MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

A brief history of forecasting competitions

Rob J Hyndman

December 2018

Working Paper no/19

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

14 December 2018

JEL classification: C22,C45,C52,C53

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 future competitions.

Keywords: blah, blah

Prediction competitions go back millenia, with rival Greek diviners competing to predict the future more accurately (Raphals 2013, p124). However, for general time series forecasting, the history is much more limited, and only goes back about 50 years.

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.

1 Early controversy

Young researchers in forecasting are often surprised to learn how controversial such competitions were when they were first conducted about 50 years ago.

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?"

[1] An early version of this article appeared as a blog post at https://robjhyndman.com/hyndsight/forecasting-competitions/.

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."

Subsequent history has shown that to be untrue provided enough data is available.

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 insults:

> "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 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 possibly 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 Makridakis & Hibon 2000) were:

1. Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones.
2. The relative ranking of the performance of the various methods varies according to the accuracy measure being used.
3. The accuracy when various methods are being combined outperforms, on average, the individual methods being combined and does very well in comparison to other methods.
4. The accuracy of the various methods depends upon the length of the forecasting horizon involved.

Remarkably, the best performing method overall 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.

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.

M3-competition

In 1998, Makridakis & Hibon ran their third competition (the second was not strictly time series forecasting), 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.

In the published results, (Makridakis & Hibon 2000) claimed that the M3 competition supported the findings of their earlier work. Yet the best two methods were not obviously "simple".

One was the "Theta" method which was described 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).

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[1]) was organized 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 general consensus in forecasting, that neural networks (and other highly non-linear and nonparametric methods) are not well suited to time series forecasting due 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[2], but as far as I know none of the results have ever been published.

[1] http://www.neural-forecasting-competition.com/NN3
[2] http://www.neural-forecasting-competition.com/

Kaggle time series competitions

Few Kaggle competitions[3] have involved time series forecasting; mostly they are about cross-sectional prediction or classification. However, there have been some notable exceptions.

- George Athanasopoulos and I organized a Tourism forecasting competition in 2010. There was a follow-up part 2 later in the same year. The best methods were described in papers published by the IJF in 2011.
- Recently, Oren Anava and Vitaly Kuznetsov organized a Web traffic competition. Here the task was to forecast future web traffic for approximately 145,000 Wikipedia articles. A paper describing the best methods is currently in progress.

One of the great benefits of the Kaggle platform (and others like it) is that it provides a leader-board 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. George Athanasopoulos and I discussed this important feature in a 2011 IJF paper.

M4-competition

Makridakis is now at it again with the M4 competition. This time there are 100,000 time series, and many more participants. New features of this competition are:

- Weekly, daily and hourly data are included, along with annual, quarterly and monthly data.
- Participants are invited to submit prediction intervals as well as point forecasts.
- There is a strong emphasis on reproducibility (a problem with earlier competitions), and competitors will be required to post their code on Github.

2 Future competitions?

The M4 competition is certainly not the end of time series competitions! There are many features of time series forecasting that have not been studied under competition conditions.

No previous time series competition has explored forecast distribution accuracy (as distinct from point forecast accuracy). The M4 competition is the first to make a start in this diretion with prediction interval accuracy being measured, but it is much richer to measure the whole forecast distribution. This was done, for example, in the GEFCom2014 and GEFCom2017 competitions for energy demand forecasting.

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.

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 interest judging by the number of submissions to the IJF I receive every week.

3 R packages

The data from many of these competitions are available as R packages.

[3] https://www.kaggle.com/competitions

- Mcomp: Data from the M-competition and M3-competition.
- M4comp2018: Data from the M4-competition.
- Tcomp: Data from the Kaggle tourism competition.
- tscompdata: Data from the NN3 and NN5 competitions.

3.1 Further reading

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

The data mining community have the annual KDD cup which has generated attention on a wide range of prediction problems and associated methods. Recent KDD cups are hosted on kaggle.

In my research group meeting today, we discussed our (limited) experiences in competing in some Kaggle competitions, and we reviewed the following two papers which describe two prediction competitions:

1. Athanasopoulos and Hyndman (IJF 2011). The value of feedback in forecasting competitions. [preprint version]
2. Roy et al (2013). The Microsoft Academic Search Dataset and KDD Cup 2013.

   Some points of discussion:

- The old style of competition where participants make a single submission and the results are compiled by the organizers is much less effective than competitions involving feedback and a leaderboard (such as those hosted on kaggle). The feedback seems to encourage participants to do better, and the results often improve substantially during the competition.
- Too many submissions results in over-fitting to the test data. Therefore the final scores need to be based on a different test data set than the data used to score the submissions during the competition. Kaggle does not do this, although they partially address the problem by computing the leaderboard scores on a subset of the final test set.
- The metric used in the competition is important, and this is sometimes not thought through carefully enough by competition organizers.
- There are several competition platforms available now including Kaggle, CrowdAnalytix and Tunedit.
- The best competitions are focused on specific domains and problems. For example, the GEFcom 2014 competitions are about specific problems in energy forecasting.
- Competitions are great for advancing knowledge of what works, but they do not lead to data scientists being well paid as many people compete but few are rewarded.
- The IJF likes to publish papers from winners of prediction competitions because of the extensive empirical evaluation provided by the competition. However, a condition of publication is that the code and methods are fully revealed, and winners are not always happy to comply.

- The IJF will only publish competition results if they present new information about prediction methods, or tackle new prediction problems, or measure predictive accuracy in new ways. Just running another competition like the previous ones is not enough. It still has to involve genuine research results.
- I would love to see some serious research about prediction competitions, but that would probably require a company like kaggle to make their data public. See Frank Diebold's comments on this too.
- 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 new time series forecasting algorithms can be easily tested against these published results. However, over-study of a single benchmark data set means that methods are probably over-fitting to the published test data. Therefore, a wider range of benchmarks is desirable.
- Prediction competitions are a fun way to hone your skills in forecasting and prediction, and every student in this field is encouraged to compete in a few competitions. I can guarantee you will learn a great deal about the challenges of predicting real data — something you don't always learn in classes or via textbooks.

References

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.

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

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.

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.

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.

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.

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