Response to reviewers

INTFOR_1812597.R1 6 March 2019

Thanks to the editor and the two reviewers for further additional comments. My response are in black, the reviewers' comments are in red.

Reviewer 1 comments

This invited contribution to the M4 special issue reads well and is very clear. i have very few quibbles. I still believe it could be improved by responding somewhat to my suggestions on the first review but I will leave it to the AE (who undoubtedly will agree with at least one point) to decide.

On p.9 there is a recommendation on using M4 as a test bed – papers should not be published unless they perform well . . . on at least some well define subset of M4. But on l.30-31 this is contradicted by the much more sensible suggestion that comparisons should be clear as to the domain to which they apply.

The statements do not contradict. In the first statement I was referring to proposals for 'a new *general* time series forecasting method'. It is reasonable to test such proposals on the M4 data, or on some well-defined subset of it. In the second statement, I am discussing the domain to which the competitions result apply. If someone is proposing a specific forecasting method for a narrow domain, then of course it makes no sense to test it on the M4 data.

P.4 I raised the question as have many others as to the issue of how representative the M3/ M4 data is of the population. Spiliotis et al explore this and in the end we are left with Hyndman's conclusions that it is not clear what population of time series the M3 conclusions apply to (p.5, l.26). This relates to the point above and merits more thorough consideration.

I now cite Spiliotis et al. (2018).

I suggested in the first round there were missing competitions to which the author's response was that the analysis was restricted to post-1980 competitions with multiple participants. OK – so why is that reasonable? Are there that many missing – not to my knowledge. The case can be made if the author had responded positively to the point about methodological innovations. I do not accept the point that a list would be hard to construct. M4's strength is I think methodological primarily, e.g. publicly available code.

There are probably hundreds (possibly thousands) of such 'competitions' if we allow any comparisons of 3 or more forecasting methods where the computations have been carried out by the authors. Besides, comparisons that involve only the authors of a paper are prone to bias because the participants will tend to favour their proposed method, or because the participants are more skilled at some methods than others. This is why Makridakis & Hibon (1979) was heavily criticised, and why the subsequent M competitions were open to multiple entrants. To extend my comparisons to cover all these papers would be totally impractical and unhelpful.

Minor issues

p.1, l.25: not 50 years. Yule is 90, Working 85. The econometricians go back to Cowles.

I've clearly defined a forecasting competition as involving multiple entrants. Such competitions go back about 50 years.

p.1, l43 'I do not cover. . . .'

Fixed

p.2 end para section 1. How do you explain Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018). Judgmental selection of forecasting models. Journal of Operations Management, 60, 34-46

My comment was in response to Jenkin's implied claim that model building cannot be represented by an algorithm or a mathematical model. I said that history had shown that to be untrue 'provided enough data are available, and the model is flexible enough to capture the variation seen in real data.' Petropoulos et al. (2018) does not refute that. Instead, they show that judgmental model selection is sometimes better than algorithms, and combining judgmental and statistical models can lead to statistically significant improvements. So it doesn't need explaining. Nevertheless, I have added a citation.

p.9 Minimizing MASE is equivalent to minimizing MAE, Note.

A comment has been added.

p.4 Do we need to say (even if we agree) that the initial explanation of Theta was highly complicated and confusing?

Yes, because it is relevant to the 'simple' claim by Makridakis.

Reviewer: 2

The author have addressed my comments. I forgot to ask the author to cite the GEFCom2017 intro paper in the first round of review. The paper is currently in press with IJF, to be published in 2019.

Tao Hong, Jingrui Xie, and Jonathan Black, 'Global Energy Forecasting Competition 2017: Hierarchical Probabilistic Load Forecasting,' International Journal of Forecasting, in press.

Now added

References

Makridakis, SG & M Hibon (1979). Accuracy of forecasting: an empirical investigation (with discussion). *Journal of the Royal Statistical Society. Series A* **142**, 97–145.

Petropoulos, F, N Kourentzes, K Nikolopoulos & E Siemsen (2018). Judgmental selection of forecasting models. *Journal of Operations Management* **60**, 34–46.

Spiliotis, E, A Kouloumos, V Assimakopoulos & S Makridakis (2018). *Are forecasting competitions data representative of the reality?* Tech. rep. 3/18. Forecasting and Strategy Unit, National Technical University of Athens.