Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

Predicting/hypothesizing the findings of the M4 Competition

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

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

ABSTRACT

Science is caught up in a replication crisis which has negative implications for published findings that cannot be reproduced by other researchers. However, such is not the case with the M4 Competition, which not only provided the means of effectively reproducing its submissions, but also preregistered ten predictions/hypotheses about its expected results two-and-a-half months before its completion. From a scientific point of view, attempting to predict the results of a study is far more powerful than merely justifying them in hindsight after they have become available. The present paper presents these ten predictions/hypotheses that the organizers of the M4 Competition made and evaluates them based on the actual results. It is shown that at least six of the ten predictions/hypotheses were entirely correct, while two were partially correct, one required additional information to be confirmed, and the remaining one was not predicted correctly.

1. Introduction

The M4 Competition (Makridakis, Spiliotis, & Assimakopoulos, 2020) is a continuation of the three previous ones organized by Spyros Makridakis, with the aim of learning from empirical evidence how to improve the forecasting accuracy and how such learning can be applied to advance the theory and practice of forecasting. These competitions were open and have attempted to be easy to reproduce, thus avoiding the replicability crisis that is affecting the social sciences (Baker, 2016; Ball, 2018; Camerer et al., 2018). They have attracted a great deal of interest both in the academic community and among practitioners, serving as a benchmark and a standard testing ground for large numbers of forecasters around the world, thus contributing to the advancement of the field. More importantly, they have provided objective evidence on the most appropriate ways of forecasting, focusing on the methods' actual accuracies rather than their mathematical properties (Hyndman, 2020).

The purpose of this paper is to present the ten predictions/hypotheses that the organizers of M4 made more than two months before its completion, confirming their confidence in the expected results (Makridakis, Spiliotis, & Assimakopoulos, 2018a) and making a clear statement of their expectations, thus avoiding the problem of rationalizing the findings after the fact. This was part of an attempt to go a step beyond the reproducibility issue and predict the actual findings of a large-scale forecasting experiment. These predictions/hypotheses have now been evaluated. The results show that, of the ten predictions/hypotheses made, at least six were entirely correct, two were partially correct, one required additional information to be properly confirmed, and the remaining one was incorrect.

2. Our predictions/hypotheses

Our ten predictions/hypotheses were deposited with the Editor-in-Chief of the *International Journal of Forecasting* (IJF), Rob J. Hyndman, two-and-a-half months before the end of the M4 Competition. In summary, our first prediction was incorrect, as the accuracies of the six top-performing methods in terms of point forecasts were considerably better than that of the Combination (Comb) benchmark (the simple arithmetic average of Single, Holt and Damped exponential smoothing) that was used for comparing all of the submitted methods. This was due to the one or two pleasant surprises that we expected

[*] Corresponding author.
E-mail address: spiliotis@fsu.gr (E. Spiliotis).

https://doi.org/10.1016/j.ijforecast.2019.02.012

according to our fourth prediction, including Smyl (2020) hybrid method, as well as a number of highly accurate methods that utilized innovative ways of determining the weights for combining the various methods and exploited information from multiple series. Six of our predictions were definitely confirmed, and the remaining three were in the grey area. Below is a more detailed evaluation of each of the predictions/hypotheses we made. Additional information about the submitted methods and their performances, as well as about the measures and benchmarks used for their evaluation, can be found in the paper by Makridakis et al. (2020).

(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.

In the M3 competition (Makridakis & Hibon, 2000), the most accurate method in terms of sMAPE (symmetric mean absolute error), namely Theta (Assimakopoulos & Nikolopoulos, 2000), was 3.8% more accurate than the Comb benchmark. Thus, we hypothesized that the difference between Comb and the top-performing method of M4 would be slightly higher. We did not expect however that the OWA (overall weighted average, used to measure the accuracy of point forecasts in M4) of Smyl's hybrid method would be 8.4% more accurate than the Comb, or that the following five methods would be more than 5% more accurate (Makridakis et al., 2018a). These results indicate that we underestimated recent advances in the fields of forecasting, computer science and machine learning (ML), which have brought some notable improvements in forecasting performance, although at the cost of greater computer execution times and higher levels of complexity.

In this regard, an interesting issue that needs to be explored further is the relationship between the forecasting accuracy and the computational time required to generate the forecasts. For this reason, we replicated so far (Makridakis, Assimakopoulos, & Spiliotis, 2018) the results of all of the submissions that we could using the code deposited at the M4 GitHub repository for 46 participating methods and benchmarks, while at the same time measuring the computation time required to forecast all 100,000 time series for each method (see Section 4.2.2, Makridakis et al., 2020). Of the 35 submissions that had been examined by the time this paper was written, 29 were classified as fully replicable, five were classified as non-replicable, and one remained unclassified due to its extremely long execution time. Note that, of the 20 top-performing methods in terms of point forecasts, 18 were accompanied by code, and 12 of these were found to be fully reproducible.

The results show that the computation time increases exponentially as sMAPE decreases, indicating that significantly more complex algorithms are required to improve the forecasting accuracy. Thus, for practical purposes, one may question whether the extent of the improvement compensates for the greater computational cost involved. In addition, it is not very clear how some of the new methods generate their forecasts, turning them into "black boxes" for forecasting users and raising the issue of whether the improved accuracy justifies the additional computing costs and the inability to understand how the forecasts are being produced. However, these are choices that should be determined by decision makers who can select the most appropriate method in relation to the cost that they are willing to assume (keeping in mind that such costs are decreasing exponentially over time), while also considering how intuitive it is for them to understand how the forecasts are made.

Although we were incorrect in our first prediction, it was a pleasant mistake, and it is our hope that forecasting accuracies will continue to improve in the future through both the discovery of new methods and the more effective implementation of existing ones.

(2) The 95% prediction intervals (PIs) will underestimate reality considerably, and this underestimation will increase as the forecasting horizon lengthens.

This was the first time that PIs had been submitted in an M Competition, meaning that no direct comparisons for evaluating potential improvements in forecasting performance over time were possible. However, many studies have discussed the inefficiency of PIs and provided some insights into the reasons why they fail to capture the real life uncertainty adequately (Bermúdez, Segura, & Vercher, 2010; Ord, Koehler, & Snyder, 1997). Some work on PIs has been done using the data from the previous M Competitions (Kolassa, 2011; Makridakis, Hibon, Lusk, & Belhadjali, 1987) using theoretically-derived PIs, but the M4 Competition enabled us to test the precision of PIs across various methods and determine their accuracy using two different measures, namely MSIS (mean scaled interval score) and ACD (absolute coverage difference). Our hypothesis was that the underestimation would be largest for yearly data, followed by quarterly and then monthly, while also increasing for longer forecasting horizons.

The results, summarized in Table 1, confirm that the forecasting horizon influences the level of underestimation strongly, with the mid- and long-term forecasts being 1.6 and 2.2 times less precise than the short-term ones, respectively, on average across all methods and frequencies. For instance, according to ACD, the coverage rates of methods vary from the initial target (95%) by an average of 3.7% for the short-term forecasts of the quarterly series, increasing to 6.9% for the long-term ones. Moreover, according to both the MSIS and ACD measures, the methods underestimate reality significantly for the case of the yearly data, followed by quarterly, then monthly. This is probably because generating long-term forecasts for monthly (13–18 months ahead) or quarterly (6–8 quarters ahead) series incorporates far less uncertainty than computing long-term forecasts for yearly series (5–6 years ahead). It is also notable that high-frequency data display rather high MSIS values but small ACD ones, indicating that the PIs submitted cover the 95% range successfully, but only by adopting rather wide bounds (the MSIS measures precision by considering both the width of the PIs and the cases where forecasts were outside the specified bounds). Undoubtedly the increased data availability and the generation of short-term forecasts affected the precision for high-frequency series. The numerical values of MSIS and ACD for all methods can be found in Appendix B of Makridakis et al. (2020).

Overall, the results of M4 confirm our prediction/hypothesis, demonstrating that much more work needs

Table 1
Average (median) performances of the M4 forecasting methods that performed better than the Naive 1 benchmark in terms of PIs.

Data frequency	MSIS			ACD		
	Short	Medium	Long	Short	Medium	Long
Yearly	22.208	39.939	56.494	0.162	0.165	0.180
Quarterly	6.388	10.270	14.776	0.037	0.054	0.069
Monthly	5.725	9.809	12.446	0.028	0.039	0.035
Weekly	12.880	20.982	25.354	0.014	0.016	0.018
Daily	17.374	29.742	47.062	0.023	0.018	0.023
Hourly	10.595	12.781	16.090	0.024	0.031	0.025

Notes: MSIS is the metric that was used for evaluating the precision of the PIs in M4, taking into consideration both the coverage rate and the width of the submitted PIs. The better the coverage rate and the narrower the width of the PIs, the lower the value of MSIS. ACD is a simplified metric that is used for measuring the coverage rates of the PIs. A value of 0.01 indicates that the average coverage rate of the PIs is 1% either higher or lower than the target set (95%), i.e., either 96% or 94%. The optimal value of ACD is zero.

Table 2
Average performance and relative rank of Simple, Holt and Damped exponential smoothing across the M4 complete dataset according to sMAPE, MASE and OWA.

Method	sMAPE	MASE	OWA	Rank
SES	13.087	1.885	0.975	3
Holt	13.775	1.772	0.971	2
Damped	12.661	1.683	0.907	1
Comb	12.555	1.663	0.898	–
Theta	12.309	1.696	0.897	–

Notes: The same information is also provided for the cases of Comb and Theta to facilitate comparisons. Rank refers to the accuracy of the SES, Holt and Damped methods.

to be done to improve the precision of PIs and capture the uncertainty better. At the same time, our prediction/hypothesis was not entirely correct, as the two top-performing methods of the competition, those of Smyl (2020) and Montero-Manso, Talagala, Hyndman, and Athanasopoulos (2020), achieved remarkable results by estimating the PIs precisely for the average of all of the M4 series. This was another pleasant surprise. We believe that the M4 findings provide an opportunity for those who underestimate reality to learn from the submissions of Smyl (2020) and Montero-Manso et al. (2020) how to improve the estimation of the PIs of their methods.

(3) The upward/downward extrapolation of the trend will be predicted more accurately when it is damped for longer horizons.

This statement was true for the M3 Competition, where Damped exponential smoothing, which damps the long-term trend of the series considerably in some cases, was found to perform better than the method of Holt, which extrapolates the trend linearly. The results of the M4 Competition, summarized in Table 2, confirm our hypothesis, showing that Damped outperforms Holt across all accuracy measures considered. According to OWA, the improvements reported from damping the linear trend of Holt are significant, reaching 6.6%. Furthermore, SES, which does not consider the trend, is less accurate than Holt, indicating that the most accurate prediction is between linear and horizontal extrapolations of the trend.

(4) The majority of pure ML methods will not be more accurate than pure statistical ones, although there may be one or two pleasant surprises where such methods are superior to statistical ones, though only by a small margin.

The nature of the economic and business series that are involved in the M4 Competition, which consist of limited numbers of observations and are characterized by considerable seasonality, some trend and a fair amount of randomness (Spiliotis, Kouloumos, Assimakopoulos, & Makridakis, 2020), was one of the reasons why we believed that pure ML methods would not be able to offer any significant advantage over statistical ones. ML methods typically require large datasets to be properly trained, which is why they are utilized commonly in applications where data availability is not an issue (e.g., in energy forecasting). The vulnerability of ML methods to overfitting further supported this belief (Makridakis, Spiliotis, & Assimakopoulos, 2018b).

The results of the competition confirmed our prediction, since none of the six pure ML methods available (two benchmarks and four submissions) managed to outperform the Comb benchmark, and only one was more accurate than the Naïve 2. On the other hand, the top two-performing methods of M4 incorporated some interesting ML features, indicating that significant improvements in forecasting accuracy are possible if we exploit the advantages of ML methods while avoiding their drawbacks. For instance, Smyl (2020) mixed exponential smoothing formulas with a recurrent neural network forecasting engine to construct an accurate hybrid model, while Montero-Manso et al. (2020) exploited the XGBoost algorithm to determine the optimal weights for combining various forecasting methods. Thus, although pure ML methods were proven to be incapable of extrapolating time series accurately, there is still the potential for ML algorithms to advance the theory and practice of forecasting within mixed forecasting frameworks. Interestingly, none of the pure ML methods were found to perform well for any of the high-frequency subsets (weekly, daily and hourly), suggesting that the accuracy of ML methods relative to statistical ones does not improve even when the number of observations increases considerably.

(5) Combinations of statistical and/or ML methods will produce more accurate results than the best individual methods.

This was another major finding of the previous M competitions, and we therefore hypothesized that it would be surprising to observe a method performing better on its own than when combined with others. It is well accepted today that forecast combinations provide considerable improvements in forecasting performance by averaging model errors, and thus reducing the uncertainty (Bates & Granger, 1969). Moreover, it is generally acknowledged that simple combinations of methods may perform similarly to or better than complex ones that involve schemes for utilizing the variances of the methods being combined (Claeskens, Magnus, Vasnev, & Wang, 2016). Thus, our prediction was that combinations of methods would outperform individual models, while in most cases simple combinations would perform equally as well as sophisticated ones.

Indeed, according to the results of the competition, only one of the ten top-performing methods was purely statistical, while the best performing one was hybrid, incorporating both ML and statistical elements, as well as several assembling levels. Moreover, it was found that simple combinations of methods can perform equally as well as sophisticated ones. For instance, the method proposed by Petropoulos and Svetunkov (2020) which utilized the median operator of four statistical models, was only 1.2% less accurate than the method of Montero-Manso et al. (2020), which involved an advanced ML algorithm for determining the optimal weights of the eight forecasting methods being used. The beneficial effect of combining simple methods is also evident from Table 2, where the Comb benchmark manages to outperform each of the individual models used for its estimation (SES, Holt and Damped).

The findings of M4 have reemphasized that combining methods is a powerful approach for enhancing the forecasting accuracy. However, what needs to be understood better is why some combinations achieve far more accurate results than others even though in most cases the models being combined are the same. It could be that the number of methods being combined and/or the way in which their weights are determined affects the forecasting performance, or maybe some other factors are even more important. For instance, it was shown that methods that determined their weights based on the characteristics of the series achieved results that were more accurate than those using completely generalized schemes. Hopefully the methods of Montero-Manso et al. (2020) and Pawlikowski and Chorowska (2020) can provide insights in that direction.

(6) Seasonality will continue to dominate the fluctuations of the time series, while randomness will remain the most critical factor influencing the forecasting accuracy.

This prediction was based on the findings of the previous M competitions, as well as more recent studies, indicating that each time series characteristic has either a positive or a negative effect on the forecasting accuracy, which may also vary significantly between methods (Petropoulos, Makridakis, Assimakopoulos, & Nikolopoulos, 2014). We investigated the influence of time series characteristics on the forecasting accuracy further by using a multiple linear regression (MLR) model to correlate the sMAPE value achieved for each series with a set of seven features (Spiliotis et al., 2020), the six proposed by Kang, Hyndman, and Smith-Miles (2017) plus the length of the training sample:

$$sMAPE_i = aF1_i + bF2_i + \cdots + fF7_i,$$

where $\mathrm{sMAPE_i}$ is the error generated for the *i*th time series of the sample by the methods examined and $\mathrm{F1_i} \ldots \mathrm{F7_i}$ are the features of randomness, trend, seasonality, frequency, linearity, stability and length, respectively (for more information, please see the study by Kang and coauthors). A different MLR model is estimated for each of the 61 methods that participated in the competition and their coefficients are visualized for each feature. Note that both the dependent (sMAPE) and independent (features) variables are 1% trimmed before estimating the models, to mitigate the effect of extreme values, and then scaled to be within the range of [0,1] so that the results are scale independent, directly comparable and easier to interpret. In this regard, a positive coefficient indicates that the feature being examined has a negative effect on the forecasting accuracy (increases the sMAPE), and vice versa.

The results of the analysis are visualized in Fig. 1, with boxplots being used to present the way in which the estimated coefficients are distributed by feature across the participating methods. As can be seen, the results confirm our prediction, indicating that randomness is the most critical factor for determining the forecasting accuracy, followed by the features of frequency and trend. On the other hand, length, stability and seasonality all have positive effects on the forecasting accuracy, meaning that linear and seasonal series are much easier to predict, especially if many observations are available for training. This last conclusion is very reasonable if we consider that linear and seasonal time series are typically less noisy. Note also that lower-frequency data are easier to forecast, probably because random variations are less likely to be observed at higher temporal levels (e.g. yearly or quarterly), resulting in series that are smoother and more predictable, at least when comparable forecasting horizons are considered (e.g. one year ahead for yearly data vs. 12 months ahead for monthly ones).

In addition, the variations observed across the boxplots of Fig. 1 confirm the general belief that there is no single method that will predict all possible types of series accurately. Some methods are more robust to randomness, others are better for extrapolating trended series, while others are better at capturing seasonality. Thus, different methods could be more appropriate for enhancing the forecasting accuracy depending on the particular characteristics of the series. Thus, this prediction was confirmed in all respects.

(7) The sample size will not be a significant factor in improving the forecasting accuracy of statistical methods.

Knowing that past M Competitions have found the effect of the sample size to be negligible for determining the forecasting accuracy, we hypothesized that the same would be true for the case of M4. We also anticipated that this conclusion would stand even for the ML methods across the economic/business data, with small improvements possibly being observed only for the high frequency series where the sample size is relatively longer.

Our hypothesis that the sample size would have no influence at all or only a small effect, and that mostly in the case of ML methods, was generally true, but not for all of the methods, as can be seen in Fig. 1. Moreover, after measuring the correlation between the sMAPE values reported by the M4 submissions for each series and their lengths, we found the forecasting accuracy to

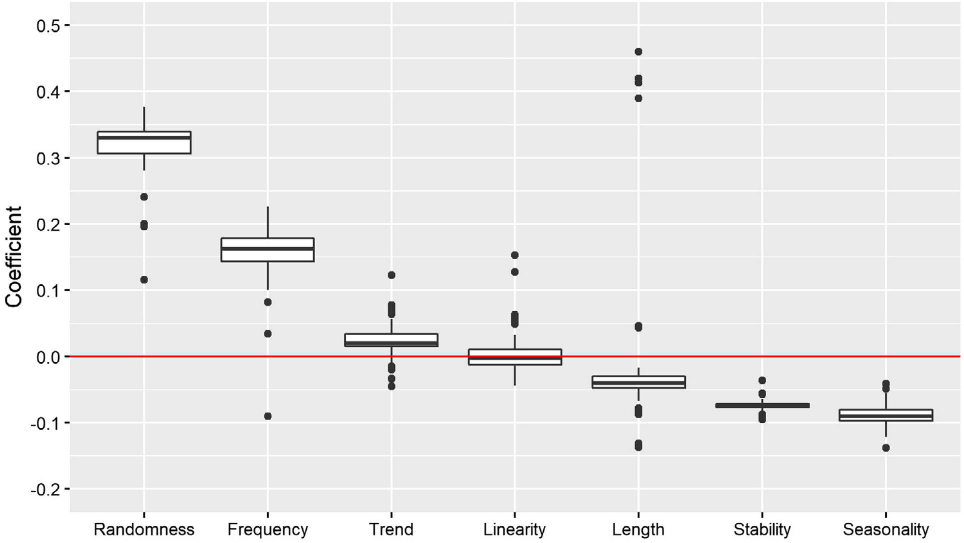

Fig. 1. The effects of time series features on the forecasting accuracy. Each boxplot represents the way in which the M4 methods are affected by a given characteristic of the series, as proposed by Kang et al. (2017). Positive coefficients indicate a negative effect, i.e., an increase in the sMAPE value based on the MLR equation, and vice versa.

be related weakly to the sample size, meaning that its positive effect is of minor importance in practice. Thus, we believe that our prediction/hypothesis about the influence of the sample size receives a weak agreement.

(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). Each method is good at capturing different time series characteristics, and the forecasting performance is affected strongly accordingly (Petropoulos et al., 2014). Moreover, time series from the same domain tend to display similar features to each other. Thus, assuming that the economic/business data of M3 are a representative sample of the real world, we believed that the same would be true for M4 (Spiliotis et al., 2020), meaning that the differences observed in the forecasting accuracies of the various methods tested between the two datasets will be small and not statistically significant.

However, excessive experiments are required in order to confirm such a hypothesis, involving the replication of all of the methods submitted to M4 on the M3 series. Given that this is a time demanding procedure, and that not all of the M4 methods are fully replicable (Makridakis et al., 2020), we proceeded by comparing the original results of the M3 methods with the respective ones that were available also in M4. The evaluation is performed only for the cases of the yearly, quarterly and monthly data, which represent the majority of the economic/business series of the competitions.

As can be seen in Table 3, the differences in forecasting accuracies between the two datasets among the seven methods examined are small. Noteworthy differences are observed only for the case of the yearly data, where most of the methods perform slightly better for the case of M4, as well as for the quarterly ones, where the accuracy of the methods decreased slightly. However, the overall average difference between the two sets is 0.84%, with those for five out of the seven methods being less than 1%. Thus, we conclude that our prediction/hypothesis about the similar performances of the M4 methods across the M3 dataset could be correct, although more testing may be needed to justify this prediction/hypothesis further.

(9) We would expect few, and 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.

At this point, we hypothesized that, if the M3 and M4 data are both representative of the real world (or close to it), their primary characteristics should display similar properties (Spiliotis et al., 2020). Having analyzed the features of the series in both datasets as proposed by Kang et al. (2017) and compared their distributions, it can be shown that the main features of the data were quite similar, displaying no statistically significant differences.

We attempted to better visualize the results of our analysis, inspect the properties of the two datasets in total and investigate their structures by generating the instance spaces of M3 and M4 for the two most significant principal components of the analysis (PC1 and PC2), which contain almost 67% of the variation in the data. The space, visualized in Fig. 2, is obtained by projecting the M4 and M3 data on the feature space of their combined dataset in order to avoid a possibly limited or bounded PC space. As can be seen, although M4 includes more highly trended time series of low spectral entropy (non-overlapping points on the right), the space obtained for the two sets is very similar, meaning that gaps that were identified previously for the M3 data remain empty for the M4 too. Thus, although M4 provided more series for spaces that previously were filled only sporadically, there are still some uninhabited areas that remain unfilled, indicating that some types of series are indeed

Table 3
Average performances (sMAPE) of the methods that participated in both the M3 and M4 competitions across their economic/business series (yearly, quarterly and monthly data).

Method	M4				M3				Avg. difference: M4 – M3
	Yearly	Quarterly	Monthly	Avg.	Yearly	Quarterly	Monthly	Avg.	
Naïve 2	16.34	11.01	14.43	14.03	17.88	9.95	16.91	15.27	−1.24%
SES	16.40	10.60	13.62	13.53	17.82	9.72	15.32	14.39	−0.86%
Holt	16.35	10.91	14.81	14.20	19.27	10.67	15.36	15.00	−0.80%
Damped	15.20	10.24	13.47	13.07	17.18	9.33	14.59	13.77	−0.70%
Comb	14.85	10.18	13.43	12.95	17.07	9.22	14.48	13.66	−0.71%
Theta	14.59	10.31	13.00	12.71	16.90	8.96	13.85	13.24	−0.53%
ARIMA	15.17	10.53	13.60	13.20	17.73	10.26	14.81	14.26	−1.06%
Average									−0.84%

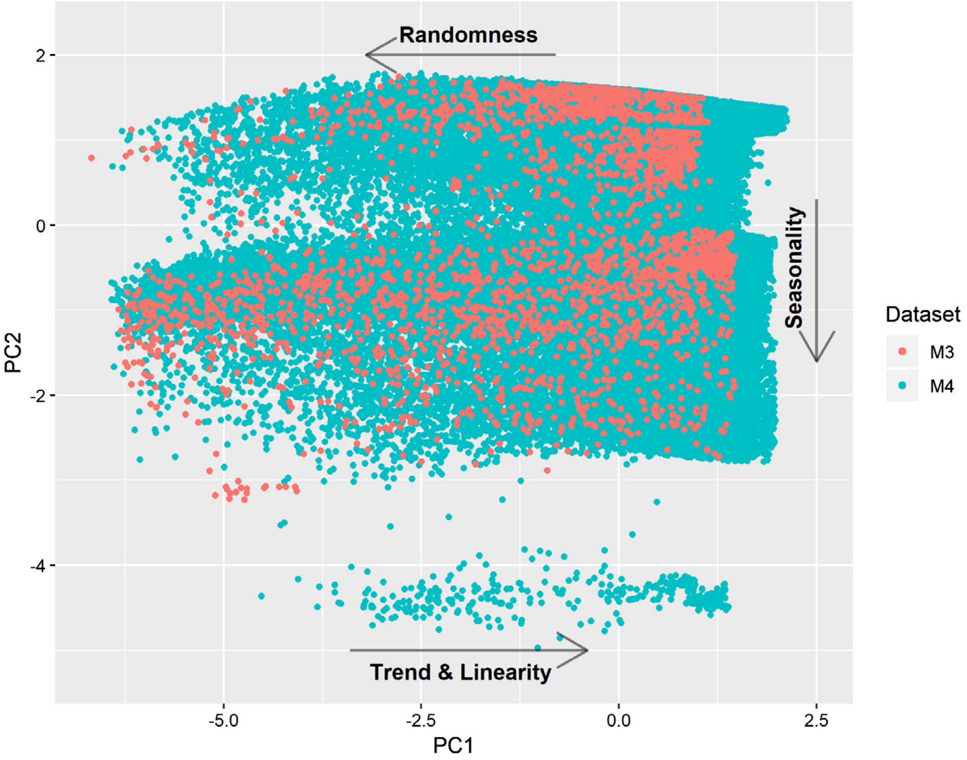

Fig. 2. The instance space of the M3 data (red) compared to that of the M4 (blue). The six time series features of Kang et al. (2017) are used to construct the space, while the combined dataset of the M3 and M4 series is used for estimating the principal components. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

observed only rarely in the real business world. Note that, unfortunately, the comparison that we have performed for all lower frequency data cannot be extended to high frequency data since such series were not available in the M3. Thus, we conclude that our prediction/hypothesis about the similarity of M3 and M4 is confirmed, at least for the frequencies of data that we could examine.

(10) There will be small, not statistically significant differences in the accuracies of the eight statistical benchmarks between the M3 and M4 datasets.

This hypothesis is derived from the previous two predictions, which indicate that the M3 and M4 data will display similar characteristics, as well as that the forecasting accuracies of the various methods will not differ significantly between the two datasets. It was also supported by the fact that most of the statistical benchmarks being utilized have remained almost identical since M3.

The results, which are available in Table 4, show that not only did the ranks of the benchmarks remain almost unchanged (Spearman's correlation coefficient was 0.95), but also the values of the errors (sMAPE, MASE and OWA) are very close. Given that time series characteristics are related closely to the performances of forecasting methods, this is a further indication that the properties of the M3 and M4 data are similar and provide a good representation of both the forecasting reality in the business world and the accuracy of the methods examined. Thus, we conclude that our prediction/hypothesis about the eight statistical benchmarks was confirmed.

Table 4
Average performances and relative ranks of the eight statistical benchmarks across the complete M3 and M4 datasets. The sMAPE, MASE and OWA values are provided for the sake of completeness.

Method	M4				M3				Difference: M4 – M3			
	sMAPE	MASE	OWA	Rank	sMAPE	MASE	OWA	Rank	sMAPE	MASE	OWA	Rank
Naive 1	14.208	2.043	1.058	7	15.702	1.788	1.069	8	−1.49	0.26	−0.01	-1
Seasonal Naive	14.657	2.057	1.078	8	15.186	1.764	1.045	7	−0.53	0.29	0.03	1
Naïve 2	13.564	1.912	1.000	6	14.702	1.669	1.000	6	−1.14	0.24	0.00	0
SES	13.088	1.885	0.975	5	13.437	1.612	0.940	4	−0.35	0.27	0.04	1
Holt	13.774	1.772	0.971	4	14.815	1.538	0.965	5	−1.04	0.23	0.01	-1
Damped	12.661	1.683	0.907	3	13.059	1.445	0.877	3	−0.40	0.24	0.03	0
Theta	12.309	1.697	0.897	1	12.819	1.421	0.862	1	−0.51	0.28	0.04	0
Comb	12.555	1.663	0.898	2	12.930	1.421	0.865	2	−0.38	0.24	0.03	0

3. Reproducibility and replicability in forecasting

Back in the late 1970s, one of the authors had the unpleasant experience of being accused that the reason why his forecasting findings were not in agreement with the prevailing theory of the time was his lack of knowledge in applying forecasting methods correctly (Makridakis & Hibon, 1979). This experience led him to establish a series of open forecasting competitions where all of the information and data required for reproducing the results would be available publicly to anyone interested in conducting research. To achieve this objective, the data from the first three M Competitions were made available at https://forecasters.org/resources/time-series-data/ and https://pkg.robjhyndman.com/Mcomp/, while those of the M4 are available at https://github.com/M4Competition.

In addition, "*participants* [of the M4] *were asked to submit a detailed description of how their forecasts were made and a source, or execution file, for reproducing the forecasts*", and a special prize was established for the best method of which the results could be fully reproduced, by asking the participants to upload at GitHub the source code of their method and some instructions regarding its use for individuals or firms who might wish to apply the method. This proves that the M Competitions were designed and implemented with the aim of avoiding the replication crisis that the social sciences face and ensuring that the forecasting research is accompanied by replicable benchmarks. The editorial published by the M4 organizers entitled "Objectivity, reproducibility and replicability in forecasting research" (Makridakis et al., 2018) discusses the difficulties that they encountered when attempting to reproduce the results of some ML forecasting studies, and suggests ways of promoting replicability in the field.

Predicting/hypothesizing the findings of the M4 Competition ahead of time was an attempt to go one step beyond the reproducibility issue and predict the actual findings of a large-scale forecasting experiment. It should be noted that the authors had no personal interest when formulating the ten hypotheses. However, their knowledge of forecasting competitions in general, and their experience with past M Competitions in particular, did influence the hypotheses. This is most obvious in the first hypothesis: "*The forecasting accuracy of simple methods … will not be too far from those of the most accurate methods*", given that the statement was true for the previous three M Competitions, which was a strong influencing factor for expecting its continuity in M4.

Overall, the main points that the authors underestimated were the submission of a few methods which involved interesting ideas such as cross-learning, and exploited the advances in computer power and ML algorithms to allow the utilization of sophisticated forecasting approaches. Would the ten hypotheses/predictions have been more accurate if the authors had known nothing about the previous M Competitions? The answer to this question is a definite negative, which implies that the findings of the M4 were believed to be much more a continuation of the previous three competitions than something completely new. Finally, now that the results of the M4 are known, would the authors have added any additional hypotheses/predictions? In hindsight, the following three hypotheses would have been added to the ten predictions already formulated:

- Hybrid methods, combining statistical and ML elements simultaneously in a common framework, are likely to improve the forecasting accuracy considerably.
- Given the advances in forecasting and computer science, a lot of new methods will perform better than traditional forecasting approaches.
- Methods that consider information from multiple series and enable cross-learning will perform considerably better than those that extrapolate the series individually.

References

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

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*(7604), 452–454.

Ball, P. (2018). High-profile journals put reproducibility test. Nature News, 27 August, 1476-4687.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *The Journal of the Operational Research Society*, *20*(4), 451–468.

Bermúdez, J. D., Segura, J. V., & Vercher, E. (2010). Bayesian forecasting with the Holt-Winters model. *The Journal of the Operational Research Society*, *61*(1), 164–171.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, *2*(9), 637–644.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762.

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

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.

Kolassa, S. (2011). Combining exponential smoothing forecasts using Akaike weights. *International Journal of Forecasting*, *27*(2), 238–251.

Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2018). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34*(4), 835–838.

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: an empirical investigation. *Journal of the Royal Statistical Society, Series A*, *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., Hibon, M., Lusk, E., & Belhadjali, M. (1987). Confidence intervals: an empirical investigation for the series in theM competition. *International Journal of Forecasting*, *3*(3/4), 489–508.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018a). The M4 competition: results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808.

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

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.

Montero-Manso, P., Talagala, T., Hyndman, R., & Athanasopoulos, G. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92.

Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, *92*(440), 1621–1629.

Pawlikowski, M., & Chorowska, A. (2020). Weighted ensemble of statistical models. *International Journal of Forecasting*, *36*(1), 93–97.

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

Petropoulos, F., & Svetunkov, I. (2020). A simple combination of univariate models. *International Journal of Forecasting*, *36*(1), 110–115.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75–85.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting competitions data representative of the reality? *International Journal of Forecasting*, *36*(1), 37–53.

Spyros Makridakis is a Professor at the University of Nicosia and the director of the Institute For the Future (IFF), as well as an Emeritus Professor at INSEAD. Spyros is the organizer of the M Competitions that started back in the early 1980s. He has authored, or co-authored, twenty-two books and more than 150 articles. He was the founding editor-in- chief of the Journal of Forecasting and the International Journal of Forecasting

Evangelos Spiliotis is a Research Fellow at the Forecasting & Strategy Unit. He graduated from School of Electrical and Computer Engineering at the National Technical University of Athens in 2013 and got his PhD in 2017. His research interests include time series forecasting, decision support systems and optimisation. He was a co-organizer of the M4 Competition.

Vassilios Assimakopoulos is a professor at the School of Electrical and Computer Engineering of the National Technical University of Athens. He has worked extensively on applications of decision support systems and he has conducted research on innovative management tools. He specialises in various fields of strategic management and forecasting. He is the author of over than 60 original publications and papers in many scientific journals and a co-organizer of the M4 Competition.