ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

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

The M4 Competition: 100,000 time series and 61 forecasting methods

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

[a] Institute for the Future, University of Nicosia, Cyprus
[b] Forecasting & Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece

ARTICLE INFO

Keywords:
Forecasting competitions
M competitions
Forecasting accuracy
Prediction intervals
Time series methods
Machine learning methods
Benchmarking methods
Practice of forecasting

ABSTRACT

The M4 Competition follows on from the three previous M competitions, the purpose of which was to learn from empirical evidence both how to improve the forecasting accuracy and how such learning could be used to advance the theory and practice of forecasting. The aim of M4 was to replicate and extend the three previous competitions by: (a) significantly increasing the number of series, (b) expanding the number of forecasting methods, and (c) including prediction intervals in the evaluation process as well as point forecasts. This paper covers all aspects of M4 in detail, including its organization and running, the presentation of its results, the top-performing methods overall and by categories, its major findings and their implications, and the computational requirements of the various methods. Finally, it summarizes its main conclusions and states the expectation that its series will become a testing ground for the evaluation of new methods and the improvement of the practice of forecasting, while also suggesting some ways forward for the field.

1. Introduction

This paper provides a detailed description of the most recent M Competition, the M4. It presents its results, discusses its findings and states its conclusions. In addition, it examines its implications for the theory and practice of forecasting and outlines some ways forward. Hyndman's (2019) excellent history of forecasting competitions elucidates their benefits and their contribution to the field of forecasting, so we do not need to add anything more, other than to agree with his conclusions and thank him for his considerable support in all stages of the M4, as well as in the design and realization of this special issue.

A major innovation of the M4 Competition was to predict/hypothesize on its findings more than two months before its completion. Our ten predictions/hypotheses were a clear statement of our expectations of such findings, rather than rationalizing its results in a post-hoc reasoning. We have now evaluated these ten predictions/hypotheses in a separate paper (Makridakis, Spiliotis & Assimakopoulos, 2019), and we are pleased to say that we were entirely correct in at least six out of the ten predictions/hypotheses made beforehand. Furthermore, we have explained where we went wrong, where we were partially correct, and where additional information is required to confirm our claims.

Another innovation of M4 was the appointment of a guest editor, Fotios Petropoulos, to supervise this special issue and assure objectivity and fairness. While certain decisions on the issue, such as its structure and length, and the number and content of invited papers and commentaries, were agreed upon between him, Rob J. Hyndman and Spyros Makridakis, the final decision about this and all other papers submitted remained solely with the guest editor. His editorial, Petropoulos and Makridakis (2019) describes the criteria used for inviting papers, the

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

https://doi.org/10.1016/j.ijforecast.2019.04.014

Please cite this article as: S. Makridakis, E. Spiliotis and V. Assimakopoulos, The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.04.014.

reviewing process and the deadlines set. In addition, it has been agreed that this paper, together with all others, would be put through a regular reviewing process before being published. Makridakis and Petropoulos (2019) conclude this special issue, while Makridakis, Hyndman and Petropoulos (2019) summarize the state of the art of forecasting in social sciences, discuss its achievements and limitations, and deliberate on its future prospects.

The present paper consists of three main parts, along with five appendices that are available online as supplementary material. The first part describes the background, organization and running of the M4 Competition. The second part presents the results of the competition, including its major findings and their implications, as well as the top performing methods, both overall and across the various subcategories, including data frequencies and application domains. In addition, it comments on the poor accuracy of the pure machine learning (ML) methods that were submitted to M4 and provides summarizing graphs and tables of various performance measures, including a graph showing the computational time required for applying a method versus its forecasting accuracy. The last part then summarizes the paper and highlights the conclusions of the competition, stating our suggestion that the 100,000 series of the M4 should become a testing ground for guiding further theoretical and practical developments by being a huge sample of time series on which the performances of new forecasting methods can be assessed. Finally, the five appendices provide comprehensive tables and graphs of the overall findings, as well as of the various subcategories, including information regarding the benchmarks used and the forecasting performances achieved by the participating methods for each forecasting horizon, frequency and domain, in terms of both point forecasts (PFs) and prediction intervals (PIs).

2. The background

Forecasting competitions have influenced the field of forecasting greatly over the years, providing a solid basis for assessing different extrapolation approaches and learning empirically how to advance forecasting theory and practice (Hyndman, 2019). Each forecasting competition has introduced some new features or data, while trying to either address possible limitations of previous ones, or focus on specific fields of application such as energy and tourism. M4 likewise involved some new features, which can be summarized as follows: (i) the introduction of high-frequency data (weekly, daily and hourly) along with low-frequency data (yearly, quarterly and monthly); (ii) the consideration of PIs as well as PFs; (iii) an emphasis on the reproducibility of the results; and, finally, (iv) the incorporation of a vast number of diverse series and benchmarks. Apart from introducing these features, though, M4 was motivated by the fact that new competitions also tend to bring to light innovative methods which can be tested against existing ones on completely unknown sets of series. This is of major importance given that M3, the previous M Competition, has been being used as a standard benchmark for comparison for almost two decades, meaning that newly proposed methods could eventually over-fit its published test sample. From our point of view, M4 has inspired the development of cutting-edge methods, as well as providing a new, larger, and therefore more difficult to over-fit dataset, thus assisting researchers and practitioners to explore and re-evaluate best practices in forecasting in greater detail.

The M4 Competition was initially announced at the beginning of November, 2017, first on the University of Nicosia's website (www.unic.ac.cy) and then on both the IIF blog (www.forecasters.org) and that of Rob J. Hyndman (www.robjhyndman.com). In addition, invitation emails were sent to all those who had participated in the previous M Competitions (Makridakis, Andersen, Carbone, Fildes, Hibon, et al., 1982; Makridakis, Chatfield, Hibon, Lawrence, Mills, et al., 1993; Makridakis & Hibon, 2000), the tourism forecasting competition (Athanasopoulos, Hyndman, Song, & Wu, 2011), the NN3 competition (Crone, Hibon, & Nikolopoulos, 2011), and recent International Symposium on Forecasting (ISF) events, as well as to those who had published relevant articles in the *International Journal of Forecasting* and other well-known forecasting, neural network and machine learning journals (*Foresight: The International Journal of Applied Forecasting*, *Journal of Forecasting*, *Technological Forecasting and Social Change*, *Expert Systems with Applications*, *Neural Networks* and *Neurocomputing*). The data were made available on 31st December, originally on the M4 website (www.m4.unic.ac.cy) and later via the *M4comp2018* R package (Montero-Manso, Netto, & Talagala, 2018) and the M4 GitHub repository (www.github.com/M4Competition). The competition ended at midnight, May 31st, 2018. A short paper with the initial results of the M4 was published in the *International Journal of Forecasting* on June 20th, 2018 (Makridakis, Spiliotis, & Assimakopoulos, 2018b), which was made available as an open access article by the generous contribution of Elsevier during the ISF conference in Boulder, Colorado, USA.

Like the previous three M Competitions, M4 was an open competition with the aim of ensuring fairness and objectivity. Moreover, in order to promote replicability in forecasting research and facilitate future work in the field (Makridakis, Assimakopoulos, & Spiliotis, 2018), the participants were encouraged to submit the code of their method, as well as a detailed description of it, to the M4 GitHub repository. Participants who claimed proprietary rights (e.g. software vendors) did not have to make their code available publicly, but provided a version of their program to the organizers of the competition to allow them to evaluate its reproducibility and measure its computational complexity. Table 7 lists the submitted methods and shows which have been replicated successfully, as well as the extent of the replication.

The rules of the competition, prizes and additional details were all made available on the M4 website. In order for participants to be eligible for the prizes, they had to provide PFs and, optionally, PIs for all 100,000 series of the competition shown in Table 1. The dataset was subdivided into six data frequencies and six application domains, and the PFs and PIs were evaluated for each subcategory, in addition to computing overall averages. Thus, the various tables in the main text, as well

Please cite this article as: S. Makridakis, E. Spiliotis and V. Assimakopoulos, The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.04.014.