International Journal of Forecasting 36 (2020) 197–200

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

Why does forecast combination work so well?

Amir F. Atiya

Department of Computer Engineering, Cairo University, Giza, Egypt

A B S T R A C T

Forecast combinations were big winners in the M4 competition. This note reflects on and analyzes the reasons for the success of forecast combination. We illustrate graphically how and in what cases forecast combinations produce good results. We also study the effects of forecast combination on the bias and the variance of the forecast.

1. Introduction

The M4 competition (Makridakis, Spiliotis, and Assimakopoulos, 2018) is the latest in a series of M-competitions (Makridakis, Andersen, Carbone, Fildes, Hibon, et al., 1982; Makridakis & Hibon, 2000) that started several decades ago. These competitions are very beneficial in helping researchers to gain an in-depth knowledge of the performances of different forecasting models, and in drawing up best practices for forecasting. The M4 competition reinforced some of the knowledge that was gained in previous competitions, such as the general superiority of simple forecasting models. However, there were also some new lessons that were learned from this competition. For a start, the top model was a hybrid approach that utilized both machine learning and statistical approaches, and hybrid statistical/nonlinear models may be a promising direction to pursue.

Another finding is that forecast combinations have been very successful. Of the 17 most accurate methods, 12 were combinations of forecasting models. Even though forecast combination is known to be a winning strategy to follow Armstrong (2001) and Clemen (1989), this is the first time it has demonstrated such strong dominance in a competition. In this note we consider the topic of forecast combination, and contemplate the reasons for its success. We present some analysis that backs up the empirical research that has been performed on this topic.

Let $u(m)$, $m = 1, \ldots, N$, be the N forecasts to be combined. The forecast combination can be written as

$$u = \sum_{m=1}^{N} w_m u(m), \tag{1}$$

where w_m is a combination weight. Usually, but not always, researchers use a convex combination, i.e., $0 \leq w_m \leq 1$ and $\sum_{m=1}^{N} w_m = 1$. The motivation behind forecast combination is the fact that forecasting problems typically possess small or rather finite histories of points. Thus, from a practical point of view, it is not possible to obtain the correct specification of the underlying data generation process. It is beneficial to hedge against the resulting inaccuracy of the derived forecasting model by considering several forecasting models and combining their forecasts. Irrespective of any estimation errors, though, there is also another explanation of why forecast combination works well, and this is illustrated graphically in the next section.

2. Graphical illustration

The following illustration explains why and in what situations forecast combinations are superior. Let the vector $y = (y_1, \ldots, y_H)^T$ be the true time series values for the horizon to be forecast (where H is the number of steps ahead to be forecast). Consider that there are five candidate forecasting models, and assume that they produce forecast vectors $u(1), \ldots, u(5)$ for the horizon considered. The components of $u(m)$, i.e., $u_i(m)$, correspond to the steps ahead being forecast. Refer to Fig. 1

E-mail address: amir@alumni.caltech.edu.

https://doi.org/10.1016/j.ijforecast.2019.03.010
0169-2070/

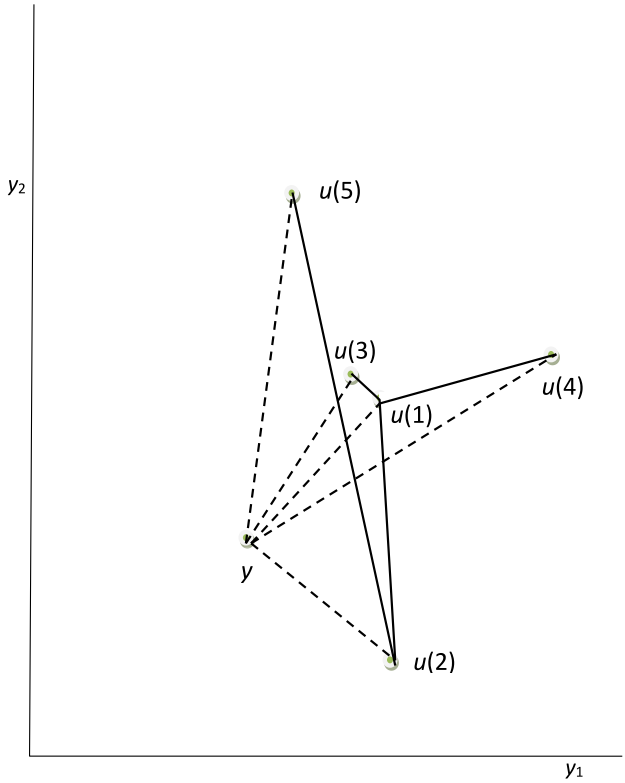

Fig. 1. An illustration of forecast combinations: The solid lines between the points (forecasts) $u(1), \ldots, u(5)$ represent the forecast combinations, while the distances along the dotted lines to the true value y are proportional to the RMSEs of the corresponding forecasts.

for an illustration for the case $H = 2$ (i.e., the y vector and the forecast vectors are 2D). The forecast root mean square error (RMSE) for some forecast $u(m)$ is given by:

$$RMSE = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(u_i(m) - y_i)^2} \tag{2}$$

$$= \frac{1}{\sqrt{H}}\text{distance}(u(m), y). \tag{3}$$

Thus, the distance in the graph is proportional to the RMSE. The forecasts $u(1)$ and $u(2)$ possess diversity, and therefore are located in different parts of the space (see Fig. 1). Moreover, they have comparable distances from y, and hence, comparable RMSEs. The line connecting them (modeled as $wu(1) + (1 - w)u(2)$, where w is a weight between zero and one) corresponds to the location of the combination of these forecasts. One can observe that the distance between y and any point on the line is smaller than the distance between y and either $u(1)$ or $u(2)$. This indicates that the RMSE of the combined forecasts is better than that of either constituent forecast. This is especially pronounced at the midpoint of the line, which corresponds to an equal weight combination.

Let us now look at the combination of the forecasts of points $u(1)$ and $u(3)$ (see Fig. 1). These forecasts are very similar (or rather highly correlated), and comparable in errors. An analysis of the distance between point y and the line between $u(1)$ and $u(3)$ shows that the forecast combination does not add much value; however, it is also not harmful.

Consider now the combination of the vectors $u(1)$ and $u(4)$. One can see that the distance between $u(4)$ and y is larger than that between $u(1)$ and y, indicating that $u(4)$ is an inferior forecast. What adds further harm is the fact that $u(4)$ is not diverse from $u(1)$, but is highly correlated. The line between $u(1)$ and $u(4)$ (corresponding to their forecast combination) possesses a larger distance from y than that of $u(1)$ alone, and hence, a worse performance. Thus, the forecast $u(4)$ is considered a drag on the performance of the forecast combination.

The forecast $u(5)$ is a similarly poor forecast. The distance between this point and y is large. However, it adds value when we combine it with the forecast $u(2)$, and we obtain a better combined forecast. This is because $u(5)$ is diverse from $u(2)$, and this makes up for its inadequate individual forecast accuracy. Note that, for the best result, the combination weight has to favor the good forecast $u(2)$ somewhat.

From these observations we can deduce the following facts, which are also verified in empirical research and in most time series forecasting competitions.

- Forecast combination should be a winning strategy if the constituent forecasts are either diverse or comparable in performance.
- One should exclude forecasts that are considerably worse than the best ones in the pool, unless they are very diverse from the rest. This also agrees with

the recommendations discussed in the literature, see Armstrong (2001) and Timmermann (2006). In addition, Kourentzes, Barrow, and Petropoulos (2019) make the point that one has to be selective in regard to the pool of forecasts to be used for forecast combination. For example, one can include in the combination pool top forecasts that are within a 10% range in performance (such as an SMAPE range from 15% to 16.5%). One may also include forecasts that are 20% or 30% worse (e.g. an SMAPE of 18% to 20%), provided that they possess the necessary diversity.

3. Diversity and forecast combination

Diversity in forecasts can be achieved by relying on variables from different sources. These could operate in different "pathways" in which they affect the variable to be forecast (see De Menezes, Bunn, & Taylor, 2000). For example, one could have one stock price forecasting model based purely on the stock price time series and another based on fundamental company information (e.g. financial statement data, such as sales, earnings, debt, etc.). The two forecasts would be considered diverse because they are based on quantities that have little direct relation. For another example, consider one GDP forecast based on economic indicators, and another based on the interest rate or interest rate changes across the yield curve. These would be fairly diverse, and it would be a good idea to pool the forecasts.

Andrawis, Atiya, and El-Shishiny (2011), Kourentzes and Petropoulos (2016) and Kourentzes, Petropoulos, and Trapero (2014) suggested a novel way of imparting diversity by combining forecasts at different time scales. For example, we could have a monthly forecast, then time-aggregate the data to produce a quarterly and/or yearly forecast. Short-, medium- and long-term dynamics are influenced by different factors, and therefore include different information.

One could also achieve diversity by using different models on the same data. If the models are sufficiently diverse, they can add value even when they are designed on the same data. For example, we could consider the following exponential smoothing models: no trend, additive trend, additive damped trend, multiplicative trend, and multiplicative damped trend. As they are each based on different assumptions, they would be varied enough. This could be helpful if, for example, the time series initially possesses no trend, but a trend commences at some later point. The fact that forecast combination takes "a middle ground" allows it to guard against unexpected changes or misspecifications. Such misspecifications are typically due to a small in-sample set, structural breaks, or parameter drift.

The only pitfall in forecast combination is when the data are rife with outliers, or have a very heavy-tailed distribution. In such cases, one may want to limit the pool of forecasts to be combined to a small number, as a large pool provides more opportunities for an outlier forecast to creep into the pool and ruin the combined forecast. An analysis of the "exotic" world of heavy-tailed distributions, their effects on prediction models, and the numerous paradoxes that they pose can be found in the work of Yousef and Kundu (2014).

4. Forecast combination and the bias–variance decomposition

The bias–variance decomposition is a fundamental concept in all estimation problems, including forecasting, classification, and parameter estimation problems. In this concept, the mean square error can be decomposed into a bias term and a variance term, as follows:

$$MSE = B^2 + V, \tag{4}$$

where MSE, B, and V represent the mean square error, the bias, and the variance, respectively. The bias represents the consistent offset of the forecast, away from the true value. For example, consider a forecasting model with some tuning parameter. If the forecast is always a given amount higher than the true value (even with different realizations of the error terms of the data generation process), then the bias is positive. It is akin to fitting a linear function to a parabola (with added noise). There will always be an offset from the true value, and this is the bias. The variance represents the variation of the forecast around its mean. Thus, if a forecasting model produces highly variable forecasts for different realizations of the error terms, it has a high variance. Simpler models tend to produce large biases and small variances. On the other hand, complex models produce small biases and large variances. An analysis of the roles of the bias and the variance in forecasting can be found in the work of Ben Taieb and Atiya (2016).

As it turns out, forecast combination tends to keep the bias about the same, or possibly improved. On the other hand, it generally decreases the variance considerably. This is shown as follows. Consider y to be the variable to be forecast, and let $u(1), \ldots, u(N)$ be the different forecasts to be combined. Let the forecast combination be

$$u = \sum_{i=1}^{N} w_i u(i), \qquad \text{with } 0 \le w_i \le 1, \quad \sum_{i=1}^{N} w_i = 1. \tag{5}$$

The bias B of the combined forecast is given by

$$B = E(u) - E(y) \tag{6}$$

$$= \sum_{i=1}^{N} w_i E(u(i)) - E(y) \tag{7}$$

$$= \sum_{i=1}^{N} w_i \Big[E(u(i)) - E(y) \Big] \tag{8}$$

$$= \sum_{i=1}^{N} w_i B_i, \tag{9}$$

where B_i is the bias for forecast i, and the expectation is over the variations of the error terms of the data generation process. Thus, the bias of the forecast combination is the weighted average of the individual biases, and if these biases are comparable, the bias of the combination will

be too. On the other hand, Hendry and Clements (2004) make the point that one would not expect to have the biases of the individual forecasts on one side only, so we typically have some bias cancellations, and this will reduce the overall bias. They mentioned the aftermath of a structural break as an exception, with the majority of the biases being all on one side, as most models would "stay behind" in similar ways.

Next, we consider the variance. Let σ_i^2 be the variance of the individual forecast i and ρ_{ij} be the correlation coefficient of forecasts i and j. The variance is given as follows:

$$V = E(u - E(u))^2 \tag{10}$$

$$= E\Big[\sum_{i=1}^{N} w_i(u(i) - \overline{u(i)})\Big]^2 \tag{11}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j E\big[(u(i) - \overline{u(i)})(u(j) - \overline{u(j)})\big] \tag{12}$$

$$= \sum_{i=1}^{N} w_i^2\sigma_i^2 + 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_i w_j \rho_{ij}\sigma_i\sigma_j \tag{13}$$

$$= \Big(\sum_{i=1}^{N} w_i\sigma_i\Big)^2 - 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_i w_j(1 - \rho_{ij})\sigma_i\sigma_j. \tag{14}$$

By the RMS-arithmetic weighted mean inequality, we can write:

$$\sum_{i=1}^{N} w_i\sigma_i \le \sqrt{\sum_{i=1}^{N} w_i\sigma_i^2}, \tag{15}$$

and hence,

$$V \le \sum_{i=1}^{N} w_i\sigma_i^2 - 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_i w_j(1 - \rho_{ij})\sigma_i\sigma_j. \tag{16}$$

The first term is the average of the variances of the individual forecasts. The second term is positive, since $\rho_{ij} \le 1$. One can therefore observe that a substantial positive term (consisting of many more terms) is subtracted from the first term (the average variance), indicating that the variance of the forecast combination tends to decrease considerably. The extent of the decrease in variance becomes larger if the correlation coefficients among the constituent forecasts are smaller, confirming the aforementioned finding that diversity improves the performance of the forecast combination. However, one must concede that the correlation coefficients of the individual forecasts are positive in most cases (for example, they are typically larger than 0.5). The reason for this is that ultimately they are all forecasting the same quantity, so they are moved together by underlying fluctuations in the data generation process. Another insightful analysis using the concept of coherence was developed by Thomson, Pollock, Onkal, and Gonul (2019).

5. Conclusion

This paper has provided a brief analysis of the reasons why forecast combinations are successful. It is a short peek into some aspects of forecast combination. There are many other very insightful works in the literature into this important topic that consider several different aspects, such as the effects of serial correlation, heteroscedasticity, structural breaks, estimation error in the combination weights, etc. We hope that our work will be one extra step in guiding researchers and practitioners towards the successful use and understanding of forecast combinations.

References

Andrawis, R., Atiya, A. F., & El-Shishiny, H. (2011). Combination of long term and short term forecasts, with application to tourism demand forecasting. *International Journal of Forecasting*, *27*, 870–886.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers.

Ben Taieb, S., & Atiya, A. F. (2016). A bias and variance analysis for multi-step-ahead time series forecasting. *IEEE Transactions Neural Networks and Learning Systems*, *27*, 62–76.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583.

De Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*, 190–204.

Hendry, D., & Clements, M. (2004). Pooling of forecasts. *The Econometrics Journal*, *7*, 1–31.

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.

Kourentzes, N., & Petropoulos, F. (2016). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*, *181*, 145–153.

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*, 291–302.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., et al. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.

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

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

Thomson, M. E., Pollock, A. C., Onkal, D., & Gonul, M. S. (2019). Combining forecasts: performance and coherence. *International Journal of Forecasting*, *35*, 474–484.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Elsevier.

Yousef, W. A., & Kundu, S. (2014). Learning algorithms may perform worse with increasing training set size: algorithm-data incompatibility. *Computational Statistics & Data Analysis*, *74*, 181–197.

Amir F. Atiya received his B.S. and M.S. degrees in 1982 and 1985 from Cairo University, and his M.S. and Ph.D. degrees in 1986 and 1991 from Caltech, Pasadena, CA, all in electrical engineering. Dr. Atiya is currently a Professor at the Department of Computer Engineering, Cairo University. His research interests are in the areas of machine learning, theory of forecasting, computational finance, and dynamic pricing. He has obtained several awards, such as the Kuwait Prize in 2005, and The Egyptian State Appreciation Award in 2018. He is currently the Handling Editor of International Journal of Forecasting.