Learning from Forecasting Competitions

The objectives of the M4 (following on from the previous M-competitions) were to "learn how to improve the forecasting accuracy" and investigate "how such learning can be applied to advance the theory and practice of forecasting". How successful have previous competitions been and can we expect anything different from the latest one? The early forecasting competitions, starting with Newbold and Granger (1974), provided a shock to the time series and statistical forecasting communities. As Hyndman (this issue) points out, commentators at the time were sceptical as to their value, with some even arguing that competitions were somehow inappropriate. However, the shock was very much temporary: by 1995, there had been almost no influence on the theoretical econometrics and time series research communities (Fildes & Makridakis, 1995), and an examination of more recent citations of the various competitions confirms that they continue to be neglected outside the forecasting community.[1] What then might we expect from the release of the M4 material, and what might we have hoped for?

There are various methodological innovations that I welcome (and I doubt I will see any dissent): the public availability of the code, the availability of the results on the web, the extension of the evaluation to include prediction intervals, and the open invitation to participate. The size of the data base has been expanded, though the purpose of this is unclear; more positive is the increase in the range of series types to include daily and hourly data. I note here that the integrity of the daily data sources is questionable (Ingel et al., this issue). The authors' attempt to include more methods, and in particular various machine learning methods, is also a key justification for another competition. Less satisfactory is the choice of error measures (see Goodwin, this issue): the results in previous competitions have always turned out to be dependent on the error measures (and the forecast horizon).

Here, I focus on the question of how the results should influence researchers and practitioners, as well as on points that are potentially controversial (at least to me).

1. M3 and M4 provide "a good representation of the forecasting reality in the business world". Two different views are prevalent: first, that a group of forecasting methods have consistent performance characteristics across business and economic time series (when the number of series used in the evaluation is large enough, as in M3 and M4). An alternative view is that each organization has its own customers and processes, and therefore its data may differ very substantially from the common population of time series, with the consequence that models' performances on their own series might well differ substantially. If the first claim is not true, an organization can best regard the results of unfocused competitions as an indication of the range of

[1] A referee queried why we would expect responses from other research communities. Predictive validation is at the heart of the scientific method, and has been proposed more recently as a core methodology for data science (Donoho, 2017). Thus, the use of appropriate benchmarks as suggested by the results of the various competitions is a requirement of good science.

methods to consider. This is not a new concern, and has been explored to a certain extent by Ord (2001) and other commentators. Most recently, Spiliotis, Kouloumos, Assimakopoulos, and Makridakis (2019) considered this question in detail as part of an analysis of the M4 Competition data. In apparent contradiction to Makridakis, Spiliotis, and Assimakopoulosthey conclude that, while M3 and M4 are similar in terms of their descriptive characteristics, they do differ from other standard data sets in several important respects, with such differences potentially explaining any differences in performance. As an illustration of how extreme the differences can be, I use daily data from Ma and Fildes (2018).[2] The short-term ETS sMAPE in the M4 data set is 1.665% from 1476 micro-daily series, while the corresponding figure in the Ma/ Fildes retail data set is 11.1% from 2000 series: more particularly, the improvement over a seasonal naïve model using ETS with the Ma/ Fildes data set is around 27% for the 1-day horizon, compared to 15% in M4: ARIMA proved the best performing univariate time method relative to ETS and Theta for both data sets. Thus, it would be inappropriate to interpret the particular retail forecasting problem facing Ma and Fildes in the light of the benchmark relative accuracies from M4, a point that Kolassa (2008) has argued vigorously; instead, each organization needs to organize its own forecasting competition for its own forecasting problems, and should not rely on even large benchmark data sets. What M4 can do is to guide us towards the methods that are worth including in the pool of methods to be considered, revising established opinion (Fildes & Lusk, 1984); these new results in M4 should certainly influence the short-list of methods, and the inclusion of two ML-based methods in the set of most accurate methods is exciting. This issue is important, not least because it affects the core methods included in commercial software.

2. ML methods generally will not be more accurate than statistical ones (hypothesis 4 of Makridakis et al., 2018; and finding 7 of Makridakis, Spiliotis, & Assimakopoulos, 2019).

Two issues are important in interpreting the findings reported by Makridakis et al.: (i) the way in which the methods were implemented, since the performances of ML methods are much affected by the pre-processing of the data, with different choices affecting the results differently; and (ii) their automatic application to the full sample of time series (admittedly part of the M4 design). Expanding on this second point, ML methods have been shown to work on specific data sets; for example, Ma and Fildes (2018) showed that the careful application of a machine learning method produces substantially better forecasts than standard benchmarks, with an improvement of over 10% relative to the univariate extrapolative benchmarks. In fact, an ML method (Smyl, this issue) produces the best forecasts, and this method also proves the most accurate over all M-Competitions. An analysis of Amazon data has demonstrated the effectiveness of ML methods on intermittent data. Further validation of the ML results using established benchmark ML procedures (a problem that I acknowledge) is necessary, but for now, the notion~~,~~ that off-the-shelf ML

Commented [CM1]: Not in reference list??

Commented [FR2R1]: I don't have the details

[2] Thanks are due to Shaohui Ma, who performed additional calculations to ensure a fair comparison with the M4 methods.

methods will *automatically* produce improvements in accuracy, which has been propagated by some software suppliers, must be regarded as fanciful.

3. Combining forecasts improves the accuracy (finding 1, Makridakis et al., 2019).

We have long known of the benefits of combining. It continues to generate both empirical (as here) and theoretical (Diebold & Shin, 2019~~8~~) interest. Nevertheless, the strong claims made here for combining, namely that individual selection (i.e., associating a particular method with a series or a sub-set of series) should not be attempted, is an incorrect interpretation of the evidence. The comparisons made in the various tables are between combining and aggregate selection (i.e., the same method applied to all series or to the particular subsets of annual, quarterly etc.). Fildes and Petropoulos (2015) demonstrated on the M3 monthly data that there are some situations, depending on the series characteristics such as predictability, seasonality, trend etc. (measured on the validation data set), that can be used to discriminate between the cases where combining is valuable and those in which selection works better. Spiliotis et al. (2019) explore this point further. A more recent perspective is to see the two alternatives as falling on a spectrum (Kourentzes, Barrow & Petropoulos, 2019), where a single model, all models or a shortlist of the most promising models are selected from a pool of candidate models and combined with equal or unequal weights. The method of Montero-Manso (this issue) effectively takes such an approach, switching between the two alternatives.

Commented [CM3]: Not in reference list??

Commented [FR4R3]: I don't have the details

Does this issue matter? The differences found by Fildes and Petropoulos were small. However, I would expect to see greater benefits from best-practice selection or combination within a restricted set of methods when using data sets with greater levels of accuracy divergence between methods.

Despite the achievements of the earlier competitions, I have become a 'competition sceptic' (Fildes, 2001). On the positive side of the balance sheet, an exploratory analysis with 100K series and 61 forecasting methods is likely to turn up some interesting features; it certainly has here with the ML evidence. In addition, though, Makridakis and colleagues are to be congratulated for improving various aspects of the methodology, making new methods and code accessible, and successfully focusing attention on prediction intervals. The conclusions have also rehabilitated the potential advantages of using more complex methods (finding 5 of Makridakis et al., 2019).

More negatively, though, there remain outstanding issues concerning the research design. Competitions require a lot of commitment from the organizers and a wide range of participants, and, perhaps as a consequence, they often lack clear objectives and research questions. In commenting on the M3 competition, I remarked that "the question researchers now face is to re-examine model selection strategies" that have been shown to be successful for homogeneous data (Fildes, Hibon, Makridakis, & Meade, 1998). Makridakis and colleagues over-state their claim in their conclusions that the results "would allow practitioners to select the most appropriate method(s) for their forecasting needs". As I argued above, the results certainly should guide the short-list, but we have no series statistics that would tell a practitioner whether, for their particular problem, they should (say) use combining, use 'Smyl', or select among Damped, Theta and ARIMA, or even an ML method.

Finally, Makridakis et al. (2019) look forward to further competitions, including a possible focus on the value, if any, of including explanatory/exogenous variables. However, we already know that the inclusion of temperature improves various demand forecasts such as energy and beer for short lead times. Likewise, we know that including a company's own marketing activities improves retail sales forecasts (e.g. Ma, Fildes, & Huang, 2016). As with the M4 experiment, I believe that the research questions in any new competition will need to be refined to determine the circumstances under which conclusions that are already available in the literature apply when set in a realistic forecasting context.

References

Diebold, F.~~rancis~~ X., ~~and ~~& Shin, M~~inchul~~. (2019~~8~~). ~~'~~Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives.~~,~~ *International Journal of Forecasting*, in press.

Donoho, D. (2017). 50 years of data science. *Journal of Computational Graphical Statistics*, *26*, 745-766.

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

Fildes, R., Hibon, M., Makridakis, S., & Meade, N. (1998). Generalising about univariate forecasting methods: Further empirical evidence. *International Journal of Forecasting*, *14*, 339-358.

Fildes, R., & Lusk, E. J. (1984). The choice of a forecasting model. *Omega* ~~-~~– *International Journal of Management Science*, *12*, 427-435.

Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time-series analysis and forecasting. *International Statistical Review*, *63*, 289-308.

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*, 1692-1701.

Goodwin, P. (~~2019~~this issue). Performance measurement in the M4 Competition: possible future research.~~,~~ *International Journal of Forecasting*, ~~(~~this issue.~~)~~

Kolassa, S. (2008). Can we obtain valid benchmarks from published surveys of forecast accuracy?~~.~~ *Foresight*, *11*, 6-14.

Kourentzes, N., Barrow, D., & Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, *209*, 226-235.

Hyndman, R. (~~2019~~this issue). A short history for forecasting competitions. *International Journal of Forecasting*,~~.~~ ~~T~~this issue.~~/~~

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Commented [CM5]: Reference not cited??

Commented [FR6R5]: Now included

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Ingel, A., Shahroudi, N., Kängsepp, M., Tättar, A., Komisarenko, V., & Kull, M. (~~2019~~this issue). Correlated daily time series and forecasting in the M4 competition. *International Journal of Forecasting*, this issue.

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*, *249*, 245-257.

Ma, S., & Fildes, R. (2018). *Customer flow forecasting with third-party mobile payment data*. ~~In.~~ Lancaster: Lancaster University.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting, 34*, 802-808.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V.~~,~~ (2019). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*,~~.~~ ~~T~~this issue.~~/~~

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combintion of forecasts. *Journal of the Royal Statistical Society, Series* ~~(~~*A*~~)~~, *137*, 131-164.

Ord, K. (2001). An introduction, some comments and a scorecard. *International Journal of Forecasting*, *17*, 537-541.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2019). Are forecasting competitions data representative of the reality?~~.~~ *International Journal of Forecasting*.

Formatted: English (United Kingdom)

Formatted: Font: Not Italic

Formatted: Font: Italic

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: (Default) Times New Roman, 11 pt

Formatted: Font: 14 pt

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic