Explaining the Results of the M3 Forecasting Competition

Michael P. Clements* and David F. Hendry†
Economics Department, Economics Department,
Warwick University. Oxford University.

July 4, 2001

Makridakis and Hibon (2000) summarize four main implications of the latest forecasting competition, which we paraphrase as: (a) 'simple methods do best'; (b) 'the accuracy measure matters'; (c) 'pooling helps'; and (d) 'the evaluation horizon matters'. We applaud the detailed empirical investigations, are unsurprised by their summary; but are surprised by the assertion that 'the strong empirical evidence, however, has been ignored by theoretical statisticians'. Having successfully published two books and more than a dozen papers across a wide range of journals, which *inter alia* analyze their four points, we refute the claim that the issue is being 'ignored', and doubt the implicit suggestion of hostility by the profession.

What must be the relationship between the world to be forecast and the models with which we forecast for conditions (a)–(d) *not* to hold? The research summarized in Clements and Hendry (1998b, 1999) (henceforth CH98 and CH99) shows that in weakly stationary processes, a congruent, encompassing model in-sample will dominate in forecasting at all horizons.[1] When the data generating process (DGP) is complicated, as is likely in economics, then so will be the dominant model, subject to possible losses from parameter estimation (CH98, ch. 12). Causal variables will dominate non-causal (CH99, ch. 1), forecast accuracy will deteriorate as the horizon increases, and there will be no forecast-accuracy gains from pooling forecasts across methods or models: indeed, pooling refutes encompassing. These are perhaps the 'optimality' claims that Makridakis and Hibon (2000) correctly doubt are empirically relevant.

The results of the forecasting competitions are manifestly at odds with such strong 'theoretical predictions'. This discrepancy between theory and practice (noted by, e.g., Fildes and Makridakis, 1995), and the systematic mis-forecasting and forecast failure that has periodically blighted macroeconomics, stimulated the research summarized in CH98 and CH99. The 'textbook' paradigm discussed in the previous paragraph offers no explanation for observed forecast failures, although they have sometimes been attributed to 'mis-specified models', 'poor methods', 'inaccurate data', 'incorrect estimation', 'data-based model selection' and so on, without those claims being proved: our research demonstrates the lack of foundation for such 'explanations'.

The reason that (a)–(d) hold in practice is that economies are non-stationary and evolving processes which are not reducible to stationarity by differencing, thereby generating moments that are

*All correspondence should be addressed to the first author.

†Financial support from the U.K. Economic and Social Research Council under grant L116251015 is gratefully acknowledged by both authors.

[1]Loosely, a well-specified model, not dominated by rival models: see Hendry (1995), chs. 9 and 14, respectively.

non-constant over time. Modern economies are regularly subject to major institutional, political, financial, legal, fashion, and technological changes which manifest themselves as structural breaks in models relative to the underlying DGP. Models are far from being facsimiles of the DGP, and even if they closely resembled it in-sample, unanticipated structural change could seriously reduce their usefulness for forecasting. Our research suggests that models which are relatively robust to, or adapt rapidly to, structural change are most likely to be successful in forecasting. Specifically, shifts in deterministic terms appear to be especially injurious to forecasting, and to be a primary factor underlying systematic forecast failure, as they cause a shift in the model's forecast mean relative to the data mean. Other breaks are surprisingly difficult to detect and have relatively benign effects on forecasts (see Hendry and Doornik, 1997, and Hendry, 2000). The remaining potential sources of forecast failure, ranging from model mis-specification, a lack of parsimony – including failure to impose restrictions such as unit roots and cointegration – inaccurate forecast-origin data, through to inefficient estimation, may all exacerbate forecast failure, but generally just play supporting roles.

Simple methods do best

Many simple methods, such as a random-walk model, for example, offer adaptability to structural change, in that the model immediately adapts to the latest level of the series. In a growing economy, second differences would be needed. But simplicity *per se* is not the key. The contrast is stark with using as a predictor, say, the simplest model of all, namely the sample mean of the data up to the forecast origin: such a predictor adapts very slowly to a shift in the mean of the process (see, e.g., Clements and Hendry, 1998a) and regularly fails. Complicated models tend to involve more that can go wrong when breaks occur, so parsimony has a role in a world of breaks: for example, vector autoregressions in growth rates (DVARs) may forecast more accurately than models which embody long-run (cointegration) constraints suggested by economic theory (Clements and Hendry, 1996). The latter may be more theoretically appealing, but are prone to an adverse performance from shifts in equilibrium means, to which DVARs are relatively immune. The forecast-error taxonomies in CH99 suggest this effect dominates in general, but analytical derivations could check the behavior under deterministic shifts of the various models deemed 'best' in forecasting by Makridakis and Hibon (2000).

The accuracy measure matters

When accuracy measures are not even invariant to linear transformations of the series to be forecast, the choice of measure must matter (see e.g., Clements and Hendry, 1993, and the ensuing discussion). CH98 ch. 3 also illustrates how the accuracy measure and evaluation horizon can interact. More generally, this is an area where the profession is active: see, in particular, the research on density evaluation (Tay and Wallis, 2000, provide a convenient survey).

Pooling helps

Different models (and measures) are differentially susceptible to different kinds of breaks, so pooling can help: Hendry and Clements (1994) provide an illustration. However, other types of pooling may be needed: in the policy context of economic forecasting after a structural break, where policy is then altered, Hendry and Mizon (2000) show the benefits of using a 'time-series' model to intercept correct an econometric system, should the latter be retained for policy analysis despite the forecast failure.

The evaluation horizon matters

Because of intermittent structural breaks in models relative to the underlying DGP, when sequences of forecasts are made and later evaluated, unanticipated breaks are likely to have occurred. Forecasts

announced prior to breaks will not fare well. However, when forecasts are made after breaks have occurred, methods which are reasonably robust to the breaks will have an advantage (a DVAR versus a cointegration model). Consider a forecast-evaluation scenario in which H observations are divided into k shorter-horizon periods of length H/k, with forecasts being made at the start of each sub-period. The multitude of forecast origins entails that a proportion of forecasts will be made after some breaks have occurred. When k is small, so that there are fewer longer-horizon forecasts, there will be fewer opportunities for robust methods to benefit from breaks that have already occurred. Our theory predicts that models that impose unit roots, such as differencing several times, should be effective when multiple breaks have occurred, as is likely for evaluation based on many short horizons. A scalar empirical case was investigated in Clements and Hendry (1998a) and found to be consistent with that theory. These findings are also consistent with the results of using a large macro-econometric system as reported in Eitrheim, Husebø and Nymoen (1999). Thus, the rankings of methods not only depend on their intrinsic properties, but also on the occurrence of breaks over the evaluation periods used, and on the lengths of horizons forecasted: so the length of the forecast horizon should indeed matter. Another surprise is that although predictability must decrease as the horizon grows, forecast accuracy need not: CH99, ch. 8 illustrates multi-step forecasts dominating 1-step after a deterministic shift.

Of course, as discussed by Fildes and Ord (2001) in their review of forecasting competitions, part of the aim of such competitions is to uncover approaches that can be applied to forecast many series with the minimum of human intervention, because this is a task that confronts many organizations. Forecasting competitions typically examine the performance of different methods for many time series, since the results for any one series could turn on idiosyncratic features of that series, and so limit the general applicability of the findings. Series are sometimes selected which share certain characteristics, with the caveat that the results might only hold for other series with those characteristics. Discovering models or methods that work well across a multitude of series reflects an emphasis that differs from our own. We would caution against the automatic adoption of a method that happens to win a particular forecasting competition: careful analysis as to why is strongly recommended. If a deterministic shift is suspected, or confirmed, then methods that are not robust to such a shift are likely to have performed poorly. But that would not necessarily preclude the continued use of a model that suffered forecast failure, subject to appropriate intercept corrections or other adjustments being used. This partly reflects our interest in macroeconomic forecasting and the interplay with policy making, rather than companies' sales forecasting. Nevertheless, we would argue that a good deal of progress has been made in the 'theoretical' literature on why some models or methods work well, and others do not. In particular, there now exists a theory relevant to forecasting non-stationary processes subject to breaks using mis-specified, data-based models, many of whose implications we have shown to be germane to interpreting the outcomes of forecasting competitions.

However, we do not claim that all the main issues have been resolved: far from it. The apparent dominance of forecast failure in determining the outcome in many situations does not preclude the importance of optimality considerations in 'normal conditions'. The classes of viable models for non-stationary processes remains tiny: unit roots on the one hand (handled by differencing and cointegration); deterministic shifts removed by indicator variables and co-breaking (see CH99, ch. 9), or modelled as switching unit roots (see e.g., Engle and Smith, 1998). More general models are urgently required; as is rigorous analysis of their properties.

References

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. (1996). Intercept corrections and structural change. *Journal of Applied Econometrics*, **11**, 475–494.

Clements, M. P., and Hendry, D. F. (1998a). Forecasting economic processes. *International Journal of Forecasting*, **14**, 111–131.

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

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

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

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

Fildes, R., and Ord, K. (2001). Forecasting competitions – their role in improving forecasting practice and research. Mimeo, Department of Management Science, Lancaster University. Forthcoming in M. P. Clements and D. F. Hendry (eds), *A Companion to Economic Forecasting*, (2001), Blackwells.

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.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

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

Hendry, D. F., and Clements, M. P. (1994). Can econometrics improve economic forecasting?. *Swiss Journal of Economics and Statistics*, **130**, 267–298.

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 Mizon, G. E. (2000). 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.

Makridakis, S., and Hibon, M. (2000). The M3-competition: Results, conclusions and implications. Discussion paper, INSEAD, Paris.

Tay, A. S., and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, **19**, 235–254.