Response to reviewers

IJF ??? 24 January 2019

1 Associate Editor's comments:

This is a very well-written paper. The two reviewers give some constructive suggestions for improving the manuscript. I especially enjoyed the quotes and anecdotes. I also have two comments - one minor and one more major:

- P4L49: I would say 'The top performing entry was . . . '

- Why not briefly touching on forecasting competitions other than time series? M2 should be briefly mentioned, I think. Also, I would add a small paragraph on the Tetlock-Superforecasters project in section 3.

2 Reviewer: 1

P1L37. I would change 'Young researchers in forecasting' to 'Researchers new to forecasting' or something not about age.

P2L36-40. While I disagree with Jenkins' original statement, I wouldn't completely dismiss it as 'untrue' and go to the other extreme. I believe building a forecasting model is a systems engineering process. Human brain has to be part of the process. I would say 'model building is best done by the human brain and computers together and is inevitably an iterative process.'

P3L19. Not sure if 'insults' is the right word there. Maybe changing it to 'questioning the competency of the analyst'. I personally believe that the competency of the analyst does affect the forecast accuracy significantly.

P4L24. Please elaborate 'the second was not strictly time series forecasting.'

P4L39. Please explain why the DSES method was not included. If it was not in a published literature, can you ask Spyros and put the answer in this paper? Otherwise, you are leaving the readers wondering why not include DSES.

A general comment to Section 2: somewhere in this section, either beginning or end, it's better mention your definition of 'time series' competition. Did you mean to say that explanatory variables were not provided?

P6L39. I don't think 'consensus' is the right word there. Many researchers are still advocating ANN and those blackbox models for time series forecasting.

P6L49. I think it's better characterize Kaggle platform for 'data mining' competitions rather than 'prediction' competitions.

P7L30-32. I don't think the info presented there is up to date. I believe Kaggle later changed it's way to calculate the leaderboard. For GEFCom2012, the public leaderboard was based on one subset of the test data, while the final leaderboard was based on a different subset of the test data. There is no overlap between the two subsets.

P7L45. 'three' –> 'five'.

P7L56. The statement is true for the wind power forecasting track. The load forecasting track asked the contestants to backcast a few sections of missing data 'and' to forecast one week after the historical period.

P8L13-17. The qualifying match of GEFCom2017 is about ISONE data. The final match involves 183 delivery point meters. See the intro paper I submitted in Dec.

A general comment to Section 4: Maybe you can add a paragraph or a few sentences to comment on how to disseminate the competition results and findings. such as the importance of formal publications after the competitions. There were many other forecasting competitions in the past, but the ones that are really

helpful to the research community are the ones organized with academic journals that can publish findings, data and insights afterwards.

3 Referee 2

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.