A brief history of forecasting competitions

Rob J Hyndman
Department of Econometrics & Business Statistics
Monash University, Clayton VIC 3800, Australia
Email: Rob.Hyndman@monash.edu

27 April 2019

A brief history of forecasting competitions

Abstract

Forecasting competitions are now so widespread that it is often forgotten how controversial they were when first held, and how influential they have been over the years. I briefly review the history of forecasting competitions, and discuss what we have learned about their design and implementation, and what they can tell us about forecasting. I also provide a few suggestions for potential future competitions, and for research about forecasting based on competitions.

Keywords: evaluation, forecasting accuracy, Kaggle, M competitions, neural networks, prediction intervals, probability scoring, time series

Prediction competitions go back millennia; for example, rival diviners in ancient Greece competed to predict the future more accurately (Raphals 2013, p124). However, for general time series forecasting (i.e., predicting the future of regularly observed data over time), the history is much more limited and goes back only about 50 years. In fact, it wasn't until computers were widely available that it became feasible for forecasting competitions to be held at all.

Time series forecasting competitions have been a feature of the *International Journal of Forecasting* and the *Journal of Forecasting* since the journals were founded in the early 1980s. This strong emphasis on large scale empirical evaluations of forecasting methods, and the need to compare newly proposed methods against existing state-of-the-art methods, has played a large part in pushing researchers to develop new methods that can be shown to work in practice (Fildes & Ord 2002).

Researchers new to forecasting are often surprised to learn how controversial such competitions were when they were first conducted about 50 years ago. I review this controversy in section 1. The influential series of Makridakis competitions are discussed in section 2, and other forecasting competitions are described in section 3. Finally, I provide a few comments on the future of forecasting competitions, and research about forecasting competitions, in section 4. I do not cover forecasting competitions that are not based around time series data.

1 Early controversy

The earliest forecasting competitions were between methods rather than people. It was not feasible, given the communication tools available at the time, to conduct a large-scale forecasting competition involving many entrants spread around the world. So the first few competitions were by individual researchers comparing the accuracy of several methods applied to multiple time series. I only include the first two of these. From 1980 onwards, my scope is restricted to competitions involving multiple entrants.

Nottingham studies

The earliest non-trivial study of time series forecast accuracy was probably by David Reid as part of his PhD at the University of Nottingham (Reid 1969). Building on his work, Paul Newbold and Clive Granger conducted a study of forecast accuracy involving 106 time series (Newbold &

Granger 1974). Although they did not invite others to participate, they did start the discussion on what forecasting methods are the most accurate for different types of time series. They presented the ideas to the Royal Statistical Society, and the subsequent discussion reveals some of the erroneous thinking of the time.

One important feature of the results was the empirical demonstration that forecast combinations improve accuracy. A similar result had been demonstrated as far back as Francis Galton in 1907 (Wallis 2014), yet one discussant (GJA Stern) stated

> "The combined forecasting methods seem to me to be non-starters ... Is a combined method not in danger of falling between two stools?"

Maurice Priestley, later to become the founding and long-serving Editor-in-Chief of the *Journal of Time Series Analysis*, said

> "The authors' suggestion about combining different forecasts is an interesting one, but its validity would seem to depend on the assumption that the model used in the Box-Jenkins approach is inadequate—for otherwise, the Box-Jenkins forecast alone would be optimal."

This reveals a view commonly held (even today) that there is some single model that describes the data generating process, and that the job of a forecaster is to find it. This seems patently absurd to me — real data comes from much more complicated, non-linear, non-stationary processes than any model we might dream up — and George Box himself famously dismissed it saying "All models are wrong but some are useful".

There was also a strong bias against automatic forecasting procedures. For example, Gwilym Jenkins said

> "The fact remains that model building is best done by the human brain and is inevitably an iterative process."

Perhaps Jenkins was reflecting the widely held view that the type of intuitive thinking and extensive experience typically involved in model building cannot be represented by an algorithm or mathematical model. Subsequent history has shown that to be untrue provided enough data are available, and the model is flexible enough to capture the variation seen in real data.

Of course human judgment still has value in forecasting, as demonstrated by Petropoulos et al. (2018) who show that combining judgment with statistical models can lead to statistically significant improvements in forecast accuracy.

2 The Makridakis competitions

Makridakis & Hibon (1979)

Five years later, Spyros Makridakis and Michèle Hibon put together a collection of 111 time series and compared many more forecasting methods. They also presented the results to the Royal Statistical Society. The resulting paper (Makridakis & Hibon 1979) seems to have caused quite a stir, and the discussion published along with the paper is entertaining, and at times somewhat shocking.

Maurice Priestley was in attendance again and was clinging to the view that there was a true model waiting to be discovered:

> "The performance of any particular technique when applied to a particular series depends essentially on (a) the model which the series obeys; (b) our ability to identify and fit this model correctly and (c) the criterion chosen to measure the forecasting accuracy."

Makridakis and Hibon replied

> "There is a fact that Professor Priestley must accept: empirical evidence is in *disagreement* with his theoretical arguments."

Many of the discussants seem to have been enamoured with ARIMA models.

> "It is amazing to me, however, that after all this exercise in identifying models, transforming and so on, that the autoregressive moving averages come out so badly. I wonder whether it might be partly due to the authors not using the backwards forecasting approach to obtain the initial errors." — *W.G. Gilchrist*

> "I find it hard to believe that Box-Jenkins, if properly applied, can actually be worse than so many of the simple methods." — *Chris Chatfield*

At times, the discussion degenerated to questioning the competency of the authors:

> "Why do empirical studies sometimes give different answers? It may depend on the selected sample of time series, but I suspect it is more likely to depend on the skill of the analyst ... these authors are more at home with simple procedures than with Box-Jenkins." — *Chris Chatfield*

Again, Makridakis & Hibon responded:

> "Dr Chatfield expresses some personal views about the first author ... It might be useful for Dr Chatfield to read some of the psychological literature quoted in the main paper, and he can then learn a little more about biases and how they affect prior probabilities."

M-competition

In response to the hostility and charge of incompetence, Makridakis & Hibon followed up with a new competition involving 1001 series. This time anyone could submit forecasts, making this the first true forecasting competition (where multiple people could submit entries) as far as I am aware. They also used multiple forecast measures to determine the most accurate method.

The 1001 time series were taken from demography, industry and economics, and ranged in length between 9 and 132 observations. All the data were either non-seasonal (e.g., annual), quarterly or monthly. Curiously, all the data were positive, which made it possible to compute mean absolute percentage errors, but was not really reflective of the population of real data.

The results of their 1979 paper were largely confirmed. The four main findings (taken from Fildes et al. 1998) were:

1. Statistically sophisticated or complex methods do not typically produce more accurate forecasts than simpler ones.
2. The ranking of the performance of the various methods varies according to the accuracy measure being used.

3. The accuracy of the combination of various methods outperforms, on average, the individual methods being combined, and does well in comparison with other methods
4. The performance of the various methods depends on the length of the forecasting horizon.

Remarkably, the best performing method overall was "DSES", which used a classical multiplicative decomposition (Hyndman & Athanasopoulos 2018) with simple exponential smoothing used to forecast the seasonally adjusted data, and a seasonal naive method used to forecast the seasonal component. The two forecasts were then combined. This extremely simple, and somewhat ad hoc approach, out-performed the best that experienced academic researchers could produce.

The paper describing the competition (Makridakis et al. 1982) had a profound effect on forecasting research. It caused researchers to:

- focus attention on what models produced good forecasts, rather than on the mathematical properties of those models;
- consider how to automate forecasting methods;
- be aware of the dangers of over-fitting;
- treat forecasting as a different problem from time series analysis.

These now seem like common-sense to forecasters, but they were revolutionary ideas in 1982. Even today, I often have to explain to other academics why forecasting is not just an application of time series analysis.

M2-competition

A few years later, a second competition was run (Makridakis et al. 1993) and used only 29 series, but with much richer contextual information, and run in real-time. Given the small sample size, and the use of additional information, few general conclusions about time series forecasting methods could be drawn.

M3-competition

In 1998, Makridakis & Hibon ran their third competition, intending to take account of new methods developed since their first competition nearly two decades earlier. They wrote

> "The M3-Competition is a final attempt by the authors to settle the accuracy issue of various time series methods… The extension involves the inclusion of more methods/researchers (in particular in the areas of neural networks and expert systems) and more series."

It is brave of any academic to claim that their work is "a final attempt"!

This competition involved 3003 time series, all taken from business, demography, finance and economics, and ranging in length between 14 and 126 observations. Again, the data were all either non-seasonal (e.g., annual), quarterly or monthly, and all were positive. Twenty-four entries were received (some from the organizers).

In the published results, Makridakis & Hibon (2000) claimed that the M3 competition upheld the findings of their earlier work, yet the results did not provide the evidence supporting the first finding (that simple methods outperform more complicated methods). The best two methods were not obviously "simple", and the Box-Jenkins' ARIMA models did much better than in the previous competitions.

The top performing entry was the "Theta" method which was described by Assimakopoulos & Nikolopoulos (2000) in a highly complicated and confusing manner. Later, Hyndman & Billah (2003) showed that the Theta method was equivalent to an average of a linear regression and simple exponential smoothing with drift. So it turned out to be relatively simple after all, but Makridakis & Hibon could not have known that in 2000.

The other method that performed extremely well in the M3 competition was the commercial software package ForecastPro. The algorithm used is not public, but enough information has been revealed that we can be sure it is not simple. The algorithm selects between an exponential smoothing and ARIMA model based on some state space approximations and a BIC calculation (Goodrich 2000).

The ForecastPro team also submitted an entry using an automatic algorithm to select an ARIMA model (labelled `BJ-automatic` in the competition), and it did much better than in any previous competitions, and better than the Holt-Winters' method. It seems that the tendency to over-fit ARIMA models had been addressed in the 20 years since the first competition (ForecastPro uses the BIC to penalize over-parametrized models).

Even after more than 20 years of forecasting competitions, the M3 competition was still generating controversy. One issue was around the statistical significance of the results (Stekler 2001), and even whether it made sense to do inference on the results when there is no well-defined population of possible time series. The M3 data constitute a convenience sample, and even if it can be established that method A produces statistically significantly better forecasts than method B, it is not clear what population of time series that conclusion applies to. This issue is explored in Spiliotis et al. (2018), building on the work of Kang, Hyndman & Smith-Miles (2017).

A second area of concern was the potential to cheat. In particular, there was partial revelation of the nature of the test sets part way through the competition (Goodrich 2001). This allowed competitors to adjust their methods on the basis of the test set. While allowing a sequence of entrants with feedback on performance is now a standard and valuable feature of many prediction competitions (Athanasopoulos & Hyndman 2011), it must be done carefully to avoid tailoring methods to fit the test set.

There were also many calls for extensions to the competition including

- evaluating prediction intervals (Goodrich 2001);
- including higher frequency data such as weekly and daily data (Goodrich 2001);
- evaluating multivariate forecasting models (Granger 2001);
- evaluating the reasons behind the differences between methods (Hyndman 2001), and
- identifying the circumstances where substantial improvements are achievable (Fildes 2001).

Finally, an issue that has arisen since the competition is the reproducibility of the results. Although the test data and the submitted forecasts are all publicly available, the computed accuracy scores do not match those in the published paper (Hyndman 2015). This is part of a broader reproducibility problem in statistics (Peng 2011) and forecasting in particular (Evanschitzky & Armstrong 2010; Boylan et al. 2015).

M4-competition

The most recent competition in this series organized by Spyros Makridakis has been the M4 competition[1] comprising 100,000 time series. Compared to the M and M3 competitions, this one involved a few new features:

- Weekly, daily and hourly data were included, along with annual, quarterly and monthly data.
- Participants were invited to submit prediction intervals as well as point forecasts.
- There was a strong emphasis on reproducibility, and competitors were required to post their code on Github.

Extensive discussion of the results of this competition are available in this issue of the *International Journal of Forecasting*, and so will not be repeated here.

3 Other competitions

Sante Fe competitions

Parallel to the series of Makridakis competitions, mathematicians and physicists were also interested in forecasting, and ran their own competition at the Santa Fe Institute, with entries closing in January 1992 (Gershenfeld & Weigend 1993). Only six time series were included, but these were much longer than those discussed above, ranging in length from 1000 observations to 34000 observations. Three of the series were from the physical or medical sciences, one was musical, and one was numerically simulated. Only one series (currency exchange data) was from the same domains as those used in other competitions. The choice of series led to a much greater focus on computationally intensive nonlinear algorithms. With such a small sample, no general conclusions were possible.

KDD cup

Since 1997, the data mining community have held an annual competition known as the KDD cup[2] organized by the Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD). It has generated attention on a wide range of prediction problems and associated methods; occasionally the competition has involved time series forecasting. For example, KDD Cup 2018[3] was a time series competition, where participants were asked to predict air pollution levels for two cities, Beijing, China, and London, UK. The competition ran for 31 days, and forecasts were to be made on each day, for the following 48 hours.

Neural network competitions

There was only one submission that used neural networks in the M3 competition, but it did relatively poorly. To encourage additional submissions, Sven Crone organized a subsequent competition (the NN3) in 2006 involving 111 of the monthly M3 series. Over 60 algorithms were submitted, although none outperformed the original M3 contestants. The paper describing the competition results (Crone, Hibon & Nikolopoulos 2011) was not published until 2011.

This supports the consensus in the forecasting community, that neural networks (and other highly non-linear and nonparametric methods) are not well suited to time series forecasting due

[1] https://www.m4.unic.ac.cy/
[2] http://kdd.org/kdd-cup
[3] https://www.kdd.org/kdd2018/kdd-cup

to the relatively short nature of most time series. The longest series in this competition was only 126 observations long. That is simply not enough data to fit a good neural network model.

There were some follow-up competitions[4], including the NN5 and NNGC1 competitions, but none of the results have ever been published.

Kaggle time series competitions

Kaggle was the first online platform dedicated to data mining competitions, established in 2010 by Australian economist Anthony Goldbloom. Most Kaggle competitions have involved cross-sectional prediction or classification, although a few have involved time series forecasting.

One of the earliest Kaggle competitions was on tourism forecasting, organized by George Athanasopoulos and me. It involved forecasting 793 monthly, quarterly and annual time series, all associated with tourism. The best methods were described in papers published by the *International Journal of Forecasting* (Athanasopoulos et al. 2011). Part of the competition (not open to outside competitors) explored the advantage of using explanatory variables (such as CPI and prices) over purely time series models. Surprisingly, the purely time series forecasts were better than the models that used explanatory variables, even ex post (with perfect knowledge of the future explanatory variables). Most likely, this was due to the relationships between the response and predictor variables changing over time.

In 2017, Oren Anava and Vitaly Kuznetsov organized a Web traffic[5] competition on Kaggle. Here the task was to forecast future web traffic for approximately 145,000 Wikipedia articles. The winning method used a recurrent neural network, trained on the entire data set, with autoregressive predictors, along with some meta-data features such as country and agent used, along with time series features such as autocorrelations. It is interesting to consider why this worked, whereas neural networks failed in the M3 and NN3 competitions. One likely factor was that the data were relatively homogeneous, and the network could be trained across all series. In both the M3 and NN3 competitions, different networks were used for each of the relatively short time series.

One of the great benefits of the Kaggle platform (and others like it) is that it provides a leaderboard and allows multiple submissions. This has been found to lead to much better results as teams compete against each other over the duration of the competition (Athanasopoulos & Hyndman 2011). Of course, the danger is that the forecasts will over-fit the test data, so several precautions are taken to limit that possibility. First, only a few submissions per day are allowed, making it difficult for teams to gain enough information to fit the test data too closely. Second, the accuracy statistics available to contestants during the competition are computed on a small subset of the actual test data, with the final accuracy computed on the remaining part of the test set only available after the competition is completed.

Global Energy Forecasting Competitions

The GEFCom series of competitions[6] were organized by Tao Hong in conjunction with the *International Journal of Forecasting*, and comprised three competitions, each involving several parts. They have been influential in focusing research attention on important issues in energy forecasting, and in providing benchmark methods and data on which to compare new methods.

[4] http://www.neural-forecasting-competition.com/
[5] https://www.kaggle.com/c/web-traffic-time-series-forecasting
[6] http://gefcom.org

A feature of the competitions has been that winning entries must post complete code implementing their methods, and submit an article to the *International Journal of Forecasting* describing their approach.

The 2012 competition (Hong, Pinson & Fan 2014) had five main aims:

1. improve the forecasting practices of the utility industry;
2. bring together state-of-the-art techniques for energy forecasting;
3. bridge the gap between academic research and industry practice;
4. promote analytics in power and energy education;
5. prepare the industry to overcome the forecasting challenges brought by the smart grid technologies and renewable integration needs.

It comprised two tracks: hierarchical load and wind power, where contestants were asked to forecast sections of missing data and (in the case of load) one week of future data. So for the most part it was not truly forecasting, and it looked at point prediction accuracy only.

The 2014 competition (Hong et al. 2016) included four tracks: load forecasting, price forecasting, wind forecasting and solar forecasting, and was the first competition to require entrants to submit complete probability distributions rather than simply point forecasts. These were submitted in the form of 99 percentiles for each forecast period. The competition used a rolling forecast process, mimicking what happens in most real forecasting tasks, where forecasts needed to be made each week for 15 consecutive weeks. In total, 581 contestants participated from 61 countries. Probabilistic forecasts were assessed using quantile scoring.

The 2017 competition (Hong, Xie & Black 2019) involved (1) hierarchical probabilistic forecasting of hourly electricity load for ISO New England, with zones and total load forecasts required; and (2) probabilistic load forecasting of 183 delivery point meters of a US utility. This time only the deciles for each forecast distribution were submitted, and again quantile scoring was used for evaluation. In total, there were 177 entrants. Most of the resulting papers are yet to be published.

4 Some thoughts on the future

There is no doubt that forecasting competitions have played an important role in advancing knowledge of what forecasting methods work. There has been much less research done on the conditions under which different methods work well. With the publication of the large data base of time series and forecasts associated with the M4 competition, researchers now have an opportunity to explore this latter issue in considerable detail. The data and the forecasts from the M4 competition are available in the R package *M4comp2018* (Montero-Manso, Netto & Talagala 2018).

Data and forecasts from the M3 competition are available in the *MComp* package (Hyndman et al. 2018), which also includes data (but not forecasts) from the M-competition. The *Tcomp* package contains data from the Kaggle tourism competition (Ellis 2018), and data from the NN3, NN5, NNGC1 and GEFCom2012 competitions are provided in the *tscompdata* package (Hyndman 2018). It would also be helpful for other forecasting competition data (and the submitted forecasts) to be made publicly available in the interests of promoting research around the efficacy of different forecasting methods. For example, what are the current limits of forecast accuracy in different application areas, and how have these varied over time? Given access

to all the forecasting entries in a competition, can the winning approach be beaten using a combination of entries? We can begin to explore these questions using the M4 data, but it would be better for all forecasting competitions to provide such data as a public good.

A nice side-effect of some competitions is that they create a benchmark data set with well-tested benchmark methods. This has worked well for the M3 data, for example, and for many years new time series forecasting algorithms have been compared to these published results. With the publication of the M4 data, there is now a useful and up-to-date set of benchmarks against which to test new forecasting methods. It could reasonably be argued that any research paper proposing a new general time series forecasting method should not be published unless it can be shown to perform at least as well as existing methods on some well-defined subset of the M4 data.

However, over-study of a single benchmark data set means that methods will eventually over-fit the published test data. I suspect this has happened with the M3 data over the past 20 years, and it is likely to happen with the M4 data, despite its much larger size. Therefore, a wider range of benchmarks is desirable, and these need to be updated regularly. Consequently, there can never be a "final forecasting competition".

One feature of the Makridakis competitions and the GEFCom competitions that has meant they have had a substantial impact on forecasting practice is their association with the *International Journal of Forecasting*. Competitions that have been subject to careful academic scrutiny, where the data and results have been the focus of subsequent research papers, have rightfully led to changes in forecasting practice and have set benchmarks for future forecasting methodological developments.

In surveying the last 40 years of forecasting competitions, it is apparent that the objective for each competition was often unclear, perhaps even to the organizers. This has led, for example, to confusion over the domain to which the results apply — if a method does well in one competition, what can be said about how well it will do on data from a specific company? It would be better if future competitions were clear about the domain to which they apply (by carefully defining the population of time series from which the sample is drawn).

The frequent use of percentage errors also makes the results somewhat difficult to apply in general. It is known that minimizing absolute percentage errors tends to support severe under-forecasts (Gneiting 2011). It would be better if future competitions used objective criteria that are based on well-recognized attributes of the forecast distribution. For example, minimizing MASE (Hyndman & Koehler 2006) leads to estimates of the median of the forecast distribution, as it is equivalent to minimzing the mean absolute error.

There are many features of time series forecasting that have not been studied under competition conditions, and future competitions will surely seek to address these.

For example, few time series competitions have explored forecast distribution accuracy (as distinct from point forecast accuracy). The M4 competition is the first general competition to make a start in this direction with prediction interval accuracy being measured, but it is much richer to measure the whole forecast distribution. The GEFCom2014 and GEFCom2017 competitions involved forecast distributions in the context of energy forecasting. It would be preferable to see the forecast distributions being used in all future forecasting competitions.

New competitions will also need to select accuracy metrics that are appropriate to the forecasting task, which are easy to explain and understand, and which are independent of the scale of the data. MAPEs, and related criteria, are clearly inappropriate when the data contain zeros, or are on an interval rather than a ratio scale. Hyndman & Koehler (2006) introduced the MASE to address this issue, but it is only relevant for point forecasting. For interval forecasting, Winkler scores (Winkler 1972) have been widely used, but are not scale-free. The scaled version of Winkler scores used to assess interval accuracy in the M4 competition seems rather ad hoc and its properties are unknown. In any case, Askanazi et al. (2018) argue that interval forecasts comparisons are problematic in several ways and should be abandoned for density forecasts. Probabilistic forecasts such as densities can be evaluated using proper scoring rules, and scale-free rules such as log density scores are available. But as these are foreign to many practising forecasts, there is a need for textbooks to introduce clear and accessible explanations of how to compute them and what they tell us.

No competition has involved large-scale multivariate time series forecasting. While many of the time series in the competitions are probably related to each other, this information has not been provided. Again, the GEFCom competitions have been ground-breaking in this respect also, by requiring true multivariate forecasts to be provided for the energy demand in different regions of the US, but the winning entries do not appear to have exploited the cross-correlations between regions. Some further thoughts on the design of multivariate forecasting competitions are provided by Fildes & Ord (2002).

I know of no large-scale forecasting competition for finance data (e.g., stock prices or returns), yet this would seem to be of great potential interest.

The frequency of the time series studied in competitions has changed from only annual, quarterly and monthly data in the competitions up to the M3 competition, to more frequent observations such as hourly, daily, and weekly in some subsequent competitions. As data are collected more frequently using automatic sensors and scanning devices, the need for higher frequency forecasts is also increasing. Higher frequency data also brings with it some new challenges including multiple seasonal patterns (De Livera, Hyndman & Snyder 2011), irregularly spaced observations, and the need to generate forecasts rapidly.

There has also been little focus on the conditions under which explanatory variables are useful when forecasting. The tourism forecasting competition included explanatory variables which proved to be unhelpful. The GEFCom competitions have all involved weather data as useful predictors for energy forecasting, but clearly these predictors also need forecasting, and over a long forecast horizon it may be better not to use them at all.

The *International Journal of Forecasting* has been at the forefront of research surrounding forecasting competitions since its inception, and the suggestions made here should provide enough ideas to fuel research in this space for many years to come.

References

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

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

Athanasopoulos, G & RJ Hyndman (2011). The value of feedback in forecasting competitions. *International Journal of Forecasting* **27**(3), 845–849.

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

Boylan, JE, P Goodwin, M Mohammadipour & AA Syntetos (2015). Reproducibility in forecasting research. *International Journal of Forecasting* **31**(1), 79–90.

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

De Livera, AM, RJ Hyndman & RD Snyder (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association* **106**(496), 1513–1527.

Ellis, P (2018). *Tcomp: Data from the 2010 Tourism Forecasting Competition*. Version 1.0.1. https://CRAN.R-project.org/package=Tcomp.

Evanschitzky, H & JS Armstrong (2010). Replications of forecasting research. *International Journal of Forecasting* **26**(1), 4–8.

Fildes, R (2001). Beyond forecasting competitions. *International Journal of Forecasting* **17**, 556–560.

Fildes, R, M Hibon, SG Makridakis & N Meade (1998). Generalizing about univariate forecasting methods: further empirical evidence. *International Journal of Forecasting* **14**, 339–358.

Fildes, R & K Ord (2002). "Forecasting competitions: their role in improving forecasting practice and research". In: *A companion to economic forecasting*. Ed. by M Clements & D Hendry. Oxford, Blackwell, pp.322–353.

Gershenfeld, NA & AS Weigend (1993). "The future of time series". In: *Time series prediction: Forecasting the future and understanding the past*. Ed. by AS Weigend & NA Gershenfeld. Reading, Mass., USA: Addison-Wesley, pp.1–70.

Gneiting, T (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**(494), 746–762.

Goodrich, RL (2000). The Forecast Pro methodology. *International Journal of Forecasting* **16**(4), 533–535.

Goodrich, RL (2001). Commercial software in the M3-competition. *International Journal of Forecasting* **17**, 560–565.

Granger, CWJ (2001). Comments on the M3 forecast evaluation and a comparison with a study by Stock and Watson. *International Journal of Forecasting* **17**, 565–567.

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

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

Hong, T, J Xie & J Black (2019). Global Energy Forecasting Competition 2017: Hierarchical Probabilistic Load Forecasting. *International Journal of Forecasting*. To appear.

Hyndman, RJ (2001). It's time to move from 'what' to 'why'. *International Journal of Forecasting* **17**, 567–570.

Hyndman, RJ (2015). *R vs Autobox vs ForecastPro vs ...* https://robjhyndman.com/hyndsight/show-me-the-evidence/ (visited on 21/12/2018).

Hyndman, RJ (2018). *tscompdata: Time series data from various forecasting competitions*. Version 0.0.1. https://github.com/robjhyndman/tscompdata.

Hyndman, RJ, M Akram, C Bergmeir & M O'Hara-Wild (2018). *Mcomp: Data from the M-Competitions*. Version 2.8. https://CRAN.R-project.org/package=Mcomp.

Hyndman, RJ & G Athanasopoulos (2018). *Forecasting: principles and practice*. 2nd edition. Melbourne, Australia: OTexts. OTexts.org/fpp2.

Hyndman, RJ & MB Billah (2003). Unmasking the Theta method. *International journal of forecasting* **19**(2), 287–290.

Hyndman, RJ & AB Koehler (2006). Another look at measures of forecast accuracy. *International journal of forecasting* **22**(4), 679–688.

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

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

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

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

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

Montero-Manso, P, C Netto & T Talagala (2018). *M4comp2018: Data from the M4-Competition*. Version 0.1.0. https://github.com/carlanetto/M4comp2018/.

Newbold, P & CWJ Granger (1974). Experience with forecasting univariate time series and the combination of forecasts (with discussion). *Journal of the Royal Statistical Society. Series A* **137**(2), 131–165.

Peng, RD (2011). Reproducible research in computational science. *Science* **334**(6060), 1226–1227.

Petropoulos, F, N Kourentzes, K Nikolopoulos & E Siemsen (2018). Judgmental selection of forecasting models. *Journal of Operations Management* **60**, 34–46.

Raphals, L (2013). *Divination and prediction in early China and ancient Greece*. Cambridge, UK: Cambridge University Press.

Reid, DJ (1969). "A comparative study of time series prediction techniques on economic data". PhD thesis. Nottingham, UK: University of Nottingham.

Spiliotis, E, A Kouloumos, V Assimakopoulos & S Makridakis (2018). *Are forecasting competitions data representative of the reality?* Tech. rep. 3/18. Forecasting and Strategy Unit, National Technical University of Athens.

Stekler, H (2001). The M3-competition: the need for formal statistical tests. *International Journal of Forecasting* **17**, 576–577.

Wallis, KF (2014). Revisiting Francis Galton's forecasting competition. *Statistical Science* **29**(3), 420–424.

Winkler, RL (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association* **67**(337), 187–191.