International Journal of Forecasting 36 (2020) 189–190

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

Performance measurement in the M4 Competition: Possible future research

Paul Goodwin

School of Management, University of Bath, United Kingdom

With 100,000 time series and 61 methods producing forecasts for multiple time horizons, the M4 competition has been a gigantic project, and the organizers are to be congratulated for both their ambition and their organisational skills in implementing it successfully. This commentary reflects on the measurement of the performances of the different methods. However, it does so with an emphasis on what researchers might usefully investigate in the future now that this huge data set is available, rather than on an evaluation of the pros and cons of the measures that were used. Such assessments can be found elsewhere (e.g. Davydenko & Fildes, 2013; Goodwin & Lawton, 1999). As the organizers acknowledge, no error measure is perfect, and the presentation of a profusion of different error and calibration measures in the initial reporting of the results would probably have been overwhelming and served only to obscure the key findings of the competition (Makridakis et al. 2020).

Nevertheless, it would clearly be interesting in the future to see how robust the findings are to alternative error measures and to the alternative loss functions that they assume implicitly. For example, the sMAPE that was used in the competition has been shown to be asymmetric in its treatment of positive and negative forecast errors (Goodwin & Lawton, 1999), while Davydenko and Fildes (2013) recommended the AvgRelMAE, which is the geometric mean of ratios of mean absolute errors, as an alternative to the MASE. In addition, an assessment of the consistency of the methods' performances across the different series and forecast horizons would be helpful. Measures of the dispersion of the error metrics would help forecasters to assess whether some methods carry significantly greater risks of poor performances than others, despite producing relatively accurate forecasts on average. It has also been shown that biases in forecasts can be at least as great a concern as inaccuracy in terms of the costs that they incur for decision makers (Sanders & Graman, 2009), so it would be useful to know whether some methods have a greater tendency toward bias than others.

The inclusion of performance measures for prediction intervals in the M4 competition is welcome, given the importance of estimating the uncertainty associated with forecasts (Goodwin, 2014). Of course, the degree to which a forecasting method has predicted two quantiles accurately – with the predictions being represented by the lower and upper bounds of the interval – is only a partial assessment of the extent to which the underlying probability distribution has been represented correctly. In the future, it would be useful to test the abilities of alternative methods to produce well-calibrated density forecasts.

The mean scaled interval score (MSIS) that was used in the competition has a number of merits (Gneiting & Raftery, 2007). As a strictly proper scoring rule, it rewards honest estimates and penalises both overly wide, and hence uninformative, intervals and cases where the intervals fail to capture the outcomes. In the latter case, the extent to which the outcome lies beyond the interval's upper or lower limit is also taken into account. The use of the average coverage difference (ACD) complements the MSIS by comparing the percentage of occasions when intervals capture outcomes to the percentage that would be achieved by perfectly calibrated intervals. However, it is worth noting that there are three non-mutually exclusive reasons why prediction intervals may capture outcomes more or less often than they should, given the stated coverage probability: (i) the interval tends to be too narrow or too wide, (ii) the location of the interval tends to be misplaced, and (iii) the interval tends to misrepresent the shape (e.g. the skewness or kurtosis) of the underlying probability distribution. For example, if the 95% prediction

E-mail address: P.Goodwin@bath.ac.uk.

https://doi.org/10.1016/j.ijforecast.2019.02.015

intervals of a method captured the outcomes only 62% of the time, this does not necessarily imply that the intervals were too narrow. Instead, it might reflect a tendency for the point forecast to be a large distance from the outcome, so that an interval has a width that is appropriate for the underlying probability distribution but is located wrongly. In this case, attempts to widen an interval – a commonly suggested response when miscalibration occurs – would be misplaced. Instead, the focus should be on improving the accuracy of the point forecast.

It would be interesting to explore the extent to which each of these three issues is associated with the less-than-perfect calibration of most of the methods in the competition. This type of analysis can be achieved easily with artificial data where the probability distribution that is used to generate the observations is known. For example, assuming a symmetric distribution, an interval with a given width can be relocated so that it is centred on the mean of the generating distribution. The percentage of observations that it would fail to capture can then be calculated so as to isolate the effect of insufficient or excessive width. Similarly, the width of an interval could be adjusted so that it is correct for the known distribution while leaving its centre unchanged, thus enabling a measure of the extent to which an error in the location of the interval accounts for any miscalibration to be determined. However, it is less easy to see how such an analysis could be carried out with the real data used in the competition. While there are huge advantages to be gained by testing methods on real data, identifying the underlying factors that account for the results is more challenging.

Nevertheless, the competition has already made a very welcome contribution to our knowledge of forecasting methods. The vast amount of data that it has generated will surely provide an invaluable resource for forecasting researchers for many years to come.

Acknowledgment

The author would like to thank Vangelis Spiliotis for providing details of the performance methods used in the competition and the reasons for their choice.

References

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, *29*(3), 510–522.

Gneiting, A. E., & Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Goodwin, P. (2014). Getting real about uncertainty. *Foresight: The International Journal of Applied Forecasting*, *33*, 4–7.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*, 405–408.

Makridakis, S., Spiliotis, E., & Assimalopoulos, V. (2020). The M4 competition: 100, 000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74.

Sanders, N. R., & Graman, G. A. (2009). Quantifying costs of forecast errors: A case study of the warehouse environment. *Omega*, *37*, 116–125.