GW | Research Program on Forecasting

What Do We Lose When We Average Expectations?

Constantin Bürgi
Department of Economics
The George Washington University
Washington, DC 20052 USA
cburgi@gwu.edu

RPF Working Paper No. 2016-013
http://www2.gwu.edu/~forcpgm/2016-013.pdf

November 15, 2016

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www2.gwu.edu/~forcpgm

What Do We Lose When We Average Expectations?[*]

Constantin Bürgi[†]

October 31, 2016

Abstract

In this paper, I use the Bloomberg Survey of forecasts to assess if evaluating the distribution of expectations will lead to important additional insights over the evaluation of the simple average. I first introduce new approaches that allow me to assess the forecast accuracy and the information rigidity at the individual level despite a large share of missing data. Applying these new approaches, I find that taking into account the distribution can significantly improve the predictive power of the survey. For example, I find that the part of uncertainty measured by disagreement can improve the prediction of recessions in a dynamic probit model relative to the simple average. On information rigidity, I find that some of the rigidity found at the aggregate level likely stems from the aggregation process. Together, my findings suggest that we should look at individual expectations whenever possible as important insights are lost by just looking at aggregate expectations.

JEL: C22, C52, C53, E17, E37

Keywords: Expectations, Bloomberg Survey, Forecast Evaluation, Uncertainty, Dynamic Probit

[*] I would like to thank Herman Stekler, Tara Sinclair, James Foster, Jane Ryngaert, Pavel Potuzak, the participants in the IAES conference and the participants in the SAGE seminar for their valuable comments and support.

[†] *The George Washington University, 2115 G Street, NW, # 340, Washington, DC 20052*

1 Introduction

Most economic models include some form of expectations. Depending on how these expectations are formed and what these expectations are, models can have quite different implications. For example, Lucas (1976) showed in his famous critique that expectations play a crucial part in how effective policy actions are. Given how crucial expectations for the models are, it is important to test the underlying assumptions made about them.

Some of the most extensive empirical work in regards to empirical work about expectations has been done in the economic forecasting literature. Economic forecasts provide future expectations that can be evaluated to assess common assumptions made in economic models like accuracy and rationality.

An unanswered question in this literature is, whether it is sufficient to assess the model assumptions by evaluating the performance of the simple average of expectations over the entire period, or whether one should instead evaluate them at the individual level and for subperiods. This is important, because evaluating the performance at the individual level can be more difficult; for example due to gaps in the individual data.

From a theoretical perspective, it is well known since at least Gorman (1961) that testing model implications at the individual level can lead to different results than testing them at the aggregate or average level.[1] This immediately leads to the question: What is the significance of this result for expectations? For example, this theoretical finding has more significance if evaluating average expectations instead of individual expectations often leads to sizeable differences in evaluation outcomes.

The empirical evidence in this regard so far has been mixed. For example, Clemen (1989) and Timmermann (2006) have found that it is difficult to improve over the simple average of expectations in terms of forecast accuracy. In contrast, Bürgi and Sinclair (2016) showed that these findings might be due to the high correlation among individual forecasters and the prevalence of missing observations at the individual level. Taking this finding into account, they are able to create a subset of better forecasters for the Survey of Professional Forecasters (SPF) that in some cases will significantly outperform the simple average in future periods.

[1] Indeed, Antonelli (1886) and Nataf (1953) had shown this property earlier.

In this paper, I will address the question whether I should look at the simple average or individual expectations by evaluating the US Bloomberg Survey of forecasts for the real GDP growth rate, the CPI inflation, the unemployment rate and 10 year treasury bond yields. I choose this survey, because it is widely used in the private sector when comparing economic releases to expectations. I will assess in three categories if there is a difference between the outcome of the evaluation at the individual level and the one at the average level.

First, I will assess the accuracy and bias of the survey at both levels. To evaluate the survey at the individual level, I develop a set of new approaches as, because of missing observations, many evaluation approaches that are commonly applied to the simple average cannot be directly applied to the individual level.[2] The first new approach allows me to assess the performance of forecasters. The second new approach allows me to show that one cannot evaluate the bias and thus the rationality of forecasts at the individual level.

I find that the simple average of the Bloomberg survey is statistically significantly more accurate than the random walk for most variables. I also find that there are significant differences in the performance of individual forecasters. Regarding rationality of expectations, I find that the survey is biased upward for bond yields at the average level and that there are systemic biases that cancel out over time for the unemployment rate and the real GDP growth rate. At the individual forecaster level, I find that applying the same method for estimating biases as at the aggregate level will lead to sizeable shares of forecasters being biased for most variables. However, I find that this result is likely due to missing observations and show that there is an identification issue. This can cause many more forecasters to be identified as biased just because of the pattern of missing observations.

Second, I try to improve the accuracy of expectations by examining the distribution and not just the simple average. Taking the significant differences among forecasters individual accuracy into account, I am able to improve over the simple average for a number of variables and horizons. I then show that the part of uncertainty that is measured by disagreement is able to predict recessions quite well. It is a better predictor of recessions than the simple average of growth

[2] For example, many of the evaluation approaches have autocorrelated errors or require covariance matrices. With more than 80% of the observations missing relative to a complete panel at the individual level, autocorrelation or covariances cannot be estimated without dropping most of the forecasters from the sample.

forecasts in a dynamic probit model for the past two recessions and much better than the latest available real GDP growth rate. Relative to the simple average, including uncertainty will correctly predict two more quarters in my sample.

Third, I will empirically test the efficiency of individual forecasters under the assumptions of the noisy information model used by Coibion and Gorodnichenko (2015) for the first time. My new approach is able to solve the problem with endogenous variables that above authors faced at the individual level.

I find some evidence of information rigidity at shorter horizons but very little evidence at longer horizons at the aggregate level. At the individual level, I am able to provide some supporting evidence for both the noisy information and sticky information models. My analysis also shows that independent of the assumption about the underlying model, there is some evidence that the aggregation process contributes to the information rigidity. This finding gives a slight edge to the sticky information model in my view.

Overall, I find that focusing on the simple average of expectations can lead to quite different results to the ones obtained at the individual level. In particular, looking at the forecasts made by individual forecasters allows me to reject and confirm certain assumptions of economic models that the simple average cannot.

The remainder of the paper is structured as follows: Section two will discuss the data set used in this paper and compare it to other surveys. Section three I will address the accuracy and the bias of expectations in the Bloomberg Survey. Section four will check if the simple average can be improved upon by applying several alternative approaches. I first look at approaches based on past performance and then use the part of uncertainty measured by disagreement among forecasters to predict recessions in a dynamic probit model. Section five evaluates the expectation formation process to check for information rigidities and section six concludes.

2 Data

The private sector uses the Bloomberg Survey very frequently to compare economic data releases to what economists had expected beforehand. Despite this popularity in the private sector, there has been very little academic research about this survey. Most of the sparse literature which analyses this survey focuses only on the ”surprises” of the data releases relative to expectations and their impact on the market (e.g. Scotti (2013) or Chen et al. (2013)).

For the US, the Bloomberg Survey was first collected in June 1993 and included variables like quarterly GDP, and quarterly averages for the unemployment rate and year-over-year CPI. It was collected at a quarterly frequency and at the end of the third month of each quarter. The forecasts where at the one and four quarter ahead horizons. Over time, the specific horizons covered changed and were extended to span five horizons from current quarter forecasts up to four quarter ahead forecasts which I will denote H0-H4.

In June 2000, there were quite a few changes made to the survey including the addition of end of quarter 10 year government bond yields variable. The survey shifted from being conducted quarterly to be conducted monthly and is now conducted at the middle of the month. While the monthly survey is available, GDP is only available at a quarterly frequency. If every survey was included in my analysis, this would create an overlap for some of the variables, because there are three surveys in any quarter. Of the three surveys per quarter, I opt to only include the one that was conducted in the last month of the quarter. This ensures that the timing of the survey remains similar to when the survey was conducted on a quarterly basis.

In addition to these changes, the survey also changed from forecasting quarterly averages to forecasting end of quarter values. While this change does not allow a comparison of the forecast performance prior to the change to the one after the change (except for GDP), the impact on the evaluation should be minor. This also implies that some data is not fully comparable to other surveys over the whole period.

The survey mainly includes forecasts from the financial services industry but also some forecasts from academia or non-financial companies. The survey also collects all names of individual forecasters as well as their institutions, allowing the 300-400 individuals or firms and sectors to be tracked over time.[3] The dataset includes many gaps for individual forecasters, as some enter and exit the survey or miss certain dates.

Altogether, the survey is similar to other macroeconomic forecast surveys like the BlueChip, the Wall Street Journal (WSJ), Consensus Economics (CE) or the Survey of Professional Forecasters (SPF), which all have been extensively

[3] The 300-400 individuals are obtained after merging individuals that were spelled differently at different times.

evaluated.[4]

Table 1: Overview over several different surveys

	SPF	BlueChip	CE	Bloomberg	WSJ
Release Date$^+$	W6	W2	W2	W2	D1
Forecasters***	40+	50+	30+	80+	60+
Anonymous	Yes$^{\#}$	No	No	No	No
Start Year	1968	1976	1989	1993	1986$^{\#\#}$
Collected	Q	M	M	M*	M*
Predicted	Q/A	A	A	Q/A	A**

* Starts at lower frequency, ** Most semi annual and GDP quarterly as well, *** Real GDP for the US, $^{\#}$Forecasters have a unique number. $^+$ Approximate release date, in weeks (W) or days (D) from the first day in the month or quarter of the survey.

Table 1 shows a comparison of the above mentioned surveys.[5] While all surveys cover the main macroeconomic variables used in this paper, there are clear differences regarding the number of forecasters included, the frequency with which it is collected and the frequency of the underlying variable. With the exception of the start year, the Bloomberg survey appears to have the most attractive dataset from these surveys. It has a very large coverage in number of forecasters regularly contributing.[6] It tracks all forecasters by name and institution which allows the user of the data to track either of the two over time. The survey is published on a monthly basis and for some indicators it is possible to obtain them even at a daily frequency. At the same time, the Bloomberg survey offers users to look both at quarterly and annual forecasts which can be important for certain economic analysis.

[4] The following papers are an incomplete list of papers in this literature: Batchelor (2007), Ager et al. (2009), Carroll (2003), Cho (1996), Greer (1999), Davies and Lahiri (1995), Dovern and Weisser (2011), and Mitchell and Pearce (2007)

[5] Not all of the surveys use the same definitions of the underlying variables and some of them are not freely available. Due to this, I cannot compare the forecast performance across these surveys

[6] The Bloomberg Survey also has a large coverage across countries similar to ConsensusEconomics, which can be important for other applications of the dataset. The other three surveys (SPF, BlueChip and WSJ) only cover the US.

For my actual data to compare the forecasts, I take the third release for GDP which is released with a one quarter delay. For the other variables, I take revised data as the revisions there are only minor. I assume the actual value is known with the first release for the unemployment rate and bond yields which is within the first week of the subsequent quarter, while the CPI release is in the middle of the first month of the quarter.

3 Evaluation of the Accuracy and Bias of Expectations

3.1 Accuracy of Expectations

3.1.1 Accuracy of the Simple Average

As a first step I assess the accuracy of the simple average of the forecasts included in the Bloomberg Survey. I assess the performance with the root mean squared error (RMSE) of the form

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(F_{t,t-i} - A_t)^2}, \tag{1}$$

where A_t is the current value at period t and $F_{t,t-i}$ is the forecast for period t made in period t-i. I compare this performance to a simple random walk (RW) of the form

$$A_t = A_{t-1} + \varepsilon_t \tag{2}$$

where A_t is the current value at period t and A_{t-1} corresponds to the value in the past quarter. A_{t-1} might not be known for long forecast horizons. Due to this, I use the most recent available release in real time instead. This allows me to test if the survey adds any information relative to the random walk.[7]

In addition to comparing the Bloomberg survey to the random walk, I also want to compare it to another benchmark survey. I use the Philadelphia Fed Survey of Professional Forecasters (SPF) for my comparison. As specified above, the definitions used in the Bloomberg survey are changing to end of quarter

[7] While I focus on the simple random walk as a comparison, for example Rubaszek and Skrzypczyński (2008) has shown that the performance of the SPF relative to a DSGE and a VAR model is not significantly different for GDP, 3-month treasury bills and the GDP deflator.

values for all variables except for GDP in June 2000. The SPF in turn reports quarterly averages for all variables. Due to this restriction, I am only able to compare the Bloomberg Survey to the SPF for GDP over an extensive sample. I compare the forecasts made in the two surveys for the dates ranging from 1993Q3 to 2013Q3.[8] The SPF is collected in the middle month of a quarter, in contrast to the last month of a quarter for the Bloomberg survey. Thus, I would expect the SPF to perform slightly worse, as the Bloomberg survey has a small information advantage. This advantage should be more pronounced at shorter horizons, as one month of additional information for the current quarter forecast is much more important than one month of additional information for the four quarter ahead forecast.

To test if the Bloomberg survey has a statistically significantly lower RMSE than the random walk or the SPF, I use a one sided test based on the Diebold and Mariano (1995) test with the adjustments from Harvey et al. (1997) (DM statistic).

As Table 2 shows, the Bloomberg survey (BBG) significantly improves over the random walk for the three macro variables GDP, unemployment and CPI at all horizons. For bond yields however, the forecasters tend to beat the random walk only at the short horizon. This result for bond yields is in line with the results found by Baghestani (2006) for the SPF and Mitchell and Pearce (2007) for the WSJ survey. One of the reasons for this could be the fact that 10 year government bond yields tend to be relatively smooth when compared with the other variables. The random walk tends to perform very poorly if there are some sharp movements in the data like a recession in GDP or Q4 2008 for the CPI.

Table 2 also shows that the SPF and the Bloomberg survey appear to have very similar forecasting power for GDP. There is a small advantage for the Bloomberg survey, which could just stem from the small information advantage. Overall, this analysis showed that the simple average of the Bloomberg survey is useful, as it adds information relative to the random walk and shows a similar RMSE to the SPF.

[8] As specified in the data section, the Bloomberg survey does not cover all horizons initially. Due to this the dates for specific horizons differ as follows: H0 and H2 only start in 2000Q2, H3 in 2000Q3 and H4 starts in 1995Q4 and ends in 2009Q1.

Table 2: RMSE of the Bloomberg survey and the random walk

	GDP				
	H0	H1	H2	H3	H4
BBG	1.34	2.01	2.19	2.32	2.68
RW	2.43***	2.98***	3.43***	3.54**	3.38***
SPF	1.69***	2.09*	2.24	2.40	2.72
	CPI				
	H0	H1	H2	H3	H4
BBG	0.51	0.78	1.14	1.28	1.33
RW	1.11***	1.29**	1.75**	1.99**	1.73*
	Unemployment				
	H0	H1	H2	H3	H4
BBG	0.20	0.33	0.59	0.82	1.04
RW	0.37***	0.58**	0.95**	1.22**	1.31**
	Bond Yield				
	H0	H1	H2	H3	H4
BBG	-	0.58	0.78	0.89	1.07
RW	-	0.71***	0.75	0.83	0.90

* significantly worse than BBG at 10% level, ** at 5% level and *** at 1% level based on DM statistic.

3.1.2 Accuracy of Expectations of Individuals

By just looking at the simple average, a lot of information might be lost. In particular, it is important to determine if all forecasters perform the same, or if there are better and worse forecasters.

The Bloomberg survey identifies individual forecasters over time by name and affiliation, similar to the WSJ survey. This allows the forecasters to be tracked over time and assess their performance relative to other forecasters. Before determining, who are the better and worse forecasters, it is necessary to check if there are better and worse forecasters. That is, I need to test if all forecasters have the same predictive accuracy.

Given the large number of missing values, I use a rank based method similar to the Friedman (1937) test and applied by Stekler (1987) and Batchelor (1990).[9] This test ranks all forecasters every period and then standardizes ranks for each forecaster according to the formula

$$z_i = \sum_{t=1}^{T} \frac{rank_{it} - (n+1)/2}{\sqrt{Tn(n+1)/12}}, \tag{3}$$

where n is the number of forecasters, T the number of periods and z_i the standardized rank for forecaster i. Under H_0, all forecasters have the same predictive ability and thus their expected standardized rank should be zero. Taking the sum of squared standardized ranks will create a variable that under H_0 follows a Chi squared distribution with n-1 degrees of freedom.

This test does not allow for missing data and thus Skillings and Mack (1981) modified it. If the Friedman (1937) test were used directly, being the best forecaster if there are few would have a much lower impact than being the best forecaster if there are many. They advise to modify the variance covariance matrix to give a larger weight to periods with fewer forecasters to compensate for this somewhat. Their method was more recently applied by Batchelor (2007) and Ashiya (2006). One problem with their test is that it requires to use the generalized inverse of a large matrix and hence can be computationally expensive.

I introduce a much simpler approach to adjust the Friedman (1937) test for missing observations by noticing how rankings are related to (discrete) uniform distributions. In a complete panel with n forecasters over t periods and under H_0, the ranks of individual forecasters are t times randomly and independently drawn from the discrete interval $[1, n]$. There are thus t draws for each individual forecaster from a uniform distribution. This uniform distribution can easily be standardized to the interval $[0, 1]$ by subtracting one from every rank and then

[9] Some of the most commonly used alternative approaches include the Diebold and Mariano (1995) test, forecast encompassing tests as found in Chong and Hendry (1986), Ericsson and Marquez (1993), Clark and McCracken (2001) or Harvey et al. (1998) and the White (2000) Reality Check test. Unfortunately, none of these tests can be applied here directly due to the very large number of forecasters relative to the number of observations and the missing data, which creates identification problems. For example, there are 353 different forecasters that contributed at least once for H1 GDP, but there are only 85 quarters in the dataset and 85.3% of observations are missing relative to a balanced panel. Also, the adjustments used in Capistrán and Timmermann (2009b), Genre et al. (2013) and Lahiri et al. (2013) to accommodate for missing data cannot be applied here without dropping most forecasters.

dividing it by (n-1). If the panel is unbalanced, n might be different for each period. However, I can standardize each period by $(n_t - 1)$ instead of (n-1). This allows me to obtained a much simpler form for my covariance matrix than the one in Skillings and Mack (1981) and can just rely on the Friedman (1937) test. The downside of this approach is that I explicitly assume that becoming first when there are two forecasters or 200 forecasters is equivalent. In such an extreme case, where there are periods with very few forecasters, this assumption might not be adequate and a different method to adjust for missing observations might be better. However, in the Bloomberg survey the discrepancy between the periods with the most forecasters relative to the least forecasters is much smaller. For example, for H1 GDP, the period with the fewest forecasters has 19 and the period with the most forecasters has 68 forecasters. Due to this, I decide to apply this approach.[10]

For every period, I rank all the forecasters participating according to their squared error for that period.[11] If several forecasters have the same prediction error, the average rank is used for all those forecasters. Then the rankings are rescaled to the interval $[0, 1]$ to avoid issues with the changing number of forecasters as described above. For each forecasters, all the scaled ranks (srank) are then added up and standardized according to the formula

$$z_i = \frac{\sum_{t=1}^{T_i} srank_{it} - T_i/2}{\sqrt{T_i/12}}, \tag{4}$$

where T_i is the number of times forecaster i is in the Bloomberg survey. As z_i is normally distributed (provided n is large), the Chi squared test for equal forecasting power with n-1 degrees of freedom becomes

$$\sum_{i=1}^{n} z_i^2 \sim \chi_{n-1}. \tag{5}$$

The results are presented in Table 3. The test statistic rejects equal forecasting power for all variables and horizons at the 95% level except for four quarter ahead GDP forecasts. This implies that forecasters do not have equal

[10] Using my approach or the approach advertised by Skillings and Mack (1981) lead to very similar results.

[11] I only include the z_is into the test statistic for forecasters that contributed at least 10 forecasts at the respective horizon, or 10 across all horizons for the joint test to avoid small sample biases.

Table 3: Chi squared test for equal forecasting power

Horizon	Unemployment	GDP	Bond Yield	CPI
H0	140.98**	142.94**	-	354.02***
	(108)	(113)		(113)
H1	291.56***	209.79***	193.63***	249.05***
	(125)	(139)	(112)	(137)
H2	253.31***	178.24***	334.57***	270.64***
	(104)	(111)	(110)	(110)
H3	271.76***	198.16***	447.70***	271.74***
	(105)	(111)	(107)	(110)
H4	235.37***	99.55	528.19***	154.96***
	(87)	(90)	(102)	(91)
Joint	1516.94***	649.30***	1490.87***	1260.94***
	(264)	(255)	(208)	(262)

Number of forecasters in brackets. * significant rejection at 10% level, ** at 5% level and *** at 1% level.

predictive accuracy ex post. By looking at the distribution of the z_i, it is possible to check the ratio of significantly better forecasters to significantly worse forecasters. I find that for the three macro variables, the number of significantly better forecasters is similar to the number of significantly worse forecasters. For bond yields, the number of significantly better forecasters is almost twice as large as the number of worse forecasters.

3.1.3 Who Has the Most Accurate Expectations Across Variables?

In evaluating whether all forecasters' expectations are equally accurate in the previous section, I was also able to determine how many forecasters were significantly better and worse than the average. The distribution of z_i can also be used to compare the accuracy of the predictions across all variables among forecasters directly and across all horizons. In particular, I can create a weighted index of the z_i for each of the four variables at a given horizon and find the forecaster who is best in predicting all four variables.

As discussed by Batchelor (2007), it is problematic to perform this rank analysis across horizons. Due to this, I opt to calculate the weighted index for the two horizons H1 and H4 separately. This requires me to show that such a weighted index is valid and if it is, how to obtain the weights for the four dimensions (i.e. variables).

As highlighted in Alkire and Foster (2011) for example, any weighted index requires comparability of the dimensions to be valid. Comparability in my case means that the z_i for GDP must be independently identically distributed to the z_i for unemployment, the ones for CPI and the ones for bond yields. While the range and variance of the z_i is the same for all variables due to the standardization, the two variables are correlated.[12] Given faster GDP growth often goes hand in hand with a fall in the unemployment rate (see Okun (1963)), making a better prediction for GDP can lead to an improvement in the unemployment prediction as well. Unless this correlation is taken into account, the index might have an unintended higher weight on those two variables, relative to other variables included in the index. Also, I require forecasters to have predicted at least 10 periods in all dimensions. After correcting for this correlation, I opt for equal weights across dimensions.

I achieve this by using a modified version of the Mahalanobis (1936) distance measure defined as

$$D = \sqrt{u_i' S^{-1} u_i}, \tag{6}$$

where u_i is the stacked vector of z_i for the four dimensions for forecaster i and S is the correlation matrix of the outright forecast errors of the simple average across the four dimensions.

Previous research like Jordà et al. (2013), Banternghansa and McCracken (2009) or Sinclair et al. (2015) used the MSE as their measure for performance. Because of this, they could not take the covariance matrix of the errors of the simple average directly. Given the high correlation among forecasters, one would otherwise almost adjust each forecaster by his or her own covariance matrix. Because of this issue, they either chose to use the covariance of the underlying variable or the covariance of some independently simulated forecasts. In my case however, accuracy is determined using a rank based method and I can thus use the prediction errors of the simple average. The performance of each

[12] If for example the range of one variable did not have an upper bound, while all others have one, this variable could dominate the index violating comparability.

Table 4: Joint Forecast performance across horizons and variables.

H1			H4		
Rank	D^2	Name	Rank	D^2	Name
1	0.690	Helaba	1	1.403	Econ. Solutions
2	0.703	Nationwide	2	1.661	Fairfield Univ.
3	0.706	Econ. Consulting	3	1.668	Briefing.com
4	0.721	Essen Hyp	4	1.753	MFR
5	0.731	RBS Greenwich	5	1.823	Essen Hyp
6	0.759	ClearView Econ.	6	1.833	RBS Greenwich
7	0.764	Raymond James	7	1.836	Credit Suisse FB
8	0.782	Societe Generale	8	1.889	Moodys
9	0.785	Anderson Economic	9	1.895	US Trust
10	0.796	US Trust	10	1.932	SwissRe
11	0.803	Barclays	11	1.981	Goldman Sachs
12	0.804	Moodys	12	1.997	ING
13	0.806	Mesirow Financial	13	2.023	Bear Stearns
14	0.822	Northern Trust	14	2.064	Eaton Vance
15	0.823	Bear Stearns	15	2.065	Daiwa
16	0.832	MacroFin Analytics	16	2.085	Deutsche Bank
17	0.834	Wayne Hummer Inv	17	2.094	Landesbank
18	0.844	Credit Suisse FB	18	2.114	National-City
19	0.845	NAR	19	2.164	Nomura
20	0.845	Lehman	20	2.173	U. of Michigan

Forecasters are required to have made at least 10 forecasts for every variable to be considered at the horizon specified.

forecaster is standardized as well and hence I do not need to correct for harder or easier to predict dimensions, just the covariances between them. I hence use the correlation matrix.

Table 4 shows the 20 forecasters with the best forecast performance for H1 and H4. It is important to note that due to the selection process, not all forecasters that are included at H1 are included at H4. Due to this, the fact that most forecasters at H1 and at H4 are different might be simply due to this issue. However, even if forecasters that do not appear at both horizons are excluded, it is still the case that quite a few forecasters that are well at predicting H1 are not doing as well at predicting H4 and vice versa.

3.2 Bias

3.2.1 Bias of the Simple Average

After assessing the accuracy of the forecasts, I want to evaluate if the simple average of all forecasters is unbiased for all variables and horizons and thus if the forecasts are rational. To test this property, I run for each of the four variables a simple Holden and Peel (1990) regression of the form

$$A_t - F_{t,t-i} = \alpha + \varepsilon_{t,t-i}; \qquad i = 0, 1, ..., 4, \tag{7}$$

where $F_{t,t-i}$ is the simple average of all forecasts for period t, made in period $t-i$, A_t is the revised actual value for the unemployment rate, yoy CPI and 10 year government bond yields and the third release for real GDP and $\varepsilon_{t,t-1}$ is the error term, which is autocorrelated for $i > 0$.[13] Due to this autocorrelation, I will use HAC errors for H1-H4. If forecasts are unbiased, the constant α should be not significantly different from zero. Table 5 presents the results from this regression. If the constant is positive, forecasters tend to over predict the underlying variable and if it is negative, forecasters tend to under predict the underlying variable.

With the exception of bond yields, the simple mean of all forecasters is broadly unbiased by this measure and thus one can conclude that the simple average is rational as well for those variables.[14] The exceptions are found for the three quarter ahead GDP forecast, and the bond yield forecasts. If the rationality assumption is to hold for these cases as well, it would require asymmetric loss functions.

Even if forecasters are not biased overall, there might still be systemic errors that cancel out over time. There has been extensive research on this topic that found that forecast errors differ across the business cycle as summarized in Fildes and Stekler (2002). Sinclair et al. (2010) and more recently Dovern and

[13] I use revised data for the first three variables, as the revisions to them tend to be very small.

[14] Similar results can be obtained using the Mincer and Zarnowitz (1969) approach for GDP, Unemployment and the CPI, where the actual value is regressed on the forecast and a constant. For bond yields, the Mincer and Zarnowitz (1969) approach will lead to a lower significance in the bias.

Table 5: Simple Bias across variables

Horizon	Unemployment	GDP	Bond Yield	CPI
H0	-0.04	0.14	-	0.02
	(0.03)	(0.19)		(0.07)
H1	-0.07	0.28	-0.23***	0.00
	(0.06)	(0.22)	(0.06)	(0.39)
H2	-0.01	-0.44	-0.42***	0.14
	(0.08)	(0.29)	(0.09)	(0.17)
H3	0.03	-0.56**	-0.60***	0.09
	(0.09)	(0.28)	(0.12)	(0.15)
H4	0.20	0.02	-0.78***	0.00
	(0.26)	(0.28)	(0.15)	(0.06)

Standard errors in brackets. * significant at 10% level, ** at 5% level and *** at 1% level based on OLS errors (H0)/HAC errors (H1-H4).

Jannsen (2015) found systemic errors over the business cycle for unemployment and GDP. Forecasts for unemployment tend to be systemically too low and forecasts for GDP tend to be systemically too high during NBER recession periods and the opposite outside recessions. To check, whether this is also the case for the Bloomberg survey, I modify equation (7) to

$$A_t - F_{t,t-i} = \alpha + \beta D_t + \varepsilon_{t,t-i}; \qquad i = 0, 1, ..., 4, \tag{8}$$

where D_t is a recession dummy variable which takes value one if the economy was in a NBER-dated recession and zero otherwise. I would expect the forecasts not to suffer from systematic errors if $\alpha = \beta = 0$. α captures the overall bias, while β captures the systematic or cyclical errors. I again use HAC errors for H1-H4 due to the autocorrelation of forecast errors. The results are presented in Table 6. They show that the recession dummy is significantly different from zero for all forecast horizons except for current quarter GDP forecasts. The Bloomberg survey thus shows a similar pattern to the Greenbook forecast, as the forecasts tend to systemically predict a lower unemployment rate and higher GDP growth during recessions and the reverse during non recession periods. When testing the systematic errors I also find that forecasters tend to under predict the unemployment rate at short horizons and over predict GDP for some horizons.[15]

Table 6: Recession Bias across variables

	Unemployment		GDP	
Horizon	α	β	α	β
H0	-0.10***	0.31***	0.13	0.03
	(0.02)	(0.05)	(0.22)	(0.50)
H1	-0.15***	0.66***	0.50**	-1.69***
	(0.02)	(0.19)	(0.21)	(0.54)
H2	-0.21***	1.06***	0.07	-2.70***
	(0.06)	(0.31)	(0.21)	(0.58)
H3	-0.19*	1.36***	-0.08	-3.16***
	(0.10)	(0.44)	(0.22)	(0.74)
H4	-0.07	1.42***	0.75**	-4.28***
	(0.16)	(0.48)	(0.37)	(0.92)

Standard errors in brackets. * significant at 10% level, ** at 5% level and *** at 1% level based on OLS errors (H0)/HAC errors (H1-H4).

At the aggregate level, there appear to by systematic errors in this dataset, but there is no extensive evidence for classical forecast biases with the exception of bond yields.

3.2.2 Biases of Individual Forecasters

After having assessed the performance of the simple average, it is important to check if these results are due to a similar pattern for individual forecasters, or if the unbiased for three of the four variables is just due to positive and negative individual biases that cancel each other out. For the biased bond yield forecast, an alternative question is if the bias stems from a large share of individual forecasters, or if it stems from a small group of strongly biased forecasters.

[15] A similar regression could be ran for bond yields and the CPI to check if errors are different for upward trending inflation and yields relative to downward trending inflation and yields. However, I refrain from running the analysis for those variables, as the timing of the periods is more ambiguous there.

At the individual level, I could estimate equation (7) and report the share of biased forecasters. Unlike the simple average, I have to deal with missing values in this case. Just estimating equation (7) without any adjustments will lead to overestimating the share of biased forecasters. This is due to gaps in the survey. For example, a forecaster might only have contributed to the survey during periods where most forecasters tended to over predict the underlying variable. This forecaster will be identified as being overall biased, even if this might only be due to correlation between α and $\varepsilon_{t,t-i}$.

The most common approach taken in the literature to avoid this sample bias is to require individual forecasters to have made at least a large number of predictions. For example, Capistrán and Timmermann (2009a) or Elliott et al. (2008) require forecasters to have made at least 30 and 20 predictions respectively. This requirement does not directly address the potential sample bias. However, one would assume that forecasters who made quite a few forecasts are less likely to only have predicted during periods when most forecasters tended to over (under) predict the underlying variable. At the same time, this method substantially reduces the sample to institutions which could cause small sample biases.[16]

To test, whether the above approaches indeed reduce the number of biased forecasters, I estimate equation (7) at the individual level, requiring forecasters to have contributed a varying number of forecasts. I report the share of biased forecasters based on the 5% level for OLS errors.[17] While I cannot directly measure the share of biased forecasters controlling for missing data, I can introduce a new approach to identify forecasters that are likely to be affected by the identification issue provided the simple average is unbiased over the entire sample. In particular, I can replace the forecasts made by a specific forecaster by the simple average. This will leave in place the pattern of missing observations, but replace the potentially biased forecasts with overall unbiased values. In addition, the simple average is likely to have the same systemic biases that cancel out over time due to the high correlation among forecasts. If I then estimate equation (7) based on this data, I either find that the simple average is biased for this specific sub sample or that it is unbiased. If it is unbiased, there is likely

[16] This could be exacerbated if some institutions have an asymmetric loss function and seek publicity as described by Laster et al. (1999). The publicity seeking could be an incentive to contribute every period.

[17] Due to missing observations, HAC errors are not feasible.

to be no correlation between α and $\varepsilon_{t,t-i}$ and the biased forecasters are correctly identified. If it is biased, it is quite likely that α and $\varepsilon_{t,t-i}$ are correlated for that forecaster and he will likely be identified as biased. This is independent from him actually being biased or not. Reporting the share of forecasters for whose sample the simple average is biased as well can thus provide an upper bound to the share of forecasters being falsely identified as being biased.[18]

I report the results in Table 7 for the Bloomberg survey and for the SPF as a cross check to ensure that the results are not survey specific. While it is the case that the numbers are broadly decreasing when the number or required contributions is increased, the decreases are not very large. What is more, the share of biased forecasters even increases sometimes when the number of required forecasts to be included is increased. For example, the 30 period restriction reduces the share of biased forecasters only by 5% for H2 and H4 relative to the two period restriction, while the number of included forecasters is reduced to less than 10 %.[19]

Next I report the share of forecasters who are missing data in such a way that the simple average is biased over the same periods for forecasters that contributed at least 10 forecasts shown in the column SA10 in Table 7. I find that there is quite a large share of forecasters that have this identification problem. Indeed they make up more than half of the biased forecasters in almost all but one cases. These results are quite similar for larger minimum numbers of contributed forecasts, especially for the SPF where the number of forecasters does not decrease as much. From this I can conclude that the estimated share of biased forecasters at an individual level is likely to be overestimated even if forecasters are required to have contributed quite a number of forecasts.

So far I have only looked at CPI forecasts. I will now turn to the other variables. As shown above, increasing the number of required forecasts in the

[18] Provided the simple average is overall unbiased and assuming that the share of forecasters being biased is independent from the simple average being biased for their sample or not, this method could be used to directly estimate the share of biased forecasters in two steps. In the first step, one would check if the simple average is unbiased for a given forecaster. If it is biased, the forecaster is dropped from the analysis as a second step. If it is unbiased, one can check if the forecaster is biased and obtain the overall share of biased forecasters in the second step. This approach would cut the share of biased forecasters roughly in half as compared without this extra step.

[19] While the row only shows the number of forecasters satisfying the contribution requirement for H4, a clear decrease can be found for the other horizons, albeit to a lesser extent.

Table 7: Share of biased CPI forecasters, when different thresholds are used to handle missing data for the SPF and the Bloomberg survey.

	Bloomberg				
	2*	10*	20*	30*	SA10**
H0	8.91	8.85	11.29	5.00	2.65
H1	25.17	25.55	15.07	6.52	21.17
H2	26.83	28.18	22.95	20.51	16.36
H3	36.36	29.09	22.58	20.51	18.18
H4	36.71	29.67	20.93	31.25	24.18
N (H4)	207	91	43	16	91
	SPF				
	2*	10*	20*	30*	SA10**
H0	18.66	18.90	12.50	12.00	11.81
H1	26.79	25.40	23.61	26.00	19.05
H2	39.02	38.10	34.29	35.42	30.95
H3	36.27	34.96	25.71	25.53	34.15
H4	40.00	39.32	29.41	31.11	35.90
N (H4)	205	117	68	45	117

Shares are based on the 5% threshold with OLS errors. *Only forecasters with at least that many forecasts are included. **This column reports the share of forecasters with at least 10 forecasts for whose sample the simple average is biased.

sample does not change the share of biased forecasters significantly. I will thus only report the share for forecasters that contributed at least 10 forecasts. Also, for the horizons and variables where the simple average does not have an overall bias I will calculate the share of forecasters who are missing periods in such a pattern that the simple average is biased over the same pattern as well.

Table 8 shows the share of biased forecasters. Similar to the findings when looking just at the CPI, I find that the share of biased forecasters is sizeable. This finding is particularly pronounced for bond yields, where I also found the overall bias. However, I also find that for a large share of forecasters this bias might simply stem from the pattern of missing data. This makes it likely that

Table 8: Percentage of biased forecasters

	Unemployment		GDP		Bond-Yield		CPI	
Horizon	10*	SA10$^{\#}$	10*	SA10$^{\#}$	10*	SA10$^{\#}$	10*	SA10$^{\#}$
H0	30.97	30.97	4.44	0	-	-	8.85	2.65
	[113]	[113]	[113]	[113]			[113]	[113]
H1	40.60	43.61	16.55	9.35	49.11	-	25.55	21.17
	[133]	[133]	[139]	[139]	[112]		[137]	[137]
H2	34.86	24.77	14.41	4.50	65.45	-	28.18	16.36
	[109]	[109]	[111]	[111]	[110]		[110]	[110]
H3	32.73	23.63	22.52	-	85.98	-	29.09	18.18
	[110]	[110]	[111]		[107]		[110]	[110]
H4	41.30	48.91	20.00	11.22	87.25	-	29.67	24.18
	[92]	[92]	[90]	[90]	[102]		[91]	[91]

Number of forecasters in square brackets. *Forecasters made at least 10 forecasts. $^{\#}$The share of forecasters with at least 10 forecasts for whose sample the simple average is biased.

the share of biased forecasters is quite over estimated.

Overall, I find that estimating equation (7) at the individual level will result in quite a sizeable number of biased forecasters. This finding does not change much by requiring forecasters to have contributed more or less forecasts. However, I found that missing data plays an important role in this finding. Many forecasters have likely been identified to be biased simply due to them missing some observations. Due to this, the share of biased forecasters is very likely to be overestimated and rendering this approach invalid. Further research might thus be warranted to specifically determine what share of individual forecasters is biased and resolve the identification problem described above.

4 Accuracy of Alternatives to the Simple Average

In this section, I look at two potential avenues on improving upon the accuracy of the simple average of expectations by taking into account the heterogeneity of individual forecasters found above. First, I identify and re-weight better and worse forecasters ex ante based on their track record using various approaches suggested in the literature. I then compare the performance of these approaches to the performance of the simple average. Second, I use the part of uncertainty measured by disagreement of forecasters to improve the prediction of recessions.

4.1 Identifying the Best Forecasters Based on Their Record

In the literature, there have been numerous studies as summarized in Clemen (1989) and Timmermann (2006) that have shown that historic performance is not always significantly improving over the simple average of all forecasters or models considered. Indeed, they come to the conclusion that the simple average is difficult to beat.

I will proceed to test if this is also the case for my dataset by comparing several commonly used alternatives to the simple average. I will test three approaches based on past performance and compare them to the simple average as well as two additional approaches that do not depend on past performance. Aside from the simple average, I will use the median and the 5% trimmed mean (Trim) as the additional approaches.[20] Both the median and the trimmed mean would perform better than the simple average if there are strong outliers in the data.

The three approaches that I will use that depend on past performance, are the recent best forecaster (RB), the inverse MSE (invMSE) as an approximation of the optimal Bates and Granger (1969) weights, and the subset approach proposed in Bürgi and Sinclair (2016).[21]

For the recent best and inverse MSE, I look at the performance of a given forecaster in the past 15 periods in real time. I calculate the MSE for each forecaster that made at least 10 forecasts during those 15 periods. This ensures that it is relatively unlikely that a forecaster will have a low MSE by chance. I then use this information to weight forecasts for future periods, taking into account only information that is known at the time it is used. This implies that the longer the horizon of the forecast, the longer the lag between assessing the performance of forecasters and the period for which this information is then used.

Bürgi and Sinclair (2016) have recently shown for the SPF that the high correlation among forecast errors and missing data can explain why it is this difficult to beat the simple average. As described above, most traditional MSE

[20] This drops the 5% highest and lowest forecasts each period.

[21] I cannot calculate the covariance matrix, as some forecasters do not overlap.

based methods for evaluating forecast performance have difficulties when there are missing data. In addition, a high correlation among forecasters will lead to a higher share of individual forecasters to perform above average by chance.

To mitigate these problems, they propose to create a subset of forecasters that have performed well in the past based on a non-parametric measure, which is less affected by missing data and high correlations. While this subset will still include forecasters that performed well by chance, the share of those forecasters in the subset is smaller than in subsets of best forecasters based on MSE performance. The non-parametric approach they propose includes forecasters in the subset that have been better than the simple average more than 52.5% in past periods. For example, if there are ten periods and the prediction of a forecaster is closer to the actual value than the simple average of predictions of all forecasters in six periods, he is included in the subset. If he was only closer in five periods, he will not be included. While the forecasters in the subset are relatively persistent, this approach also allows the set to change over time.

I will apply the same approach to my dataset, to see if that approach performs well with the Bloomberg dataset as well. Similar to Bürgi and Sinclair (2016), I will also use an expanding window and require forecasters to have made at least ten forecasts in the past.

Table 9 shows the performance of the various methods relative to the simple average. The RMSE of the simple average is normalized to one. From the table, it is clear that aside from the current quarter forecasts (H0), there are very few occasions where any of the five approaches tested performs significantly better than other approaches. Both the inverse MSE weighted approach and the subset approach improve significantly over the simple average in four cases. However, the subset shows much larger gains than the inverse MSE. The other three approaches are never significantly better than the simple average in my dataset. This confirms that the best forecaster in the past is not better than the simple average in the future.

While the subset approach does not perform particularly well for GDP, it still has a lower MSE than any of the other approaches based on past performance in 12 out of 19 cases. When comparing the results from Table 3 to Table 9 the subset was able to significantly improve for the most significant rejections of equal forecasting power.

Similarly to Bürgi and Sinclair (2016) I also find that the subset shows

Table 9: RMSE relative to the simple average

	GDP				
	Trim	Median	Subset	invMSE	RB
H0	1.00	0.99*	1.00	0.98***	1.08
H1	1.00	1.01	1.04	1.01	1.12
H2	1.00	1.00	1.03	1.02	0.97*
H3	1.00	1.01	1.00	1.02	1.10
H4	1.00	1.01	1.04	1.02	1.03
	CPI				
	Trim	Median	Subset	invMSE	RB
H0	1.00	1.05	0.75***	0.91***	1.13
H1	0.99	1.00	1.07	0.97**	1.07
H2	1.00	1.01	0.97*	0.98	1.26
H3	1.00	0.98*	1.00	0.99	1.01
H4	0.96	0.95*	1.01	1.01	1.21
	Unemployment				
	Trim	Median	Subset	invMSE	RB
H0	1.00	1.04	0.95***	0.97***	1.02
H1	1.00	0.99	0.98	1.00	1.02
H2	1.00	1.01	0.97	0.99	0.97
H3	1.00	1.00	1.00	1.01	1.08
H4	1.00	1.00	0.98	1.00	0.99
	Bond Yield				
	Trim	Median	Subset	invMSE	RB
H1	1.00	0.99	1.01	1.01	1.01
H2	0.99	1.00	0.94**	1.00	0.98
H3	0.99	1.01	0.96	0.98	0.95
H4	0.99	0.99	0.92**	0.99	0.99

* significant improvement over the simple average at 10% level, ** at 5% level and *** at 1% level based on DM statistic.

the most significant gains for bond yields in the Bloomberg survey. This is important for investors who trade based on these forecasts.[22] One reason, why this method works particularly well for bond yields could be that unlike the other variables there are many more significantly better forecasters than significantly worse forecasters for bond yields as discussed in the section on accuracy of the expectations of individuals.

4.2 Predicting Recessions

In the literature, it has been shown that professional forecasters have difficulties predicting recessions (see Zarnowitz (1986), Fintzen and Stekler (1999) and Sinclair et al. (2010)). As shown in the bias section above, forecasters in the Bloomberg survey appear to have similar difficulties. Given this inability to predict recessions, researchers have ventured into the distribution of and uncertainty around forecasts. During uncertain times, the distribution of forecasters might be different from relatively more certain times and this pattern might be related to recessions (e.g. Lahiri and Teigland (1987) and Zarnowitz and Lambros (1987)).

At the same time, there has been an extensive literature on how uncertainty can translate into lower investment and thus be related to business cycles (e.g. Bloom (2009) or Christiano et al. (2014)). Due to this, there is the possibility that uncertainty can be used as a leading indicator for predicting recessions as well.

In the literature, there are two broad sources for measuring uncertainty. The first source is market or news data, which allows to come up with a measure of uncertainty (e.g. Bloom (2009) uses the VIX). While these measures tend to be at a very high frequency and forward looking, they tend to be quite volatile and event based. As the events might not be particular to any country, these indices might be better at capturing global uncertainty.

The other source is the use of economic forecasts to infer the level of uncertainty. Some surveys ask respondents about the uncertainty around their forecasts. This subjective measures of uncertainty can then be aggregated into

[22] However, as shown in the previous section, the accuracy is not better than a random walk. But, investors often look at the directional accuracy instead of the MSE (e.g. see Mitchell and Pearce (2007)). While the directional accuracy for both the simple average and the subset is better than 50% for H2 and H4 (albeit not statistically significantly different), the subset is more accurate for both horizons as well (two periods for H2 and one period for H4).

an overall measure of uncertainty (e.g. see Giordani and Söderlind (2003) or Engelberg et al. (2009)).

Alternatively, it can be decomposed into two types of uncertainty as emphasized in Batchelor and Dua (1993), Bomberger (1996), Lahiri and Sheng (2010) and Ozturk and Sheng (2016). The first type of uncertainty is the common uncertainty, which can be measured through (G)ARCH models (e.g.Engle (1982)). Broadly speaking if forecast errors of the simple average were large over a period of time, there was larger common uncertainty than in periods with overall smaller errors. The second type of uncertainty is idiosyncratic uncertainty or disagreement. Disagreement can be measured by the interquartile range of forecasts or their cross sectional standard deviation.

While it would be optimal to use the full uncertainty, the common uncertainty requires the actual data as well and thus is not as forward looking as the idiosyncratic uncertainty. I will thus focus my analysis on disagreement. In the forecasting literature, it has also been established that the spread between forecasters tend to help predicting the underlying variable (e.g. Driver et al. (2013)). While I have a relatively short sample in the Bloomberg survey with GDP forecasts starting only after the 90s recession, it is still the case that GDP was expected to grow faster for some periods than in others. This could lead to a higher disagreement purely due to faster growth. To avoid this, I use a modified coefficient of variance mCV defined as

$$mCV = \frac{stdev}{\max(1, abs(SA))}, \tag{9}$$

that is if the absolute value of the underlying variable is less than one, the denominator of the coefficient of variance is set equal to one. I proceed to calculate the mCV at the H1 horizon for GDP and use this as a predictor if there will be a recession in the next quarter. Also, the H1 horizon has the longest history, which is why I chose this horizon.

There is also a literature on predicting recessions independent from uncertainty. For example, Dueker (1997) and Proaño and Theobald (2014) showed that dynamic probit models are able to predict recessions quite well. Due to these findings, I will also run a dynamic probit regression of the form

$$\phi_t = \alpha + \beta y_{t-1} + \gamma x_t + \varepsilon_t, \tag{10}$$

where ϕ_t is the recession dummy that takes value one if there is a recession and value zero otherwise, y_{t-1} is the previous period real GDP growth rate and x_t

are additional variables included in the regression.[23] Given that there are only two recession periods with a total of 11 recession quarters in my dataset, out of sample testing is not possible.

My first specification (1) will only use past real GDP growth and thus $\gamma = 0$. My second specification (2) uses real GDP and the simple average of the Bloomberg survey as x_t. The last specification (3) replaces the simple average by the standard deviation of the forecasts made for GDP. I will assess the performance of the three models based on three different measures. The first two measures are mean absolute and root mean squared errors.[24] The third measure I use is the Theil inequality coefficient, defined as

$$Theil = \frac{\sqrt{\sum_{t=1}^{T}(A_t - F_{t,t-h})^2}}{\sqrt{\sum_{t=1}^{T} A_t^2} + \sqrt{\sum_{t=1}^{T} F_{t,t-h}^2}}, \tag{11}$$

where $F_{t,t-h}$ is the forecast made in period t-h for period t and A_t the actual data at period t.[25]

I report the results in Table 10 and find that the in sample performance improves clearly, when either the simple average or the modified CV of forecasters is included into the specification. This improvement might in part be due to forecasters already having some information about the current quarter (e.g. unemployment or ISM), which is not incorporated in the GDP release for the previous month.

While there is a clear improvement when including the either the simple average or the mCV, there is also a small improvement of the mCV model over the simple average model. In terms of miss specified quarters, the difference is around two. This implies that using the mCV model will lead to two more quarters correctly being specified relative to the simple average model.

While I am only able to do in sample testing, I can plot the mCV series and highlight recessions. As figure 4.2 shows, the signal obtained from the mCV is a strong one, as the shaded areas have a much higher mCV value than most of the periods outside.

[23] At the time the survey is conducted, the second release of the real GDP growth rate of the previous quarter is available, which is the one I will use.

[24] Lahiri and Wang (2013) call the root mean squared errors quadratic probability score and Brier (1950) score.

[25] Often, the measure is stated as averages instead of sums, which is equivalent as the $1/T$ cancels out.

Table 10: Probit results

	(1)	(2)	(3)
MAE	0.21	0.10	0.08
RMSE	0.32	0.23	0.17
Theil	0.62	0.37	0.26

In sample measures of fit for the dynamic probit model (1), (1) with the simple average of real GDP forecasts (2) and (1) with the modified CV (3).

5 Information Rigidity

In this section, I look into one aspect of how expectations are created. In particular, many economic models rely on some form of rigidity. I delve deeper into the information rigidity based on how forecasters revise their expectations.

5.1 Information Rigidity Based on the Simple Average

Given the survey collects forecasts across different horizons, it is possible to investigate if forecast errors are uncorrelated with forecast revisions. If revisions are significantly positively (negatively) correlated, the simple average will have under (over) incorporated all new information.

There have been proposed several models to explain, why forecasters might under or over revise their forecasts as found for some of the variables and horizons. Ehrbeck and Waldmann (1996) show that over or under adjustment might be due to concerns about the reputation of a forecaster in a game theory frame work. For under adjustment, the reasoning is that forecasters might smooth their forecasts over time as changing a forecast often might be interpreted as the traits of bad forecasters. For over adjustment, the reasoning is that forecasters might get individual private signals. Forecasters whose private signal does not have a lot of noise will significantly update their forecasts. Due to this, forecasters with noisier signals might imitate the ones with better signals, leading to an over adjustment overall.

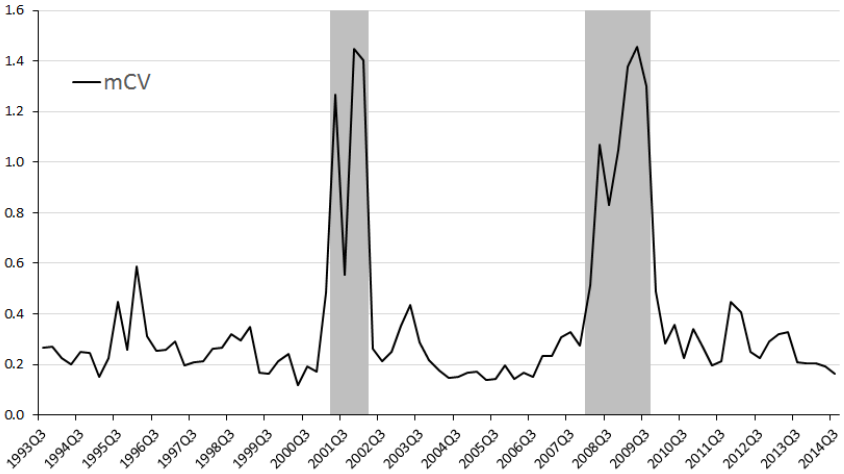

Figure 1: mCV for H1 GDP forecast with recession shading for the forecasted period.

Coibion and Gorodnichenko (2015) and Andrade and Bihan (2013) offer an alternative explanation for forecast under adjustment. They linked it to information rigidity in the form of two models. The first model is sticky information as suggested by Mankiw and Reis (2002). In this model, agents do not update their information set every period. In every period, they only receive new information with probability $1 - \lambda$. In a forecasting setting, this implies that their forecasts are only updated with that probability and remain unchanged with probability λ, as their information set has not changed. When information is updated however, agents rationally update their forecasts. This implies that for the aggregate, assuming in infinite number of forecasters,

$$F_{t,t-i} = (1 - \lambda) \sum_{j=0}^{\infty} \lambda^j E_{t-j-i} A_t, \tag{12}$$

where $E_{t-j-i}A_t$ is the expected value of A_t taken in period $t - j - i$. That is, the average forecast made in period $t - i$ is equal to the weighted average of current and past rational forecasts. Taking the first difference and rearranging, one obtains the equation

$$A_t - F_{t,t-i} = \frac{\lambda}{1 - \lambda}(F_{t,t-i} - F_{t,t-i-1}) + \varepsilon_{t,t-i}; \qquad i = 0, 1, ..., 3. \tag{13}$$

The second model is an imperfect information model, based on Lucas (1972) and Finn E. Kydland (1982). In this model, agents update their information sets every period, but information is noisy. This noisy information will cause agents to put some weight $1-G$ on old information. The new forecast will then be the weighted average between the old forecast and new information. This model might be set up in the following way: Assume the underlying variable follows an AR(1) with the iid error ε_t and forecasters j receive a noisy signal x_{jt-i} of the form

$$x_{jt-i} = A_{t-i} + \nu_{jt-i}, \tag{14}$$

where ν_{jt} is an iid error. Each forecaster then can generate a forecast based on this signal using the Kalman (1960) filter

$$F_{jt,t-i} = (1-G)x_{jt-i} + GF_{jt,t-i-1}. \tag{15}$$

Given the AR(1) assumption on the underlying model, this can be rearranged to

$$A_t - F_{jt,t-i} = \frac{G}{1-G}(F_{it,t-i} - F_{jt,t-i-1}) - \nu_{jt-i} + \sum_{k=1}^{i} \rho^{i-k}\varepsilon_{t-i}. \tag{16}$$

If this expression is averaged across individuals, ν_{jt-i} will cancel out due to the iid assumption and it simplifies to

$$A_t - F_{jt,t-i} = \frac{G}{1-G}(F_{it,t-i} - F_{jt,t-i-1}) - \sum_{k=1}^{i} \rho^{i-k}\varepsilon_{t-i}. \tag{17}$$

This expression is very similar to equation (13), but here the error term is auto correlated and the slope coefficient can be interpreted as the weight given to old information relative to new information. This expression is also very similar to the more general Nordhaus (1987) test for weak forecast efficiency as specified in the following equation.

$$A_t - F_{t,t-i} = \alpha + \beta(F_{t,t-i} - F_{t,t-i-1}) + \varepsilon_{t,t-i}; \qquad i = 0, 1, ..., 3, \tag{18}$$

Both the sticky information and noisy information model restrict α to be zero and β to be positive. There is no restriction on the sign of β for the reputation models. Given the Nordhaus (1987) test is a more general as it allows for over adjustment as well, I estimate the model based on equation (18) and will use HAC errors throughout, as the forecasts are serially correlated.

Table 11: Nordhaus efficiency test across variables and horizons

	Unemployment		GDP		Bond-Yield		CPI	
Horizon	α	β	α	β	α	β	α	β
H0	-0.04	0.31*	0.24	0.39***	-	-	-0.01	0.35***
	(0.03)	(0.17)	(0.14)	(0.10)			(0.04)	(0.11)
H1	-0.08*	0.67**	-0.16	-0.11	-0.24***	-0.06	0.08	0.26**
	(0.05)	(0.29)	(0.23)	(0.41)	(0.07)	(0.12)	(0.10)	(0.13)
H2	-0.07	0.84***	-0.44	0.25	-0.50***	-0.51***	0.11	0.12
	(0.08)	(0.31)	(0.27)	(0.78)	(0.12)	(0.18)	(0.17)	(0.19)
H3	-0.04	-0.74***	-0.75*	2.18	-0.63***	-0.21	0.22	-0.05
	(0.03)	(0.06)	(0.43)	(2.33)	(0.14)	(0.17)	(0.31)	(0.31)

Standard errors in brackets. * significant at 10% level, ** at 5% level and *** at 1% level based on HAC errors.

I present the results in table 11 for the simple average and find mixed results, similar to Dovern and Weisser (2011), Lahiri and Sheng (2008) and Messina et al. (2015). There is little evidence of under or over adjustment based on new information for the simple average of GDP or CPI forecasts at longer horizons. For the short horizon, I find that forecasters tend to under revise their forecasts. For unemployment, it appears that new information is under adjusted at the shorter horizons but over adjusted at the very long horizons.

For bond yields, it appears that forecasters tend to systemically over predict bond yields. This finding is not very surprising as this is consistent with the upward bias found in the previous section. At the same time, it appears that new information is over adjusted at H2.

Due to the mixed results at the level of the simple average, it is clear that the behaviour of forecasters is different for different horizons. At the same time, both information rigidity models have the same implication at the aggregate level which makes it difficult to make the case for either of these models. However, all these models have some implications at the individual level as well, which I will address next.

5.2 Information Rigidity Based on Individual Forecasters

Any under adjustment found at the aggregate level could stem from individual forecasters under adjusting or the aggregation process. That is, forecasters might efficiently update their forecast but not change their forecasts every period. Due to this, it would look as if they under adjust at the aggregate level even if each individual forecaster efficiently updates his or her forecast. As the models described above have different implications for the individual level, I can test the models at that level as well.

The imperfect information model suggests that individual forecasters should have a positive β, just like the aggregate level. This means that the under adjustment does not arise from the aggregation process, but is due to individual forecasters under adjusting. At the same time, they should update their forecasts every time there is new information and there should be relatively few periods, where forecasters do not revise their forecasts.

The sticky information model has a different implication for individual forecasters. Here, individual forecasters are assumed to be rational, but they only periodically update their forecasts. Indeed in its strictest form, this model only allows forecasters to update their model with probability $1 - \lambda$. This implies that if equation (18) was run for individual forecasters, one would expect $\beta = 0$ for most forecasters. The aggregate smoothness of forecasts then is mainly due to the aggregation process and not individuals smoothing their forecasts.

The reputation model allows for either over adjustment or under adjustment. However, similarly to the case with the imperfect information model, all forecasters are expected to show the same behaviour. Due to this, aggregation should not matter for this model either and the coefficients found at the aggregate level should be the same as found at the individual level.

Coibion and Gorodnichenko (2015) already did some analysis to determine if the sticky information model or the noisy information model more accurately explains the pattern in the CPI data. However, they only focused their analysis on the simple average. They concluded due to different coefficients for different variables and noisier series having higher information rigidity that imperfect information models are more likely to explain the information rigidity observed in their data. To remain relatively close to their paper, I will also focus exclusively on CPI forecasts.

One reason why they did not repeat the analysis at the individual level is ν_{jt-i} in equation (16). This error is the noise in the signal received by the forecasters. As the forecasts are based on the noisy signal, it is quite likely that this noise is correlated with the updated forecast. Due to this, equation (16) cannot be estimated by OLS and requires an instrument that is correlated with the revision but not with this individual error. I found a natural instrument that allows this analysis at the individual level. The average revision will not be correlated with the individual error by assumption but is correlated with the individual revision. This allows me to use the average revision as an instrument for the individual revision to obtain an unbiased estimate for $\frac{1-G}{G}$. Unfortunately, this IV approach cannot simultaneously test assumptions of the sticky information model, because there the information rigidity arises from aggregation and thus the first stage regression is not valid. This will require me to run two separate regressions for these two models. In turn, the sticky information regression can also test the reputation model, as there is no issue that requires an IV approach (see Ehrbeck and Waldmann, 1996).

As mentioned in previous sections above, missing data makes testing these three models also more complicated. To ensure that the coefficients are accurately estimated, I require forecasters to have contributed at least 30 data points to estimate either of the two models for them. I first calculate the simple average for the reduced number of forecasters to ensure that they still exhibit expectation smoothing at the aggregate level.[26] The results of this analysis is shown in the first column of table 12.

I then proceed in calculating the coefficients based on the individual forecasters. I use a panel data approach similar to Davies and Lahiri (1995). I estimate the model either using a standard random effects model with HAC errors or a panel IV regression. The first thing I note is that forecasters tend to over adjust their expectations differently across horizons. At short horizons CPI forecasters under adjust most while at long horizons they over adjust. This pattern is more pronounced for the sticky information model relative to the noisy information model. Given the small sample several individuals have in my dataset, I do not report the corresponding standard errors for the median forecaster.

From the sticky information regression in columns two and four of Table 12, it is clear that individual forecasters do not smooth their forecasts to the same degree as the aggregate. Indeed, it appears to be the case that the aggregation process adds roughly 30 basis points to the average of the individual forecasters for all horizons. This finding is consistent with the sticky information model,

[26] While there is a sizeable number of forecasters for H0-H2 that contributed at least 30 forecasts, there are only two for H3. Due to this, I do not report the results.

Table 12: Efficiency of individual forecasters

	30 Observations	All			
	Simple Average	Panel	Panel IV	Median	Median IV
H0	0.35***	0.01	0.32***	-0.01	0.18
	(0.11)	(0.03)	(0.05)		
H1	0.24**	-0.07**	0.21***	-0.09	0.29
	(0.12)	(0.03)	(0.07)		
H2	-0.01	-0.24***	-0.07	-0.27	0.05
	(0.15)	(0.04)	(0.08)		
H3				-0.51	-0.29

Standard errors in brackets. * significant at 10% level, ** at 5% level and *** at 1% level based on HAC errors. Simple average is the Nordhaus test based on the average of forecasters with at least 30 observations. Panel is the panel estimate of the common coefficient and panel IV is the coefficient when estimated using IV. Median is the median of the individual coefficients and Median IV is the median when estimated using IV.

but not with the reputation model. At the same time, it appears that aside from the very short horizon, forecasters do over adjust their forecasts.

From the noisy information IV regression, I can learn that the behaviour of forecasters is broadly in line with that model as well. Indeed, the average coefficient is broadly in the area of the coefficient of the average of all forecasters with at least thirty observations. The median for all forecasters in column five of Table 12 shows some bigger differences relative to the coefficient of the average reported in Table 11. The individual weight appears to be lower than the aggregate weight, which implies that the aggregation process also causes some of the forecast smoothing, contrasting with the noisy information model.[27]

An alternative way of testing these models is to look at the share of periods, where forecasters do not change or only minimally change their forecasts. Independent of the horizon, forecasters do not change their forecasts 10%-15%

[27] Given that I use the simple average as my instrument, one would expect that if there is an aggregation effect, it would already be mostly captured by the instrumented variable. Due to this, the aggregation effect would be underestimated.

of times and minimally change it (by up to 0.1) roughly 25%-30% of times.[28] To get the 25-35 basis point aggregation effect found in the data, it is required that roughly 25% of forecasters do not change their forecast under the sticky information model. This implies that aggregation appears to play an important role in explaining the information rigidity.

I am also able to confirm the finding of Dovern et al. (2015) in my data set. That is, forecasters tend to make larger revisions at shorter horizons, relative to longer horizons. Indeed, the share of unchanged forecasts is 10% for current horizon forecasts and 15% for three quarter horizon forecasts. Similarly 33% of forecasters change their forecasts by less than 0.2 at the current quarter horizon and 46% change it by the same amount at H3. This implies that about half of the revisions at the longer horizon are larger than 0.1, while two thirds of the revisions at the shorter horizon are larger than 0.1.[29] This finding is at odds with both sticky information and noisy information. For both models one would expect smaller revisions as the forecast horizon becomes shorter.

I also repeated the analysis with the SPF and the results are very similar to the ones found with the Bloomberg survey.

Together, my findings show that while there are information rigidities, they alone cannot explain the pattern found in the data. My regression results cannot reject either of the two models for information rigidities, but there is some evidence that there is an aggregation component to it. This would point more towards a sticky information model, rather than a noisy information model and somewhat contrast with the results in Coibion and Gorodnichenko (2015), who favoured the noisy information explanation for the information rigidity. Due to the aggregation component, my evidence also points against reputational models.

At the same time, I find that revisions become smaller the shorter the horizon, which neither of these models can explain. Also, I find that forecasters tend to over adjust their forecasts at longer horizons. Both these things imply that these models miss some important factors that explain the forecast revision and thus information updating process found in the data.

[28] As Burgi and Guo (2016) show, this measure might only capture between one third to two thirds of unchanged forecasts.

[29] All forecasts are made up to a tenth.

6 Conclusion

In this paper, I wanted to assess if the additional information learned from evaluating individual expectations provides important insights that warrant the additional costs and difficulties this entails. To answer this question, looked into three aspects of the US Bloomberg Survey of forecasts for the real GDP growth rate, the CPI inflation, the unemployment rate and 10 year Treasury bond yields in four categories.

First, I found that the simple average of the Bloomberg survey is statistically significantly more accurate than the random walk for most variables. I also found that there are significant differences in the performance of individual forecasters based on my newly introduced rank based approach. With this approach, I was also able to rank individual forecasters in my sample based on their joint performance across the four variables after correcting for the correlation among the variables.

Regarding rationality of expectations, I found that the survey is biased upward for bond yields but not for the other variables at the aggregate level and that there are systemic biases that cancel out over time for the unemployment rate and the real GDP growth rate. At the individual forecaster level, I found that applying the same method for estimating biases as at the aggregate level will lead to sizeable shares of forecasters being biased for most variables. However, I find that this result is likely due to missing observations and show that there is an identification issue. Some forecasters might only have contributed in periods, when forecasters tended to over predict the underlying variable. This can cause many more forecasters to be identified as biased just because of the pattern of missing data, independent of having contributed few or many forecasts.

Second, I tried to improve the accuracy of expectations by looking at the distribution as well and not just the simple average. Weighting forecasters based on their past performance applying the Bürgi and Sinclair (2016) approach, allowed me to improve over the simple average for a number of variables and horizons.

Also, I found that the part of uncertainty that is measured by disagreement is able to predict recessions quite well. It is a better predictor of recessions than the simple average of growth forecasts in a dynamic probit model for the past two recessions and much better than the latest available real GDP growth rate. Relative to the simple average, including uncertainty will correctly predict two more quarters in my sample.

Third, I looked into the expectation formation process or forecast efficiency and find some evidence of information rigidity for shorter horizons but very little evidence for longer horizons at the aggregate level. Indeed, forecasters tend to over adjust their forecasts at longer horizons. At the individual level, I am able to provide some supporting evidence for both the noisy information and sticky information models. Under the sticky information assumption, my empirical results point towards efficient revisions, while under the noisy information assumption, my model suggests that individuals behave similarly to the aggregate. In both cases however, the aggregation process appears to play some role even though not always statistically significant. This finding gives a slight edge to the sticky information model in my view.

Overall, I found that focusing on aggregate expectations can lead to quite different results to the ones obtained at the individual level. In particular, looking at the forecasts made by individual forecasters allows me to reject and confirm certain assumptions of economic models that the simple average cannot.

References

Ager, P., Kappler, M., and Osterloh, S. (2009). The accuracy and efficiency of the consensus forecasts: A further application and extension of the pooled approach. *International Journal of Forecasting*, 25(1):167 – 181.

Alkire, S. and Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics*, 95(7-8):476–487.

Andrade, P. and Bihan, H. L. (2013). Inattentive professional forecasters. *Journal of Monetary Economics*, 60(8):967 – 982.

Antonelli, G. B. (1886). *Sulla Teorie Matematica Della Economie Politica.* Nella Tipografia del Folchetto, Pisa. Reprinted in Giornale degli Economisti e Annali di Economia, Nuova Serie, 10, 233-263, 1951.

Ashiya, M. (2006). Forecast accuracy and product differentiation of japanese institutional forecasters. *International Journal of Forecasting*, 22(2):395 – 401.

Baghestani, H. (2006). An evaluation of the professional forecasts of u.s. long-term interest rates. *Review of Financial Economics*, 15(2):177–191.

Banternghansa, C. and McCracken, M. W. (2009). Forecast disagreement among fomc members. *Federal Reserve Bank of St. Louis Working Paper No.*

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

Batchelor, R. (2007). Bias in macroeconomic forecasts. *International Journal of Forecasting*, 23(2):189 – 203.

Batchelor, R. and Dua, P. (1993). Survey vs arch measures of inflation uncertainty. *Oxford Bulletin of Economics and Statistics*, 55(3):341–353.

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

Bloom, N. (2009). The impact of uncertainty shocks. *econometrica*, 77(3):623–685.

Bomberger, W. A. (1996). Disagreement as a measure of uncertainty. *Journal of Money, Credit and Banking*, 28(3):381–392.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Burgi, C. and Guo, X. (2016). Mechanical forecast changes. *Unpublished*.

Bürgi, C. and Sinclair, T. (2016). A Nonparametric Approach to Identifying a Subset of Forecasters that Outperforms the Simple Average. *Empirical Economics*, (Forthcoming).

Capistrán, C. and Timmermann, A. (2009a). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking*, 41(2-3):365–396.

Capistrán, C. and Timmermann, A. (2009b). Forecast combination with entry and exit of experts. *Journal of Business & Economic Statistics*, 27(4):428–440.

Carroll, C. D. (2003). Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics*, 118(1):269–298.

Chen, L. H., Jiang, G. J., and Wang, Q. (2013). Market reaction to information shocks - does the bloomberg and briefing.com survey matter? *Journal of Futures Markets*, 33(10):939–964.

Cho, D. W. (1996). Forecast accuracy: Are some buisiness economists consistently better than others? *Business Economics*, 31(4):45–49.

Chong, Y. Y. and Hendry, D. F. (1986). Econometric evaluation of linear macroeconomic models. *The Review of Economic Studies*, 53(4):671–690.

Christiano, L. J., Motto, R., and Rostagno, M. (2014). Risk shocks. *The American Economic Review*, 104(1):27–65.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110.

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

Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–78.

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.

Dovern, J., Fritsche, U., Loungani, P., and Tamirisa, N. (2015). Information rigidities: Comparing average and individual forecasts for a large international panel. *International Journal of Forecasting*, 31(1):144 – 154.

Dovern, J. and Jannsen, N. (2015). Systematic Errors in Growth Expectations over the Business Cycle. Kiel Working Papers 1989, Kiel Institute for the World Economy.

Dovern, J. and Weisser, J. (2011). Accuracy, unbiasedness and efficiency of professional macroeconomic forecasts: An empirical comparison for the {G7}. *International Journal of Forecasting*, 27(2):452 – 465.

Driver, C., Trapani, L., and Urga, G. (2013). On the use of cross-sectional measures of forecast uncertainty. *International Journal of Forecasting*, 29(3):367–377.

Dueker, M. (1997). Strengthening the case for the yield curve as a predictor of U.S. recessions. *Federal Reserve Bank of St. Louis Review*, 79(2):41–51.

Ehrbeck, T. and Waldmann, R. (1996). Why are professional forecasters biased? agency versus behavioral explanations. *The Quarterly Journal of Economics*, pages 21–40.

Elliott, G., Komunjer, I., and Timmermann, A. (2008). Biases in macroeconomic forecasts: irrationality or asymmetric loss? *Journal of the European Economic Association*, 6(1):122–157.

Engelberg, J., Manski, C. F., and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, 27(1):30–41.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.

Ericsson, N. R. and Marquez, J. (1993). Encompassing the Forecasts of U.S. Trade Balance Models. *The Review of Economics and Statistics*, 75(1):19–31.

Fildes, R. and Stekler, H. (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics*, 24(4):435–468.

Finn E. Kydland, E. C. P. (1982). Time to build and aggregate fluctuations. *Econometrica*, 50(6):1345–1370.

Fintzen, D. and Stekler, H. (1999). Why did forecasters fail to predict the 1990 recession? *International Journal of Forecasting*, 15(3):309–323.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

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.

Giordani, P. and Söderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, 47(6):1037–1059.

Gorman, W. M. (1961). On a class of preference fields. *Metroeconomica*, 13(2):53–56.

Greer, M. R. (1999). Assessing the soothsayers: An examination of the track record of macroeconomic forecasting. *Journal of Economic Issues*, 33(1):77. Zuletzt aktualisiert - 2013-02-23.

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

Harvey, D. S., Leybourne, S. J., and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics*, 16(2):254–259.

Holden, K. and Peel, D. A. (1990). On testing for unbiasedness and efficiency of forecasts. *The Manchester School*, 58(2):120–127.

Jordà, O., Knüppel, M., and Marcellino, M. (2013). Empirical simultaneous prediction regions for path-forecasts. *International Journal of Forecasting*, 29(3):456–468.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Lahiri, K., Peng, H., and Zhao, Y. (2013). Machine Learning and Forecast Combination in Incomplete Panels. Discussion Papers 13-01, University at Albany, SUNY, Department of Economics.

Lahiri, K. and Sheng, X. (2008). Evolution of forecast disagreement in a bayesian learning model. *Journal of Econometrics*, 144(2):325 – 340.

Lahiri, K. and Sheng, X. (2010). Measuring forecast uncertainty by disagreement: The missing link. *Journal of Applied Econometrics*, 25(4):514–538.

Lahiri, K. and Teigland, C. (1987). On the normality of probability distributions of inflation and gnp forecasts. *International Journal of Forecasting*, 3(2):269 – 279.

Lahiri, K. and Wang, J. G. (2013). Evaluating probability forecasts for GDP declines using alternative methodologies. *International Journal of Forecasting*, 29(1):175 – 190.

Laster, D., Bennett, P., and Geoum, I. S. (1999). Rational bias in macroeconomic forecasts. *The Quarterly Journal of Economics*, 114(1):293–318.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19 – 46.

Lucas, R. J. (1972). Expectations and the neutrality of money. *Journal of Economic Theory*, 4(2):103–124.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

Mankiw, N. G. and Reis, R. (2002). Sticky information versus sticky prices: A proposal to replace the new keynesian phillips curve. *Quarterly Journal of Economics*, 117(4).

Messina, J. D., Sinclair, T. M., and Stekler, H. (2015). What can we learn from revisions to the Greenbook forecasts? *Journal of Macroeconomics*, 45(C):54–62.

Mincer, J. A. and Zarnowitz, V. (1969). The Evaluation of Economic Forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, NBER Chapters, pages 3–46. National Bureau of Economic Research, Inc.

Mitchell, K. and Pearce, D. K. (2007). Professional forecasts of interest rates and exchange rates: Evidence from the wall street journal's panel of economists. *Journal of Macroeconomics*, 29(4):840 – 854.

Nataf, A. (1953). Sur des questions d'agrégation en économétrie. *Publications de l'Institute de Statistique de l'Université de Paris*, pages 5–61.

Nordhaus, W. D. (1987). Forecasting efficiency: concepts and applications. *The Review of Economics and Statistics*, pages 667–674.

Okun, A. M. (1963). *Potential GNP: its measurement and significance*. Yale University, Cowles Foundation for Research in Economics.

Ozturk, E. and Sheng, X. S. (2016). Measuring global and country-specific unvertainty. *Working Paper*.

Proaño, C. R. and Theobald, T. (2014). Predicting recessions with a composite real-time dynamic probit model. *International Journal of Forecasting*, 30(4):898–917.

Rubaszek, M. and Skrzypczyński, P. (2008). On the forecasting performance of a small-scale DSGE model. *International Journal of Forecasting*, 24(3):498 – 512.

Scotti, C. (2013). Surprise and uncertainty indexes: real-time aggregation of real-activity macro surprises. International Finance Discussion Papers 1093, Board of Governors of the Federal Reserve System (U.S.).

Sinclair, T. M., Joutz, F., and Stekler, H. (2010). Can the fed predict the state of the economy? *Economics Letters*, 108(1):28 – 32.

Sinclair, T. M., Stekler, H. O., and Carnow, W. (2015). Evaluating a vector of the fed's forecasts. *International Journal of Forecasting*, 31(1):157–164.

Skillings, J. H. and Mack, G. A. (1981). On the use of a friedman-type statistic in balanced and unbalanced block designs. *Technometrics*, 23(2):171–177.

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.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Zarnowitz, V. (1986). The record and improvability of economic forecasting. Working Paper 2099, National Bureau of Economic Research.

Zarnowitz, V. and Lambros, L. A. (1987). Consensus and Uncertainty in Economic Prediction. *Journal of Political Economy*, 95(3):591–621.