A brief history of forecasting competitions - review

A reviewer always has to face up to whether the author should be left to write his or her own paper, or should be required to write the reviewer's paper for him. Having writing a somewhat similar paper 15 years ago, I obviously am glad to see a new interpretation of the questions arising from forecasting competitions. But there are some important things in this paper I think are missing, in particular there are two issues I would like to see much more fully discussed – the objectives of the different competitions and how they have changed (if at all) and the criticisms levelled. The latter issue is regularly visited in the paper but not I think fully faced up to. At the time of the Fildes & Ord paper the commentary on the M3 competition was not available. The section on p.5 picks up on 5 points focussing more on extensions. However, a quick review of the M3 discussion reveals a number of additional and I would say important points. This links to the lack of clarity as to the objectives of the different competitions. Granger and Newbold were clear (even if misguided) but many of the later competitions have been opaque.

Many of the objections to the competitions related to a lack of a clear population from which the *series are drawn. As I commented with Spyros in 1995,*

This criticism is valid, but the emphasis is misplaced. "Many (if not all) field experiments suffer from the same limitation of using a non-random sample experimental base. It is overcome by the experimenter increasing the sample size and the diversity of the set (of selected time series) and by ensuring that the non-random components in the experiment are of little importance to the possible outcome. As Johnstone (1989) argues, the statistical testing of hypotheses is not dependent on randomness in the sample, only lack of systematic bias. Inexplicable findings should lead to a revised definition of the population under study and a revised view of the results expected."

I would therefore see the discussion on significance p.5 as perpetuating a misunderstanding. It also suggests that more series (111, 1001, 3000,100K - M4) is better. Well it depends doesn't it?

Other issues are the survey of past competitions. Again the objectives need consideration. I note that Fildes and Ord included Meese, R. A., & Geweke, J. (1984). A comparison of autoregressive univariate forecasting procedures for macroeconomic time series *Journal of Business and Economic Statistics, 2* 191-200.

Also there are competitions which are not fully focussed on univariate time series sata but have relevance, e.g. Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research, 249*, 245-257. This covers 926 SKUs and in passing shows the inadequacy of ETS compared with a causal models. I don't think there's a lot of multivariate comparisons out there (Hyndman's own of course) but this could be used as an example rather than requiring a full survey. We have a paper which is barely passed the conference presentation stage, again on promotional data but it includes all the univariate method as well as extensions to include promotional indicators. TBATS wins with promotional data included.

The discussion on replication could be usefully expanded to include papers published in the IJF, e.g. Boylan.

The final point of what I appreciate is a rather rambling review is that in the discussion of the response to the competitions the Fildes and Makridakis paper shows that the theoretical time series literature up to 1994 did not respond. I haven't looked again at this issue apart from a superficial check now but I don't think there's any evidence of the finding being taken on board by for example time series statisticians.

To make my suggestions for revision more focused, I suggest a table that covers objectives and a table/ list that covers methodological changes seen. In addition the various points listed above need further thought and discussion.