but less extreme, reversal occurs with horizons 1 and 2. Such values can be easily explained with a single series, by having some strange values at the start, but an explanation is not so easy with 1428 series.

Concerning future research, consideration of trends is certainly important, even if we cannot define a trend, but emphasis on random walks, and unit roots, which have been major topics in econometrics for a decade, are likely to be replaced by consideration of breaks as in Clements and Hendry (1999).

It is also quite clear that M4 has to involve multivariate forecasting methods, as economics, finance, and demographics, all involve strong relationships. Many of the programs involved in M3 can deal with multivariate situations. Of course, the design of the experiment is much more difficult. The only attempt of which I am aware, by Magnus and Morgan (1999), involved eight or nine separate groups of econometricians using the same data, but the forecasting aspect was downplayed. The question should not be do multivariate forecasts beat univariate ones, but by how much?

References

Clements, M., & Hendry, D. F. (1999). *Forecasting non-stationary economic time series*, MIT Press, Boston, MA.

Engle, R. F., & White, H. (1999). *Cointegration, causality, and forecasting: a Festschrift in honour of Clive W.J. Granger*, Oxford University Press, Oxford.

Magnus, R. J., & Morgan, M. S. (1999). *Methodology and tacit knowledge. two experiments in economics*, Wiley, New York.

Stock, J. H., & Watson, M. W. (1999). *A Comparison of linear and non-linear university models for forecasting economic time series*: Chapter 1 of Engle and White (1999).

It's time to move from 'what' to 'why'

Rob Hyndman

Monash University, Clayton, Victoria, Australia

I would like to congratulate Makridakis and Hibon for another seminal contribution to the empirical forecasting literature. The M-Competitions have been highly influential in guiding forecasting practice and in motivating new forecasting research, and this latest competition will provide additional impetus to the push for forecasting methods that stand up empirically.

However, we now need to go beyond the simple comparison of forecasting methods, and the tabulation of which does better for different types of series and forecasting horizons. What is required is a careful analysis of *why* some methods perform better than others under different conditions. Is it possible to characterize the features of the best performing methods?

For example, it has long been recognized that single exponential forecasting (SES) is equivalent to an ARIMA(0,1,1) model (e.g., Makridakis, Wheelwright & Hyndman, 1998). The additional flexibility of ARIMA models may be thought to lead to more accurate empirical forecasts. However, Table 13 of MH shows that there is virtually no improvement in forecasting accuracy using ARIMA models (labeled B–J automatic). This is interesting, but has been widely known since at least the time of the first M-Competition. It is comforting that the previous findings continue to hold, but our understanding of forecasting has not been advanced much by its confirmation. Instead, I would like to see more research effort in *explaining* such empirical phenomena.

Perhaps part of the explanation for the remarkable performance of SES forecasts is found in Rosanna and Seater (1995) who provide empirical evidence that many aggregated economic series with relatively low sampling frequencies can be approximated by an ARIMA(0,1,1) process. Probably this is a consequence of the series being generated by random walks (due to the efficient markets hypothesis[1]) and observed with error (see Harvey, 1989). The success of SES is then a happy side-effect of the ubiquity of random-walk-like behaviour with observational error.

Furthermore, SES forecasts are now known to be optimal for a much larger class of models than previously recognized (see Harvey, Ruiz & Sentana, 1992; Harvey & Koopman, 2000; Chatfield, Koehler, Ord, & Snyder, 2001). Thus, SES forecasts may be inherently more robust than ARIMA forecasts because they are applicable to a larger class of stochastic processes than an ARIMA process.

A further factor leading to the success of SES forecasting over classical ARIMA models may be due to the lack of robustness in model-selection when identifying ARIMA models. To test this conjecture, I have carried out a small Monte-Carlo study based on the following ARIMA(0,1,1) process:

$$Y_t = Y_{t-1} + e_t - 0.5e_{t-1} \qquad (1)$$

where e_t is a Gaussian white noise series with zero mean and unit variance. Consequently, Box–Jenkins ARIMA modelling and SES should both be optimal for this process. I generated 1000 such series, each of length 30, and estimated an ARIMA(0,1,1) model for each one using only the first 20 observations. The remaining 10 observations were forecast and the forecast errors computed. The squared errors were then averaged across the 1000 series to obtain estimated MSE values for each forecast horizon. Note that the true model order was assumed, so the only source of error in the forecasts (apart from the error e_t process) was due to estimation.

For the same 1000 series, I implemented a restricted form of Box–Jenkins ARIMA modelling by fitting the models ARIMA(0,1,1), ARIMA(1,1,0) and ARIMA(1,1,1) with and without a constant term. From these six fitted models, the 'best' was chosen using Akaike's Information Criterion, and it was used to produce forecasts for the last 10 observations. These forecasts are then subject to both estimation error and model selection error.

The resulting MSE values are plotted in Fig. 1. Also shown is the optimal MSE (assuming the true underlying model) given by $(3+h)/4$ where h is the forecast horizon. Clearly, the estimation error is having minimal effect — the MSE from SES is close to optimal. However, the model-selection error results in a substantial increase in MSE. Thus, the additional flexibility of the Box–Jenkins approach increases the MSE because of some incorrect model selections.

To further study the robustness of ARIMA modelling, I generated series according to (1) but with $e_t = \delta_t z_t + (1 - \delta_t)a_t$ where z_t is standard Gaussian white noise with mean zero and variance 1, a_t is Gaussian white noise with mean zero and variance $\sigma^2 \geq 1$, and δ_t is an iid binary sequence taking value 1 with probability 0.9 and value 0 with probability 0.1. Thus, Y_t follows an ARIMA(0,1,1) process with a mixed error distribution which allows occasional large innovation outliers.

[1] More precisely, the efficient markets hypothesis leads to martingale series, of which random walks are a special case. See Campbell, Lo and MacKinlay (1997, pp. 20–33).

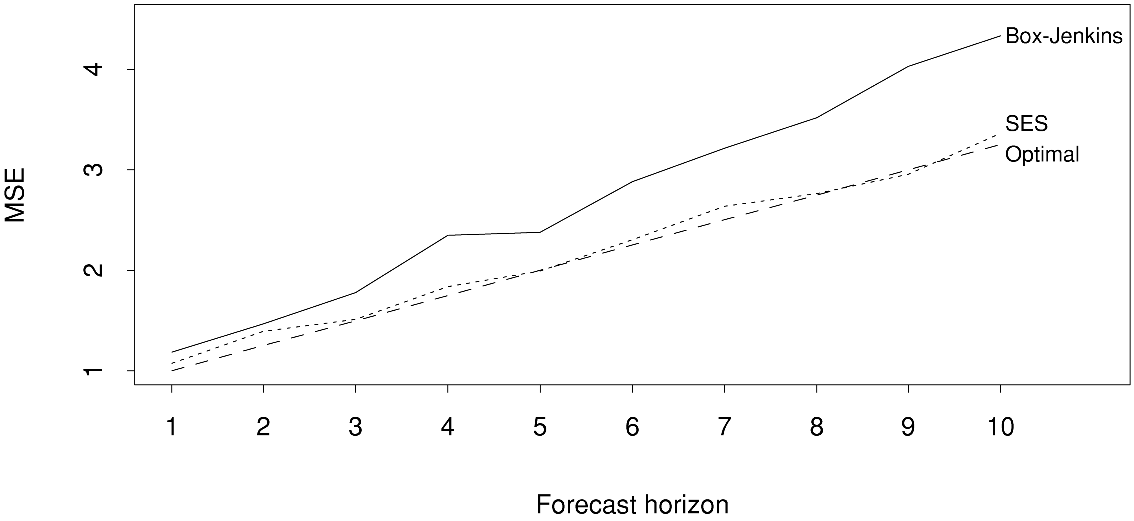

Fig. 1. Forecast MSE for the ARIMA(0,1,1) process. 'Optimal' shows the MSE assuming the true underlying model. 'SES' shows the MSE using SES (obtained by fitting an ARIMA(0,1,1) model) and 'Box–Jenkins' shows the MSE obtained from the best-fitting first-order ARIMA model.

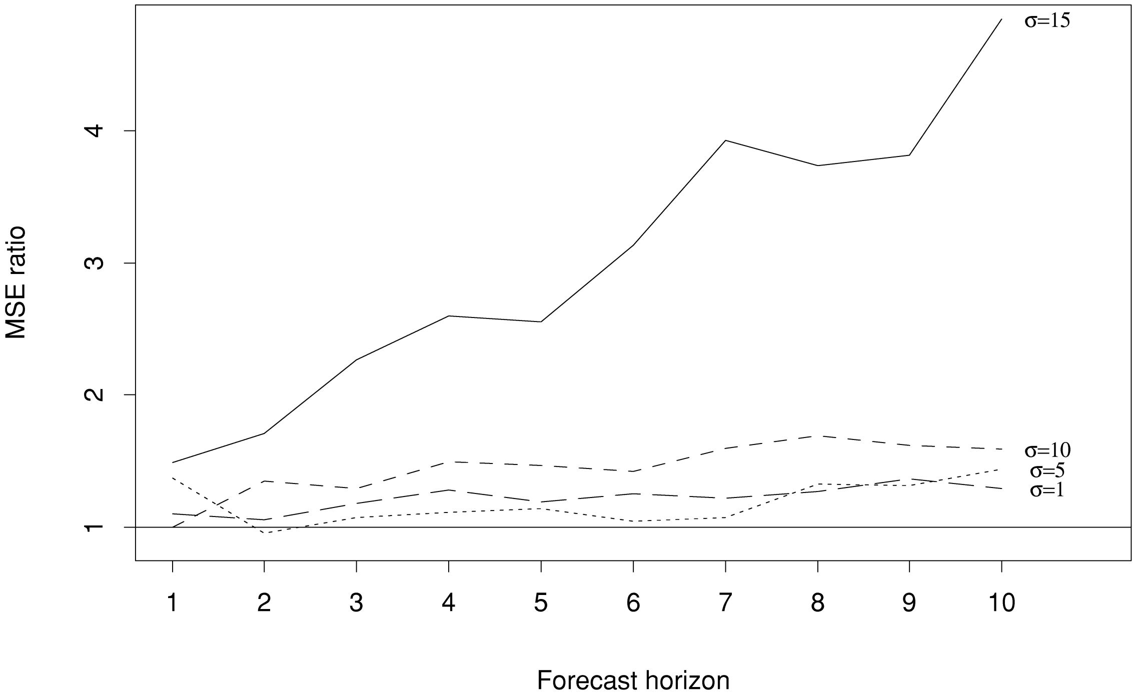

Fig. 2. The ratio of Box–Jenkins MSE to SES MSE for different values of σ.

I generated 1000 series for each of $\sigma=1$, $\sigma=5$, $\sigma=10$ and $\sigma=15$. Applying the same modelling procedure outlined above, I computed the MSE for SES and the restricted form of Box–Jenkins modelling. The ratios of MSEs (Box–Jenkins to SES) are shown in Fig. 2. Note that the non-normality of the errors has a large effect on the model-selection error of the Box–Jenkins procedure.

This simple Monte-Carlo study demonstrates that part of the relatively poor performance of Box–Jenkins methods is due to model selection errors in the larger model space, and that such errors are much worse when the data generating process is non-Gaussian. This suggests that a fruitful line of research would be to develop more robust model-selection methods for ARIMA modelling.

In summary, the M-Competitions have been invaluable in focusing attention on *empirical* forecasting performance rather than what might be possible on well-behaved data under ideal conditions. Makridakis and Hibon (2000) show us *what* works well and what does not. Now it is time to identify *why* some methods work well and others do not.

References

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The econometrics of financial markets*, Princeton University Press, Princeton, NJ.

Chatfield, C., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2001) Modelling for exponential smoothing: a review of recent developments. *The Statistician*, to appear.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Cambridge.

Harvey, A. C., & Koopman, S. J. (2000). Signal extraction and the formulation of unobserved component models. *Econometrics Journal 3*, 84–107.

Harvey, A. C., Ruiz, E., & Sentana, E. (1992). Unobserved component time series models with ARCH disturbances. *Journal of Econometrics 52*, 129–157.

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

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: methods and applications*, 3rd ed., John Wiley & Sons, New York.

Rosanna, R., & Seater, J. (1995). Temporal aggregation and economic time series. *Journal of Business and Economic Statistics 13*, 441–451.

The asymmetry of the sAPE measure and other comments on the M3-Competition

Anne B. Koehler

Miami University, Oxford, OH, USA

1. Introduction

Comparing the effectiveness of forecasting methods on real data is a worthy endeavor. Spyros Makridakis and Michèle Hibon conducted the M3-Competition for forecasting methods with a wealth of experience from past competitions. They have tried to address past criticisms of their studies by the selection of the data, software, and error measures. Since there