INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Future of Forecasting Competitions: Design Attributes and Principles

Spyros Makridakis, Chris Fry, Fotios Petropoulos, Evangelos Spiliotis

To cite this article:
Spyros Makridakis, Chris Fry, Fotios Petropoulos, Evangelos Spiliotis (2022) The Future of Forecasting Competitions: Design Attributes and Principles. INFORMS Journal on Data Science 1(1):96-113. https://doi.org/10.1287/ijds.2021.0003

http://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 1, No. 1, April–June 2022, pp. 96–113
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

The Future of Forecasting Competitions: Design Attributes and Principles

Spyros Makridakis,[a] Chris Fry,[b] Fotios Petropoulos,[c] Evangelos Spiliotis[d]

[a] Institute for the Future, University of Nicosia, Engomi 2417, Nicosia, Cyprus; [b] Google, Inc., Mountain View, California 94043; [c] School of Management, University of Bath, Bath BA2 7AY, United Kingdom; [d] Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece

Contact: makridakis.s@unic.ac.cy, https://orcid.org/0000-0003-2519-8095 (SM); chrisfry@google.com (CF); f.petropoulos@bath.ac.uk, https://orcid.org/0000-0003-3039-4955 (FP); spiliotis@fsu.gr, https://orcid.org/0000-0002-1854-1206 (ES)

Received: January 16, 2021
Revised: May 19, 2021
Accepted: June 27, 2021
Published Online in Articles in Advance: November 12, 2021

https://doi.org/10.1287/ijds.2021.0003

Abstract. Forecasting competitions are the equivalent of laboratory experimentation widely used in physical and life sciences. They provide useful, objective information to improve the theory and practice of forecasting, advancing the field, expanding its usage, and enhancing its value to decision and policymakers. We describe 10 design attributes to be considered when organizing forecasting competitions, taking into account trade-offs between optimal choices and practical concerns, such as costs, as well as the time and effort required to participate in them. Consequently, we map all major past competitions in respect to their design attributes, identifying similarities and differences between them, as well as design gaps, and making suggestions about the principles to be included in future competitions, putting a particular emphasis on learning as much as possible from their implementation in order to help improve forecasting accuracy and uncertainty. We discuss that the task of forecasting often presents a multitude of challenges that can be difficult to capture in a single forecasting contest. To assess the caliber of a forecaster, we, therefore, propose that organizers of future competitions consider a multicontest approach. We suggest the idea of a forecasting-"athlon" in which different challenges of varying characteristics take place.

History: Nick Street served as the senior editor for this article.

Keywords: data science • business analytics • competitions • organization • design • forecasting

1. Introduction

Hyndman (2020) reviews the history of time series forecasting competitions and discusses what we have learned from them as well as how they have influenced the theory and practice of forecasting. In this article, we provide a systematic approach to the design of forecasting competitions, focusing on forecasting competitions that allow participants to submit their forecasts, thus excluding early studies in which all the methods and approaches were provided by the researchers (see, for example, Newbold and Granger 1974, Makridakis and Hibon 1979).

In this sense, the first time series forecasting competition, M (Makridakis et al. 1982), was held in 1981 with seven participants, all known to the organizer and invited personally by telephone or regular mail, and the M5 (Makridakis et al. 2020c, e), hosted by Kaggle in 2020 and run over the internet, attracted in its two tracks (accuracy and uncertainty) 8,229 participants from 101 countries around the world, offering $100,000 in prizes to the winners. There is little doubt, therefore, that forecasting competitions have changed a great deal and have become big events, attracting large numbers of participants from diverse backgrounds and with varying reasons to join. As time passes, however, there is a need to question the way forecasting competitions are structured and consider improvements in their design and the objectives they strive to achieve in order to attain maximum benefits from their implementation. Also, as presented in the encyclopedic overview of Petropoulos et al. (2020), the applications of forecasting expand to many social science areas, such as economics, finance, healthcare, climate, sports, and politics, among others. As such, there is also the need to consider new application areas for future forecasting competitions beyond operations, supply chain, and energy, which have been the main case till now.

Forecasting competitions are the equivalent of the laboratory experimentation widely used in physical and life sciences. They are used to evaluate the forecasting performance of various approaches and determine their accuracy and uncertainty. Their broad purpose is to provide objective, empirical evidence to aid policy and decision makers about the most appropriate forecasting approach to use to realize their specific needs.

There have been many commentaries over time on the design and limitations of such competitions (see, for instance, discussions and commentaries of issues 17:4 and 36:1 of the *International Journal of Forecasting* for the case of the M3 and M4 forecasting competitions). However, given the large number of forecasting competitions conducted over the last decade, organized from academic teams and also companies and organizations, a structured analysis of their design attributes seems to be necessary. Moreover, we deliberate about the future of forecasting competitions and what should be done to improve their value and expand their usefulness across application domains. In this regard, we provide a systematic review of past forecasting competitions, determining their main attributes and key innovations while also proposing how future competitions could be designed so that we better learn from data and "data analysis" (Donoho 2017) of the competitions' results, the circumstances under which a forecasting method is expected to work best, instead of just focusing on the winners.

The paper consists of six sections and a conclusion. After this short introduction, Section 2 summarizes the conclusions of Hyndman's influential paper about past time series forecasting competitions, an interest of the present discussion, and enumerates his suggestions about the characteristics of future ones. Section 3 describes various types of forecasting competitions, considering their scope, the type of data used in terms of diversity and representativeness, structure, granularity, availability, the length of forecasting horizon, and several other attributes, including performance measures and the need for benchmarks. Consequently, Section 4 identifies the commonalities as well as design gaps of past forecasting competitions by mapping the designs of indicative, major ones to the attributes described previously and mentioning the advantages and drawbacks of each. Section 5 focuses on outlining the proposed features of some "ideal" forecasting competitions that would avoid the problems of past ones while filling existing gaps in order to improve their value and gain maximum benefits from their implementation. Section 6 presents some thoughts about institutionalizing the practice of forecasting competitions and systematizing the way they are conducted, moving from running single competitions to structuring them across multiple forecasting challenges in the way that pentathlons are run with single winners in each challenge and an overall one across all. Finally, the conclusion summarizes the paper and proposes expanding the competitions beyond business forecasting to cover other social science areas in need of objective information to improve policy and decision making.

2. A Brief History of Time Series Forecasting Competitions

In his paper, Hyndman (2020) concludes that time series forecasting competitions play an important role in advancing our knowledge of what forecasting methods work and how their performance is affected by various influencing factors. He believes that in order to improve objectivity and replicability, the data and the submitted forecasts of competitions must be made publicly available in order to promote research and facilitate the diffusion and the usage of their findings in practice. At the same time, their objectives must be clear, and the extent to which their findings can be generalized must be stated. According to him, future competitions should carefully define the population of data from which the sample is drawn and the possible limitations of generalizing the findings to other situations. The usage of instance spaces (Kang et al. 2017) could provide a way to specify the characteristics of the data included and allow comparisons to other competitions or data sets with well-known properties (Fry and Brundage 2020, Spiliotis et al. 2020a). Moreover, a nice side effect of time series competitions is that they introduce popular benchmarks, allowing the evaluation of performance improvements and comparisons among competitions for judging the accuracy and uncertainty of the submitted methods, including the assessment and replication of their findings over time. Furthermore, as new competitions emerge and the benchmarks are regularly updated, the effect of developing methods that overfit published data are mitigated, and new, robust forecasting methods can be effectively identified.

On the negative side, Hyndman (2020) expresses concerns about the performance measures used, stating that these should be based on well-recognized attributes of the forecast distribution. This is particularly true for the case of the prediction intervals, stating that the widely used Winkler (1972) scores are not scale-free and that their scaled version used to assess the interval performance in the M4 competition (Makridakis et al. 2020b) seems rather ad hoc with unknown properties. Consequently, he cites the work of Askanazi et al. (2018), who assert that comparisons of interval predictions are problematic in several ways and should be abandoned for density forecasts. Probabilistic forecasts, such as densities, could be evaluated instead using proper scoring rules and scale-free measures, such as log density scores, as done in M5 and in some energy competitions (Hong et al. 2016, 2019). There is, therefore, a need to reconsider how such probabilistic forecasts are made and evaluated in future competitions to avoid the criticisms that they are inadequate. However, no matter how such

evaluations are done, Hyndman (2020) suggests that it would be desirable that forecast distributions are part of all future forecasting competitions. Another issue he raises is whether explanatory/exogenous variables improve forecasting performance over that of time series methods. For instance, in the tourism forecasting competition (Athanasopoulos et al. 2011), explanatory/exogenous variables were helpful only for one-step-ahead forecasts, and in some energy competitions (Hong et al. 2014, 2016, 2019), using temperature forecasts was beneficial for short-term forecasting, in which weather forecasts were relatively accurate, with the results being mixed for longer forecasting horizons. On the other hand, explanatory/exogenous variables whose values can be specified, such as the existence of promotions, day of the week, holidays, and days of special events such as the Super Bowl, are generally considered to be helpful for improving forecasting performance and should be, therefore, included in the forecasting process (Makridakis et al. 2020a).

A major suggestion of Hyndman (2020), previously discussed by the commentators of the M3 competition (Fildes 2001, Hyndman 2001), is that future time series competitions should focus more on the conditions under which different methods work well rather than simply identifying the methods that perform better than others. Doing so presents a significant change that is particularly relevant for breaking the black box of machine and deep learning forecasting methods that is necessary to better understand how their predictions are made and how they can be improved by concentrating on the factors that influence accuracy and uncertainty the most. In addition, he believes that future time series competitions should involve large-scale, multivariate forecasting challenges while focusing on irregularly spaced and high-frequency series, such as hourly, daily, and weekly data that is nowadays widely recorded by sensors, systems, and the internet of things. Finally, Hyndman (2020) states that he does not know of any large-scale time series forecasting competition that has been conducted using finance data (e.g., stock and commodity prices and/or returns) and that such a competition would seem to be of great potential interest to the financial industries and investors in general.

3. Design Attributes of Forecasting Competitions

In this section, we identify and discuss 10 key attributes that should be considered when designing forecasting competitions even if some of them might not be applicable to all of them. Table 1 provides a summary description of these attributes, which are then discussed in detail in the next sections.

3.1. Scope

The first decision in designing a forecasting competition relates to its scope, which can be defined based on (i) the focus of the competition, (ii) the type of the submissions it will attract, and (iii) the format of the required submissions.

Table 1. Summary Description of the Design Attributes of Forecasting Competitions

	Design attribute	Description
1	Scope	Focus of competition (domain or application); type of submission (numeric or judgment); format of submission (point forecasts, uncertainty estimates, or decisions)
2	Diversity and representativeness	Degree to which the findings and insights obtained can be generalized and applied to other settings or data sets
3	Data structure	Degree the data are connected or related; explanatory or exogenous variables used for supporting the forecasting process
4	Data granularity	The most disaggregated level, cross-sectional or temporal, at which data are available
5	Data availability	Amount of information provided by the organizers for producing the requested forecasts (e.g., number of series contained in the data set and historical observations available per series)
6	Forecasting horizon	Length of time into the future for which forecasts are requested
7	Evaluation setup	Number of evaluation rounds (single versus rolling origin); live versus concealing data competitions
8	Performance measurement	Measures used for evaluating performance in terms of forecasting accuracy or uncertainty, utility, or cost
9	Benchmarks	Standards of comparisons used for assessing performance improvements
10	Learning	What can be learned for advancing the theory and practice of forecasting; replicability or reproducibility of results; making and evaluating hypotheses about the findings of the competitions; challenging and confirming the results of past competitions

Regarding the focus, there is a spectrum of possibilities ranging from generic to specific competitions. Generic competitions feature data from multiple domains that represent various industries and applications as well as from various frequencies. Examples include the M, M3, and M4 forecasting competitions that include data from different domains (micro, macro, industry, demographic, finance, and others) and various frequencies (yearly, quarterly, monthly, weekly, daily, hourly, and others). Although the results of such competitions identify the methods performing best on each data domain or frequency, they typically determine the winners based on their average performance across the complete data set. Thus, although their main findings may not be necessarily applicable to all the domains or frequencies examined, they help us effectively identify best forecasting practices that hold for diverse types of data.

Specific competitions feature data of a particular domain or frequency, a particular industry or company or organization. Examples of such competitions include the global energy ones and the majority of those hosted on Kaggle (Bojer and Meldgaard 2020), including M5. Although these competitions may be more valuable for specific industries or organizations, replicating real-world situations, their findings are restricted to the specific data set and cannot be generalized to other situations. Finally, semispecific competitions feature data that, although referring to a particular domain, include instances from various applications of that domain, which may, therefore, require the utilization of significantly different forecasting methods. For example, a semispecific energy competition may require forecasts for renewable energy production, demand, and prices with the winners being determined based on their average performance across these tasks. In this case, factors that influence forecasting performance in the examined domain can be effectively identified although the key findings of the competitions can be applicable to several forecasting tasks of that domain.

Apart from the focus, when deciding on the scope of a competition, organizers need to think about the types of submissions that they will receive, particularly if these submissions are based on automatic statistical algorithms or human judgment. Although most competitions do not state this explicitly, the type of submission is usually implied based on the number of inputs required. In a large-scale forecasting competition, in which one has to provide many thousands of inputs, automatic algorithms might be the only feasible way. In smaller scale competitions, judgment can be used in predicting events, and in cases in which data are insufficient or even unavailable, judgment may be the only possible way to produce forecasts and estimate uncertainty. Consider, for instance, challenges similar to the ones posed within the Good Judgment project[1] and questions such as "What is the possibility that humans will visit Mars before the end of 2030?". In such cases, the focus of the competition is the events examined, and the required submissions have to be made judgmentally.

A third decision on the scope of a competition has to do with the format of the submissions requested from the participants. Although some of the forecasting competitions so far have asked for the submission of point forecasts only, it is preferable that submissions of uncertainty should be required too. This can be obtained by the submission of prediction intervals for one or multiple indicative quantiles or, even better, the submission of a fine grid of quantiles, including the extreme tails. If the event to be forecast has a discrete number of possible solutions, uncertainty can be provided in a form of predicted probabilities (e.g., 90% certainty) or categorical answers (e.g., low, moderate, and high confidence). Note also that it can be the case that a forecasting competition does not ask for forecasts (or estimates of uncertainty) per se, but the decisions to be made when using such forecasts. Examples include setting the safety stock in an inventory system, the selection of a portfolio of stocks in investing, or betting amounts for future events given their odds.

Finally, we believe that there is a need to have clear objectives and hypotheses before the commencing of a forecasting competition and define its scope-related attributes accordingly. Recent forecasting competitions have followed this example (see, for instance, Makridakis et al. 2020d), thus avoiding the problem of HARKing (Kerr 1998, Lishner 2021) by rationalizing the findings after the fact in perfect foresight (hindsight bias) or being driven by findings that are directly biased by the design of the competition itself. This is standard practice in other fields and is closely linked with the promotion of open research and the avoidance of "*p*-hacking." For example, psychological studies are often preregistered, an increasingly popular requirement for many academic journals.

3.2. Diversity and Representativeness

Regardless if the focus of a competition is generic or not, it is important that the events considered have a reasonable degree of diversity that allows for generalization of the findings and insights obtained. Diversity effectively refers to the heterogeneity of the events to be predicted. In the case of the forecasting competitions that provide historical information in the form of time series, diversity is usually determined by visualizing spaces based on time series features (Kang et al. 2017) that may include the strength of predictable patterns (trend, seasonality, autocorrelations, etc.), the degree of predictability (coefficient of variation, signal-to-noise ratio, entropy, etc.), the degree of intermittence and

sparseness (fast versus slow-moving items) as well as the length and periodicity of the data, among others. In time series, such features can be endogenously measured although in competitions in which past data are not provided, diversity can be appreciated with regards to the intent of the events under investigation and the implicit requirements from a participant's perspective in analyzing and producing forecasts or uncertainty for such events.

Diversity can also include the country of origin of the data, the type of data domains, the frequencies considered, the industries or companies investigated, and the time frame covered. For example, the results of a competition such as M5 that focused on the sales of 10 U.S. stores from a global grocery in 2016 would not necessarily apply to a grocery retailer in China in the same year or another U.S. grocery retailer in 2021. Similarly, they may not apply to other types of retailers, such as fashion, pharmaceutical, or technology or to firms operating online or providing different discounts and promotions strategies. Diversifying the data set of the competition so that multiple events of different attributes are considered is a prerequisite for designing competitions to represent reality realistically and ensure that its findings can be safely generalized across the domain(s), frequencies, or application(s) being considered.

Other competitions could be based on forecasting data of unknown or undisclosed sources as well as competitions based on forecasting synthetic time series (i.e., time series data generated through simulations). Such competitions would allow identifying the conditions under which particular forecasting models perform well, including time series characteristics, such as seasonality, trend, noise, and structural changes as well as decisions such as the forecasting horizon considered. These competitions would enable learning more from their results, understanding how methods and models that obey particular theoretical properties and assume certain distributions would work under real-life empirical settings.

3.3. Data Structure

Although, in some competition settings, it is possible that no data are provided at all, in most competitions, historical data are made available. Such data may be individual time series that are not somehow connected to one another. In such cases, although series are typically forecast separately, participants may attempt to apply cross-learning techniques to improve the accuracy of their solutions as was the case with the two top-performing solutions in the M4 competition (see for example Montero-Manso and Hyndman 2020, Montero-Manso et al. 2020, Semenoglou et al. 2020). It is also possible that competition data are logically organized to form hierarchical structures (Hyndman et al. 2011). Such structures do not have to be uniquely defined necessarily. For instance, in competitions such as M5, the sales of a company may be disaggregated by regions, categories, or both if grouped hierarchies are assumed. Given that, in many forecasting applications, hierarchies are present and information exchange between the series is possible, deciding on the correlation of the data provided is critical for determining under which circumstances the findings of the competition will apply.

Alternatively, the provided time series data may or may not be supported by additional information. For example, in competitions such as M4 in which the existence of time stamps may have led to information leakage about the actual future values of the series, dates should not be provided. However, when this information is indeed available, then multivariate settings may also be considered. Also, although data availability might be limited to the variables for which forecasts are required, explanatory/exogenous variables can also be provided. Information for such variables may match the time window of the dependent variables, part of it, or even exceed it. Explanatory or exogenous variables may be either provided by the organizers of the competition to its participants directly or collected by them through various external sources. In any case, it is important that the explanatory or exogenous variables used for producing the forecasts only refer to information that would have been originally available at the time the forecasts were produced and not after that point to make sure that no information about the actual future is leaked. For example, short-term weather forecasts may be offered as an explanatory variable for predicting wind production but no actual future weather conditions measured either onsite or at a nearby meteorological station.

3.4. Data Granularity

Data granularity refers to the most disaggregated level at which data is available and may refer to both cross-sectional and temporal aggregation levels (Spiliotis et al. 2020b). In most cases, the granularity of the data matches that of the variable to be forecast, but this does not have to always be the case. If, for example, a competition focuses on the sales of a particular product in the European Union, then country-level or even store-level sales might be helpful in improving forecasting performance. Similarly, smart-meter data may enhance the predictions of energy consumption at city level with hourly measurements being also useful in predicting daily demand. This is particularly true in applications in which data appear in mixed frequencies. For instance, in econometric regression, a quarterly time series may be used as an external regressor in forecasting a monthly time series.

Temporal granularity is more relevant when the data under investigation is organized over time (time series data). Increasingly, forecasting competitions are focusing on higher frequency data such as daily and weekly series, but this should not be considered a panacea for all future competitions. The choice of the frequency needs to be linked with the scope of the competition as low-frequency data is naturally used for supporting strategic decisions and high-frequency ones for supporting operations. For instance, daily data are not available for macroeconomic variables compared with monthly, quarterly, or yearly frequencies. Similarly, daily or hourly data would be more relevant in forecasting the sales of fresh products for store replenishment purposes. Finally, special treatment should be given in instances in which seasonality is not an integer number as, for example, when using weekly frequency data.

3.5. Data Availability

Data availability refers to the amount of information provided by the organizers for producing the requested forecasts. For time series competitions, this includes the number of historical observations available per series as well as the number of series contained in the data set. Note that both dimensions of data availability may be equally important in determining the performance of the submitted forecasts. For instance, in time series competitions, methods can be trained in both a series-by-series fashion, in which a large number of historical observations is desirable per series, and in a cross-learning one, in which data sets of multiple series are preferable for building appropriate models. In general, relatively large data sets are more advantageous over smaller ones so that the participants are capable of effectively training their models by extracting more information from the data. In addition, the probability of a participant winning the competition by luck rather than skill is effectively reduced. For example, in competitions of the size of the M4, which involved 100,000 series, it is practically impossible to win by making random choices (Spiliotis et al. 2020a).

Data availability can be also driven by the scope of the competition and the type of the events to be predicted. For example, if the competition focuses on new product or technological forecasting, data availability is naturally limited over time. Similarly, if the competition focuses on the sales of a manufacturer that produces a limited number of products, data availability is naturally bound over the series, requiring more manufacturers of the same industry to be included in the data set to expand its size and improve its representativeness. Moreover, data availability may be influenced by the frequency of the series, especially when multiple periodicities are observed. Hourly electricity consumption data, for instance, may contain three seasonal cycles: daily (every 24 hours), weekly (every 168 hours), and yearly (every 12 months). In addition, when dealing with seasonal data, it is generally believed that a minimum of three seasonal periods are required in order for the seasonal component of the series to capture the periodic patterns existing across time.

Certain domain-specific future forecasting competitions may not offer any data at all. In the era of big data and instant access to many publicly available sources of information, participants are usually in a position to gather the required data by themselves and also to complement their forecasts by using any other publicly available information. However, in the case that the organizers decide not to provide data, there is still a benefit to specifying a "default" data set to be used for evaluation purposes. Finally, in non–time series forecasting competitions, such as the Good Judgment project, quantitative data may not only not be provided but it may not be available at all.

3.6. Forecasting Horizon

The forecasting horizon may vary from predicting the present situation (also known as nowcasting, especially popular in predicting macroeconomic variables) to immediate-, short-, medium-, and long-term planning horizons. The exact definition of each planning horizon may differ with regards to the frequency of the data under investigation. For instance, for hourly data, 1–24 hours ahead is usually considered short-term forecasting. At the same time, one to three months ahead can also be regarded as "short-term" when working with monthly data. Accordingly, the forecasting horizon can be naturally bound based on the frequency of the series. For daily data, for example, it is probably unreasonable to produce forecasts for the following three years, a request that is reasonable for quarterly data.

The choice of the appropriate forecasting horizon is a function of various factors that may include the importance of the specific planning horizons for the application data, the user of the forecasts, and the hierarchical level of the forecast. Short-term horizons are suitable for operational planning and scheduling; mid-term horizons are appropriate for financial, budgeting, marketing, and employment decisions; and long-term forecasts are associated with strategic decisions that include technological predictions as well as business and capacity planning.

It is not uncommon in forecasting competitions to require forecasts for multiple periods ahead with the performance usually being measured as the average across all horizons. However, for some applications, such as store replenishment or production, it is more relevant to consider the cumulative forecast error (difference between the sum of actual values and the sum

of the forecasts for the lead time) rather than the average of the forecast errors across all horizons. In other applications, specific forecast horizons may be more important than others, so averaging across all horizons may not be so useful.

3.7. Evaluation Setup

In time series forecasting competitions, the most common design setup is to use historical data and conceal part of it to be used as test data to evaluate the performance of the submitted forecasts. The setup of concealing data may be further expanded to a number of rolling evaluation rounds. In single-origin evaluation, participants do not receive feedback on their performance, which is based on a single time window, which may not be representative of the entire series. For example, in electricity load forecasting in which three strong seasonal patterns are typically observed across the year, evaluating submissions by considering only one particular day, week, or month is not appropriate. Similarly, we found this to be a drawback of the evaluation setup used in M5.

To avoid the disadvantage of a single origin, the competition can be rolling (Tashman 2000), revealing some more of the hidden data each time and asking for new forecasts at each rolling iteration, providing the participants the opportunity to learn and improve their performance over time. A potential disadvantage of rolling-origin competitions is that they require more inputs and energy by the participants who may wish or have to adjust their models at each new round. For this reason, rolling-origin competitions display higher dropout rates, excluding also participants that are interested in participating but missed some early rounds and those that cannot be committed for a long period of time. An alternative could be a rolling-origin evaluation setup in which the participants provide the code for their solutions and then the organizers produce forecasts automatically for multiple origins as required. Yet, even if a forecasting competition does not have a rolling-origin evaluation design, participants may still decide to perform rolling-origin evaluation on the available (not the concealed) data to develop their algorithms, validate their performance under different settings, and select proper hyperparameters. This is closely related to the concept of time series cross-validation.

Instead of concealing data, a competition can be designed to take place on a real-time basis (live competition) with forecasts being evaluated against the actual data once they become available. The major advantage of live competitions is that participants can incorporate current information into their forecasts in real time, meaning that data and external variables could be fetched by the participants themselves based on their preferences and methods used. Also, information leakage about the actual future values becomes impossible, and the competition represents reality perfectly. Its disadvantage is that it is much more difficult to run (e.g., data must be collected in real time and evaluations must be accordingly updated) while taking some time until the actual values become available. A real-time competition may have a single-submission origin or multiple, rolling ones. In the latter case, feedback is explicitly provided to the participants in real time, allowing learning with each additional rolling iteration. Its major disadvantage is that it would be much more difficult to run and would require great motivation to participate given the considerable effort to keep informed and update the forecasts each time.

In some cases, when historical information is not available, concealing data are not an option. In such cases, the real-time design is the only alternative. Examples include elections and sports forecasting, in which a single evaluation origin typically is possible. However, participants may be also allowed to submit multiple forecasts (or revise previously submitted forecasts) until a particular point in time in live submission setups that include, for instance, prediction markets.

3.8. Performance Measurement

Another important decision in designing a competition is how the performance is measured and evaluated. It is common that the performance of the (point) forecasts is evaluated using statistical error measures. The choice of such measures should be based on a variety of factors, such as their theoretical foundation, applicability, and interpretability. Nowadays, relative and scaled error measures are generally preferred to percentage ones (Hyndman and Koehler 2006); however, the latter are still dominant in practice by being more intuitive. The evaluation of the estimation of the uncertainty around the forecasts can be performed using interval scores and proper scoring rules (Makridakis et al. 2020e). Proper scoring rules can address both sharpness and calibration, which is relevant in estimating the performance under fat tails. In all cases, however, robust measures with well-known statistical properties should be preferred to interpret the results and be confident of their value.

In cases in which the importance (volume and value) of the predicted events varies, performance measurements may include weighting schemes that account for such differences. This is especially true when evaluating the performance of hierarchical structured data in which some aggregation levels may be more important than others based on the decisions that the forecasts support. For instance, product–store forecasts may be considered more important than regional ones when used for supply chain management purposes with the opposite being true in cases in which forecasts are used for budgeting purposes. Similarly, forecasts

that refer to more expensive or perishable products may be weighted more than those that refer to inexpensive, fast-moving ones.

Whenever possible, instead of measuring the performance of the forecasts, one should measure their utility value directly. For instance, if the forecasts refer to investment decisions, the actual profit/loss from such investments can be measured. If the forecasts are to be used in a supply chain setting, then inventory-related costs, achieved service levels, and/or the variance of the forecasted variable can be useful measurements of their utility (Petropoulos et al. 2019). If more than two performance indicators need to be considered, then multicriteria techniques can be used to balance the performance across the chosen criteria. A simpler approach would be to assume equal importance across criteria and apply a root mean square evaluation measure. Care should be used to address any double counting that can arise when evaluating hierarchical series with multiple related levels.

Another critical factor in evaluating forecasts is the cost relating to various functions of the forecasting process, including data collection, costs related to preprocessing and cleansing the raw data, computational resources required to produce the forecasts (Nikolopoulos and Petropoulos 2018), and personnel time that is needed to revise or finalize such forecasts when judgment is needed. In standard forecasting competitions in which data are provided and the submission format usually refers to automatic forecasts, the computational cost can be easily measured by sharing the code used for their production and reproducing them. Once the computational cost is determined, it is important to contrast any improvements in performance against any additional costs. Effectively, this becomes a forecast value added (FVA) exercise (Gilliland 2013, 2019), accepting that computational time is often subject to programming skills and optimization techniques, making its correct estimate a considerable challenge.

3.9. Benchmarks

An important decision in designing competitions similar to selecting the performance measurements has to do with the choice of appropriate benchmarks. Such benchmarks should include both traditional and state-of-the-art models and algorithms that are suitable for the competition based on its scope and particularities of the data. Usually, benchmarks that include individual methods that have performed well in previous, similar competitions, are considered standard approaches for the forecasting task at hand or display a performance that is considered a minimum for such a task. For example, ARIMAX, linear regression, and decision tree–based models can be used as benchmarks in competitions that involve explanatory or exogenous variables; Croston's method in competitions that refer to inventory forecasting; the winning methods of the first three M competitions for the fourth one; and a random walk model for the performance of a major index, such as S&P 500 or Financial Times Stock Exchange in a stock market competition. Simple combinations of state-of-the-art methods are also useful benchmarks, especially given the ample evidence on their competitive performance (Makridakis et al. 2020a). It is good practice that the implementation of the benchmark methods is fully specified. This allows participants to obtain a valid starting point for their investigation and facilitates transparency and reproducibility, indicating the additional value added of a proposed method over that of an appropriate benchmark.

3.10. Learning

Regardless of the design of the competition, its objective should not be just to determine the winners of the examined forecasting task, but also to learn how to advance the theory and practice of forecasting by identifying the factors that contribute to the improvement of the forecasting accuracy and the estimation of uncertainty. This is the case for the competitions organized by academics but not in all others. In order to allow for such learning, sufficient information is required about how the forecasts are made by the participants with the code used (when applicable) being also published to facilitate replicability or reproducibility of the results (Boylan et al. 2015, Makridakis et al. 2018). For instance, this was true with the M4 competition, in which the vast majority of the methods were effectively reproduced by the organizers, but not with the M5 in which this was only done with the winners that were obliged to provide a clear description of their method along with their code as well as a small number of the top 50 submissions that complied with the repeated requests of the organizers to share such information.

Another idea would be for the organizers to make specific hypotheses before launching the competitions in order to test their predictions once the actual results become available, thus learning from their successes and mistakes. Such an approach would highlight the exact expectations of the competition and clarify its objectives, avoiding the problem of rationalizing the findings after the fact and allowing the equivalent of the scientific method, widely used in physical and life sciences, to be utilized in forecasting studies. This practice was followed in the M4 competition with positive results (Makridakis et al. 2020d) and has been repeated with the M5.

Finally, future forecasting competitions should challenge the findings of previous ones, testing the replicability of their results and trying to identify new, better forecasting practices as new, more accurate methods become available. For example, combining the forecasts of more than one method has been a

consistent finding of all competitions that has also flourished with machine and deep learning methods in which ensembles of numerous individual models are used for producing the final forecasts. Another critical finding, lasting until the M4 competition, was that simple methods were at least as accurate as more sophisticated ones. This finding was reversed with the M4 and M5 as well as the latest Kaggle competitions, indicating the need for dynamic learning by which new findings may reverse previous ones as new concepts (such as cross-learning) and more accurate methods are outperforming existing ones.

4. Mapping the Design Attributes of Past Competitions

In this section, we map the design attributes discussed in Section 3 to past, major forecasting competitions with the aim to identify their commonalities and design gaps while also highlighting the advantages and drawbacks of each. We focus on the major competitions organized by the community of the International Institute of Forecasters and also on recent competitions hosted on Kaggle. In total, we consider 17 forecasting competitions, which are listed in the rows of Tables 2–4 with the columns of the table presenting the various design attributes discussed in the previous section. Table 5 offers citations to the relevant papers and links to the data and the winning submissions when available. From the total of the 17 competitions conducted in the last 40 years, seven were hosted by Kaggle, five were M competitions, three were energy ones, and there was a single tourism and a sole neural network one.

There are several common attributes characterizing practically all 17 competitions. First, the submissions required were all numerical except for M2, which asked, in addition, for judgmental inputs from the forecasters. Second, in all but three competitions (M2, GEF2012, and GEF2017), the submission setup involved a fixed origin evaluation on concealed data. Third, there were only three live competitions (M2, GEF2017, and Web Traffic Time Series Forecasting), which were also limited in a small number of evaluation rounds. Fourth, the majority of the competitions (15 out of the 17) required point forecasts although five also demanded uncertainty estimates, ranging from two quantiles in M4 to 99 in GEF2014. Fifth, although there is a balance between generic, specific, and semi-specific competitions, we observe that specific ones focus on tourism, energy, and retail forecasting applications with the majority of the specific ones including high-frequency, hierarchically structured series and explanatory/exogenous variables and the generic ones focusing on lower frequency data, such as yearly, quarterly, and monthly, that were not accompanied by additional information. Moreover, there seems to be a trend toward more detailed data sets as more recent competitions move from individual time series to hierarchically structured ones that may be influenced by explanatory or exogenous variables. Sixth, none of the competitions required submissions in the form of decisions or evaluated their performance in terms of utility or cost-based measures, utilizing various statistical measures that build on absolute, squared, and percentage errors. Finally, with the exception of the competitions that were organized by academics, little emphasis was given to the element of learning and how to improve forecasting performance, and few noncompetitive benchmarks were considered for evaluating such improvements. For instance, the M and energy competitions included several variations of naive approaches, combinations of exponential smoothing models, ARIMA, the Theta method, and simple machine learning or statistical regression methods, and Kaggle ones featured only naive methods and dummy submissions (e.g., all forecasts are set equal to the global average or median or zero).

These observations reveal both a consensus to apply what has worked in the past and that is easy to implement in practice as well as a desire for experimentation. What is clear from Tables 2–4 is the difference between the top 12 competitions organized by the academic community and the last five ones hosted by Kaggle. In the former, the emphasis is on learning by publishing the results in peer-reviewed journals and providing open access to the data and forecasts so that others can comment on the findings, respond to their value, and suggest improvements for future ones. Thus, it is not surprising that the number of citations received by the former (close to 6,300 as of October 2021) are significantly more than that of the latter (probably limited to less than 100). Citations are an integral part of learning as other researchers read the cited work and become aware of its findings that they then try to extend in additional directions. At the same time, the Kaggle approach encourages cooperation among competitors, for example, in the form of forum discussions and code exchange, to come up with the best solution to the problem at hand without concern for the dissemination of the findings to the wider data science community. Equally important, Kaggle involves public leaderboards that provide instant feedback to the participants in order for them to revise their methods and resubmit forecasts, thereby encouraging competition and driving innovation (Athanasopoulos and Hyndman 2011). A clear breakthrough will come by combining the academic and Kaggle approaches by exploiting the advantages of both as there is no reason that Kaggle scientists will not be willing to share their knowledge so that others can also learn from their experience nor for the academics not to benefit from leaderboards, public

Table 2. Mapping the Design Attributes of Past Competitions: Scope

	Scope							
	Focus			Type of submission		Format of submission		
Competition (year)	Generic	Semispecific	Specific	Numerical	Judgmental	Point forecasts	Uncertainty estimates	Decisions
M or M1 (1982)	Macro, micro, industry, and demographic			✓		✓		
M2 (1993)	Micro and macro			✓	✓	✓		
M3 (2000)	Micro, macro, industry, demographic, finance, and other			✓		✓		
NN3 (2006)		Industry		✓		✓		
Tourism (2011)		Tourism		✓		✓	Four quantiles (computed only for benchmarks)	
GEFCom 2012 (2012)			Load/wind	✓		✓		
GEFCom 2014 (2014)		Load/wind/solar/price		✓			Ninety-nine quantiles	
GEFCom 2017 (2017)			Load	✓			Five quantiles (nine in qualifying match)	
M4 (2018)	Micro, finance, macro, industry, demographic, and other			✓		✓	Two quantiles	
M5 (2020)			Retail sales	✓		✓	Nine quantiles	
Walmart Recruiting: Store Sales Forecasting (2014)			Retail store sales	✓		✓		
Walmart Recruiting II: Sales in Stormy Weather (2015)			Retail sales of weather-sensitive products	✓		✓		
Rossmann Store Sales (2015)			Drug store sales	✓		✓		
Grupo Bimbo Inventory Demand (2016)			Bakery goods sales	✓		✓		
Web Traffic Time Series Forecasting (2017)			Traffic of web pages	✓		✓		
Corporación Favorita Grocery Sales Forecasting (2018)			Grocery store sales	✓		✓		
Recruit Restaurant Visitor Forecasting (2018)			Restaurant visits	✓		✓		

Table 3. Mapping the Design Attributes of Past Competitions: Diversity and Representativeness (Specified Based on the Origin and Size of the Data Set and the Length of the Period Examined), Data Structure, Data Granularity, and Data Availability

		Data structure		Data granularity		Data availability	
Competition (year)	Diversity and representativeness	Hierarchies	Exogenous variables	Cross-sectional	Temporal	Number of events	Observations per event (min-median-max)
M or M1 (1982)	Moderate			From country to company	Monthly, quarterly, and yearly	1,001	Monthly 30-66-132, Quart 10-40-106, Year 9-15-52
M2 (1993)	Low		✓	Country and company	Monthly and quarterly	29	45-82-225 (monthly), 167-167-167 (quarterly)
M3 (2000)	Moderate			From country to company	Monthly, quarterly, yearly, and other	3,003	Monthly 48-115-126, quart 16-44-64, year 14-19-41, others 63-63-96
NN3 (2006)	Low			From country to region	Monthly	111	50-116-126
Tourism (2011)	Moderate		✓	From country to company	Yearly, quarterly, and monthly	1,311	7-23-43 (yearly), 22-102-122 (quarterly), 67-306-309 (monthly)
GEFCom 2012 (2012)	Moderate	Hierarchical / ✗	✓ (only for train data) / ✓	Utility zone/wind farm	Hourly	44,397	38,070/19,033
GEFCom 2014 (2014)	Moderate		✓	Utility/ wind farm/solar plant/zone	Hourly	1/10/3/1	Round-based: 50,376-60,600/6,576-16,800/8,760-18,984/21,528-25,944
GEFCom 2017 (2017)	Moderate	Hierarchical	✓	Delivery point meters (zones in qualifying match)	Hourly	161 out of 169 (eight in qualifying match)	2,232-61,337 (Round based 119,904-122,736 in qualifying match)
M4 (2018)	High			From country to company	Monthly, quarterly, yearly, daily, hourly, and weekly	100,000	42-202-2,794 (monthly), 16-88-866 (quarterly), 13-29-835 (yearly), 93-2,940-9,919 (daily), 700-960-960 (hourly) & 80-934-2,597 (weekly)
M5 (2020)	Moderate	Grouped	✓	Store-product	Daily	30,490	96-1,782-1,941
Walmart Recruiting: Store Sales Forecasting (2014)	Moderate	Grouped	✓	Store-department	Weekly	3,331	1-143-143
Walmart Recruiting II: Sales in Stormy Weather (2015)	Moderate	Grouped	✓	Store-product	Daily	4,995	851-914-1,011

Table 3. (Continued)

Competition (year)	Data structure			Data granularity		Data availability	
	Diversity and representativeness	Hierarchies	Exogenous variables	Cross-sectional	Temporal	Number of events	Observations per event (min-median-max)
Rossmann Store Sales (2015)	Moderate	Grouped	✓	Store	Daily	1,115	941-942-942
Grupo Bimbo Inventory Demand (2016)	Moderate	Hierarchical		Store-product	Weekly	26,396,648	1-2-7
Web Traffic Time Series Forecasting (2017)	Moderate	Grouped		Page and traffic type	Daily	145,063	803
Corporación Favorita Grocery Sales Forecasting (2018)	Moderate	Grouped	✓	Store-product	Daily	174,685	1-1,687-1,688
Recruit Restaurant Visitor Forecasting (2018)	Moderate	Grouped	✓	Restaurant	Daily	829	47-296-478

discussions, or code exchange. In our view, such a breakthrough is inevitable in the near future.

5. Future Forecasting Competitions

5.1. Proposed Principles

In the previous sections, we discuss the design aspects of forecasting competitions and map these to the past ones. Then, we elaborate on the design opportunities, that is, the gaps that past forecasting competitions have left. In this section, we propose some principles for future competitions.

5.1.1. Replicability. One crucial aspect of any research study is that its results should be able to be replicated. This has been an increasing concern across sciences (Goodman et al. 2016), including the forecasting field (Boylan et al. 2015, Makridakis et al. 2018). To be most useful to the forecasting community, competitions should ideally be transparent and allow for the full replicability of the results. One way to achieve this is by requiring submission of the source code (or at least an executable file) of the participating solutions coupled with sufficient descriptions and open libraries for benchmarks and performance measures. Reproducibility also allows those interested to test if the results of a forecasting competition hold for other data sets, performance measures, forecasting horizons, and testing periods while also enabling computational cost comparisons. In addition, it would enable rolling-origin evaluation to be done in an automated fashion, reflecting the realistic situation when forecasting models are built and then run repeatedly without the opportunity to tweak them each time an output is generated.

5.1.2. Representativeness. If possible, organizers of forecasting competitions should aim for a diverse and representative set of data. A high degree of representativeness (Spiliotis et al. 2020a) allows for a fuller analysis of the results, enabling us to understand the conditions under which some methods perform better than others. Moving away from "the overall top-performing solution wins it all," we are able to effectively understand the importance of particular features (including frequencies) and gain insights into the performance of various methods for specific industries or organizations. One strategy to improve representativeness could be to look at the feature space for time series included in a competition in comparison with other samples from the relevant population of series (Kang et al. 2017, Fry and Brundage 2020, Spiliotis et al. 2020a).

Forecasting under highly stable conditions offers little challenge. Therefore, competition organizers should consider evaluating forecasts across a range of conditions, including conditions in which past patterns or relationships are bound to fail (e.g., structural changes,

Table 4. Mapping the Design Attributes of Past Competitions: Forecasting Horizon, Evaluation Setup, Performance Measurement, Benchmarking, and Learning

		Evaluation setup		Performance measurement				
Competition (year)	Forecasting horizon	Live	Rounds	Forecast	Utility	Cost	Benchmarks	Learning
M or M1 (1982)	1-18 (monthly), 1-8 (quarterly), and 1-6 (yearly)		1	MAPE, MSE, AR, MdAPE, and PB			Naive and ES	✓
M2 (1993)	1-15 (monthly) and 1-5 (quarterly)	✓	2	MAPE			Naive, ES, ARIMA, and combination of ES	✓
M3 (2000)	1-18 (monthly), 1-8 (quarterly), 1-6 (yearly), and 1-8 (other)		1	sMAPE			Naive, ES, combination of ES, and ARIMA	✓
NN3 (2006)	1-18		1	sMAPE			Naive, ES, theta, combination of ES, expert systems, ARIMA, vanilla NNs, and SVR	✓
Tourism (2011)	1-4 (yearly), 1-8 (quarterly), and 1-24 (monthly)		1	MASE and coverage			Naive, ES, theta, ARIMA, expert systems, and models with explanatory variables	✓
GEFCom 2012 (2012)	1-168/1-48		1/1 of 157 periods	RMSE			Vanilla MLR/naive	✓
GEFCom 2014 (2014)	1-24 for price and (1-31)*24 for the rest		15	PL improvement over benchmark, adjusted for simplicity and quality			Naive	✓
GEFCom 2017 (2017)	8,784 ((1-31)*24 in qualifying match)	✓	1 (6 in qualifying match)	PL improvement over benchmark			Vanilla MLR	✓
M4 (2018)	1-18 (monthly), 1-8 (quarterly), 1-6 (yearly), 1-14 (daily), 1-48 (hourly), and 1-13 (weekly)		1	OWA and MSIS			Naive, ES, theta, combination of ES, ARIMA & vanilla NNs	✓
M5 (2020)	1-28		1	WRMSSE/WSPL			Naive, ES, ARIMA, Croston and variants, combinations, NNs, RTs	✓
Walmart Recruiting: Store Sales Forecasting (2014)	1-39		1	WMAE			All zeros	

Table 4. (Continued)

		Evaluation setup		Performance measurement				
Competition (year)	Forecasting horizon	Live	Rounds	Forecast	Utility	Cost	Benchmarks	Learning
Walmart Recruiting II: Sales in Stormy Weather (2015)	1-25 (in an interpolation fashion)		1	RMSLE			All zeros	
Rossmann Store Sales (2015)	1-48		1	RMSPE			All zeros, median day of week	
Grupo Bimbo Inventory Demand (2016)	1-2		1	RMSLE			All sevens	
Web Traffic Time Series Forecasting (2017)	3-65	✓	1	sMAPE			All zeros	
Corporación Favorita Grocery Sales Forecasting (2018)	1-16		1	NWRMSLE			Mean item sales, last year sales, all zeros	
Recruit Restaurant Visitor Forecasting (2018)	1-39		1	RMSLE			Median visit, all zeros	

Note. MAPE, mean absolute percentage error; MSE, mean squared error; AR, average ranking; MdAPE, median absolute percentage error; PB, percentage best; sMAPE, symmetric mean absolute percentage error; MASE, mean absolute scaled error; RMSE, root mean squared error; PL, pinball loss; OWA, overall weighted average; MSIS, mean scaled interval score; WRMSSE, weighted root mean squared scaled eError; WSPL, weighted scaled pinball loss; WMAE, weighted mean absolute error; RMSLE, root mean squared logarithmic error; RMSPE, root mean square percentage error; NWRMSLE, normalized weighted root mean squared logarithmic error.

fat tails, recessions, pandemics) to identify methods that are more robust under such conditions in order to offer valuable insights and enhance our understanding toward managing such situations. Moreover, competitive actions and reactions should be included as this is the reality in which modern companies operate. Future competitions could also explore the possibility of multivariate (but not hierarchically structured) sets of data that also include information directly coming from online data devices, including nowcasting.

5.1.3. Robust Evaluation. For the results of a competition to be meaningful, robust evaluation strategies must be considered. We suggest moving away from evaluating forecasts produced from a single origin, especially when the data set considered is homogeneous, and introduce rolling evaluation schemes. This would be particularly relevant for seasonal time series, in which evaluation periods should cover many different times within the calendar year if not one or more complete years. This would mitigate against sampling bias caused by evaluating forecasts over a short interval. Organizers could also consider evaluating holdout sets for representativeness using the principles discussed. Future competitions could also offer multiple evaluation rounds in a live setup. Undoubtedly, this would add a more pragmatic dimension to forecasting competitions.

5.1.4. Measuring Impact on Decisions. Reflection to reality may include how a forecasting solution is indeed implemented in practice, but it also offers metrics of performance measurement that are directly linked to decisions. For example, in inventory forecasting, Petropoulos et al. (2019) map the forecasting performance of various forecasting methods to their inventory performance, measured in terms of holding cost, achieved service levels, and variance of forecasts. We argue that future forecasting competitions may need to shift the focus to measure the utility of the forecasts or uncertainty directly. In many applications, the translation from point and probabilistic forecasts to their decision-making implications is a big step and a formidable challenge as utility can be not only nonlinear but also nonmonotonic. Whenever possible, such utility should be expressed in monetary terms that would allow comparing meaningful trade-offs. Such trade-offs could include conflicting optimization criteria (such as inventory holdings versus service levels) and also would allow for a more systematic value-added analysis of the complexity of the participating solutions and their computational (or otherwise) cost. However, we should be careful to distinguish between evaluating the impact of forecasts on decisions versus evaluating the impact of decisions themselves.

Table 5. Past Forecasting Competitions: Citations and Additional Information

Competition	Citation	Links to data and/or winning methods
M or M1 (1982)	Makridakis et al. (1982)	https://forecasters.org/resources/time-series-data/ https://cran.r-project.org/package=Mcomp
M2 (1993)	Makridakis et al. (1993)	https://forecasters.org/resources/time-series-data/
M3 (2000)	Makridakis and Hibon (2000)	https://forecasters.org/resources/time-series-data/ https://cran.r-project.org/package=Mcomp
NN3 (2006)	Crone et al. (2011)	http://www.neural-forecasting-competition.com/NN3
Tourism (2011)	Athanasopoulos et al. (2011)	https://www.kaggle.com/c/tourism1 https://www.kaggle.com/c/tourism2 https://github.com/robjhyndman/tscompdata
GEFCom 2012 (2012)	Hong et al. (2014)	https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting http://www.drhongtao.com/gefcom/2012
GEFCom 2014 (2014)	Hong et al. (2016)	http://www.drhongtao.com/gefcom/2014
GEFCom 2017 (2017)	Hong et al. (2019)	http://www.drhongtao.com/gefcom/2017
M4 (2018)	Makridakis et al. (2020b)	https://forecasters.org/resources/time-series-data/ https://github.com/Mcompetitions/M4-methods https://github.com/carlanetto/M4comp2018
M5 (2020)	Makridakis et al. (2020c, e)	https://www.kaggle.com/c/m5-forecasting-accuracy https://www.kaggle.com/c/m5-forecasting-uncertainty https://forecasters.org/resources/time-series-data/ https://github.com/Mcompetitions/M5-methods
Walmart Recruiting: Store Sales Forecasting (2014)		https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting
Walmart Recruiting II: Sales in Stormy Weather (2015)		https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather
Rossmann Store Sales (2015)		https://www.kaggle.com/c/rossmann-store-sales
Grupo Bimbo Inventory Demand (2016)		https://www.kaggle.com/c/grupo-bimbo-inventory-demand
Web Traffic Time Series Forecasting (2017)		https://www.kaggle.com/c/web-traffic-time-series-forecasting
Corporación Favorita Grocery Sales Forecasting (2018)		https://www.kaggle.com/c/favorita-grocery-sales-forecasting
Recruit Restaurant Visitor Forecasting (2018)		https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting

5.1.5. Showcase FVA. Forecasting competitions need to clearly demonstrate the added value of a proposed solution over the state-of-the-art methods and benchmarks. The choices of benchmarks are wide and can include top-performing methods from previous competitions. For instance, a future large-scale generic forecasting competition could have as a benchmark the winning method of Smyl (2020) in the M4 competitions or N-Beats (Oreshkin et al. 2019). Also, a future competition on retail forecasting should include as benchmarks the top methods from M5 or other Kaggle competitions. Finally, the inclusion of past winning approaches as benchmarks can act as a way of measuring improvements from new competitions and determining the value they have added in forecasting performance. We suggest that an FVA analysis should be multifold and include not only the performance of the point forecasts, but also the performance in estimating uncertainty, dealing with fat tails, and the computational cost and complexity of each method. The last two aspects (complexity and cost) are increasingly important for the acceptance and successful implementation of a method, particularly when millions of forecasts or estimates of uncertainty are needed on a weekly basis.

5.1.6. Enhancing Knowledge. We want to see future competitions focus on contributing new learning and insights to the forecasting community, moving away from a horse-race exercise toward bridging the gap between theory and practice. If possible, forecasting competitions should not focus on picking a winning team, but rather understanding what constitutes a winning method, that is, a successful underlying mechanism, and how the results could be transferable to other settings. For instance, cross-learning is proven to be an effective method in the M5 competition setup in which grouped series are forecast, and combination is a winning approach for univariate time series forecasting tasks. They should be able to show how the results can be implemented to improve the baseline and what the consequences are of the forecasting accuracy or uncertainty on decision making. Although not an objective of all past forecasting competitions, learning must become an integral part of all future ones to maximize their expected value by making their

findings widely known to anyone wishing to utilize them to improve the theory or practice of forecasting. The current trend toward open access of knowledge must be applied to forecasting competitions as its findings will improve the much-talked-about circular economy by eliminating waste and achieving optimal results across a wide variety of operational and strategic areas.

5.1.7. Merging the Academic and Kaggle Approaches.

There is much to gain and nothing to lose by combining the academic approach of disseminating learning and achieving high citations with that of Kaggle encouraging high collaboration and open participation by the participating groups. Facilitating learning by widely disseminating the findings of Kaggle competitions benefits the entire data science community and avoid concerns about their relevance (see Chawla 2020). At the same time, stimulating a more supportive collaborative spirit in academic competitions can encourage innovation and foster team effort as long as some clever ways of supporting collaborative work can be adopted.

We note that one strategy that enables both replicability and robust evaluation is the use of code-only competitions, in which the organizers of the competition use the submitted codes to produce forecasts for multiple origins. Participants may be given the option to alter their code in key points; for instance, the participants may resubmit their codes every quarter when forecasts are produced and evaluated every week. Such a strategy also reflects the real-world situation in that a forecasting model used in practice may not be able to benefit from manual tweaking between each subsequent forecast generation, leading also to unreasonably higher costs in terms of postperformance analysis and reengineering. In a code-only competition, an additional requirement could be that the code must run within a specific time limit given a specific data set and a computer architecture. That would focus attention on finding methods that achieve optimal outcomes within a computationally constrained environment, limiting the need to consider metrics related to the computational cost.

5.2. Toward Forecasting Athlons

The capabilities of forecasters to make and use forecasts have progressed significantly during the last four decades based in part on the findings of forecasting competitions that, as Hyndman (2020) mentions, contribute a great deal to improve the theory and practice of forecasting and provide considerable value to business firms using such predictions to improve their operations. Forecasting competitions can be further expanded beyond business applications to other social science areas to provide objective information and improve policy and decision making. In addition, uncertainty needs to receive attention among academicians and practitioners alike. It must be accepted that uncertainty always exists and cannot be avoided or reduced no matter if we would like to live in a world without uncertainty. What we have to do is to understand its risk implications and consider what actions to take to minimize the negative consequences involved. Directly linking forecasting competitions with decision-making aspects and the utility of the forecasts is also very important and allows us to gain further insights on the use of forecasts in practice.

Attempting to incorporate all of these considerations into a single forecasting contest or evaluation can be difficult if not impossible. Therefore, we suggest that some future forecasting competitions could move from featuring a single challenge to multiple ones with a winner in each challenge and an overall winner for the entire competition. For example, a future forecasting competition could be a pentathlon (or hexathlon or heptathlon …), in which the various challenges are organized around domain skills, such as (i) forecasting of univariate series with no exogenous information, (ii) forecasting of multivariate series, (iii) forecasting of series with exogenous information (e.g., weather, price, promotion activity, competitor actions, etc.), (iv) long-range forecasting with market or competitor uncertainties, (v) forecasting of intermittent series, (vi) lifecycle forecasting, etc. We view this structure as valuable for a comprehensive forecasting competition for several reasons.

First, as noted, it may be impossible to cover all of the ideal aspects and core forecasting skills in a single challenge. Second, the use of multiple challenges also allows for greater diversity of application domains. Third, this enables evaluation of participants in multiple skill domains and reduces the randomness in the final results and rankings.

Another possibility is the organization of challenges around applications. For example, within a manufacturing company, that could include forecasting for (i) inventory, (ii) scheduling, (iii) budget, (iv) cash flow, (v) long-range planning, and (vi) human resources, among others. Such challenges better reflect reality and showcase FVA because, in real life, in order for an organization to thrive, accurate forecasts and correct estimates of uncertainty are required for multiple aspects of its strategy-, planning-, and operations-related decisions.

Future domain-specific competitions could focus on new application areas, covering the economy (gross domestic product, monetary policies, interest rates), finance (stocks, commodities), operations (new products, promotional forecasting, spare parts, predictive maintenance, reverse logistics), healthcare (epidemics, healthcare management, mortality, preventable medical errors), climate, sports, elections, call centers, big and

megaprojects, transportation, and online commerce, among others. Finally, the increasing role of judgment in various aspects of the forecasting process, such as adjusting or finalizing forecasts or even selecting between models, calls for further investigations and should be further explored in future competitions.

Overall, we foresee that forecasting competitions still have much to offer if they are designed in a way to represent reality even closer. If forecasting competitions are done systematically and consistently, they allow for comparisons and assessing improvements over time while covering also various areas of applications and time horizons.

6. Conclusions

Forecasting competitions, the equivalent of laboratory experimentation in physical and life sciences, provide useful, objective information to improve the theory and practice of forecasting, advancing the field and enhancing decision and policy making. This paper describes all major past forecasting competitions, discussed their design attributes, and identified those of ideal competitions, extending their coverage to a multitude of applications and social science areas, echoing Hyndman's suggestion that the main objective of competitions is learning as much as possible rather than identifying winners.

The main part of the paper describes 10 design attributes to be considered by the organizers of competitions who need to decide those relevant for their own, considering trade-offs between optimal choices and practical concerns, such as costs, as well as elements related to the time and effort required to participate in them. Next, the paper maps all pertinent past competitions in respect to the described design attributes, identifying similarities and differences between the competitions, as well as design gaps and making suggestions about the attributes that future competitions should consider, putting a particular emphasis on learning as much as possible from their implementation in order to help improve forecasting accuracy and uncertainty.

The majority of past competitions concentrate on point forecasts. Our proposal is that all future competitions should also request probabilistic forecasts for a sufficient number of quantiles so that both the main part of the uncertainty distribution and its tails are effectively captured. This is of critical importance because both point forecasts and uncertainty estimates need to be considered in all future-oriented decisions. Another concentration of past competitions is the usage of the single-origin, concealed-data evaluation setup, which is the easiest to implement and requires the least time to participate. This practice has to change by first expanding the evaluation setup to several rolling origins and then potentially moving to rolling live competitions that may be the hardest to run but provide a great value as they run on a real-time basis in which all information is currently available and judgmental inputs can be directly incorporated. Clearly, there are trade-offs that need to be considered between the number of rolling origins used and the amount of effort required to complete the competition with the same trade-offs deliberated between live- and concealed-data ones. Competitions are costly to run, requiring a considerable amount of effort both to be implemented and participate. Their advantage is the objective evidence they provide to improve the theory and practice of forecasting. As such, they must continue, and maybe their costs are financed by a joint industry or specific group effort in search of solutions to improve the accuracy and uncertainty of their specific predictions. Whatever the solution, the practice of forecasting competitions must expand in the future to gain the maximum benefits from their findings.

The final section of the paper ends with the observation that the task of forecasting presents a multitude of challenges for organizations and societies. Business firms, for instance, must predict the level of their inventories for the large number of items sold in their stores, schedule their production and workforce, and estimate their budget requirements and their long-term strategic plans, including competitive and technological forecasts. Moreover, economic forecasting is also necessary at the societal level as well as energy, climate, and health predictions. Such a multitude of challenges cannot be met with a single competition. Instead, a number of them would be demanded as in a pentathlon in which different challenges take place, identifying the winner of each and also the overall one that would contribute the most to the overall forecasting effort among various areas or even industries with varying characteristics.

Endnote

[1] See https://goodjudgment.com/.

References

Askanazi R, Diebold FX, Schorfheide F, Shin M (2018) On the comparison of interval forecasts. *J. Time Series Anal.* 39(6):953–965.

Athanasopoulos G, Hyndman RJ (2011) The value of feedback in forecasting competitions. *Internat. J. Forecasting* 27(3):845–849.

Athanasopoulos G, Hyndman RJ, Song H, Wu DC (2011) The tourism forecasting competition. *Internat. J. Forecasting* 27(3):822–844.

Bojer CS, Meldgaard JP (2020) Kaggle forecasting competitions: An overlooked learning opportunity. Internat. J. Forecasting 37(2): 587–603.

Boylan JE, Goodwin P, Mohammadipour M, Syntetos AA (2015) Reproducibility in forecasting research. *Internat. J. Forecasting* 31(1): 79–90.

Chawla V (2020) How much is Kaggle relevant for real-life data science? Accessed January 5, 2021, https://analyticsindiamag.com/how-much-is-kaggle-relevant-for-real-life-data-science/.

Crone SF, Hibon M, Nikolopoulos K (2011) Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *Internat. J. Forecasting* 27(3): 635–660.

Donoho D (2017) 50 years of data science. *J. Comput. Graphical Statist.* 26(4):745–766.

Fildes R (2001) Beyond forecasting competitions. *Internat. J. Forecasting* 17(4):556–560.

Fry C, Brundage M (2020) The M4 forecasting competition—A practitioner's view. *Internat. J. Forecasting* 36(1):156–160.

Gilliland M (2013) FVA: A reality check on forecasting practices. *Foresight: Internat. J. Appl. Forecasting* 29:14–18.

Gilliland M (2019) The value added by machine learning approaches in forecasting. *Internat. J. Forecasting* 36(1):161–166.

Goodman SN, Fanelli D, Ioannidis JPA (2016) What does research reproducibility mean? *Sci. Translational Medicine* 8(341):341ps12.

Hong T, Pinson P, Fan S (2014) Global energy forecasting competition 2012. *Internat. J. Forecasting* 30(2):357–363.

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

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

Hyndman RJ (2001) It's time to move from "what" to "why." *Internat. J. Forecasting* 17(4):567–570.

Hyndman RJ (2020) A brief history of forecasting competitions. *Internat. J. Forecasting* 36(1):7–14.

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. *Internat. J. Forecasting* 22(4):679–688.

Hyndman RJ, Ahmed RA, Athanasopoulos G, Shang HL (2011) Optimal combination forecasts for hierarchical time series. *Comput. Statist. Data Anal.* 55(9):2579–2589.

Kang Y, Hyndman RJ, Smith-Miles K (2017) Visualising forecasting algorithm performance using time series instance spaces. *Internat. J. Forecasting* 33(2):345–358.

Kerr NL (1998) HARKing: Hypothesizing after the results are known. *Personality Soc. Psych. Rev.* 2(3):196–217.

Lishner DA (2021) HARKing: Conceptualizations, harms, and two fundamental remedies. *J. Theoretical Philosophical Psych.* Forthcoming.

Makridakis S, Hibon M (1979) Accuracy of forecasting: An empirical investigation. *J. Roy. Statist. Soc. Ser. A* 142(2):97–145.

Makridakis S, Hibon M (2000) The M3-competition: Results, conclusions and implications. *Internat. J. Forecasting* 16(4):451–476.

Makridakis S, Assimakopoulos V, Spiliotis E (2018) Objectivity, reproducibility and replicability in forecasting research. *Internat. J. Forecasting* 34(4):835–838.

Makridakis S, Hyndman RJ, Petropoulos F (2020a) Forecasting in social settings: The state of the art. *Internat. J. Forecasting* 36(1): 15–28.

Makridakis S, Spiliotis E, Assimakopoulos V (2020b) The M4 competition: 100,000 time series and 61 forecasting methods. *Internat. J. Forecasting* 36(1):54–74.

Makridakis S, Spiliotis E, Assimakopoulos V (2020c) The M5 accuracy competition: Results, findings and conclusions. Preprint, submitted October 6, https://www.researchgate.net/publication/344487258_The_M5_Accuracy_competition_Results_findings_and_ conclusions.

Makridakis S, Spiliotis E, Assimakopoulos V (2020d) Predicting/hypothesizing the findings of the M4 competition. *Internat. J. Forecasting* 36(1):29–36.

Makridakis S, Chatfield C, Hibon M, Lawrence M, Mills T, Ord K, Simmons LF (1993) The M2-competition: A real-time judgmentally based forecasting study. *Internat. J. Forecasting* 9(1):5–22.

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. *J. Forecasting* 1(2):111–153.

Makridakis S, Spiliotis E, Assimakopoulos V, Chen Z, Gaba A, Tsetlin I, Winkler RL (2020e) The M5 uncertainty competition: Results, findings and conclusions. Preprint, November 30, https://www.researchgate.net/publication/346493740_The_M5_Uncertainty_competition_Results_findings_and_conclusions.

Montero-Manso P, Hyndman RJ (2021) Principles and algorithms for forecasting groups of time series: Locality and globality. *Internat. J. Forecasting* 37(4):1632–1653.

Montero-Manso P, Athanasopoulos G, Hyndman RJ, Talagala TS (2020) FFORMA: Feature-based forecast model averaging. *Internat. J. Forecasting* 36(1):86–92.

Newbold P, Granger CWJ (1974) Experience with forecasting univariate time series and the combination of forecasts. *J. Roy. Statist. Soc. Ser. A* 137(2):131–165.

Nikolopoulos K, Petropoulos F (2018) Forecasting for big data: Does suboptimality matter? *Comput. Oper. Res.* 98:322–329.

Oreshkin BN, Carpov D, Chapados N, Bengio Y (2019) N-beats: Neural basis expansion analysis for interpretable time series forecasting. Preprint, submitted May 24, https://arxiv.org/abs/1905.10437.

Petropoulos F, Wang X, Disney SM (2019) The inventory performance of forecasting methods: Evidence from the M3 competition data. *Internat. J. Forecasting* 35(1):251–265.

Petropoulos F, Apiletti D, Assimakopoulos V, Babai MZ, Barrow DK, Ben Taieb S, Bergmeir C, et al. (2020) Forecasting: Theory and practice. Preprint, submitted December 4, https://arxiv.org/abs/2012.03854.

Semenoglou AA, Spiliotis E, Makridakis S, Assimakopoulos V (2020) Investigating the accuracy of cross-learning time series forecasting methods. *Internat. J. Forecasting* 37(3):1072–1084.

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

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

Spiliotis E, Petropoulos F, Kourentzes N, Assimakopoulos V (2020b) Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. *Appl. Energy* 261:114339.

Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: An analysis and review. *Internat. J. Forecasting* 16(4):437–450.

Winkler RL (1972) A decision-theoretic approach to interval estimation. *J. Amer. Statist. Assoc.* 67(337):187–191.