International Journal of Forecasting 38 (2022) 1279–1282

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Editorial

Introduction to the M5 forecasting competition Special Issue

1. The forecasting field and the value of forecasting competitions

COVID-19, which was unknown before 2019, has, so far, inflicted close to 475 million people around the world, killing more than 6 million people, and causing havoc in practically all aspects of our lives, including work, financial matters, and social and professional relationship. Yet, decisions to minimize its impact, such as lockdowns, vaccines, working from home, or wearing masks, must be made based on future, uncertain events and unknown, new developments. Unfortunately, for novel events like COVID-19, past experience cannot help us improve much the performance of our forecasts, allowing us to make better decisions that are based on objective information and influenced as little as possible from judgmental biases and irrational thinking. The last major pandemic, the Spanish flu, occurred more than 100 years ago, under completely different conditions, and cannot provide us with much help to more successfully cope with COVID-19. The value of the forecasting field comes from studying repetitive events and being able to identify established patterns and/or relationships to base its predictions and estimates of uncertainty.

Forecasting research focuses on improving a particular forecasting model, method, or approach, experimenting with variations of existing techniques, or proposing brand-new ones. No matter what the methodological innovation might be, the proposed solution is consequently tested empirically on some data using one or many performance measurements. It can be argued that such data and measures are arbitrarily selected by the researchers of the particular study. Follow-up studies may use different data and measures, making it hard to track the added value of each solution under a common framework. Other decisions regarding the empirical design, such as the length of the forecasting horizon or the use of single versus multiple evaluation origins, will further complicate comparisons across different propositions in the literature. Journals, including the *International Journal of Forecasting*, are increasingly ask for the provision of code and the availability of data that would allow other researchers to replicate (if not reproduce) the results of a specific study, expanding its applicability beyond the specific set of data and methodological design (Hyndman, 2010; Makridakis et al., 2018).

Forecasting competitions offer robust and predefined empirical frameworks that allow for comparisons across many different models, methods, and approaches using a common framework (empirical design, data, measures, etc.). Thus, they provide an effective, relative ranking for evaluating different solutions while also keeping a historical record over time. However, by focusing on a single set of data, the results of a single forecasting competition may not be generalizable to other sets. To tackle this limitation, organizers of past forecasting competitions focused on either diverse sets of data or a challenge that focused on a particular context. Examples of the former include the M3 and M4 forecasting competitions (Makridakis & Hibon, 2000; Makridakis et al., 2020) that used large sets of diverse time series, covering a considerable number of data domains and frequencies (Spiliotis et al., 2020). Examples of the latter include the tourism forecasting competition (Athanasopoulos et al., 2011) and the global energy forecasting competitions (Hong et al., 2014, 2016, 2019).

This special issue is dedicated to the latest forecasting competition organized by Professor Spyros Makridakis and his team, the M5 forecasting competition. The fifth installment in the Makridakis competitions diverges from the norm of the previous four ones in that it is no longer a generic time series challenge but focuses on the context of retail/sales forecasting. Additionally, the competition features a number of innovations that were integrated by the organizing team to directly address the feedback received from fellow academics and forecasting practitioners on the design of previous forecasting competitions. These innovations are discussed in the next section.

The M5 forecasting competition was the first M competition to run through the Kaggle platform, rendering it widely available to the data science and machine learning communities. As a result, the M5 competition was the largest forecasting competition to date, with almost 6,400 participating teams (5,507 in its "Accuracy" challenge and

https://doi.org/10.1016/j.ijforecast.2022.04.005

another 892 in the "Uncertainty" challenge). The organizers of the competition offered prizes totaling $100,000 USD.

Undeniably, the amount of knowledge gained from organizing such a competition is high. This special issue aims to report on the findings, summarize key insights, and discuss their contribution to the forecasting theory and practice.

2. The value of feedback

Since the first M forecasting competition (Makridakis et al., 1982), Makridakis and his team have received meaningful, constructive feedback from the forecasting community that enabled them to design even more robust forecasting competitions that are closer to reality. In particular to the M5 forecasting competition, there are five key innovations emerging directly or indirectly from the valuable feedback of researchers and practitioners in previous iterations of the M forecasting competitions. These innovations are listed below:

1. The use of high-frequency (daily) data for the entire set is considered. Presently, many companies, including but not limited to supply chain contexts, make use of such granular data to support their decisions.
2. The use of data that are intermittent in nature. So far, the M competitions have focused on time series for which the values are positive. In the M5 competition, there is a high degree of intermittence for the majority of the time series which translates to zero values (zero sales) for some observations. This creates an additional challenge in terms of forecasting: one needs to forecast not only "how much", but also "when" to achieve the best performance.
3. Previous M competitions did not offer any links across the series, apart from labeling each series as part of one of five large data domains (micro, macro, industry, demographic, or finance). The M5 competition offers a clear hierarchical structure, where the most granular level corresponds to the sales of one product at a particular store. Then, such sales can be aggregated in terms of stores and states or products, departments, and categories.
4. The M5 competition is the first in the Makridakis competitions to explicitly offer information regarding exogenous variables, such as product prices and special events.
5. The use of well-defined and fit-to-purpose measures to evaluate the performance of the point forecasts, as well as the probabilistic ones that approximate the uncertainty distributions in terms of nine quantiles.

Despite the many innovations, we believe that any forecasting competition can be regarded as a "work in progress" and that any results are limited by the data used and, in this case, the particular retail sales application. Regardless, we also believe that such competitions have a lot to offer to our understanding of how forecasts perform under different settings and how uncertainty behaves.

3. In this issue…

Given the focus of the M5 competition on forecasting retail sales, the special issue starts with a review paper on "Retail forecasting: Research and practice" by Fildes et al. Even if their review article was not originally prepared for this special issue, the authors accepted our invitation for inclusion in the special issue and also prepared a postscript that offers their most up-to-date insights and considerations in the COVID-19 era.

The next four papers are authored by Makridakis and his colleagues and offer details on the organization and the results of the competition. In more detail, the first paper provides the background and motivation toward organizing a competition that focused on retail data, design, and implementation details, and how relevant decisions were made. In the second paper, the authors offer their hypotheses and predictions of the findings of the M5 competition. The third and fourth papers focus on the results, findings, and conclusions of the two challenges of the M5 competition, "Accuracy" and "Uncertainty", but also discuss the implications of these results for practice and research. In their conclusions for the "Accuracy" challenge, Makridakis, Spiliotis, and Assimakopoulos argue that what is important is "*the integration of statistics and data science into a unique field covering all academic aspects of forecasting and uncertainty, as well as determining how to increase the usage of forecasting in organizations by persuading executives of the benefits of systematic forecasting for improving their bottom line*".

Next, the special issue includes a series of invited methodological papers and short notes written by participants that made the best performing submissions in the M5 competition or developed solutions that diversified methodologically compared to other top-performing submissions.

In and Jung offer the details of their winning submission in the "Accuracy" challenge, which was based on multiple direct and recursive LightGBM models. Bandara et al. provide a variation of that submission that combines LightGBM with pooled regression, an approach that finished in the 17th position and within the top 1% of the solutions. Anderer (2nd position; "Accuracy" challenge) with Li offer the methodological details of an approach based on LightGBM for the granular, intermittent series, a deep learning forecasting model, N-BEATS, for the higher-level, continuous series, and a simple bottom-up variant for hierarchical alignment. Jeon and Seong (3rd position; "Accuracy" challenge) describe their deep learning approach, which was based on DeepAR and trained, instead of using actual historical data, by considering data sampled from a trained distribution.

Lainder and Wolfinger's top-performing submission in the "Uncertainty" challenge was based on LightGBM models, and the authors describe how they used cross-validation and data augmentation to tune the parameters of such models, select input variables, and better generalize their solution. Mamonov et al. (2nd position; "Uncertainty" challenge) combined machine learning methods with nonparametric estimation of probability distributions and techniques for time series analysis, producing predictions for the median and other quantiles

separately. In a similar fashion, Chiew and Choong (13th position; "Uncertainty" challenge) also estimated these quantiles separately, using a hybrid of LightGBM and DeepAR. Nasios and Vogklis' probabilistic approach (3rd position; "Uncertainty" challenge) was based on machine learning models that integrated gradient boosted trees and neural networks. Finally, de Rezende et al. (6th position; "Uncertainty" challenge) took a different, white-box approach and estimated the quantiles by using a multi-stage state-space model and employing Monte Carlo simulations.

The methodological papers are followed by a series of discussion papers.

Seaman and Bowman offer their perspective of the results and how these are applicable at Walmart. Their discussion paper is organized among four themes, namely data (and the associated challenges in forecasting tasks), M5 competition's accuracy and uncertainty metrics (and their relevance in practice), and the applicability of the leading solutions. They conclude: "*The role of the statistician/data scientist seems to us to be transitioning from that of a craftworker to those of a machine operator and an engineer. [...] It will be interesting to see what the future holds for the skills required to be an effective data scientist practitioner.*"

Januschowski et al. (Amazon Web Services) provide reasons for the good performance of tree-based solutions in the M5 competition. Such reasons include the ease in implementing different loss functions in LightGBM/gradient-boosting solutions, their strength in handling data that contain outliers or zero values, and their computational efficiency and interpretability (compared to neural networks). They also discuss how such solutions could be further improved.

Wellens et al. focus on the computational cost of the top-performing solutions of the M5 competition and argue that such costs could be a limitation from widely adopting these solutions in practice. They propose the use of transfer learning to reduce (by one-fourth) the cost of tree-based solutions without deteriorating the forecasting performance.

Ma and Fildes argue against the use of single-origin evaluations for the final rankings in forecasting competitions. They empirically substantiate their arguments showing that a global bottom-up approach, similar to the top-performing solutions of the M5 competition, does not produce robust forecasts across multiple evaluation windows.

Theodorou et al. explore the representativeness of the M5 competition data. They use time series features and feature spaces to compare the retail data used in the M5 competition against the data of two other retail companies, Corporación Favorita (a data set publicly available from another Kaggle competition) and a major Greek supermarket chain. Their analysis shows that there exist only small differences between the three data sets, implying that M5 forecasting competition results could be directly applicable to other retail companies.

As the M5 competition was hosted by Kaggle, one of its features was an increased social interaction through online notebooks and discussion forums. Li et al. explore the social influence of the virtual community and make several interesting observations that include the popularity of topics, such as LightGBM and Tweedie, that were part of the approaches of many of the top-performing submissions. The authors also perform social network analysis to study cooperation versus competitive dynamics and get insights regarding the flow of sharing information and knowledge.

Alroomi et al. take a more philosophical stance and question why one expects their method to work well, why increasingly complex methods outperform simpler ones, why each forecasting competition has a different winner, and how forecasting is performed in practice. To shed some light on these "whys" and "hows", they performed surveys and interviews. They conclude: "*The 'why' remains unanswered and more importantly very often even the forecast providers do not know why their methods are winning or losing*" and suggest that future competitions should ask participants to submit explanations as to why they expect their methods to perform well.

Focusing on the "Uncertainty" challenge of the competition, Ord enumerates the contributions of the M5 competition. Among others, Ord discusses the generalization of the results provided different margins of data availability, the importance of exogenous variables and how their effects should be integrated into forecasting software, and how prediction intervals constructed on one level can be propagated to other levels in a compatible manner.

Chen et al. offer an analysis of the results of the "Uncertainty" challenge of the competition with regards to the performance of the submitted probabilistic forecasts for different hierarchical aggregation levels and different probability levels. Their findings show the forecasts to be well calibrated, a deterioration of performance on more granular levels, and an inverted U-shaped function with regards to the performance at different quantile levels.

Ziel's discussion paper also focuses on the uncertainty results of the M5 competition. He focuses on the unique challenge presented by the data in terms of distribution and intermittence. He proposes a distributional forecasting approach that uses low-dimensional distribution and illustrates its application on the M5 data.

Bojer presents a framework that aims to aid researchers and practitioners toward mapping and comparing different machine learning methods. The framework maps machine learning solutions based on five distinct areas, namely pre-processing, data construction, model training and validation, post-processing, and ensembling. Bojer applies the framework to two top-performing approaches of the "Uncertainty" challenge to demonstrate how it can be used to compare different methods and understand their performance.

Kolassa's discussion paper argues that the performance of simple methods, such as exponential smoothing, is understated. While he notes that Walmart is not a typical retailer (given the "every-day low price" strategy), he notes that "*only 7.5% of teams manage to beat an extremely*

simple benchmark! [...] *One could interpret this provocatively as a very rough prior probability: a priori, you only have a 7.5% chance of outperforming bottom-up Exponential Smoothing.*" Kolassa also discusses the need for appropriate benchmarks for the "Uncertainty" challenge of the competition that can handle intermittent demand patterns and the need to balance added complexity against any benefits in utility performance.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*(3), 822–844.

Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, *30*(2), 357–363.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913.

Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, *35*(4), 1389–1399.

Hyndman, R. J. (2010). Encouraging replication and reproducible research. *International Journal of Forecasting*, *26*(1), 2–3.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153.

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. (2000). The M3-competition: results, conclusions and implications.. *International Journal of Forecasting*, *16*(4), 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.

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
Institute for the Future (IFF), University of Nicosia, Cyprus
University of Nicosia, Cyprus

Fotios Petropoulos *
University of Bath, United Kingdom
E-mail address: f.petropoulos@bath.ac.uk.

Evangelos Spiliotis
Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece
National Technical University of Athens: Ethniko Metsobio Polytechneio, Greece

* Corresponding editor.