Journal of Business Research 69 (2016) 3951–3962

Contents lists available at ScienceDirect

Journal of Business Research

When to choose the simple average in forecast combination

Sebastian M. Blanc *, Thomas Setzer

Karlsruhe Institute of Technology, Fritz-Erler-Straße 23, 76131 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 1 October 2015
Received in revised form 17 May 2016
Accepted 21 May 2016
Available online xxxx

Keywords:
Forecast combination
Simple average
Optimal weights
Structural changes

ABSTRACT

Numerous forecast combination techniques have been proposed. However, these do not systematically outperform a simple average (SA) of forecasts in empirical studies. Although it is known that this is due to instability of learned weights, managers still have little guidance on how to solve this "forecast combination puzzle", i.e., which combination method to choose in specific settings. We introduce a model determining the yet unknown asymptotic out-of-sample error variance of the two basic combination techniques: SA, where no weightings are learned, and so-called optimal weights that minimize the in-sample error variance. Using the model, we derive multi-criteria boundaries (considering training sample size and changes of the parameters which are estimated for optimal weights) to decide when to choose SA. We present an empirical evaluation which illustrates how the decision rules can be applied in practice. We find that using the decision rules is superior to all other considered combination strategies.

1. Introduction

The combination of forecasts has been subject to research in economics since the pioneering work of Reid (1968) and Bates and Granger (1969). Numerous studies show that the combination of forecasts often results in increased accuracy in comparison to any of the forecasts alone (Makridakis et al., 1982; Clemen, 1989; Makridakis & Hibon, 2000; Fildes & Petropoulos, 2015). Various techniques aiming at deriving a weighting of individual forecasts which minimizes errors out-of-sample have been proposed.

Bates and Granger (1969) introduced the so-called optimal weights (OW). The weights are determined in a least squares estimation using available past forecast error data. They are referred to as optimal as they minimize the in-sample error variance; by design, OW outperforms any other linear weighting approach in-sample. However, the out-of-sample performance is not necessarily superior since the estimated weights are strongly fitted to the training data and are consequently subject to sampling-based variance.

As a consequence, alternative weight estimation approaches have been proposed. Clemen (1989); Diebold and Lopez (1996), and Timmermann (2006) provided thorough literature reviews of the various approaches to forecast combination. Approaches include variants of optimal weights constrained to the interval [0,1], shrinkage towards the average, Bayesian outperformance probabilities, and several more approaches. Each of the alternative approaches outperformed OW as well as other approaches out-of-sample in some evaluations, but are outperformed in others. As no model exists to decide which of the approaches to choose and empirical results are ambiguous, there is no clear consensus on which forecast combination method can be expected to perform best in a particular situation.

A surprising observation of the reviews was, however, that amongst the approaches under study, the simple average (SA) was not systematically outperformed by any other approach in out-of-sample evaluations. Stock and Watson (2004) coined the term "forecast combination puzzle" for this phenomenon. Besides model-based forecasting, SA is also competitive when combining expert predictions. For instance Genre, Kenny, Meyler, and Timmermann (2013) found that for forecasts of unemployment rate and GDP growth, only few combination methods outperform SA, while their results caution against any assumption that the identified improvements would persist in the future.

The forecast combination puzzle is in line with the more general phenomenon that simpler forecasting procedures usually outperform more complex techniques. Green and Armstrong (2015) reviewed 97 studies comparing simple and complex methods, concluding that "none of the papers provide a balance of evidence that complexity improves the accuracy of forecasts out-of-sample". Simplicity in forecasting procedures corresponds to using models where few different cues are used and/or few parameters have to be estimated. Likewise, in forecast combination, where weights of forecasts instead of cues are chosen, SA is the simplest model as – in contrast to more complex models such as OW – no parameters are estimated at all.

Brighton and Gigerenzer (2015) argued that the benefits of simplicity are often overlooked because of a "bias bias", where the importance of the bias component of the error is inflated. In contrast, the variance component, resulting from oversensitivity to different samples from the same population, is often ignored. Simpler approaches are typically

* Corresponding author.
E-mail addresses: sebastian.blanc@kit.edu (S.M. Blanc), thomas.setzer@kit.edu (T. Setzer).

http://dx.doi.org/10.1016/j.jbusres.2016.05.013
0148-2963/© 2016 Elsevier Inc. All rights reserved.

more robust against different samples as the variance component is directly related to model complexity.

Simple averaging strategies have also been shown to be highly competitive in applications besides forecast combination. For instance, for venture capital decisions, Woike, Hoffrage, and Petty (2015) found that the decision quality when using equally weighted binary cues is comparable to more complex strategies, but even more robust. Graefe (2015) argued that estimating coefficients (weights) of predictors in multivariate models is only reasonable for large and reliable datasets and few predictors. For small and noisy datasets and a large number of predictors, the authors argued that including all relevant variables is more important than the weighting.

In forecast combination, the robustness of SA has been an important research topic and a considerable body of literature examines the forecast combination puzzle theoretically and empirically. As will be discussed in Section 2, results indicate that the robustness of SA stems from unstable weight estimates from small training samples or diverging forecast error characteristics between the training and the evaluation samples. In a broader sense, these findings support the "‘Golden Rule of Forecasting", stating that forecasts are to be conservative (Armstrong, Green, & Graefe, 2015). That is because increasing asymmetry of weights results in higher sensitivity to the results of one individual forecast that is less counterbalanced by others.

Although these qualitative relations are known, managers still have little guidance on which method to choose in a particular setting. More specifically, we are not aware of any comprehensive quantitative decision guidance on when to choose OW or SA.

In this paper, we introduce a model for the expected out-of-sample error variance of a forecast combination, in particular when using SA and OW. Using the model, we derive multi-criteria decision boundaries determining whether OW or SA will lead to lower asymptotic error variance in a specific setting. Practitioners can furthermore use the thresholds to assess the robustness of a decision. We show that existing empirical guidelines can largely be explained by the model. Furthermore, in an empirical study with data from the M3 competition, we demonstrate that the recommendations and the thresholds can be used to implement successful combination decision strategies in practical settings.

2. Related work

A substantial amount of research has been conducted on the performance and robustness of SA in comparison to other forecast combination methods. A basic and intuitive finding is that the performance of SA depends on the ratio of the error variances of the forecasts as well as on their correlation. SA can be expected to perform well in case of similar error variances and low or medium error correlations (Bunn, 1985; Gupta & Wilton, 1987), since the weights which are optimal in the evaluation sample then approach equal weights. However, as shown by Dickinson (1973); Winkler and Clemen (1992), and Smith and Wallis (2009), SA can outperform other methods even for differing error variances or strongly correlated errors because of instable weight estimates. Elliott (2011) found that gains from using OW instead of SA are often too small to balance estimation errors. Claeskens, Magnus, Vasnev, and Wang (2016) showed that weight estimation can even introduce biases in combinations of unbiased forecasts.

Monte Carlo simulations by Kang (1986) and Gupta and Wilton (1987) confirmed that unstable weight estimates are key to the high competitiveness of SA. Evaluations on real-world data, for instance for U.S. money supply forecasts (Figlewski & Urich, 1983) or GNP forecasts (Kang, 1986; Clemen & Winkler, 1986) showed similar results.

Some guidelines to help decision-makers in selecting a combination method have been proposed. In the case of two forecasts, Schmittlein, Kim, and Morrison (1990) recommended SA for small sample sizes and for errors with similar variances and weak correlation. De Menezes, Bunn, and Taylor (2000) recommended SA only for approximately equal error variances and OW for large samples and low error correlation. In other cases, they suggested using outperformance probabilities (with small samples and unequal error variances), optimal weights constrained to the interval [0,1] (with medium or large samples and correlation over 0.5), or OW calculated with a correlation of zero instead of the estimated correlation, i.e., assuming uncorrelated errors (with medium sample sizes and correlations below 0.5). Thresholds for similarity/dissimilarity of error variances and sample size were, however, not quantified.

Both guidelines assume equal characteristics (error variances and covariances) of known training and unknown (future) observations. However, these characteristics might change over time because of structural changes in time series, which might influence the performance of OW and SA very differently. Miller, Clemen, and Winkler (1992) showed that SA can, in comparison to OW and other approaches, benefit from several types of structural breaks such as location shifts. Diebold and Pauly (1987) found that structural changes generally tend to impact complex approaches more than simpler ones as the estimated weights tend to increasingly differ from the ones that would minimize error in the evaluation sample.

In this paper, in contrast to existing guidelines, we propose an analytical model to determine whether SA will asymptotically outperform OW in a specific setting. We derive decision rules based on statistical considerations that do not only consider sample size and variance/covariance estimates, but also how much those values are allowed to divergence between training and evaluation sample for a decision to stay optimal. These thresholds are key to assessing the robustness of a decision but have received scant attention in the literature so far.

3. Forecast combination

Given two forecasts $\hat{y}_A$ and $\hat{y}_B$ for an event y, a combined forecast can be calculated by weighting both forecasts. The most common approach is a linear combination of the forecasts using weight w to derive a novel forecast $\hat{y}_C = w\hat{y}_A + (1-w)\hat{y}_B$. Assuming unbiased individual forecasts with errors $e_A = y - \hat{y}_A \sim \mathcal{N}(0, \sigma_A^2)$, $e_B = y - \hat{y}_B \sim \mathcal{N}(0, \sigma_B^2)$ and a correlation ρ between e_A and e_B, Bates and Granger (1969) proposed optimal weights (OW) minimizing the error variance of $\hat{y}_C$ in-sample. The original definition as well as an alternative one using the ratio of error standard deviations $\phi = \sigma_A/\sigma_B$ and the assumption $\sigma_A = 1$ (which, in combination, does not change the estimate) are presented in Eq. (1).

$$w = \frac{\sigma_B^2 - \rho\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B} = \frac{1-\rho\phi}{1+\phi^2-2\rho\phi} \tag{1}$$

The in-sample error variance of a forecast combination with different weights is illustrated in Fig. 1. The individual error variances $\sigma_A^2 = 1$ and $\sigma_B^2 = 4$ are indicated by the dotted horizontal lines. The graph shows the error variance resulting from combining forecasts with OW, SA, and with static weights set to $1/\phi$ (2:1 in the example).

When using OW, the combined error variance never exceeds the lower of the two error variances. In contrast, the combined error variances with SA and a static 2:1 weighting are lowest for an error correlation of −1 and linearly increase with error correlation. At some level of error correlation, the combined error variance exceeds the one of the better forecast ($\sigma_A^2 = 1$, in our case). However, the combined error variance still never exceeds the higher error variance — in our case $\sigma_B^2 = 4$. In summary, the difference between error variance with fixed weights (SA or 2:1) and OW is small for strong negative correlations and strictly increases with error correlation.

While OW combination leads to lower in-sample error variance than any other weighting scheme (especially weightings that ignore error variances and error correlation in the training data), we reconsider that SA often outperforms OW out-of-sample, indicating that the estimated weights do not always fit unknown observations well.

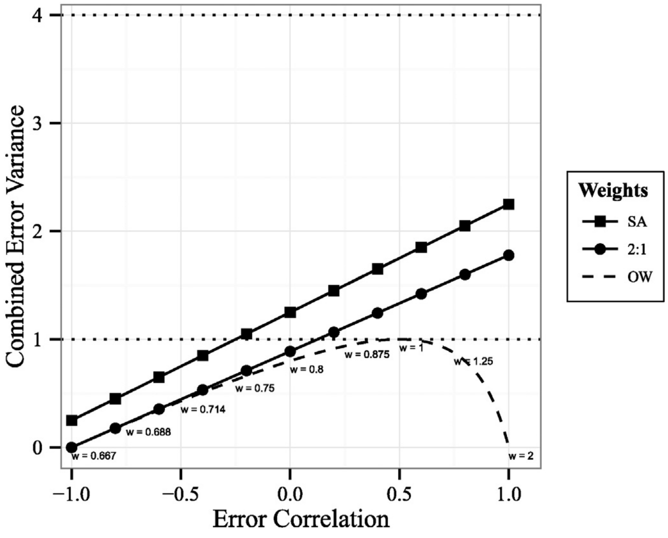


Fig. 1. The plot shows the in-sample error variance of different combinations of two forecasts (with error variances $\sigma_A^2 = 1, \sigma_B^2 = 4$) for different error correlations. Optimal weights (dashed line) always performs at least as well as the better forecast (lower dotted line). Error variances for a SA or 2:1 combination strongly depend on the error correlation.

Generally, issues with OW can be decomposed into sampling issues due to small training samples and changes of ϕ or ρ between training and evaluation sample. In the example above, as SA and OW have comparable in-sample error variances for negative error correlations, small errors in weight estimations can easily negate the benefits of OW in case of low and negative error correlations. For (strong) positive error correlations, weights learned with OW are highly sensitive to error correlation, where small correlation estimation errors can dramatically increase the error variance of the combination. Consider, for instance, the example above and error correlations approaching $+1$. With OW, we obtain a weight of 2 for $\hat{y}_A$ (and consequently -1 for $\hat{y}_B$). The weight learned with OW strongly decreases with decreasing error correlation, for example to 1.5 when error correlation decreases to 0.9. As a result, the potential benefits from using OW can easily be eliminated by small changes in error variance or error correlation between the two samples.

We will now introduce a statistical model to determine the expected out-of-sample error variance of linear combinations of forecasts, with OW and SA as special cases. Using the model, one can anticipate whether OW will (asymptotically) lead to lower error variance compared to SA in order to decide which of both combination techniques to use. We will derive multi-criteria decision boundaries regarding training sample size, ratio of error standard deviations, and error correlation as well as regarding the deviations of error variances and correlation between the training and evaluation sets.

4. Error variance of forecast combination and decision boundaries

The *training sample* (T) and the *evaluation sample* (E) are two independent bivariate samples of forecast errors. T has size n, a ratio of error standard deviations ϕ_T and error correlation ρ_T. Optimal weights $\hat{w}$ are estimated from T and are then applied to E (with a potentially different ratio of error standard deviations ϕ_E and error correlation ρ_E). The error of $\hat{y}_C$ in E is $e_C^E = \hat{w}e_A^E + (1-\hat{w})e_B^E$ with $E[e_C^E] = 0$. For our theoretical analyses, we assume $\sigma_A = 1$ and focus on ϕ, as this reduces one parameter and does not influence a decision between OW and SA. We denote the error variance of $\hat{y}_C$ (in E) by ξ. ξ depends on the characteristics of E in terms of ϕ_E and ρ_E, as well as on the expectation and variance of the weights estimated on the training set T. ξ is computed as shown in Eq. (2). The step-by-step development of ξ is provided in Appendix A.

$$\xi(E[\hat{\omega}], Var[\hat{\omega}], \phi_E, \rho_E) = \frac{1 + \phi_E^2 - 2\rho_E\phi_E}{\phi_E^2}\left(Var[\hat{\omega}] + (E[\hat{\omega}])^2\right) + \frac{2(\rho_E\phi_E - 1)}{\phi_E^2}E[\hat{\omega}] + \frac{1}{\phi_E^2} \tag{2}$$

For Bates and Granger's optimal weights, OW, $E[\hat{w}] = \frac{1-\rho_T\phi_T}{1 + \phi_T^2 - 2_{\rho T\phi T}}$ and, as shown by (Winkler & Clemen, 1992), $Var[\hat{w}] = \frac{\phi_T^2(1-\rho_T^2)}{(n-3)(1 + \phi_T^2 - 2\rho_T\phi_T)^2}$. Plugging both into Eq. (2) directly gives ξ_{OW}, the expected error variance of OW in E shown in Eq. (3).

$$\xi_{OW}(n, \phi_T, \phi_E, \rho_T, \rho_E) = \frac{1 + \phi_E^2 - 2\rho_E\phi_E}{\phi_E^2\left(1 + \phi_T^2 - 2\rho_T\phi_T\right)^2}\left(\frac{\phi_T^2(1-\rho_T^2)}{n-3} + (1-\rho_T\phi_T)^2\right) - \frac{2(1-\rho_T\phi_T)(1-\rho_E\phi_E)}{\phi_E^2\left(1 + \phi_T^2 - 2\rho_T\phi_T\right)} + \frac{1}{\phi_E^2} \tag{3}$$

For SA, we have $E[\hat{w}] = 0.5$ and $Var[\hat{w}] = 0$ by definition. Plugging both into Eq. (2) and simplifying gives ξ_{SA}, the expected error variance of SA in E shown in Eq. (4).

$$\xi_{SA}(\phi_E, \rho_E) = \frac{1 + \phi_E^2 + 2\rho_E\phi_E}{4\phi_E^2} \tag{4}$$

Assuming ϕ_E and ρ_E are known, the decision rule would simply be to choose SA instead of OW if (and only if) $\xi_{OW} \geq \xi_{SA}$. However, in practical settings the parameters ϕ_E and ρ_E are unknown (and can differ from ϕ_T and ρ_T observed in the training data), which prevents a straight-forward application of this decision rule.

We proceed as follows. We will relate Eqs. (3) and (4) to derive critical parameter values (thresholds) for deciding whether OW can be expected to outperform SA. This allows us to define minimum margins to the thresholds for selecting OW, which provides a definable level of robustness for a decision of choosing OW. First, we derive the critical training sample size n, assuming no changes of ϕ and ρ between training and evaluation sample. Second, we derive critical values for ϕ_E (assuming $\rho_T = \rho_E$) and ρ_E (assuming $\phi_T = \phi_E$). Thresholds for both ϕ_E and ρ_E are then used to define the multi-criteria decision boundaries. Third, we will relate the derived thresholds to ϕ_T and ρ_T to analyze the maximum changes of both parameters – our margins– for a decision to stay optimal.

4.1. Threshold for training sample size

Small sample sizes of T can lead to large parameter estimation errors and higher error variance with OW in comparison to SA. To determine the critical training sample size $\tilde{n}$ depending on $\phi = \phi_T = \phi_E$ and $\rho = \rho_T = \rho_E$, we set $\xi_{OW}(\tilde{n}, \phi, \phi, \rho, \rho) = \xi_{SA}(\phi, \rho)$ and solve for the critical sample size $\tilde{n}$ shown in Eq. (5). The derivation of Eq. (5) is presented in Appendix B.

$$\tilde{n} = \left\lceil \left(\frac{2\phi}{\phi^2 - 1}\right)^2 (1 + \rho^2) + 3 \right\rceil \tag{5}$$

OW can be expected to perform at least as well as SA for all training samples with at least $\tilde{n}$ observations, given unchanged error variances and error correlation.

4.2. Thresholds for changes in error correlation and error variances

We now derive the thresholds for ϕ_E and ρ_E that separate decision for OW or SA. We will keep one of both characteristics fixed and determine a threshold for the other.

Regarding changes in error correlation: assuming error variances do not change and the training sample is large, OW can be expected to perform better than SA in case of unchanged error correlation. An increase in error correlation can also be expected to be beneficial for OW combination since a higher error correlation would demand an even stronger weighting. That is because the weights which are optimal in the evaluation sample would then be further away from equal weights than the estimated OW. Hence, only a decreasing error correlation can make SA the better choice.

By keeping $\phi = \phi_T = \phi_E$ fixed and solving $\xi_{OW}(n, \phi, \phi, \rho_T, \tilde{\rho}_E) = \xi_{SA}(\phi, \tilde{\rho}_E)$ for $\tilde{\rho}_E$, we derive the critical value for ρ_E shown in Eq. (6). For the complete derivation see Appendix C.

$$\tilde{\rho}_E = \frac{\left(1-\phi^2\right)\left(\phi^4-4\rho_T\phi^3+4\rho_T\phi-1\right)(n-3)+4\phi^2\left(1+\phi^2\right)\left(1-\rho_T^2\right)}{8\phi^3\left(1-\rho_T^2\right)+2\phi\left(1-\phi^2\right)^2(n-3)} \tag{6}$$

Note that some values of ρ_T and ϕ can lead to $\tilde{\rho}_E < -1$. As correlations below -1 do not exist, there is no critical value in these cases and OW combination is robust against all kinds of changes in error correlation. We provide a detailed discussion of Eq. (6) in the next section.

Regarding changes in error variances: as discussed earlier in this article, SA is hard to beat for increasingly similar error variances, especially for $\tilde{\phi}_E \approx 1$, Thus, a change of the ratio of error standard deviations from ϕ_T to ϕ_E towards one can lead to SA performing better than OW. In contrast, if the difference in error variances increases, OW is still beneficial. Setting $\rho = \rho_T = \rho_E$ and solving $\xi_{OW}(n, \phi_T, \tilde{\phi}_E, \rho, \rho) = \xi_{SA}(\tilde{\phi}_E, \rho)$ for $\tilde{\phi}_E$ yields the critical value shown in Eq. (7), where, for reasons of convenience, we set $\eta_1 = 3 + \phi_T^2 - 4\rho\phi_T$, $\eta_2 = 1 + 3\phi_T^2 - 4\rho\phi_T$, $\psi = \phi_T^2(1-\rho^2)$ and $m = (\phi_T^2 - 1)(n-3)$. The derivation of Eq. (7) is provided in Appendix D.

$$\tilde{\phi}_E = \frac{4\rho\psi+\rho\left(\phi^2-1\right)m \pm \sqrt{\rho^2\left(4\psi+\left(\phi_T^2-1\right)m\right)^2-(4\psi-\eta_1 m)(4\psi+\eta_2 m)}}{4\psi-\eta_1 m} \tag{7}$$

We will provide a detailed discussion and illustrations of Eq. (7) in the next section, but will now already highlight one surprising conclusion that can be drawn from Eq. (7). If we find that $4\rho\psi + \rho(\phi^2-1)m \geq \sqrt{\rho^2(4\psi+(\phi_T^2-1)m)^2-(4\psi-\eta_1 m)(4\psi+\eta_2 m)}$, two different valid, i.e., positive, solutions exist; hence, two different thresholds exist for $\tilde{\phi}_E$. The first solution is reached by changing ϕ towards 1. The second solution, however, is reached when ϕ decreases to a value close to zero, in which case OW will also lead to higher asymptotic error than SA. While the first solution is expected and in line with our previous discussion, the second solution is less intuitive. This second threshold can be explained by estimated OW to the weights which are optimal in the evaluation sample. For example, assuming $\rho_T = \rho_E = 0.9$ and $\phi_T = 0.6$ decreasing to $\phi_E = 0.01$, the OW estimate is 1.643, while the weight minimizing the error variance in the evaluation sample is $1.009 \approx 1$. In this example, equal weights are “closer” than the OW estimate – even though ϕ decreased – which is generally beneficial for OW.

Since both sampling issues due to small sample sizes and changing ϕ or ρ parameters lead to weight estimates which are not well fit to the evaluation sample, we will henceforth isolate the effect of diverging sample characteristics from sampling-based weight estimation errors. For this purpose, we assume an infinitely large training sample and analyze the derived critical values for ϕ and ρ under $n \to \infty$, which we will refer to as $\tilde{\rho}_E^\infty$ and $\tilde{\phi}_E^\infty$. The thresholds are shown in Eqs. (8) and (9).

$$\tilde{\rho}_E^\infty = \lim_{n\to\infty} \tilde{\rho}_E = 2\rho_T - \frac{\phi^2+1}{2\phi} \tag{8}$$

$$\tilde{\phi}_E^\infty = \lim_{n\to\infty} \tilde{\phi}_E = \frac{\rho\left(1-\phi_T^2\right) \pm \sqrt{\rho^2\left(1-\phi_T^2\right)^2+\left(3+\phi_T^2-4\rho\phi_T\right)\left(1+3\phi_T^2-4\rho\phi_T\right)}}{3+\phi_T^2-4\rho\phi_T} \tag{9}$$

4.3. Illustration and discussion

We now illustrate and discuss the critical values derived in the previous section, before we compare them to results and recommendations from the literature. For this purpose, we assume that $\phi_T \leq 1$, which can always be achieved by switching the two forecasts. The critical training sample size $\tilde{n}$ is depicted in Fig. 2 as a function of ρ and for different ϕ between 0.5 and 0.9. First, the figure shows that $\tilde{n}$ increases with ϕ — the training sample size required to put OW in favor of SA increases with the difference between error variances. Second, for fixed ϕ, the higher the absolute value of ρ, the smaller the required training sample. For example, choosing OW requires the largest training samples with uncorrelated errors.

The first finding is consistent with the literature and intuitively clear; the closer the error variances of the forecasts, the more training observations are required to ensure that OW (in these cases close to equal weights) are beneficial. Considering the uncertainty in weight estimation, the probability of estimating OW that are more appropriate in the evaluation sample than SA decreases with increasing ϕ. For instance, for ϕ over 0.8 and uncorrelated errors, a training sample of at least 25 observations is required. For ϕ exceeding 0.9, more than 75 observations are necessary to favor OW over SA.

The second relationship, that $\tilde{n}$ decreases with $|\rho|$, can – for positive error correlations – be directly concluded from Fig. 1. Reconsidering that a positive error correlation leads to OW exceeding one for one forecast and a negative weight for the other, fewer observations $\tilde{n}$ are required in the training sample for OW to perform better than SA.

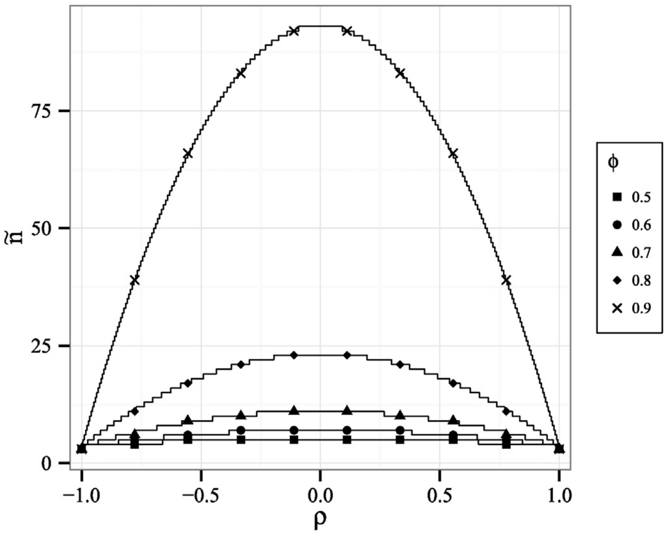


Fig. 2. Minimal training sample size $\tilde{n}$ to favor OW over SA. Forecasts with ϕ close to 1 require larger training samples. Strong correlation of forecast errors lowers the required size of the training sample.

The relationship between $\tilde{n}$ and ρ for negative values of ρ is less obvious since error variance with SA also decreases with ρ decreasing from 0 to -1. However, as also illustrated in Fig. 1, OW differ only slightly between negative values of ρ. That makes OW more robust against small estimation errors of ρ. In the example in Fig. 1, assuming that $\hat{\phi}_T = 0.5$, OW are $\hat{w} = 0.714$ for $\hat{\rho}_T = -0.5$ and $\hat{w} = 0.667$ for $\hat{\rho}_T = -1$, a difference in weighting of only 7% as a result of a correlation changing by 0.5. Similarly, estimation errors of ϕ do not lead to large OW differences. For instance, assuming that $\hat{\rho}_T = -0.95$, OW will estimate $\hat{w} = 0.527$ with $\hat{\phi}_T = 0.9$, while $\hat{\phi}_T = 0.8$ leads to an OW estimated weights of $\hat{w} = 0.557$, a difference of only 6%.

The critical change in error correlation, $\tilde{\rho}_E^\infty - \rho_T$ (based on the critical value $\tilde{\rho}_E^\infty$ introduced in Eq. (8)), is depicted in Fig. 3 as a function of ρ_T. For each curve, ϕ is kept constant at a value between 0.5 and 0.9. It is clear from the figure that only decreases of error correlation are critical while the stronger the (positive) error correlation, the smaller the critical change. Furthermore, the different values of ϕ lead to critical changes differing by a constant factor, which is independent of ρ_T. OW is least robust against changes for ρ_T and ϕ close to 1.

Both dependencies are intuitive. As illustrated above, strong positive error correlations can lead to extreme OW estimates. In these cases, small decreases of the error correlation result in equal weights being closer to the weights which are optimal in the evaluation sample. The reasoning for the increased robustness for lower ϕ is similar: the lower ϕ, the less likely SA performs as well as OW considering the differing error variances of the forecasts. As a consequence, high changes of ρ are required to make OW perform worse than SA.

The critical change of the ratio of error standard deviations, $\tilde{\phi}_E^\infty - \phi_T$, based on Eq. (9), is depicted in Fig. 4 as a function of ϕ_T. The figure shows four curves for different values of ρ between -0.99 and 0.99 (for each curve we assume equal error correlation in both samples, $\rho = \rho_T = \rho_E$).

The critical change in ϕ decreases with increasing ϕ_T and ρ. In particular, with ρ close to 1 (the 0.99 line) and similar error variances ($\phi \approx 1$), a minimal change in ϕ leads to a better performance with SA in comparison to OW. In contrast, with extreme differences between error variances, substantial changes in error variances are required to justify a decision in favor of SA, increasingly so with higher error correlation.

So far we discussed the critical values that a parameter is allowed to change assuming the other remains constant. To provide further insights, we will now present numerical results for $\xi_{SA} = \xi_{OW}$ depending on multiple parameter changes. Results for selected ϕ_T, ρ_T, and n are depicted in Fig. 5.

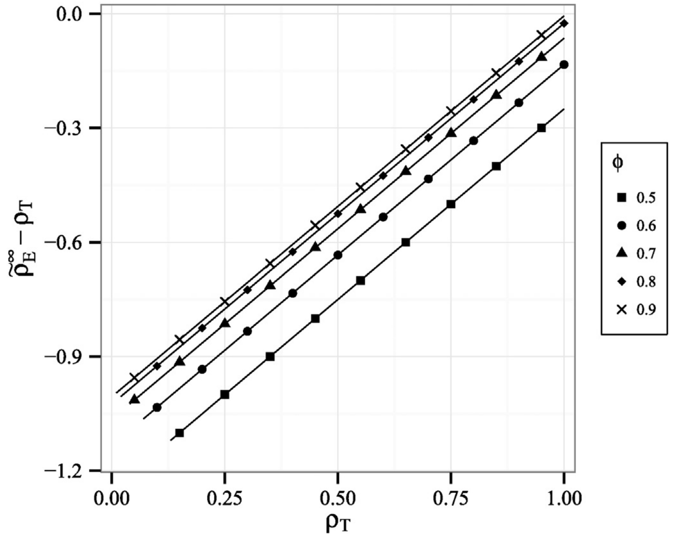


Fig. 3. Critical changes of error correlation depending on the error correlation in T and the ratio of error standard deviations ϕ. An OW combination of forecasts which are highly correlated in T is less robust to changes than a combination of forecasts with weak correlations. Larger differences in accuracy (lower ϕ) between forecasts increase the robustness of choosing OW.

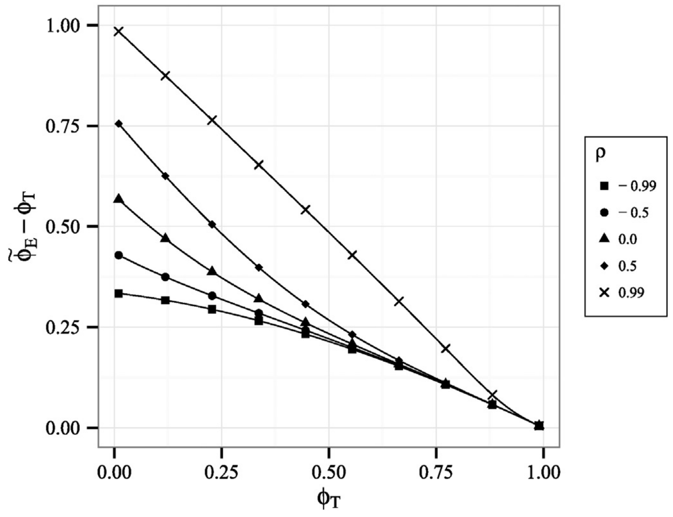


Fig. 4. The figure shows critical changes of the ratio of error standard deviations depending on the ratio in the training sample and on the error correlations. The critical change of ϕ increases with ϕ_T as well as with ρ. The robustness of OW consequently strongly decreases for higher error correlations.

Each graph shows critical thresholds (decision boundaries) for $\rho_E - \rho_T$ and $\phi_E - \phi_T$ with different combinations of ρ_T and ϕ_T. The curves in a plot correspond to threshold lines with different numbers of training observations. A point in each of the plots indicates zero changes of ρ and ϕ.

For parameter changes exceeding a threshold line SA would be advantageous. Taking $\rho_T = 0.9, \phi_T = 0.6$ (plot on the upper-right of the figure) with $n = 50$ as an example, SA can be expected to perform better if, for instance, ρ decreases by 0.22, ϕ increases by 0.28, or ρ decreases by 0.2 and ϕ increases by 0.04.

As the curves in Fig. 5 confirm our previous discussions, we will now focus on interesting issues regarding small sample sizes and high correlations. First, taking $\rho_T = 0.75, \phi_T = 0.8$ as an example, the “no change” point is on the right-hand sides of the $n = 10$ and $n = 25$ lines. This indicates a too small training sample, making the estimates on the training sample too unstable and SA the better decision. For these small training samples, OW would only have lower asymptotic error variance if ϕ and/or ρ change significantly between training and evaluation sample *in favor* of OW. Second, as previously discussed, not only increases but also strong decreases of ϕ lead to lower error variance with SA than OW. This can be seen in the plots with $\rho_T = 0.9$ and low ϕ_T, where the area of beneficial OW combinations is convex. This result especially occurs for small sample sizes, again indicating issues with small samples. Before we study beneficial margins to the decision boundaries to learn a more robust decision rule, we now first compare the decision boundaries to recommendations and results in the literature on forecast combination.

4.4. Comparison to recommendations in Schmittlein et al. (1990)

First, we compare our decision boundaries to the recommendations in Schmittlein et al. (1990). Based on mean squared errors (MSEs) with SA and OW observed in Monte Carlo simulations for different combinations of n, ϕ, and ρ, Schmittlein et al. analyzed when to use which combination method. They considered 19 values for ρ, 20 values for ϕ, and training sample sizes of 10, 25, 50, and 100.

The results of the replicated simulation experiment with 100 runs are depicted in Fig. 6. The four plots show which of the two models leads to lower mean MSE across runs per parameter combination. Filled circles

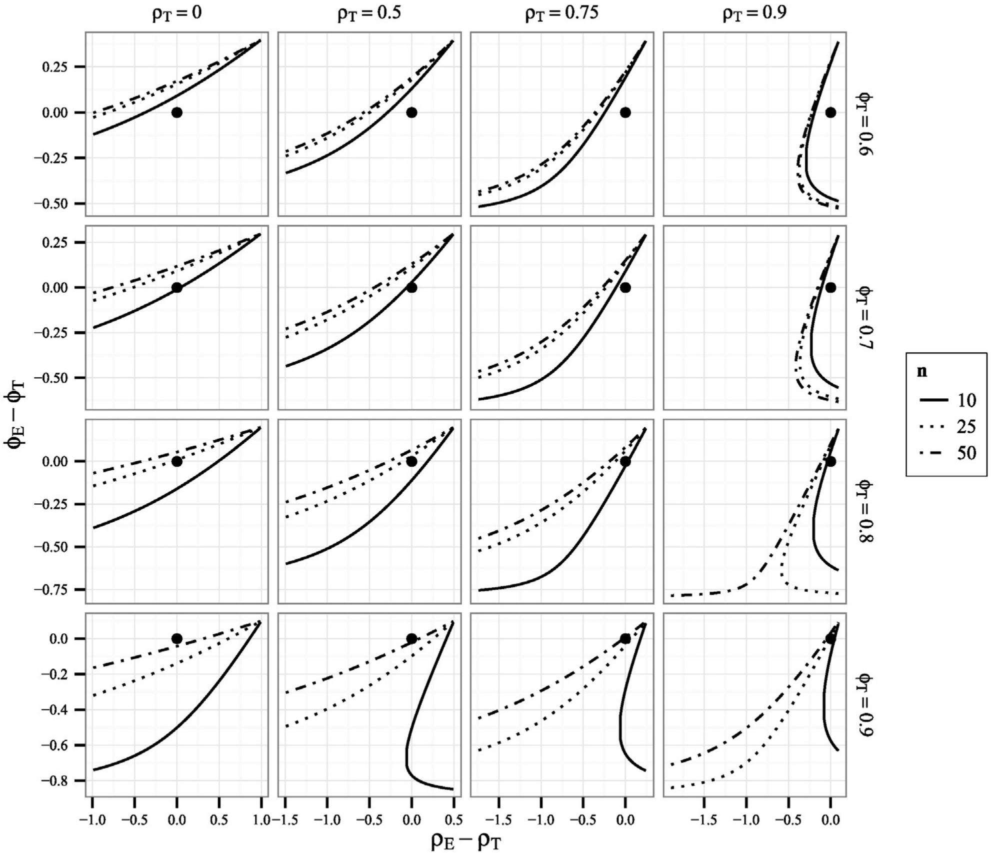


Fig. 5. Decision boundaries for changes of ϕ and ρ from training to evaluation sample depending on ϕ_T, ρ_T, and training sample size n. The dot in each plot indicates unchanged parameters. SA is beneficial for combinations on the top-left of the individual boundaries.

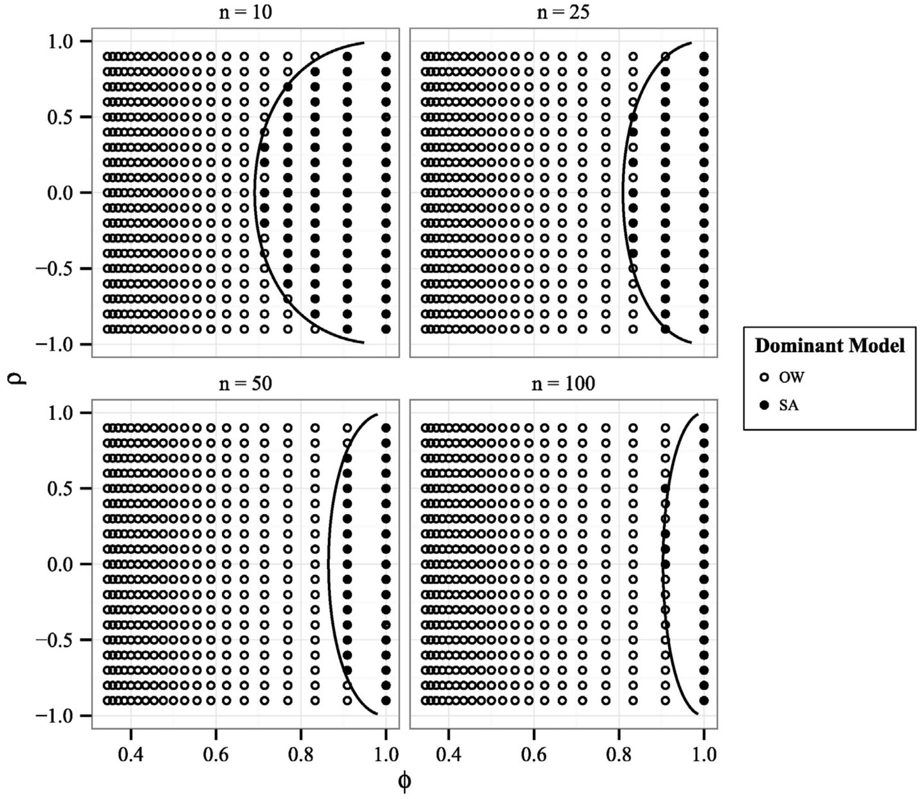


Fig. 6. Decision boundaries (OW vs. SA) based on (ϕ,ρ), for different training sample sizes n. The plots also show which model leads to lower mean MSE in Monte Carlo simulations. The simulation described in Schmittlein et al. (1990) was replicated and outcomes of our simulations as well as our decision boundaries conform with the findings in Schmittlein et al. (1990). Our analytically derived decision boundaries separate the classes well, with mis-classifications only very close to the boundary.

Table 1
Treatments regarding ϕ and ρ before and after the structural break in the experiments adapted from Miller et al. (1992). The last column shows the critical values for ϕ or ρ analytically derived with our model for $n = 29$. For changes of ϕ, only the variance increase treatment (second row) in Set 1 is expected to favor SA. The variance decrease treatment in Set 2 has a changed value only marginally differing from the threshold; SA and OW are thus expected to perform similarly. Regarding changes of ρ, we expect that the correlation decreases for both Set 1 and Set 2 as well as the correlation increase for Set 1 are critical.

Set	Treatment	Initial properties ($t<30$)	Changed property ($t\geq30$)	Critical value
Set 1	Var. increase	$\rho=0.8, \phi=0.949$	$\phi=0.728$	$\tilde{\phi}=0.846$
	Var. decrease	$\rho=0.8, \phi=0.728$	$\phi=0.949$	$\tilde{\phi}=0.857$
Set 2	Var. increase	$\rho=0.8, \phi=0.837$	$\phi=0.624$	$\tilde{\phi}=0.884$
	Var. decrease	$\rho=0.8, \phi=0.624$	$\phi=0.837$	$\tilde{\phi}=0.838$
Set 1	Cor. increase	$\rho=0.4, \phi=0.949$	$\rho=0.8$	$\tilde{\rho}=0.907$
	Cor. decrease	$\rho=0.8, \phi=0.949$	$\rho=0.4$	$\tilde{\rho}=0.935$
Set 2	Cor. increase	$\rho=0.4, \phi=0.837$	$\rho=0.8$	$\tilde{\rho}=0.403$
	Cor. decrease	$\rho=0.8, \phi=0.837$	$\rho=0.4$	$\tilde{\rho}=0.715$

indicate parameter combinations where SA outperformed OW, while non-filled circles indicate a recommendation for OW.[1] As a solid line, we additionally show the decision boundary resulting from our model.

Recommendations given in Schmittlein et al. (1990), for instance to choose SA if $|\rho|<0.6$, $\phi>0.83$, and $n\leq10$, or if error standard deviations differ by at most 10% ($\phi_T>0.91$), $|\rho|<0.4$ and $n=25$ are all well captured by the boundaries.

Furthermore, for smaller sample sizes ($n = 10$ and $n = 25$), the decision boundary separates both regions precisely, with only few misclassifications when the parameter combinations approach the boundary. In these cases, however, the difference between errors with both combination methods approaches zero with high randomness regarding the sign. The decision boundaries for larger samples sizes of 50 and 100 at the bottom of the figure also separate the cases for and against SA precisely. Results with higher numbers of simulation runs (for instance 1000), which are excluded for reasons of brevity, lead to even more accurate separations.

4.5. Comparing the impact of structural breaks with Miller et al. (1992)

We also compare the derived boundaries to the results of Miller et al. (1992), who analyzed the impact of breaks in error time series on different combination methods. For this purpose, we re-run their simulation experiments but with two instead of three individual forecasts. In the experiment, two forecast error time series of length 100 are generated and combined using SA and OW. The experiments start at time interval eight, when OW are first estimated on all past error observations. Then, the size of the training sample is increased and revised OW are estimated, again on all past observations. At time 30, one characteristic of the errors (either ϕ or ρ) increases or decreases from its initial value. Based on the treatments of Miller et al. (1992), changes of the characteristics are defined for two basic sets of properties (referred to as *Set 1* and *Set 2*), as shown in Table 1.

For each set and treatment, the initial values of ρ and ϕ are shown together with the changed characteristic after the break at time 30. Based on the initial characteristics and $n=29$, we calculated the critical values $\tilde{\phi}$ and $\tilde{\rho}$, for which OW and SA can be expected to perform similarly. In most cases this value is the level a characteristic is *allowed to change* for OW to remain the better choice afterwards. However, in the correlation decrease treatment for both sets, this value is the level the characteristic would *have to change* to make OW perform at least as well as SA. Changed characteristics exceeding (or not exceeding, depending on the threshold) the critical value are printed in bold. As a result, the variance decrease and the correlation increase treatments for Set 1 as well as the correlation decrease for both sets can be expected to result in SA performing better than OW. The variance decrease for Set 2 can be expected to result in very similar performance. It is important to note that OW learns new weights with each (seen) observation since the training period continuously grows and weights are adjusted after the structural break.

The average root mean squared error (RMSE) outcomes of the simulation experiments over 3000 runs are shown per time interval in Fig. 7. In the figure, the RMSE over all runs is plotted for each time interval per parameter set and treatment. Curves are smoothed with smoothing splines for visualization purposes, as done by Miller et al. (1992).

Results indicate that our derived thresholds properly differentiate between uncritical and critical structural breaks, and show which combination model should be chosen after the break to achieve lower error variance. All cases where the new value of the changed characteristic is (in the right direction) far away from the critical value are indeed uncritical. On the other hand, we observe a higher RMSE for OW after the critical variance decrease and (with a small difference between SA and OW) the correlation increase for Set 1 as well as for the critical correlation decrease in both sets.

An interesting case is the variance decrease for Set 2, where the critical value and the new value of ϕ only differ by 0.001 (approximately 0.1%). In this case, the average RMSE in Fig. 7 directly after the structural break is indeed very similar for SA and OW.

5. Application of the decision boundaries to real-world data

In this section, we assess the applicability of the proposed decision rules to empirical data. We use the out-of-sample error variances of SA and OW estimated using the proposed model and the derived thresholds to implement different strategies for deciding between SA and OW.

As empirical data set, we use the time series data of the M3 Competition (Makridakis & Hibon, 2000). We limit our analysis to monthly time series (1426 of the 3003 time series) to ensure a sufficient length of the time series. We preprocess the data as follows. First, we merge the training and test time series data of the competition and then split the resulting (longer) time series into three separate samples as illustrated in Fig. 8. The first 36 months of time series data (sample I) are used as training data to calibrate two statistical forecasting methods. As example prediction models, auto-ARIMA ($\hat{y}_A$) and Damped Trend Exponential Smoothing ($\hat{y}_B$) provided by Hyndman (2015) for (R Core Team, 2015) are used. Rolling one-month-ahead forecasts are then calculated for all months in samples II and III. Using the resulting errors in sample II, OWs are estimated. These weights are applied to the forecasts for the last 24 months (sample III), our evaluation set for which the error variance is calculated. It should be noted that the size n of sample II varies between time series because of different overall time series lengths. These differences allow analyzing the robustness against different n.

5.1. Distributions of ϕ and ρ

To assess whether we can expect to find a superior performance of SA in comparison to OW at least in some cases, we now analyze

[1] Note that results for equal error variances ($\phi=1$) are missing in the plots in Schmittlein et al. (1990) leading to slightly differing results.

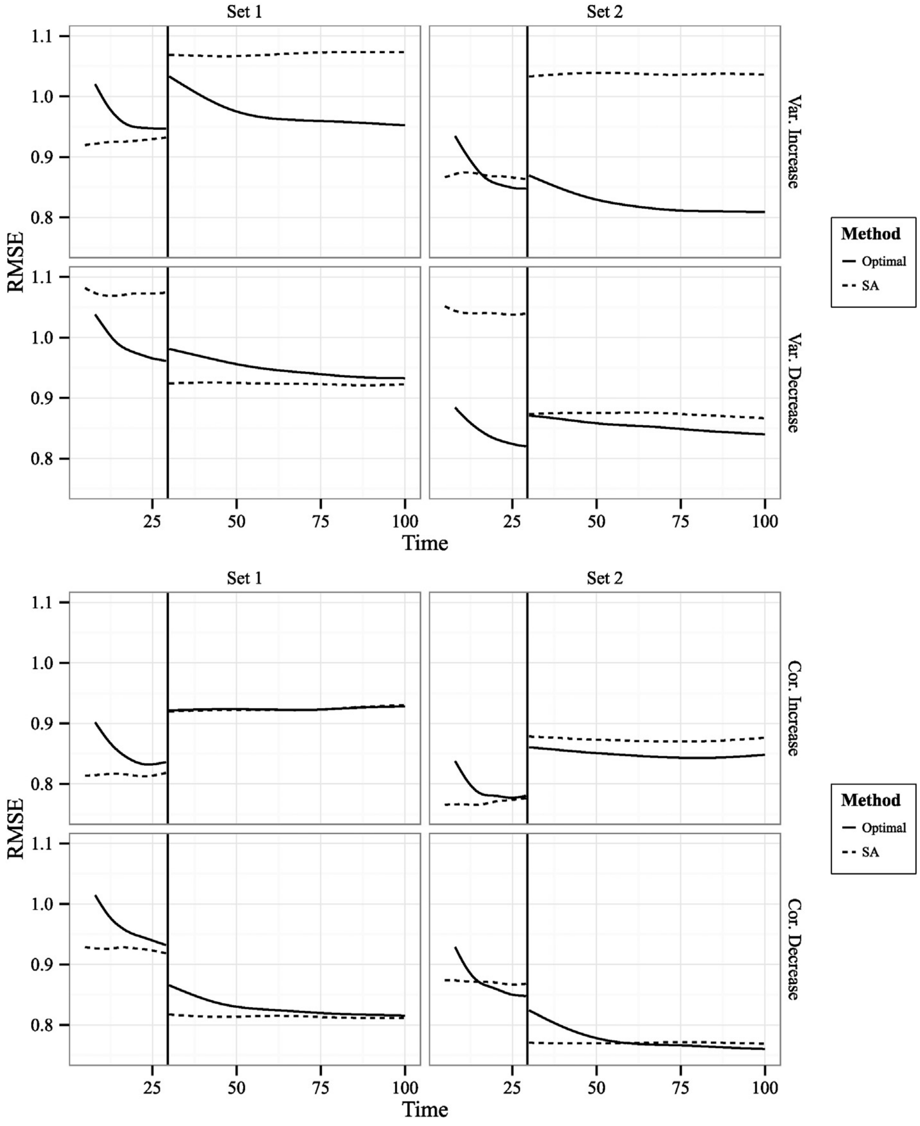
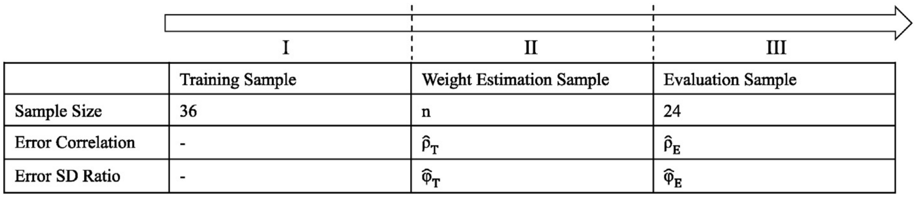


Fig. 7. Mean RMSE per time interval before and after a break (vertical line) for different parameter sets and treatments. The performance of OW improves before the break as weight estimates are stabilizing. Immediately after the break, SA outperforms OW in four of the eight cases, as indicated by our critical values. Results of OW and SA are very similar for the variance decrease treatment in Set 2, as also indicated by the critical value.

the distributions of ϕ and ρ in sample II. The two plots in Fig. 9 show the respective distributions, with $\hat{\phi}_T$ distributed around 1 with most values close to 1 and $\hat{\rho}_T$ strongly skewed with most values also close to 1. Due to the high frequency of values close to 1 for both $\hat{\phi}_T$ and $\hat{\rho}_T$, SA can be expected to be preferable in a substantial number of cases. Decision strategies utilizing the model and the thresholds introduced in this paper will be evaluated in the next section.

5.2. Evaluation of decision strategies

As discussed before, the optimal choice between OW and SA would require knowledge of ϕ_E and ϕ_E, which are unknown parameters in empirical settings. The only information available are estimates of the characteristics $\hat{\rho}$ and $\hat{\phi}$ which are derived from a sample from a population with ρ_T and ϕ_T. As already discussed, the estimators $\hat{\rho}$ and $\hat{\phi}$ will be almost certainly deviate from the population characteristics of the



	Training Sample	Weight Estimation Sample	Evaluation Sample
Sample Size	36	n	24
Error Correlation	-	$\hat{\rho}_T$	$\hat{\rho}_E$
Error SD Ratio	-	$\hat{\varphi}_T$	$\hat{\varphi}_E$

Fig. 8. Each of the time series of the M3 Competition is split into three samples. Sample I is only used as training data for two statistical forecasting models. Rolling forecasts (and corresponding errors) for the second sample (Sample II) are used to derive optimal weights. OW and SA are then applied to combine forecasts in Sample III, the evaluation data set held out.

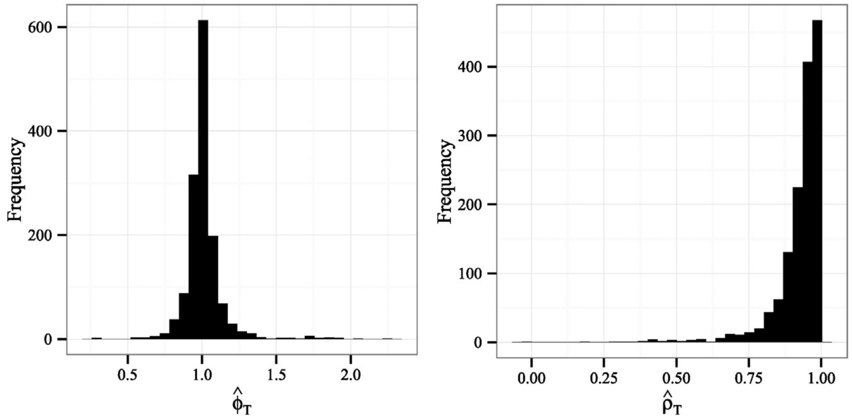


Fig. 9. Distribution of the estimated ratio of error standard deviations and correlation in Sample II. ϕ_T is centered around 1. ρ_T is skewed and most observations are approaching 1.

evaluation sample because of the sampling-related variance and potential changes from training to evaluation set.

While the sampling-related variance is already penalized in our formulae to calculate the expected out-of-sample error variance, we do not know how much the estimate differs from the population error characteristics of the evaluation set. This aspect is closely related to structural changes that can also cause a difference of parameters which can make SA the better choice and the thresholds can consequently be used. If the thresholds of a decision in favor of OW are too narrow, a switch to SA should be considered since small deviations between the available estimates and the population characteristics of the evaluation set could change the optimal decision. In such cases, a decision for SA – as the more robust option – is likely to be beneficial.

This is illustrated in Fig. 10, where the expected difference in combined error variance (calculated as difference between Eqs. (3) and (4)) between SA and OW is depicted, depending on changes of ρ (upper plots) and ϕ (lower plots), for different values of the other characteristic.

The upper plots in the figure indicate the asymmetry of expected error variance between OW and SA by color density, assuming different ϕ_T increasing from left ($\phi_T = 0.5$) to right ($\phi_T = 0.9$), depending on ρ_T (increasing from zero to one in each of the plots). Reconsidering that a correlation decrease benefits SA and vice versa, the plots show that even small decreases in error correlation can result in large differences regarding the asymptotic error variance for high ρ_T. Especially for similar error variances ($\phi_T = 0.8$ and $\phi_T = 0.9$), negative deviations from the

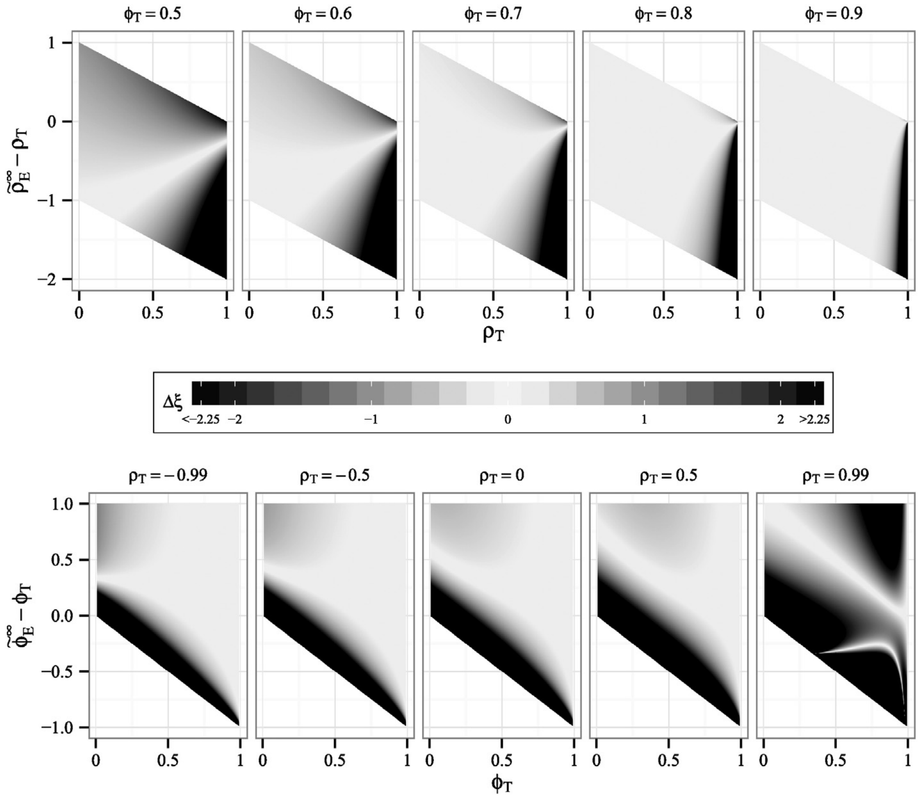


Fig. 10. Difference in expected combined error variance for changes of ρ (upper plots) and ϕ (bottom plots). In particular in cases with large ρ_T, relatively small decreases of ρ can easily result in large disadvantages when using OW instead of SA. On the other hand, small increases (in favor of OW) lead to much lower advantages when using OW instead of SA. An asymmetric relationship is also observed when ϕ changes, in particular for higher values of ϕ_T, but we observe rather moderate moderate differences in error variance for moderate changes, except for cases with high values of ρ.

horizontal no-change line ($\tilde{\rho}_E^\infty - \rho_T = 0$) result in steep increases of differences in asymptotic error variance, to the disadvantage of OW. As correlation increases, we observe that the imbalance in favor of OW increases only moderately, overall resulting in a considerable asymmetry regarding the effect of parameter deviations. Asymmetric error variance relationships are also observed when ϕ changes (lower plots), in particular with higher levels of ϕ_T, but we observe rather moderate increases of the disadvantage with OW when ϕ decreases only slightly, and steep increases later on at higher levels of ϕ changes. These insights are important to derive appropriate decision strategies.

In light of this asymmetry, an appropriate decision rule might be to select OW when the proposed model suggests to do so, but only if pre-defined margins to the decision boundary are not violated. The intuition is that small changes can easily negate the expected benefits of OW, which ought to be considered in a decision to ensure a certain degree of robustness.

In our empirical evaluation, we will apply this decision strategy with different thresholds regarding changes in ρ and ϕ. We will study the application of either narrow minimum margins of 0.01 – distances of 1% to the thresholds – denoted by *L*, or higher margins of 0.05, denoted by *H*. A decision strategy that, for instance, uses L as threshold for ϕ and H for ρ, is denoted by *Threshold L–H*. The threshold-based rules will be benchmarked against the static strategies of always choosing one method (*OW* and *SA*) as well as against a strategy based on the model recommendations, ignoring the thresholds (*Recommendation*).

As an evaluation criterion, we use the relative MSE regret and compute different quantiles over all decisions with a strategy. Relative MSE is defined as the relative difference between the MSE of a combination and the MSE of the (ex-post) optimal combination — hence, the relative MSE that could have been avoided when always making the ex-post better decision. The relative MSE regret is used as MSE values are not directly comparable across the time series of our case study, which are at very different levels. Clearly, regret is zero if all decisions are correct and increases with every wrong decision.

Table 2 displays the quantiles of regret across all time series of our case study.

The results shows that the *OW* and *Recommendation* strategies result in the highest regret values of all strategies regarding all quantiles of the regret distributions. The strategies are clearly the most aggressive ones, as, in the spirit of the bias–variance theory and the forecast combination puzzle, they clearly have the lowest in-sample bias but the highest variance. From this point of view, these results are well in line with the “Golden Rule of Forecasting” that forecasts should be conservative, as a simple average clearly dominates both approaches.

However, the results also indicate that the “Threshold” strategies are highly competitive, as all of them outperform all alternative strategies for most quantiles but the maximum regret. Consequently, the thresholds are an essential basis for robust decisions on which forecast combination method to choose. The best performing strategy amongst the strategies under study is Threshold L–H with a narrow minimum threshold for error variances and a wide minimum threshold for error correlation. A lower threshold for changes in error correlations compared to error variance ratios is well in agreement with our previous theoretical discussion on the impacts of parameter changes in Fig. 10. As a consequence, strategies involving the thresholds should put more emphasis on thresholds regarding changes of error correlation than error variances.

6. Conclusion and implications

The “forecast combination puzzle” refers to the recurring empirical finding that more sophisticated weight learning models typically do not outperform a simple average (SA) in forecast combination. It is known that estimates of the error variances of individual forecasts and their covariances, the parameters used for weighting the forecasts, are often too unstable because of small training samples or changes in the underlying time series and the corresponding error characteristics. However, models quantifying the error variance with a particular forecast combination method do not exist and managers still have little guidance on which forecast combination method to choose in a specific situation.

We introduced a model for the expected out-of-sample error variance of a forecast combination, given the characteristics of the training and evaluation sample. We used the model to calculate decision boundaries to determine when to select the simple average (SA) or weights that minimize the error variance in the training sample (OW). We focused on the combination of two individual forecasts for two reasons, which in most cases apply for the prediction of business figures in enterprises. Typically, a judgmental forecast and one that is derived using purely statistical means are available and corporate planning can be based on one of the forecast or a combination of both forecasts, where additional forecasts cannot be expected to introduce as much additional information. Furthermore, focusing on the two-forecast case allowed us to provide a variety of in-depth analyses. The challenge of extending the model and the decision boundaries to a larger, arbitrary number of forecasts is subject to future research.

In addition, we proposed robust decision rules when to choose OW based on the introduced minimum margins (of the model recommendation) to the derived decision boundaries (critical values of error correlation and variances in the evaluation sample). The intuition is to consider changes of error correlation and variances between training and test samples, which often affect the asymptotic error variance of OW much more than the one of SA. With the margins, we determined against which degrees of change of error variances and correlation between training and test time series a decision in favor of OW must be robust, i.e., still dominating SA. In an empirical study with data from the M3 competition, we showed that the thresholds are key to the implementation of successful strategies for deciding between simple and complex combinations in practical settings.

Overall, this work provides the means to better understand and analytically solve an important facet of the forecast combination puzzle by deriving a model for the asymptotic error variance of a combination, considering – besides bias – the sampling-based variance component of the error. Managers can apply the model and decision boundaries (with the L–H thresholds) to their specific settings to make profound decisions on whether to use a simple average or estimated weights.

Table 2
Quantiles of the relative MSE regret (computed across all time series) resulting from different decision strategies. Strategies using the model-based recommendation considering thresholds are named “Threshold” followed by a combination of two letters (H or L), where the first (second) letter indicates if a high or low ϕ (ρ) threshold is used. Most of the threshold-based strategies outperform the alternative strategies, including the “SA” strategy, which is the most conservative approach.

Strategy	Quantile										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
SA	0%	0%	0%	0%	0%	0%	0.30%	1.30%	3.43%	9.41%	117.85%
OW	0%	0%	0%	0%	0%	0.22%	1.53%	4.24%	10.53%	34.42%	32,810.09%
Recommendation	0%	0%	0%	0%	0%	0.09%	0.81%	2.78%	6.85%	26.08%	32,810.09%
Threshold H–H	0%	0%	0%	0%	0%	0%	0.17%	0.94%	2.87%	7.11%	163.54%
Threshold H–L	0%	0%	0%	0%	0%	0%	0.30%	1.22%	3.31%	9.54%	261.27%
Threshold L–H	0%	0%	0%	0%	0%	0%	0.06%	0.75%	2.41%	5.88%	163.54%
Threshold L–L	0%	0%	0%	0%	0%	0%	0.30%	1.28%	3.47%	10.16%	261.27%

Appendix A. Derivation of Eq. (2)

With errors e_A, e_B and corresponding error variances σ_A^2, σ_B^2 in the evaluation sample:

$$
\begin{aligned}
Var\ [\epsilon] &= Var\ [\hat{w}e_A] + Var[(1-\hat{w})e_B] + 2Cov(\hat{w}e_A, (1-\hat{w})e_B) \\
&= Var\ [\hat{w}]Var[e_A] + (E[e_A])^2 Var[\hat{w}] + (E[\hat{w}])^2 Var[e_A] + Var\ [1-\hat{w}]Var[e_B] + (E[e_B])^2 Var[1-\hat{w}] + (E[1-\hat{w}])^2 Var[e_B] \\
&\quad + 2Cov(\hat{w}e_A, e_B) - 2Cov(\hat{w}e_A, \hat{w}e_B) \\
&= \sigma_A^2 Var[\hat{w}] + \sigma_A^2 (E[\hat{w}])^2 + \sigma_B^2 Var[\hat{w}] + \sigma_B^2 (1-E[\hat{w}])^2 + 2\rho_E\sigma_A\sigma_B E[\hat{w}] - 2(\rho_E\sigma_A\sigma_B)\left((E[\hat{w}])^2 + Var[\hat{w}]\right) \\
&= \sigma_A^2 Var[\hat{w}] + \sigma_A^2 (E[\hat{w}])^2 + \sigma_B^2 Var[\hat{w}] + \sigma_B^2 - 2\sigma_B^2 E[\hat{w}] + \sigma_B^2 (E[\hat{w}])^2 + 2\rho_E\sigma_A\sigma_B E[\hat{w}] - 2(\rho_E\sigma_A\sigma_B)\left((E[\hat{w}])^2 + Var[\hat{w}]\right) \\
&= (\sigma_A^2 + \sigma_B^2 - 2\rho_E\sigma_A\sigma_B)Var[\hat{w}] + 2(\rho_E\sigma_A\sigma_B - \sigma_B^2)E[\hat{w}] + (\sigma_A^2 + \sigma_B^2 - 2\rho_E\sigma_A\sigma_B)(E[\hat{w}])^2 + \sigma_B^2 \\
&= \frac{1+\phi_E^2 - 2\rho_{E\phi E}}{\phi_E^2} Var[\hat{w}] + \frac{1+\phi_E^2 - 2\rho_{E\phi E}}{\phi_E^2}(E[\hat{w}])^2 + \frac{2\left(\rho_{E\phi E} - 1\right)}{\phi_E^2} E[\hat{w}] + \frac{1}{\phi_E^2} \\
&= \frac{1+\phi_E^2 - 2\rho_{E\phi E}}{\phi_E^2}(Var[\hat{w}]) + (E[\hat{w}])^2 + \frac{2\left(\rho_{E\phi E} - 1\right)}{\phi_E^2} E[\hat{w}] + \frac{1}{\phi_E^2}
\end{aligned}
$$

Appendix B. Derivation of Eq. (5)

$$
\begin{aligned}
0 &= \xi_{OW}(\tilde{n}, \phi, \phi, \rho, \rho) - \xi_{SA}(\phi, \rho) \\
0 &= \frac{1-\rho^2}{(\tilde{n}-3)(1+\phi^2-2\rho\phi)} - \frac{(1-\rho\phi)}{\phi^2(1+\phi^2-2\rho\phi)} + \frac{4(1-\phi^2)}{4\phi^2} - \frac{1+\phi^2-2\rho\phi}{4\phi^2} \\
0 &= \frac{1-\rho^2}{\tilde{n}-3} - \frac{4(1-\rho\phi)^2 - 4(1-\rho\phi) + (1+\phi^2-2\rho\phi)^2}{4\phi^2} \\
0 &= \frac{1-\rho^2}{\tilde{n}-3} - \frac{(\phi^2-1)^2}{4\phi^2} \\
\tilde{n} &= \frac{4\phi^2}{(\phi^2-1)^2}(1-\rho^2) + 3 \\
\tilde{n} &= (\frac{2\phi}{\phi^2-1})^2(1-\rho^2) + 3
\end{aligned}
$$

Appendix C. Derivation of Eq. (6)

$$
\begin{aligned}
0 &= \xi_{OW}(n, \phi, \phi, \rho_T, \tilde{\rho}_E) - \xi_{SA}(\phi, \tilde{\rho}_E) \\
0 &= -\frac{2\tilde{\rho}_E\phi}{\phi^2(1+\phi^2-2\rho_T\phi)^2}\frac{\phi^2(1-\rho_T^2)}{n-3} - \frac{2\tilde{\rho}_E\phi}{\phi^2(1+\phi^2-2\rho_T\phi)^2}(1-\rho_T\phi)^2 + \frac{2\tilde{\rho}_E\phi(1-\rho_T\phi)}{\phi^2(1+\phi^2-2\rho_T\phi)} - \frac{2\tilde{\rho}_E\phi}{4\phi^2} \\
&\quad + \frac{1+\phi^2}{\phi^2(1+\phi^2-2\rho_T\phi)^2} + \frac{\phi^2(1-\rho_T^2)}{n-3} + \frac{(1+\phi^2)(1-\rho_T\phi)^2}{\phi^2(1+\phi^2-2\rho_T\phi)^2} + \frac{1}{\phi^2} - \frac{2(1-\rho_T\phi)}{\phi^2(1+\phi^2-2\rho_T\phi)} - \frac{1+\phi^2}{4\phi^2} \\
0 &= \tilde{\rho}_E\left(-\frac{2}{\phi(1+\phi^2-2\rho_T\phi)^2}\frac{\phi^2(1-\rho_T^2)}{n-3} - \frac{2(1-\rho_T\phi)^2}{\phi(1+\phi^2-2\rho_T\phi)^2} + \frac{2(1-\rho_T\phi)}{\phi(1+\phi^2-2\rho_T\phi)} - \frac{2}{4\phi}\right) \\
&\quad + \frac{1+\phi^2}{\phi^2(1+\phi^2-2\rho_T\phi)^2}\frac{\phi^2(1-\rho_T^2)}{n-3} + \frac{(1+\phi^2)(1-\rho_T\phi)^2}{\phi^2(1+\phi^2-2\rho_T\phi)^2} + \frac{1}{\phi^2} - \frac{2(1-\rho_T\phi)}{\phi^2(1+\phi^2-2\rho_T\phi)} - \frac{1+\phi^2}{4\phi^2} \\
0 &= \tilde{\rho}_E\left(-\frac{2}{\phi(1+\phi^2-2\rho_T\phi)^2}\frac{\phi^2(1-\rho_T^2)}{n-3} - \frac{(1-\phi^2)^2}{2\phi(1+\phi^2-2\rho_T\phi)^2}\right) \\
&\quad + \frac{1+\phi^2}{\phi^2(1+\phi^2-2\rho_T\phi)^2}\frac{\phi^2(1-\rho_T^2)}{n-3} + \frac{(1+\phi^2)(1-\rho_T\phi)^2}{\phi^2(1+\phi^2-2\rho_T\phi)^2} + \frac{1}{\phi^2} - \frac{2(1-\rho_T\phi)}{\phi^2(1+\phi^2-2\rho_T\phi)} - \frac{1+\phi^2}{4\phi^2} \\
0 &= \tilde{\rho}_E\left(-\frac{4\phi^2(1-\rho_T^2)}{n-3} - (1-\phi^2)^2\right) + \frac{2}{\phi}\left(\frac{\phi^2(1+\phi^2)(1-\rho_T^2)}{n-3} + (1+\phi^2)(1-\rho_T\phi)^2 + (1+\phi^2-2\rho_T\phi)^2\right. \\
&\quad \left. -2(1-\rho_T\phi)(1+\phi^2-2\rho_T\phi) - \tfrac{1}{4}(1+\phi^2)(1+\phi^2-2\rho_T\phi)^2\right) \\
0 &= \tilde{\rho}E\left(-\tfrac{4\phi^2(1-\rho_T^2)}{n-3} - (1-\phi^2)^2\right) + \frac{2}{\phi}\left(\frac{\phi^2(1+\phi^2)(1-\rho_T^2)}{n-3} + \phi^2(\phi-\rho)^2 + \phi^2(1-\rho_T\phi)^2 - \frac{1}{4}(1+\phi^2)(1+\phi^2-2\rho_T\phi)^2\right) \\
0 &= \tilde{\rho}E\left(-\tfrac{4\phi^2(1-\rho_T^2)}{n-3} - (1-\phi^2)^2\right) + \tfrac{1-\phi^2}{2\phi}\left(\tfrac{4\phi^2(1+\phi^2)(1-\rho_T^2)}{(1-\phi^2)(n-3)} + \phi^4 - 4\rho_T\phi^3 + 4\rho_T\phi - 1\right) \\
0 &= \tilde{\rho}E(-4\phi^2(1+\phi^2) - (1-\phi^2)^2(n-3)) + \tfrac{1}{2\phi}(8\phi^3(1+\phi^2)(1-\rho_T^2) + 2\phi(1-\phi^2)(\phi^4 - 4\rho_T\phi^3 + 4\rho_T\phi - 1)(-3)) \\
\rho &= \frac{4\phi^2(1+\phi^2)(1-\rho_T^2) + (1-\phi^2)(\phi^4 - 4\rho_T\phi^3 + 4\rho_T\phi - 1)(n-3)}{4\phi^2(1+\phi^2) + (1-\phi^2)^2(n-3)}
\end{aligned}
$$

Appendix D. Derivation of Eq. (7)

$$0=\xi\text{ow}\left(n,\phi_T,\tilde{\phi}_E,\rho,\rho\right)-\xi_{SA}\left(\tilde{\phi}_E,\rho\right)$$

$$0=\frac{1+\tilde{\phi}_E^2-2\rho\tilde{\phi}_E}{\left(1+\phi_T^2-2\rho\phi_T\right)^2}\left(\frac{\phi_T^2(1-\rho^2)}{n-3}+(1-\rho\phi_T)^2\right)+1-\frac{2(1-\rho\phi_T)\left(1-\rho\tilde{\phi}_E\right)}{1+\phi_T^2-2\rho\phi_T}-\frac{1}{4}\left(1+\tilde{\phi}_E^2+2\rho\tilde{\phi}_E\right)$$

$$0=\left(1+\tilde{\phi}_E^2-2\rho\tilde{\phi}_E\right)\left(\frac{\phi_T^2(1-\rho^2)}{n-3}+(1-\rho\phi_T)^2\right)+\left(1+\phi_T^2-2\rho\phi_T\right)^2-2(1-\rho\phi_T)\left(1-\rho\tilde{\phi}_E\right)\left(1+\phi_T^2-2\rho\phi_T\right)-\frac{1}{4}\left(1+\tilde{\phi}_E^2+2\rho\tilde{\phi}_E\right)\left(1+\phi_T^2-2\rho\phi_T\right)^2$$

$$0=\tilde{\phi}_E^2\left(\frac{\phi^2(1-\rho^2)}{n-3}+(1-\rho\phi_T)^2-\frac{1}{4}\left(1+\phi_T^2-2\rho\phi\right)^2\right)+\tilde{\phi}_E\left(-2\rho\frac{\phi^2(1-\rho^2)}{n-3}-2\rho(1-\rho\phi_T)^2+2\rho(1-\rho\phi_T)\left(1+\phi_T^2-2\rho\phi_T\right)-\frac{1}{2}\rho\left(1+\phi_T^2-2\rho\phi_T\right)^2\right)$$
$$+\frac{\phi^2(1-\rho^2)}{n-3}+(1-\rho\phi_T)^2\quad+\left(1+\phi_T^2-2\rho\phi\right)^2-2\rho(1-\rho\phi_T)\left(1+\phi_T^2-2\rho\phi_T\right)-\frac{1}{4}\left(1+\phi_T^2-2\rho\phi\right)^2$$

$$0=\tilde{\phi}_E^2\left(\frac{\phi^2(1-\rho^2)}{n-3}+\frac{1}{4}\left(3+\phi_T^2-4\rho\phi_T\right)\left(1-\phi_T^2\right)\right)\quad-2\tilde{\phi}E\rho\left(\frac{\phi^2(1-\rho^2)}{n-3}+\frac{1}{4}\left(1-\phi_T^2\right)^2\right)\quad+\frac{\phi^2(1-\rho^2)}{n-3}+\frac{1}{4}\left(\phi_T^2-1\right)\left(1+3\phi_T^2-4\rho\phi_T\right)$$

$$0=\tilde{\phi}_E^2\left(\phi^2(1-\rho^2)+\frac{1}{4}\left(3+\phi_T^2-4\rho\phi_T\right)\left(1-\phi_T^2\right)(n-3)\right)-2\tilde{\phi}_E\rho\left(\phi^2(1-\rho^2)+\frac{1}{4}\left(\phi_T^2-1\right)^2(n-3)\right)+\phi^2(1-\rho^2)+\frac{1}{4}\left(\phi_T^2-1\right)\left(1+3\phi_T^2-4\rho\phi_T\right)(n-3)$$

$$0=\tilde{\phi}_E^2\left(\psi-\frac{1}{4}\eta_1 m\right)-2\tilde{\phi}E\rho\left(\psi+\frac{1}{4}\left(\phi_T^2-1\right)m\right)+\phi^2(1-\rho^2)+\frac{1}{4}\eta_2 m$$

$$\tilde{\phi}_E=\frac{4\rho\psi+\rho\left(\phi_T^2-1\right)m\pm\sqrt{\rho^2\left(4\psi+\left(1-\phi_T^2\right)m\right)^2-(4\psi-\eta_1 m)(4\psi+\eta_2 m)}}{4\psi-\eta_1 m}$$

References

Armstrong, J. S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research*, *68*(8), 1717–1731.

Bates, J., & Granger, C. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research*, *68*(8), 1772–1784.

Bunn, D. W. (1985). Statistical efficiency in the linear combination of forecasts. *International Journal of Forecasting*, *1*(2), 151–163.

Claeskens, G., Magnus, J., Vasnev, A., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Clemen, R., & Winkler, R. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, *4*(1), 39–46.

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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*(1), 190–204.

Dickinson, J. P. (1973). Some statistical results in the combination of forecasts. *Journal of the Operational Research Society*, *24*(2), 253–260.

Diebold, F., & Lopez, J. (1996). Forecast evaluation and combination. In G.S., M., C.R., & R. (Eds.), *Statistical methods in finance. Vol. 14 of handbook of statistics*. (pp. 241–268). Elsevier.

Diebold, F., & Pauly, P. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, *6*(1), 21–40.

Elliott, G. (2011). Averaging and the optimal combination of forecasts. *Tech. rep.*. San Diego: University of California.

Figlewski, S., & Urich, T. (1983). Optimal aggregation of money supply forecasts: Accuracy, profitability and market efficiency. *The Journal of Finance*, *38*(3), 695–710.

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692–1701.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121.

Graefe, A. (2015). Improving forecasts using equally weighted predictors. *Journal of Business Research*, *68*(8), 1792–1799.

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, *68*(8), 1678–1685.

Gupta, S., & Wilton, P. (1987). Combination of forecasts: An extension. *Management Science*, *33*(3), 356–372.

Hyndman, R. J. (2015). *forecast: Forecasting functions for time series and linear models. R package version 6.1*.

Kang, H. (1986). Unstable weights in the combination of forecasts. *Management Science*, *32*(6), 683–695.

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

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., ... Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153.

Miller, C., Clemen, R., & Winkler, R. (1992). The effect of nonstationarity on combined forecasts. *International Journal of Forecasting*, *7*(4), 515–529.

Reid, D. (1968). Combining three estimates of gross domestic product. *Economica*, 431–444.

Schmittlein, D., Kim, J., & Morrison, D. (1990). Combining forecasts: Operational adjustments to theoretically optimal rules. *Management Science*, *36*(9), 1044–1056.

Smith, J., & Wallis, K. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355.

Stock, J., & Watson, M. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430.

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, *1*. (pp. 135–196).

Winkler, R. L., & Clemen, R. T. (1992). Sensitivity of weights in combining forecasts. *Operations Research*, *40*(3), 609–614.

Woike, J. K., Hoffrage, U., & Petty, J. S. (2015). Picking profitable investments: The success of equal weighting in simulated venture capitalist decision making. *Journal of Business Research*, *68*(8), 1705–1716.