The M4 Competition: Results, findings, conclusion and way forward

Article *in* International Journal of Forecasting · June 2018

DOI: 10.1016/j.ijforecast.2018.06.001

CITATIONS
35

READS
5,584

3 authors:

Spyros Makridakis
University of Nicosia
204 PUBLICATIONS **9,391** CITATIONS
SEE PROFILE

Evangelos Spiliotis
National Technical University of Athens
30 PUBLICATIONS **112** CITATIONS
SEE PROFILE

Vassilis Assimakopoulos
National Technical University of Athens
84 PUBLICATIONS **997** CITATIONS
SEE PROFILE

ARTICLE IN PRESS

International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

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

The M4 Competition: Results, findings, conclusion and way forward

Spyros Makridakis [a,b,*], Evangelos Spiliotis [c], Vassilios Assimakopoulos [c]

[a] University of Nicosia, Nicosia, Cyprus
[b] Institute For the Future (IFF), Nicosia, Cyprus
[c] Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Zografou, Greece

ARTICLE INFO

Keywords:
Forecasting competitions
M Competitions
Forecasting accuracy
Prediction intervals (PIs)
Time series methods
Machine Learning (ML) methods
Benchmarking methods
Practice of forecasting

ABSTRACT

The M4 competition is the continuation of three previous competitions started more than 45 years ago whose purpose was to learn how to improve forecasting accuracy, and how such learning can be applied to advance the theory and practice of forecasting. The purpose of M4 was to replicate the results of the previous ones and extend them into three directions: First significantly increase the number of series, second include Machine Learning (ML) forecasting methods, and third evaluate both point forecasts and prediction intervals. The five major findings of the M4 Competitions are: 1. Out Of the 17 most accurate methods, 12 were "combinations" of mostly statistical approaches. 2. The biggest surprise was a "hybrid" approach that utilized both statistical and ML features. This method's average sMAPE was close to 10% more accurate than the combination benchmark used to compare the submitted methods. 3. The second most accurate method was a combination of seven statistical methods and one ML one, with the weights for the averaging being calculated by a ML algorithm that was trained to minimize the forecasting. 4. The two most accurate methods also achieved an amazing success in specifying the 95% prediction intervals correctly. 5. The six pure ML methods performed poorly, with none of them being more accurate than the combination benchmark and only one being more accurate than Naïve2. This paper presents some initial results of M4, its major findings and a logical conclusion. Finally, it outlines what the authors consider to be the way forward for the field of forecasting.

1. Introduction

The M4 Competition ended on May 31, 2018. By the deadline, we had received 50 valid submissions predicting all 100,000 series and 18 valid submissions specifying all 100,000 prediction intervals (PIs). The numbers of submissions are a little over 20% and 7%, respectively, of the 248 registrations to participate. We assume that 100,000 series was a formidable number that discouraged many of those who had registered from generating and submitting forecasts, especially those who were utilizing complex machine learning (ML) methods that required demanding computations. (A participant, living in California, told us of his huge electricity bill electricity bill from using his 5 home computers running almost constantly over 4.5 months.)

After a brief introduction concerning the objective of the M4, this short paper consists of two parts. First, we present the major findings of the competition and discuss their implications for the theory and practice of forecasting. A special issue of the IJF is scheduled to be published next year that will cover all aspects of the M4 in detail, describing the most accurate methods and identifying the reasons for their success. It will also consider the factors that had negative influences on the performances of the

* Corresponding author.
E-mail address: makridakis.s@unic.ac.cy (S. Makridakis).

https://doi.org/10.1016/j.ijforecast.2018.06.001
0169-2070/

majority of the methods submitted (33 of the 50), whose accuracies were lower than that of the statistical benchmark used, and will outline what we consider to be the way forward for the field. Second, we briefly discuss some predictions/hypotheses that we made two and a half months ago (detailed in a document e-mailed to R. J. Hyndman on 21/3/2018), predicting/hypothesizing about the actual results of the M4. A detailed appraisal of these predictions/hypotheses will be left for the special issue.

We would like to emphasize that the most important objective of the M4 Competition, like that of the previous three ones, has been to "learn how to improve the forecasting accuracy, and how such learning can be applied to advance the theory and practice of forecasting", thus providing practical benefits to all those who are interested in the field. The M4 is an open competition whose series have been available since the end of last year, both through the M4 site (https://www.m4.unic.ac.cy/the-dataset/) and in the *M4comp2018* R package (Montero-Manso, Netto, & Talagala, 2018). In addition, the majority of the participants have been requested to deposit the code used for generating their forecasts in GitHub (https://github.com/M4Competition/M4-methods), along with a detailed description of their methods, while the code for the ten M4 benchmarks has been available from GitHub since the beginning of the competition. This will allow interested parties to verify the accuracy and/or reproducibility of the forecasts of most of the methods submitted to the competition (for proprietary methods/software, the checking will be done by the organizers). This implies that individuals and organizations will be able to download and utilize most of the M4 methods, including the best one(s), and to benefit from their enhanced accuracy. Moreover, academic researchers will be able to analyze the factors that affect the forecasting accuracy in order to gain a better understanding of them, and to conceive new ways of improving the accuracy. This is in contrast to other competitions, like those organized by Kaggle, where there is actually a "horse race" that aims to identify the most accurate forecasting method(s) without any consideration of the reasons involved with the aim of improving the forecasting performance in the future. To recapitulate, the objective of the M Competitions is to learn how to improve the forecasting accuracy, not to host a "horse race".

2. The five major findings and the conclusion of M4

Below, we outline what we consider to be the five major findings of the M4 Competition and advance a logical conclusion from these findings.

- The combination of methods was the king of the M4. Of the 17 most accurate methods, 12 were "combinations" of mostly statistical approaches.
- The biggest surprise was a "*hybrid*" approach that utilized both statistical and ML features. This method produced both the most accurate forecasts and the most precise PIs, and was submitted by Slawek Smyl, a Data Scientist at Uber Technologies. According to sMAPE, it was close to 10% more accurate than the combination (*Comb*) benchmark of the competition (see below), which is a huge improvement. It is noted that the best method in the M3 Competition (Makridakis & Hibon, 2000) was only 4% more accurate than the same combination.
- The second most accurate method was a combination of seven statistical methods and one ML one, with the weights for the averaging being calculated by a ML algorithm that was trained to minimize the forecasting error through holdout tests. This method was submitted jointly by Spain's University of A Coruña and Australia's Monash University.
- The most accurate and second most accurate methods also achieved an amazing success in specifying the 95% PIs correctly. These are the first methods we are aware of that have done so, rather than underestimating the uncertainty considerably.
- The six pure ML methods that were submitted in the M4 all performed poorly, with none of them being more accurate than Comb and only one being more accurate than Naïve2. These results are in agreement with those of a recent study that was published in PLOS ONE (Makridakis, Spiliotis, & Assimakopoulos, 2018).

Our conclusion from the above findings is that the accuracy of individual statistical or ML methods is low, and that hybrid approaches and combinations of method are the way forward for improving the forecasting accuracy and making forecasting more valuable.

2.1. The comb benchmark

One innovation of the M4 was the introduction of ten standard methods (both statistical and ML) for benchmarking the accuracy of the methods submitted to the competition. From those ten, we selected one against which to compare the submitted methods, namely Comb, a combination based on the simple arithmetic average of the Simple, Holt and Damped exponential smoothing models. The reason for such a choice was that Comb is easy to compute, can be explained/understood intuitively, is quite robust, and typically is more accurate than the individual methods being combined. For instance, for the entire set of the M4 series, the sMAPE of Single was 13.09%, that of Holt was 13.77%, and that of Damped was 12.66%, while that of Comb was 12.55%. Table 1 shows the sMAPE and the overall weighted average (OWA) – for definitions, see M4 Team (2018) – of the Comb benchmark, the 17 methods that performed better than such a benchmark, and the two most accurate ML submissions. The last two columns of Table 1 show the percentage that each method was better/worse than Comb in terms of both sMAPE and OWA.

2.2. The six methods with the lowest OWA and the six ML ones

We planned originally to discuss only the five best methods submitted for the competition, but later decided to increase the number to six, as the difference between the fifth and sixth was miniscule (0.005). Of these six methods, the most accurate one was the hybrid approach proposed by Smyl of Uber, which mixed exponential smoothing formulas with a "black-box" recurrent neural network (RNN)

Please cite this article in press as: Makridakis, S., et al., The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2018.06.001.

Table 1
The performances of the 17 methods submitted to M4 that had OWAs at least as good as that of the Comb.

Type	Author(s)	Affiliation	sMAPE[c]					OWA[d]					Rank[a]	% improvement of method over the benchmark	
			Yearly (23k)	Quarterly (24k)	Monthly (48k)	Others (5k)	Average (100k)	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Others (5k)	Average (100k)			
Benchmark for methods[b]			14.848	10.175	13.434	4.987	12.555	0.867	0.890	0.920	1.039	0.898	19	sMAPE	OWA
Hybrid	Smyl, S.	Uber Technologies	13.176	9.679	12.126	4.014	11.374	0.778	0.847	0.836	0.920	0.821	1	9.4%	8.6%
Combination	Montero-Manso, P., Talagala, T., Hyndman, R. J. & Athanasopoulos, G.	University of A Coruña & Monash University	13.528	9.733	12.639	4.118	11.720	0.799	0.847	0.858	0.914	0.838	2	6.6%	6.7%
Combination	Pawlikowski, M., Chorowska, A. & Yanchuk, O.	ProLogistica Soft	13.943	9.796	12.747	3.365	11.845	0.820	0.855	0.867	0.742	0.841	3	5.7%	6.3%
Combination	Jaganathan, S. & Prakash, P.	Individual	13.712	9.809	12.487	3.879	11.695	0.813	0.859	0.854	0.882	0.842	4	6.8%	6.2%
Combination	Fiorucci, J. A. & Louzada, F.	University of Brasilia & University of São Paulo	13.673	9.816	12.737	4.432	11.836	0.802	0.855	0.868	0.935	0.843	5	5.7%	6.1%
Combination	Petropoulos, F. & Svetunkov, I.	University of Bath & Lancaster University	13.669	9.800	12.888	4.105	11.887	0.806	0.853	0.876	0.906	0.848	6	5.3%	5.6%
Combination	Shaub, D.	Harvard Extension School	13.679	10.378	12.839	4.400	12.020	0.801	0.908	0.882	0.978	0.860	7	4.3%	4.2%
Statistical	Legaki, N. Z. & Koutsouri, K.	National Technical University of Athens	13.366	10.155	13.002	4.682	11.986	0.788	0.898	0.905	0.989	0.861	8	4.5%	4.1%
Combination	Doornik, J., Castle, J. & Hendry, D.	University of Oxford	13.910	10.000	12.780	3.802	11.924	0.836	0.878	0.881	0.874	0.865	9	5.0%	3.7%
Combination	Pedregal, D.J., Trapero, J. R., Villegas, M. A. & Madrigal, J. J.	University of Castilla-La Mancha	13.821	10.093	13.151	4.012	12.114	0.824	0.883	0.899	0.895	0.869	10	3.5%	3.2%
Statistical	Spiliotis, E. & Assimakopoulos, V.	National Technical University of Athens	13.804	10.128	13.142	4.675	12.148	0.823	0.889	0.907	0.975	0.874	11	3.2%	2.7%
Combination	Roubinchtein, A.	Washington State Employment Security Department	14.445	10.172	12.911	4.436	12.183	0.850	0.885	0.881	0.992	0.876	12	3.0%	2.4%
Other	Ibrahim, M.	Georgia Institute of Technology	13.677	10.089	13.321	4.747	12.198	0.805	0.890	0.921	1.079	0.880	13	2.8%	2.0%
Combination	Kull, M., et al.	University of Tartu	14.096	11.109	13.290	4.169	12.496	0.820	0.960	0.932	0.883	0.888	14	0.5%	1.1%
Combination	Waheeb, W.	Universiti Tun Hussein Onn Malaysia	14.783	10.059	12.770	4.039	12.146	0.880	0.880	0.927	0.904	0.894	15	3.3%	0.4%
Statistical	Darin, S. & Stellwagen, E.	Business Forecast Systems (Forecast Pro)	14.663	10.155	13.058	4.041	12.279	0.877	0.887	0.887	1.011	0.895	16	2.2%	0.3%
Combination	Dantas, T. & Oliveira, F.	Pontifical Catholic University of Rio de Janeiro	14.746	10.254	13.462	4.783	12.553	0.866	0.892	0.914	1.011	0.896	17	0.0%	0.2%
Best ML	Trotta, B.	Individual	14.397	11.031	13.973	4.566	12.894	0.859	0.939	0.941	0.991	0.915	23	−2.7%	−1.9%
2nd Best ML	Bontempi, G.	Université Libre de Bruxelles	16.613	11.786	14.800	4.734	13.990	1.050	1.072	1.007	1.051	1.045	37	−11.4%	−16.4%

[a] Rank: rank calculated considering both the submitted methods (50) and the set of benchmarks (10).

[b] Com: the benchmark is the average (combination) of Simple, Holt and Damped exponential smoothing.

[c] sMAPE: symmetric mean absolute percentage error.

[d] OWA: overall weighted average of the relative sMAPE and the relative MASE.

forecasting engine. It is also interesting that this forecasting method incorporated information from both the individual series and the whole dataset, thus exploiting data in a hierarchical way. The remaining five methods all had something in common: they utilized various forms of combining mainly statistical methods, and most used holdout tests to figure out the optimal weights for such a combination. The differences in forecasting accuracy between these five methods were small, with the sMAPE difference being less than 0.2% and the OWA difference being just 0.01. It is also interesting to note that although these five methods used a combination of mostly statistical models, some of them utilized ML ideas for determining the optimal weights for averaging the individual forecasts.

The rankings of the six ML methods were 23, 37, 38, 48, 54, and 57 out of a total of 60 (50 submitted and the 10 benchmarks). These results confirm our recent findings (Makridakis et al., 2018) regarding the poor performances of the ML methods relative to statistical ones. However, we must emphasize that the performances of the ML methods reported in our paper, as well as those found by Ahmed, Atiya, Gayar, and El-Shishiny (2010) and those submitted in the M4, used mainly generic guidelines and standard algorithms that could be improved substantially. One possible direction for such an improvement could be the development of hybrid approaches that utilize the best characteristics of ML algorithms and avoid their disadvantages, while at the same time exploiting the strong features of well-known statistical methods. We believe strongly in the potential of ML methods, which are still in their infancy as far as time series forecasting is concerned. However, to be able to reach this potential, their existing limitations, such as preprocessing and overfitting, must be identified and most importantly accepted by the ML community, rather than being ignored. Otherwise, no effective ways of correcting the existing problems will be devised.

2.3. Prediction intervals (PIs)

Our knowledge of PIs is rather limited, and much could be learned by studying the results of the M4 Competition. Table 2 presents the performance of the 12 methods (out of the 18 submitted) that achieved mean scaled interval scores (MSISs), (see M4 Team, 2018, for definitions) that were at least as good as that of the Naïve1 method. The last two columns of Table 2 show the percentages of cases in which each method was better/worse than the Naive1 in terms of MSIS and the absolute coverage difference (ACD); i.e., the absolute difference between the coverage of the method and the target set of 0.95. The performances of two additional methods, namely ETS and auto.arima in the R *forecast* package (Hyndman et al., 2018), have also been included as benchmarks, but Naïve1 was preferred due to its simplicity, low computational requirements and intuitive understanding. Some of the results in Table 2 are impressive. The improvements in the performances of the top five methods range from almost 50% for the first to nearly 34% for the fifth. Moreover, the coverages achieved by the methods of Smyl and Montero-Manso et al. are 0.948 and 0.960 respectively (compared to a perfect score of 0.95), meaning that both of them did an excellent job of estimating the forecasting uncertainty. Once again, the hybrid and combination approaches led to the most correct PIs, with the statistical methods displaying significantly worse performances and none of the ML methods managing to outperform the benchmark. However, it should be mentioned that, unlike the best ones, the majority of the methods underestimated reality considerably, especially for the low frequency data (yearly and quarterly). For example, the average coverages of the methods were 0.80, 0.90, 0.92 and 0.93 for the yearly, quarterly, monthly and other data, respectively. These findings demonstrate that standard forecasting methods fail to estimate the uncertainty properly, and therefore that more research is required to correct the situation.

3. The hypotheses

We now present the ten predictions/hypotheses we sent to the IJF Editor-in-Chief two and a half months ago about the expected results of the M4. A much fuller and more detailed appraisal of these ten hypotheses is left for the planned special issue of IJF. This section merely provides some short, yet indicative answers.

1. **The forecasting accuracies of simple methods, such as the eight statistical benchmarks included in M4, will not be too far from those of the most accurate methods.**
   This is only partially true. On the one hand, the hybrid method and several of the combination methods all outperformed Comb, the statistical benchmark, by a percentage ranging between 9.4% for the first and 5% for the ninth. These percentages are higher than the 4% improvement of the best method in the M3 relative to the Comb benchmark. On the other hand, the improvements below the tenth method were lower, and one can question whether the level of improvement was exceptional, given the advanced algorithms utilized and the computational resources used.
2. **The 95% PIs will underestimate reality considerably, and this underestimation will increase as the forecasting horizon lengthens.**
   Indeed, except for the two top performing methods of the competition, the majority of the remaining methods underestimated reality by providing overly narrow confidence intervals. This is especially true for the low frequency data (yearly and quarterly), where longer-term forecasts are required, and therefore the uncertainty is higher. Increases in data availability and the generation of short-term forecasts may also contribute to the improvements displayed in the case of the high frequency series. It is also verified that underestimation increased on average at longer forecasting horizons for the yearly, quarterly and monthly data, while it remained rather constant for the rest of the data.
3. **The upward/downward extrapolation of the trend will be predicted more accurately when it is damped for longer horizons.**

Please cite this article in press as: Makridakis, S., et al., The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2018.06.001.

Table 2
The performances of the 12 methods submitted to M4 with MSISs at least as good as that of the Naïve.

Type	Author(s)	Affiliation	MSIS[c]					ACD[d]					Rank[a]	% improvement of method over the benchmark	
			Yearly (23k)	Quarterly (24k)	Monthly (48k)	Others (5k)	Average (100k)	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Others (5k)	Average (100k)			
Naïve 1: Benchmark for methods[b]			56.554	14.073	12.300	35.311	24.055	0.234	0.084	0.023	0.019	0.086	15	MSIS	ACD
Hybrid	Smyl, S.	Uber Technologies	23.898	8.551	7.205	24.458	12.230	0.003	0.004	0.005	0.001	0.002	1	49.2%	97.4%
Combination	Montero-Manso, P., Talagala, T., Hyndman, R. J. & Athanasopoulos, G.	University of A Coruña & Monash University	27.477	9.384	8.656	32.142	14.334	0.014	0.016	0.016	0.027	0.010	2	40.4%	88.8%
Combination	Doornik, J., Castle, J. & Hendry, D.	University of Oxford	30.200	9.848	9.494	26.320	15.183	0.037	0.029	0.054	0.039	0.043	3	36.9%	50.0%
Combination	Fiorucci, J. A. & Louzada, F.	University of Brasilia & University of São Paulo	35.844	9.420	8.029	26.709	15.695	0.164	0.056	0.028	0.005	0.065	5	34.8%	24.6%
Combination	Petropoulos, F. & Svetunkov, I.	University of Bath & Lancaster University	35.945	9.893	8.230	27.780	15.981	0.171	0.064	0.035	0.012	0.072	6	33.6%	16.4%
Combination	Roubinchtein, A.	Washington State Employment Security Department	37.706	9.945	8.233	29.866	16.505	0.167	0.049	0.021	0.004	0.061	7	31.4%	29.4%
Statistical	Talagala, T., Athanasopoulos, G. & Hyndman, R. J.	Monash University	39.793	11.240	9.825	37.222	18.427	0.165	0.074	0.056	0.048	0.085	8	23.4%	0.9%
Other	Ibrahim, M.	Georgia Institute of Technology	45.028	11.284	9.390	52.604	20.202	0.211	0.086	0.050	0.011	0.094	10	16.0%	−9.1%
Statistical	Iqbal, A.,Seery, S. & Silvia, J.	Wells Fargo Securities	53.469	11.295	11.159	32.721	22.001	0.214	0.047	0.048	0.050	0.086	11	8.5%	0.1%
Combination	Reilly, T.	Automatic Forecasting Systems, Inc. (AutoBox)	53.998	13.537	10.344	34.680	22.367	0.293	0.102	0.057	0.046	0.121	12	7.0%	−41.1%
Statistical	Wainwright, E., Butz, E. & Raychaudhuri, S.	Oracle Corporation	58.596	12.426	9.714	31.036	22.674	0.295	0.100	0.056	0.028	0.120	13	5.7%	−39.6%
Combination	Segura-Heras, J.-V., Vercher-González, E., Bermúdez-Edo, J. D. & Corberán-Vallet, A.	Universidad Miguel Hernández & Universitat de Valencia	52.316	15.058	11.220	33.685	22.717	0.132	0.039	0.014	0.052	0.049	14	5.6%	43.1%

[a] Rank: rank calculated considering both the submitted methods (18) and the set of benchmarks (3).

[b] Naive: the benchmark is the Naïve 1 method.

[c] MSIS: mean scaled interval score.

[d] ACD: absolute coverage difference, i.e., the absolute difference between the coverage of the method and the target (0.95).

This is verified through the performances of the respective M4 benchmarks. Holt exponential smoothing displays an OWA of 0.971 (30th position), while Damped has an OWA of 0.907 (21st position), thus improving the forecasting accuracy by more than 6.5% when extrapolating using damped instead of linear trends.

4. **The majority of ML methods will not be more accurate than statistical ones, although there may be one or two pleasant surprises where such methods are superior to statistical ones, though only by some small margin.**
   This was one of the major findings of the competition, the demonstration that none of the pure ML used managed to outperform Comb and that only one of them was more accurate than the Naive2. The one pleasant surprise was the hybrid method introduced by Smyl, which incorporated features from both statistical and ML methods and led to a considerable improvement in forecasting accuracy.
5. **The combination of statistical and/or ML methods will produce more accurate results than the best of the individual methods combined.**
   This is a major finding, or better, a confirmation, from this competition, the demonstration that combination is one of the best forecasting practices. Since the majority of the participants who combined various methods used more or less the same pool of models (most of which were the M4 statistical benchmarks), an interesting question is, "which features make a combination approach more accurate than others?"
6. **Seasonality will continue to dominate the fluctuations of the time series, while randomness will remain the most critical factor influencing the forecasting accuracy.**
   Having correlated the sMAPE values reported per series for the 60 methods utilized within M4 with the corresponding estimate of various time series characteristics (Kang, Hyndman, & Smith-Miles, 2017), such as randomness, trend, seasonality, linearity and stability, our initial finding is that, on average, randomness is the most critical factor for determining the forecasting accuracy, followed by linearity. We also found that seasonal time series tend to be easier to predict. This last point is a reasonable statement if we consider that seasonal time series are likely to be less noisy.
7. **The sample size will not be a significant factor in improving the forecasting accuracy of statistical methods.**
   After correlating the sMAPE values reported for each series with the lengths of these series, it seems that the forecasting accuracy is related weakly to the sample size. Thus, we conclude that statistical methods are not influenced heavily by data availability (though this finding needs further confirmation).
8. **There will be small, not statistically significant, differences in the forecasting accuracies of the various methods on the economic/business series in the M3 and M4 datasets.**
   This prediction/hypothesis is related closely to the following two claims (9 and 10). If the characteristics of the M3 and M4 series are similar, then the forecasting performances of the various participating methods will be close, especially for economic/business data, where the data availability and frequency of the two datasets are comparable. However, excessive experiments are required in order to answer this question, and therefore no firm conclusion can be drawn at this point. We can only state that this is quite true for the case of the statistical M4 benchmarks.
9. **We would expect few, not statistically important, differences in the characteristics of the series used in the M3 and M4 Competitions in terms of their seasonality, trend, trend-cycle and randomness.**
   An analysis of the characteristics of the time series (Kang et al., 2017) in each of the two datasets demonstrated that the basic features of the series were quite similar, though the M4 data were slightly more trended and less seasonal that those of M3. However, a more detailed analysis will be performed to verify these preliminary findings, taking into consideration possible correlations that may exist between such features. This conclusion is also supported through claim 10, as time series characteristics are related closely to the performances of forecasting methods (Petropoulos, Makridakis, Assimakopoulos, & Nikolopoulos, 2014).
10. **There will be small, not statistically significant, differences in the accuracies of the eight statistical benchmarks between the M3 and M4 datasets.**
    Similarities were evident when the performances of the M4 benchmarks were compared across the M3 and M4 datasets. Not only did the ranks of the benchmarks remain almost unchanged (Spearman's correlation coefficient was 0.95), but also the absolute values of the errors (sMAPE, MASE and OWA) were very close. This is an indication that both the M3 and M4 data provide a good representation of the forecasting reality in the business world.

4. The way forward

Forecasting competitions provide the equivalent of lab experiments in physics or chemistry. The problem is that sometimes we are not open-minded, but instead put our biases and vested interests above the objective evidence from the competitions (Makridakis, Assimakopoulos et al., in press; Makridakis & Hibon, 1979). The M1 Competition (Makridakis, Andersen, Carbone, Fildes, Hibon, et al., 1982) proved that simple methods were at least as accurate as statistically sophisticated ones. M2 (Makridakis, Chatfield, Hibon, Lawrence, et al., 1993) demonstrated that the use of additional information and judgment did not improve the accuracy of time series forecasting. M3 reconfirmed the findings of M1 and introduced a couple of new, more accurate methods. Furthermore, even after it

Please cite this article in press as: Makridakis, S., et al., The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2018.06.001.

ended, it still led other researchers to experiment with the M3 data and propose additional, more accurate forecasting approaches. The major contributions of M4 have been (a) the introduction of a "hybrid" method, (b) the confirmation of the superiority of the combinations and the incorporation of ML algorithms for determining their weighting, and thus improving their accuracy, and (c) the impressive performances of some methods for determining PIs.

Rationality suggests that one must accept that all forecasting approaches and individual methods have both advantages and drawbacks, and therefore that one must be eclectic in exploiting such advantages while also avoiding or minimizing their shortcomings. This is why the logical way forward is the utilization of hybrid and combination methods. We are confident that the 100,000 time series of the M4 will become a testing ground on which forecasting researchers can experiment and discover new, increasingly accurate forecasting approaches. Empirical evidence must guide our efforts to improve the forecasting accuracy, rather than personal opinions.

References

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, *29*(5–6), 594–621.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., & O'Hara-Wild, M. et al., (2018). Forecast: Forecasting functions for time series and linear models. R package version 8.3.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*(2), 345–358.

M4 Team (2018). M4 competitor's guide: prizes and rules. See https://www.m4.unic.ac.cy/wp-content/uploads/2018/03/M4-Competitors-Guide.pdf.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., et al. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.

Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2018). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34*(3) (in press).

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., et al. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*(1), 5–22.

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: an empirical investigation. *Journal of the Royal Statistical Society, Series A (General)*, *142*(2), 97–145.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: concerns and ways forward. *PLOS ONE*, *13*(3), 1–26.

Montero-Manso, P., Netto, C., & Talagala, T. (2018). M4comp2018: Data from the M4-Competition. R package version 0.1.0.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). 'Horses for courses' in demand forecasting. *European Journal of Operational Research*, *237*(1), 152–163.

Please cite this article in press as: Makridakis, S., et al., The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2018.06.001.