On the M4.0 forecasting competition

Nikolopoulos, Konstantinos; Thomakos, D.D.; Katsagounos, Illias; Alghassab, Waleed

International Journal of Forecasting

DOI:
10.1016/j.ijforecast.2019.03.023

Published: 01/03/2020

Peer reviewed version

On the M4.0 forecasting competition: can you tell a 4.0 earthquake from a 3.0?

Abstract

Twenty years on from the publication of the results of the well-celebrated M3 competition, and right about the time we got used to the idea that there will be no more M-type competitions, the M4 competition came in 2019. A 4.0 earthquake is 10 times ‘stronger’ than a 3.0, and that was what M4.0 was aspiring to; mission accomplished?

Keywords: M4 competition; Hybrid method; Combination; Benchmark; Intermittence;

On the M4.0 forecasting competition: can you tell a 4.0 earthquake from a 3.0?

First cut is the deepest

First cut is the deepest and probably no new forecasting competition will ever have the impact of M1 (Makridakis et al. 1982). There are 1360 citations to date in that article and many IIF members argue that the whole discipline is practically an offspring of M1.

The first time you say the story that simplicity matters, and simple models can be as accurate, more robust than complex ones, it breaks the waves: you definitely feel that 'scientific earthquake'. Nevertheless, as Makridakis himself admitted in an interview to Fildes and Nikolopoulos (2006): "*I don't know if there is more work to be done on this type of competitions*".

Strangely enough, the M2 competition that was very different (Makridakis et al. 1993): focusing on non-disguised data and comparing real experts, working in real series, been able to search for whatever information they wanted and even using their judgment to forecast; that is the one that got the least attention (288 citations to date).

Thinner, Lighter, Faster

M1 was almost ten times bigger than its predecessor was. The M3 competition was three times bigger than M1, with more methods and metrics employed. It was indeed impossible to run in real time 3003 long series in the early 80s, but that was definitely doable (over a weekend actually) in one expensive PC in late 90s; today you can probably do that in less than a minute in a 100$ laptop. However, a forecasting competition is not meant to be like a new iPad: thinner, lighter, and faster; it must every time redefine expectations on how empirical forecasting evaluations should be performed.

A new competition

A new forecasting competition cannot just be ten times bigger than the previous one (M3, Makridakis and Hibon, 2000). In order to claim the 4.0 in the long history of forecasting competitions (Hyndman, 2019), M4 brought in new things: far more series, more categories, prediction intervals, replicability, and full transparency. In addition, industry participation for the first time was a major plus; and an open invitation to the machine learning community to really take part in M4.

Reality matters and more can be done

One fundamental question remains unanswered: does M4 represents reality? How do companies really produce forecasts? There is evidence (Fildes and Goodwin, 2007), that forecasts are prepared in practically no time, for thousands of time series, with forecasters being familiar with only a few SKUs, in outdated systems that users often do not trust and override continuously.

Reality matters and our personal take is that blind and static competitions are not fit-for-purpose any more. We need competitions with real series, for products and services known to the participants. We need participants to provide point forecasts and prediction intervals regularly every 2-3 months. That needs commitment, but people in real life do that at much higher frequencies, and they are committed, so it is definitely doable.

It is also very important to focus on the series that really matter in real life. We tend to forecast in vacuum and think that it does not matter if we forecast ‘apples’ or ‘oranges’: but it does. For example, in finance for an investment bank to take investment decisions a set of time series needs to be forecasted regularly. From personal communications with an investment bank based in London we know that many economic and monetary series are monitored in a financial forecasting context. Every trading house uses obviously more or less series, but this is the common denominator in the financial sector. Therefore, size does not matter in the design of the ‘finance’ subset of the next forecasting competition; we need less and named series if we are to move forward, rather than more (and collinear) anonymous series.

Sins of commission

What else real life is? Real life is intermittent: 60% of any inventory consists of spare parts, and these are not cheap to stock. So we do have 60% of SKUs in any warehouse that present intermittent demand patterns but we have decided to ignore such series from our forecasting competitions for the last 40 years. There must be a rational for not including such series, but it looks more like a sin of commission rather than one of omission.

The winner takes it all

The team from Uber led by Smyl is the winner, by a good margin (see the results in table 4 of Makridakis et al. 2019). From second to sixth position we find five different combinations: this is something we expected, maybe not to that extend; in fact in the top-25 positions we find 15 combinations.

In the past M-competitions big private organizations have not participated. They have had in other types of competitions, but not the M-type ones. This time M4 got the attention of the likes of Uber and Amazon and Microsoft, even if not all of them formally participated. The win of Uber also advocates for the fact that there is a lot of forecasting expertise in the practitioners' community. This expertise and research taking place in industry is not scholarly reported in IJF. Uber's method was impressive by itself – a hybrid method, state of the art technically; and intuitively appealing as it exploits properties of the entire dataset every time forecasts are produced for an individual time series.

We also notice that Forecast pro outperforms all benchmarks including the Theta method (Assimakopoulos & Nikolopoulos, 2000), the latter being the only method that performed better than it in M3. There were no articles or announcements in the recent years about any change in the core algorithm of Forecast Pro. The more forecasts needed, the more accurate Forecast Pro becomes, and the selection algorithm it employs eventually outperforms individual methods – even the ones not included in its engine. This is a sign of robustness and consistency and this is all good news for the Forecast Pro team. This is also good news for the entire commercial Forecasting Support Systems development community. We also must congratulate the company for always been willing to test their software in real blind competitions, and face the respective publicity that comes with it.

Omelettes and Eggs

Given that there were so many submissions in the 'combinations' category, and performed so well, it is inevitable to ask the obvious question: who gets the credit? So if someone does an equal weighted combination of Theta method, ARIMA and ETS should the credit go to the one combining, or to those developed those three methods, to both, or none? As the famous football manager Jose Mourinho[1] has nicely once put it:

" *'Omelettes and Eggs': you can not make a good omelette without good eggs…*"

Time is of the Essence

Despite the cloud services and the unlimited computing power than one can buy nowadays, time is still of the essence. It was more of an issue 20 years ago for the M3. Nevertheless, if a method needs 3 days to run in an i7 laptop, while another method runs in 7 minutes or 7 seconds, this arguably constitutes a competitive advantage. A major retailer has only a window a few hours every night in order to forecasts 100K to150K SKUs. Of the M4 more advanced benchmarks, Theta method seems to have the edge running in 12.7 mins for the entire 100K series of M4 dataset in Amazon Web Services with 8 cores, ETS coming second with 888 mins, and ARIMA third with 3030 mins.

The one to beat

Over the years, the IIF community has seen many forecasting studies that proposed new methods that could only outperform Naïve, a moving average or just ETS; this is methodologically wrong, and we should as an academic community work towards banishing this phenomenon. It has been obvious for the last two decades that there is a series of very accurate methods, which are computationally cheap and free in R and Python packages for example Hyndmans's forecast package.

[1] https://www.youtube.com/watch?v=hgGE3VH_LpE

M4 results corroborated grossly to this; in any empirical forecasting investigation, the following methods should be employed as benchmarks - in order of performance in M4 (table 4, Makridakis et al., 2019): the Theta method – even the basic model used in M3 and not one of the advanced ones (Nikolopoulos and Thomakos, 2019), ARIMA, Damped ES and ETS. In addition, combinations should be employed starting with the average of Simple, Holt, and Damped exponential smoothing.

We also propose that we should also use the mean and median of the combination of: Theta method, ARIMA, ETS, and Damped ES. Any newly proposed forecasting method, in order to be publishable, should be on par or better than these 'fast and cheap' benchmarks – and probably even more advanced methods like the awarded MAPA method (Kourentzes et al., 2014); c'est la vie!

Verdict

We really felt this 4.0 'scientific earthquake'.

References

Assimakopoulos, V. and Nikolopoulos, K. (2000). "The Theta Model: A Decomposition Approach to Forecasting", *International Journal of Forecasting* 16 (4): 521-530.

Fildes, R. and Nikolopoulos, K. (2006) "Spyros Makridakis: An Interview with the International Journal of Forecasting". *International Journal of Forecasting* 22(3): 625-636.

Fildes, R. and Goodwin, P. (2007), " Against Your Better Judgment? How Organizations Can Improve Their Use of Management Judgment in Forecasting", Interfaces 37(6), 570-576.

Hyndman, R. J. (2019), "A brief history of forecasting competitions" *International Journal of Forecasting*, forthcoming

Kourentzes,N., Petropoulos,F., & Trapero, J.R. (2014). "Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30(2), 291–302.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting, 1, 111–153.

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, 5–23.

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

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2019), "The M4 Competition: 100,000 Time Series And 61 Forecasting Methods", *International Journal of Forecasting*, forthcoming

Nikolopoulos, K. and Thomakos, D. D. (2019), *Forecasting with the Theta Method: Theory Applications*., Wiley: New Jersey.