Research Program on Forecasting

A Nonparametric Approach to Identifying a Subset of Forecasters that Outperforms the Simple Average

Constantin Bürgi
The George Washington University
cburgi@gwu.edu

Tara M. Sinclair
The George Washington University
tsinc@gwu.edu

RPF Working Paper No. 2015-006
http://www.gwu.edu/~forcpgm/2015-006.pdf

December 31, 2015

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www.gwu.edu/~forcpgm

A Nonparametric Approach to Identifying a Subset of Forecasters that Outperforms the Simple Average *

Constantin Bürgi and Tara M. Sinclair †

December 31, 2015

JEL: C22, C52, C53

Keywords: Forecast combination; Forecast evaluation; Multiple model comparisons; Real-time data; Survey of Professional Forecasters

Abstract

Empirical studies in the forecast combination literature have shown that it is notoriously difficult to improve upon the simple average despite the availability of optimal combination weights. In particular, historical performance-based combination approaches do not select forecasters that improve upon the simple average going forward. This paper shows that this is due to the high correlation among forecasters, which only by chance causes some individuals to have lower root mean squared errors (RMSE) than the simple average. We introduce a new nonparametric approach to eliminate forecasters who perform well based purely on chance as well as poor performers. This leaves a subset of forecasters with better performance in subsequent periods. It improves upon the simple average in the SPF for bond yields where some forecasters may be more likely to have specialized knowledge.

*We would like to thank Neil Ericsson, Tatevik Sekhposyan, Herman Stekler and Benjamin Williams for their valuable comments and support. We would also like to thank participants in the Federal Forecasters Conference, the Georgetown Center for Economic Research (GCER) conference, and the GWU SAGE seminar series.

† *cburgi@gwu.edu, tsinc@gwu.edu The George Washington University, 2115 G Street, NW, # 340, Washington, DC 20052*

1 Introduction

Since Bates and Granger (1969), it has become well established that combinations of forecasts perform better than individual forecasts. In particular, they introduced optimal weights to combine individual foresters based on the variances of and covariances between individual forecast errors. However, empirical studies summarized by Clemen (1989) and Timmermann (2006) showed several drawbacks of using optimal weights. In particular, it appears to be quite difficult to improve upon the simple average of individual forecasts using optimal weights. This result is often attributed to the estimation of these weights. In addition, other combination approaches like past performance do not select forecasters that improve upon the simple average going forward. A more recent study of the variables in the ECB Survey of Professional Forecasters by Genre et al. (2013) found similar results using a wide array of combination approaches.[1]

Due to these findings, it is not surprising that many surveys collecting forecasts report the simple average of forecasts as the benchmark.[2] However, as Blix et al. (2001) showed for Consensus Economics forecasts, there are individuals that consistently have smaller root mean squared errors (RMSE) than this simple average even over extensive periods.

The existence of individual forecasters who outperform the simple average based on RMSE immediately leads to the question: Why is it difficult to find the best forecasters and to improve upon the simple average? This paper examines this issue using the Survey of Professional Forecasters (SPF) conducted by the Federal Reserve Bank of Philadelphia, and shows that it is linked to the correlation among forecasters. A high correlation among forecast errors can lead to many individual forecasters outperforming the simple average merely by chance. This immediately implies that selecting the best forecaster based

[1] Note that we are focusing in this paper on point forecasts. Jore et al. (2010) and Kascha and Ravazzolo (2010) show that the estimation of time-varying weights provide gains over equal weights for density forecasts. Kenny et al. (2015) evaluate macroeconomic density forecasts and show that for many forecasters it is possible to systematically improve their forecast performance. Lahiri et al. (2015) show that the density forecasts can be used to create a subset of forecasts that improve over the simple average.

[2] In our analysis here we focus on the mean of the forecasters, which is typically used in the forecast combination literature. For the SPF the median is what is often used. Our results are robust to using the median instead and the results are available from the authors upon request.

on recent RMSE performance might not always lead to a superior forecast afterwards. The forecaster's past good performance might just have been due to pure chance. Taking this finding into account, a new approach is introduced that eliminates a greater number of forecasters that outperform the average due to chance. This leaves a subset of forecasters who are more likely to outperform the average in subsequent periods. This new approach is subsequently applied to the SPF forecasts of CPI, unemployment and the bond yield. This approach yields statistically significant improvements upon the simple average for several forecasts, particularly for the bond yield where some forecasters may have better information and/or pay more attention than other individuals.

The layout of the remainder of the paper is as follows. In section 2 we explore the role of pairwise correlation with an example of individual foreasts of CPI inflation from the SPF and present a simulation under simple assumptions to show the percentage of forecasters that beat the simple average by chance. Next, in section 3 we present our new nonparametric approach as well as the popular alternative approaches. Section 4 details our application to SPF data where we compare the new approach to the alternatives and examine the reasons why the new approach works particularly well for bond yield forecasts. Finally, section 5 concludes.

2 The Role of Correlation

2.1 Pairwise Correlation

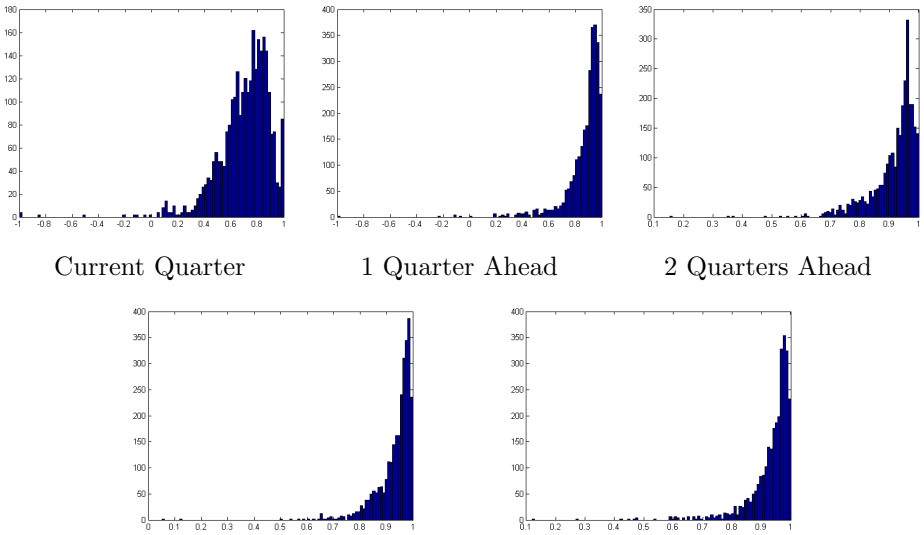
As an example, we begin by considering individual forecasts for the quarter-on-quarter CPI percent change from the Survey of Professional Forecasters (SPF) from the Federal Reserve Bank of Philadelphia for the period 1992Q1 - 2013Q3.

Around 40% of the CPI inflation forecasters with at least 10 forecasts have lower RMSE than the simple average at every horizon. This result is consistent with Blix et al. (2001), who showed that there are individuals that beat the simple average for extended periods of time within the Consensus Economics forecasts. At the same time, the histograms of the pairwise correlation matrix of SPF CPI forecasters with at least 20 forecasts[3] in this period show the distribution is very skewed towards high correlations for all five horizons (Figure

[3]The higher threshold here is necessary to have forecasters overlap more over the entire sample.

1). It is also noteworthy that longer horizons are more skewed than shorter horizons. The average pairwise correlation for the current quarter forecasters is 0.65 and around 0.8-0.85 for the other horizons. This high correlation is a well known fact as shown by Clark and McCracken (2010) for example. Given the distribution of correlations shown in Figure 1, the simple average is not an optimal weighting of forecasters. As Timmermann (2006) highlighted, equal weight cannot be optimal if the pairwise correlations among forecasters is not the same.

Figure 1: Distribution of pairwise correlations among CPI forecasters



Current Quarter | 1 Quarter Ahead | 2 Quarters Ahead

3 Quarters Ahead | 4 Quarters Ahead

2.2 Simulation

To illustrate the relationship between the high correlation among individual forecasters and by chance outperformance of the simple average, assume that forecast errors take the simple form

$$\nu_{it} = \delta\gamma_t + (1-\delta)\varepsilon_{it}, \tag{1}$$

where the individual forecast errors ν_{it}, are the weighted sum of a common forecast error γ_t and an error ε_{it}, which are iid with $(0, \sigma_\gamma^2 < \infty)$ and $(0, \sigma_\varepsilon^2 < \infty)$,

respectively[4]. This forecast error form is very similar to Davies and Lahiri (1995). By construction, the variance of ν_{it} for $t \to \infty$ is given by

$$\delta^2\sigma_\gamma^2 + (1-\delta)^2\sigma_\varepsilon^2, \tag{2}$$

while the asymptotic variance of the simple average is

$$\delta^2\sigma_\gamma^2 < \delta^2\sigma_\gamma^2 + (1-\delta)^2\sigma_\varepsilon^2. \tag{3}$$

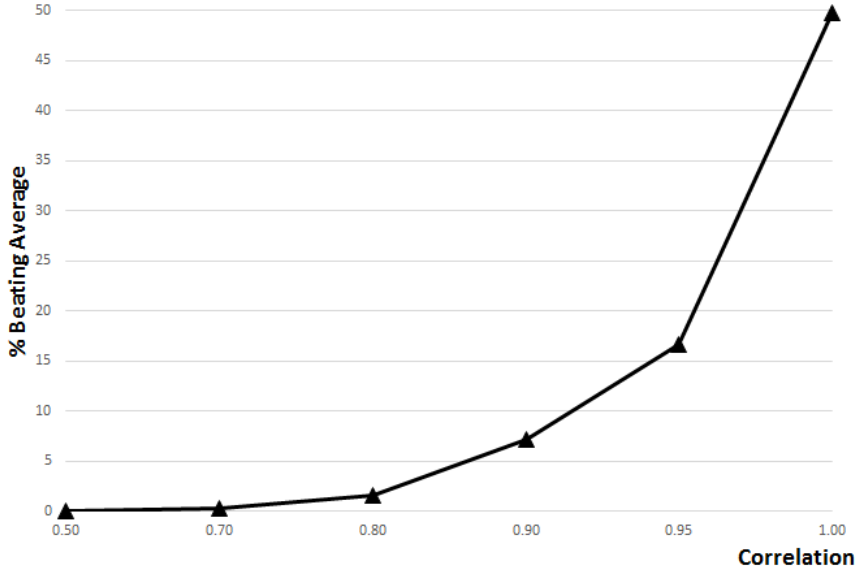
Under these assumptions no forecaster has any idiosyncratic information and thus in infinite samples there are never any individuals that beat the simple average for any correlation $\rho < 1$ and $\sigma_\varepsilon^2 > 0$. However, this result does not hold in finite samples for high correlations due to the Law of Large Numbers. In particular, there is a probability that a forecaster will have a lower RMSE than the simple average by chance, even over several periods. This probability is higher when the number of forecast periods is small, and thus the percentage of forecasters beating the simple average by chance becomes more sizeable for shorter samples. As the gains from averaging are smaller at high correlations (high δ), there is also a positive relationship between the correlation among forecast errors and the percentage of forecasters beating the simple average by chance.

A simulation is conducted to capture the extent to which individual forecasters should be expected to beat the simple average by chance in a finite sample[5]. To roughly match the dimensions of the SPF dataset used, 54 forecasters will be simulated over 80 periods for various values of δ which map to different pairwise correlations. Both ν_{it} and γ_t are assumed to be independently normally (0,1) distributed and the simulation is run 10,000 times. Figure 2 shows that as the correlation approaches 1, around 50% of forecasters should beat the simple average by chance over the entire sample based on RMSE. This implies that approaches that use past performance based on RMSE to select a subset of forecasters are likely to include several forecasters into the subset that performed well by chance.

[4]This result would not change, if forecasters had different variances but are still independent from each other.

[5]We do not assume different regimes or more complicated environments in this paper. Zhao (2015) conducts simulations for a number of different environments for a range of approaches.

Figure 2: Simulated percentage of forecasters beating the simple average by chance for various pairwise correlations with 80 periods



3 Approaches to Outperform the Simple Average

3.1 Our New Nonparametric Approach

Based on our simulation above, even over 80 periods, there is a sizeable percentage of individual forecasters who have a lower RMSE by chance when there are high pairwise correlations among them. To improve upon the simple average, it is therefore necessary to find a selection approach that identifies fewer forecasters that beat the simple average by chance. In addition to this property, the selection approach should be able to deal with forecasters (re-)entering and exiting the survey. It should be a relative criterion to avoid issues that are associated with changing forecast variance.

We construct a performance rank based on a nonparametric real-time approach to obtain a subset of best forecasters similar to Stekler (1987) and Batchelor (1990).[6] To construct the subset for a given period, only information about

[6]This approach is similar to the method used by Gamber et al. (2014). They want to

the past forecast performance available to all forecasters at that point in time is used. In particular, for each forecaster a new variable is calculated that takes value 1 in a given period, if that forecaster has a lower squared error in that period than the simple average and 0 otherwise. This means that the forecaster is ranked a better forecaster this period than the average. The average of this binary variable over time gives the percentage share of times each individual forecaster has beaten the simple average in the past. If a forecaster has beaten the simple average more often than a certain percentage threshold p, that forecaster is included in the subset for the next forecasting period. To obtain a single forecast for the set of best forecasters, the simple average over all forecasters in the subset is calculated.

As this nonparametric selection criterion is relative to the simple average for each period, it should be less influenced by periods that are easier (harder) to forecast and hence would have had smaller (larger) forecast errors. This is in stark contrast to absolute criteria like the RMSE. Also, this rank based approach does not depend on the magnitude of the difference between the individual errors and that of the simple average. This avoids the situation where a forecaster that is much better in one or two periods by chance is weighted too heavily. These properties allow this recursively employed approach to easily handle gaps in the dataset[7].

A second simulation is conducted based on the new nonparametric approach. Figure 3 shows the simulated percentage of forecasters beating the simple average by chance based on this rank based approach for various values of δ and thus pairwise correlations among forecast errors and the thresholds $p = 50, p = 52.5$ and $p = 55$[8]. The simulation shows that the new approach would indeed se-

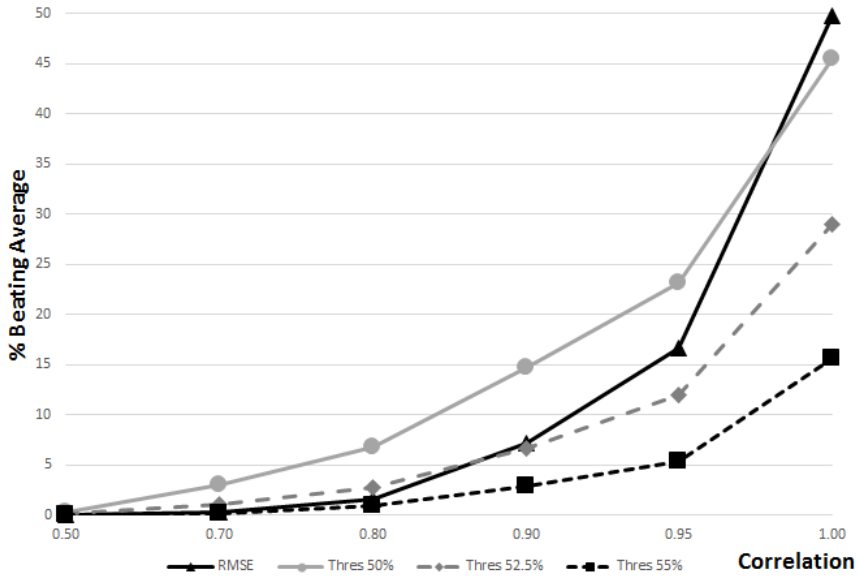
compare a subset of the best forecasters in the SPF to the Federal Reserve forecasts, which is very different from our objective. To obtain the subset, they used a static RMSE percentile performance threshold in the first stage instead of the rank based performance threshold relative to the mean we use. The second selection stage is the same as ours. While the first stage of their approach will guarantee to have a subset in every period unlike ours, their approach does not require forecasters in the subset to perform well relative to the mean. Assuming the performance remains the same before and after the selection into the subset, their approach might not perform better than the mean for certain distributions of forecasters.

[7] As an alternative to this nonparametric threshold approach, one could also use an estimated approach like impulse indicator saturation as described in Ericsson and Reisman (2012), but this would require more modifications to the data set due to the large gaps.

[8] While it might be desirable to increase the threshold until the percentage of individual forecasters beating it by chance reaches a very small number, this would also decrease the

lect fewer forecasters by chance than selecting forecasters that improve over the simple average based on RMSE for high correlations.

Figure 3: Simulated percentage of forecasters beating the simple average by chance for various pairwise correlations with 80 periods and different thresholds for the nonparametric approach



3.2 Alternative Approaches

Before applying the approach described above to our data set, we will discuss some alternative approaches of creating subsets used in the literature to be able to assess the performance of those approaches compared to the new approach. As highlighted by Lahiri et al. (2014), Genre et al. (2013), Poncela et al. (2011) or Conflitti et al. (2015) it is important to take into account the gaps in the surveys for the alternative approaches[9]. While many of the above authors reduce the number of forecasters to the forecasters who most consistently contributed forecasts to the survey, this is not possible for the 20 year span of the SPF data used, as it would discard more than 80% of the forecasters. At the same time,

probability of detecting moderately better forecasters creating a trade off between the two. We will evaluate this in the next section.

[9] As specified above, the new approach is much less affected by gaps in the survey

we do not want forecasters that only contributed for very few periods to impact our results. We only include forecasters once they made 10 forecasts, similar to the threshold suggested by D'Agostino et al. (2012).

As comparison with our results, we will consider five alternative approaches: median, trimmed mean (5% Trim), the single recent best forecaster (RB), the equal weight of the five recent best forecasters (RB5) and the weighted average of the inverse MSE as used in Capistrán and Timmermann (2009) (invMSE). We do not consider regression based approaches, because due to the long time frame, some forecasters do not have any overlap in forecasts and because there is a large number of forecasters relative to the number of periods. Unlike the other approaches, the median and the trimmed mean do not require any information about past performance. For this analysis, we choose a trimmed mean that drops the 5% of forecasters that have the highest and lowest forecast each period.

The other three approaches require us to evaluate past performance recursively, taking into account that forecasts for the SPF are collected in the middle of the month. That is, by the time the current quarter forecasts (H0) are surveyed, the individual performance in the previous quarter can be calculated for all variables. This information is then used to construct the subsets. For longer horizons (H1-H4), the information about the performance of individual forecaster in the previous quarter is not available. This leads to an information lag about forecasting performance, which increases with the horizon.

We opt for a rolling window approach[10], where we evaluate the forecast performance over the past 15 available quarters, provided forecasters contributed at least 10 forecasts to the survey during that period. This 10 forecast requirement means that the forecast evaluation can only start with the Q4 1994 for current quarter forecasts. Given that there are gaps in the data, there might be issues as some time periods have unanticipated outliers leading to larger MSE than other periods. This could lead to a forecaster looking much better than other forecasters, because he did not contribute at that date. Due to this, when comparing performance, we divide by the MSE of the simple average, which should avoid this issue.[11]

Once we have calculated the relative MSE for all forecasters and periods, we can create the respective subsets and compare the RMSE performance of each

[10] The three approaches tend to perform worse on an expanding window basis

[11] As a crosscheck, we also compared MSE directly, which led to a worse performance of the subsets alternative approaches.

subset relative to the RMSE of the simple average and to the other measures.

For the new nonparametric approach proposed in this paper, the threshold to include a given forecaster based on the historic performance needs to be specified. To ensure that the subset includes at least a few forecasters every period and only about 30 % of forecasters that beat the simple average by chance for very high correlations based on the simulation results in Figure 3 above, the threshold is set at $p = 52.5\%$[12].

4 Application

4.1 Data

For our application we consider the individual forecasts conributed to the Survey of Professional Forecasters by the Federal Reserve Bank of Philadelphia for the period 1992Q1 - 2013Q3. We focus on three key variables in the survey: quarter-on-quarter CPI percent change, average quarterly unemployment rate, and average quarterly 10-year treasury bond yield. These variables are chosen because the values of the published first release are very little revised and can be used as the actual values in a forecast evaluation. When series such as GDP are revised, it is necessary to determine which vintage of the data are to be used. This problem is avoided here. In the SPF, bond yield forecasts only start in 1992, which is the date chosen for all three series. The sample ends in Q1 2013 and includes the Great Recession period. For the actual series, the actual values from the SPF are used.

4.2 Results

The overall RMSE of the various subsets of forecasters over the three variables and five horizons (current quarter forecast (H0) to the four-quarter-ahead prediction (H4)) are shown in Table 1, all divided by the RMSE of the simple average. This means that if an approach has a value below 1, it has a lower RMSE relative to the simple average and a higher RMSE for values above 1. The Diebold and Mariano (1995) test with quadratic loss function and the adjustment made by Harvey et al. (1997) (DM-test) is used to determine the

[12] A sensitivity analysis of this threshold can be found in the next section.

significance of these results with the error adjustment for horizon 1-5 forecasts.[13]

Similar to Timmermann (2006), taking the recent best forecaster (Top1) does not perform very well relative to the other approaches. Overall, the alternative approaches tend to show very limited gains relative to the simple average and the only statistically significant gains can be found for bond yield forecasts.

Table 1: Comparison of several combination approaches

CPI	**Subset**	**Median**	**5% Trim**	**invMSE**	**Top1**	**Top5**
H0	0.74**	0.98	1.00	0.94	1.04	0.94
H1	1.01	1.01	1.00	1.00	1.00	1.02
H2	1.02	1.00	1.00	1.01	1.02	1.02
H3	0.99	1.00	1.00	1.00	1.10	1.02
H4	1.01	1.00	1.00	0.99	0.99	1.00
Unemployment						
H0	0.94*	0.96*	0.99	1.01	1.17	1.01
H1	0.95	1.01	1.00	1.03	1.20	1.03
H2	0.94**	0.99	1.00	1.03	1.11	1.02
H3	0.96	1.01	1.00	0.99	0.98	0.95
H4	1.02	0.99	1.00	1.03	1.09	1.01
10-yr Treasury						
H0	0.87***	0.93***	0.97***	0.91***	1.03	0.91**
H1	0.96*	1.00	1.00	0.99	1.09	0.98
H2	0.97	1.00	1.00	0.99	1.04	0.98
H3	0.92**	1.01	1.00*	0.96*	1.01	0.94**
H4	0.95**	0.99	0.99**	0.97**	1.00	0.95***

The RMSE of the various approaches are relative to the simple average. * significant improvement at 10% level, ** at 5% level and *** at 1% level over simple average based on one sided DM-test.

Overall, the new nonparametric approach (Subset) shows larger and mostly more significant gains than the alternative approaches tested. Indeed, the subset

[13] We also performed a modified Stekler (1987) and Batchelor (1990) test for equal forecast performance following Bürgi and Stekler (2015) and found that equal performance was strongly rejected for all variables at all horizons.

outperforms the best approach in 9 out of 15 cases and comes second for the remaining cases, where any approach gets more than a 1% improvement over the simple average.

The single largest gain of the new approach is found in the current quarter CPI forecasts where the RMSE of the subset chosen improves by 26% over the SPF average. Similar to the other approaches, there are no gains for any of the other horizons.

The unemployment estimates yield mixed results in terms of significance. There is a larger gain over the simple average at most horizons compared to the other approaches. However only the new approach's two-quarter-ahead forecasts are significantly better than the simple average at the 5% level.

The forecasts of the 10-year government bond yields show the most significant gains of all three variables and approaches. This could stem from the specialized knowledge of some forecasters, as many forecasters in the SPF might focus more on GDP, inflation and unemployment rather than bond yields. In particular at the very short horizon and the longer term, the subset and alternative approaches show a significant improvement over the overall average, while the gains are less significant for medium horizons[14]. Compared with the other approaches, the new approach outperforms the others for most horizons.

Given that the new approach tends to outperform the simple average by more than any of the alternative approaches, we will focus on the new approach for the rest of the paper.

Table 2 presents the percentage of periods in which the subset of best forecasters selected from previous periods beats the simple average. The percentage of periods in which the subset of individuals is more accurate than the average is significant for the bond yield forecasts at all horizons.

Taken together, there is clear evidence that the nonparametric approach found some individuals who could significantly outperform the simple average when forecasting treasury yields. It did not find similarly conclusive evidence for the other variables. While this finding is similar for the alternative approaches, the new nonparametric approach tends to show larger gains than the best alternative approach.

[14]Note that while both the subset and the simple average are biased for 10-year bond yields, the subset is clearly less biased.

Table 2: Share of forecasts made by the subset that beat the simple average

	CPI	Unemployment	10-yr Treasury
H0	0.63***	0.57	0.64***
H1	0.47	0.65***	0.63**
H2	0.38	0.64***	0.77***
H3	0.55	0.53	0.72***
H4	0.58	0.40	0.67***

* significant at 10% level, ** at 5% level and *** at 1% level based on a one sided coin flip test.

4.3 Robustness check: Different thresholds and alternative time periods

This section presents two checks for the robustness of the above results. These are the sensitivity of the results to different values of the percentage threshold p and the performance over different time periods. The sensitivity analysis for bond yields is shown because forecasts of that variable showed the most improvement. As stricter thresholds might lead to periods without any forecasters in the subset, the simple average replaces the subset for those periods.[15]

Table 3 shows the percentage improvement of the subset relative to the simple average based on RMSE for different threshold values of p. The results show that individuals who were able to beat the simple average 45-52.5% of the previous periods generally perform significantly better in future periods also.

The gains for bond yields are not much different for different time periods. While the gains appear to be most consistent in the first part of the sample and a bit more volatile recently, there is no clear pattern during or outside NBER recessions or whether bond yields are increasing or decreasing.

4.4 Are poorly performing forecasters driving the results?

D'Agostino et al. (2012) found evidence based on MSE that there are no innately better forecasters, however there are groups of forecasters that perform

[15] While it might be preferable to use the subset of the previous period instead of the simple average, that subset might also not contain any forecast for the current period, due to (re-) entry and exit of forecasters.

Table 3: 10-yr Treasury percent RMSE improvement for different threshold values of p

	0.45	0.50	0.525	0.55	0.60
H0	-13.44***	-15.16***	-12.69***	-7.80**	0.73
H1	-3.97**	-4.73***	-3.77**	-2.69*	-4.24
H2	-3.97**	-4.04**	-2.76	0.65	4.71
H3	-5.14**	-7.14**	-7.66**	-8.38**	-1.69
H4	-4.24*	-4.27**	-4.51**	-2.89	5.22

* significant at 10% level, ** at 5% level and *** at 1% level based on a one sided DM-test. Negative sign is an improvement.

very poorly. They based their analysis on the distribution of forecast errors in the SPF for GDP and the GDP deflator. Therefore it is important to check whether the new approach just excludes particularly poorly performing forecasters or if it is indeed able to detect better forecasters with potential specialized knowledge. That is, there might be a handful of bad forecasters in the sample, which deteriorate the performance of the simple average. If those forecasters were removed, one might find that the remaining forecasters have equal performance and all gains are simply due to removing the poorly performing ones. Alternatively, there could be better forecasters as well, which could be identified with the new approach. Given the most significant gains for the new approach are for bond yields, this property is only shown for bond yield forecasts[16].

To test this property, firstly poorly performing forecasters need to be identified and removed. To obtain real-time poorly performing forecasters for a given period, the MSE for the available past is calculated for every forecaster j. Forecasters whose MSE is i times larger than the simple average $\bar{X}$ are dropped for that period. This creates a new set of forecasters that all satisfy

$$MSE_j \leq MSE_{\bar{X}} * i \qquad \forall j. \tag{4}$$

Secondly, for every period, the new average $\bar{X}_i$ is calculated using the forecasters in this set. It is then checked for every period, if forecasters that have already

[16]Results for the other variables are similar (i.e. if the subset significantly beats the simple average for a given horizon, it also tends to significantly beat the average excluding poorly performing forecasters).

made at least 10 forecasts in that set beat the new average $\bar{X}_i$ more often than the threshold of 52.5%. Forecasters that satisfy this criterion are then included in the new subset S_i. Lastly, the performance of the new average $\bar{X}_i$ is compared to the simple average without dropping poorly performing forecasters $\bar{X}$ and the subset S_i is compared to the new average $\bar{X}_i$.

Table 4: Percentage RMSE improvement after dropping poorly performing 10-yr Treasury forecasters

.	$\bar{X}_2$	S_2	$\bar{X}_3$	S_3
H0	-10.76***	-4.42	-7.69***	-6.49**
H1	-0.67	-3.12*	-0.15*	-3.63**
H2	0.30	-3.05*	-0.30	-2.47
H3	-1.11*	-6.63**	-0.59	-7.11**
H4	-1.65***	-2.91**	-0.97**	-3.57***

* significant improvement at 10% level, ** at 5% level and *** at 1% level based on one sided DM-test. X_i is the new average's improvement over simple average. S_i is the subset's improvement over $\bar{X}_i$, excluding forecasters with i times larger MSE than the simple average

Following D'Agostino et al. (2012), we select the value for i based on the confidence interval for the worst 5% of forecasters relative to the median forecaster, which corresponds to $i \in \{2, 3\}$. As the goal is to drop particularly poorly performing forecasters, one would assume that they should be much worse at predicting outcomes than the simple average. Table 4 shows that most of the gains for the current quarter (H0) are due to dropping poorly performing forecasters. However, most of the gains for all other horizons are due to selecting well performing forecasters. In addition, the significance of the improvement of the subset S_3 relative to the new average $\bar{X}_3$ is essentially unchanged from Table 1 and for longer horizons, the subset S_2 is still significantly better than $\bar{X}_2$.

4.5 Specialized knowledge

This new combination approach showed largest improvements in the accuracy of bond yield forecasts as opposed to those of CPI inflation or unemployment. At the same time, the latter two variables are very closely watched by all forecasters while not every forecaster might pay as much attention to bond yields. It is thus possible that a few individuals closely watch bond yields and might have specialized knowledge about them, while others do not. This could explain why the gains for bond yields are much more significant than for the other variables.

If some forecasters do indeed have specialized knowledge that helps them forecast well, this knowledge may be helpful for predicting multiple horizons. We check this property using the subset of forecasters for one-quarter-ahead forecasts[17].

Table 5: Percentage RMSE improvement relative to simple average, using the best one-quarter-ahead forecasters

	10-yr Treasury
H2	-4.76***
H3	-5.06**
H4	-5.94**

* significant improvement at 10% level, ** at 5% level and *** at 1% level based on one sided DM-test.

As Table 5 shows for bond yields, forecasters who perform well one-quarter-ahead tend to perform better at longer horizons as well based on RMSE. This could suggest that some forecasters have superior skills or knowledge over other forecasters.

In addition to the performance across horizons, the composition of the subset also shows evidence for specialized knowledge. Based on the industry classification used in the SPF, one would assume that forecasters at financial institutions are more likely to watch bond yields closely as compared with forecasters at other institutions. Indeed while forecasters in the financial sector are about 46% of the overall sample, they make up 55% of the forecasters selected to be

[17]Note that this would also reduce the lag between forecasts being made and evaluated for longer horizons.

included in the one-quarter-ahead subsets.

5 Conclusions

In this paper we show that the high correlation among forecast errors can lead to a sizeable percentage of individual forecasters outperforming the simple average merely by chance. Due to this, selecting the best forecasters based on past RMSE might not always identify the individuals who can make the most accurate forecasts in the future.

Subsequently a new approach to select a subset of best forecasters was developed. This approach is able to find individuals whose forecasts of 10-year government bond yields were significantly more accurate than the simple average. While alternative approaches yield higher gains for bond yields as well, the gains by the new nonparametric approach are larger and more significant. This result holds across several forecast evaluation approaches and robustness checks. While some of the gains could be due to removing poorly performing forecasters at shorter horizons, this is not the case for longer horizons. In addition, there is strong evidence that bond yield forecasters that are in the subset for one quarter ahead forecasts tend to perform better at other horizons as well. This provides further evidence that it is likely that specialized knowledge makes some individuals perform better than the consensus forecast.

Further research might be able to test this approach on other data sets as well or determine why there appear to be gains across horizons in bond yields, but much more limited gains for CPI and unemployment.

References

Batchelor, R. (1990). All forecasters are equal. *Journal of Business & Economic Statistics*, 8(1):143–44.

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

Blix, M., Wadefjord, J., Wienecke, U., and Adahl, M. (2001). How good is the forecasting performance of major institutions? *Sveriges riksbank economic review*, pages 38–68.

Bürgi, C. and Stekler, H. (2015). Forecast Evaluation with re-entry of Forecasters. *Manuscript*.

Capistrán, C. and Timmermann, A. (2009). Forecast Combination With Entry and Exit of Experts. *Journal of Business & Economic Statistics*, 27(4):428–440.

Clark, T. E. and McCracken, M. W. (2010). Averaging forecasts from vars with uncertain instabilities. *Journal of Applied Econometrics*, 25(1):5–29.

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

Conflitti, C., Mol, C. D., and Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, 31(4):1096 – 1103.

D'Agostino, A., Mcquinn, K., and Whelan, K. (2012). Are some forecasters really better than others? *Journal of Money, Credit and Banking*, 44(4):715–732.

Davies, A. and Lahiri, K. (1995). A new framework for analyzing survey forecasts using three-dimensional panel data. *Journal of Econometrics*, 68(1):205 – 227.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Ericsson, N. and Reisman, E. (2012). Evaluating a Global Vector Autoregression for Forecasting. *International Advances in Economic Research*, 18(3):247–258.

Gamber, E. N., Smith, J. K., and McNamara, D. C. (2014). Where is the fed in the distribution of forecasters? *Journal of Policy Modeling*, 36(2):296 – 312.

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

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2):281–291.

Jore, A. S., Mitchell, J., and Vahey, S. P. (2010). Combining forecast densities from vars with uncertain instabilities. *Journal of Applied Econometrics*, 25(4):621–634.

Kascha, C. and Ravazzolo, F. (2010). Combining inflation density forecasts. *Journal of Forecasting*, 29(1-2):231–250.

Kenny, G., Kostka, T., and Masera, F. (2015). Density characteristics and density forecast performance: a panel analysis. *Empirical Economics*, 48(3):1203–1231.

Lahiri, K., Peng, H., and Zhao, Y. (2014). On-line learning and forecast combination in unbalanced panels. *Forthcoming in Econometric Reviews*.

Lahiri, K., Peng, H., and Zhao, Y. (2015). Testing the value of probability forecasts for calibrated combining. *International Journal of Forecasting*, 31(1):113–129.

Poncela, P., Rodríguez, J., Sánchez-Mangas, R., and Senra, E. (2011). Forecast combination through dimension reduction techniques. *International Journal of Forecasting*, 27(2):224–237.

Stekler, H. O. (1987). Who forecasts better? *Journal of Business & Economic Statistics*, 5(1):pp. 155–158.

Timmermann, A. (2006). Chapter 4 forecast combinations. In G. Elliott, C. G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1 of *Handbook of Economic Forecasting*, pages 135 – 196. Elsevier.

Zhao, Y. (2015). Robustness of forecast combination in unstable environment: A monte carlo study of advanced algorithms. *Research Program on Forecasting Working Paper No. 2015-005*.