Barrell, R. (2001). Forecasting the world economy. In Hendry and Ericsson (2001), pp. 149–169.

Bewley, R. (2000). Controlling spurious drift in macroeconomic forecasting models. mimeo, University of New South Wales.

Bewley, R. A. (1979). The direct estimation of the equilibrium response in a linear model. *Economics Letters*, **3**, 357–361.

Bhansali, R. J. (2001). Multi-step forecasting. In Clements and Hendry (2001a), pp. 206–221.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994). ARCH models. In Engle, R. F., and McFadden, D. (eds.) , *The Handbook of Econometrics, Volume 4*, pp. 2959–3038: North-Holland.

Burns, T. (1986). The interpretation and use of economic predictions. *Proceedings of the Royal Society*, **A407**, 103–125.

Calzolari, G. (1981). A note on the variance of ex post forecasts in econometric models. *Econometrica*, **49**, 1593–1596.

Camba-Mendez, G., Kapetanios, G., Weale, M. R., and Smith, R. J. (2001). The forecasting performance of the OECD composite leading indicators for France, Germany, Italy and the UK. In Clements and Hendry (2001a), pp. 386–408.

Chevillon, G. (2000). Multi-step estimation for forecasting non-stationary processes. MPhil Thesis, Economics Department, University of Oxford.

Chong, Y. Y., and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, **53**, 671–690. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, 591–605.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, **39**, 841–862.

Chu, C.-S., Stinchcombe, M., and White, H. (1996). Monitoring structural change. *Econometrica*, **64**, 1045–1065.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5**, 559–583.

Clements, M. P. (1995). Rationality and the role of judgement in macroeconomic forecasting. *Economic Journal*, **105**, 410–420.

Clements, M. P. (1997). Evaluating the rationality of fixed-event forecasts. *Journal of Forecasting*, **16**, 225–239.

Clements, M. P., and Hendry, D. F. (1993). On the limitations of comparing mean squared forecast errors. *Journal of Forecasting*, **12**, 617–637. With discussion.

Clements, M. P., and Hendry, D. F. (1995). Forecasting in cointegrated systems. *Journal of Applied Econometrics*, **10**, 127–146. Reprinted in T. C. Mills (ed.) *Economic Forecasting*. The International Library of Critical Writings in Economics, Edward Elgar.

Clements, M. P., and Hendry, D. F. (1996a). Intercept corrections and structural change. *Journal of Applied Econometrics*, **11**, 475–494.

Clements, M. P., and Hendry, D. F. (1996b). Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics*, **58**, 657–684.

Clements, M. P., and Hendry, D. F. (1997). An empirical study of seasonal unit roots in forecasting. *International Journal of Forecasting*, **13**, 341–356.

Clements, M. P., and Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Clements, M. P., and Hendry, D. F. (1999a). *Forecasting Non-stationary Economic Time Series*. Cambridge, Mass.: MIT Press.

Clements, M. P., and Hendry, D. F. (1999b). On winning forecasting competitions in economics. *Spanish Economic Review*, **1**, 123–160.

Clements, M. P., and Hendry, D. F. (eds.) (2001a). *A Companion to Economic Forecasting*. Oxford: Blackwells.

Clements, M. P., and Hendry, D. F. (2001b). Explaining forecast failure in macroeconomics. In Clements and Hendry (2001a), pp. 539–571.

Clements, M. P., and Hendry, D. F. (2001c). Explaining the results of the M3 forecasting competition. *International Journal of Forecasting*, **Forthcoming**.

Clements, M. P., and Hendry, D. F. (2001d). Forecasting with difference-stationary and trend-stationary models. *Econometrics Journal*, **4**, S1–S19.

Clements, M. P., and Hendry, D. F. (2001e). An historical perspective on forecast errors. *National Institute Economic Review*, **177**, 100–112.

Clements, M. P., and Hendry, D. F. (2001f). Modelling methodology and forecast failure. *Econometrics Journal*, forthcoming.

Clements, M. P., and Hendry, D. F. (2001g). Structural change and economic forecasting. Unpublished typescript, Economics Department, University of Oxford.

Clements, M. P., and Krolzig, H.-M. (1998). A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP. *Econometrics Journal*, **1**, C47–75.

Clements, M. P., and Mizon, G. E. (1991). Empirical analysis of macroeconomic time series: VAR and structural models. *European Economic Review*, **35**, 887–932.

Clements, M. P., and Smith, J. (2000a). Evaluating forecasts from SETAR models of exchange rates. *Journal of International Money and Finance*, **20**, 133–148.

Clements, M. P., and Smith, J. (2000b). Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment. *Journal of Forecasting*, **19**, 255–276.

Clements, M. P., and Taylor, N. (2000). Evaluating prediction intervals for high-frequency data. mimeo,

Department of Economics, University of Warwick.

Clements, M. P., and Taylor, N. (2001). Robust evaluation of fixed-event forecast rationality. *Journal of Forecasting*, **20**, 285–295.

Cooper, J. P., and Nelson, C. R. (1975). The ex ante prediction performance of the St. Louis and FRB-MIT-PENN econometric models and some results on composite predictors. *Journal of Money, Credit, and Banking*, **7**, 1–32.

De Gooijer, J. G., and Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing and forecasting. *International Journal of Forecasting*, **8**, 135–156.

Dhrymes, P. J., *et al.* (1972). Criteria for evaluation of econometric models. *Annals of Economic and Social Measurement*, **1**, 291–324.

Diebold, F. X. (1998). The past, present and future of macroeconomic forecasting. *The Journal of Economic Perspectives*, **12**, 175–192.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts: With applications to financial risk management. *International Economic Review*, **39**, 863–883.

Don, F. J. H. (2001). Forecasting in macroeconomics: A practitioner's view. *De Economist*, **149**, 155–175.

Eitrheim, Ø., Husebø, T. A., and Nymoen, R. (1999). Equilibrium-correction versus differencing in macroeconometric forecasting. *Economic Modelling*, **16**, 515–544.

Emerson, R. A., and Hendry, D. F. (1996). An evaluation of forecasting using leading indicators. *Journal of Forecasting*, **15**, 271–291.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.

Engle, R. F., and Smith, A. D. (1998). Stochastic permanent breaks. *Review of Economics and Statistics*, **81**, 553–574.

Ericsson, N. R. (2001). Forecast uncertainty in economic modeling. In Hendry and Ericsson (2001), pp. 68–92.

Ericsson, N. R., and Irons, J. S. (1995). The Lucas critique in practice: Theory without measurement. In Hoover, K. D. (ed.) , *Macroeconometrics: Developments, Tensions and Prospects*. Dordrecht: Kluwer Academic Press.

Fildes, R., and Ord, K. (2001). Forecasting competitions – their role in improving forecasting practice and research. In Clements and Hendry (2001a), pp. 322–253.

Fildes, R. A., and Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review*, **63**, 289–308.

Franses, P. H., and Van Dijk, D. (2000). *Non-linear Time Series Models in Empirical Finance*. Cambridge: Cambridge University Press.

Goldfeld, S. M., and Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, **1**, 3–16.

Granger, C. W. J. (2001). Evaluation of forecasts. In Hendry and Ericsson (2001), pp. 93–103.

Granger, C. W. J., and Pesaran, M. H. (2000a). A decision-theoretic approach to forecast evaluation. In Chon, W. S., Li, W. K., and Tong, H. (eds.) , *Statistics and Finance: An Interface*, pp. 261–278. London: Imperial College Press.

Granger, C. W. J., and Pesaran, M. H. (2000b). Economic and statistical measures of forecasting accu-

racy. *Journal of Forecasting*, **19**, 537–560.

Granger, C. W. J., and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Granger, C. W. J., White, H., and Kamstra, M. (1989). Interval forecasting: An analysis based upon ARCH-quantile estimators. *Journal of Econometrics*, **40**, 87–96.

Hall, A. (1989). Testing for a unit root in the presence of moving average errors. *Biometrika*, **76**, 49–56.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

Hamilton, J. D. (1993). Estimation, inference, and forecasting of time series subject to changes in regime. In Maddala, G. S., Rao, C. R., and Vinod, H. D. (eds.) , *Handbook of Statistics*, Vol. 11: Amsterdam: North–Holland.

Hansen, B. E. (2001). The new econometrics of structural change: Understanding and dating changes in US productivity. mimeo, Department of Economics, University of Wisconsin.

Hansen, H., and Johansen, S. (1998). Some tests for parameter constancy in cointegrated VAR-models. Mimeo, Economics Department, European University Institute.

Hatch, N. (2001). Modelling and forecasting at the Bank of England. In Hendry and Ericsson (2001), pp. 124–148.

Hendry, D. F. (1995). A theory of co-breaking. Mimeo, Nuffield College, University of Oxford.

Hendry, D. F. (2000a). A general forecast-error taxonomy. Mimeo, Nuffield College, Oxford.

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

Hendry, D. F. (2001). How economists forecast. In Hendry and Ericsson (2001), pp. 15–41.

Hendry, D. F. , and Clements, M. P. (2001g). Pooling of forecasts. Unpublished typescript, Economics Department, University of Oxford.

Hendry, D. F., and Doornik, J. A. (1997). The implications for econometric modelling of forecast failure. *Scottish Journal of Political Economy*, **44**, 437–461. Special Issue.

Hendry, D. F., and Ericsson, N. R. (eds.) (2001). *Understanding Economic Forecasts*. Cambridge, Mass.: MIT Press.

Hendry, D. F., and Juselius, K. (2000). Explaining cointegration analysis: Part I. *Energy Journal*, **21**, 1–42.

Hendry, D. F., and Juselius, K. (2001). Explaining cointegration analysis: Part II. *Energy Journal*, **22**, 75–120.

Hendry, D. F., and Massmann, M. (2000). Macro-econometric forecasting and co-breaking. Mimeo, Economics Department, Oxford University.

Hendry, D. F., and Mizon, G. E. (2000a). On selecting policy analysis models by forecast accuracy. In Atkinson, A. B., Glennerster, H., and Stern, N. (eds.) , *Putting Economics to Work: Volume in Honour of Michio Morishima*, pp. 71–113. London School of Economics: STICERD.

Hendry, D. F., and Mizon, G. E. (2000b). Reformulating empirical macro-econometric modelling. *Oxford Review of Economic Policy*, **16**, 138–159.

Hillmer, S. (1984). Monitoring and adjusting forecasts in the presence of additive outliers. *Journal of Forecasting*, **3**, 205–215.

Hoogstrate, A. J., Palm, F. C., and Pfann, G. A. (1996). To pool or not to pool: Forecasting international

output growth rates. Research Memorandum 96/025, Meteor, University of Limburg, Maastricht.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**, 231–254. Reprinted in R.F. Engle and C.W.J. Granger (eds), Long-Run Economic Relationships, Oxford: Oxford University Press, 1991, 131–52.

Johansen, S., Mosconi, R., and Nielsen, B. (2000). Cointegration analysis in the presence of structural breaks in the deterministic trend. *Journal of Econometrics*, forthcoming.

Judge, G. G., and Bock, M. E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North Holland Publishing Company.

Kennedy, P. E. (2001). Sinning in the basement: What are the rules? The Ten Commandments of applied econometrics. Mimeo, Economics Department, Simon Fraser University.

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

Klein, L. R. (1971). *An Essay on the Theory of Economic Prediction*. Chicago: Markham Publishing Company.

Koop, G., and Potter, S. M. (2000). Nonlinearity, structural breaks, or outliers in economic time series. In Barnett, W. A., Hendry, D. F., Hylleberg, S., Teräsvirta, T., Tjöstheim, D., and Würtz, A. (eds.) *Nonlinear Econometric Modeling in Time Series Analysis*, pp. 61–78. Cambridge: Cambridge University Press.

Krolzig, H.-M., and Toro, J. (2000). Testing for cobreaking and superexogeneity in the presence of deterministic shifts. Economics discussion paper 2000-W35, Nuffield College, Oxford.

Lu, M., and Mizon, G. E. (1991). Forecast encompassing and model evaluation. In Hackl, P., and Westlund, A. H. (eds.) , *Economic Structural Change, Analysis and Forecasting*, pp. 123–138. Berlin: Springer-Verlag.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In Brunner, K., and Meltzer, A. (eds.) , *The Phillips Curve and Labor Markets*, Vol. 1 of *Carnegie-Rochester Conferences on Public Policy*, pp. 19–46. Amsterdam: North-Holland Publishing Company.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., *et al.* (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, **1**, 111–153.

Makridakis, S., and Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, **16**, 451–476.

Marsland, J., and Weale, M. (1992). The leading indicator in a VAR model of the UK. Unpublished paper, Downing College and Clare College, University of Cambridge.

Massmann, M. (2001). Co-breaking in macroeconomic time series. Unpublished paper, Economics Department, Oxford University.

McNees, S. K. (1990). The accuracy of macroeconomic forecasts. In Klein, P. A. (ed.) , *Analyzing Modern Business Cycles*, pp. 143–173. London: M. E. Sharpe Inc.

Mincer, J., and Zarnowitz, V. (1969). The evaluation of economic forecasts. In Mincer, J. (ed.) , *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Newbold, P. (1993). Comment on 'On the limitations of comparing mean squared forecast errors', by M.P. Clements and D.F. Hendry. *Journal of Forecasting*, **12**, 658–660.

Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *Review of Economics and Statistics*, **69**, 667–674.

Osborn, D. (2001). Unit root versus deterministic representations of seasonality for forecasting. In

Clements and Hendry (2001a), pp. 409–431.

Osborn, D. R., Sensier, M., and Simpson, P. W. (2001). Forecasting and the UK business cycle. In Hendry and Ericsson (2001), pp. 104–123.

Pesaran, M. H., and Skouras, S. (2001). Decision-based methods for forecast evaluation. In Clements and Hendry (2001a), pp. 241–267.

Phillips, P. C. B. (1994). Bayes models and forecasts of Australian macroeconomic time series. In Hargreaves, C. (ed.) , *Non-stationary Time-series Analysis and Cointegration*, pp. 53–86. Oxford: Oxford University Press.

Phillips, P. C. B. (1996). Econometric model determination. *Econometrica*, **64**, 763–812.

Richard, J.-F., and Zhang, W. (1996). Econometric modelling of UK house prices using accelerated importance sampling. *Oxford Bulletin of Economics and Statistics*, **58**, 601–613.

Sampson, M. (1991). The effect of parameter uncertainty on forecast variances and confidence intervals for unit root and trend stationary time-series models. *Journal of Applied Econometrics*, **6**, 67–76.

Singer, M. (1997). Thoughts of a nonmillenarian. *Bulletin of the American Academy of Arts and Sciences*, *51*(2), 36–51.

Stock, J. H. (1996). VAR, error correction and pre-test forecasts at long horizons. *Oxford Bulletin of Economics and Statistics*, **58**, 685–701.

Stock, J. H., and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER Macro-Economic Annual*, 351–409.

Stock, J. H., and Watson, M. W. (1993). A procedure for predicting recessions with leading indicators: Econometric issues and recent experience. In Stock, J. H., and Watson, M. W. (eds.) , *Business Cycles, Indicators and Forecasting*, pp. 95–156: Chicago: University of Chicago Press.

Stock, J. H., and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics*, **14**, 11–30.

Stock, J. H., and Watson, M. W. (1999a). A comparison of linear and nonlinear models for forecasting macroeconomic time series. In Engle, R. F., and White, H. (eds.) , *Cointegration, Causality and Forecasting*, pp. 1–44. Oxford: Oxford University Press.

Swanson, N. R., and White, H. (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting*, **13**, 439–462.

Tay, A. S., and Wallis, K. F. (2001). Density forecasting: A survey. Mimeo, Department of Economics, University of Warwick, UK. Forthcoming in M. P. Clements and D. F. Hendry (eds), *A Companion to Economic Forecasting*, (2001), Blackwells.

Theil, H. (1966). *Applied Economic Forecasting*. Amsterdam: North-Holland.

Turner, D. S. (1990). The role of judgement in macroeconomic forecasting. *Journal of Forecasting*, **9**, 315–345.

Wallis, K. F. (1986). Forecasting with an econometric model: The 'ragged edge' problem. *Journal of Forecasting*, **5**, 1–14.

Wallis, K. F. (1989). Macroeconomic forecasting: A survey. *Economic Journal*, **99**, 28–61.

Wallis, K. F., and Whitley, J. D. (1991). Sources of error in forecasts and expectations: UK economic models, 1984–1988. *Journal of Forecasting*, **10**, 231–253.