Received: 8 December 2022 | Revised: 10 June 2024 | Accepted: 21 July 2024

DOI: 10.1002/for.3178

RESEARCH ARTICLE

WILEY

Shapley-value-based forecast combination

Philip Hans Franses[1] | **Jiahui Zou**[2] | **Wendun Wang**[1]

[1]Econometric Institute, Erasmus School of Economics, Rotterdam, the Netherlands

[2]School of Statistics, Capital University of Economics and Business, Beijing, China

Correspondence
Philip Hans Franses, Econometric Institute, Erasmus School of Economics, POB 1738, NL-3000 DR Rotterdam, the Netherlands.
Email: franses@ese.eur.nl

Abstract

This paper puts forward a new and simple method to combine forecasts, which is particularly useful when the forecasts are strongly correlated. It is based on the Mincer Zarnowitz regression, and a subsequent determination using Shapley values of the weights of the forecasts in a new combination. For a stylized case, it is proved that such a Shapley-value-based combination improves upon an equal-weight combination. Simulation experiments and a detailed illustration show the merits of the Shapley-value-based forecast combination.

KEYWORDS

forecast combination, forecasting

1 | INTRODUCTION

The combination of forecasts is common practice. The idea of combining forecasts goes back a long time, and it is widely recognized that the seminal paper is Bates and Granger (1969). The frequently cited survey papers, Clemen (1989) and Timmermann (2006), have also contributed to the popularity of combining forecasts. In practice, it is often observed that combined forecasts outperform individual forecasts. Highlights of knowledge on forecast combinations are presented in Section 2.6 of Petropoulos et al. (2022), and a recent extensive review appears in Wang et al. (2023).

A key issue in combining forecasts concerns the determination of the weights of the individual forecasts in the combination. Although there are many methods around, like those based on the historical performance of the forecasts or based on time series features, it is frequently found that a simple arithmetic average of the forecasts (equal weights, EW) is hard to beat; see Claeskens et al. (2016) and Smith and Wallis (2009) for explanations.

The beneficial feature of taking the arithmetic average is of course that no weights must be estimated, whereas the potential disadvantage is not accounting for the relative quality of the individual forecasts. As an example, one may not want to include all forecasts before taking an arithmetic average. To meet this concern, there are methods for prior selection, see for example Diebold and Shin (2019). It can also be argued that it is the diversity of the forecasts that makes the combination work well; see Kang et al. (2022); hence, one may wish to preselect along that dimension.

Even though the literature on the combination of forecasts is vast, in this paper, a new method is proposed to estimate the weights. This new method is motivated by the fact that, in practice, a typical feature of individual forecasts is that they can be strongly correlated. Think of the surveys of macroeconomic forecasters, which lead to consensus forecasts, where these forecasters in part base their predictions on information commonly known to all, and where correlations between forecasts are close to 1. One may thus wish to somehow incorporate these correlations and consider a method that may come close to

This revised version: June 10, 2024

Data statement
Data used can be obtained from the authors upon request.

the arithmetic average while addressing the correlations. Note that when there is a strong correlation, the application of regression-based weights (RW) becomes problematic, as we will show in a small illustration in Section 2.

Thus motivated, this paper puts forward a new and simple forecast combination method that is based on Shapley values (Shapley weights, SW). These values take account of (potentially substantial) collinearity across forecasts. The method draws upon the Mincer and Zarnowitz (1969) regression and derives the Shapley values from all R^2 values of all combinations of forecasts. See Shapley (1953) for the original idea of Shapley values and for example Lipovetsky and Conklin (2001) and Aas et al. (2021) for recent detailed and general accounts.

Our paper proceeds as follows. In Section 2, we provide a small empirical illustration for two forecasts to sketch the setting, where we will see that RW is not adequate. In Section 3, a stylized situation is analyzed, and we theoretically show that combining correlated forecasts with equal weights is suboptimal to using Shapley values weights. Section 4 provides the methodology for K forecasts. Section 5 presents the results from an extensive Monte Carlo simulation. Section 6 provides a detailed empirical illustration to examine the relative merits of the Shapley-value-based forecast combination. The final section concludes with a discussion, with limitations and avenues for further research.

2 | A SIMPLE EMPIRICAL EXAMPLE

To introduce the new forecast combination method, consider the task of modeling and predicting the data depicted in Figure 1. These are the average departure delays per day (in minutes) for KLM Royal Dutch Airlines flights from Schiphol Airport Amsterdam, for the period January 01, 2015, to November 28, 2017.

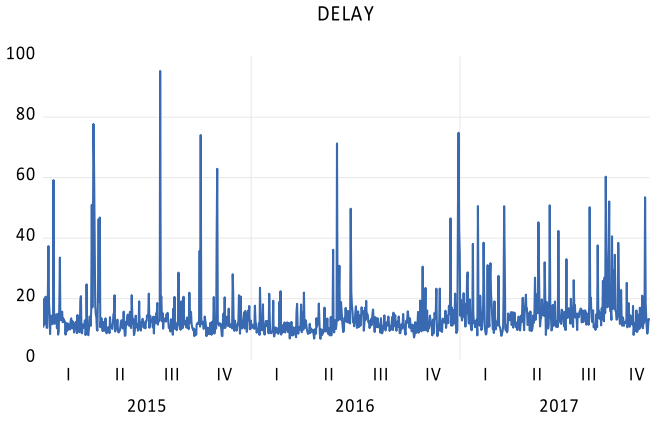

FIGURE 1 Average departure delays per day (in minutes) for KLM flights from Schiphol Airport Amsterdam, January 01, 2015, to November 28, 2017.

Without any prior testing or searching, we decompose the sample into a first sub-sample from January 01, 2015, to December 31, 2016, a second sub-sample from January 01, 2017, to June 30, 2017, and a third sub-sample from July 01, 2017, to November 28, 2017 (this is where the available full sample ends). The first sample is used to design the models. The second sample is used to find the weights in combined forecasts. The third sample is used for out-of-sample evaluation of forecast accuracy.

It is often found that delays can be predicted by past delays and by some seasonal factors; see for example, Khan et al. (2021) and Campanelli et al. (2016). Furthermore, bad weather conditions shall also impact delays. Examples of studies that show the relevance of weather conditions for airline delays are Kim (2016), McCrea et al. (2008), Ye et al. (2020), and Palin et al. (2016).

Denote y_t as the daily average departure delays of all flights. Denote $x_{1,t}$ as the average daily precipitation (in millimeters), $x_{2,t}$ as the average daily visibility (in miles), and $x_{3,t}$ as the average daily wind speed (in knots) and call m_t the 1/0 dummy variable for the Monday.

A simple linear time series regression model is

$$y_t = \delta_0 + \delta_1 m_t + \delta_2 y_{t-1} + \delta_3 y_{t-2} + \delta_4 x_{1,t} + \delta_5 x_{2,t} + \delta_6 x_{3,t} + \varepsilon_t. \quad (1)$$

Including the current values of $x_{1,t}, x_{2,t}$, and $x_{3,t}$ assumes that at the beginning of a new day, quite accurate forecasts are available for the weather conditions, and this is indeed usually the case. The estimated errors ε_t have no sign of residual autocorrelation. The ordinary least squares (OLS) based estimation results for the first sample are presented in the left-hand panel of Table 1,

TABLE 1 Estimation results for linear time series models. Estimated standard errors are in parentheses.

	Dependent variable			
	Delays		Log (delays)	
Intercept	21.599	(1.883)	2.109	(0.130)
Monday	1.619	(0.778)	0.074	(0.033)
Lag 1	0.234	(0.035)	0.245	(0.035)
Lag 2	0.098	(0.035)	0.115	(0.035)
Precipitation	0.025	(0.006)	0.001	(0.000)
Visibility	−2.699	(0.303)	−0.112	(0.013)
Wind speed	0.196	(0.070)	0.010	(0.003)

and clearly, all variables are statistically significant, even much below a 0.05 level. For the one-step-ahead forecasts for the second sub-sample, we have a root mean squared prediction error (RMSPE) of 7.497 for this linear model.

We also consider a linear model for the natural logs of delays (to dampen the variance in the dependent variable and to reduce the potential impact of influential observations), that is,

$$\log(y_t) = \delta_0 + \delta_1 m_t + \delta_2 \log(y_{t-1}) + \delta_3 \log(y_{t-2}) + \delta_4 x_{1,t} + \delta_5 x_{2,t} + \delta_6 x_{3,t} + \varepsilon_t. \quad (2)$$

The estimation results appear in the right-hand panel of Table 1, and again, all variables are statistically significant. The RMPSE for the 181 observations in the second sub-sample is 7.652.

To illustrate the forecast combination with two candidate forecasts, we consider the familiar Mincer Zarnowitz regression using the second sub-sample

$$y_t = \alpha + \beta_1 f_{1,t} + \beta_2 f_{2,t} + \varepsilon_t. \quad (3)$$

Upon using OLS, we obtain $\widehat{\beta}_1 = 1.428\,(0.730)$ and $\widehat{\beta}_2 = -1.097\,(1.047)$, with standard errors in parentheses. These parameter estimates and the standard errors immediately show the consequences of potentially strongly correlated forecasts, as the correlation here is 0.984. The sign of β_2 is odd, and its value seems insignificant. Notably, individual regressions $y_t = \alpha + \beta_1 f_{1,t} + \varepsilon_t$ and $y_t = \alpha + \beta_2 f_{2,t} + \varepsilon_t$ give $\widehat{\beta}_1 = 0.675\,(0.130)$ and $\widehat{\beta}_2 = 0.918\,(0.188)$, respectively. These latter outcomes show the differences in quality between the forecasts, even though the correlation is large.

2.1 | Estimating weights in case of two forecasts

Setting equal weights would give $\beta_1 = \beta_2 = 0.5$. To obtain Shapley-value-based weights, consider a general multiple regression model for a continuous variable y with just two variables, X_1 and X_2, that is,

$$y_t = \alpha + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \varepsilon_t, \quad (4)$$

and assume that OLS can be used to estimate the parameters. The overall fit of this model is measured by the R^2_{12}, where the subscript 12 indicates the inclusion of the two predictors. When the model would include only X_1, one gets R^2_1, and when it includes only X_2, one has R^2_2.

The contribution of X_1 to the overall fit, taking into account the correlation between X_1 and X_2, is now defined as

$$\frac{1}{2}R^2_1 + \frac{1}{2}(R^2_{12} - R^2_2), \quad (5)$$

and the associated Shapley value after scaling by R^2_{12} is

$$s_1 = \frac{\frac{1}{2}R^2_1 + \frac{1}{2}(R^2_{12} - R^2_2)}{R^2_{12}}. \quad (6)$$

Similarly, for X_2, we have

$$s_2 = \frac{\frac{1}{2}R^2_2 + \frac{1}{2}(R^2_{12} - R^2_1)}{R^2_{12}}, \quad (7)$$

Clearly, $s_1 + s_2 = 1$.

Returning to the illustration above, for the regression in (3), we obtain $R^2_{12} = 0.136$. When we consider the forecasts for the levels only (the model in (1)), the $R^2_1 = 0.130$, and when we only include the forecasts for the log levels (the model in (2)), the $R^2_2 = 0.117$. This implies that the Shapley values are $s_1 = 0.548$ and $s_2 = 0.452$. Note that these values are close to 0.5.

The RMSPE for the third out-of-sample period of the equally weighted combined forecast is 7.529, whereas that of the Shapley-weighted combined forecast is 7.522. Here, the Shapley-valued-combined forecasts are slightly better than the equal-weights combination.

3 | THEORY

This section considers a stylized case with two forecasts for which it is derived that, at least asymptotically, a Shapley-value-based forecast combination can be useful in the case of correlated forecasts. Detailed derivations appear in a technical web appendix.

Suppose there are two forecasts $f_{1,t}$ and $f_{2,t}$ for the same variable y_t (assume a zero mean for all variables for convenience). Consider the forecast combination $\beta_1 f_{1,t} + \beta_2 f_{2,t}$. An often-applied forecast combination adopts that $\beta_1 = \beta_2 = 0.5$. An alternative approach is to run the regression

$$y_t = \beta_1 f_{1,t} + \beta_2 f_{2,t} + \varepsilon_t, \quad (8)$$

and use OLS to retrieve the weights. A potential problem with OLS occurs when the forecasts are strongly

correlated. In what follows, we provide theoretical proof that in the case of such correlation, the R^2 of the regression of the variable on the forecast combination, that is,

$$y_t = \alpha_3(\beta_1 f_{1,t} + \beta_2 f_{2,t}) + \varepsilon_t, \tag{9}$$

is always larger when the weights β_1 and β_2 are based on the Shapley values than when they are set at $\beta_1 = \beta_2 = 0.5$.

Consider the following stylized case with

$$y_t = v_t \, with \, v_t \sim N(0,1),$$

and two unbiased forecasts for this variable as

$$f_{1,t} = y_t + w_t \, with \, w_t \sim N(0,1),$$

$$f_{2,t} = f_{1,t} + kz_t \, with \, z_t \sim N(0,1),$$

with v_t, w_t, z_t mutually independent and all independent from y_t, and with $k \geq 0$. When $k = 0$, there is a perfect correlation across the two forecasts, and when k is small, there is a strong correlation and when k gets larger, the correlation goes to 0.

For this stylized case, the relevant variances are covariances of the variables are

$$\text{var}(y_t) = 1,$$

$$\text{var}(f_{1,t}) = 2,$$

$$\text{var}(f_{2,t}) = 2 + k^2,$$

$$\text{cov}(f_{1,t}, f_{2,t}) = \text{var}(f_{1,t}) = 2,$$

$$\text{cov}(y_t, f_{1,t}) = \text{var}(y_t) = 1,$$

$$\text{cov}(y_t, f_{2,t}) = \text{var}(y_t) = 1,$$

and

$$corr(f_{1,t}, f_{2,t}) = \frac{cov(f_{1,t}, f_{2,t})}{\sqrt{var(f_{1,t}) var(f_{2,t})}} = \frac{2}{\sqrt{2(2+k^2)}} = \sqrt{\frac{2}{2+k^2}}. \tag{10}$$

When $k = 0$, this correlation is 1. When $k = 2$, this correlation is 0.577, when $k = 0.5$, this correlation is 0.943.

We consider five regression models, where the parameters are all estimated using OLS.

Model 1: $y_t = \alpha_1 f_{1,t} + \varepsilon_t$

OLS gives

$$\widehat{\alpha}_1 = \frac{cov(y_t, f_{1,t})}{var(f_{1,t})} = \frac{1}{2}.$$

To compute the fit of this model we replace α_1 by $\widehat{\alpha}_1$, and we obtain

$$R_1^2 = 1 - \frac{var(\varepsilon_t)}{var(y_t)} = 1 - var(\varepsilon_t) = 1 - \frac{1}{2} = \frac{1}{2}.$$

Model 2: $y_t = \alpha_2 f_{2,t} + \varepsilon_t$

OLS gives

$$\widehat{\alpha}_2 = \frac{cov(y_t, f_{2,t})}{var(f_{2,t})} = \frac{1}{2+k^2},$$

and

$$R_2^2 = 1 - \frac{var(\varepsilon_t)}{var(y_t)} = 1 - \left(1 - \frac{1}{2+k^2}\right) = \frac{1}{2+k^2}.$$

Model 3: $y_t = \alpha_1 f_{1,t} + \alpha_2 f_{2,t} + \varepsilon_t$

Upon using $(X'X)^{-1}X'y$ in matrix notation, we get

$$\begin{pmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} var(f_{1,t}) & cov(f_{1,t}, f_{2,t}) \\ cov(f_{1,t}, f_{2,t}) & var(f_{2,t}) \end{pmatrix}^{-1} \begin{pmatrix} cov(f_{1,t}, y_t) \\ cov(f_{2,t}, y_t) \end{pmatrix}.$$

With the numbers above this becomes

$$\begin{pmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2+k^2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

or

$$\begin{pmatrix} \widehat{\alpha}_1 \\ \widehat{\alpha}_2 \end{pmatrix} = \frac{1}{2k^2} \begin{pmatrix} 2+k^2 & -2 \\ -2 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}.$$

So, $\widehat{\alpha}_2 = 0$, and hence, Model 3 boils down to Model 1, irrespective of the value of $k > 0$.

Then, it also holds that the R^2 of the model with two variables and the R^2 of the model with just $f_{1,t}$ are equal, that is, $R_{12}^2 = R_1^2$. Hence, for Model 3 with the two forecasts, we have the Shapley values

$$s_1 = \frac{\frac{1}{2}R_1^2 + \frac{1}{2}\left(R_{12}^2 - R_2^2\right)}{R_{12}^2} = 1 - \frac{\frac{1}{2}R_2^2}{R_{12}^2} = 1 - R_2^2 = 1 - \frac{1}{2+k^2}. \tag{11}$$

And so,

$$s_2 = \frac{\frac{1}{2}R_2^2 + \frac{1}{2}\left(R_{12}^2 - R_1^2\right)}{R_{12}^2} = \frac{1}{2+k^2}. \tag{12}$$

Now, consider Model 4: $y_t = \alpha_3\left(\frac{1}{2}f_{1,t} + \frac{1}{2}f_{2,t}\right) + \varepsilon_t$
For this model, we can derive that

$$\widehat{\alpha}_3 = \frac{1}{2+\frac{1}{4}k^2}.$$

If $k=2$, then this is $\frac{1}{3}$. If $k=1$, then this is 0.444, and if $k=0.5$, then it becomes 0.485. The R^2 for this regression with equal-weighted forecasts is

$$R^2 = \frac{1}{2+\frac{1}{4}k^2}. \tag{13}$$

Finally, consider Model 5: $y_t = \alpha_3\left(s_1 f_{1,t} + s_2 f_{2,t}\right) + \varepsilon_t$, where the weights are the Shapley values. OLS gives

$$\widehat{\alpha}_3 = \frac{1}{2+\left(\frac{k}{2+k^2}\right)^2}.$$

If $k=2$, then this is 0.486. If $k=1$, then this is 0.474. And,

$$R^2 = \frac{1}{2+\left(\frac{k}{2+k^2}\right)^2}. \tag{14}$$

Finally, the R^2 based on the Shapley values in (14) is larger than the R^2 based on the equal valued weights regression in (13) if

$$\left(\frac{k}{2+k^2}\right)^2 < \frac{1}{4}k^2,$$

which is always the case for $k>0$.

One might now consider Principal Components Analysis (PCA) to retrieve proper weights, but this is not going to help. When we apply PCA to the correlation matrix

$$\begin{pmatrix} 1 & \sqrt{\frac{2}{2+k^2}} \\ \sqrt{\frac{2}{2+k^2}} & 1 \end{pmatrix}, \tag{15}$$

the eigenvalues are $\lambda_1 = 1 + \sqrt{\frac{2}{2+k^2}}$ and $\lambda_2 = 1 - \sqrt{\frac{2}{2+k^2}}$. The first eigenvalue is larger than 1. The eigenvectors are $(1,1)$ and $(1,-1)$, which again leads to equal weights.

The above analysis shows, at least for for a stylized case and asymptotically, that in the case of correlated forecasts, it is best to consider Shapley-value-based forecast combinations, instead of equal weights.

Shapley-value-based forecast combination for K forecasts

In the case of K forecasts, one can consider the Mincer Zarnowitz regression, that is,

$$y_t = \alpha + \beta_1 f_{1,t} + \beta_2 f_{2,t} + \ldots + \beta_K f_{K,t} + \varepsilon_t. \tag{16}$$

In that case, we can compute

$$SH_j = \sum_{\substack{S \subseteq K \\ j \in S}} \frac{(s-j)!(k-s)!}{k!}\left[R^2(S) - R_2(S\{j\})\right], \tag{17}$$

where $R^2(S)$ is the R^2 of the model with all forecasts in a set $S \subseteq K$, see for example Chantreuil and Trannoy (2011). Denoting $R^2_{12..K}$ as the R^2 with all forecasts, the Shapley weights in the new combined forecast

$$s_1 f_{1,t} + s_2 f_{2,t} + \ldots + s_K f_{K,t} \tag{18}$$

are found as

$$s_j = \frac{SH_j}{R^2_{12..K}}. \tag{19}$$

There are studies in the literature that propose alternative notations for the computation of Shapley values, see Lipovetsky and Conklin (2001) and recently Aas et al. (2021). To illustrate the above expression for $K=4$, consider the web appendix.

4 | MONTE CARLO SIMULATION

To illustrate the finite sample performance of the Shapley-value-based combination, we generate the data via the following model:

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1,2,...,n,$$

where $x_i = \left(1, x_{i1}, ..., x_{i(d-1)}\right)^T \sim N(0,V)$ with $V = \left(\rho^{|j-k|}\right)_{(d-1)\times(d-1)}$ and ε_i is drawn independently from a standard normal distribution. We generate each element of the coefficient vector from a uniform distribution, namely, $\beta_l \sim U(-1,1)$ for $i = 1,...,d$. In our experiments, we set $d = 12$, and the parameter ρ is chosen from $\{0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.93, 0.96, 0.99\}$ to mimic different degrees of correlation. We consider a sample size $n = 1400$, of which we use the first 300 observations to estimate the slope parameter β, the next 100 observations to compute the Shapley-value-based weights, and the final 1000 observations to evaluate the forecasts.

As we are particularly interested in the case where candidate forecasts are correlated with each other, we follow Hansen (2007) to construct the candidate models in a nested manner that first sorts the regressors based on the magnitude of their bivariate correlation with the dependent variable and then includes regressors sequentially. Our method to construct nested candidate models can be regarded as a model screening based on bivariate correlation in a similar spirit to the "sure independence screening" proposed by Fan and Lv (2008).

We compare the performance of the Shapley-value-based combination (SW) with two commonly used combination methods, namely, EW and regression-based weights (RW) computed by regressing the actual value on the candidate forecasts using OLS. We evaluate these methods using the Diebold Mariano (DM) test for the root mean squared forecast error (RMSFE), the mean absolute deviation (MAD), and the RMSFE itself, based on 1000 replications.

Table 2 presents the percentage of cases where a method performs the best in terms of DM tests, the average MAD, and RMSFE over the replications. The DM test shows that SW significantly outperforms the other two combination methods in most cases. The average MAD and RMSFE of SW are also generally lower than those of RW and EW. Interestingly, as the correlation among regressors ρ increases, leading to a stronger dependence among candidate models, the proportion of DM tests that SW performs the best increases. This result confirms the theoretical finding that the Shapley-value-based combination is particularly useful when combining highly correlated candidate forecasts.

5 | REAL DATA EXAMINATION

To further examine the potential usefulness of the Shapley-value-based forecast combination, we apply the proposed method to forecast macroeconomic variables, using the comprehensive dataset of FRED-MD monthly panel of US macroeconomic variables, constructed by McCracken and Ng (2016). It contains monthly observations from January 1987 to December 2021, namely, 420 time series observations.

We focus on forecasting real personal income (RPI) and use 10 popular determinants that cover output, employment, inflation, and financial markets indicators. Table 3 provides the variable description. The first column "T-code" shows which transformation is applied to the original variable z_t with the following lables: (1) no transformation, (2) $\triangle z_t$, (3) $\triangle \log(z_t)$, and (4) $\triangle^2 \log(z_t)$. The second column "Fred" gives the mnemonics in Fred, and the third column provides a short description.

TABLE 2 Simulation results.

	Best in DM test (%)			MAD			RMSFE		
ρ	SW	RW	EW	SW	RW	EW	SW	RW	EW
0.99	91.071	4.383	4.545	0.828	1.973	0.843	1.037	2.477	1.056
0.96	78.571	21.008	0.432	0.861	1.319	0.909	1.079	1.653	1.139
0.93	69.657	30.211	0.132	0.878	1.748	0.956	1.100	2.189	1.199
0.90	63.939	36.071	0.000	0.888	0.993	1.314	1.113	1.245	1.656
0.85	58.485	41.525	0.000	0.899	1.048	1.597	1.137	1.301	1.998
0.80	53.838	46.172	0.000	0.906	1.070	1.363	1.146	1.340	1.718
0.75	50.828	49.182	0.000	0.912	1.093	2.275	1.143	1.370	2.848
0.70	48.404	51.606	0.000	0.917	1.111	1.464	1.159	1.392	1.845
0.65	46.747	53.263	0.000	0.921	1.125	1.399	1.155	1.410	1.742
0.60	45.172	54.838	0.000	0.924	1.146	1.373	1.169	1.424	1.721

TABLE 3 Empirical application: Variable description.

T-code	Fred	Description
3	RPI	Real Personal Income
3	NDMANEMP	All Employees, Nondurable Goods
3	CLAIMSx	Initial claims
4	INVEST	Securities in Bank Credit, All Commercial Banks
3	CMRMTSPLx	Real Manufacturing and Trade Industries Sales
3	S&P 500	S&P's Common Stock Price Index: Composite
1	T5YFFM	5-Year Treasury C Minus FEDFUNDS
4	DTCTHFNM	Total Consumer Loans and Lease Outstanding
2	S&P div yield	S&P's Composite Common Stock: Dividend Yield
1	CES0600000007	Avg weekly hours: Goods producing
3	PAYEMS	All employees: Total nonfarm

Note: The first column "T-code" shows which transformation is applied to the original variable z_t with the following labels: (1) no transformation; (2) $\triangle z_t$, (3) $\triangle \log(z_t)$, and (4) $\triangle^2 \log(z_t)$. The second column "Fred" gives the mnemonics in Fred, and the third column provides a short description.

To facilitate computation, we employ a similar procedure to construct candidate models in a nested manner as in the simulation experiments. We split the entire time span into three subsamples, used for estimating the slope parameters, determining combination weights and evaluation, respectively. To examine how the performance varies across different sizes of training samples, we fix the size of the evaluation subsample as 20% of the entire periods but vary the training samples, such that the proportion of each subsample is as follows: (1) 0.65:0.15:0.2, (2) 0.7:0.1:0.2, (3) 0.75:0.05:0.2, and (4) 0.78:0.02:0.2.

Table 4 presents the values of ρ and Table 5 presents the DM test, MAD, and RMSFE of the three combination methods. The DM test compares each method with respect to SW, and a negative value indicates that SW outperforms its rival. We find that the performance of competing methods varies across different sizes of subsamples used to determine the combination weights. Particularly, when there are relatively sufficient observations to estimate the weights, say 15% of the entire sample, RW produces the most accurate combination, but its superiority to SW is not significant. As the subsamples for weight estimation decrease, the performance of RW deteriorates, whereas SW becomes the most accurate method when the weight estimation subsample is 10% and 5%. This result suggests that the requirement of samples for

TABLE 4 The values of ρ.

ρ	RPI	NDMANEMP	CLAIMSx	INVEST	CMRMTSPLx	S&P 500	T5YFFM	DTCTHFNM	S&P div yield	CES0600000007	PAYEMS
RPI	1.000	−0.832	0.172	0.959	0.994	0.941	0.189	0.960	−0.659	0.496	0.971
NDMANEMP	−0.832	1.000	−0.128	−0.925	−0.799	−0.841	−0.247	−0.920	0.589	−0.353	−0.687
CLAIMSx	0.172	−0.128	1.000	0.084	0.137	−0.045	0.203	0.124	0.312	−0.516	0.229
INVEST	0.959	−0.925	0.084	1.000	0.936	0.951	0.203	0.956	−0.629	0.517	0.868
CMRMTSPLx	0.994	−0.799	0.137	0.936	1.000	0.941	0.163	0.954	−0.688	0.507	0.979
S&P 500	0.941	−0.841	−0.045	0.951	0.941	1.000	0.141	0.923	−0.731	0.577	0.875
T5YFFM	0.189	−0.247	0.203	0.203	0.163	0.141	1.000	0.182	−0.215	0.027	0.166
DTCTHFNM	0.960	−0.920	0.124	0.956	0.954	0.923	0.182	1.000	−0.694	0.442	0.889
S&P div yield	−0.659	0.589	0.312	−0.629	−0.688	−0.731	−0.215	−0.694	1.000	−0.567	−0.623
CES0600000007	0.496	−0.353	−0.516	0.517	0.507	0.577	0.027	0.442	−0.567	1.000	0.471
PAYEMS	0.971	−0.687	0.229	0.868	0.979	0.875	0.166	0.889	−0.623	0.471	1.000

TABLE 5 Empirical application: Forecasting performance

Sample proportion		SW	RW	EW
	DM test	NA	16.78	−4.88***
0.65:0.15:0.20	MAD	5837.66	5942.62	3475.36
	RMSFE	5939.62	6013.65	3507.10
	DM test	NA	−44.13***	−92.52***
0.70:0.10:0.20	MAD	5488.88	5858.79	9551.65
	RMSFE	5578.78	5931.83	9566.91
	DM test	NA	−70.40***	−25.24***
0.75:0.05:0.20	MAD	2958.24	3290.33	6884.78
	RMSFE	3045.49	3360.68	6906.83
	DM test	NA	0.61	21.05
0.78:0.02:0.20	MAD	2959.17	2791.94	2264.98
	RMSFE	3025.34	2860.84	2956.86

Note: The DM test provides the test statistics comparing each method with SW. A negative value shows that SW outperforms the rival.
***Significance of test at 1%.
**Significance of test at 5%.
*Significance of test at 10%.

weight determination is generally higher for RW than for SW. When we further reduce the weight estimation subsamples to 2%, neither of the two weight estimation methods outperforms the simple average EW, which does not require any weight estimation, although the difference between SW and EW is not significant. This finding is in line with the bias-variance tradeoff in the forecast combination puzzle that the benefit of using optimal weights may be offset by the estimation uncertainty of unknown weights, especially when large estimation noise is present. See Claeskens et al. (2016) for a comprehensive theoretical discussion. Overall, we find that SW provides rather accurate combined forecasts. Even in the cases where SW is not the best, it is often the second-best method and not significantly worse than the best one.

6 | CONCLUSION

We proposed a simple way of combining forecasts, which is particularly useful when the forecasts are strongly correlated. We saw some potential gain in forecast accuracy in a simple case study. For a stylized case, it was proved that indeed Shapley weights improve upon equal weights when combining correlated forecasts. Simulation experiments and a detailed empirical illustration showed that the Shapley-value-based forecast combinations do have merit.

When there are many forecasts to combine, the number of computations can become very large. Preselection methods can then be relevant. A fruitful other avenue for future research can be simpler computational methods, which can be based on nested or sequential versions of Shapley value regression; see Shorrocks (2013).

ACKNOWLEDGMENTS

Thanks are due to an anonymous reviewer, Dick van Dijk, and Christiaan Heij for extensive comments and to Chen (Doris) Xiong for her help with collecting the data. Data and detailed estimation results can be obtained upon request.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Philip Hans Franses https://orcid.org/0000-0002-2364-7777

REFERENCES

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual prediction when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468. https://doi.org/10.1057/jors.1969.103

Campanelli, B., Fleurquin, P., Arranz, A., Etxebarria, I., Ciruelos, C., Equiluz, V., & Ramasco, J. J. (2016). Comparing the modeling of delay propagation in the US and European air

traffic networks. *Journal of Air Transport Management*, *56*(9), 12–18. https://doi.org/10.1016/j.jairtraman.2016.03.017

Chantreuil, F., & Trannoy, A. (2011). Inequality decomposition values. *Annals of Economics and Statistics*, *101/102*(January/-June), 13–36. https://doi.org/10.2307/41615472

Claeskens, G., Magnus, J. R., Vasney, A. L., & Wang, W. (2016). The forecasting combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762. https://doi.org/10.1016/j.ijforecast.2015.12.005

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting*, *35*(4), 1679–1691. https://doi.org/10.1016/j.ijforecast.2018.09.006

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, *70*(5), 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Hansen, B. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175–1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x

Kang, Y., Cao, W., Petropoulos, F., & Li, F. (2022). Forecast with forecasts: Diversity matters. *European Journal of Operational Research*, *301*(1), 180–190. https://doi.org/10.1016/j.ejor.2021.10.024

Khan, W. A., Ma, H.-L., Chung, S.-H., & Wen, X. (2021). Hierarchical integrated machine learning model for predicting flight departure delays and duration in series. *Transportation Research Part C: Emerging Technologies*, *129*(August), 103225. https://doi.org/10.1016/j.trc.2021.103225

Kim, M. S. (2016). Analysis of short-term forecasting for flight arrival time. *Journal of Air Transport Management*, *52*(April), 35–41. https://doi.org/10.1016/j.jairtraman.2015.12.002

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*(4), 319–330. https://doi.org/10.1002/asmb.446

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589. https://doi.org/10.1080/07350015.2015.1086655

McCrea, M., Sherali, H. D., & Tranic, A. A. (2008). A probabilistic framework for weather-based rerouting and delay estimations within an airspace planning model. *Transportation Research Part C: Emerging Technologies*, *16*(4), 410–431. https://doi.org/10.1016/j.trc.2007.09.001

Mincer, J. A., & Zarnowitz, V. (1969). The evaluation of economic forecasts. In J. Mincer (Ed.), *Economic forecasts and expectations: Analysis of forecasting behavior and performance* (pp. 3–46). National Bureau of Economic Research, Inc.

Palin, E. J., Scaife, A. A., Wallace, E., Pope, E. C. D., Arribas, A., & Brookshaw, A. (2016). Skillful seasonal forecasts of winter disruption to the U.K. transport system. *Journal of Applied Meteorology and Climatology*, *55*(2), 325–344. https://doi.org/10.1175/JAMC-D-15-0102.1

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taleb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Luiz, F., Oliveira, C., De Baets, S., ... Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Shapley, L. (1953). A value for N-person games. In *Contributions to the theory of games, volume 2* (pp. 307–317). RAND Corporation. https://doi.org/10.1515/9781400881970-018

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *Journal of Economic Inequality*, *11*(1), 99–126. https://doi.org/10.1007/s10888-011-9214-z

Smith, J. P., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355. https://doi.org/10.1111/j.1468-0084.2008.00541.x

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1) (pp. 135–196). Elsevier. https://doi.org/10.1016/S1574-0706(05)01004-9

Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A methodology for predicting aggregate flight departure delays in airports based on supervised learning. *Sustainability*, *12*(7), 2749. https://doi.org/10.3390/su12072749

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, *39*(4), 1518–1547. https://doi.org/10.1016/j.ijforecast.2022.11.005

AUTHOR BIOGRAPHIES

Philip Hans Franses is a professor of Applied Econometrics. His research interests concern forecasting, marketing research, empirical finance, and time series econometrics.

Dr. Jiahui Zou is an associate professor at Capital University of Economics and Business, China. He obtained his doctorate from the University of Chinese Academy of Sciences in China. He was also a visiting scholar at City University of Hong Kong. His research interests include model averaging, optimal subsampling, and machine learning.

Dr. Wendun Wang is an associate professor at Econometric Institute, Erasmus University Rotterdam. His primary interests are in forecasting, panel data models, model selection and averaging.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Franses, P. H., Zou, J., & Wang, W. (2024). Shapley-value-based forecast combination. *Journal of Forecasting*, 1–9. https://doi.org/10.1002/for.3178