International Journal of Forecasting 36 (2020) 203–205

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Discussion

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

Konstantinos Nikolopoulos [a,*], Dimitrios D. Thomakos [b], Ilias Katsagounos [b], Waleed Alghassab [c]

[a] *forLAB, the Forecasting Laboratory www.forLAb.eu, Bangor Business School, Prifysglo Bangor University, LL57 2DG, Bangor, Gwynedd, Wales, UK*
[b] *Department of Economics, University of Peloponnese, Tripolis, 22100, Greece*
[c] *University of Hail Building, Agrigultural Plan Road, 2440, Hail 11415, Saudi Arabia*

ABSTRACT

Twenty years on from the publication of the results of the celebrated M3 competition and we were just about used to the idea that there would be no more M-type competitions, when the M4 competition came along in 2019. A 4.0 earthquake is 10 times 'stronger' than a 3.0, and that was what M4.0 was aspiring to; so was its mission accomplished?

### *First cut is the deepest*

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

The first time you say that simplicity matters, and that simple models can be as accurate as complex ones, and more robust, it makes waves: you definitely feel the '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 competition in the M-series that has received the least attention (288 citations to date) is the M2 competition, which was very different from the others (Makridakis et al., 1993): it focused on non-disguised data and compared real experts, working in real series, who were able to search for whatever information they wanted and even use their judgment to forecast.

### *Thinner, lighter, faster*

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

### *A new competition*

A new forecasting competition cannot merely be ten times as big as the previous one (M3, Makridakis & Hibon, 2000). In order to claim the 4.0 in the long history of forecasting competitions (Hyndman, 2020), M4 introduced 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, as was the open invitation to the machine learning community to really take part in M4.

### *Reality matters and more can be done*

However, one fundamental question remains unanswered: does M4 represent reality? How do companies

[*] Corresponding author.
*E-mail address:* k.nikolopoulos@bangor.ac.uk (K. Nikolopoulos).

https://doi.org/10.1016/j.ijforecast.2019.03.023

really produce forecasts? There is evidence (Fildes & Goodwin, 2007) that in many cases forecasts are prepared in practically no time, for thousands of time series, of which the forecasters are familiar with only a few SKUs, in outdated systems that users often do not trust and therefore override continuously.

Reality matters, and our personal take is that blind and static competitions are no longer fit-for-purpose. We need competitions with real series, for products and services that are known to the participants. We need participants to provide point forecasts and prediction intervals regularly every 2–3 months. That takes commitment, but people in real life do it at much higher frequencies, and they are committed, so it is definitely doable.

It is also very important to focus on the sorts of series that really matter in real life. We tend to forecast in vacuum and think that it does not matter whether we are forecasting 'apples' or 'oranges', but it does. For example, in finance, an investment bank must forecast a set of time series regularly in order to make investment decisions. 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. Obviously, different trading houses use different numbers of series, but this is the common denominator in the financial sector. Thus, size is not important in the design of the 'finance' subset of the next forecasting competition; we need fewer and named series if we are to move forward, rather than more (and collinear) anonymous series.

### Sins of commission

What else is real life like? Real life is intermittent: 60% of any inventory consists of spare parts, and these are not cheap to stock. Thus, 60% of SKUs in any warehouse present intermittent demand patterns, but we have decided to ignore such series consistently in our forecasting competitions over the last 40 years. There must be a rationale 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, Spiliotis, & Assimakopoulos, 2020). Then, from second to sixth positions, we find five different combinations. This is something that we expected, though maybe not to that extent; in fact, we find 15 combinations in the top 25 positions.

Big private organizations have not participated in past M-competitions. They have participated in other types of competitions, but not the M-series ones. This time, though, M4 got the attention of the likes of Uber and Amazon and Microsoft, even if not all of them participated formally. The win of Uber also proves that there is a lot of forecasting expertise in the practitioners' community. This expertise and research that is taking place in industry is not reported in scholarly publications like *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 that forecasts are produced for an individual time series.

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

### Omelettes and eggs

Given that there were so many submissions in the 'combinations' category, and that they performed so well, it is inevitable to ask the obvious question: who should get the credit? That is, if someone does an equal-weighted combination of Theta method, ARIMA and ETS, for example, should the credit go to the one combining, to those who developed the three constituent methods, to both, or to no one? As the famous football manager Jose Mourinho[1] once nicely put it:

> " *'Omelettes and Eggs': you cannot make a good omelette without good eggs. . .*"

### Time is of the essence

Despite the cloud services and the unlimited computing power that one can buy nowadays, time is still of the essence. It was more of an issue 20 years ago for the M3; nevertheless, if one method takes three days to run on an i7 laptop while another method runs in seven minutes or seven seconds, it could be argued that this constitutes a competitive advantage for the latter. A major retailer has a window of only a few hours every night in order to forecast 100k to 150k SKUs. Of the M4's more advanced benchmarks, the Theta method seems to have the edge, running in 12.7 min for the entire 100k series of the M4 dataset in Amazon Web Services with 8 cores, while ETS came second at 888 min, and ARIMA third at 3030 min.

### The one to beat

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

The M4 results grossly corroborated this; in any empirical forecasting investigation, the following methods should be employed as benchmarks, in order of their

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

performance in M4 (Table 4, Makridakis et al., 2020): the Theta method – even just the basic model used in M3 rather than one of the advanced ones (Nikolopoulos & 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 the use of the mean and median of the combination of the Theta method, ARIMA, ETS, and Damped ES. Thus, in order to be publishable, any newly proposed forecasting method should have to *be on par with or better than* these 'fast and cheap' benchmarks – and probably even more advanced methods like the awarded MAPA method (Kourentzes, Petropoulos, & Trapero, 2014); c'est la vie!

### *Verdict*

We really felt this 4.0 'scientific earthquake'.

## References

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*(4), 521–530.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? how organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*(6), 570–576.

Fildes, R., & Nikolopoulos, K. (2006). Spyros makridakis: an interview with the international journal of forecasting. *International Journal of Forecasting*, *22*(3), 625–636.

Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36*(1), 7–14.

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., et al. (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., et al. (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. (2020). The M4 competition: 100, 000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74.

Nikolopoulos, K., & Thomakos, D. D. (2019). *Forecasting with the theta method: theory applications*. New Jersey: Wiley.