WORKING PAPERS

Col.lecció d'Economia E15/320

Combine to compete: improving fiscal forecast accuracy over time

Laura Carabotta
Peter Claeys

Universitat de Barcelona

UB Economics Working Papers 2015/320

Combine to compete: improving fiscal forecast accuracy over time

Abstract: Budget forecasts have become increasingly important as a tool of fiscal management to influence expectations of bond markets and the public at large. The inherent difficulty in projecting macroeconomic variables – together with political bias – thwart the accuracy of budget forecasts.
We improve accuracy by combining the forecasts of both private and public agencies for Italy over the period 1993-2012. A weighted combined forecast of the deficit/ ratio is superior to any single forecast. Deficits are hard to predict due to shifting economic conditions and political events. We test and compare predictive accuracy over time and although a weighted combined forecast is robust to breaks, there is no significant improvement over a simple RW model..

JEL Codes: G12, C14, E43, E62, H62, H63.

Keywords: deficit, forecast accuracy, fiscal forecasting, forecast comparison, forecast combination, fluctuation test.

Laura Carabotta
Universitat de Barcelona

Peter Claeys
Université libre de Bruxelles

Acknowledgements: We thank Barbara Rossi, Domenico Giannone and participants at the Halle Macroeconometrics workshop, the III Time Series Conference in Zaragoza, and AQR-IREA for useful comments and suggestions. The usual disclaimer applies.

ISSN 1136-8365

1. Introduction

Budget forecasts are increasingly becoming a tool of fiscal management as the Financial Crisis led directly and indirectly to a fiscal meltdown in developed economies. Budget deficits that were rather contained in all industrialised economies before 2007 quickly gave way to deep budget deficits due to stimulatory tax cuts and spending hikes, financial bailouts, and the dragging on of the economic crisis. Forecasts of how the budget deficit will evolve are at the centre of political discussion in the US, Europe and Japan. In Europe, budgetary forecasts now play a key role in the preparation of economic measures under the European Semester, and in the monitoring of excessive deficits under the enhanced Stability and Growth Pact. Budget forecasts have always been a crucial part of the democratic policy process, but they are now becoming a key input of informed decision-making, and a tool to manage expectations of fiscal responsibility both towards financial markets and the public at large.

Evidence tells us that at present, budget forecasts are a poor guide to correctly assessing the fiscal outlook, especially if forecasts are produced by governments. The projections often paint a too rosy picture of reality, and are consistently biased towards too low deficits, especially when confronted with comparable predictions made by international institutions. Projections of fiscal adjustments are usually pushed forward over time, and revised when the decision nears (Beetsma and Giuliodori, 2010). A large literature argues that this bias in prediction performance is the consequence of setting politically motivated targets rather than realistic economic projections.

Research into better forecasting practices of budget variables has not come to more conclusive findings. The bottom-line of most applied work is that results depend on the choice of the forecasting procedure, the consistency of macroeconomic and fiscal forecasts, the forecast horizon and the level of disaggregation of fiscal forecasts. Fiscal forecasts may require as much judgement and expertise on budgetary developments as econometric or modelling techniques (Leal et al., 2008). The unsatisfactory implication is that little 'technical' improvement is possible as any progress depends on better knowledge of the dark box of the budget process.

The unfortunate deduction is often that forecasting the budget deficit is more of an art than a science, as there are as many forecasts as there are forecasters. The fortunate implication, we argue in this paper, is that we can exploit the information contained in individual budget forecasts. We are helped by efforts in recent years to make many budget forecasts publicly available.

In this paper, we use the judgment and expertise of many forecasters to construct a more accurate budget forecast. The way we do this is by averaging forecasts from different sources in a variety of ways. It is an established finding in the forecasting literature that averaging improves upon the forecast of any single model. We include simple as well as more advanced averaging techniques, which account for past forecasting performance, to compute a combined forecast.

Our main finding is that different combinations of budget forecasts result oftentimes in more accurate forecasts. This is particularly the case for a weighted forecast combination that values the more accurate forecasts.

Constant follow-up of forecast performance helps in improving accuracy. We show that predictive accuracy changes over time, and no single forecast is superior at all times, not even the combined forecast. Applying the predictive fluctuation test, we are able to establish that the weighted forecast combination is able to outperform other predictors in all years. Yet, the improvement in accuracy is not significant, even when compared to a random walk prediction.

Fiscal projections that are useful for government decisions are as much an art as a science, but substantial improvements are possible by using a set of budget forecasts and checking their performance. Our procedure advances over existing indicators with a 10% gain in the year ahead forecasts and a 5% improvement in accuracy of the current year forecast.

The paper is structured as follows. We first review in Section 2 several techniques for combining forecasts, and ways to evaluate and compare forecasts (over time). Data are discussed in Section3. In Section 4, we first discuss the tests to compare combination of forecasts to other forecast models, and then consider their evolution over time and in Section 5 we conduct some robustness checks. Section 6 concludes.

2. Combining forecasts

2.1 Forecast combinations

A vast literature shows that the combination of various forecasts of a single variable result in improved prediction performance (Clemens, 1989).[3] The reason for the improved performance is

[3] Many authors have approached the combination of forecasts. Zarnowitz (1967) noted that the published averages of inflation and GNP growth forecasts was better than the individual ones. Bates and Granger (1969) discovered that the simple average outperform the forecasts taken individually. The idea was also to use the relative combination of variances and covariances to construct a weighted average of the forecasts that minimizes the mean square error of the combined forecasts. Likewise, Nelson (1972) and Cooper and Nelson (1975) showed that the combination of forecasts with ARIMA estimates produces a smaller error compared to the models alone. The suggested reasons for the better performance of ARMA models in their paper are the incapacity of econometric models to arrange structural changes in the economy. Granger and Newbold (1973) also start from the similar point in terms of forecast evaluation. Makridakis (1982, 1983) studied a large variety of time series forecasting methods which were applied to 1,001 different economic time series. The forecast performance was measured using various error summary measures. Two different combining schemes were studied: both of these combinations performed well relative to the individual techniques, with the simple average having the better performance of the two.Clemens (1989) provided a very deep review of the methods used in combining and confirming these results. Clemens and Winker (1989) give root to a combination in their philosophical approach.

that single forecasts are the product of a specific forecasting model, which may include a set of econometric techniques and personal judgment, each with a specific error due to some aspect of the model. Combining the forecasts averages out the errors. Also, models used in forecasting are reflecting stable relationships, but in the real world, political events and crises change economic relations continuously. Combination levels out the instability and any structural breaks. Further, there are many macroeconomic variables that are endogenous in the economic cycle. This means that to forecast these variables, the methods that are used for forecasts use other variables to explain the former. These proxies introduce a systematic bias in the measurement of the real value and reduce forecast accuracy. In this case, combining reduces the risk of bias. Finally, some forecasting models are constructed to minimize past errors. Combination tends to avoid the selection of the best model by this process (Timmermann, 2006).

The aim of combination is to make forecasting practices robust to the different types of uncertainty and show robustness in various scenarios, rather than selecting a "true model" with the hope of explaining the most likely scenario in the future.

There are just two studies that support the claim that combination works well for budget projections. Marcellino (2002, 2004) studies the minimum mean squared forecast error (MSFE) and the mean absolute error (MAE) of a pooled forecast of IMF and OECD projections of the deficit ratio for G7 countries. Their results indicate that, on average, combination methods work well. Ozcan (2011) makes projections of the US deficit ratio from 1970 to 2005 and shows that forecast combinations of ADL models provide forecast gains relative to a simple AR model.

We compute simple combination models that average different forecasts (simple average, geometric average, harmonic average and median), and weighted models that give different weights to each forecasts by some criterion. A combined forecast is of the form:

$$\mathrm{Y}^*_{\mathrm{t+h}} = \alpha_{\mathrm{t}} + \sum_{\mathrm{t}} \beta_{\mathrm{it}} \dot{\hat{\mathrm{Y}}}_{\mathrm{it+h}} \qquad (1)$$

Where $\mathrm{Y}^*_{\mathrm{it+h}}$ is the combined forecast at h periods ahead of the variable obtained from the *i* individual sources. A considerable amount of research has been undertaken to determine how best to choose the coefficients, α_{t} and β_{it}.

Evidence suggests that the simple approach of averaging the individual predictions works well (Lupoletti and Webb, 1986; Clemen and Winkler, 1986; Clemens, 1989). In this case, β_{it} is equal to 1/n on all individual predictions. Similarly, the geometric mean and harmonic mean and the median can be used as a summary. The simple average has often been found to be the most robust forecast for a set of macro-economic variables, showing that forecasters are on average right (Clemens, 1989).

A weighted combination gives weights β_{it} that depend on the past performance of forecasters. A couple of possibilities exist. One is to construct the weights from a regression of the actual series on each of the forecasts. The coefficients on each forecast are then the weights. As explained in the article of Fildes and Stekler (2002) considering that there is no "best" forecast model, a more useful approach is to look for a combination of these models that may be able to provide better results.

Different weights can even result in better performance. We follow Stock and Watson (1999ab, 2004) and give to each individual predictor a weight that is inversely proportional to the predictor's Mean Square Forecast Error. A discount factor δ (0.90, 0.95 or 0.99) is applied to attach greater weight to the recent predictive ability of the individual predictor. The weights for (1) are then given as follows:

$$\beta_{\mathrm{i,t+h}} = {m_{\mathrm{i,t}}^{-1}}^{-1}_{\mathrm{i,t}} \Big/ \sum_{f=1}^{n} m_{\mathrm{f,t}}^{-1} \quad \text{where} \quad m_{\mathrm{i,t+h}} = \sum_{t}^{t-h} \delta^{t-h} \left(Y_{it+h}^{*} - \widehat{Y}_{it+h}\right)^{2} \qquad (2)$$

Rather than applying a discount factor, we alternatively cut off the past performance after some relevant period of time so as to exclude outdated versions of forecasting models.
Other combined model used is the "Rbest". It is related with the most recently best, and the weights on the individual forecast have the lowest average squared forecast error over the previous four periods.
This 'most recently best', which as implemented here places all weight on the individual forecast that has the lowest average squared forecast error over the previous four periods.

2.2 Forecasting test

2.2.1 Test of predictive accuracy

Following the approach of Artis and Marcellino (2001) and Keereman (1999) we test the accuracy of each fiscal forecast, a combination of those and a simple random walk. We analyse if a combination of forecasts outperforms the performance of any individual forecast, or of the AR model. The test used in the analysis are the common ones in the literature: RMSE (root mean squared error), MSE (Mean squared error), MAD (Mean absolute deviation), and the Theil's test which compares each forecast with a naïve no-change forecast. We also apply the Diebold-

Mariano test of predictive accuracy that compares a (single or combined) forecast with each other and with a simple random walk model.[4] The DM test (1995) is:

$$DM = \frac{\frac{1}{T}\sum_{1}^{t} T\{g(e_{1t+h})-g(e_{RW,t+h})\}}{\sqrt{2\pi}\ \hat{f}(0)/T} \quad (3)$$

With $g(e_{1t})$ and $g(e_{2t})$ denoting the loss from forecast error evolving from a prediction model and the random walk. The null hypothesis tested is that $H_0: E(e_{1,t+h}) = (e_{RW,t+h})$.

2.2.2 Fluctuation test

The DM test is inadequate for carrying on forecast evaluation in an environment characterized by instability. As argued by Stock and Watson (2003), forecasts based on individual indicators are unstable. Finding an indicator that predicts well in one period is no guarantee that it will predict well in later periods. We protect ourselves to such instability by looking at the best weighted forecast combination after a cut off.

It is possible to test if predictive accuracy changes over time. The predictive accuracy of a model relative to a competitor forecaster appears very much connected to some specific period of time, after which the relative accuracy reverses. Giacomini and Rossi (2010) formally test this idea with the fluctuation test. This test examines the fluctuations in relative predictive performance of forecasting methods over time by comparing the MSFE provided by two different models computed over rolling windows. Giacomini and Rossi (2010) derive the critical values (Table I, page 601) for testing the null hypothesis that the local relative MSFE equals zero at each point in time.[5] One can employ a rolling scheme as a sequence of a window of *P* observations of out- of- sample forecasts loss differences, where θ are the parameters.

$$\{\Delta L_t(\hat{\beta}_{t-h,R,}\hat{\theta}_{t-h,R})\}_{t=R+h}^{T} \quad (5)$$

that depend on the realizations of the variable and on the in-sample estimates for each model re-estimated at each time *t=R+h,...T* over a window of size *R*. The local relative loss for the two models is defined over centred rolling windows of size *m* as:

$$\frac{1}{m}\sum_{j=t-m/2}^{t+\frac{m}{2}-1} \Delta L_j(\hat{\beta}_{j-h,R,}\hat{\theta}_{j-h,R}) \quad (6)$$

The fluctuation test statistic is then defined as:

$$F_{t,m} = \hat{\theta}^2 m^{1/2} \sum_{j=t-m/2}^{t+\frac{m}{2}-1} \Delta L_j(\hat{\beta}_{j-h,R,}\hat{\theta}_{j-h,R}) \quad (7)$$

where the null hypothesis is that

$$H_0 = E\Delta L_t(\hat{\beta}_{t-h,R,}\hat{\theta}_{t-h,R}) = 0 \quad (8)$$

We compute the MSFE differences over rolling windows and testing the null hypothesis that MSFE is equal to zero between the combination models, and for each combination model relative to an

[4] The Theil statistic could be misleading insomuch as the differences among each forecast could not be significant from a statistical point of view.
[5] The DM test is a special case as it averages over the whole sample (Diebold and Mariano, 1995; West, 1996).

AR(1). If the relative MSFE exceeds the critical value in some part of the sample, we reject the null hypothesis and we conclude that there are periods during the sample that one model outperforms the other.

3. Data

We use deficit forecasts from both private and public forecasters both for the current year ($d_{f,t}$) and the year ahead ($d_{f,t+1}$) over the sample period 1993-2012. As in previous studies, the public forecasts come from four institutions: the OECD Economic Outlook, the IMF Forecast, the EC Economic Forecasts, and the Italian Ministry of Economy and Finance (MEF). Generally speaking, these agencies produce projections twice a year (Spring and Autumn) although their exact timing differs somewhat.

In recent years, different datasets have become available that include more forecasts on the deficit. One of those datasets is Consensus Economics Forecasts, Inc. (CEF). This company conducts a survey in several OECD countries among professional economists working for commercial or investment banks, industry, government based agencies, and university departments. Most of the surveyed experts are at domestic institutions that provide forecasts for a single country only; a few work for international financial institutions or research institutes that provide forecasts for several countries simultaneously.

Unlike other surveys, individual forecasts in the CEF should not suffer a bias owing to the release of strategic forecasts, as often happens for official forecast released by governmental agencies (Ottaviani and Sorensen, 2006). CEF data are public, which prevents a participant from reproducing others' forecasts and also limits the possibility of herding (Trueman, 1994). Analysts are bound in their survey answers by their recommendations to their clients, and discrepancies between the survey and their private recommendation would be hard to justify (Keane and Runkle, 1990). In addition, and unlike other surveys, professional economists who participate in the CEF poll not only take a stance on the direction of the expected change of a macroeconomic variable, but also forecast the level of the macroeconomic variable. Evidence shows that CEF forecasts are less biased and more accurate than other surveys.[6]

CEF has gradually expanded the scope and coverage of the survey by including several variables for some OECD countries. We focus on Italy, with data covering the period from January 1993 to

[6] Batchelor (2001) shows that CEF forecasts are less biased and more accurate in terms of mean absolute error and root mean square error than OECD and IMF forecasts. Dovern and Weisser (2011) also find that the participants in the CEF poll provide rational and unbiased inflation and growth forecasts for the G7 countries.

December 2011. Overall, CEF includes 42 forecasters in Italy. Our sample is a small subset of these respondents. Despite the gradual expansion of the dataset, fiscal forecasts have not always received the same attention by forecasters over time. Some forecasters stopped producing projections for the budget balance over time, while others that were initially included, left the sample owing to closure, mergers, or other reasons. Moreover, new forecasters joined the CEF survey only at a later stage. Therefore, we do not consider those forecasters that have participated just a few times in the survey. In particular, any forecaster participating less than 12 consecutive months in the CEF survey is excluded from our sample. This reduces the panel to a selection of five forecasters from Italian banks and research institutes.

The survey enquires respondents every first week of each month about current and year-ahead forecasts for a number of macroeconomic variables. The forecasts are then published early in the second week of the same month.[7] The forecasts require some transformation before they can be used in the empirical analysis.

CEF asks respondents for a forecast of the overall balance in nominal terms.[8] In order to transform this forecast into one of the budget balance as a ratio to GDP, we divide the forecast of the nominal balance (surplus) for year *t+1* in a certain month *m* by the GDP forecast for the same year. As the CEF only provides forecasts of GDP *growth rates*, we compute the year-ahead nominal GDP forecast by applying the CEF growth rate to the latest available estimate for the same year GDP. The latter is taken from IMF WEO (see Appendix A for more details).

We consider in total nine forecasters: five private and four public (OECD, EU, IMF, MEF). Table 1 shows the names of the public forecasters, and the abbreviations of the private forecasting institutions. The 'N' stands for National forecaster. As the private institutions produce forecasts every month, data restrictions come from the month in which the different public forecasters publish their predictions. The OECD publishes its forecasts twice a year in June and December for both d_{ft} and d_{ft+1} in the OECD Economic Outlook. IMF forecasts are published in the IMF's World Economic Outlook in the "World Economic and Financial Surveys" and forecasts of European Commission are released in May ($d_{f,t}$) and in October ($d_{f,t+1}$). The publication of forecasts by the Italian Ministry of Economy and Finance is included in the document "Economic and Financial Planning Document (DPEF)" from 1992 to 1997, and the "Forecast and Planning Report (RPP)" from 1998 to 2011. The forecasts are produced in October, July and June for both d_{ft} and d_{ft+1}.

To match the timing of the forecasts from these sources with the five CEF forecasts that are produced very month, we select only four (May, June, October and December). For the purpose of

[7] Further information on how the survey is conducted is available via the Internet: www.consensuseconomics.com.
[8] For Italy, specialists forecast the general budget balance for the calendar (end of the) year.

combining and comparing forecasts provided by different respondents, we use in our analysis the forecasts published in Spring (May and June) as the December forecast is too close to the end of the year to lead to divergent forecasts of the budget deficit for that year.[9]

Table 1. Timing of release of deficit forecast.

Month	Current Year Forecast d_{ft}	Year ahead Forecast d_{ft+1}
May	EC	EC
	IMF	IMF
	Private forecasters	Private forecasters
June	OECD	OECD
	Private forecasters	Private forecasters
October	MEF	MEF
	EC	EC
	IMF	IMF
	Private forecasters	Private forecasters
December	OECD	OECD
	Private forecasters	Private forecasters

Source: Our elaborations on official databases. *Note*: MEF projections are published as follow: 1992-1995 July, 1996-1997 June, 1998-2012 October. OECD projections are published in December during 1992-2011

In particular, for the analysis we create a database which includes the information from the months of May (or June) for public institutions (EC, OECD, IMF, MEF) and May for the forecasters from the CEF database. In the latter case, there are some missing values in May and we added forecasts in two ways. First, we include the forecasts from the earlier month in the sample. So for instance, if there are missing values in May of a specific year for one private forecast, we will use information from April which is the closest month if present, if not, we use the next closest month etc. These adjustment are computed in order to have a more complete data profile.

We then compare the forecast to the realised deficit to GDP ratio for that period. Realised deficit ratio/ GDP d_t at time t comes from the OECD Economic Outlook.[10]

[9] Results not reported, but available upon request.
[10] Note that the last four years could still be object of adjustment.

Figure 1 shows a graph of the different forecasts over time.

Figure 1. Realised deficit ratio/ GDP, and current- year forecasts, sample 1993-2012.

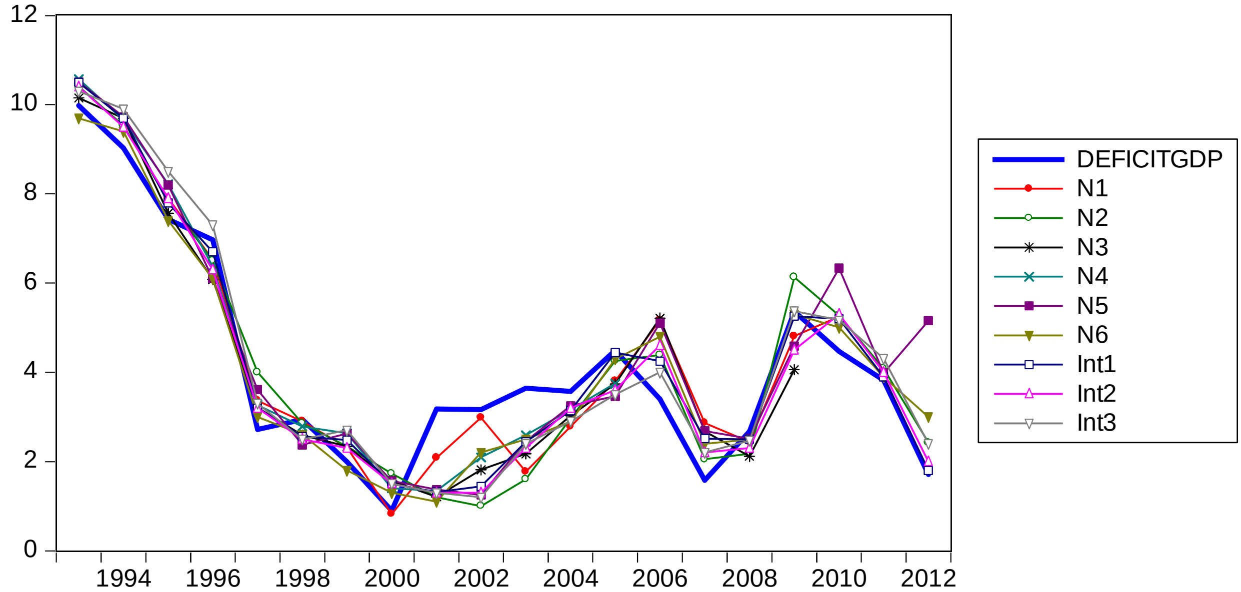
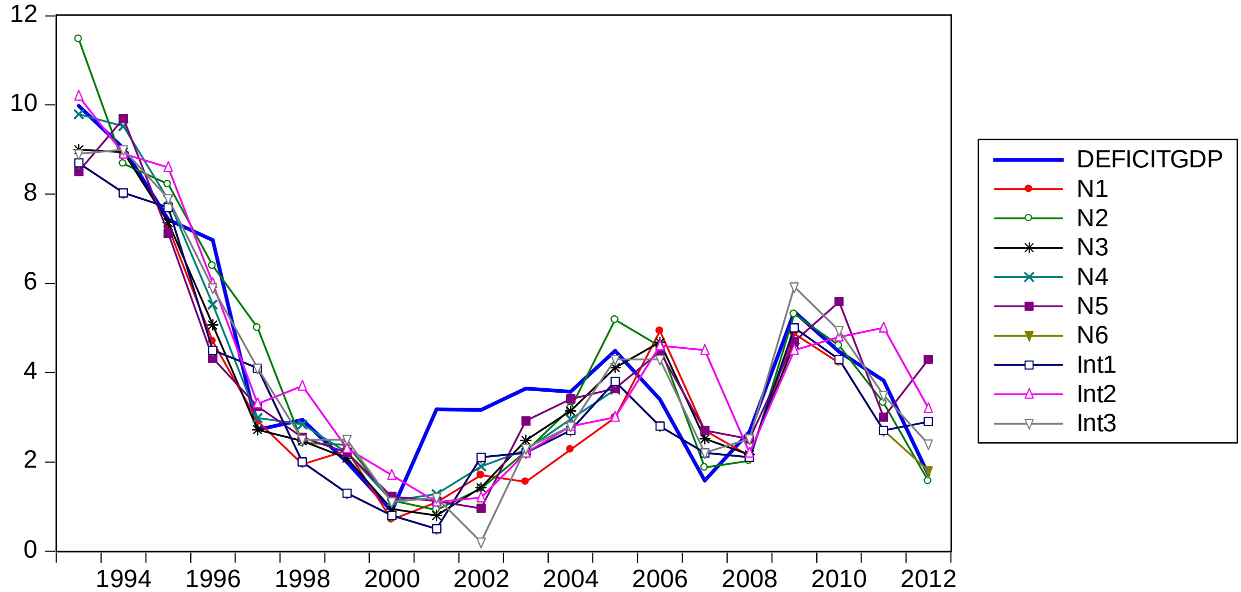


Figure 2. Realised deficit ratio/ GDP, and year-ahead forecasts, sample 1993-2012.



In Figure 2, we show the deficit forecasts one year ahead provided by the same respondents. In both Figures, the forecasts are close to the actual value of the deficit ratio from 1992 to 2001. This impression may be imposed by the sharp decline in deficits in the run up to the start of EMU. Afterwards, the forecast tends to be less accurate for each forecaster. In Figure 1, the values are very different with much variability of forecasts between them.

Figure 3 Realised deficit ratio/ GDP with the nine combined forecasts for current year

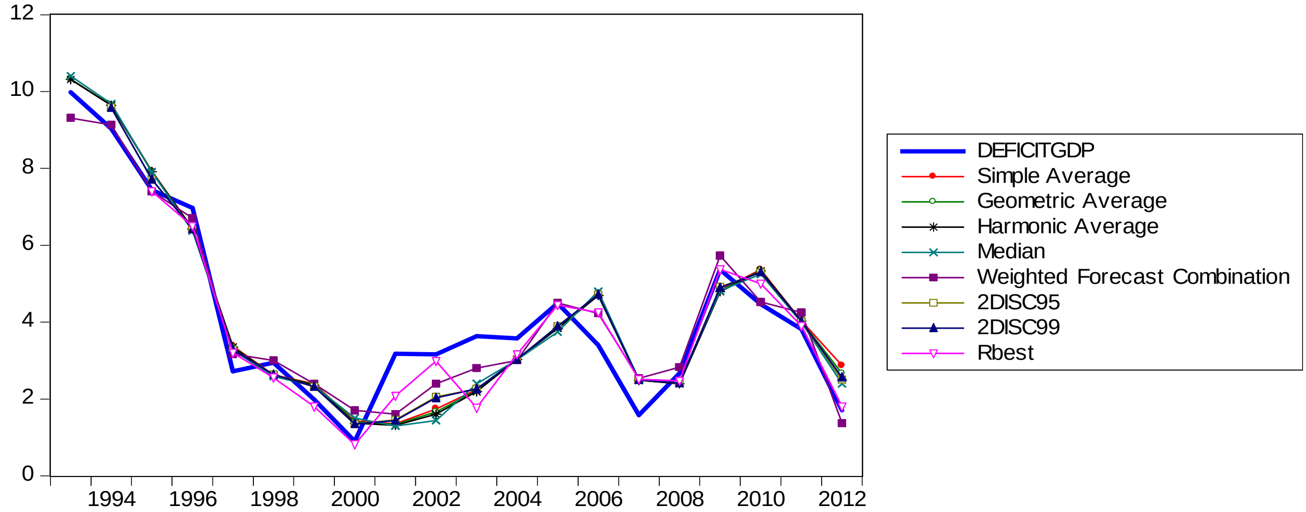


Figure 4 Realised deficit ratio/ GDP with the nine combined forecasts for year ahead

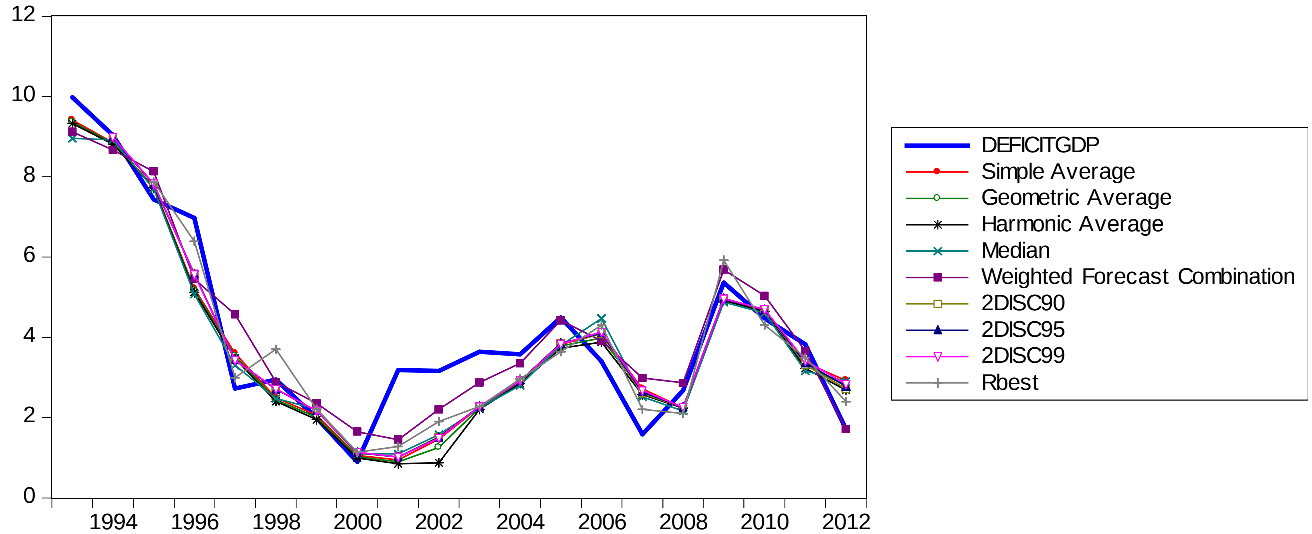


Figure 3 displays the actual deficit together with the nine combined forecasts that we compute for the current year forecast, and Figure 4 does the same for the year ahead forecast. To the aim of facilitate the reading we simplify the nominations of the forecasts combined model as following: Weighed Forecast Combination, 2disc90, 2disc95, 2disc99 e Rbest.

All combined forecasts track closely the deficit over the first part of the sample (up to 2001). Afterwards, there is a tendency to deviate from the deficit for a couple of years. Comparison of Figures 1-2 to 3-4 shows that combination forecasts are less variable than the single forecasts. Indeed, the stability of the forecasts provided by combinations permits less variability in the forecasts compared to the forecasts of individual methods. A simple view of Figures 3 and 4 suggest that the weighted forecast combination as well as the Rbest combination are closest to the actual data in 2001-2002 when all the agencies tended to make large forecasting mistakes.[11]

4. Combining forecasts and predictive accuracy

4.1 Accuracy of forecast errors

We see from Table 2 that all forecasters do much better than a simple RW model would suggest. For the current-year forecast the weighted forecast combination and the Rbest are more accurate than any single forecaster, or any other combination of forecasts. By contrast, in Table 3, the combination models do not generally outperform the single forecast with the exception of weighted forecast combination that is more accurate in terms of all the test of accuracy results.

Table 2. Accuracy test of single and combination forecasts, for current year forecasts of deficit ratio.

Forecast accuracy	RMSE	MAD	MSE
National 1	3.73	0.72	1.93
National 2	4.10	0.74	2.02
National 3	4.22	0.84	2.05
National 4	3.45	0.74	1.86
National 5	4.66	0.97	2.16
National 6	3.37	0.60	1.84
OECD	3.48	0.66	1.86
EU	3.88	0.78	1.97
IMF	3.89	0.76	1.97
Simple average	3.73	0.72	1.93
Harmonic average	3.80	0.72	1.95
Geometric average	3.77	0.72	1.94
median	3.82	0.71	1.95
Weighted Forecast combination	2.78	0.48	1.67
2DISCMSFE95	3.51	0.69	1.87
2DISCMSFE99	3.54	0.70	1.88
RBEST	2.73	0.44	1.65
AR	6.99	1.20	2.64

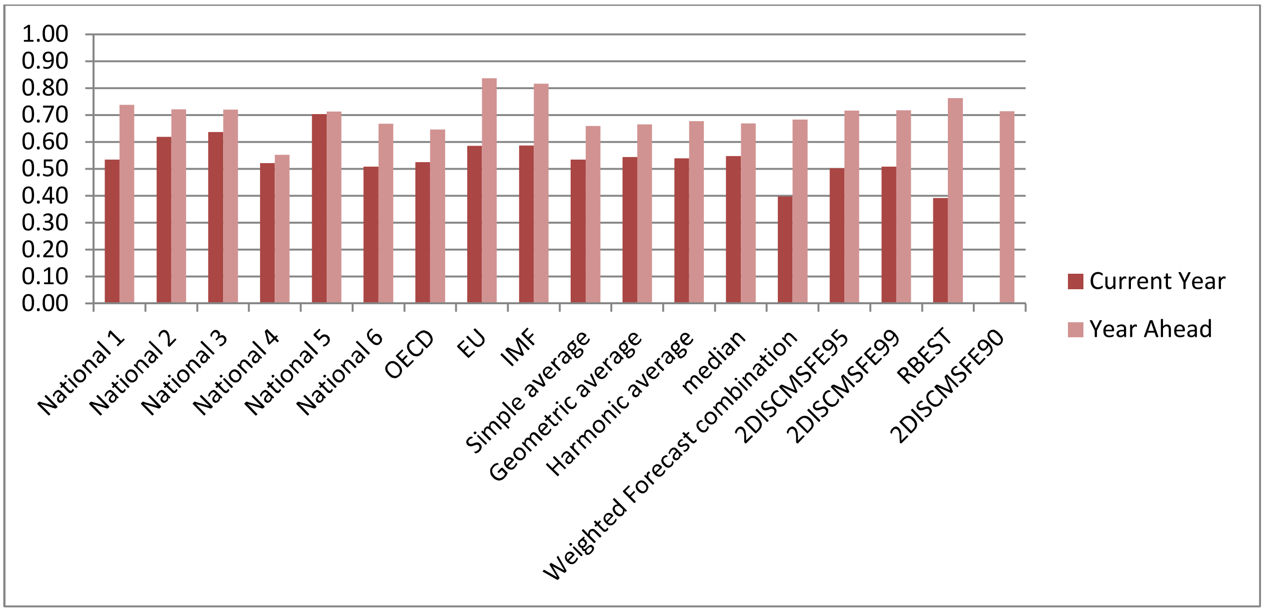
[11] For the coefficient of the regression of weighted forecast combination (WFC) of forecast current year and year ahead see the appendix A- additional tables.

Table 3. Accuracy test of single and combination forecasts, for year ahead forecasts of deficit ratio.

Forecast accuracy	**RMSE**	**MAD**	**MSE**
National 1	7.15	1.39	2.67
National 2	6.99	1.68	2.64
National 3	6.98	1.34	2.64
National 4	5.36	1.25	2.31
National 5	6.92	1.22	2.63
National 6	6.47	1.17	2.54
OECD	6.27	1.09	2.50
EU	8.11	1.64	2.85
IMF	7.91	1.50	2.81
Simple average	6.60	1.24	2.57
Geometric average	6.65	1.25	2.58
Harmonic average	6.78	1.25	2.60
median	6.69	1.25	2.59
Weighted Forecast combination	6.83	1.38	2.61
2DISCMSFE95	7.16	1.39	2.68
2DISCMSFE99	7.18	1.39	2.68
RBEST	7.62	1.52	2.76
2DISCMSFE90	7.14	1.39	2.67
AR	10.00	1.84	3.16

The Theil test (1958) shows us in figure 5 and 6 more formally the improvement in performance relative to a RW model. Any single forecast, or a combination of them, does much better. Unsurprisingly, the accuracy is also generally better for individual and combination models in the current year as compared to a forecast in the year ahead. While a weighted forecast combination or the Rbest procedure does improve considerably the accuracy, this is not generally the case for the year-ahead forecast. In fact, respondent N4 improves by 10% over the best combined forecast (the simple average in this case).

Figure 5. Theil test of single and combination forecasts, for current year and year ahed forecasts of deficit ratio.

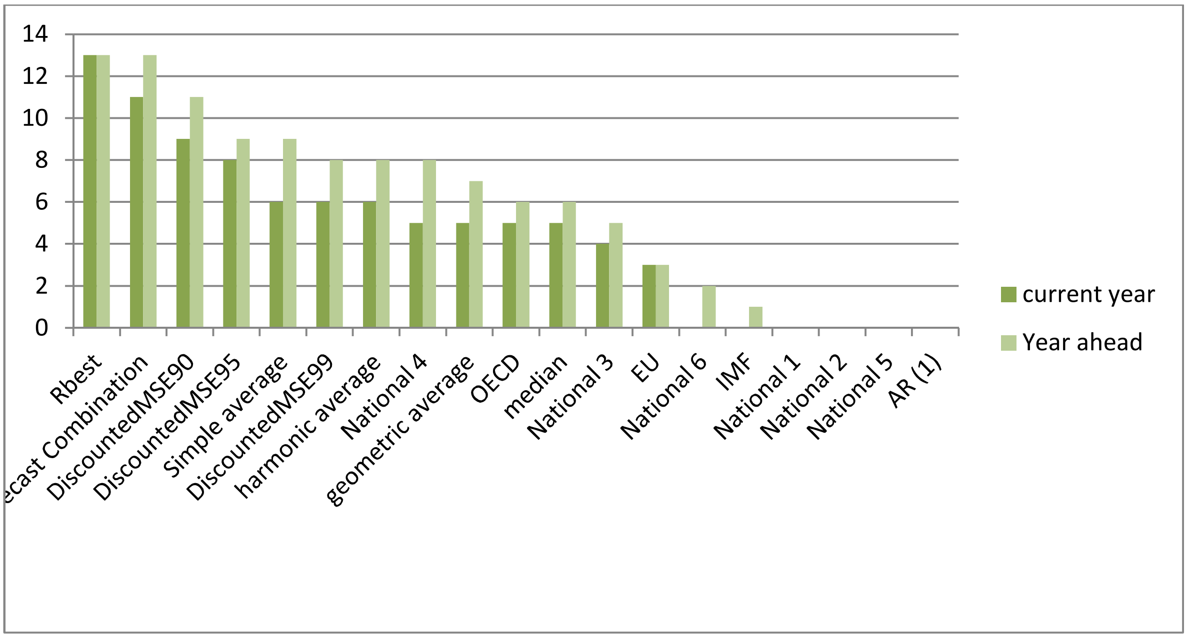


4.2 Results of predictive test

Table 4 summarises the results of the DM test and shows in the first panel the comparison of all forecasts in the current year, and in the second panel for the year ahead. In each cell, we put the winner of the contest between the combined and the single forecast. Panel 1 of Table 4 confirms the findings in Table 2 and 3 that the combination outperforms the individual models often, although some individual forecasters sometimes do better. To summarize these results, we synthesize the number of time each model outperforms the others. of the performance of each as shown in the Figure 7 for current year and Figure 8 for year ahead. As we can see in the both figures, the Rbest and Weighted forecast combination[12] do always better than any other forecasters.

[12] The results of Figure 5 and 6 show that Rbest and Weighted forecast combination are the best performer respectively, with 13 win for current year and year ahead and 11 win for current year and 13 for year ahead.

Figure 6. The number of best performance of each model (single and combined) – Forecast Current Year and Year Ahead.



The results in the second panel show a similar picture: any combination of forecast is often better than a single forecaster, although we cannot exclude that some forecasters does always better, like forecaster N4.

Another important result of Table 4 is that - consistent with existing literature (Artis and Marcellino 2001, Ozcan 2011) – the naïve AR (1) model is always performing worse than any combination.

Among the combination forecasts, the most accurate one is the weighted average forecast, both in the current-year and year-ahead data.

Table 4. Diebold-Mariano test comparison for current year and year ahead

Current Year	National 1	National 2	National 3	National 4	National 5	National 6	OECD	EU	IMF	AR(1)	WFC	2disc95	2disc99	rbest
Results	-0.06	-0.1	-0.11	0.007	-0.31	0.13	0.14	0.01	-0.002	-0,12	0,12			
Simple average	SIMPL AVERAGE	SIMPL AVERAGE	SIMPL AVERAGE	N4	SIMPL AVERAGE	N6	OECD	EU	SIMPL AVERAGE	SIMPL AVERAGE	WFC			
pvalue	0.52	0.06	0.19	0.89	0.01	0	0.07	0.71	0.97	0,26	0,26			
Results	0.08	-0.09	-0.1	0.01	-0.31	0.12	0.13	-0.01	-0.008	-0,1	0,1			
geometric average	N1	GEOM MEAN	GEOM MEAN	N4	GEOM MEAN	N6	OECD	GEOM MEAN	GEOM MEAN	GEOM MEAN	WFC			
pvalue	0.43	0.07	0.21	0.81	0.01	0	0.05	0.74	0.91	0,3	0,3			
Results	0.14	-0.09	-0.1	0.02	-0.32	0.14	0.13	0,007	-0.01	-0,09	0,09			
Harmonic average	N1	HARMON MEAN	HARMON MEAN	N4	HARMON MEAN	N6	OECD	EU	HARMON MEAN	HARMON MEAN	WFC			
pvalue	0.3	0.06	0.2	0.78	0.02	0	0.05	0.79	0.85	0,34	0,34			
Results	-0.02	-0.42	-0.29	-0.015	-0.86	-0.16	-0,21	-0.06	-0.02	-0,09				
WFC	WFC	WFC	WFC	WFC	WFC	WFC	WFC	WFC	WFC	WFC	-			
pvalue	0,03	0,05	0,08	0,5	0.04	0,12	0,31	0,75	0,9	0,51				
Results	0.1	-0.08	-0.08	0.04	-0.31	0.13	0.14	0.01	-0.005	-0,08	0,08			
Median	N1	MED	MED	N4	MED	N6	OECD	EU	MED	MED	WFC			
pvalue	0.38	0.1	0.34	0.53	0.03	0	0.01	0.44	0.94	0,27	0,27			
Results	0.2	-0.17	-0.19	-0.05	-0.39	0.064	0.08	-0.05	-0.08	-0.55	0.21		-0.007	0.25
2disc95	N1	2disc95	2disc95	2disc95	2disc95	N6	OECD	2disc95	2disc95	2disc95	WFC	-	2disc95	rbest
pvalue	0.8	0.01	0.005	0.35	0.0036	0.03	0.3	0.23	0.4	0.01	0		0.008	0
Results	0.02	-0.17	-0.19	-0.04	-0.38	0.07	0.09	-0.04	-0.07	-0.54	-0.22			0.2656
2disc99	N1	2disc99	2disc99	2disc99	2disc99	2disc99	OECD	2disc99	2disc99	2disc99	2disc99	-	-	rbest
pvalue	0.78	0.01	0.005	0.4	0.0036	0.01	0.27	0.32	0.46	0.02	0			0
Results	-0.21	-0.42	-0.43	-0.28	-0.66	-0.19	-0.16	-0.31	-0.32	-0.82	-0.5			
rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest			
pvalue	0.02	0	0	0	0.0001	0.0006	0.05	0	0.0008	0.0008	0.25			

Year Ahead	National 1	National 2	National 3	National 4	National 5	National 6	OECD	EU	IMF	AR(1)	WFC	2disc95	2disc99	rbest
Results	-0.27	-0.14	-0.01	0.07	-0.24	-0.01	-0.21	-0.29	-0.04	-0,12	0,12			
Simple average	SIMPL AVERAGE	SIMPL AVERAGE	SIMPL AVERAGE	N4	SIMPL AVERAGE	SIMPL AVERAGE	SIMPL AVERAGE	SIMPL AVERAGE	SIMPL AVERAGE	SIMPL AVERAGE	WFC			
pvalue	0.01	0.12	0.84	0.21	0.05	0.85	0.12	0.01	0.6	0,26	0,26			
Results	-0.26	-0.11	0.005	0.1	-0.22	0.003	-0.2	-0.27	-0.02	-0,1	0,1			
geometric average	GEOM MEAN	GEOM MEAN	N3	N4	GEOM MEAN	N6	GEOM MEAN	GEOM MEAN	GEOM MEAN	GEOM MEAN	WFC			
pvalue	0.01	0.27	0.92	0,07	0.08	0.95	0.15	0.02	0.71	0,3	0,3			
Results	-0.19	-0.08	0.03	0.14	-0.2	-0.12	-0.17	-0.26	-0.01	-0,09	0,09			
Harmonic average	HARMON MEAN	HARMON MEAN	N3	N4	HARMON MEAN	HARMON MEAN	HARMON MEAN	HARMON MEAN	HARMON MEAN	HARMON MEAN	WFC			
pvalue	0.01	0.45	0.55	0.02	0.14	0.44	0.22	0.03	0.88	0,34	0,34			
Results	-2.03	-3.86	-3.7	3.93	-5.44	-6.18	-5.66	-5.46	-5.93	-0,09				
WFC	WFC	WFC	WFC	N4	WFC	WFC	WFC	WFC	WFC	WFC	-			
pvalue	0,05	0.03	0	0.08	0.01	0	0.02	0.02	0.02	0,51				
Results	-0.26	-0.12	0.03	0.08	-0.19	0.03	-0.16	-0.24	0.006	-0,08	0,08			
Median	MED	MED	N3	N4	MED	N6	MED	MED	IMF	MED	WFC			
pvalue	0	0.26	0.14	0.04	0.1	0.48	0.31	0.01	0.95	0,27	0,27			
Results	-0.25	-0.11	0.0025	0.0375	-0.24	-0.16	-0.22	-0.36	-0.05	-0.51	0.07		-0.007	0.08
2disc95	2disc95	2disc95	N3	N4	2disc95	2disc95	2disc95	2disc95	2disc95	2disc95	WFC	-	2disc95	rbest
pvalue	0.0008	0.17	0.95	0.51	0.07	0.27	0.1	0.0001	0.47	0.03	0.58		0.0001	0.05
Results	-0.24	-0.11	0.01	0.04	-0.23	-0.15	-0.21	-0.35	-0.04	-0.5	0.08			0.09
2disc99	2disc99	2disc99	N3	N4	2disc99	2disc99	2disc99	2disc99	2disc99	2disc99	WFC	-	-	rbest
pvalue	0.001	0.17	0.7789	0.45	0.08	0.3	0.12	0.0006	0.54	0.04	0.55			0.03
Results	-0.32	-0.2	-0.06	-0.02	-0.31	-0.21	-0.27	-0.46	-0.15	-0.57	0.01			
rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest	rbest	WFC			
pvalue	0	0.08	0.2	0.72	0.05	0.19	0.07	0	0.16	0.03	0.95			
Results	-0.26	-0.11	-0.004	0.035	-0.25	-0.17	-0.23	-0.37	-0.06	-0.51	0.069			
2disc90	2disc90	2disc90	2disc90	N4	2disc90	2disc90	2disc90	2disc90	2disc90	2disc90	WFC			
pvalue	0.0008	0.18	0.92	0.55	0.06	0.24	0.08	0.0006	0.4	0.03	0.61			

Note: for the coefficient of the regression (1) of the Weighted Forecast Combination (WFC) see Appendix A.

4.3 Stable prediction performance

We observed in Figures 1 and 2 that forecasting performance changed over time. Up to 2001, most forecasters performed quite well, and projections were mostly aligned with actual budget outcomes. Afterwards, performance has diverged. Even expert forecasters are unable to anticipate all economic and political changes. This seems to suggest that the models have to be adaptive, but also in this way represent the best state of knowledge. A combination of forecasts aggregates information and reduces uncertainty by eliminating judgment errors on structural changes. The outcome is still based on the global performance of forecasters, however, and not on the change in performance over time of different competing forecasts.

We apply the fluctuation test and the one-time reversal test – proposed by Giacomini and Rossi (2011) – to analyse the evolution over time of the models' relative performance over the sample. We do this by comparing the nine simple forecasts to the eight combined forecasts and the simple AR model.

Table 6: Fluctuation Test

Forecast	CWp-values
gBOA	7.1586e-005
gCER	0.00074427
gENI	0.00045058
gFIA	0.0018211
gPRO	0.00011781
gMEF	0.00019944
gOEC	0.00018055
gEUC	0.00015129
gIMF	0.0001085
gWFC	5.4412e-005
gMED	0.00016378
gMEN	0.00014401
gGEO	0.00018236
gARM	0.00020812
gW95	0.00012304
gW99	0.00012082
gRBE	2.583e-005
gW90	0.00012534
Forecast	GWp-values
gBOA	0.21041
gCER	0.18962
gENI	0.12256
gFIA	0.13077
gPRO	0.25056
gMEF	0.1602
gOEC	0.17323
gEUC	0.23474

gIMF	0.12918
gWFC	**0.035254**
gMED	0.088731
gMEN	0.10018
gGEO	**0.099479**
gARM	0.12191
gW95	0.064389
gW99	0.066099
gRBE	**0.042219**
gW90	0.062587

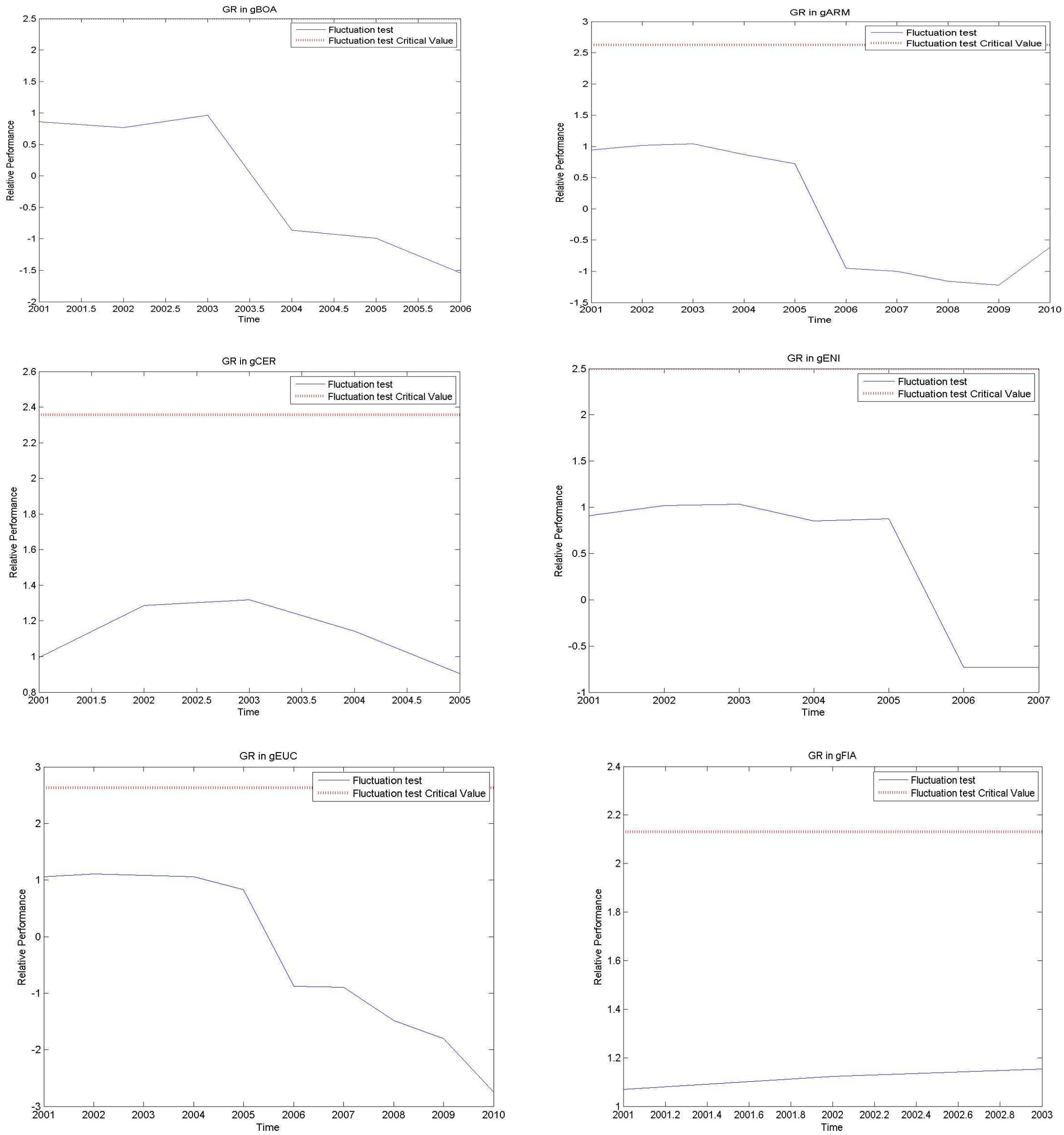
With the CW test, we would reject that the forecast has a better performance than the RW. This rejection is very strong. With the Giacomini Rossi test instead, we see that over the full sample, only the WFC and Rbest are the forecast (combinations) that outperform the RW at 5% significance level.

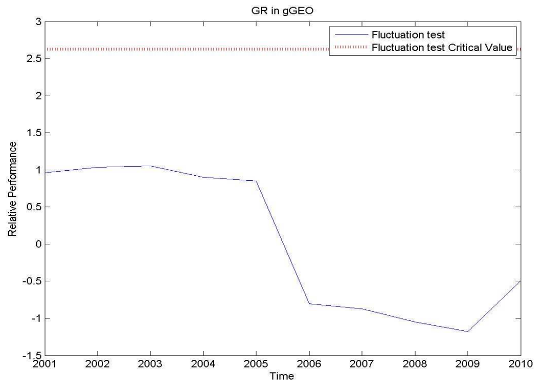
Given that there may be instability over time, we also run the Fluctuation Test. The following graphs plot the test statistic, together with the critical value. Positive values indicate that the forecaster performs better than a random walk, and a negative number that the random walk would be preferred. All forecasters perform considerably better than the random walk from 2001 up to 2006. From that year onwards, performance drops and no forecaster would have done better than a random walk. This is true both for the private institutions as the public forecasters. The result shows that the deficit forecast, whether made by private or public forecasters, is not robust to large shifts. In this case, the Financial Crisis led to a strong revision of the deficit numbers and although it has often been argued that the successive Italian governments managed to keep the deficit under control (REF&), forecasters had problems in evaluating the deficit.

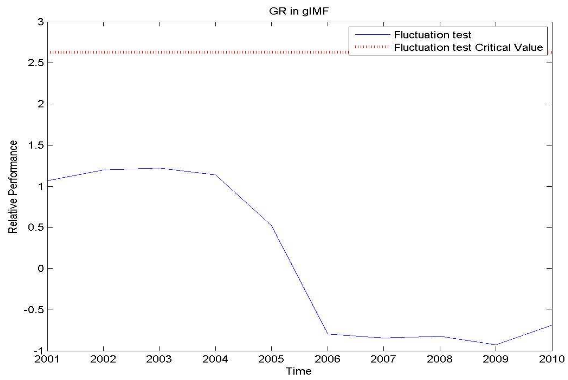
As all forecasters make mistakes around the same break, the combined forecasts are suffering from the same problem. The only exception here is the WFC. The reason for its consistently good performance is that the combination puts less weight on the less stable predictions in the latter period. As the Italian government did not change behaviour so much, the model predicts the deficit relatively well.

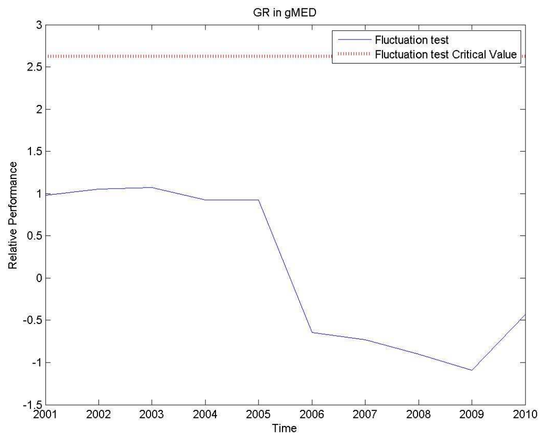
Consistent with the evidence of the table, none of the forecasters is doing significantly better than a RW model at any point in time. Even the WFC, which was found to outperform the RW over the full sample, is not structurally stable enough to be better than a RW.

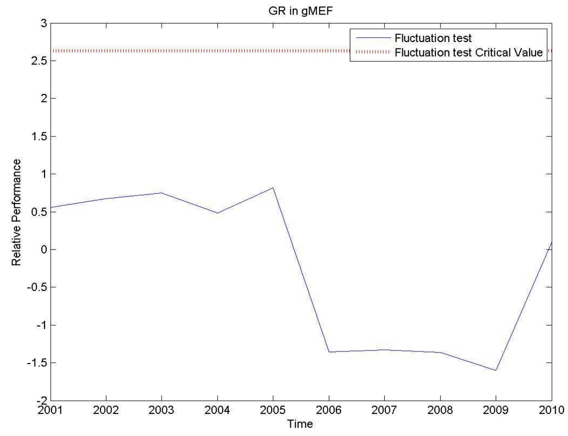
Figures 7: The fluctuation tests for individual and combination

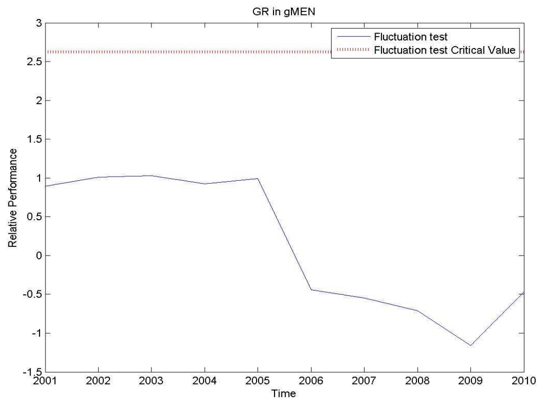


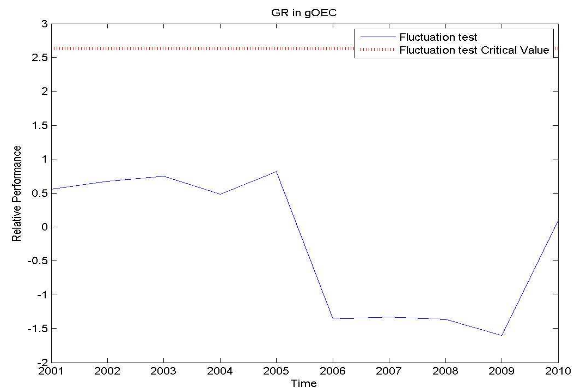

GR in gGEO
Fluctuation test
Fluctuation test Critical Value
Relative Performance
Time

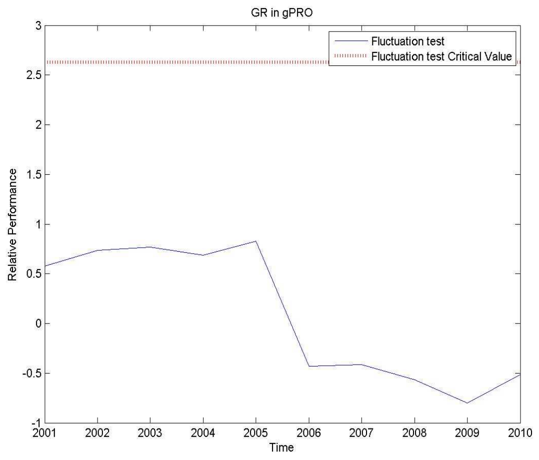

GR in gIMF
Fluctuation test
Fluctuation test Critical Value
Relative Performance
Time

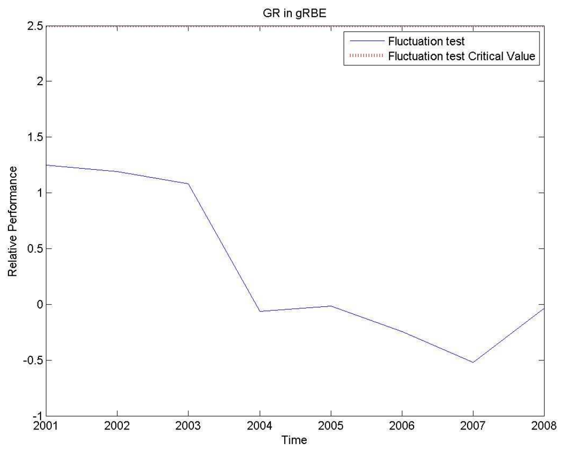

GR in gMED
Fluctuation test
Fluctuation test Critical Value
Relative Performance
Time

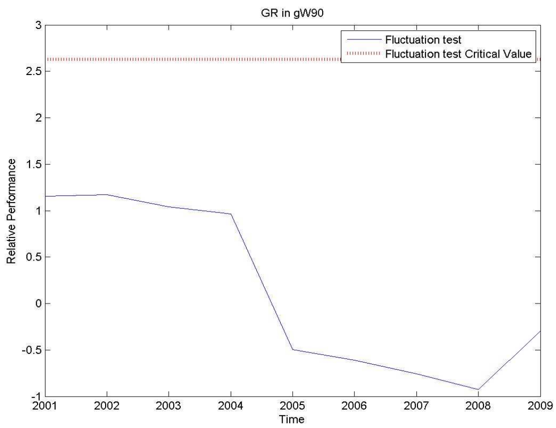

GR in gMEF
Fluctuation test
Fluctuation test Critical Value
Relative Performance
Time

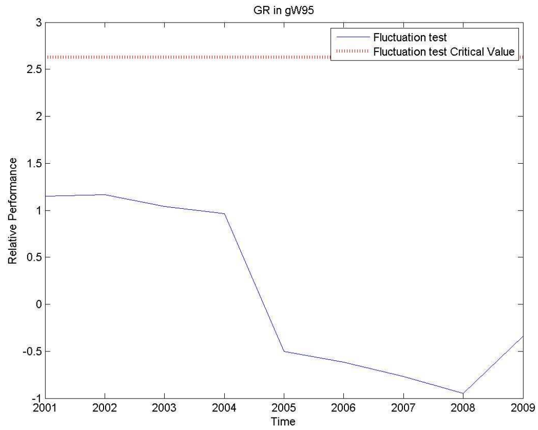

GR in gMEN
Fluctuation test
Fluctuation test Critical Value
Relative Performance
Time

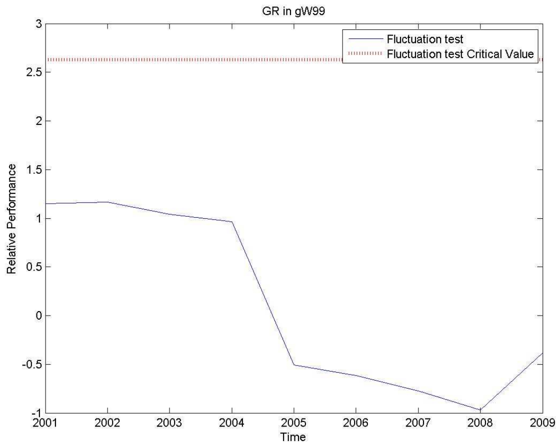

GR in gOEC
Fluctuation test
Fluctuation test Critical Value
Relative Performance
Time

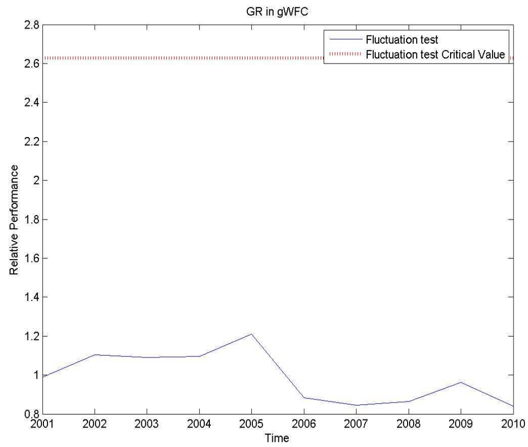

GR in gPRO
Fluctuation test
Fluctuation test Critical Value
Relative Performance
Time


GR in gRBE
Fluctuation test
Fluctuation test Critical Value
Relative Performance
Time









6. Robustness check

6.1. Alternative dataset

We have so far used a baseline dataset but repeat the same analysis above also on other two databases for the current-year and year-ahead. So far, we used in our analysis only the forecasts published in Spring (May and June). As private forecasters produce a monthly forecast, and the public forecasters also a forecast in Autumn it is possible to evaluate these other forecasts too. The closer the December forecast, however, the more information goes into the forecast which leads to a convergence of forecasts of the budget deficit for that year. We try to look as a robustness check at two additional databases. The first one include all months May and June. The second one includes the months of October and December.

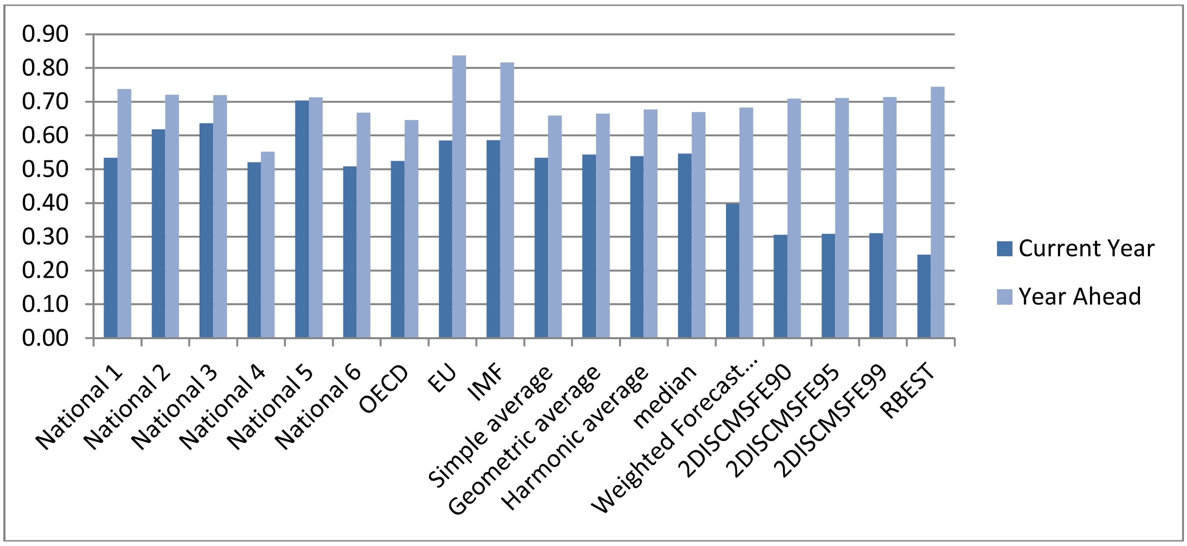
Table 7. Database without adjustment of fiscal forecast data 1992-2012.

Categories without adjustment	Current Year Forecast (d_{ft})	Year ahead Forecast (d_{ft+1})
May/June	EU	EU
	IMF	IMF
	OECD	OECD
	Private forecasters	Private forecasters
October/December	MEF	MEF
	EU	EU
	OECD	OECD
	Private forecasters	Private forecasters

Source: Our elaborations on official databases.
Notes: EU, IMF and private agencies provide forecasts in May, the OECD in June, October is the month in which Mef (projections are published as follow: 1992-1995 July, 1996-1997 June, 1998-2012 October) and EU provide forecasts and OECD in December. It isn't necessary to make an adjustments database for the Database C because there is already enough information provided by the forecasters at this time of the year therefore the data profile is more complete than in the Database B.

Results show in the figure 8 that summarise the Index Theil for the database B are consistent with the previous results.

Figure 8. Theil test of single and combination forecasts, for current year and year ahead forecasts of deficit ratio. (Database B)



The results of the accuracy tests of these two databases B and C are consistent with the previous results. The accuracy of combination models improve. Unsurprisingly, during this last month of the year, forecasts in general are more accurate than in the rest of the year. For this reason the combination of these forecasts provide more accurate results. Also, for both current-year and year-ahead forecasts, every combination model outperforms the naive model.

If we compare all combination models between them, we can conclude that also using the databases B and C the results are consistent.

7. Conclusions

Despite the growing importance of fiscal projections in the short-term to inform policy-makers, control fiscal monitoring and manage expectations, practitioners seem to require a lot of judgment in making better fiscal projections. We show that exploiting the information from many different forecasters can still lead to substantial gains in predictive accuracy. Datasets that have become available in recent years, such as CEF, allow combining forecasts in several ways. Applying eight different combination techniques to the current year and year ahead forecasts of the Italian budget deficit over the period 1993 to 2011 results in substantial gains in forecast accuracy. In particular, the combination models are more accurate than individual models for 65% in the year ahead and 54% in current year and, in any case, each combination model is better than an AR(1). Given the changes in forecast performance over time, no single model is to be preferred at any time, and a combination guards against the weighted forecast combination model that provides the best performance compared by the other methods. But even then, no single combination can beat National 4. Still, combining forecasts can result in substantial gains in predictive accuracy when set off against current standards.

References

Artis, M. and Marcellino M., (2001), "Fiscal Forecasting: The track record of the IMF, OECD and EC", Econometrics Journal, Volume 0, pp.1-23;

Batchelor, R., (2001), 'How useful are the forecasts of intergovernmental agencies? The IMF and OECD versus the consensus', Applied Economics, 33, 225 – 35.

Bates, J. M., and C. W. J. Granger, (1969), "The combination of forecasts", Operational Research Quarterly, 20, 451-468.

Beetsma, R. and M. Giuliodori (2010). Fiscal Adjustment to Cyclical Developments in the OECD: An Empirical Analysis Based on Real-Time Data, Oxford Economic Papers 62, 3, 419-441.

Clemen, Robert T. (1989), "Combining forecasts: A review and annotated bibliography," International Journal of Forecasting, 5, 559-583.

Clemen, R.T., and Winkler, R.L. (1986), "Combining economic forecasts", Journal of Business and Economic Statistics,4, 39-46.

Cooper, J.P., and Nelson, CR. (1975). "The ex-ante prediction performance of the St. Louis and FRB-MIT-PENN econometric models, and some results on composite predictions", Journal of Money, Credit and Banking.

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, 452 –65.

Diebold, F. X. and R. S. Mariano, (1995), "Comparing predictive accuracy", Journal of Business and Economic Statistics, 13, 253-263.

Fildes, R. and Stekler, H. (2002) "The state of macroeconomic forecasting", Journal of Macroeconomics, 24, 435- 68.

Giacomini, R. *and* Rossi, B. *(*2010*). '*Forecast comparisons in unstable environments*',* Journal of Applied Econometrics*, Vol.* 25*,* 595–620*.*

Granger, C.W.J., and Newbold, P. (1973) “Some comments on the evaluation of forecasts”, Applied Economics, 5, 35-47.

Keane, M. P., and Runkle, D. E. (1990), “Testing the Rationality of Price Forecasts: New Evidence From Panel Data,”American Economic Review , 80, 714–735.

Leal, T., Pérez, J. J., Tujula, M., & Vidal, J.-P. (2008). Fiscal forecasting: Lessons from the literature and challenges. Fiscal Studies ,29, 347–386

Lupoletti, V.M., and Webb, R.H. (1986), “Defining and improving the accuracy of acroeconomic forecasts: Contributions from a VAR model”, Journal of Business 5,9 , 263-285.

Makridakis, S., (1989). "Why Combining Works?," International Journal of Forecasting, 5, 4, 601-603.
Winkler, R.L., and Makridakis, S. (1983) “The combination of forecasts”, Journal of the Royal Statistical Socie{v,Serie.q A, 146, 150-157.

Makridakis, S., and Winkler, R.L., (1983). "Averages of Forecasts: Some Empirical Results," Management Science), 987-996.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. (1982), “The accuracy of extrapolation (time series) methods: Results of a forecasting competition”, Journal of Forecasting, 1, 111-153.

Marcellino M. (2002), “Forecast pooling for short time series of macroeconomic variables”, CEPR WP 3313.

Marcellino M. (2004), “Forecast Pooling for European Macroeconomic Variables”. Oxford Bulletin of Economics and Statistics, 66(1): 91-112.

Nelson, C.R. (1972), “The prediction performance of the F.R.B.-M.I.T.-PENN model of the U.S. economy”, American Economic Review, 62, 902-917.

Ottaviani M, Sorensen PN. (2006), “The strategy of professional forecasting”, Journal of Financial Economics ,81:441–466.

Ozcan N. (2011), “A first step to develop a forecasting model for U.S. government budget deficit: A regression-based method with quarterly economic indicators”, University of North Caroline at chapel hill, department of economics.

Stock JH, Watson MW. (1999a), “Forecasting inflation”, Journal of Monetary Economics”, 44, 293–335.

Stock, JH, Watson, MW. (1999b), “A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In Cointegration, Causality and Forecasting”, A Festschrift for Clive W.J. Granger, Engle R, White H (eds). Oxford University Press: Oxford; 1–44.

Stock J.H., Watson M.W. (2003), “Forecasting output and inflation: the role of asset prices”, Journal of Economic Perspectives, 41, 788–829.

Stock J. and M.Watson (2004). “Combination Forecasts of Output Growth in a Seven-Country Data Set". Journal of Forecasting, 23: 405-430.

Theil, H., (1958), Economic Forecasts and Policy, (Amsterdam: North Holland), Ch. 2.

Timmermann, A. (2006), “Forecast Combinations”. In Handbook of Economic Forecasting, Elliot, G., Granger, C. and A. Timmermann (eds.). North-Holland Vol I; 136-196.

Trueman, B.(1994), “Analyst Forecasts and Herding Behaviour”, Review of Financial Studies, 7(1): 97–124.

Zamowitz, V. (1967). “An appraisal of short-term económic forecasts” (National Bureau of Economic Research, New York). F, A, E. On pages 123-125.

Appendices

Appendix A. Additional tables

Table 1.The coefficient regressions of the "Weighted forecast combination" for current year and year ahead)[13] Annex.1

A	Forecast Current Year (1)	Forecast Year Ahead (1)
C	0.64	0.76
National 5	-0.52	0.01
National 6	0.69	1.54
OECD	0.10	0.06
EU	0.13	-0.93
IMF	0.47	0.25

Note: the R^2 of Forecast Current Year (1) is 0.93 Forecast Year Ahead (1) 0.96

Appendix B. Calculation of the forecasted budget balance (as a ratio of GDP)

CEF provides forecasts for the total deficit only in nominal values (local currency). Hence, we follow Heppke-Falk and Hüfner (2004) and Poplawski-Ribeiro and Rülke (2011) to construct a forecast measure of deficit ratio to GDP (percentage of GDP). For that, we cannot simply scale the nominal value deficit forecast by the GDP forecast, since the CEF surveys for growth rates only, and not for the GDP in nominal value.

We construct a measure of the expected nominal year-ahead GDP forecast of forecaster *i* at month *m* and year *t* as follows. In the first step, we take a real-time measure of real GDP in levels for a particular year *t*. We use the real-time forecast of the same-year real GDP (in levels) coming from the most recent IMF World Economic Outlook (WEO) vintage available at any particular month *m* of year *t*. The IMF WEOs are published either in April or October, hence from May to October we use the April issue, and the October issue in the other months.

The second step is to compute the year-ahead GDP forecast in nominal value. We multiply the real-time (WEO) measure of same-year real GDP (in levels), $E_{WEO,t}\left[y_t\right]$, by the year-ahead market (Consensus) forecasts for GDP *growth*, $E_{i,t,m}\left[\Delta y_{t+1}\right]$, and inflation, $E_{i,t,m}\left[\pi_{t+1}\right]$, for each forecaster *i* at a particular month *m* of year *t*. The expected year-ahead nominal GDP value for each country is then

[13] To compute the combination model Weighted forecast combination throght the regression (1) we use the forecasts available from 1992-2012 without any missing value. So we use 5 forecasters: 2 National and 3 international from database A) B) and C)

$$E_{i,t,m}\left[y_{t+1}\right] = E_{WEO,t}\left[y_t\right] \times \left(1 + E_{i,t,m}\left[\Delta y_{t+1}\right] + E_{i,t,m}\left[\pi_{t+1}\right]\right). \tag{A.1}$$

The year-ahead expected budget balance for each country is then:

$$E_{i,t,m}\left[b_{t+1}\right] = \frac{E_{i,t,m}\left[b_{t+1}^{nom}\right]}{E_{i,t,m}\left[y_{t+1}\right]}, \tag{A.2}$$

where $E_{i,t,m}\left[b_{t+1}^{nom}\right]$ is the (CEF) forecast of the nominal budget balance by forecaster *i* in month *m* of year *t* for one year-ahead *t+1*.