Commentary on Fotios Petropoulos & Enno Siemsen: "Criteria for Selecting Forecasts: What About Forecast Representativeness?"

Nigel Harvey[1] and Shari De Baets[2]

1. **Nigel Harvey** (n.harvey@ucl.ac.uk) is Professor of Judgment and Decision Research at University College London. His research deals with the many ways in which judgment is involved in the forecasting process.

2. **Shari De Baets** (Shari.DeBaets@UGent.be) is a senior post-doctoral researcher in the Faculty of Economics and Business Administration at Ghent University. She works at the interface of operations management and behavioural science with a particular focus on judgmental forecasting and the human perception of algorithms in work life.

Fotios and Enno introduce us to the representativeness criterion for selecting between different statistical forecasting models. To use this criterion, samples are taken from historic datasets and then the patterns in successive chunks of those data are compared with the pattern in a similar length chunk provided by the forecast sequence. The less similar the pattern in the forecast sequence chunk is to the patterns in the historical data chunks, the less representative the forecast sequence is of the data series: in other words, the bigger the representativeness gap. To obtain a value of the representativeness criterion (REP), the representativeness gap is added to the performance gap, the in-sample difference between forecasts and outcomes. The model with the lowest REP value is the one selected.

Fotios and Enno compare the success of the REP criterion with that of two established approaches, the Akaike information criterion (AIC) and cross-validation (CV), in selecting between models from the exponential smoothing family. To do this, they used forecasts produced by the models for real series drawn from those used in forecasting competitions. REP produced higher forecast accuracy than the other two approaches. Even using the representativeness gap alone excluding the performance gap produced higher forecast accuracy than use of the information criterion.

These results are impressive. How generalizable are they? Petropoulos and Siemsen (2022) showed that REP is more effective than AIC and CV not just in selecting between alternative exponential smoothing models but also in selecting between alternative ARIMA models. Furthermore, they went on to demonstrate that it is more effective than CV in selecting between heterogeneous model types (exponential smoothing, ARIMA, Theta model) for lower frequency data (yearly, quarterly, monthly, weekly) and similar to CV for higher frequency data (daily, hourly). Thus the REP approach generalizes well across model types.

What about generalization across series types? In their empirical work, Petropoulos and Siemsen used 103,830 series, some of which were drawn from those used in the M and the M3 forecasting competitions but most of which were drawn from the M4 forecasting competition. These series were drawn from a wide variety of different domains; for example, M4 series were taken from industries, services, tourism, imports & exports, demographics, education, labour & wage, government, households, bonds, stocks, insurances, loans, real estate, transportation, and natural resources and the environment. In fact, Spiliotis, Kouloumos, Assimakopoulos and Makridakis (2020) provide evidence that the M4 series are statistically representative of real world series. So we can say that the findings reported by Petropoulos and Siemsen (2020) hold, *on average*, for the types of series from which models are used to make forecasts in the real world.

Of course, this does not mean we can be certain that Fotios and Enno's conclusions will hold for a particular type of series. We know from the Makridakis forecasting competitions that there is no overall optimal way to make forecasts; instead, particular approaches suit particular types of series more than others. Similarly, different ways of selecting between forecasting models may be appropriate for different types of series. We'll consider two types of series that might be interesting to focus on in future work: non-linear series and trended series. This is because, in both these cases, patterns in past data may not be a perfect guide to the patterns we should expect in future data.

Generally, M4 series showed a high degree of linearity (Spiliotis et al., 2020) but there are specific domains (finance, meteorology) where series are typically better characterized as non-linear (Mandelbrot and Hudson, 2008). Differences between non-linear series can be characterized by their Hurst exponent (H). The H exponent for most financial series varies between 0.3 and 0.7 (Sang et al., 2001). An H value of 0.5 corresponds to a random walk; an H value greater than 0.5 means that the series changes direction less frequently than a random walk – it is persistent; an H value less than 0.5 means that it changes more frequently than a random walk – it is anti-persistent. For example, daily stock returns for Caterpillar can be characterised by a Hurst exponent of 0.329 (Sang et al., 2001). This means that trends in the series are likely to continue for a while before reversing and continuing in the opposite direction, only to eventually reverse again. The correlogram shows a long slow decline, indicating long-range dependencies in the data (e.g., Baillie, 1996).

How should we proceed? It may look as if such series are describing a non-deterministic seasonal process but these patterns are transient: they cannot be used to make adequate long-range forecasts. The best approach would be to start by establishing that the series is indeed non-linear and determining the nature of its non-linearity by, for example, extracting its Hurst exponent. However, this requires analysis of a very long data series that, except in some applications such as high-frequency trading, is typically not available. We can still use established forecasting models and expect them to produce adequate short-term forecasts; for example, with a Hurst exponent of 0.329, we know there is a good chance of the current trend continuing into the future over the next few periods.

Would REP provide the best way of selecting between models to provide forecasts from non-linear series? Assuming, as Petropoulos and Siemsen (2022) do, that the data generation process is unknown to forecasters, the length of the forecast window would be set equal to the apparent periodicity in the data series and the candidate models (e.g., from the exponential smoothing family) would each provide forecasts for that window. The representativeness gap between each set of forecasts and data in each of the previous windows would then be calculated in the way that Fotios

and Enno describe. Given the apparent 'seasonal' patterns in non-linear series, the representativeness gap may be smallest with the exponential smoothing models that include seasonal component. Also, given the transience of these 'seasonal' patterns, such models may fit better when the discount factor is higher (i.e., when greater emphasis is placed on comparing forecasts with more recent time windows).

What about the AIC approach? Petropoulos and Siemsen (2022) provide examples of where this approach fails: forecasts did not follow trends in the data series because of the penalty applied to complexity. Complex models are likely to be needed in attempts to account for the behaviour of non-linear series and so the complexity penalty could result in AIC selecting a simpler model than that chosen by REF in this case as well. What about the CV approach? It does not penalize model complexity. But it is likely to be affected by the transient nature of apparent trends and seasonalities in non-linear series: patterns present in the training sets may not remain present in validation sets. As a result, CV may also select a simpler model than REP. The question would then be whether non-linear series are better forecast by the simpler exponential models selected by AIC and CV or by the more complex ones selected by REP. Here, our arguments about the types of model that might be favoured by different selection criteria are, of course, speculative but they are designed to support our point that it would be worth carrying out empirical research to compare different ways of selecting between models for particular types of series – in this case, non-linear series.

We now turn to trended series. In their paper, Petropoulos and Siemsen (2022) point out that their development of the REP criterion was at least partly triggered by findings from judgmental forecasting research that indicate that people try to make their sequence of forecasts look like (i.e., have the same characteristics) as the data series. Can other results from that area of research add to the development of REF as a way of selecting between forecasting models?

One major finding is that, when making forecasts from trended series, people tend to place their forecasts below upward trend lines and above downward ones (Eggleton, 1982; Harvey and Reimers, 2013). It appears that when forecasting from linear or exponentially increasing trends in data series, people take account not just the data in front of them but also their knowledge of the real world. That knowledge tells them that nothing continues to increase or decrease for ever (entropy apart). Instead, trends eventually asymptote or turn into cycles. Forecasters factor this knowledge into their judgments as trend damping. Statistical forecasters have also discovered the value of trend damping: Gardner and McKenzie (1985) showed that adding an *ad hoc* damping term can improve performance of exponential smoothing models that take trends into account. In summary, the

steepness of trends in past data cannot be considered to be a perfect guide to the steepness of trends in future data.

It would be possible to add an ad hoc damping term to forecast sequences that contain trends. In Figure 1 of their paper, the trend-only and the trend & seasonal forecasts could be subject to a small amount of damping to improve their accuracy. However, this would have two effects. First, it would make the forecast sequence (window 6 in Figure 1) less representative of the data in the windows of the data series (windows 1-5 in Figure 1). As a result, the representativeness gap would increase. However, on the basis of Gardner and McKenzie's (1985) findings, we would expect the performance gap to decrease. If this latter effect is larger, REP would be lower when a damping term is used to improve forecasts. So our question is whether REP can select an exponential model with damping over one without damping. The AIC approach is less likely to do this because addition of the damping term adds to the model and will therefore be subject to a complexity penalty.

Finally, Fotios and Enno have described how the innovative work that they describe in this paper was at least partially triggered by research on judgmental forecasting. We can ask how their ideas might influence work on judgmental forecasting.

There has long been a debate about the relative effectiveness of purely judgmental, purely statistical, and hybrid approaches to forecasting. Lawrence, Edmundson and O'Connor (1985, p 25) concluded from their studies that "judgmental extrapolation is on average no less accurate than statistical forecasting and, in a number of subgroups of the time series, was the most accurate". In the M2 competition (Makridakis, Chatfield, Hibon, Lawrence, Mills, Ord and Simmons, 1993), forecasts produced by purely statistical methods were compared with those produced by people who had access to statistical methods but who could judgmentally adjust the forecasts produced by those methods to take account of additional information. There was little difference between these two types of forecast at short horizons but purely statistical forecasts tended to be superior at longer ones. The authors conclude that "no judgmental revisions ought to be made to the quantitative forecasts without making sure beforehand about the need and value of such revisions" (pp 17-18).

This debate is unlikely to be resolved easily; circumstances in which judgment adds quality to forecasting processes still need to be identified more precisely. However, it should be possible to use the REP criterion to select not just between purely statistical approaches but also between those approaches and both pure judgmental forecasting and hybrid approaches that combine statistical and judgmental contributions in various ways.

References

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics, 73*(1)*,* 5-59.

Eggleton, I. R. C. (1982). Intuitive time-series extrapolation. *Journal of Accounting Research, 20*(1)*,* 68–102.

Gardner, E. S., Jr., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science, 31*(10)*,* 1237–1246

Harvey, N. & Reimers, S. (2013). Trend damping: Under-adjustment, experimental artefact, or adaptation to features of the natural environment? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2)*,* 598-607.

Lawrence, M. J., Edmundson, R. H. & O'Connor, M. J. (1985). An examination of the accuracy of judgmental extrapolation of time-series. *International Journal of Forecasting, 1*(1)*,* 25-35.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K. & Simmons, L.F. (1993). The M-2 competition: A real-time judgmentally based forecasting study*. International Journal of Forecasting, 9*(1)*,* 5-22.

Mandelbrot, B. B. & Hudson, R. L. (2008). *The (mis)behaviour of markets: A fractal view of risk, ruin and reward.* London: Profile Books.

Petropoulos, F. & Siemsen, E. (2022). Forecast election and representativeness. *Management Science*, In press.

Sang H-W, Ma, T. & Wang, S-Z. (2001). Hurst exponent analysis of financial time series. *Journal of Shanghai University (English edition), 5*(4), 269–272.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V. & Makridakis, S. (2020). Are forecasting competitions data representative of reality? *International Journal of Forecasting, 36*(1)*,* 37-53.